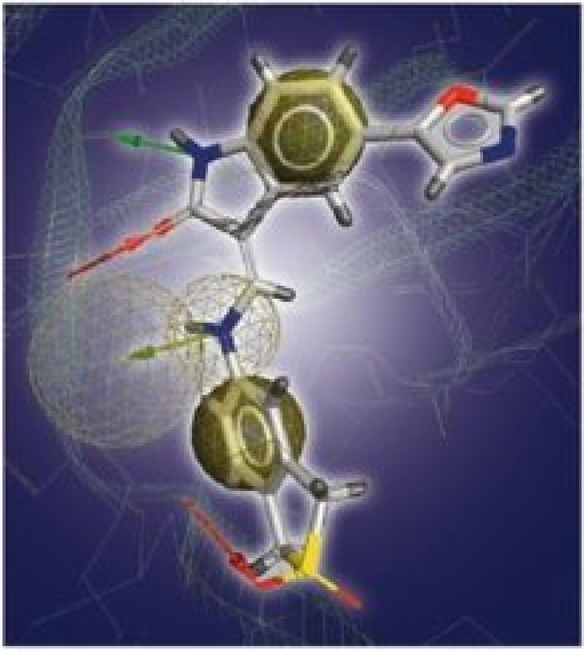
Methods and Principles in Medicinal Chemistry

Edited by
Thierry Langer and Rémy D. Hoffmann

WILEY-VCH

Pharmacophores and Pharmacophore Searches



Volume 32

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

Pharmacophores and Pharmacophore Searches

Edited by
Thierry Langer
and Rémy D. Hoffmann

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

Methods and Principles in Medicinal Chemistry

Previous Volumes of this Series:

H. Kubinyi, G. Müller (eds.)
Chemogenomics in Drug Discovery
Vol. 22
2004, ISBN 3-527-30987-X

T. I. Oprea (ed.)
Chemoinformatics in Drug Discovery
Vol. 23
2005, ISBN 3-527-310753-2

R. Seifert, T. Wieland (eds.)
G-Protein Coupled Receptors as Drug Targets
Vol. 24
2005, ISBN 3-527-30819-9

O. Kappe, A. Stadler
Microwaves in Organic and Medicinal Chemistry
Vol. 25
2005, ISBN 3-527-31210-2

W. Bannwarth, B. Hinzen (eds.)
Combinatorial Chemistry, 2nd Ed.
Vol. 26
2005, ISBN 3-527-30693-5

G. Cruciani (ed.)
Molecular Interaction Fields
Vol. 27
2005, ISBN 3-527-31087-8

M. Hamacher, K. Marcus, K. Stühler, A. van Hall, B. Warscheid, H. E. Meyer (eds.)
Proteomics in Drug Design
Vol. 28
2005, ISBN 3-527-31226-9

D. Triggle, M. Gopalakrishnan, D. Rampe, W. Zheng (eds.)
Voltage-Gated Ion Channels as Drug Targets
Vol. 29
2006, ISBN 3-527-31258-7

D. Rognan (ed.)
GPCR Modelling and Ligand Design
Vol. 30
2006, ISBN 3-527-31284-6

D. A. Smith, H. van de Waterbeemd, D. K. Walker
Pharmacokinetics and Metabolism in Drug Research, 2nd Ed.
Vol. 31
2006, ISBN 3-527-31368-0

Pharmacophores and Pharmacophore Searches

Edited by
Thierry Langer and Rémy D. Hoffmann

WILEY-VCH Verlag GmbH & Co. KGaA

The Editors

Prof. Dr. Raimund Mannhold
Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstrasse 1
40225 Düsseldorf
Germany
mannhold@uni-duesseldorf.de

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

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

Volume Editors

Prof. Dr. Thierry Langer
Institute of Pharmacy
Leopold-Franzens-Universität Innsbruck
Innrain 52A
6020 Innsbruck
Austria
thierry.langer@uibk.ac.at

Dr. Rémy D. Hoffmann
Accelrys, SARL
Parc Club Orsay Université
20 Rue Jean Rostand
91898 Orsay Cedex
France
remy@accelrys.com

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

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.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>

Typsetting K+V Fotosatz GmbH, Beerfelden
Printing Strauss GmbH, Mörlenbach
Binding Litges & Dopf GmbH, Heppenheim
Cover Design Grafik-Design Schulz, Fußgönheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN-13: 978-3-527-31250-4
ISBN-10: 3-527-31250-1

Contents

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

Preface

The idea is very straightforward: find and define all locations in space at a certain time of all substituents of a bioactive molecule that contribute to its biological activity. The readout would be a three-dimensional map – with respect to structure – that represents a minimal set of substituents which would adapt to a negative casting mold of the target binding site. By estimating or calculating the electronic and geometric properties of the substituents at their locations you would expand the 3D map to multiple dimensions. You call it a pharmacophore. After that, theoretically, you would walk through the Periodic Table and create a set of substituents, tied together by an appropriate backbone to fulfill all electronic and steric requirements of the pharmacophore. Finally, you obtain a new chemical entity with good prospects for activity at the target of choice.

But you get more. A "map" is a tool that relates objects to each other. These relations may be distances as they appear on a roadmap, it may be frequencies or densities on a web exploration map or it may be metabolism–emotion relationships in a brain map. Hence the pharmacophoric map can be used as a filter by matching the property vectors and a library of synthetic and/or virtual ligands, sorting out putative binders.

Well, "*Before the gates of excellence the high gods have placed sweat; long is the road thereto and rough and steep at first*" (Hesiod, *Work and Days*).

In the present book, Thierry Langer and Rémy Hoffmann give us a description of the long road with a firm sight on what can be done now and what is still to be achieved. Camille Wermuth, a doyen of the field, starts the arc of contributions shaping the history of the pharmacophore concept. The subsequent chapters are grouped into two major parts: "Pharmacophore Approaches" and "Pharmacophores for Hit Identification and Lead Profiling: Applications and Validation". Much attention is devoted to the problem of alignment and cost of energy. The contributions face the problems not only from the small molecule, the ligand's view, but also from the complementary side, the receptor's binding site. Experience from both industrial research and development laboratories and academic research is covered, especially in the applications and validation part, which gives the reader a feeling for the feasibility and implementation of the approaches and bridges the gap between theory and practice.

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

The series editors are indebted to the authors and the editors who devoted much of their time to educational purposes and rendered this exciting issue possible.

We also want to express our gratitude to Renate Doetzer and Frank Weinreich of Wiley-VCH for cooperative and easy collaboration and their invaluable support in this project.

April 2006

Hugo Kubinyi, Weisenheim am Sand
Gerd Folkers, Zürich
Raimund Mannhold, Düsseldorf

A Personal Foreword

Pharmacophores! Behind this simple word and concept that may be seen somehow reductionist, a vast amount of information about bioactive molecules and their structure–activity relationships is hidden, but available. Both of us had the privilege of having been exposed first to these important tools in medicinal chemistry by Professor Camille-Georges Wermuth some 20 years ago at the faculty of Pharmacy of the Université Louis Pasteur in Strasbourg. In this academic laboratory, several drug molecules have been developed that were successfully brought to the market. The pharmacophore concept was used always keeping in mind the need to understand, explain and predict molecular interactions with the targets in addition to structure–activity relationships. Its practical applicability for medicinal chemists made it an excellent communication tool between modelers and synthetic chemists. We are therefore grateful to Professor Wermuth, who has kindly accepted to write the first chapter of this book.

Since that time, we have been working in the context of using and developing tools and methods for rational molecular design, in both academic and industrial environments. We have seen several key changes in paradigms, such as combinatorial chemistry and associated HTS techniques, structure-based design strongly related to the ever-increasing number of characterized 3D structures of target proteins and the emerging virtual screening technologies. Pharmacophores have somehow been neglected in the last decade, although some gold standard tools were already available to the research community that have unfortunately not been further developed. However, as the hype about both structure-based design and large-scale HTS has flattened, a new area for pharmacophore tools obviously has begun.

As outlined in this book, several innovative tools and approaches for pharmacophore-based modeling and screening have emerged recently in the literature. Since the last textbook on pharmacophores and their usage in drug discovery, edited by Osman F. Güner in 2000, considerable progress has been achieved and also a large number of success stories in different application areas have clearly demonstrated the power of this approach. We felt that now was the right time to summarize these developments and their applicability. Therefore, we are grateful to the series editors, Professors Hugo Kubinyi, Gerd Folkers and Raimund Mannhold, for having invited us to edit a book focusing on this exciting research area. Starting with an introductory historical overview, ligand-based

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

approaches, including 3D pharmacophores and 4D QSAR, are discussed, and also the concept and application of pseudoreceptors. Another section on structure-based approaches includes pharmacophores from ligand–protein complexes, FLIP and a chapter on 3D protein-ligand binding interactions. The whole is rounded off with a complete section devoted to applications and examples, including modeling of ADME properties.

The intention of this book is to provide the reader with the different aspects of pharmacophores and pharmacophore-based screening in the drug discovery and development context. Each chapter is written by well-recognized experts in their respective fields. We take the opportunity to thank them all for their contributions to this book. It was a privilege to interact with them in order to bring this ambitious project to fruition. We hope that this book will contribute to stimulating further developments in this area, since we feel that there is still room for new technologies and improvements around pharmacophores. Happy reading!

Innsbruck and Paris, March 2006

Thierry Langer
Rémy D. Hoffmann

List of Contributors

Francine Acher
Laboratoire de Chimie et Biochimie
Pharmacologiques et Toxicologiques
Université René Descartes – Paris V
UMR 8601 – CNRS
45 rue des Saints-Pères
75270 Paris Cedex 06
France

Stefano Alcaro
Dipartimento di Scienze
Farmacobiologiche
Università di Catanzaro
“Magna Græcia”
Complesso Ninì Barbieri
88021 Roccelletta di Borgia (CZ)
Italy

Hughes-Olivier Bertrand
Accelrys
Parc Club Orsay Université
20 rue Jean Rostand
91898 Orsay Cedex
France

Maurizio Botta
Dipartimento Farmaco Chimico
Tecnologico
Università degli Studi di Siena
Via Alcide de Gasperi, 2
53100 Siena
Italy

Ruth Brenk
Department of Pharmaceutical
Chemistry
UCSF – QB3–501C
Box 2550
1700 4th Street
San Francisco, CA 94143
USA

Cheng Chang
Department of Pharmaceutical
Sciences
University of Maryland
20 Penn Street
Baltimore, MD 21201
USA

Claudio Chuaqui
Computational Drug Design Groups
Department of Research Informatics
Biogen Idec
12 Cambridge Center
Cambridge, MA 02142
USA

Zhan Deng
Computational Drug Design Groups
Department of Research Informatics
Biogen Idec
12 Cambridge Center
Cambridge, MA 02142
USA

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

Sean Ekins
Department of Pharmaceutical
Sciences
University of Maryland
20 Penn Street
Baltimore, MD 21201
USA
and
GeneGo, Inc.
500 Renaissance Drive, Suite 106
St. Joseph, MI 49085
USA

Uli Fechner
Johann-Wolfgang-Goethe-Universität
Institut für Organische Chemie und
Chemische Biologie
Max-von-Laue-Straße 7
60439 Frankfurt am Main
Germany

Sally Hindle
BioSolveIT GmbH
An der Ziegelei 75
53757 Sankt Augustin
Germany

Rémy D. Hoffmann
Accelrys, SARL
Parc Club Orsay Université
20 Rue Jean Rostand
91898 Orsay Cedex
France

Prabha Karnachi
Johnson & Johnson Pharmaceutical
Research and Development
1000 Route 202
P.O. Box 300
Raritan, NJ 08869
USA

Thomas Klabunde
Aventis Pharma Deutschland GmbH
Scientific & Medical Affairs,
Drug Design
Building G838
65926 Frankfurt am Main
Germany

Gerhard Klebe
Institute of Pharmaceutical Chemistry
University of Marburg
Marbacher Weg 6
35032 Marburg
Germany

Robert Kosara
University of North Carolina
at Charlotte (UNCC)
Department of Computer Science
College of Information Technology
9201 University City Blvd
Charlotte, NC 28223
USA

Amit Kulkarni
Accelrys Inc.
9685 Scranton Road
San Diego, CA 92121
USA

Thierry Langer
Institut für Pharmazie/
Abt. Pharmazeutische Chemie
Leopold-Franzens-Universität
Innsbruck
Innrain 52
6020 Innsbruck
Austria

Tien Luu
Accelrys Ltd.
334 Cambridge Science Park
Cambridge CB4 0WN
UK

Patrick Maaß
Center for Bioinformatics Hamburg
(ZBH)
University of Hamburg
Bundesstraße 43
20146 Hamburg
Germany

Fabrizio Manetti
Dipartimento Farmaco Chimico
Tecnologico
Università degli Studi di Siena
Via Alcide de Gasperi, 2
53100 Siena
Italy

Günther Metz
Santhera Pharmaceuticals AG
Im Neuenheimer Feld 518–519
69120 Heidelberg
Germany

Francesco Ortuso
Dipartimento di Scienze
Farmacobiologiche
Università di Catanzaro
"Magna Græcia",
Complesso Ninì Barbieri
88021 Roccelletta di Borgia (CZ)
Italy

Konstantin Poptodorov
Accelrys Ltd.
334 Cambridge Science Park
Cambridge CB4 0WN
UK

Matthias Rarey
Center for Bioinformatics Hamburg
(ZBH)
University of Hamburg
Bundesstraße 43
20146 Hamburg
Germany

Steffen Renner
Johann-Wolfgang-Goethe-Universität
Institut für Organische Chemie und
Chemische Biologie
Max-von-Laue-Straße 7
60439 Frankfurt am Main
Germany

Christian Rummey
Santhera Pharmaceuticals AG
Im Neuenheimer Feld 518–519
69120 Heidelberg
Germany

Klaus-Jürgen Schleifer
BASF Aktiengesellschaft
Computational Chemistry and Biology
Carl-Bosch-Straße 38
67056 Ludwigshafen
Germany

Gisbert Schneider
Johann-Wolfgang-Goethe-Universität
Institut für Organische Chemie und
Chemische Biologie
Max-von-Laue-Straße 7
60439 Frankfurt am Main
Germany

Juswinder Singh
Computational Drug Design Groups
Department of Research Informatics
Biogen Idec
12 Cambridge Center
Cambridge, MA 02142
USA

Wolfgang Sippl
Institute of Pharmaceutical Chemistry
Martin-Luther-Universität
Halle-Wittenberg
Wolfgang-Langenbeck-Straße 4
06120 Halle (Saale)
Germany

Andrea Tafi
Dipartimento Farmaco Chimico
Tecnologico
University of Siena
Via Aldo Moro
53100 Siena
Italy

Nicolas Triballeau
Accelrys
Parc Club Orsay Université
20 rue Jean Rostand
91898 Orsay Cedex
France

Camille G. Wermuth
Prestwick Chemical
Boulevard Gonthier d'Andernach
67400 Illkirch Cédex
France

Gerhard Wolber
Inte:Ligand GmbH
Mariahilferstrasse 74B/11
1070 Vienna
Austria

Marc Zimmermann
Fraunhofer Institute for Algorithms
and Scientific Computing (FhI-SCAI)
Schloss Birlinghoven
53754 Sankt Augustin
Germany

Part I
Introduction

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

1
Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist

Camille G. Wermuth

Since the appearance of computer-aided structure–activity studies, the term "pharmacophore" has become one of the most popular words in medicinal chemistry. However, depending on their scientific background and/or traditions, the different medicinal chemistry groups attribute various meanings to this term. Therefore, it appeared necessary to devote a brief paragraph to the definition of the word pharmacophore, and this is followed by a historical perspective and finally by some comments from a medicinal chemistry practitioner.

1.1
Definitions

Many authors use the term "pharmacophores" to define functional or structural elements possessing biological activity. This does not correspond to the official definition elaborated by an IUPAC working party and published in 1998 [1]: *A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response*. As a consequence:

1. The pharmacophore describes the essential, steric and electronic, function-determining points necessary for an optimal interaction with a relevant pharmacological target.
2. The pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure.
3. Pharmacophores are not specific functional groups (e.g. sulfonamides) or "pieces of molecules" (e.g. dihydropyridines, arylpiperazines).

A pharmacophore can be considered as the highest common denominator of a group of molecules exhibiting a similar pharmacological profile and which are recognized by the same site of the target protein. However, despite the official

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

definition and the remarks made above, many medicinal chemists continue to call pharmacophores some specific functional groups, especially if they appear to be often associated with biological activity.

1.1.1 Functional Groups Considered as Pharmacophores: the Privileged Structure Concept

The retrospective analysis of the chemical structures of the various drugs used in medicine led medicinal chemists to identify some molecular motifs that are associated with high biological activity more frequently than other structures. Such molecular motifs were called privileged structures by Evans et al. [2], to represent substructures that confer activity to two or more different receptors. The implication was that the privileged structure provides the scaffold and that the substitutions on it provide the specificity for a particular receptor. Two monographs deal with the privileged structure concept [3, 4].

Among the most popular privileged structures, historical representatives are arylethylamines (including indolylethylamines), diphenylmethane derivatives, tricyclic psychotropics and sulfonamides. Dihydropyridines [5], benzodiazepines, [2, 5], *N*-arylpiperazines, biphenyls and pyridazines [6] are more recent contributions.

A statistical analysis of NMR-derived binding data on 11 protein targets indicates that the biphenyl motif is a preferred substructure for protein binding [7].

1.2 Historical Perspective

1.2.1 Early Considerations About Structure–Activity Relationships

In his interesting Edelstein award lecture, presented at the 224th American Chemical Society Meeting in Boston, MA, in August 2002 and entitled “To Bond or Not to Bond: Chemical Versus Physical Theories of Drug Action”, John Parascandola [8] relates the early history of structure–activity relationships.

Regarding drug selectivity, he cites Earles, who states: “The fact that drugs may exert a selective action on specific organs of the body had long been recognized empirically and expressed vaguely in the traditional designation of certain remedies as cordials (acting on the heart), hepatics (acting on the liver), etc.” [9].

One of the earliest to recognize structure–activity relationships was Robert Boyle in 1685, who tried to explain the specific effects of drugs in terms of mechanical philosophy by suggesting that since the different parts of the body have different textures, it is not implausible that when the corpuscles of a substance are carried by the body fluids throughout the organism, they may, according to their size, shape and motion, be more fit to be detained by one organ than another [10].

Later, at the turn of the 20th century, the German scientist Sigmund Fränkel argued that the selective action of drugs can only be understood by assuming that certain groups in the drug molecule enter into a chemical union with the cell substance of a particular tissue. Once fixed in the cell in this manner, the drug can exert its pharmacological action [11].

Despite this pioneering view, the understanding of the nature of chemical bonding and of cellular structure and function was still in its infancy at the beginning of the 20th century. Thus there was significant controversy over whether the physical or the chemical properties of a substance could best explain its pharmacological action and over the value of attempts to relate the physiological activity of a drug to its chemical structure. As an example, in 1903 Arthur Cushny, Professor of Materia Medica and Therapeutics at the University of Michigan, published a paper in the *Journal of the American Medical Association* entitled "The pharmacologic action of drugs: is it determined by chemical structure or by physical characters?" [12]. To a chemist today, such a question might seem odd. Finding convincing answers to it became possible only after the discovery of the existence and role of pharmacological receptors.

1.2.2 Early Considerations About the Concept of Receptors

The idea that drugs act upon receptors began with Langley in 1878 [13], who introduced the term "receptive substance" [14]. However, the word "receptor" was introduced later, by Paul Ehrlich [15, 16]. During the first half of the 20th century, several observations highlighted the critical features associated with the concept of receptors [17].

"Three striking characteristics of the actions of drugs indicate very strongly that they are concentrated by cells on small, specific areas known as receptors. These three characteristics are (i) the high dilution (often 10^{-9} M) at which solutions of many drugs retain their potency, (ii) the high chemical specificity of drugs, so discriminating that even D- and L-isomers of a substance can have different pharmacological actions, and (iii) the high biological specificity of drugs, e.g. adrenaline has a powerful effect on cardiac muscle, but very little on striatal muscle." [17].

1.2.3 Ehrlich's "Magic Bullet"

Selective interaction of a drug molecule with the corresponding receptor was not always accepted. One of the most brilliant demonstrations came from Paul Ehrlich's discovery of salvarsan, which gave rise to the concept of a chemotherapeutic "magic bullet" against specific infectious organisms. Beginning with dyes and later extending his studies to include arsenical compounds, Ehrlich modified the chemical structure of numerous molecules to produce effective drugs against trypanosome and later spirochete infections. They tested hundreds of compounds before they came upon one, number 606, that Ehrlich thought was the chemotherapeutic agent he was searching for. Clinical tests confirmed the potential of the drug in treating syphilis and trypanosomiasis. The discovery was announced in 1910. Ehrlich named the drug salvarsan. The German physician, bacteriologist and chemist Paul Ehrlich shared the Nobel Prize in 1908 with Ilya Metchnikoff for their contributions to immunity.

1.2.4 Fischer's "Lock and Key"

Ehrlich's seminal discoveries reinforced the assertion made in 1894 by another brilliant German chemist, Emil Fischer. In a publication dealing with the effect of glucoside conformation on the interaction with enzymes, he wrote: "Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen, um eine chemische Wirkung auf einander ausüben zu können" (To illustrate, I would like to say that enzyme and glucoside must fit together like lock and key, in order to have a chemical effect on each other) [18]. The image of "lock and key" is still used today, even if it suggests a rigid structure of the receptor or enzyme protein. Probably another image, such as "hand in a glove", would be more accurate. Effectively, in addition to the steric complementarity, it would account for chirality and receptor flexibility.

1.3 Pharmacophores: the Viewpoint of a Medicinal Chemist

Even before the advent of computer-aided drug design, simple pharmacophores were described in the literature and considered as tools for the design of new drug molecules. Initial structure–activity relationship considerations were accessible in the 1940s thanks to the knowledge of the bond lengths and the van der Waals sizes which allowed the construction of simple two-dimensional model structures. With the availability of X-ray analysis and conformational chemistry, access to three-dimensional models became possible in the 1960s.

1.3.1 Two-dimensional Pharmacophores

1.3.1.1 Sulfonamides and PABA

The recognition of the quantitatively almost unmatched ability of *p*-aminobenzoic acid (PABA) to oppose the bacteriostatic efficiency of the sulfonamides led Woods and Fildes [19, 20] to formulate the fundamentals of the theory of metabolite antagonism (Fig. 1.1).

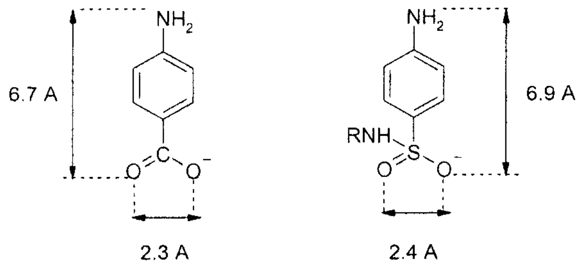


Fig. 1.1 PABA and *p*-aminobenzenesulfonamide show similar critical distances. The incorporation of the sulfonamide instead of PABA inhibits the biosynthesis of tetrahydrofolic acid.

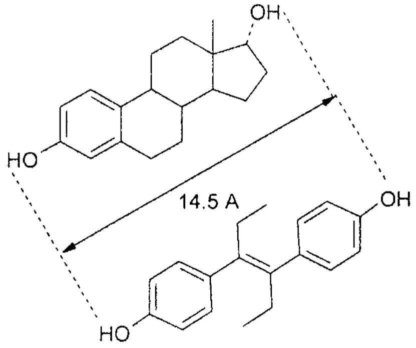


Fig. 1.2 Analogy between estradiol and *trans*-diethylstilbestrol.

1.3.1.2 **Estrogens**

Another early achievement (Fig. 1.2) was the synthesis and the pharmacological evaluation of *trans*-diethylstilbestrol as an estrogenic agent showing similarities with estradiol [21]. Here again the proposed model was two-dimensional [22], despite the fact that the non-planar conformation of estradiol was already known.

1.3.2
An Early Three-dimensional Approach: the Three-point Contact Model

When an asymmetric center is present in a compound, it is thought that the substituents on the chiral carbon atom make a three-point contact with the receptor. Such a fit insures a very specific molecular orientation which can only be obtained for one of the two isomers (Fig. 1.3). A three-point fit of this type was first suggested by Easson and Stedman [23], and the corresponding model proposed by Beckett [24] in the case of (*R*)-(–)-adrenaline [= (*R*)-(–)-epinephrine]. The more active natural (*R*)-(–)-adrenaline establishes contacts with its receptor through the three interactions shown in Fig. 1.3.

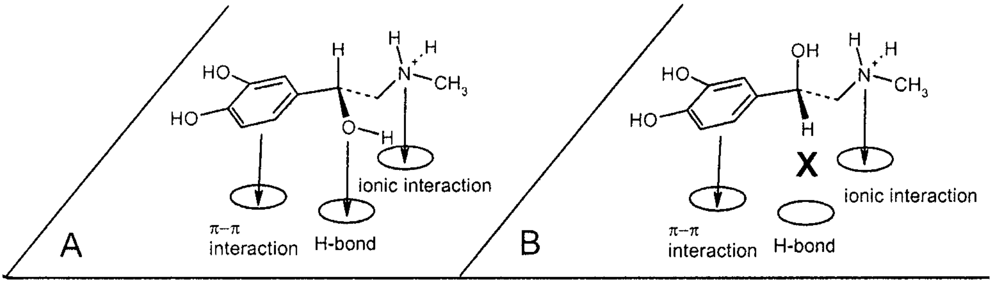


Fig. 1.3 Interaction capacities of the natural (*R*)-(–)-epinephrine and its (*S*)-(+)-antipode.

In simply assuming that the natural (*R*)-(–)-epinephrine establishes a three-point interaction with its receptor (A), the combination of the donor–acceptor interaction, the hydrogen bond and the ionic interaction will be able to generate energies of the order of 12–17 kcal mol^{-1}, which corresponds [25] to binding constants of 10^{-9}–10^{-12}. The less active isomer, (*S*)-(+)-epinephrine, may establish only a two-point contact (B). The loss of the hydrogen bond interaction equals ~3 kcal mol^{-1}, hence this isomer should possess an ~100-fold lesser affinity. Experience confirms this estimate. If we consider less abstract models, it becomes apparent that the less potent enantiomer also is able to develop three intermolecular bonds to the receptor, provided that it approaches the receptor in a different manner. However, the probability of this alternate binding mode to trigger the same biological response is close to zero.

1.3.2.1 Clonidine and Its Interaction with the *α*-Adrenergic Receptor

In the early 1970s, it was accepted that the hypotensive activity of clonidine was due to its direct interaction with the central norepinephrine receptor [26]. To trigger the *α*-adrenergic receptor, it was accepted that norepinephrine binds to its receptor by means of three bonds [27, 28]:

1. an ionic bond between the protonated amino function and an anion (carboxylate, phosphate) of the receptor active site;
2. a hydrogen bond between the secondary alcoholic hydroxyl and a, NH–CO function of the receptor;
3. a stacking (or charge transfer?) between the aromatic ring and an electron-deficient ring such as a protonated imidazole of a histidine residue.

In addition, it was known that the phenolic hydroxyls are not essential for *α* activity and that the cationic head should not be too bulky.

Pullmann et al. [29], in their model of the *α*-adrenergic receptor, found the following critical intramolecular distances: $D = 5.1$–5.2 Å from N$^+$ to the center of the aromatic ring and $H = 1.2$–1.4 Å for the elevation of the positive charge to the plane of the aromatic ring (Fig. 1.4).

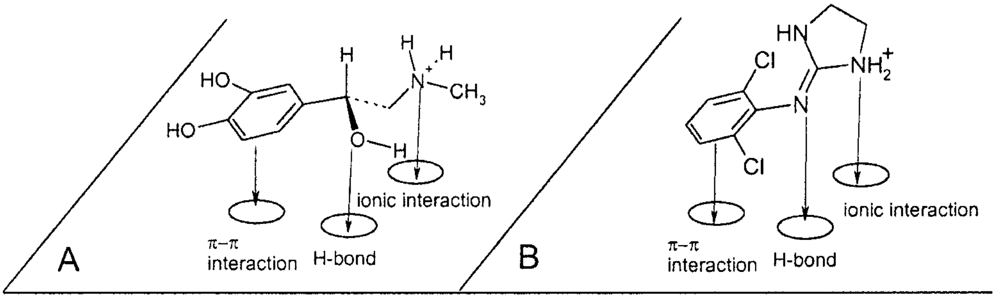


Fig. 1.4 In clonidine (B) the restricted rotation resulting from *o*- and *o*′-substitution imposes a quasi-perpendicular orientation of the imidazolic ring towards the phenyl ring. As a result, clonidine can yield the same kind of interactions than norepinephrine (A).

At first glance, the similarity between clonidine and norepinephrine was not evident; However an NMR structural study of clonidine demonstrated the restricted rotation resulting from *o*- and *o*′-substitution and imposing a quasi-perpendicular orientation of the imidazolic ring towards the phenyl ring [30]. As a result, clonidine can yield the same kind of interactions as norepinephrine.

Taken together, the examples shown above illustrate typically some pre-computer attempts to elucidate pharmacophoric patterns usable as guides for the design of new drugs. They prepared the minds for Garland Marshall's seminal publications (see references in [31, 32]) on computer-aided pharmacophore identification and all the derived applications that will be presented in the following chapters.

1.3.3 Criteria for a Satisfactory Pharmacophore Model [32]

To be recognized as a useful tool, a pharmacophore model has to provide valid information for the medicinal chemist exploring structure–activity relationships.

1. First, it has to highlight the functional groups involved in the interaction with the target, the nature of the non-covalent bonding and the different intercharge distances. This means that worthless images of ribbon and spaghetti models [33], without indication of the molecular features of the interacting partners, have to be avoided. This is true also for many unnecessary and opaque theoretical digressions. The model also has to show some *predictive power* and lead to the design of new, more potent compounds or, even better, of totally novel chemical structures, not evidently deriving from the translation of structural elements from one active series into the other. An interesting aspect of pharmacophore-based analogue design is referred to as scaffold hopping. It consists in the design of functional analogues by searching within large virtual compound libraries of isofunctional structures, but based on a

different scaffold. The objective is to escape from a patented chemical class in identifying molecules in which the central scaffold is changed but the essential function-determining points are preserved and form the basis of a relevant pharmacophore [34].

2. The second criterion for a valid pharmacophore model is that it should discriminate stereoisomers. Stereospecificity is one of the principal attributes of pharmacological receptors and a perfect stereochemical complementarity between the ligand and the binding-site protein is an essential criterion for high affinity and selectivity. A convincing example of enantiomeric discrimination was observed for GABA-A receptor antagonists [35].
3. In a similar manner, the ideal model should distinguish between agonists and antagonists. This is relatively easy for the specific category of antagonists which, according to Ariëns et al. theory [36], derive from the agonists simply through the addition of some supplementary aromatic rings which play the role of additional binding sites (e.g. the passage from muscarinic agonists to muscarinic antagonists [37] or from GABA agonists to GABA antagonists [35]). The discrimination between the two categories becomes less evident when the passage from agonist to antagonist relies on relatively subtle changes such as one observes for glutamate, oxotremorine and benzodiazepine antagonists.
4. Sometimes a good pharmacophore model can *explain* apparently *paradoxical observations*, e.g. the unexpected affinity reversal found in *R*- and *S*-enantiomers of the sulpiride series on changing *N*-ethyl to *N*-benzyl derivatives [38].
5. Finally, it has to account for the *lack of activity* of certain analogues of the active structures. The knowledge of structural or electronic parameters leading to poorly active or inactive compounds is a cost-lowering factor that allows the number of compounds to be synthesized to be reduced.

1.3.4 Combination of Pharmacophores

Some highly specific mono-target drugs have clearly proven the usefulness of mono-target medicine. Examples are phosphodiesterase 5 inhibitors such as sildenafil, the α-1a antagonist drugs such as tamsulosine, selective COX-2 inhibitors such as celecoxib and kinase-specific anticancer drugs such as imatinib. However, in addition to one-target drugs, clinicians are more and more convinced that modulating a multiplicity of targets can be an asset in treating a range of disorders. An extreme example of a multi-target drug is clozapine, which exhibits nanomolar affinities for more than a dozen different receptors.

As a consequence of this trend, an increasing number of publications reflect an awakening of interest in the rational design of multiple ligands and may suggest an ongoing re-evaluation of the "one disease, one drug" paradigm which has dominated thinking in the pharmaceutical industry for the last few decades. Although there is little chance of switching back to the animal-centric approach of the past, it is now widely recognized that high specificity for a single target may not deliver the required efficacy versus side-effect profile and, in many cases, a balanced activity at several targets may produce a superior effect.

In a recent paper, entitled “From magic bullets to designed multiple ligands”, Morphy et al. [39] discuss the opportunity and the advantages attached to the design of ligands acting on two (or more) specific targets, such *intentionally* designed multiple ligands (DM ligands) being opposed to *serendipitous* multiple ligands. It is highly probable that computer-driven combinations of two pharmacophores can lead to the design of new active entities combining in one molecule the critical structural elements of two partners.

1.4 Conclusion

For medicinal chemistry practitioners, the term “pharmacophore” covers two different meanings: “pieces of molecules conferring activity, often referred too as privileged structures” and “the highest common denominators of a group of molecules exhibiting a similar pharmacological profile and which are recognized by the same site of the target protein”. The knowledge of the first meaning and its daily use belong to the medicinal chemists’ “culture générale”.

The second meaning aims to approach drug design by rational, computer-aided reasoning. It usefulness covers three major domains. The first is the establishment of a relevant pharmacophore model, consistent with structure–activity relationships in a series of molecules and allowing the design of optimal ligands. The second is scaffold hopping, which consists in the design of functional analogues by searching within large virtual compound libraries of isofunctional structures, but based on a different scaffold. The third deals with computer-driven combinations of two pharmacophores in the hope of designing new active entities combining in one molecule the critical pharmacophoric elements of two partners. All these applications will be presented and discussed in the following chapters of this book.

References

1 Wermuth, C.G., Ganellin, C.R., Lindberg, P., Mitscher, L.A., Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1997). *Annu. Rep. Med. Chem.* **1998**, *33*, 385–395.

2 Evans, B.E., Rittle, K.E., Bock, M.G., DiPardo, R.M., Freidinger, R.M., Whitter, W.L., Lundell, G.F., Veber, D.F., Anderson, P.S., Chang, R.S., Lotti, V.J., Cerno, D.J., Chen, T.B., Kling, P.J., Kunkel, K. A., Springer, J.P., Hirshfield, J., Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

3 Trainor, G., Privileged structures – an update. *Annu.Rep. Med. Chem.* **2000**, *35*, 289–298.

4 Sheridan, R.P., Miller, M.D., A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915–924.

5 Thompson, L. A., Ellman, J. A., Synthesis and applications of small molecule libraries. *Chem. Rev.* **1966**, *96*, 555–600.

6 Wermuth, C. G., Search for new lead compounds: the example of the chemical and pharmacological dissection of aminopyridazines. *J. Heterocycl. Chem.*, **1998**, *35*, 1091–1100.

7 Hajduk, P. J., Bures, M., Praestgaard, J., Fesik, S. W., Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **2000**, *43*, 3443–3447.

8 Parascandola, J., To bond or not to bond. *Bull. Hist. Chem.* **2003**, 28.

9 Earles, M. P., Early theories of the mode of action of drugs and poisons. *Ann. Sci.* **1961 (publ. 1963)**, *17*, 97–110.

10 Boyle, R., *Of the Reconcileableness of Specific Medicines to the Corpuscular Philosophy.* Samuel Smith, London, **1685**, pp. 72–75.

11 Fränkel, S., *Die Arzneimittel-Synthese auf Grundlage der Beziehungen zwischen chemischem Aufbau und Wirkung.* Julius Springer, Berlin, **1901**, pp. 13–41.

12 Cushny, A. R., The pharmacologic action of drugs: is it determined by chemical structure or by physical characters? *J. Am. Med. Assoc.* **1903**, *41*, 1252–1253.

13 Langley, J. N., On the physiology of the salivary secretion. Part II. On the mutual antagonism of atropin and pilocarpin, having especial reference to their relations in the sub-maxillary gland of the cat. *J. Physiol.* **1878**, *1*, 339–369.

14 Langley, J. N., On the reaction of cells and nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari. *J. Physiol.* **1905**, *33*, 374–413.

15 Ehrlich, P., Morgenroth, J., Über Haemolysine. Dritte Mitteilung. *Berl. Klin. Wochnschr.* **1900**, *37*, 453–457.

16 Maehle, A. H., Prull, C. R., Halliwell, R. F., The emergence of the drug receptor theory. *Nat. Rev. Drug Discov.* **2002**, *1*, 637–641.

17 Albert, A., *Selective Toxicity. The Physicochemical Basis of Therapy.* Chapman and Hall, London, **1979**, p. 23

18 Fischer, E., Einfluss der Konfiguration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985–2993.

19 Woods, D. D., The relation of *p*-aminobenzoic acid to the mechanism of the action of sulphonamide. *Br. J. Exp. Pathol.* **1940**, *21*, 74–90.

20 Woods, D. D., Fildes, P., The anti-sulphanilamide activity (*in vitro*) of *p*-aminobenzoic acid and related compounds. *Chem. Ind.* **1940**, *59*, 133–134.

21 Dodds, E. C., Lawson, W., Molecular structure in relation to oestrogenic activity. Compounds without phenanthrene nucleus. *Proc. R. Soc. London, Ser. B* **1938**, *125*, 122–132.

22 Schueler, F. W., Sex hormonal action and chemical constitution. *Science* **1946**, *103*, 221–223.

23 Easson, L. H., Stedman, E., Studies on the relationship between chemical constitution and physiological action. V. Molecular dissymmetry and physiological activity. *Biochem. J.* **1933**, *27*, 1257–1266.

24 Beckett, A. H., Stereochemical factors in biological activity. In *Fortschritte der Arzneimittel Forschung*, Birkhäuser Verlag, Basel, **1959**, pp. 455–530.

25 Farmer, P. S., Ariëns, E. J., Speculations on the design of nonpeptide peptidomimetics. *Trends Pharmacol. Sci.* **1982**, *3*, 362–365.

26 Anden, N. E., Corrodi, H., Fuxe, K., Hoekfelt, B., Hoekfelt, T., Rydin, C., Svensson, T., Evidence for a central noradrenaline receptor stimulation by clonidine. *Life Sci.* **1970**, *9*, 513–523.

27 Barlow, R. B., *Introduction to Chemical Pharmacology*, 2nd edn. Methuen, London, **1964**.

28 Belleau, B., An analysis of drug-receptor interactions. In *Modern Concepts in the Relationship Between Structure and Pharmacological Activity*, Brunings, K. J. (ed.). Pergamon Press, Oxford, **1963**, pp. 75–99.

29 Pullmann, B., Coubeils, J. L., Courrière, P., Gervois, J. P., Quantum mechanical study of the conformational properties of phenethylamines of biochemical and medicinal interest. *J. Med. Chem.* **1972**, *15*, 17–23.

30 Wermuth, C. G., Schwartz, J., Leclerc, G., Garnier, J. P., Rouot, B., Conformation de la clonidine et hypothèses sur son interaction avec un récepteur alpha-adrénergique. *Chim. Thér.* **1973**, *1*, 115–116.

31 Marshall, G. R., Binding-site modeling of unknown receptors. In *3D QSAR in Drug Design, Theory Methods and Applications*, Kubinyi, H. (ed.). ESCOM, Leiden, **1993**, pp. 80–116.

32 Wermuth, C. G., Langer, T., Pharmacophore identification. In *3D QSAR in Drug Design. Theory Methods and Applications*, Kubinyi, H. (ed.). ESCOM, Leiden, **1993**, pp. 117–136.

33 Wermuth, C. G., The impact of QSAR and CADD methods in drug design. In *Rational Approaches to Drug Design*, Hoeltje, H. D., Sippl, W. (eds.). Prous Science, Barcelona, **2001**, pp. 3–20.

34 Schneider, G., Giller, T., Neidhart, W., Schmid, G., "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896.

35 Rognan, D., Boulanger, T., Hoffmann, R., Vercauteren, D. P., André, J. M., Durant, F., Wermuth, C. G., Structure and molecular modeling of GABAA antagonists. *J. Med. Chem.* **1992**, *35*, 1969–1977.

36 Ariëns, E. J., Rodrigues de Miranda, J. F., Simonis, A. M., The pharmacon-receptor–effector concept: a basis for understanding the transmission of information in biological systems. In *The Receptors*, O'Brien, R. D. (ed.). Plenum Press, New York, **1979**, pp. 33–91.

37 Wermuth, C. G., Aminopyridazines–an alternative route to potent muscarinic agonists with no cholinergic syndrome. *Farmaco* **1993**, *48*, 253–274.

38 Rognan, D., Sokoloff, P., Mann, A., Martres, M. P., Schwartz, J. C., Costentin, J., Wermuth, C. G., Optically active benzamides as predictive tools for mapping the dopamine D2 receptor. *Eur. J. Pharmacol. Mol. Pharmacol. Sect.* **1990**, *3*, 59–70.

39 Morphy, R., Kay, C., Rankovic, Z., From magic bullets to designed multiple ligands. *Drug Discov. Today* **2004**, *9*, 641–651.

Part II
Pharmacophore Approaches

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

2
Pharmacophore Model Generation Software Tools

Konstantin Poptodorov, Tien Luu, and Rémy D. Hoffmann

2.1
Introduction

Although the concept of pharmacophores constituting a simple representation of molecules and chemical groups in certain order was introduced nearly a century ago [1], there has been increasing interest and focus on pharmacophores in recent years following the advances in computational chemistry research. The historical development of the pharmacophore concept has recently been reviewed [2].

Often, all alignment-based methods and molecular field and potential calculations are classified as pharmacophore perception techniques. We will include most of these methods in this review; however, when using the term pharmacophore model, we will be referring mainly to one specific type of perception, namely three-dimensional feature-based pharmacophore models represented by geometry or location constraints, qualitative or quantitative. An extrapolation of the pharmacophore approach to a set of multi-dimensional descriptors (pharmacophore fingerprints) has been developed mostly for library design and focusing purposes [3–8].

At the beginning of this chapter we will look into the different automated alignment methods as correct alignment is the first and most important prerequisite for a successful pharmacophore identification process. Further, we will elaborate how essential issues of pharmacophore modeling such as conformational search, pharmacophore feature definitions, compounds structure storage and screening are handled by various available software packages.

Various ways of perceiving pharmacophores have been explored, known issues with pharmacophore modeling have been addressed in one way or another and several computer-based applications with a pharmacophore focus have been created since the 1980s. Many of these programs are not intensively used today, but we consider that they should be mentioned in this review: ALADDIN [9], DANTE [10–13], APOLLO [14], RAPID [15], SCREEN and its PMapper from ChemAxon [16] and ChemX fingerprints [3] from Chemical Design (now Accelrys).

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

This review is based on literature data and on the personal experience of the authors and should not represent a direct comparison between packages but rather a snapshot of the current developments in pharmacophore perception technology from our perspective.

2.2 Molecular Alignments

Although the terms molecular alignment and superposition and pharmacophore elucidation are often used interchangeably, it is probably more accurate to differentiate alignment as providing a prerequisite to pharmacophore development. Conversely, some alignment methods require a pharmacophore as a starting point [17–19]. In this section, we briefly overview the molecular alignment methods available; extensive reviews and summaries of different superposition algorithms over the last 10 years are available elsewhere [20–22]. Of course, molecular alignment is not limited to just providing a basis for pharmacophore elucidation; it can also be used to derive 3D-QSAR models that potentially can estimate binding affinities, in addition to indirectly providing insight into the spatial and chemical nature of the receptor–ligand interaction of the putative receptor. Essentially, an alignment endeavors to produce a set of plausible relative superpositions of different ligands, hopefully approximating their putative binding geometry.

Many of the issues and concerns in the generation of pharmacophore models are inherent in different alignment methods. These issues can be used to differentiate or categorize the plethora of available algorithms.

2.2.1 Handling Flexibility

Primary among these issues is that of ligand flexibility, vital in the determination of the relevant binding conformation for each of the ligands concerned. Alignment methods can be considered rigid, semi-flexible or flexible. Rigid methods, while generally simpler and faster, require a presumption of the bioactive conformation of the ligands; this is often not possible and also removes the impartiality of the method. Semi-flexible methods are those that are fed with pre-generated conformers which are processed in either a sequentially, iterative or combinatorial manner. These methods lead to a further series of considerations such as whether the weighting, number and spread of conformers are determined by energy cut-offs or Boltzmann probability distributions and whether solvation models should be used. Flexible methods are considered to be those in which the conformational analysis is performed on-the-fly and these are generally the most time consuming as they require rigorous optimization.

2.2.2 Alignment Techniques

The fundamental nature of alignment methods can be broadly described as being either point or property based. In point-based algorithms, pairs of atoms or pharmacophores are usually superposed using a least-squares fitting. These algorithms often use clique detection methods [23, 24], which are based on the graph-theoretical approach to molecular structure, where a clique is a maximum completely connected subgraph, to identify all possible combinations of atoms or functional groups to identify common substructures for the alignment. The greatest limitation of these algorithms is the need for predefined anchor points, as the generation of these points can become problematic in the case of dissimilar ligands.

Property-based algorithms, often also termed field-based, make use of grid or field descriptors, the most popular of which are those obtained from the program GRID, developed by Goodford [25]. These are generated by defining a three-dimensional grid around a ligand and calculating the energy of interaction between the ligand and a given probe at each grid point. These diverse descriptors include various molecular properties such as molecular shape and volume, electron density, charge distribution such as molecular electrostatic potentials and even high-level quantum mechanical calculations.

These algorithms are commonly broken down into three stages, which are subject to much variation. First, each ligand is represented by a set of spheres or Gaussian functions displaying the property or properties of interest. Usually the property is first calculated on a grid and subsequently transformed to the sphere or Gaussian representation. A number of random or systematically sampled starting configurations are then generated depending on the degrees of freedom considered, rotational, translational and conformational. Finally, local optimizations are performed with some variant of the classical similarity measure of the intermolecular overlap of the Gaussians as the objective function. While earlier property-based alignment methods were commonly grid-based, these have been surpassed by Gaussian molecular representation and Gaussian overlap optimization. These provide high information contents and avoid the dependence on additional parameters such as grid spacing while also providing a substantial increase in speed.

Variations on these algorithms have included the application of Fourier space methods to optimize the electron density overlap, similar to the molecular replacement technique in X-ray crystallography [26] and differentially weighted molecular interaction potentials or field terms [27, 28]. Another interesting alternative has been to apportion the conformational space of the ligands into fragments, compute the property field on pairs of fragments and determine the alignment by a pose clustering and incremental build-up procedure of retrieved fragment pairs [29].

2.2.3 Scoring and Optimization

All alignment methods require some quantitative measure or fitness function, to assess the degree of overlap between the ligands being aligned and to monitor the progression of that optimization. This is most often manifested as a molecular similarity score or alignment index [22].

Typically in point-based algorithms, the optimization process endeavors to reduce the root-mean-square (RMS) deviation of the distances between the points or cliques by least-squares fitting. However, interesting variations have been developed including the use of distance matrices to represent any given conformation of a ligand [30]. Simulated annealing is used to optimize the fitness function, which is a quantification of the sum of the elements of the difference distance matrix created by calculating the magnitude of the difference for all corresponding elements of two matrices.

Another optimization method, related to the least-squares fitting used in point-based algorithms, is the directed tweak method [31]. This is a torsional space optimizer, in which the rotatable bonds of the ligands are adjusted at search time to produce a conformation which matches the 3D query as closely as possible. As directed tweak involves the use of analytical derivatives, it is very fast and allows for an RMS fit to consider ligand flexibility.

In property-based alignments where the molecular fields are represented by sets of Gaussian functions, the intermolecular overlap of the Gaussians is used as the fitness function or similarity index. The two most common optimization methods are Monte Carlo and simulated annealing [32, 33]. Other straightforward optimization algorithms include gradient-based methods and the simplex method, which seeks the vector of parameters corresponding to the global extreme (maximum or minimum) of any *n*-dimensional function, searching through the parameter space [34].

Further, more sophisticated, optimization algorithms include neural networks and genetic algorithms which mimic the process of evolution as they attempt to identify a global optimization [35]. In an alignment procedure chromosomes may encode the conformation of each ligand in addition to intramolecular feature correspondences, orientational degrees of freedom, torsional degrees of freedom or other information such as molecular electrostatic potential fields. During the optimization the chromosome undergo manipulation by genetic operators such as crossover and mutation.

Alignment methods are also known to combine different optimization methods, such as a genetic algorithm and a directed tweak method [36].

Although this summary has highlighted the most common differentiators that can be used to categorize the plethora of available algorithms, further issues are significant to the alignment dilemma. Such issues include the degree of human intervention required, how to address the relative significance or weighting of some ligands over other ligands and how some algorithms generate multiple alignment solutions rather than an optimum superposition.

2.3 Pharmacophore Modeling

A general workflow for the generation of pharmacophores from multiple ligands is outlined in Fig. 2.1. The field-based methods, although important and worth mentioning in this chapter, will be treated here separately from the "classical" pharmacophore modeling as defined above.

2.3.1 Compound Structures and Conformations

The generation of the correct compound structures is a critical step in which different components such as atomic valences, correct bond orders and properly defined aromaticity have to be considered carefully. In addition, the correct stereochemistry flags need to be added for a correct treatment of stereochemistry. Most of the current pharmacophore generation packages include compound builders, but users can also import them from external sources using common file formats, for example SMILES, MOL, SD or MOL2.

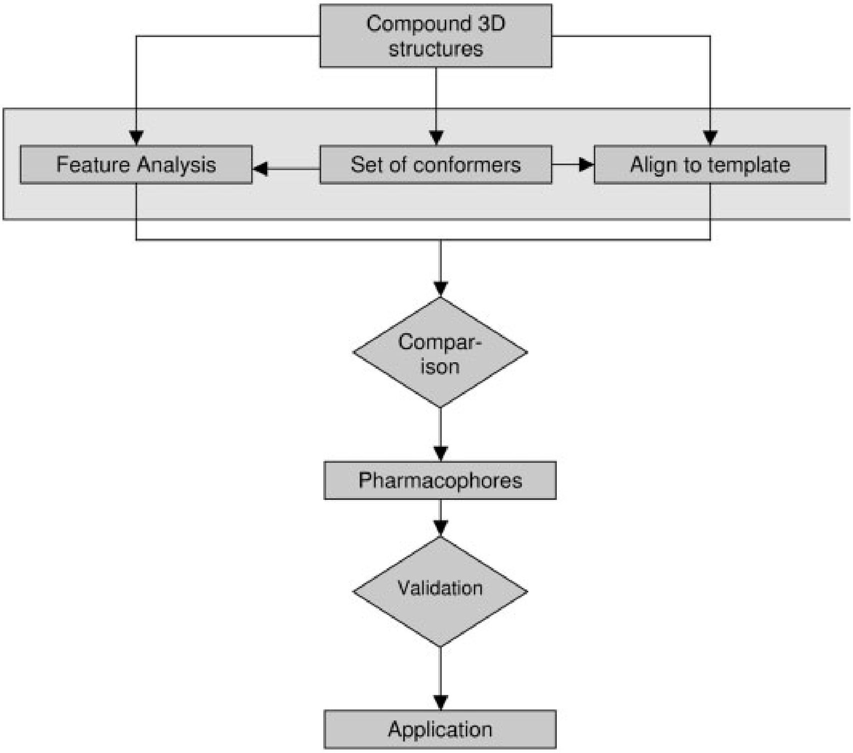


Fig. 2.1 Pharmacophore modeling workflow.

2.3.2 Representation of Interactions in the Pharmacophore Models

The representation of pharmacophores varies from one package to another and includes the nature of the pharmacophore points (fragments, chemical features) and the geometric constraints connecting these points (distances, torsions, three-dimensional coordinate location constraints).

The interpretation of the chemical structures of the molecules (Fig. 2.1, Feature Analysis) can be done at two levels:

1. Substructural, where molecules can be decomposed into different fragments, each fragment carrying certain specifications (e.g. basic nitrogen or aromatic ring).
2. Functional, where an abstraction of the structure is made such that each molecular fragment of the compounds is expressed by the general property it carries. In the current stage, the properties mapped on the fragments are chemical properties, e.g. hydrophobic or ionic interactions or hydrogen bonding features. The characterization of the chemical properties of compounds requires these functions to be accessible for the interaction with the binding partner (receptor, enzyme or nucleic acid), so in case the bioactive conformation of the ligand is not known, a conformational expansion analysis is a necessary step in order to identify a conformation which makes those functions available for interaction with the macromolecular target.

2.3.3 Conformational Expansion

This is probably the most critical step, since the goal here is not only to have the most representative coverage of the conformational space of a molecule, but also to have either the bioactive conformation as part of the set of generated conformations or at least a cluster of conformations that are close enough to this the bioactive conformation. Here we divide the methods that can be used for this purpose roughly into four categories: systematic search in the torsional space, optionally followed by clustering, stochastic methods, e.g. Monte Carlo, sampling, e.g. Poling [37], and molecular dynamics. The resulting set of conformations can be further optimized using minimization with or without solvent.

There are numerous references in the literature, e.g. [17], showing the effects of various sets of conformational models on pharmacophore generation; however, the goal of this chapter is not to describe and analyze the different approaches.

Marshall et al. described the so-called Active Analog Approach [38], in which the conformational space of flexible molecules is constrained to the geometry of a reference molecule (generally active and as rigid as possible). Pharmacophore models are then derived from the set of resulting alignments. This approach has been successfully used since the mid-1980s and still forms the basis of many existing automated pharmacophore modeling techniques.

2.3.4 Comparison

This step constitutes the pharmacophore generation itself and represents the major focus of this chapter. The majority of pharmacophore generation packages generate qualitative pharmacophores that do not consider the activity of the molecules (potency), so in general equipotent molecules have to be used. Most of these methods are based on minimizing the RMS superposition error between conformations of various compounds while trying to increase the three-dimensional overlay of pharmacophores. The result is generally multiple pharmacophore solutions, ranked according to different metrics depending on the package used. To our knowledge, currently only two packages are capable of generating SAR models on-the-fly by using directly activity values (K_i or IC_{50}): Catalyst® HypoGen [39, 40] and Apex3D [41].

2.3.5 Pharmacophores, Validation and Usage

After performing pharmacophore analysis on a set of compounds, typically the user will have to select the model(s) with biological and/or statistical relevance, often from multiple possible solutions and use for further research purposes. The validation of the pharmacophore models is therefore a critical aspect of the pharmacophore generation process. A review of the validation methods applicable to the field of pharmacophore generation is described elsewhere in this book [42].

In a nutshell, these validation methods can be ordered into three categories:

1. Statistical significance analysis, randomization tests.
2. Enrichment based methods. These focus on recovering active molecules from a test database in which a small number of known actives have been hidden in a large database of randomly selected compounds. Database mining and the utilization of receiver operating characteristic (ROC) curves [43] can be included in this category.
3. Biological testing of a selection of compounds.

The main utility of pharmacophores is their use as screening tools. Many examples in the literature show their successful usage in finding new scaffolds [44–51].

2.4 Automated Pharmacophore Generation Methods

In order to have an objective view of the different available pharmacophore perception software tools, we chose to analyze these using the following criteria whenever possible:

1. Compound builder: is there a molecular builder? Which file formats are supported?
2. Stereochemistry: how is the stereochemistry of molecules handled in the program?
3. 3D conformations: does the program contain conformer generation methods?
4. Pharmacophore generation engine: which type of pharmacophore perception engine is implemented in the program?
5. Fitness function: how are the pharmacophores evaluated?
6. Alignment method: does the program require pre-alignment of molecules? On what basis are the molecules aligned together?
7. Pharmacophore definition: description of the type of pharmacophore locations and associated functions. Can other descriptors be added to the pharmacophore alignment?
8. Database searching: is a database search engine implemented in the program? Which types of database searches are possible, e.g. substructure, pharmacophore, shape, exclusion?
9. Scoring of hits: Can database search hits be ranked?

The currently available pharmacophore perception methods are reviewed here in three major categories: geometry- and feature-based methods, field-based methods and pharmacophore fingerprints. Finally, the methods that do not fall into any of the above categories are described in an additional section.

2.4.1 Methods Using Pharmacophore Features and Geometric Constraints

2.4.1.1 DISCO, GASP and GALAHAD

The above programs are all currently implemented and marketed by Tripos and were developed by either the pharmaceutical industry or academic institutions and in cooperation with Tripos. All are integrated into the Sybyl® environment [52] and use it for visualization and molecule construction.

DISCO (DISCOtech)

Even though the original authors of DISCO do not consider it to be an automated pharmacophore identification program [53], we decided to include the method in this review because of its considerable influence over the development of modern pharmacophore modeling tools.

By design, no conformational engine was implemented in DISCO, based on the assumption that at the time, no universal force fields and methods suitable for all types of compounds were available [53]. However, the commercial distributor Tripos provides access to 3D converters and conformational search engines such as Concord® and Confort® via the Sybyl interface. These algorithms will not be reviewed here as strictly seen they are not part of any pharmacophore identification program. The distance geometry approach has been used successfully for subsequent pharmacophore modeling with DISSCO by the authors of DISCO and other researchers [53, 54].

DISCO considers three-dimensional conformations of compounds not as coordinates but as sets of interpoint distances, an approach similar to a distance geometry conformational search. Points are calculated between the coordinates of heavy atoms labeled with interaction functions such as HBD, HBA or hydrophobes. One atom can carry more than one label. The atom types are considered as far as they determine which interaction type the respective atom would be engaged in. The points of the hypothetical locations of the interaction counterparts in the receptor macromolecule also participate in the distance matrix. These are calculated from the idealized projections of the lone pairs of participating heavy atoms or H-bond forming hydrogens. The hydrophobic points are handled in a way that the hydrophobic matches are limited to, e.g., only one atom in a hydrophobic chain and there is a differentiation between aliphatic and aromatic hydrophobes. A minimum constraint on pharmacophore point of a certain type can be set, e.g. if a certain feature is known to be required for activity [53, 54].

DISCO relies on the Bron–Kerbosch clique detection algorithm for interdistance comparison. In DISCO, multiple conformations per compound are considered in the alignment and the stereochemistry is preserved. However, there is no mechanism for selecting conformations within the algorithm apart from the alignment to other structures, hence the user has to provide a conformational model that contains only the desired (low-energy) conformations. As a direct consequence, conformationally restrained compounds should be the preferred input for the program, provided that they carry the same activity as the more flexible analogues and the performance of the program tends to decrease with increasing flexibility of the input compounds [53].

During a DISCO run, one compound is taken as reference and each conformation of the remaining compounds is aligned on to the reference conformation in order to find a pharmacophore match. Typically, the least flexible compound serves as a reference as this reduces the pharmacophore space to explore and the number of results left to evaluate. The result is scored multiple pharmacophore solutions rather than a single model. The score is based on the number of participating molecules, number of features and the interfeature distances. Higher model quality is achieved by automatically reiterating through a number of variables such as distance tolerance specified as minimum, maximum and increment, number of features and compounds used in the analysis [53, 55]. The resulting pharmacophores are required to match all features in all compounds.

The pharmacophore points in the Tripos implementation of DISCO, currently marketed under the name DISCOtech^{TM}, can be represented as Tripos UNITY$^{®}$ [56] query features and the models can be used directly for UNITY database searches or in combination with 3D QSAR such as CoMFA as described in [57].

GASP

GASP stands for Genetic Algorithm Superposition Program and, as suggested by its name, it uses a genetic algorithm for pharmacophore identification. GASP was developed by Jones, Willet and Glenn in the mid-1990s [35]. The methods used in GASP are similar to those in the leading docking application GOLD, developed by the University of Sheffield, GlaxoSmithKline and CCDC [58, 59].

Unlike other pharmacophore identification routines, the conformational search is performed on-the-fly in GASP and represents an integral part of the program. Each compound is input a single, low-energy conformation and random rotations and translations are applied in order to explore the conformational variation prior to superposition.

The first step in the pharmacophore generation process with GASP is the determination of the pharmacophore features: rings, donors (protons) and acceptors (lone pairs). The atoms defined as HBA carriers can be aliphatic and aromatic nitrogens, alcohols, ethers, carbonyl, carboxyl oxygens and halogens; HBD carriers include amines and hydroxyls [60]. Projection points for the hydrogen bonding features are considered during pharmacophore analysis. GASP considers only aromatic structures as hydrophobic and there is no option to modify any of the pharmacophore feature definitions or introduce new ones [54]. If a training set consists of N compounds, a chromosome will consist of $2N - 1$ strings: N binary encoding the conformational information about each compound and $N - 1$ integer strings representing the feature mappings of the training set members to a single reference (base) molecule. The length of each integer string equals the number of features in the respective molecule. The compound with the least pharmacophore features is selected as base molecule [60]. No more than one molecule can be used for that purpose. Essentially, the program tries to maximize the mappings using a least-squares method while trying to satisfy a fitness function comprising three components: the similarity score of the mapped features, the volume integral of the aligned structures and the internal steric energy of the participating conformers, where the weighting of each contribution can be adjusted by the user.

GASP uses two genetic operators, crossover (two parents produce two children) and mutation (one parent produces one child) to evolve models with a maximum fitness score and therefore the highest quality structural alignment. The similarity score for the overlaid molecules is the sum of the scores of the similarity match between donors, acceptors and aromatic rings. The volume integral is determined as the mean volume integral per molecule with the base molecule. Finally, the internal van der Waals energy is calculated as Lennard–Jones 6–12 potential and represented as the difference from the preceding conformer. [60]. All features of all molecules must match in the alignment, hence no outliers are allowed and sometimes subsetting may be required during the training set preparation phase in order to separate out compounds that carry somewhat different pharmacophoric information [54]. Owing to the nature of the algorithm, each run may result in a slightly different solution. Several solutions can then be collected, ranked according to fitness score and analyzed visually in order to find the most suitable answer [54].

Similarly to DISCO, the alignments coming from GASP can be used as a starting point for CoMFA studies [61].

GALAHAD

GALAHAD [62] is a joint development between Tripos, the University of Sheffield, Novo Nordisk and Biovitrum. At the time of writing this review, there was little public information available about this new program; however, the underlying methods have been described earlier [63, 64]. The program uses a modified GA and seems to address the limitations of GASP in terms of increasing performance, reducing bias towards a single template (base) molecule, introducing partial matching and an improved multi-objective Pareto scoring function.

GALAHAD allows the use of pre-generated conformations as a starting point, which increases the speed of the calculation. Each molecule is represented as a core and set of torsions. In the alignment phase, a new method is used, where each molecule is compared with each other, hence no template is required. Pharmacophore similarity rather than feature mappings is used for the comparison, which should result in shorter run times. The fact that not all features are required to map contributes to the ability of the models to accommodate more diverse structures. Unlike GASP, GALAHAD reports multiple solutions from a single run which are ranked according to their scores and can be resubmitted for refinement if desired [65].

All of the programs discussed in this section can be used a database search queries using the Tripos database mining utility UNITY. UNITY allows 3D searching using queries not only from pharmacophores but also using surface and excluded volumes, queries derived from receptor binding sites and 2D substructure and similarity searching. The 3D flexible searching is based on the Directed Tweak method [66]. Conformational flexibility during searching with UNITY is handled on-the-fly. UNITY provides an interface to Oracle® and thereby relational database querying features [56].

2.4.1.2 **Catalyst**

Catalyst® [67] was launched 1992 by BioCAD (now Accelrys) as a tool for automated pharmacophore pattern recognition in a collection of compounds based on chemical features correlated with three-dimensional structure and biological activity data.

Catalyst models (hypotheses) consist of sets of abstract chemical features arranged at certain positions in the three-dimensional space. The feature definitions are designed to cover different types of interactions between ligand and target, e.g. hydrophobic, H-bond donor, H-bond acceptor, positive ionizable, negative ionizable. Except in some special cases, different chemical groups that lead to the same type of interaction, and thus to the same type of biological effect, are handled as equivalent. The directions of the H-bonds are usually determined and are given by vectors. Distinct chemical features in a particular conformation of a compound must be located within the tolerance constraints in order to satisfy the model. These models can be used directly as three-dimensional database search queries in the Catalyst environment.

The pharmacophore identification process as implemented in the Catalyst package involves 3D structure generation, followed by conformational search and definition of the pharmacophore points consistent with the training set.

Molecular structure editor

For the construction of molecular structures, a 2D formula editor is provided in combination with 3D conversion. Standard potential energy minimization is performed using the modified parameter set of the CHARMm force field [68]; the conformational models are built using Monte Carlo conformational analysis together with poling as described in the next section.

Conformational analysis in Catalyst

(i) Overview

While many common methods attempt to identify one global minimum energy conformation and other local minima as representative of the space, the approach to conformational analysis taken within Catalyst claims a broad coverage of bioaccessible conformational space of the molecules within a user-specified energy threshold. This implies that the representative conformers generated by Catalyst are not necessarily at local minima on the potential surface but are distributed widely over the space. This approach to conformational analysis is emerging from the consideration that, in many cases, the bound conformation of a small molecule to a receptor may not be the lowest energy conformation. Furthermore, the global minimum predicted by a force field could be incorrect owing to solvation effects or approximation errors in the force field.

A common difficulty accompanying the representation of the conformational space by sampling is the redundancy among conformers. Usually many hundreds or thousands of conformers are generated and then reprocessed to pick out families representative of the whole space. After a local minimization, many of these conformers may fall into the same conformation, reproduced several times. Therefore, Catalyst focuses on the coverage of all possible bioactive conformations of a compound compared with methods that represent conformational space as a collection (clusters) of local minima.

Another issue that should be addressed briefly is the relationship between size and resolution of a conformational model particularly in terms of coverage of the low-energy regions of the accessible conformational space. The coverage should at any rate be consistent with the precision of the application which uses the conformational model. During three-dimensional pharmacophore generation for database search purposes, the restriction is given by the tolerance of the pharmacophore query. It has been shown in principle that a limited number of conformers is sufficient to represent the low-energy conformational space of small- to medium-sized molecules [69, 70].

Catalyst addresses conformational flexibility by storing compounds as multiple conformers per molecule. Given that one has to generate and search through a very large number of conformers that may be in fact similar enough to can be treated as identical when mapped on to a pharmacophore hypothesis, the need for variation with a simultaneous reduction in the number of conformers becomes evident. The Poling algorithm of Smellie et al. [37] implemented in Catalyst is intended to solve many of these problems.

(ii) Conformational search in Catalyst: catConf/ConFirm

Two types of conformational search, BEST and FAST, are employed in Catalyst. Both methods emphasize adequate coverage of the conformational space, each with specific advantages. The FAST method delivers a reasonable model within a short time and is utilized primarily for database generation purposes, whereas the BEST method is intended to build more precise conformational models of molecules for hypothesis generation. Both methods use Poling by default, BEST for all molecules and FAST depending on the size and flexibility of the compounds in question. For smaller, less flexible compounds, the FAST method uses systematic search in the torsional space instead of Poling. Poling and various aspects of conformational search parameters are user adjustable and can be turned off if required. Stereochemistry is handled in an exhaustive manner with the options to specify explicit, relative and unknown chirality. Specified explicit and relative chirality will always be preserved during conformational search and pharmacophore analysis, whereas for compounds with chirality marked as unknown, mirror images will be considered unless this is not desired by the user.

Conformational models generated by other programs can be used for pharmacophore generation and in Catalyst databases by importing multiconformer structures stored, e.g., in SD file format.

Pharmacophore modeling with Catalyst

Catalyst provides two algorithms for automated pharmacophore arrangement search. HypoGen uses biological assay data (e.g. IC_{50} or K_i) to derive hypotheses that can predict quantitatively the activity of compounds, whereas HipHop seeks a common three-dimensional configuration of chemical features shared among a set of active molecules. In the case of HypoGen, similarly to 3D QSAR, all members of the training set must possess the same binding mode; the second method optionally allows automatic elimination of compounds that may have a different molecular site of action. The resulting models undergo a complex evaluation process by the program and the top scoring results are reported to the user.

(i) HipHop

The HipHop algorithm [71] attempts to produce an alignment of compounds expressing certain activity against a particular target and by superposition of diverse conformations to find common three-dimensional arrangements of features shared between them. Even though HipHop does not use activity data as input, it is a good idea to select highly active chemically diverse compounds when composing training sets whenever possible.

HipHop identifies common features by a pruned exhaustive search, starting with the simplest possible (two-feature) arrangements and expanding the model to three, four, five features and so on until no more common configurations can be found. This includes a search through two large spaces – the conformational space of the training set and the pharmacophore domain. HipHop does not need a particular reference conformation. If required, HipHop will attempt consecutively to align with each other all conformers of every training set member. Still, at least one molecule as the entire conformational model (principal compound) must be specified as a reference. Which exact conformer will then be present in the alignment depends on the remaining compounds and their conformational diversity and also on the conditions of the run.

First, the program identifies matches and distribution of the chosen features among the training set members, followed by the alignment procedure. The features are considered superimposed when of each of them lies within a specified distance (tolerance) from the ideal location, and at the same time the RMS deviation for the configuration as a whole is measured. The quantitative estimation of the goodness of match between a molecule and a configuration of features (Fit) can be pursued similarly to a scoring function to rank virtual screening results.

In the ideal case, superposition of all input molecules is desired. Sometimes it could be of advantage to permit some molecules, up to a specified number, to miss one, one particular or more than one of the features of a configuration in order to map all the remaining features. The benefit from such an option is that it allows one to work with compounds that may have a different binding mode or show activity in a particular assay as a result of an alternative mechanism of action or experimental errors.

In most cases, the result of a HipHop run will be numerous configurations of features so there is a need to score and rank them. For instance, the input molecules may often share feature arrangements widespread among drug molecules or there may be configurations common for the training set but rare in general. The ranking of the HipHop models is therefore based on rarity [71]. Maximizing the score of a configuration will minimize the probability that the training set molecules map the model by chance, making the pharmacophore specific

(ii) HypoGen

The HypoGen algorithm is designed to correlate structure and activity data for pharmacophore model generation.

HypoGen consists of three phases: constructive, subtractive and optimization. Generally, the constructive phase is similar to the proceeding of the HipHop algorithm. The training set is divided into two subsets, “active” and “inactive” compounds. First, all pharmacophores shared between the first two most active compounds are identified by overlaying systematically all their conformations, then only hypotheses that fit a minimum subset of features present in the remaining active compounds are kept.

In the subtractive phase, the program inspects the hypotheses already created and removes those most common to the inactive part of the training set. Compounds are considered inactive when their activity lies 3.5 logarithmic units (this value is user adjustable) below that of the most active compound.

The subtractive phase is followed by an optimization phase where simulated annealing is used to improve the predictive power of the hypotheses. Small changes are made to the models and they are scored according to the accuracy in activity estimation. Finally, the simplest models that correctly estimate activity are selected (Occam's Razor) and the top *N* solutions are reported to the user. The method has been described in more detail elsewhere [39, 40].

An important assumption that is made within both HipHop and HypoGen is that more contacts to the receptor and therefore more features per molecule lead to enhanced activity. It is well known from practice that often this is not true, e.g. large and feature-rich compounds may be barely active because of unfavorable steric interactions. An extension to the HypoGen algorithm, HypoRefine, is intended to help in solving this problem by placing the exclusion volume in key locations derived from atoms of well-fitting but inactive compounds. On the other hand, when insufficient activity or only HTS data are present, the HipHop Refine algorithm allows the use of "negative" information from inactive compounds matching the pharmacophore in order to generate a grid-based exclusion volume which eliminates false-positive vHTS hits and increases enrichment rates [72, 73].

(iii) Compound databases and database searching in Catalyst

Essentially there are two approaches to address the problem of conformational flexibility during pharmacophore screening: the use of multiple stored conformations and on-the-fly conformer calculation [74]. Catalyst offers a combination of both solutions within the Fast and Best Flexible Search algorithms. Catalyst databases consist of compounds stored as multiple conformations. When executing Fast Flexible Search, the search is performed using only conformations already existing in the database and Fast Search tries to find one fitting the pharmacophore among those available. The algorithm used with Best Flexible Search Databases/Spreadsheets can modify the conformation of a molecule during the computation to enforce a fit within a given energy threshold.

The database search process starts with a rapid screening process within which molecules possessing properties required from potential hits are sorted out from those that can be excluded *a priori*. The screen involves substructure match followed by screens matching three-dimensional pharmacophore features, molecular shapes or exclusion volumes and text constraints (1D properties) if present in the query (through Oracle). All this greatly reduces the number of potential hit compounds in the database. The next step of the search process tries a rigid fit of each conformation of each compound to the corresponding features. Compounds are selected as hits after the first successful mapping of all features and once all compounds have passed the procedure a hit list is obtained [40].

The Best database search first identifies all potentially suitable compounds by using loosened constraints, thus including those that would fail a rigid search. Within this preliminary list, the algorithm attempts to modify additionally the conformers so that they can fit the original query while remaining below a certain energy overflow [40]. The use of a Best search is justified when one has to deal with too small hit lists.

Once a hit list has been obtained, Catalyst provides the possibility to compute fit values that can be used for scoring.

Here, we consider it appropriate to mention briefly the so-called shape-based methods for flexible compound searching. Although these are not strictly seen in any relationship with the functional pharmacophore perception, the shape and size of compounds obviously influence activity and, in some cases, may be the main factors determining biological action.

A typical example of such a 'volume-searching' application is the shape-based methodology introduced by Hahn in 1997 [75]. Essentially, this approach involves the computation of the van der Waals surface enclosing a single or multiple structures and representation of the volume enclosed in this surface as a grid with a default size of 1 Å. The enclosed volume together with the surface represents the query. The searching procedure involves passing number of filtering constraints in order to identify quickly the most suitable molecules, whereas the actual match is done by comparing the Tanimoto similarity of the intersection divided by the sum of the volumes of the query and the target conformation of the candidate compound. The similarity score is computed after aligning the query and the candidate structure by their principal axis [75]. Earlier studies by the same authors describe Receptor Surface Models and their utility in QSAR analysis [76, 77].

Examples of volume-based approaches by other commercial software distributors are FlexS from Tripos [78, 79] and ROCS from OpenEye [80, 81].

5.4.1.3 **Phase**

Phase is the pharmacophore generation module provided by Schrödinger [82]. Like other modules available from Schrödinger, Phase uses the Maestro interface [83] as the visualization tool. Maestro provides a molecule sketcher and all the common molecular file formats are supported.

The pharmacophore generation module in Phase generates pharmacophore models using a four to five step procedure described below.

Ligand preparation

Molecule construction and 2D to 3D conversion are performed by using the LigPrep application in the Maestro modeling environment [84, 85]. Ionization at a given pH or neutralization, tautomer enumeration and stereoisomer enumeration are also supported. Stereoisomers can be treated either as being separate or identical molecules.

The molecule preparation step includes also conformational expansion using a torsional search or a combined Monte Carlo Multiple Minimum/Low Mode search. During the search, the intramolecular hydrogen bonds are not considered. Molecules can be minimized, OPLS-2005 or MMFF force fields [86, 87] are available, and also two continuum solvation models (distance-dependent dielectric or GB/SA). A double criterion is used to eliminate redundant conformations; it uses distances between pairs of corresponding atoms within a 1 kcal mol^{-1} energy window.

Using all compounds chosen to participate in a pharmacophore analysis, a molecular spreadsheet can be created and the user can manually select the molecules that will belong to the set that will define the reference pharmacophore space (active set).

Creating the pharmacophoric sites

Similarly to other software packages such as DISCO and Catalyst, Phase uses chemical features (hydrophobic, H-bond acceptors, H-bond donors, negative charge, positive charge, aromatic ring) to define the pharmacophore points called sites. These features are encoded in SMARTS and can be edited. H-bonding features are vectorized features (their directionality is considered).

Finding common pharmacophores

Using the sites defined in the previous step, pharmacophores common either to all or to a user-defined number of the selected active molecules will be generated, Phase uses a tree-based partitioning algorithm for that purpose, which places pharmacophore configurations in multi-dimensional boxes and groups them according to their inter-site distances The user has control over the size of the pharmacophore models (maximum number of features), and also the inter-pharmacophore point spacing. Pharmacophores containing between three and seven sites can be generated.

A given pharmacophore can be edited (feature addition or removal) and the excluded volume can be added in order to add some more information based on inactive molecules.

Scoring the pharmacophores

All ligands will be aligned on the models. The model ranking is performed using a user-weighted scoring function consisting of:

1. the quality of the alignment (RMSD in the site-point positions);
2. the definition of a vector score that measures the angle deviation (average cosine) between the vectorized features on the molecules;
3. the definition of a volume similarity (common/total) score based on the overlap of steric models of heavy atoms in each pair of molecules.

Partial mapping of the molecules on a pharmacophore model is allowed. At this stage, pharmacophore models and alignments can be visualized. Excluded volumes can be added manually after having aligned the inactive molecules on the pharmacophore models.

Building a QSAR model

The generation of a QSAR model is done as a post-processing step of pharmacophore generation. This is conceptually different from the Catalyst/HypoGen approach, in which SAR data are used actually to build the pharmacophore models, and this is reflected in these models. In the QSAR approach of Phase, molecular structures are aligned on the pharmacophore, a rectangular grid that encompasses the aligned molecules is created (generating uniformly sized cubes) and partial least-squares (PLS) is used for the regression. As in CoMFA, favorable and unfavorable regions can be visualized. Both atoms and pharmacophores can be used for the models.

Phase also has its own database management system, with the possibility of either storing single conformations for the molecules or storing different sets of conformations.

As part of the processing within this system, molecules can be cleaned (chirality, ionization). Conformers can also be generated on-the-fly when performing the database search. In addition to conformations, indexing of a database can be done by adding pharmacophore sites. Partial match of the hits on a pharmacophore query is allowed. The pharmacophore search hits are ranked using a fitness function.

2.4.1.4 Pharmacophores in MOE

MOE (Molecular Operating Environment) [88] is the modeling platform developed by the Chemical Computing Group. This platform allows access to different sets of computational tools ranging from bioinformatics, protein modeling, structure-based design to pharmacophore perception. All these applications have been integrated using the Scientific Vector Language (SVL) [89, 90].

The pharmacophore models built in MOE are qualitative. There is no possibility of using the SAR of a set of molecules in the building of the models.

The workflow that is used in MOE can be divided into four main steps:

1. generate annotations for all ligand conformations;
2. create a pharmacophore query;
3. database search;
4. edit the pharmacophore query for refinement and search the database again.

Generate annotations

In the MOE environment, molecules are stored in a database with their associated set of conformations. Several methods can be applied to expand the conformational space of organic molecules ranging from molecular dynamics to stochastic methods and systematic search. A fragment-based high-throughput methodology is provided for the construction of conformation databases. For each molecular conformation, an annotation can be generated using a so-called Pharmacophore Pre-Processor. The goal is to encode all the possible structural features (H-bond donors and acceptors including tautomers, anions and cations, including resonance forms and hydrophobic and aromatic areas) that describe the ligand's pharmacophore. This tool recognizes the different conformations of a molecule by molecular graph comparison. However, its use is optional and annotation can be performed during the database search (with the obvious consequence of increased search times). Molecules can be then visualized using the Database Viewer.

Create a pharmacophore query

The definition of pharmacophores is done manually by applying so-called schemes using a Pharmacophore Query Editor. A template molecule is generally used for this purpose. In the MOE environment, a scheme is a collection of functions that define how each ligand is annotated. This is accessed via an SVL function. The default scheme is called PCH (Polarity-Charged-Hydrophobicity). New schemes can be created to represent certain molecules better, e.g. Planar-Polar-Charged-Hydrophobicity [91].

If the structural information of a receptor is not available, molecule alignments can be performed using an all-atom flexible alignment procedure that combines a force field and a 3D similarity function based on Gaussian descriptions of shape and pharmacophore features to produce an ensemble of possible alignments of a collection of small molecules [92]. Pharmacophore queries can be derived from the resulting set of aligned conformations of known actives.

Currently, there is no automated tool in MOE that can generate pharmacophore models from a set of active/inactive molecules. As a consequence, there is no pharmacophore scoring or ranking or a validation method implemented in the program.

Database search

The so-generated pharmacophore is then used for database mining. In MOE, molecules are stored in databases with pre-calculated conformations. No new conformations are generated during a database mining experiment. Compounds are aligned with the query using a rigid-body superposition, with no flexible adjustment of the rotatable bonds. Full or partial mapping of the pharmacophore features can be obtained, with user control of the pharmacophore matching rules. Excluded volumes can be used to refine a query further.

Editing the pharmacophore query for refinement

The built-in query editor allows the user to refine a previously built pharmacophore model further.

2.4.2 Field-based Methods

As we have already mentioned in the Introduction to this chapter, we felt that we needed to include the traditional and well-validated field-based methods in this review. Our view is, however, that these methods do not fall directly into the classical definition of pharmacophores, which we rather associate with feature-based alignments and geometric constraints. This perception certainly involves a degree of oversimplification, yet it allows easier coverage of different conformational states, which may otherwise result in completely different fields. Furthermore, we do appreciate its relationship to the traditional understandings in medicinal chemistry and hence its helpfulness in the discovery process. On the other hand, the high complexity of 3D descriptors, the dependence on the binding mode and the alignment associated with the field-based methods makes these accurate but labor-intensive 3D QSAR methods less suitable in a virtual screening process, but undoubtedly useful tools for compound optimization.

In this review, we focus mainly on the classical CoMFA® methodology, but also mention other, more recent, developments.

2.4.2.1 CoMFA

Comparative Molecular Field Analysis (CoMFA) was introduced in 1988 and has since established itself as a recognized industry standard 3D QSAR method. It is patented and is commercially distributed by Tripos. As the name suggests, CoMFA uses molecular fields to characterize 3D structure–activity relationships within a set of molecules. The initial CoMFA studies involved two types of fields: steric derived from the Lennard–Jones potentials and electrostatic from the electrostatic potentials calculated against an ion probe. Later these were extended with several other types such as hydrogen bonding, indicator and parabolic fields, available within the Tripos Advanced CoMFA® module. More complex probes are used in GRID by Goodford [25]. Comparative molecular Shape analysis is similar to CoMFA but uses a Gaussian function for field assessment [65, 93].

The first, most important step in the CoMFA analysis workflow is the establishment of a meaningful molecular alignment hypothesis. However, this does not imply that the alignment should necessarily be very similar to the relative orientation of the ligands in a receptor binding pocket. This may seem confusing, but one has to be aware of the level of abstraction involved in this type of modeling and the purpose of the alignment. In the case of the typical ligand-based CoMFA method, the aim is not the close reproduction of binding modes, but finding the structural regions of the compounds in question suitable for modification in order to achieve an improved activity profile. In other words, the information here is gained based on the 3D structural comparison between ligands rather than a comparison between ligands and a receptor, hence a direct relationship to a receptor binding pocket should not always be expected. Similar considerations also apply to a certain extent to all ligand-based methods and to the pharmacophore modeling methods described in this chapter. In fact, alignments obtained from bound ligand conformations often lead to less predictive models [93].

Typically, the best alignments for CoMFA analysis are obtained manually by closely superimposing similar chemical groups. However, when working with conformationally flexible compounds, the alignment task may become a too complex or simply tedious task and automatic alignments may become methods of choice. These may be substructure-based or common feature-based pharmacophore alignments such as GASP or Catalyst.

The CoMFA field calculations are performed at each point in a typically rectangular grid with a spacing of 2 Å. The resulting field descriptors are used as input for PLS analysis in order to obtain a QSAR model [93]. PLS carries out regression using latent variables from the independent and dependent data that are along their axes of greatest variation and is typically applied when the independent variables are highly correlated or the number of independent variables exceeds the number of observations [94]. The resulting models are typically subjected to a validation procedure using either leave-one-out combinations of training sets or a completely external test data set [93].

2.4.2.2 XED

As an alternative to describing molecules by their structural features (substructural elements, functional groups) and similarly to CoMFA, this approach uses field points to describe the van der Waals and electrostatic minima and maxima that surround molecules and compares these field points. The field points that are used are derived from molecular electrostatic potential descriptors. The XED model is marketed by Cresset Biomolecular and forms the basis for the proprietary virtual screening technology FieldPrint^{TM} [95].

The eXtended Electron Distribution (XED) force field was first described by Vinter [96]. This force field proposes a different electrostatic treatment of molecules to that found in classical molecular mechanics methods. In classical methods, charges are placed on atomic centers, whereas the XED force field explicitly represents electron anisotropy as an expansion of point charges around each atom. The author claims that it successfully reproduces experimental aromatic π stacking. Later, others made similar observations [97]. This force field is now available in Cresset BioMolecular's software package [95]. Apaya et al. were the first to describe the applicability of electrostatic extrema values in drug design, on a set of PDE III inhibitors [98].

Conformational expansion of molecules (also called conformation hunting in Cresset's XedeX^{TM} software module) applies a Monte Carlo approach combined with fast molecular dynamics for ring conformations. The minimization of the conformations is done using the XED force field, in order to assign correct charges. Based on the results obtained by Boström [99], this method performs comparably to other available methods when considering the RMS difference between the bound conformation and the closest conformation found considering the number of conformations found with an RMS value between 0.0 and 1.0 Å [99].

Three types of field points can be calculated with XED: positive and negative extrema and van der Waals points (also called "sticky" points). These points are calculated by moving probes on a grid of points placed above the van der Waals molecular surface. Extrema values are found using a 3D simplex algorithm and coincident positions are filtered out [100]. The field points are color coded and their radius reflects the depth of the energy well. Pairwise molecule comparison can be performed by using these field points only. A score reflecting the degree of similarity of the two sets of field points is calculated. This avoids having to pre-align the molecules as is the case for other field-based methods (CoMFA). As an extension to this, Cresset developed the technology FieldPrint [101] to encode a molecule's complex 3D field pattern in a 1D string and store it in a database. This database can be searched with the field print of any molecule and retrieve compounds that do not necessarily belong to the same chemical class. Cresset's database contains over 1 500 000 distinct commercially available compounds [102].

A recent paper illustrates the use of this technology to design pyrrole- and imidazole-based CCK2 antagonists [97].

2.4.3
Pharmacophore Fingerprints

Here we define pharmacophore fingerprints as the binary encoded information (key) about the presence or absence of pharmacophore features and distances in a single molecule or a compound collection. This concept can be extended to include the occurrence counts of distinct pharmacophores. Usually the focus is put on two to four point fingerprints but larger number can be used and the utilization of up to nine point pharmacophores has been described [6]. Pharmacophore triplets are widely used as traditionally they have been considered to be most effective in terms of information content versus complexity. The pharmacophore space is binned and the method of binning and the bin size are of significant importance. The most common application of pharmacophore fingerprints is in the area of diversity and similarity calculations, compound library focusing and selection, but 3D pharmacophore descriptors can also be used for the analysis of structure–activity relationships, in decision trees and QSAR. Fingerprint focusing methods commonly use similarity coefficients such as Tanimoto to retrieve or classify compounds of interest out of a typically large collection.

2.4.3.1 **ChemX/ChemDiverse, PharmPrint, OSPPREYS, 3D Keys, Tuplets**

Numerous examples of 3D fingerprint methods have been described in the literature, but in this review we focus only on those which can be classified as software packages or parts of them.

One of the most popular applications is ChemX/ChemDiverse of Chemical Design/Oxford Molecular (now Accelrys). Details of the approaches used there are included in this review by Mason et al. [3]. Another example, of an in-house pharmacophore fingerprint construction, is PharmPrint by Affymax [4, 5].

The Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS) approach, introduced by Martin and Hoeffel [6], is in software terms an extension of CCG's MOE package, written using SVL. The 3D oriented substituent pharmacophores are aimed towards better representation of diversity and similarity in combinatorial libraries in the 3D pharmacophore space. Combinatorial library design often operates only on substituents rather than on the final products as the complications related to the conformational coverage in the 3D space and the scaffold dependency limit the product-based approaches to smaller libraries. The 3D oriented substituent pharmacophores add two more points and the corresponding distances to each substituent pharmacophore which represent the relationship of the substituents in the product with only little additional information. The fingerprints permit the creation of property space by multidimensional scaling (MDS) and, since scaffold independent, can be stored separately and applied to different libraries [6].

The Accelrys implementation of pharmacophore fingerprint descriptors is called 3D Keys. This application is based on standard Catalyst feature definitions and is a part of the Cerius^2 software package [7].

The collection of all combinations of three (triplets) or four (quadruplets) features in 3D space over all conformations of all compounds in the supplied data set is computed. Each triplet or quadruplet is characterized by a set of feature types and the corresponding inter-feature distances. Optionally, appearance counts can be included in a fingerprint. Using these fingerprints, the property space of molecules can explored on the basis of pharmacophore diversity after MDS. These fingerprint descriptors can be used for diverse and similar selections, clustering, library comparison and optimization or applied to decision trees and QSAR. Finally, relevant pharmacophore hypotheses can be extracted from the keys and used for database mining. 3D Keys can be derived both from small molecules and from three-dimensional receptor binding site features.

Another, similar application is Tripos Tuplets, which handles two to four point fingerprints, with the option of requesting the presence of certain features or substructures in the fingerprint. Tuplets can be used for clustering, can provide the basis for similarity selections and can utilize both ligand and target information. Tuplets can be applied for the purpose of identifying alternative binding modes as well as for deriving hypotheses from compounds, UNITY queries or binding pockets, which then can be analyzed using multiple similarity measurements [8].

2.5 Other Methods

2.5.1 SCAMPI

Most of the above-mentioned pharmacophore generation techniques use a small number of user selected molecules, commonly called a dataset, to derive the pharmacophore models. With the advent of high-throughput methods [103], there was a need to extract pharmacophore information from much larger datasets.

SCAMPI (Statistical Classification of Activities of Molecules for Pharmacophore Identification) is a program developed in C language by Chen et al. [104]. According to the authors, it allows the use of datasets of approximately 1000–2000 compounds. The SCAMPI program's implementation has been done to allow users to visualize the molecules and the generated pharmacophores in the Sybyl environment.

Two different, but connected, spaces are searched by the program:

1. the conformational space, representing all possible conformations of the compounds;
2. the correspondence space, representing all the possible correspondences of chemical features and configurations among different compounds.

As opposed to other pharmacophore generation methods that treat the conformer expansion and pharmacophore identification phases separately, SCAMPI combines the two searches and lets them depend on each other. Figure 2.2 illustrates the workflow used by SCAMPI.

SCAMPI reads multiple MOL2 files containing structures and a data file containing the biological activities. The conformational expansion of the molecules is done by applying random search techniques, with no post-clustering. This search is performed in Cartesian and internal coordinate space.

The pharmacophore points are represented by chemical features, in addition to specific atoms such as nitrogen, oxygen, sulfur, phosphorus, fluorine and other halogens). The correspondence search uses a recursive partitioning algorithm, comparable to the FIRM and SCAM programs [104]. The split criterion used by SCAMPI to partition the whole data set in multiple subsets uses a Student's *t*-test corrected by the Bonferroni *p*-value. The test is based on the presence or absence of a feature also called molecular descriptor. The absence of a feature means that this feature could not be identified in any of the generated conformations. The molecular descriptor that gives the highest *t* value is the one selected for the split. Both a substructure and a rule-based search method have been implemented for the detection of features represented by groups of atoms and features represented by single atoms.

The pharmacophore build-up procedure is similar to that in Catalyst HipHop. Two-point pharmacophores characterized by the two features and a binned dis-

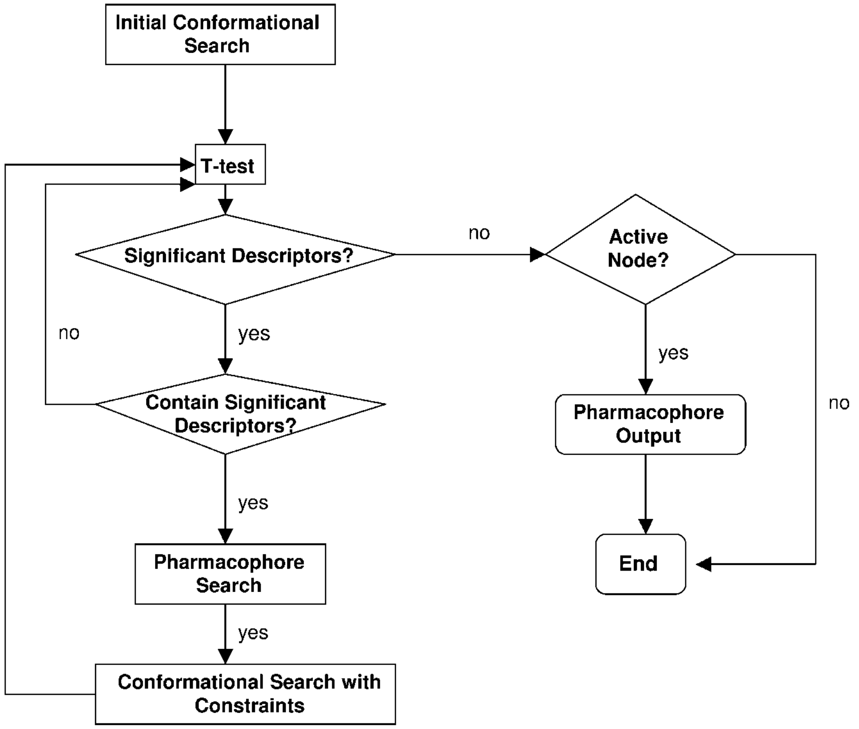


Fig. 2.2 SCAMPI workflow [104].

tance are searched first. A new point is added only if found, and the process continues until no more pharmacophore points can be found. Pharmacophores are already recorded in the conformer generation phase. The activity of molecules is handled in a semi-quantitative manner.

The authors illustrated the approach by using two datasets: 1650 MAO inhibitors from Abbott and 114 ACE inhibitors. The pharmacophores identified by the program match some known SAR, especially the multiple mechanism of action in the MAO series [104].

2.5.2 THINK

THINK (To Have Information aNd Knowledge) is a modular system developed by Treweren Consultants [105] to assist with lead generation and optimization. This system allows structure-based virtual screening, data analysis and pharmacophore profiling and is organized in different modules. Around a core module that provides chemical structure reading and writing, command scripts for batch and server jobs, there is a 2D module for data analysis and *de novo* derivative generation capabilities, a 3D module for 3D coordinate generation and conformer generation, a pharmacophore module for pharmacophore perception, a Microsoft GUI module only available for the Microsoft® Windows version of the program and a Screening Database module consisting of Treweren's current collection of drug-like molecules.

In THINK, molecules can be built using a 2D editor and the program reads MOL, SD, SMILES and PDB files. Three-dimensional coordinates of molecules, when not available from the input file, are generated automatically by the program itself.

Two classical methods are available in THINK to perform the conformational expansion of molecules: systematic search and random search. When the systematic search option is used, the use of contacts check avoids high-energy conformations and reduces the overall processing time. The random method uses a random number generator to select the conformations from within the estimated total number of conformations. The implementation of the program does not prohibit identical conformations to be output resulting from symmetry. These conformations are used in the pharmacophore generation and site search modules.

The so-called pharmacophore centers use classical chemical functions such as donors, acceptors, acids, bases, hydrophobic and positive and negative charges functions. Metal ions and electron donor lone pairs are possible centers. The users can also define their own functions.

THINK considers fuzzy two-, three- or four-center pharmacophores. If a given molecule contains more than three or four centers, then all possible groups of two to four centers are taken. The distances (including a tolerance) between the pharmacophore centers are measured exactly and then allocated to distance bins, each distance being represented by the bin into whose range it falls. The distance bins are used to transform the distances within each pharmacophore into a set of integers that give a more compact representation of the pharmacophores.

For example, let us consider a two-center pharmacophore that has a distance of 4.85 Å between the centers. Applying a distance tolerance of 0.25 Å gives a range of possible distances between 4.6 and 5.1 Å. The default distance bins that cover this range are 4.0–5.0 and 5.0–6.0 Å: 80% of the distance range lies within the 4.0–5.0 Å bin and 20% within the 5.0–6.0 Å bin. Hence this two-center pharmacophore would generate two fractional pharmacophores, one with a count of 0.8 and the other with a count of 0.2.

Pharmacophore profiles are defined that represent the set of all the pharmacophores found across the conformers of a series of conformers or series of molecules. Each pharmacophore added to the profile has to be unique. This profile will help in showing the spread of pharmacophores across the conformational space of a molecule or a series of molecules. No sum of the exhibited pharmacophores or normalization is done. There is no direct graphical representation of pharmacophore models. The pharmacophores can be saved to a file in CSV format that can be imported into a MySQL or Oracle database. This approach permits the use of standard SQL queries to extract common pharmacophores within sets of molecules, helping to discriminate between active and inactive compounds.

The Receptor Site search module of THINK uses these pharmacophores to eliminate quickly conformations of molecules that cannot bind to a receptor site.

2.5.3 Feature Trees

Feature trees have been described by Rarey and Dixon [106] as a new way of analyzing the similarity of molecules. This approach is based on building trees that represent molecules. These trees describe the major building blocks of molecules, in addition to their overall arrangement. They are conformation independent. Different types of pairwise comparison algorithms are available to compare trees of different molecules.

2.5.4 ILP

Inductive logic programming (ILP) is not a pharmacophore generation method by itself, but a subfield of the machine learning approach. In this field, other methods such as hidden Markov models, Bayesian learning, decision trees and logic programs are available.

Sternberg and Muggleton described the use of ILP to analyze the SAR of a series of 28 ACE inhibitors [108]. This type of approach learns from observations (examples) which often are chemical structures. Both active and inactive molecules can be used, since each of them will help in defining rules. Properties such as hydrophobicity, chemical connectivity and spatial relationships can be encoded. An algorithm will then identify the property combinations that cover most of the actives while covering the smallest number of inactive molecules. The resulting rule can be saved and used either for refinement (with a new set of active molecules not used in the first instance) or for prediction. No conformer generation or alignment of molecules is necessary in order to formulate the rule. However, this approach does not handle numerical calculations so quantitative SAR cannot be modeled and only semi-quantitative (class-based) models can be derived.

2.6 Conclusions

In this chapter, we have tried to demonstrate the great diversity of software tools available to the researcher in the area of ligand-based pharmacophore modeling. With the expansion of combinatorial chemistry techniques and the need to manipulate very large amounts of real or virtual chemical data, pharmacophore-based techniques have proved their potential in the areas of database mining with pharmacophore queries and library design using pharmacophore fingerprints.

A lot of effort has been invested over the past 20 years in the optimization of the different steps of pharmacophore generation: molecular editing and 3D representation, combinatorial enumeration, conformational expansion and pharmacophore perception methodologies for small drug-like data sets. However, we note that today there are still some areas with potential for improvement in the field of ligand-based pharmacophore modeling:

- Validation: most of the available packages only approach validation from a given angle. The problems of validation are addressed elsewhere in this book [42].
- Chemical space coverage: it can be considered a limitation of the majority of today's ligand-based approaches that only small-sized sets of chemical structures – training (learning) sets – are used to derive pharmacophore models. Consequently, these learning sets cover only a small portion of the chemical space and the performance of the resulting models generally tends to decrease when evaluating large datasets or within other chemical classes of compounds. As there is no unique answer to complex problems such as multiple independent data, large and diverse datasets or receptor flexibility issues, so-called ensemble pharmacophores consisting of multiple models generated from different subsets of large sets of chemical structures could represent an approach that should be pursued in the future.

References

1 Ehrlich, P., *Ber. Dtsch. Chem. Ges.*, **1909**, *42*, 17.

2 Guner, O. F., *Curr. Top. Med. Chem.*, **2002**, *2*, 13121–1332.

3 Mason, J. S., Good, A. C., Martin, E. J., *Curr. Pharm. Des.*, **2001**, *7*, 567–597.

4 McGregor, M. J., Muskal S. M., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 569–574.

5 McGregor, M. J., Muskal, S. M., *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 117–125.

6 Martin, E. J., Hoeffel, T. J., *J. Mol. Graph. Model.*, **2000**, *18*, 383–403.

7 *Cerius^2*. Accelrys Software, San Diego, CA; http://www.accelrys.com/.

8 *Tuplets*. Tripos, St. Louis, MO; http://www.tripos.com/

9 Van Drie, J. H., Weininger, D., Martin, Y. C., *J. Comput.-Aided Mol. Des.*, **1989**, *3*, 225–251.

10 VanDrie, J. H., *J. Comput.-Aided Mol. Des.*, **1996**, *10*, 623.

11 VanDrie, J. H., *J. Comput.-Aided Mol. Des.*, **1997**, *11*, 39–52.

12 Van Drie, J. H., *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 38–42.

13 Van Drie, J. H., Nugent, R. A., *SAR QSAR Environ. Res.*, **1998**, *9*, 1–21.

14 Snyder, J. P., Rao, S. N., Koehler, K. F., Vedani, A., Pellicciari, R., in *Trends in QSAR and Molecular Modeling*, C. G. Wermuth, Y. Rival (eds.). **1993**, Elsevier: Amsterdam. p. 367–403.

15 Finn, P. W., Karraki, L. E., Latombe, J.-C., Motwani, R., Shelton, C., Venkatasubramanian, S., Yao, A., RAPID: randomized pharmacophore identification for drug design, in: *Proc of the 3rd annual symposium on computational geometry*, Nice, France, ACM Press, p. 324–333.

16 *SCREEN*. http://www.jchem.com/index.html?content=doc/user/Screen.html.

17 Good, A.C., Cheney, D.L., *J. Mol. Graph. Model.*, **2003**, *22*, 23–30.

18 Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B., Marshall, G.R, *J. Comput.-Aided Mol. Des.*, **1995**, *9*, 491–499.

19 Marshall, G.R., Barry, C.D., Bosshard, H.D., Dammkoehler, R.D., Dunn, D.A., *Comput.-Assist. Drug Des.*, **1979**, *112*, 205–222.

20 Klebe, G., in *3D QSAR in Drug Design. Theory, Methods and Applications*, H. Kubinyi (ed.). ESCOM, Leiden, **1993**, pp. 173–199.

21 Bures, M.G., in: *Practical Application of Computer-Aided Drug Design*, P.S. Charifson (ed.). Marcel Dekker, New York, **1997**, pp. 39–72.

22 Melani, F., Gratteri, P., Adamo, M., Bonaccini, C., *J. Med. Chem.*, **2003**, *46*, 1359–1371.

23 Brint, A.T., Willett, P., *J. Comput.-Aided Mol. Des.*, **1989**, *2*, 311–320.

24 Gardiner, E.J., Artymiuk, P.J., Willett, P., *J. Mol. Graph. Model.*, **1997**, *15*, 245–253.

25 Goodford, P., *J. Med. Chem.*, **1985**, *28*, 849–857.

26 Nissink, J.W.M., Verdonk, M.L., Kroon, J. Mietzner, T., Klebe, G., *J. Comput. Chem.*, **1997**, *18*, 638–645.

27 Barbany, M., Gutiérrez-de-Terán, H. Sanz, F., Villà-Freixa, J., *Proteins: Struct. Funct. Bioinf.*, **2004**, *56*, 585–594.

28 Mestres, J., Rohrer, D.C., Maggiora, G.M., *J. Comput. Chem.*, **1997**, *18*, 934–954.

29 Pitman, M.C., Huber, W.K., Horn, H., Krämer, A., Rice, J.E. Swope, W.C., *J. Comput.-Aided Mol. Des.*, **2001**, *15*, 587–612.

30 Mills, J.E.J., de Esch, I.J.P., Perkins, T.D.J., Dean, P.M., *J. Comput.-Aided Mol. Des.*, **2001**, *15*, 81–96.

31 Hurst, T., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 190–196.

32 Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M,P., *Science*, **1983**, *220*, 671–680.

33 Cerny, V., *J. Optim. Theory Appl.*, **1985**, *45*, 41–51.

34 Spendley, W., Hext, G.R., Himsworth, F.R., *Technometrics*, **1962**, 441–461.

35 Jones, G., Willett, P., Glen, R.C., *J. Comput.-Aided Mol. Des.*, **1995**, *9*, 532–549.

36 Handschuh, S., Wagener, M., Gasteiger, J., *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 220–232.

37 Smellie, A., Teig, S., Towbin P., *J. Comput. Chem.*, **1994**, *16*, 171–187.

38 Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A., Dunn, D.A., in *Computer-Assisted Drug Design*, E.C. Olson, R.E. Christoffersen (eds.) American Chemical Society, Washington, DC, **1979**, p. 205–226.

39 Guner, O.F. (ed.), *Pharmacophore Perception, Development and Use in Drug Design.* IUL Biotechnology Series, Vol. 1. IUL, La Jolla, CA, **2000**.

40 Kugori, Y., Güner, O.F., *Curr. Med. Chem.*, **2001**, *8*, 1035–1055.

41 Golender, V., Vesterman, B, Vorpagel, E., *APEX-3D Expert System for Drug Design*; http://www.netsci.org/Science/Compchem/feature09.html.

42 Triballeau, N., Bertrand, H.-O., Acher, F., in *Pharmacophores and Pharmacophore Searches,* T. Langer and R. Hoffmann (eds.), Wiley-VCH, Weinheim, **2006**, p. 325–362.

43 Triballeau, N., Acher, F., Brabert, I., Pin J.-P., Bertrand, H.-O., *J. Med. Chem.*, **2005**, *48*, 2534–2547.

44 Krovat, E.M., Fruhwirth, K.H., Langer, T., *J. Chem. Inf. Model.*, **2005**, *45*, 146–159.

45 Lengauer, T., Lemmen, C., Rarey, M., Zimmermann, M., *Drug Discov. Today*, **2004**, *9*, 27–34.

46 Rollinger, J.M., Haupt, S., Stuppner, H., Langer, T., *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 480–488.

47 Steindl, T.M., Crump, C.E., Hayden, F.G., Langer, T., *J. Med. Chem.*, **2005**, *48*, 6250–6260.

48 Good, A.C., Cho, S.J., Mason, J.S., *J. Comput.-Aided Mol. Des.*, **2004**, *7*, 523–527.

49 Christmann-Franck, S., Bertrand, H.-O., Goupil-Lamy, A., der Garabedian, P.A., Mauffret, O., Hoffmann, R., Fermandjian, S., *J. Med. Chem.*, **2004**, 47*,* 6840–6853.

50 Evers, A., Hessler, G., Matter, H., Klabunde, T., *J. Med. Chem.*, **2005**, *48*, 5448–5465.

51 Klabunde, T., Hessler, G., *ChemBiochem.*, **2002**, *3*, 928–944.

52 *Sybyl*. Tripos, St. Louis, MO; http://www.tripos.com/.

53 Connolly Martin, Y., in *Pharmacophore Perception, Development and Use in Drug Design*, O. Guner (ed.). IUL Biotechnology Series, Vol. 1. IUL, La Jolla, CA, **2000**, p. 49–68.

54 Patel, Y., Gillet, V.J., Bravi, G., Leach, A.R., *J. Comput.-Aided Mol. Des.*, **2002**, *16*, 653–681.

55 *DISCOtech*. Tripos, St. Louis, MO; http://www.tripos.com/.

56 *Unity*. Tripos, St. Louis, MO; http://www.tripos.com/.

57 Jung, D.F., J., Gund, T.M., *J. Comput. Chem.*, **2004**, *25*, 1385–1399.

58 Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor R., *J. Mol. Biol.*, **1997**, *267*, 727–748.

59 *GASP*. Tripos, St. Louis, MO; http://www.tripos.com/.

60 Jones, G., Willett, P., Glen, R.C., in *Pharmacophore Perception, Development and Use in Drug Design*, O. Guner (ed.). IUL Biotechnology Series, Vol. 1. IUL, La Jolla, CA, **2000**, p. 87–106.

61 Yuan, H., Kozikowski, A.P., Petukhov, P.A., *J. Med. Chem.*, **2004**, *47*, 6137–6143.

62 *GALAHAD*. Tripos, St. Louis, MO; http://www.tripos.com/.

63 Cottrell, S.J., Gillet, V.J., Taylor, R., Wilton, D., *J.Comput.-Aided Mol. Des.*, **2004**, *18*, 665–682.

64 Richmond, N.J., Willet, P., Clark, R.D., *J. Mol. Graph. Model.*, **2004**, *23*, 199–209.

65 *GALAHAD*, Tripos, St. Louis, MO; http://www.tripos.com/data/SYBYL/GALAHAD_9-7-05.pdf.

66 Hurst, T., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 190–196.

67 *Catalyst*. Accelrys Software, San Diego, CA; http://www.accelrys.com.

68 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., *J. Comput. Chem.*, **1983**, *4*, 187–217.

69 Smellie, A., Kahn, S.D, Teig, S.L., *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 285–294.

70 Smellie, A.K., Kahn, S.D, Teig, S.L, *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 295–304.

71 Barnum, D., Greene, J, Smellie, A., Sprague, P., *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 563–571.

72 Maynard, A.J., HypoGenRefine and HipHopRefine: pharmacophore refinement using steric information from inactive compounds. Presented at the ACS National Meeting, Spring, **2004**.

73 Toba, S., Srinivasan, J., Maynard, A.J., Sutter, J., *J. Chem. Inf. Model.*, **2005**, *46*, 728–735.

74 Sprague, P., Hoffmann, R., in *Computer-Assisted Lead Finding and Optimization*, B.T. H. van de Waterbeemd, G. Folkers (eds.). Wiley-VCH, Weinheim, **1997**, pp. 225–240.

75 Hahn, M., *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 80–86.

76 Hahn, M., *J. Med. Chem.*, **1995**, *38*, 2080–2090.

77 Hahn, M., Rogers, D., *J. Med. Chem.*, **1995**, *38*, 2091–2102.

78 *FlexS*. Tripos, St. Louis, MO; http://www.tripos.com/.

79 Lemmen, C., Lengauer, T., Klebe, G., *J. Med. Chem.*, **1998**, *41*, 4502–4520.

80 *ROCS*. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

81 *Grant*, J.A., Gallardo, M.A., Pickup, B., *J. Comput. Chem.*, **1996**, *17*, 1653.

82 *Phase*. Schrödinger, Portland, OR; http://www.schrodinger.com/.

83 *Maestro*. Schrödinger, Portland; OR, http://www.schrodinger.com/.

84 *LigPre*. Schrödinger, Portland; OR, http://www.schroedinger.com/.

85 *Maestro*. Schrödinger, Portland; OR, http://www.schroedinger.com/.

86 Jorgensen, W.L., Maxwell, D.S. Tirado-Rives, J., *J. Am.Chem. Soc.*, **1996**, *118*, 11225–11236.

87 Halgren, T.A., *J. Comput. Chem.*, **1998**, *17*, 490–519.

88 *Molecular Operating Environment*. Chemical Computing Group, Montreal, QC; http://www.chemcomp.com/.

89 CCG: Methodology Development and Deployment; http://www.chemcomp.com/software-mdd.htm.

90 SVL Exchange; http://svl.chemcomp.com/.

91 Lin, A., *Overview of pharmacophore applications in MOE*; http://www.chemcomp.com/journal/ph4.htm.

92 Labute, P., *Flexible alignment of small molecules*; http://www.chemcomp.com/journal/malign.htm.

93 Clark, R. D., Leonard, J. M., Strizhev, A., in *Pharmacophore Perception, Development and Use in Drug Design*, O. Guner (ed.). IUL Biotechnology Series, Vol. 1. IUL, La Jolla, CA, **2000**, p. 151–167.

94 Livingstone, D., *Data Analysis for Chemists*. Oxford University Press, Oxford, **1995**.

95 *eXtended Electron Distribution (XED)*. Cresset BioMolecular Discovery, Letchworth; http://www.cresset-bmd.com/.

96 Vinter, J.G., *J. Comput.-Aided Mol. Des.*, **1994**, *8*, 653–668.

97 Chessari, G., Hunter, C. A., Low, C. M., Packer, M. J., Vinter, J. G., Zonta, C., *Chem. Eur. J.*, **2002**, *8*, 2860–2867.

98 Apaya, R. P., Lucchese, B., Price, S. L., Vinter, J. G., *J Comput.-Aided Mol. Des.*, **1995**, *9*, 33–43.

99 Boström, J., *J. Comput.-Aided Mol. Des.*, **2001**, *15*, 1137–1152.

100 Vinter, J. G., Trollope, K. I., *J. Comput.-Aided Mol. Des.*, **1995**, *9*, 297–307.

101 *FieldPrint*$^{\text{TM}}$. Cresset Biomolecular Discovery, Letchworth; http://www.cresset-bmd.com/.

102 Cresset Database, Cresset BioMolecular Discovery, Letchworth; http://www.cressetbmd.com/technology5.html.

103 Sittampalam, G. S., Kahl, S. D., Janzen, W. P., *Curr. Opin. Chem. Biol.*, **1997**, *1*, 384–391.

104 Chen, X., Rusinko, A., Tropsha, A., Young S., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 887–896.

105 *THINK*. Treweren Consultants, Eversham; http://www.treweren.com/.

106 Rarey, M., Dixon, J. S., *J. Comput.-Aided Mol. Des.*, **1998**, *12*, 471–490.

107 Sternberg, M. J. E., Muggleton, S. H., *QSAR Comb. Sci.*, **2003**, *22*, 527–532.

3
Alignment-free Pharmacophore Patterns – A Correlation-vector Approach*

Steffen Renner, Uli Fechner, and Gisbert Schneider

3.1
Introduction

According to the Medicinal Chemistry Section of IUPAC, a pharmacophore is the "ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" [1]. The concept of a pharmacophore regards ligand–receptor interactions as a function of individual functional group contributions. The respective functional groups are also termed "potential pharmacophore points" (PPPs), to stress that we do not know *a priori* which functional group actually contributes to the ligand–receptor interaction. A pharmacophore thus reflects the way medicinal chemists characterize the binding ability of molecular structures to a particular biological target. As ligand–receptor interactions take place in the three-dimensional (3D) space, pharmacophore models that are based on observed 3D interaction patterns represent the most intuitive choice. However, if we do not know a receptor-relevant ligand conformation or conformation ensemble, quantitative structure–activity relationship (QSAR) studies that are based on 3D pharmacophore models can be misleading [2]. In addition to the process of conformer generation, an often limiting time-consuming step in pharmacophore matching methods is the 3D alignment of molecular features, e.g. matching a screening molecule to a given pharmacophore model. To permit rapid database searching, the explicit alignment step can be avoided by an *alignment-free* representation of pharmacophoric patterns. The idea is to convert the spatial or topological distribution of PPPs and other molecular features taking account of the shape and surface electrostatic properties to a vector representation. Such vectors are referred to as "fingerprints", "bitstrings", "correlation vectors" (CVs), or "spectra", depending on the type of information stored and the particular method that was used for their generation. The trick is to compare these reduced molecular representations instead of explicit 3D feature alignment. By this, significant reduction of calcula-

* Please find a "List of Abbreviations" at the end of this chapter.

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

tion time is gained for virtually screening large compound libraries for potential ligands. In this chapter we will highlight the correlation-vector approach as an example of alignment-free pharmacophore pattern matching.

The use of atom pairs as the basis of molecular descriptors has a long history in the field of cheminformatics and a pioneering publication dates back to 1985 [3]. An atom-pair descriptor encodes information about a molecule by enumerating all possible combinations of two atoms and their pairwise distances. Atoms are characterized by one or more selected properties such as element type, hybridization state, partial charge or pharmacophore properties. The distance between a pair of atoms can be measured either in bonds (topological distance) or as a through-space distance (topographical distance). Well-known descriptors consider atom triplets or even atom quartets instead of atom pairs, as detailed elsewhere in this book (for an extensive recent review of pharmacophore concepts, see, e.g., Ref. [4]). PPP triplets allow for a more detailed representation of the distribution of PPPs than pairs of PPPs. PPP quartets additionally consider the chirality of molecules, which is not possible when relying on PPP pairs. However, the number of features increases significantly with the consideration of triplets and quartets. This inevitably leads to the mere binary count of features (bitstring vector) because of storage and computational limitations, which in turn requires careful maximization of their signal-to-noise ratio [5]. However, even the computation of an atom-pair descriptor typically yields a high-dimensional vector. The dimension of this vector depends on the number of features that are assigned to atoms or PPPs and the handling of distances. Since the topological or topographical atom-pair properties of a molecular structure are mapped on a vector, the calculation of similarity between two structures is the mere calculation of similarity between the two respective vectors. In other words, similarity calculations based on atom-pair descriptors do not require an explicit alignment of structures. The alignment-free similarity calculation of atom-pair descriptors renders a fast virtual screening campaign of huge databases possible.

Pharmacophore descriptors do not consider the element types of ligand atoms but their generalization according to potential interaction types with a biological target. This generalization paves the way for an interesting facet of pharmacophore descriptors: Their inherent suitability for "scaffold hopping". Scaffold hopping or "lead hopping" is the identification of isofunctional structures with different backbone architectures [6]. The ability to move to new scaffolds during the drug development process may be desirable owing to, for example, intellectual property constraints, poor synthetic accessibility or pharmacological profile of a lead compound. The CATS (*c*hemically *a*dvanced *t*emplate *s*earch) descriptor, a topological pharmacophore descriptor, was originally developed with its proposed scaffold-hopping ability in mind [6], and will be reviewed in detail here. Since pharmacophore descriptors are often characterized by their ability for scaffold-hopping [6–8], a discussion of this aspect is part of this chapter.

3.2 The Correlation-vector Approach

3.2.1 The Concept

Spatial autocorrelation is a quantitative measure of the probability of finding objects of defined properties within a distance of interest [9, 10]. The concept of autocorrelation is mainly applied in fields such as geography, economics, ecology or meteorology to describe the spatial distribution of features. The idea of a molecular descriptor based on the autocorrelation concept was first introduced into the field of cheminformatics by Moreau and Broto in 1980 [11] with the ATS (*a*utocorrelation of a *t*opological *s*tructure) descriptor. For this approach, the atoms of a molecule were represented by properties such as atomic mass or partial charge. The distance between atoms was measured as the number of bonds between the respective atoms (topological distance).

The ATS descriptor for a given topological distance *d*, ATS_d, is calculated by

$$ATS_d = \sum_{i=1}^{A} \sum_{i=1}^{A} \delta_{ij,d}(w_i w_j) \tag{1}$$

where w is the atomic property, A is the number of atoms in the molecule and $\delta_{\mathrm{ij,d}}$ (Kronecker delta) = 1 for all pairs of atoms with distance d.

To obtain the full descriptor, the ATS autocorrelation is calculated over all defined distances and concatenated to a vector $\{ATS_0,\ ATS_1,\ ATS_2,\ \ldots,\ ATS_D\}$, where D is the maximum distance considered. Moreau et al. [12] were the first to apply this approach to the three-dimensional conformation of a molecule. For the 3D approach, the topological distance was replaced by the spatial Euclidean distance between two atoms. Pairs of atoms were clustered into groups with distances falling into predefined distance ranges (bins). All atom pairs within one bin were treated as having the same distance. Gasteiger and coworkers extended this approach to the spatial autocorrelation of the partial charges calculated for surface points [9, 13]. The resulting vector values were normalized by dividing the raw counts by the number of atom pairs in each distance range.

In 2000, Pastor et al. [14] presented GRIND (*gr*id-*in*dependent *d*escriptors), an approach very similar to the autocorrelation descriptors. The GRIND descriptor is calculated from force field-based interaction energies calculated for GRID [15] points surrounding a molecule. Instead of summing up all products of interaction energies for pairs of GRID points within a distance range, only the most favorable energy contribution is stored for each distance range. Given a descriptor vector, pairs of grid points can be identified that are sensible for each descriptor value. Such a trace back from the descriptor to the underlying pairs of grid points is usually not amenable to other autocorrelation approaches.

In 1985, Carhart et al. [3] introduced a topological atom-pair descriptor using atom types instead of atom property values: each atom is assigned to one atom type class instead of an atom property value. Atom types are defined by their element, the number of neighboring non-hydrogen atoms and their number of π-electrons. The employment of these atom types leads to a distinction of chemical elements according to the atom environment. Binary values are assigned to each atom, i.e. an atom does or does not belong to a specific atom type. Consequently and in contrast to the Moreau–Broto approach, the resulting autocorrelation vector for an atom type is equivalent to a histogram counting the frequencies of the atom pairs of the considered atom type over the different atom-atom distances. Calculation of the autocorrelation between pairs of atoms of different atom types is referred to as "cross-correlation". The Carhart descriptor vector consists of the autocorrelation vectors for all atom types and the cross-correlation vectors of all pairs of different atom types.

In 1996, Sheridan et al. [16] were the first to use pharmacophoric atom types for an autocorrelation approach. This technique is suited to characterize ligand–receptor interactions in a general way, allowing for more different but equally interacting molecules to be identified as similar. Sheridan et al. also extended the topological Carhart approach to the 3D case, and this was soon followed up by a binary representation of such a descriptor [17]. In 2003, Stiefl and Baumann [18] reported an autocorrelation approach using surface points representing pharmacophoric features.

The work of Schneider et al. [6] first focused on the scaffold-hopping ability of autocorrelation descriptors, in this case topological pharmacophores. The general description of the atoms with pharmacophore atom types in combination with the decomposition of molecules into atom pairs was shown to be especially successful in finding new molecules with significant different molecular scaffolds, maintaining the desired biological effect.

The following discussion considers three PPP pair descriptors: CATS (topological PPP pairs), CATS3D (spatial PPP pairs) and SURFCATS (surface PPP pairs). Figure 3.1 provides a graphical overview.

3.2.2
Comparison of Molecular Topology: CATS

The CATS descriptor belongs to the class of topological atom-pair descriptors. The CATS descriptor does not characterize the atoms of each atom-pair by their chemical element type. Instead, atoms are assigned to PPP types. The employment of the 2D molecular structure as the basis for the calculation is a crude simplification of reality. However, even though the interaction between a ligand and its binding partner is clearly a 3D event, the two-dimensional structure captures much about the physical properties and reactivity of a molecule [19]. A clear advantage with topological descriptors is that they circumvent the problem of conformational flexibility inherent to all 3D descriptor methods.

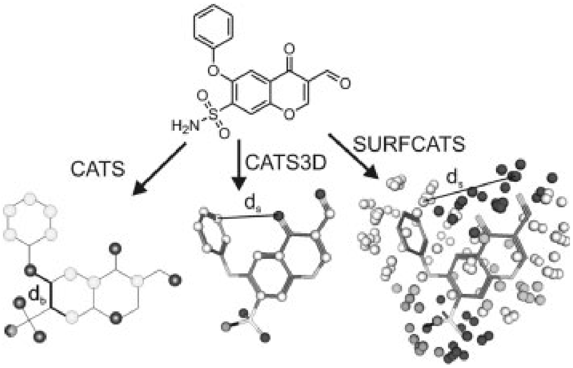
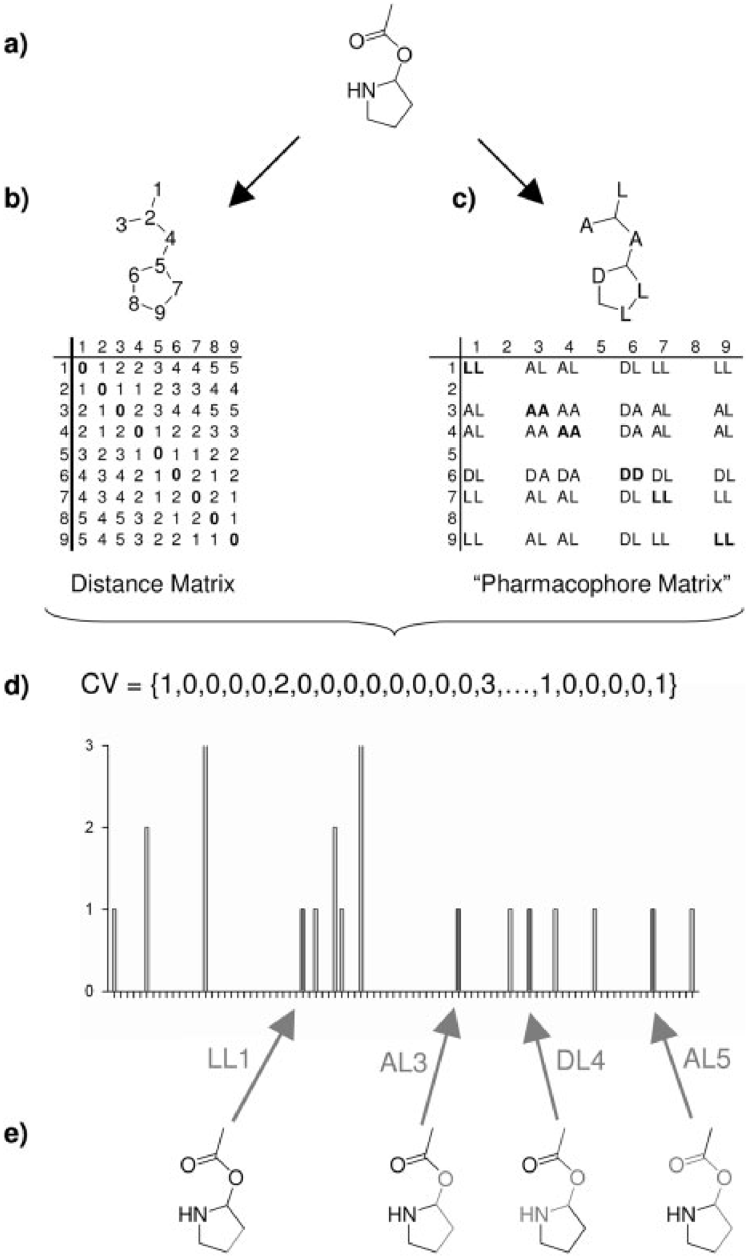


Fig. 3.1 The CATS family of descriptors: CATS, CATS3D and SURFCATS. The degree of abstraction from the atomic molecular structure is assumed to be SURFCATS > CATS3D > CATS. All descriptors are based on a PPP-type description of the underlying molecule. For each descriptor, pairs of PPPs are transformed into a correlation vector. CATS is calculated from the topological distances of atom-based PPP pairs. For CATS3D, the spatial distances between atom-based PPPs are used instead. SURFACTS uses the spatial distances between PPPs on the contact surface of a molecule. Here the PPPs represent the atom types of the nearest atom to each surface point.

The hydrogen-depleted molecular graph represents the basis for computing the CATS descriptor (Fig. 3.2). Two types of information are derived from the molecular graph: the topological distance matrix and the assignment of PPPs to the nodes (atoms) of the graph which ultimately yields a "pharmacophore matrix" (see below). The topological distance matrix of a molecular graph contains the minimal number of edges (bonds) between all pairs of vertices in the graph. The entries d_{ij} of the distance matrix D hold the shortest path measured as the number of bonds between vertex i and vertex j. (Note: an algorithm that calculates D must guarantee that the shortest path between all pairs of vertices is always found. This is of major importance if the molecular graph is cyclic and different edges can be passed to make a connection between two vertices.)

Our implementation of the CATS descriptor applies a breadth-first algorithm to compute the distance matrix. The concept of this algorithm can be illustrated with the canal-water analogy: if the graph represents a system of canals and water is filled in this canal at one point, the water would spread out uniformly in this system. The water would always "decide" to take the shortest path from the starting point to all other points in the system. The algorithm uses a data structure termed queue. A queue stores data according to the FIFO (first-in first-out) principle. Details about this particular data structure can be found elsewhere [20]. (Note: a variety of other all-pairs shortest path algorithms have been described. One approach is the deployment of a single-source shortest path algorithm. Such an algorithm finds the shortest path from a single vertex



	1	2	3	4	5	6	7	8	9
1	**0**	1	2	2	3	4	4	5	5
2	1	**0**	1	1	2	3	3	4	4
3	2	1	**0**	2	3	4	4	5	5
4	2	1	2	**0**	1	2	2	3	3
5	3	2	3	1	**0**	1	1	2	2
6	4	3	4	2	1	**0**	2	1	2
7	4	3	4	2	1	2	**0**	2	1
8	5	4	5	3	2	1	2	**0**	1
9	5	4	5	3	2	2	1	1	**0**

	1	2	3	4	5	6	7	8	9
1	**LL**		AL	AL		DL	LL		LL
2									
3	AL		**AA**	AA		DA	AL		AL
4	AL		AA	**AA**		DA	AL		AL
5									
6	DL		DA	DA		**DD**	DL		DL
7	LL		AL	AL		DL	**LL**		LL
8									
9	LL		AL	AL		DL	LL		**LL**

to all other vertices in the graph. An iteration over all vertices in the graph leads to an all-pairs shortest path solution. Among the popular single-source shortest graph algorithms is Dijkstra's [21]. Algorithms that are specifically tailored for the all-pairs shortest path provide a better running time. The Floyd–Warshall algorithm is such an example [22].)

The derivation of the topological distance matrix from the molecular graph is followed by the assignment of PPPs to the nodes of the graph. The following list provides chemical definitions of the five PPP types that are implemented in the CATS descriptor. The upper-case letter in parentheses is the abbreviation of each PPP type. Additionally, a functional group description is paired with its corresponding SMARTS in square brackets:

1. hydrogen-bond donor (D)
   - oxygen atom of an OH-group – [#6H]
   - nitrogen atom of an NH or NH$_2$ group – [#7H,#7H2]
2. hydrogen-bond acceptor (A)
   - oxygen atom – [#6]
   - nitrogen atom not adjacent to a hydrogen atom – [#7H0]
3. positive (P)
   - atom with a positive charge – [*+]
   - nitrogen atom of an NH$_2$-group – [#7H2]
4. negative (N)
   - atom with a negative charge – [*–]
   - carbon, sulfur or phosphorus atom of a COOH, SOOH or POOH group – [C&$(C(=O)O),P&$(P(=O)O),S&$(S(=O)O)]
5. lipophilic (L)
   - chlorine, bromine or iodine atom – [Cl,Br,I]
   - sulfur atom adjacent to exactly two carbon atoms (C–S–C) – [S;D2;$(S(C)(C))]
   - carbon atom adjacent only to carbon atoms – [SMARTS omitted owing to complexity].

Fig. 3.2 Schematic of the CATS descriptor calculation. (a) The hydrogen-depleted two-dimensional molecular graph provides the input. (b) The graph is simplified for the distance matrix computation: different bond orders are not considered (unweighted graph) and all element types are disregarded. The algorithm starts at an arbitrary chosen atom and visits all nodes of the graph in a breadth-first manner, thereby building up the distance matrix. The numbers at the vertices are used to reference individual atoms in the distance matrix. These numbers also illustrate the visiting order of the algorithm during graph traversal. Values on the first bisecting line of the distance matrix are shown in bold. (c) Potential pharmacophore points (PPPs) are assigned to the atoms (D, hydrogen donor; A, hydrogen acceptor; L, lipophilic). These assignments are then employed to set up the pharmacophore matrix. (d) Finally, corresponding elements of the distance and the pharmacophore matrix are combined to yield the CATS descriptor. The descriptor is depicted numerically as a correlation vector (CV) and shown graphically as a histogram. (e) Four examples of PPP–PPP distance tuples and their respective occurrence in the histogram.

According to these definitions, each atom of a molecule is assigned to no, one or two PPPs. Since the descriptor is based on atom pairs, a "pharmacophore matrix" is built up. The entries p_{ij} of the pharmacophore matrix P hold the PPP pair of vertex i and vertex j. If an atom is not a member of any PPP group, the row and column that correspond to the atom remain empty. A single atom can also belong to more than one PPP group. In this case, the entry p_{ij} of the pharmacophore matrix P holds more than one PPP pair. All possible pairing combinations of the five PPPs result in 15 pairs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL).

The information of the distance matrix and the pharmacophore matrix is then combined to yield the CATS descriptor. Each entry p_{ij} of the pharmacophore matrix P is associated with the corresponding entry d_{ij} of the distance matrix D. In other words: Each PPP pair is related to the respective distance between the two PPPs, and the PPP–PPP distance occurrences (PPP pair frequencies) are counted. Usually, we consider a minimum distance between a PPP pair of zero bonds, i.e. an atom is correlated with itself, and a maximum distance of nine bonds. Thus, the final result of the CATS descriptor calculation is a 150-dimensional correlation vector arising from 15 possible PPP pairs related to 10 different distances (zero to nine). The final step is the application of a scaling scheme to the descriptor vector. We developed three different schemes: no scaling (raw counts), division by the number of non-hydrogen atoms in the molecule and division of each of the 15 possible PPP pairs by the added occurrences of the two respective PPPs. The latter reflects the idea that rare PPP types might contribute in a special way to a ligand–receptor interaction, whereas abundant types, e.g. lipophilic atoms, should be downweighted to avoid dominance of the descriptor.

A particular property of the topological CATS descriptor is its speed of calculation. Thereby, the program qualifies for applications that deal with very large numbers of compounds, e.g. virtual screening campaigns in early stages of the drug discovery process.

3.2.3
Comparison of Molecular Conformation: CATS3D

Like many 2D descriptors, CATS has a counterpart in 3D space: the CATS3D descriptor. While the topological pharmacophore approach has the advantage that the time-consuming calculation of conformations can be avoided, the binding event is nevertheless a three-dimensional interaction between a ligand and its receptor. Accordingly, it should be advantageous to exploit such information if available.

The main difference in the correlation vector representation of a 3D conformation in comparison with a topological representation of a molecule is that the distances between the atoms are no longer shortest paths. Instead, Euclidean distances between all atoms are used. Distances between atoms are not restricted to integer values, so the distances have to be partitioned into a set of distance bins. Several such binning schemes have been proposed [9, 16, 17]. For CATS3D we generally employ 20 distance bins that cover distances from 0 to 20 Å in steps of 1 Å.

For CATS3D we used the modified PATTY atom types [23] available with the pH4_aType function in MOE [24]. Other PPP assignment schemes could also be employed. PATTY provides six PPP types: cation, anion, hydrogen-bond acceptor, hydrogen-bond donor, polar (hydrogen-bond acceptor *and* hydrogen-bond donor) and hydrophobic. Whereas the topological CATS descriptor allows assignments of more than a single PPP type to one atom, the CATS3D descriptor employs a single PPP type per atom.

Using 20 distance bins for each of the 21 possible combinations of PPP pairs resulted in a descriptor of 420 dimensions. The value stored in each bin is scaled by the added incidences of the two respective features. Each dimension ("bin") of the CATS3D CV is calculated according to the equation

$$CV_d^T = \frac{1}{N_1 + N_2} \sum_i \sum_j \frac{1}{2} \delta_{ij,d}^T \tag{2}$$

where i and j are atom indices, d is a distance range, T is the pair of PPP types of atoms i and j, N_1 and N_2 are the total number of atoms of types of i and j present in a molecule and δ_d^T (Kronecker delta) = 1 for all pairs of atoms of type T within the distance range d. The factor of 0.5 in the sum avoids double counting of pairs. Pairs of atoms with themselves are not considered.

3.2.4
Comparison of Molecular Surfaces: SURFCATS

The SURFCATS approach is a further extension of the CATS3D concept. The interaction between ligand and receptor is mediated by the surface between the two molecules. Accordingly, it should be advantageous to describe molecules by their surface properties. Generally, it is assumed that a surface-based description of a molecule is less dependent on the scaffold of the ligand than a topological or atom-based representation, and consequently should have an enhanced scaffold-hopping capability [25, 26].

The first application of a surface-based pharmacophore correlation vector was reported by Stiefl and Baumann in 2003 [18] with the MaP (*ma*pping *p*roperty distributions of molecular surfaces) descriptor. They applied their MaP descriptor for QSAR applications. To our knowledge, an application of this descriptor to similarity searching has not been reported.

The surface points for the calculation of SURFCATS are taken from the contact surface (we usually employ the Gauss–Connolly function in MOE with a spacing of 2 Å). Each surface point is then assigned to the PPP type of the nearest atom. Equation (1) is used to calculate the CV with surface points instead of atoms. In contrast to MaP, the surface points are not equally distributed on the surface of a molecule. The effect of this circumstance has not yet been analyzed in detail. We expect only a minor effect on descriptor performance. In fact, the original and very successful surface autocorrelation approach of Wagener et al. [9] did not employ equidistant surface points.

3.3 Applications

3.3.1 Retrospective Screening Studies

Chemical similarity searching can support the identification of novel molecules that reveal similar biological activity as one or more query structures. Ligand-based similarity indices allow chemical similarity searching in case of an absent structure for the biological target of interest. This concept is frequently and successfully employed for compiling activity-enriched subsets in early-phase virtual screening and compound library design [27–30]. Fundamentally, these methods rely on

- a representative reference structures (also termed "query" or "seed" structures)
- molecular descriptors that are correlated with biological activity (e.g., a pharmacophore descriptor)
- an appropriate similarity metric (for an overview, see Ref. [31]).

"Retrospective screening" provides a means of evaluating these factors. The basic idea is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximize the number of known actives in the subset, thereby forming a "focused library" [32]. Subset selection is based on the pairwise chemical similarity between the query structure and each molecule in the pool. The result of this calculation is a list ranked according to descending similarity. Such a retrospective screening experiment can be rated by the enrichment factor, *ef* [32, 33]:

$$ef = \left(\frac{S_{act}}{S_{all}}\right) \Bigg/ \left(\frac{P_{act}}{P_{all}}\right) \tag{3}$$

where P_{all} is the total number of compounds in the database (pool), S_{all} is the number of molecules in the subset, P_{act} is the number of "active" molecules in the pool and S_{act} is the number of actives found in the subset. A method that is superior to a random selection of compounds returns an $ef > 1$. The enrichment factor can be visualized by an enrichment curve: S_{all}/P_{all} is plotted on the abscissa and S_{act}/P_{act} on the ordinate. A well-performing similarity search should result in a curve above the diagonal line.

Pharmacophore CV descriptors are highly modifiable. Several parameters can be altered and tweaked. Examples are the chemical definitions of the PPPs, the considered minimum and maximum distance between an atom pair, the binning scheme in case of a 3D descriptor or the scaling of the final CV. During our work with the different flavors of the pharmacophore CV descriptor we carried out several studies to evaluate parameter settings. Most of these experiments employed 12 selected subsets of the COBRA dataset for retrospective virtual screening. The COBRA dataset is a collection of reference molecules for ligand-based library design compiled from recent scientific literature [34], which were divided into non-overlapping subsets. The 12 datasets were made up of a set of active compounds (query structures) and the respective remainder of the COBRA dataset as "inactive compounds" (virtual screening library). The sets of active compounds contained ligands that bind to angiotensin-converting enzyme (ACE, 44 compounds), cyclooxygenase 2 (COX2, 93 compounds), corticotropin releasing factor (CRF antagonists, 63 compounds), dipeptidylpeptidase IV (DPP, 25 compounds), G-protein coupled receptors (GPCR, 1642 compounds), human immunodeficiency virus protease (HIVP, 58 compounds), nuclear receptors (NUC, 211 compounds), matrix metalloproteinase (MMP, 77 compounds), neurokinin receptors (NK, 188 compounds), peroxisome proliferator-activated receptor (PPAR, 35 compounds), β-amyloid converting enzyme (BACE, 44 compounds) and thrombin (THR, 188 compounds).

In a first study, we investigated the influence of the dataset and the descriptor on ligand-based virtual screening [33]. We employed the 12 different datasets compiled from the COBRA dataset and two different correlation vector descriptors, namely CATS and CATS3D, for which a single conformation was calculated for each molecule of the dataset with the program CORINA [35]. With the exception of the GPCR dataset, considerable enrichment factors of up to 26 for the first percentile of the similarity-ranked datasets were yielded with all three descriptors. A comparison of the descriptors revealed that none of them is superior for all 12 datasets, but for some datasets there is a preferred one. The suitability of the descriptors depends on the underlying dataset, i.e. the binding patterns of a specific ligand–receptor pair. Distinct performances of the descriptors were expected, as the CATS2D encodes topological information of PPPs, and the CATS3D spatial information of PPPs.

Irrespective of the descriptor, the approximate classification accuracy seems to be determined by the dataset. Some target classes yield better enrichment factors than others. We deduced two possible reasons for this behavior. First, the descriptors may cover the essential binding pattern of particular datasets to a different extent. Second, the individual datasets are defined at different levels of specificity. Some include sets of ligands binding to individual receptor subtypes (e.g., BACE, THR) whereas others comprise very loosely defined classes of bioactive agents (e.g., GPCR, NUC). Whichever of these two reasons might hold true, in either case the dataset with its inherent properties has a major influence on the outcome of a virtual screening experiment.

Since the enrichment factor discriminates only between active and inactive compounds, we further investigated *which* active compounds were retrieved by the two descriptors among the top-ranking ones. Figure 3.3 depicts this for the first five percentiles of three dataset by means of Euler–Venn diagrams. It is no-

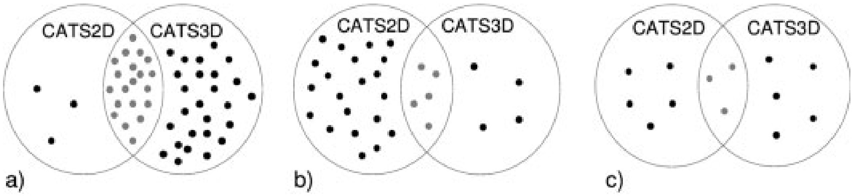


Fig. 3.3 Elements of the Euler–Venn diagrams represent compounds that were found among the first 5% of the similarity-ranked list that results from retrospective screening with the (a) COX2, (b) HIV protease and (c) MMP datasets of the COBRA dataset. The Manhattan distance was employed as a distance measure. Membership indicates that the respective compound was retrieved by retrospective screening with the corresponding descriptor. The diagrams reveal that the three descriptors complement one another to different extents depending on the underlying dataset.

teworthy that although the enrichment factors with different descriptors were approximately the same, the active compounds among the top-ranking ones varied. Figure 3.3 shows that a large number of compounds were exclusively retrieved with one descriptor, and that the intersection sizes of the descriptors were rather small. These two observations sustain the hypothesis that each descriptor covers a certain, and to a varying extent different, aspect of the ligand–receptor binding pattern. Moreover, the information contents of the 2D and the 3D descriptors complement each other. The extent of completion can be measured by computation of the “cumulative percentages”: For a given dataset the two descriptors gave rise to two different similarity-ranked lists. The active compounds among the first 5% of these lists were extracted to obtain two sets of active compounds. The sets were then united according to the union operator of set theory. Finally, the number of elements of the united set was related to the total number of active compounds of the particular dataset. Cumulative percentages facilitated the retrieval of additional 5–51% of active compounds compared with the exclusive employment of the topological CATS descriptor. Hence it may be appropriate to unite the information encoded by different descriptors if a similarity search is performed to cover more facets of the ligand–receptor binding pattern under investigation. Willett and co-workers came to similar conclusions from their retrospective screening studies and recommend a “data fusion” strategy for the combination of ranked lists [36, 37].

In a subsequent study, we examined the influence of seven similarity indices on the enrichment of actives using the topological CATS descriptor and the 12 COBRA datasets [31]. In particular, we evaluated to what extent different similarity measures complement each other in terms of the retrieved active compounds. Retrospective screening experiments were carried out with seven similarity measures: Manhattan distance, Euclidian distance, Tanimoto coefficient, Soergel distance, Dice coefficient, cosine coefficient, and spherical distance. Apart from the GPCR dataset, considerable enrichments were achieved. Enrichment factors for the same datasets but different similarity measures differed only slightly. For most of the datasets the Manhattan and the Soergel distance yielded the overall highest enrichment factors. One might deduce that if only a single distance measure is applied, the Manhattan distance should be preferred owing to its computational simplicity and altogether above-average performance.

To what extent are the top-ranking active compounds identical if different similarity metrics are applied? Each of the 12 datasets yielded seven similarity-ranked lists obtained with the seven similarity metrics. For each dataset, the cumulative percentages were calculated for the first 5% of these lists. This procedure led to the retrieval of significantly more hits than found by any single similarity metric. The increase of the cumulative percentages for all seven metrics compared with the employment of only the Manhattan distance ranged from additional 5 to 28% with an average of 19% over all 12 datasets. Our descriptor comparison study [33] suggests that different descriptors complement each other in terms of the top-ranking active compounds. The comparison of seven similarity indices led to the conclusion that they complement each other in the same way. Therefore, it might be advantageous to employ several molecular descriptors and similarity metrics in parallel and thereby benefit from a unification of the various definitions of "chemical similarity".

Descriptor Scaling

Which influence do different scaling methods have on the performance of the topological CATS descriptor? We addressed this question with a comparison of three different ways of scaling the correlation vector descriptor [38]:

- No normalization. The values of the vector represent raw counts ("*counts*").
- Division by the number of non-hydrogen atoms in the molecule ("*normalization1*").
- Division of each of the 15 possible PPP pairs by the added occurrences of the two respective PPPs ("*normalization2*").

The three scaling methods were assessed by enrichment factors that resulted from retrospective screening campaigns. Retrospective screening was performed with 12 different datasets, each of which was a subset of the COBRA dataset. Altogether, *normalization2* exhibited the highest performance of the three scaling approaches: it achieved superior enrichment factors for 10 of the 12 subsets and comparable values to the other two scaling methods for the remaining two subsets. Differences in the enrichment factors were up to 173% compared with the second-best scaling method. The *normalization2* scaling method can be regarded as the most "sensitive" one: each of the 15 possible PPP pairs is scaled individually, thereby taking into account the unequal occurrences of the respective PPPs within a molecule. This procedure guarantees a balanced scaling for each PPP pair. A division of the complete descriptor by the number of non-hydrogen atoms (*normalization1*) puts less frequent PPP pairs in danger of becoming minuscule. Thus, *normalization1* may lead to an unintentional emphasis on more common PPPs. According to our observations, this holds true in particular for lipophilic centers. If less frequent PPPs play a crucial role in the interaction pattern of a ligand and its binding partner, the inherent emphasis of more frequent PPPs seems especially disadvantageous. The study clearly demonstrated that appropriate descriptor scaling can tweak a similarity search. We decided to employ the *normalization2* scaling for the topological CATS correlation vector descriptor in future applications. Similar experiments led to comparable conclusions for the CATS3D descriptor (unpublished data). However, one should always keep in mind that each virtual screening campaign presents novel challenges and requires careful selection of all parameters.

In the following, we summarize the outcome of several studies that addressed further questions related to appropriate descriptor calculation.

"Fuzzy" Binning

We explored the application of a "fuzzy" binning scheme for the CATS descriptor [38]. Given the occurrence *c* of a specific PPP pair spaced *n* bonds apart, then the counters of the bins that are associated to the same PPP pair in the distance $n+1$ and n–1 are incremented by *bc*, where *b* has values between zero (no fuzzy binning) and 1. The increase *b* was performed in steps of 0.1, and *normalization2* was employed. Again, evaluation took place in terms of the enrichment factor for the 12 subsets of the COBRA dataset. To our surprise, the enrichment factors were only insignificantly affected with respect to the estimated error margins. We concluded that it seems to be reasonable not to apply a fuzzy binning scheme for the topological CATS descriptor if enrichment factors are of interest. This result is in contrast to studies with three-dimensional descriptors where "fuzzification" had proven to be useful for similarity searching [16, 17, 39, 40].

"Binarization"

The number of virtually screened compounds is often very large in early stages of the hit- or lead-finding process. The application of ligand-based similarity searching at this point in drug discovery requires the calculation of many pairwise compound similarities. These calculations can be speeded up with binary encoded descriptors since binary operations are computationally less expensive than numerical operations. Moreover, binary encoded descriptors occupy less space than descriptors composed of integers or floating-point numbers on internal and external storage devices such as random access memory and disks. Individual values of the topological CATS and the CATS3D descriptor vector are floating-point numbers. To generate binary CVs, we converted the "holographic" (i.e. real-valued) vectors to a binary representation: Each value of the descriptor vector was set to one if its numerical value was greater than zero. Otherwise, i.e. if the value was zero, it was left unchanged. This "binarization" was motivated by the outcome of a prior neuro-fuzzy analysis that was aimed at the classification of active from inactive compounds [41]. Again, the 12 datasets compiled from the COBRA dataset were employed. Retrospective screening studies were performed to assess the influence of the binarization on the enrichment factor [42].

Holographic and binary representations of the CATS and CATS3D descriptor were analyzed in detail. A deviation was termed "significant" if the enrichment factor between the two representations differs by more than 20%. With the topological CATS descriptor we obtained equal enrichment factors for three datasets, a better performance of the holographic representation for seven datasets (among which were three significant differences) and one non-significant better performance of the binary vector. Retrospective screening with the CATS3D descriptor led to similar results: equal enrichment factors for two datasets, greater enrichment factors of the holographic representation for five datasets (among which were three significant differences) and four non-significant better performances of the binary vector. The holographic descriptor vector seems to be advantageous for the two CATS descriptors. However, significant performance gains of the holographic descriptor compared with its binary counterpart were only achieved for three of the 11 datasets. These performance gains ranged from 22 to 43% with most being less than 27%.

The overall correlation between the holographic data and the binary data for the first 2% of the screened dataset was 0.94 and 0.92 for the topological CATS and the CATS3D descriptor, respectively. These high correlations provide additional evidence of similarity between the holographic and the binary pharmacophore-based descriptor vectors. The "binarized" CATS and CATS3D descriptor can be employed for rapid similarity searching without losing significant enrichment of actives in the virtual hit lists. It might even be rewarding to convert other non-binary pharmacophore descriptors to their binary counterpart when large numbers of compounds impede the application in chemical similarity searching.

Conformation Dependency

Finally, we examined the impact of molecular flexibility on virtual screening with CATS3D [43]. Using a descriptor based on the 3D conformation of a molecule (e.g., CATS3D), one might assume that it is essential for new molecules to be presented in a conformation near to the conformation of the reference to be considered as similar. Consequently, it is often the strategy to calculate a set of multiple conformations per molecule of a database. This is based on the observation that for most molecules multiple conformations exist with comparable energies. One of these usually binds to the receptor, but not necessarily the one with the lowest energy [44]. However, calculating multiple conformations can be rather time consuming. On the other hand, CATS3D has been shown to perform better than CATS for some classes of molecules, using only a single conformation [33]. To test the impact of multiple conformations, co-crystal structures of 11 target classes served as queries for virtual screening of the COBRA database. Different numbers of conformations were calculated for the COBRA database with the programs CORINA [35] and ROTATE [45] for the purpose of retrospective screening. We found that using only a single conformation already results in a significant enrichment of isofunctional molecules. This observation was also made for ligand classes with many rotatable bonds. The impact of using multiple conformations on the enrichment of actives was generally low. Only for some classes of molecules was considerable improvement in the enrichment of active molecules observed when multiple conformations were considered. We conclude that CATS3D provides a 3D virtual screening approach that is only moderately dependent on the presence of conformations that are close to the "bioactive" conformation of a molecule to estimate its biological activity.

3.3.2
Scaffold-hopping Potential

The ultimate goal in virtual screening is to find the maximum number of maximally diverse active compounds from a given chemical subspace. There are several reasons for seeking a set of diverse structures. Diverse structures offer the medicinal chemist a choice in terms of chemical accessibility and prospects for lead optimization. Multiple leads ("backup" compounds) lower the chance of drug development attrition in case of undesirable ADMET (absorption, distribution, metabolism, excretion and toxicity) properties. One criterion for a diverse set of molecules is the presence of different scaffolds. This concept is based on the idea that drug-like molecules are built up from a scaffold (framework) and side-chains [46]. A recently published method for scaffold classification inspired us to tackle the question of the scaffold-hopping capability of different virtual screening methods [8]. The program Meqi (*m*olecular *eq*uivalence *i*ndices), devised by Xu and Johnson [47], was used to classify the scaffolds. First, the full molecular representation was simplified to a scaffold representation or to a reduced scaffold representation (Fig. 3.4). Subsequently, an equivalence number for each scaffold or reduced scaffold was calculated with a modified Morgan algorithm [48]. To assess the scaffold-hopping ability of the CATS family pharmacophore pair descriptors, we used the MACCS keys as a second class of descriptors (based on a substructure fingerprint) in the retrospective screening experiment [65]. To summarize: we employed 10 different datasets, four descriptors

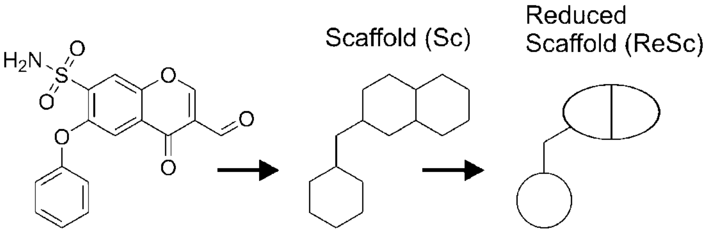


Fig. 3.4 Definition of scaffold (Sc) and reduced scaffold (ReSc). In this work we defined the scaffold of a molecules as the side-chain depleted molecular graph without annotation of atom types. A reduced scaffold is a more general representation which does not discriminate between rings consisting of different numbers of heavy atoms, but systems containing different numbers of rings are still not considered being equal.

(CATS, CATS3D, SURFCATS, MACCS) and three molecular representations (full, scaffold and reduced scaffold representation).

Intuitively, one would assume that the scaffold-hopping capability would be best for the descriptors that encode molecules at a high level of abstraction from the chemical structure. According to this hypothesis, the resulting order should be SURFCATS > CATS3D > CATS. The conceptually different substructure-based MACCS keys might be assumed to be the most conservative similarity searching method in terms of scaffold hopping: Substructures represent "exact" molecular fragments – not allowing for ambiguities. However, it has already been shown that the MACCS keys can be superior to 3D pharmacophore pair descriptors for the task of clustering actives within compound databases [17]. On the other hand, this behavior might have resulted from sets of structurally very similar active molecules.

Retrospective screening was performed in the same manner as by Fechner et al. [33], leaving out the two very general classes GPCR and nuclear receptors. The average enrichment factors for the first 5% of the database are shown in Fig. 3.5.

As stated previously for the topological CATS descriptor [31], the influence of different similarity metrics on the overall enrichment is marginal. For the full

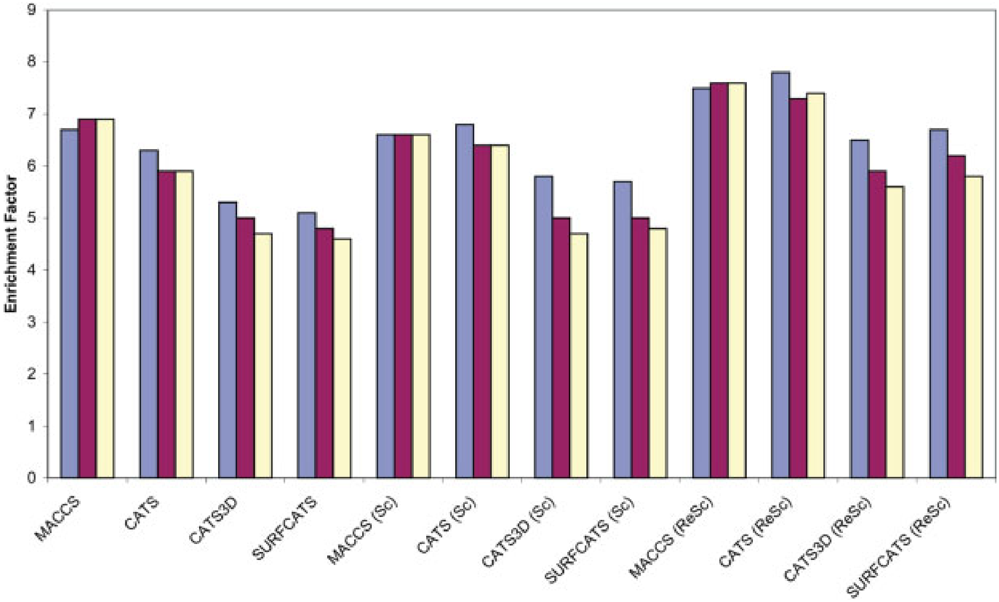


Fig. 3.5 Averaged enrichment factors over 10 ligand classes from the COBRA database (top 5%). Comparison of the performances of MACCS, CATS, CATS3D and SURFCATS for full molecular representations, scaffolds (Sc) and reduced scaffolds (ReSc). Three similarity metrics were applied: the Tanimoto similarity (blue), the Euclidean distance (red) and the Manhattan distance (yellow).

molecular representations, the order of the methods in terms of the enrichment factors was found to be MACCS > CATS > CATS3D > SURFCATS. This order is exactly the reverse of that *intuitively* expected for the enrichments of scaffold or reduced scaffold representations: SURFCATS > CATS3D > CATS > MACCS. Regarding the enrichment of scaffolds and reduced scaffolds, CATS performs comparably to MACCS. An explanation for the high performance of the MACCS keys in scaffold enrichment might be that the connectivity of the substructures is not accounted for in the descriptor. This might lead to an effective retrieval of molecules with slightly different scaffolds but the same side-chain decoration.

A different outcome can be observed for the enrichment of single activity classes. Figure 3.6 shows enrichment curves for three selected classes: COX-2, HIV protease and neuraminidase. One can see that the descriptor performance depends on the class of ligands. For all three examples, the shapes of the enrichment curves for full molecular representations and the respective reduced scaffold representations were similar. This might lead to the conclusion that generally none of the descriptors focuses on the molecular scaffolds *per se*. As a consequence, enrichment of different scaffolds should be most likely with descriptors performing well in full molecule enrichment. The reverse, on the other hand, seems not necessarily true, i.e. substructure searching might result in a high enrichment of actives in a database of many molecules comprising this particular substructure, while not finding other actives with a different scaffold. We wish to stress that such conclusions should be treated with caution since only retrospective studies were carried out using compound sets representing artificially compiled activity classes. Typically, these structures are the result of a limited number of lead optimization projects and might therefore not adequately represent the “drug universe”.

The mutual complementation of the different methods was examined in more detail for four selected molecules: rofecoxib (COX-2), celecoxib (COX-2), indinavir (HIV protease), and lanepitant (neurokinin receptor). The results are shown in the form of Euler–Venn diagrams in Fig. 3.7. Apparently the methods complement each other. Each method was able to retrieve actives which were not found by the other methods. Interestingly, the performance of the different descriptors varied significantly within one class of ligands (compare, e.g., rofecoxib and celecoxib).

Investigating the capability of several methods for the enrichment of scaffolds or reduced scaffolds, we found only marginal differences from the enrichment of full molecular representations. Most important, only small improvements in scaffold hopping occurred using more general descriptors for the molecules in comparison with less general descriptors. It appeared that the MACCS keys were most successful in both retrieving active molecules and finding a diverse set of actives. It remains a matter of debate whether it is a reasonable assumption to leave out all information about the connectivity of the fragments. Additionally, bioisosteric replacements might be more difficult to find with the MACCS keys.

Summarizing, for the CATS family of descriptors we found that a higher abstraction level does not automatically lead to a higher percentage of retrieved scaffolds. Only in the case of CATS3D and SURFCATS did a more general description lead to slightly higher enrichment factors for reduced scaffolds.

3.3.3 Prospective Virtual Screening

An overview of successful prospective virtual screening campaigns using CV methods is given in Fig. 3.8. The first prospective application of the CATS descriptor was a virtual screening study aiming at finding novel cardiac T-type Ca^{2+} channel-blocking agents [6]. Using mibefradil (**1**, $IC_{50}=1.7\ \mu M$) as a reference structure, the 12 highest ranking molecules were tested experimentally. Nine of these compounds showed an IC_{50} below 10 μM. The best hit was clopimozid (**2**) with an IC_{50} below 1 μM. Clopimozid had a significantly different scaffold than the reference structure.

Naerum et al. used a very similar descriptor termed "CATS2", composed of a slightly different PPP type definition, for the identification of novel glycogen synthase kinase-3 (GSK-3) inhibitors [49]. Using the high-throughput screening hit **3** as reference structure, a new inhibitor **4** with an IC_{50} of 1.2 μM was found. Further experimental lead optimization led to molecule **5** with an IC_{50} of 0.39 μM. This result demonstrates that CATS-based similarity searching is suited for finding novel lead candidates. These initial hits can thereafter be optimized using more focused virtual screening or molecular modeling methods which take into account specific interactions that are relevant for the target under consideration.

Applications of CATS in ligand based *de novo* design with the program TOPAS (*top*ology-*a*ssigning *s*ystem) were also reported [50, 51]. TOPAS implements an evolution strategy to assemble molecular fragments via a defined set of virtual reactions. The molecular fragments were derived from retro-synthetic fragmentation [52] of the *World Drug Index* [53]. The newly assembled molecules are scored with their CATS similarity to a reference ligand for the biological target of interest. For the first application of this approach, a potent potassium channel blocker with an IC_{50} of 0.11 μM (**6**) served as a reference. One of the designs led to the new inhibitor **7** from a different chemical class with an IC_{50} of 7.34 μM. Again, a slight modification of this new lead recovered an activity within the order of magnitude of the reference compound, namely molecule **8** with an IC_{50} of 0.47 μM.

Another study reported the identification of novel cannabinoid receptor ligands using a combination of fragment-based *de novo* design and parallel synthesis [54]. In a first experiment, small libraries of similar molecules to the reference structure **9** ($K_i=0.11\ \mu M$) were generated. Since the cannabinoid receptors are part of the class of GPCRs, a fragment library tailored for GPCR ligand design was employed. The two recurring motifs **10** and **11** were used as templates for the parallel synthesis. For each of the templates the core structure

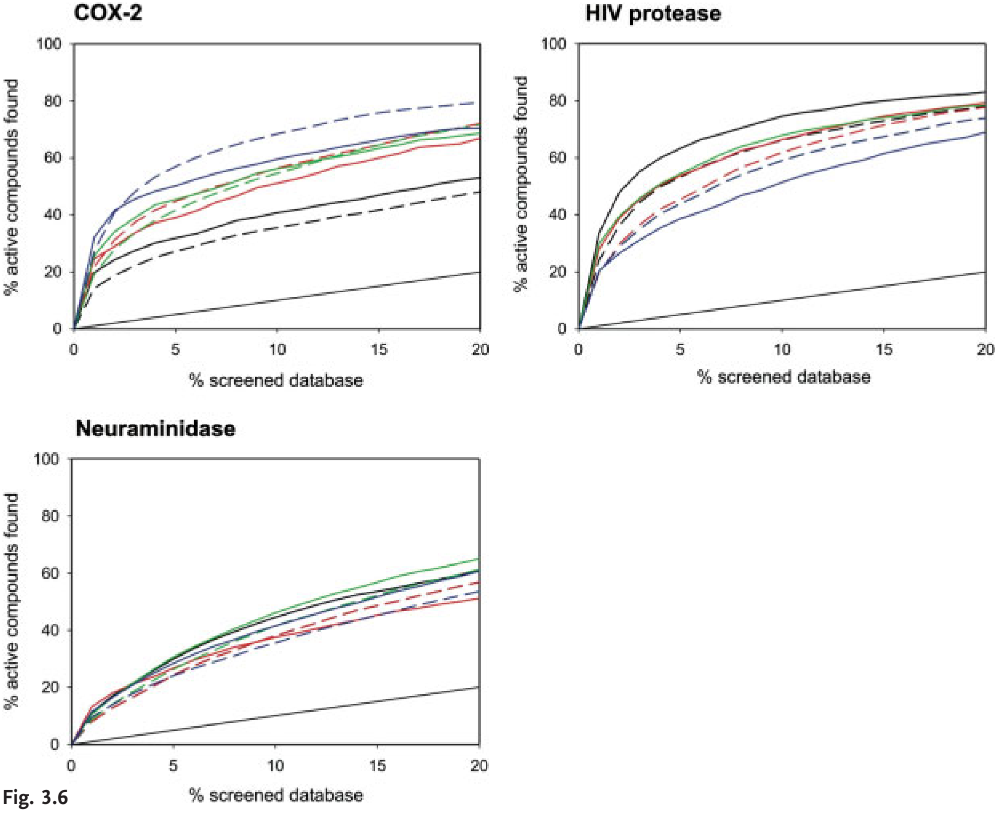

COX-2
HIV protease
Neuraminidase
% active compounds found
% screened database
0
20
40
60
80
100
0
5
10
15
20

Fig. 3.6

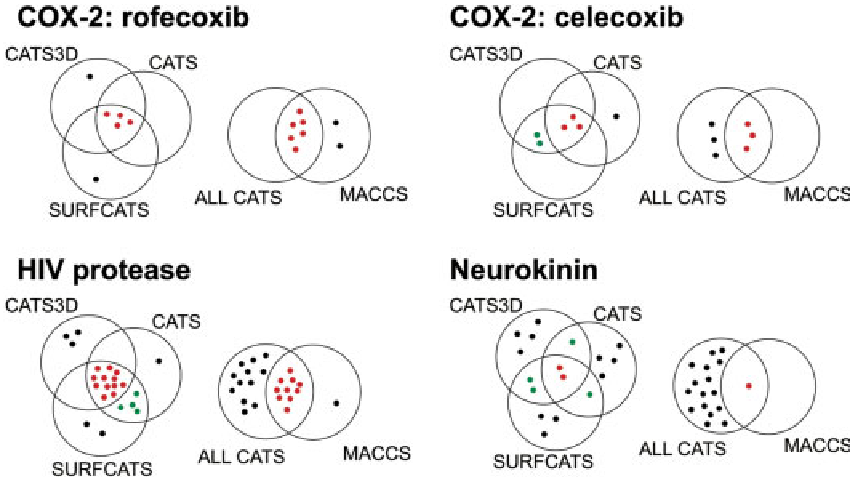

COX-2: rofecoxib
COX-2: celecoxib
HIV protease
Neurokinin
CATS3D
CATS
SURFCATS
ALL CATS
MACCS

Fig. 3.7

was maintained and two positions were identified for combinatorial optimization. The best hits **12** and **13** exhibited a K_i of 2.0 and 0.3 µM, respectively.

We recently applied CATS3D similarity searching to find novel metabotropic glutamate receptor 5 (mGluR5) modulators. This resulted in eight out of 29 experimentally tested molecules with a K_i below 50 µM [55]. All hits showed different scaffolds compared with the reference molecules.

Seven known antagonists of mGluR5 (**14–20**) with sub-micromolar IC_{50} were used as reference ligands (Fig. 3.9). A hypothesis about the receptor-bound conformation of these ligands was generated with the flexible alignment tool in MOE [24]. The 20 000 most drug-like compounds – as predicted by an artificial neural network approach [56] – of the Asinex [57] vendor database were screened with each of the seven molecules as a reference. From the resulting hit lists, a set of 29 high-scoring molecules were selected and tested in a binding assay. To determine the specificity of the hits the K_i for the receptor most similar to mGluR5, mGluR1, was also measured. Unfortunately, most of the ligands were only moderately specific for mGluR5. Ligand **22** exhibited even a higher K_i for mGluR1 than for mGluR5. Sometimes small structural modifications within the molecules determine specificity or trigger "jumps" in activity. Such "sensitive" structure–activity relationships can hardly be modeled by general approaches like the CATS descriptor family, because they lack the incorporation of any structure activity data.

In the mGluR5 study, all hits had different scaffolds than the reference molecules. To estimate the degree of uniqueness of the hits and the degree of scaffold hopping, we compared the average distance of each of the 29 hits from the virtual screening campaign with its respective nearest reference compound ($\langle D_{\text{lib}} \rangle$) with the average distance between the reference molecules ($\langle D_{\text{ref}} \rangle$). We applied three measures of molecular similarity: CATS3D with the Manhattan

Fig. 3.6 Enrichment curves of the average enrichment of actives for COX-2, HIV protease and neuraminidase. Solid lines denote the enrichment of the reduced scaffold representations of the molecules and dashed lines designate the enrichment of full molecular representations. CATS enrichment curves are shown in black, CATS3D in red, SURFCATS in green and MACCS in blue. The thin straight line (black) denotes the theoretical retrieval of actives, assuming an even distribution of actives over the database.

Fig. 3.7 Euler–Venn diagram showing the mutual complementation in the retrieval of reduced scaffolds with CATS, CATS3D, SURFCATS and MACCS. The first percentile of the database was considered. Four exemplary selected molecules, rofecoxib (COX-2), celecoxib (COX-2), indinavir (HIV protease) and lanepitant (neurokinin receptor), were used. The left Euler–Venn diagram in each example shows the mutual complementation of the three CATS methods. Red dots represent reduced scaffolds found by all three methods, green dots were found by two of the methods and black dots were found by only one of the methods. The right diagrams show the complementation between all CATS methods (ALL CATS) and the MACCS fingerprints. Red dots represent reduced scaffolds found by at least one CATS method and MACCS. Black were fond either solely by MACCS or solely by at least one of the three CATS approaches.

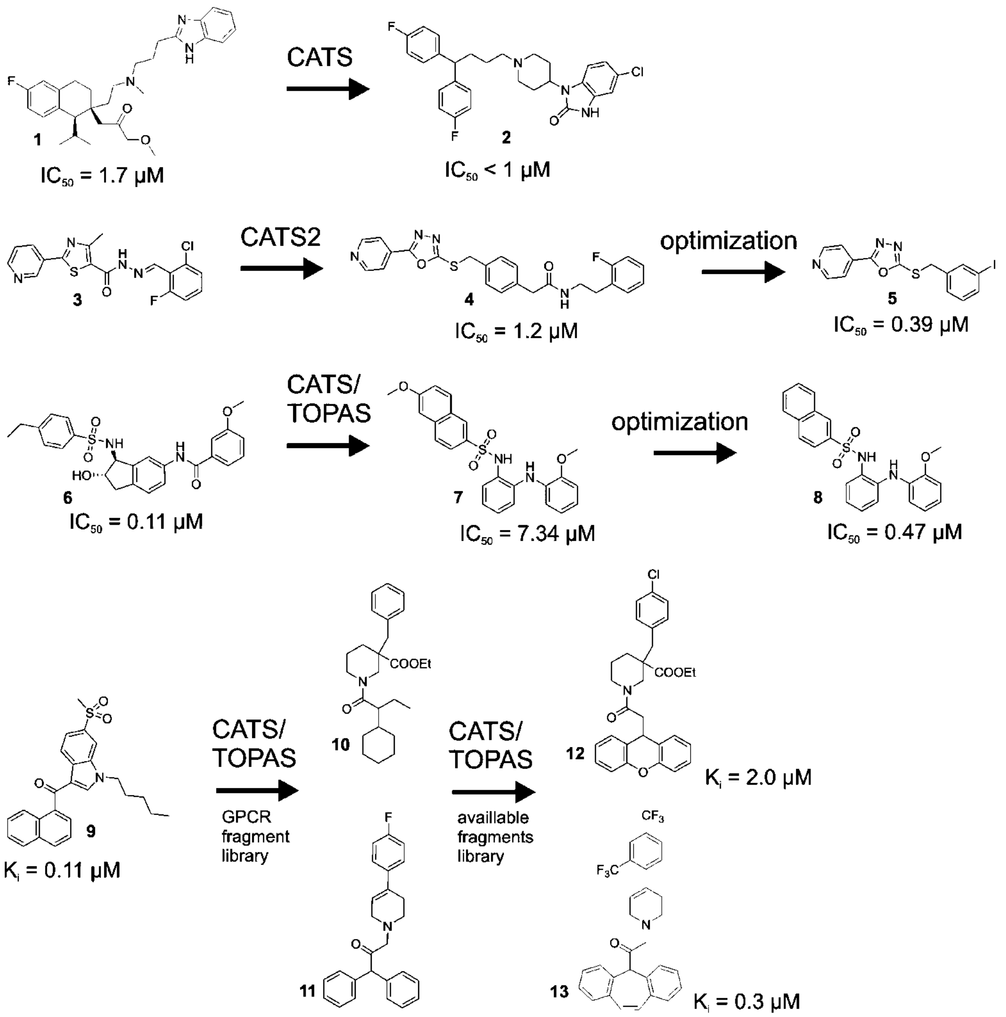


Fig. 3.8 Prospective screening examples using the CATS descriptor. CATS2 is a similar approach using a slightly different definition of PPPs. TOPAS was used in two cases for the *de novo* assembly of molecules.

distance, topological CATS with the Manhattan distance and the MACCS keys with the Tanimoto similarity. The average CATS3D distance of the virtual screening hits to their reference molecules was significantly smaller than the average distance between the reference molecules ($\langle D_{\text{lib}} \rangle = 1.41 \pm 0.45$, $\langle D_{\text{ref}} \rangle = 2.66 \pm 0.89$). In contrast, $\langle D_{\text{lib}} \rangle$ was only marginally smaller than $\langle D_{\text{ref}} \rangle$ for the topological CATS (3.31 ± 1.48 versus 3.6 ± 1.4). Employment of the MACCS keys resulted in $\langle D_{\text{lib}} \rangle$ being smaller (less similar) than $\langle D_{\text{ref}} \rangle$ (0.33 ± 0.11 versus 0.39 ± 0.15). This indi-

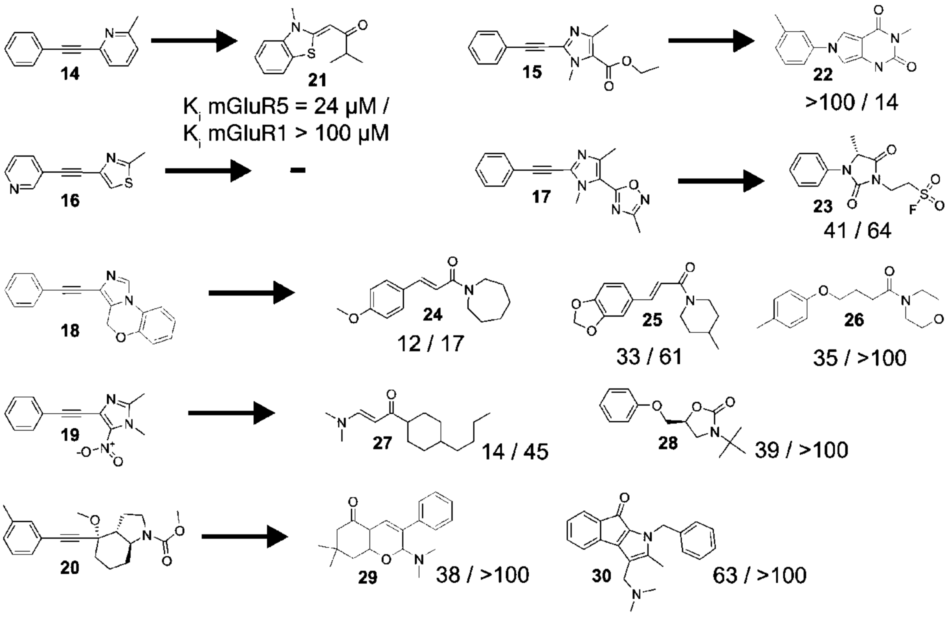


Fig. 3.9 Prospective screening for modulators of the mGluR5 modulators with CATS3D. Seven reference molecules with reported low nanomolar activity were used (**14**–**20**).

cates a greater structural similarity among the reference set than between the virtual screening hits and the reference molecules. Moreover, it is demonstrated that the compiled library contains molecules that are different from the reference structures – as estimated by MACCS substructure fingerprints – but are still considered isofunctional by the two CATS pharmacophore approaches. It is noteworthy that a substructure fingerprint Tanimoto similarity threshold of 0.85 is usually used for similarity searching [58, 59].

CATS3D was not only successful in scaffold hopping on the basis of the definition above (Meqi). We also observed a "substructure hopping", which might be seen as an equivalent to more traditional bioisosteric replacement strategies. It seems that the CATS descriptor family represents molecules in a way that allows a combination of scaffold hopping and "substructure hopping" at once. This can result in a selection of molecules which would not be considered similar by other methods such as the MACCS keys.

3.4 New Methods Influenced by the Correlation-vector Approach

3.4.1 "Fuzzy" Pharmacophores: SQUID

Using pairs, triplets or even quartets of atoms as PPPs is one possibility for the construction of a CV descriptor. An extension to this approach is to use pairs of larger and more general objects, which might result in a more generalized and abstract description of the molecule. The SQUID (*s*ophisticated *qu*antification of *i*nteraction *d*istributions) fuzzy pharmacophore is such an approach where pairs of Gaussian probability densities are used for the descriptor calculation [60]. The Gaussians represent clusters of atoms comprising the same pharmacophoric feature within an alignment of several active reference molecules. The incorporation of multiple aligned ligands within the SQUID approach resembles conceptual similarity to the traditional idea of a pharmacophore model [61].

Based on an alignment of active molecules, tolerances for the features are usually estimated to compensate for ligand and receptor flexibility. Pharmacophoric features that are present in many of the reference molecules result in a high probability and features which are sparse in the underlying molecules result in a low probability. Tolerances of the features which are considered by this approach might be better represented by Gaussian densities than by rigid spheres. For the resulting fuzzy pharmacophore models, different degrees of fuzziness can be defined, e.g. the model can be very generalizing or more restricted to the underlying distribution of atoms from the alignment. The fuzziness can be affected by the cluster radius, a variable which determines the radius within which atoms are clustered into PPPs.

For virtual screening, the 3D distribution of Gaussian densities is transformed into a two-point correlation vector representation which describes the probability density for the presence of atom pairs, comprising defined pharmacophoric features. This representation is independent of translation and rotation like the atom-pair descriptors, which renders rapid database screening possible without the necessity explicitly to align the molecules with the pharmacophore model in a pairwise fashion. Hence the fuzzy pharmacophore CV is useful for ranking 3D pharmacophore-based CV representations of molecules, namely CATS3D descriptors. Consequently, SQUID can be characterized as a hybrid approach between conventional pharmacophore searching, similarity searching and fuzzy modeling.

Figure 3.10, adapted from Palomer et al. [62], shows an alignment of the COX-2 inhibitors M5, SC-558 and rofecoxib. According to these authors, essential interactions for specific COX-2 inhibition are mediated by the aromatic rings A and B and the sulfonyl group. A set of pharmacophore model representations was calculated with different cluster radii, resulting in models with different degrees of fuzziness. The model with 1 Å cluster radius resulted in the most detailed representation of the underlying alignment, correspondingly with

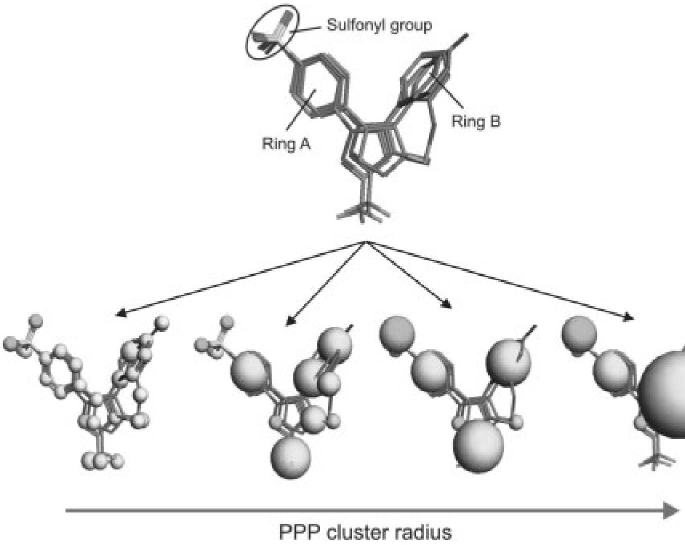


Fig. 3.10 SQUID fuzzy pharmacophore model for COX-2 inhibitors. Using a larger cluster radius results in more general models. From left to right: 1 Å, 1.5 Å, 2.5 Å, 3.5 Å.

the lowest abstraction from the scaffolds of the molecules in the alignment. Using larger cluster radii leads to pharmacophore models with higher degrees of generalization until, like the model resulting from 3.5 Å, the underlying alignment is only marginally visible. For virtual screening it must be tested for each target and each set of molecules in an alignment which degree of fuzziness results in molecules that are most likely to be active or possess some desired characteristics. Retrospective screening for known active molecules using models with different resolutions can serve this purpose.

In Fig. 3.11, the CV of the best found COX-2 fuzzy pharmacophore model with a cluster radius of 1.4 Å is shown in comparison with the scaled CATS3D vectors of the underlying molecules from the alignment. As one can see, the fuzzy and thus "generalizing" representation of the underlying molecules from the alignment is retained in the CV. It becomes clear that the SQUID CV and the CATS3D CVs differ significantly in the meaning of their content. The SQUID CV describes a broad range of descriptor areas which are favorable for the desired biological activity, whereas the CATS3D descriptor contains a smaller subset of the actual occurrences of atom pairs in a specific ligand. Consequently, commonly used similarity indices such as the Euclidean distance or the Tanimoto index, which are based on the assumption that both descriptors which are to be compared represent objects in the same way, cannot be used to assess the activity of the molecules under consideration. To overcome this problem, a SQUID similarity score was developed:

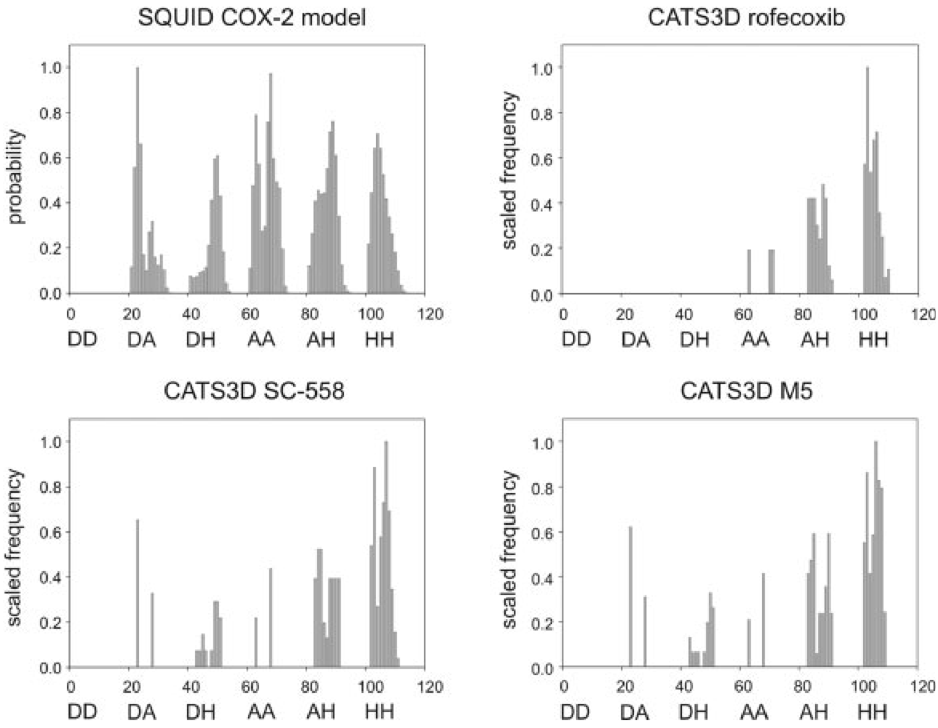


Fig. 3.11 Comparison of the correlation vectors of a SQUID COX-2 pharmacophore model and the CATS3D correlation vectors of the three molecules used for the calculation of the pharmacophore model.

$$S(a,b) = \frac{\sum_{i=1}^{n} a_i b_i}{1 + \sum_{i=1}^{n} [(1 - a_i) b_i]} \tag{4}$$

where a_i is the value of the ith element of the SQUID CV, b_i is the value of the ith element of a molecule CV and n is the total number of dimensions. The value a_i may be considered as the idealized probability of the presence of features in b_i. This results in high scores for molecules with many features in regions of the query descriptor which have a high probability. To penalize the presence of such atom-pairs in regions with a low probability, the denominator weights the presence of atom pairs with the inverted probabilities of the descriptor of the pharmacophore model (Note: a value of 1 was added to the denominator to avoid division by zero and high scores resulting from a very low value in the denominator of the term.) Accordingly, the SQUID scores for the CVRs from the

COX-2 inhibitors shown in Fig. 3.11 decrease in the order M5 > SC-558 > rofecoxib.

Using SQUID, it had turned out that the overall probabilities for the presence of pharmacophoric type pairs (e.g., the probability to find a hydrophobic–polar pair, irrespective of the distance between the atoms) were often not optimal to retrieve isofunctional ligands. For that purpose, a training step was included to optimize these probabilities prior to the final virtual screening experiment. In retrospective screening, an optimized COX-2 SQUID fuzzy pharmacophore model outperformed CATS3D similarity searching using the single molecules from the molecular alignment (Fig. 3.12).

SQUID fuzzy pharmacophores were applied to prospective virtual screening with the aim of retrieving molecules inhibiting the Tat–TAR RNA interaction which is crucial for HIV replication [63]. The pharmacophore model was built up from one ligand (acetylpromazine, IC_{50}=500 µM) and a fragment of another known ligand (CGP40336A), which was assumed to bind with a comparable binding mode as acetylpromazine. Using the optimized pharmacophore model in Fig. 3.13 a, the 20 000 most drug-like molecules from the Specs database [64] were screened for Tat–TAR ligands. A set of 10 molecules was selected for experimental testing. In a fluorescence resonance energy transfer (FRET) assay, the best hit showed an IC_{50} value of 46 µM, which represents an approximately 10-fold improvement over the reference acetylpromazine. An alignment of this hit with the reference alignment is shown in Fig. 3.13 b. For comparison, CATS3D similarity searching was also applied using the two reference molecules from the pharmacophore. This resulted in a best ligand with an IC_{50} comparable to acetylpromazine (IC_{50}=500 µM).

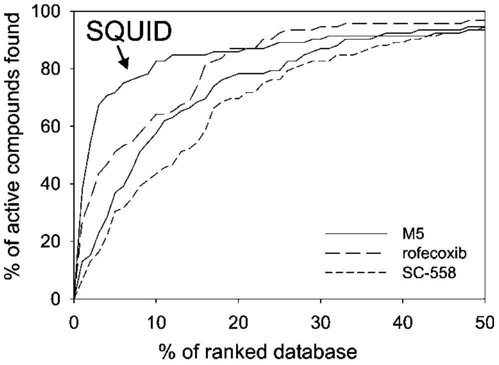


Fig. 3.12 Comparison of the enrichment curves of the SQUID COX-2 model with CATS3D retrospective screening with the molecules used for pharmacophore model calculation.

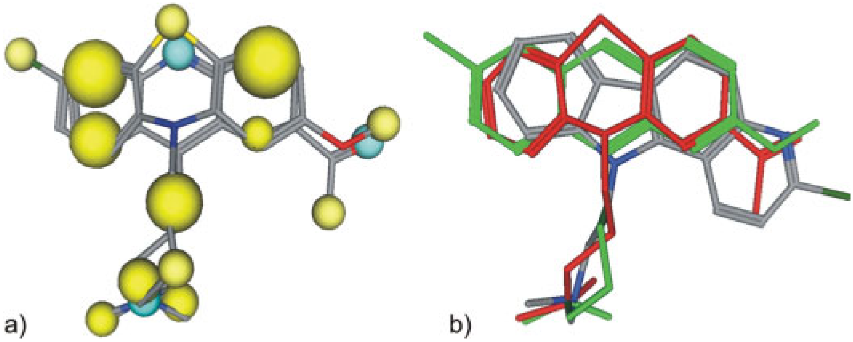


Fig. 3.13 (a) SQUID fuzzy pharmacophore model of Tat–TAR interaction inhibitors. (b) Alignment of the best hit (colored by atom types) found with SQUID with the reference alignment (red, acetylpromazine; green, fragment of CGP40336A).

3.4.2 Feature Point Pharmacophores: FEPOPS

Recently, Jenkins et al. reported another approach, based on the pre-calculation of more generalized representations of substructures, which are utilized for chemical similarity computation [7]. In the FEPOPS (*fe*ature *po*int *p*harmacophore*s*) approach, atoms of single molecules are clustered into four *representative* PPPs using the *k*-means algorithm. Each atom is then assigned to the nearest PPP. For each of the four PPPs, the sum of the partial charges and the sum of the atomic log*P* (Alog*P*) values are calculated. The presence of hydrogen-bond acceptors and donors is represented by two binary values. The four points are sorted by increasing sum of partial charges, to obtain a defined alignment rule for the quartets. The values of the four points and the values of all six distances are then combined to a vector representation of the molecule. Ultimately, the vectors are mean centered and scaled to unit variance. Similarity searching was performed using the Pearson correlation coefficient as a distance measure. The method was extensively tested in retrospective screening studies. Using scaffold definitions with the program Meqi [47], FEPOPS performed well in finding scaffolds that are different from the reference molecules in comparison with several other established methods.

3.5 Conclusions

In a growing number of studies, correlation-vector representations of pharmacophoric features of molecules have been proven to be useful in similarity searching and *de novo* design projects. They allow for rapid database screening, which is gained by avoiding an explicit pairwise alignment step. Such methods seem to be particular useful in early stages of the lead-finding process when only one or some few reference compounds ("templates", "seeds") are known. Owing to limitations of current definitions of "pharmacophoric features", these methods are rather coarse-grained, i.e. they perform similarity estimations between molecules without taking into consideration worked-out structure–activity relationship models. The CV encoding scheme represents a compromise between accuracy and speed, which renders most of these approaches unsuitable for lead optimization. Keeping these limitations in mind, alignment-free CV methods have found their place in pharmacophore-based virtual screening. They complement existing techniques by retrieving additional hit and lead candidates which would not be found otherwise. Future developments of CV approaches will have to provide more elaborate weighting schemes of pharmacophore points and a method for representing *quantitative* structure–activity relationship models.

Acknowledgments

This work was supported by the Fonds der Chemischen Industrie and the Beilstein-Institut zur Förderung der Chemischen Wissenschaften. S.R. and U.F. are grateful for fellowships from Merz Pharmaceuticals and Aventis Pharma Deutschland, respectively.

Abbreviations

2D	two-dimensional
3D	three-dimensional
ADMET	absorption, distribution, metabolism, excretion, toxicity
ATS	autocorrelation of a topological structure
CATS	chemically advanced template search
CV	correlation vector
FEPOPS	feature point pharmacophores
GPCR	G-protein coupled receptor
mGluR	metabotropic glutamate receptor
PPP	potential pharmacophore point
QSAR	quantitative structure–activity relationship
SQUID	sophisticated quantification of interaction distributions

References

1 C.G. Wermuth, C.R. Gannelin, P. Lindberg, L.A. Mitscher, *Pure Appl. Chem.* **1998**, *70*, 1129.

2 A.M. Doweyko, *J. Comput.-Aided Mol. Des.* **2004**, *7–9*, 587.

3 R.E. Carhart, D.H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.

4 S. Picket, in *Protein–Ligand Interactions – From Molecular Recognition to Drug Design* (Eds H.-J. Böhm, G. Schneider), Wiley-VCH, Weinheim, **2003**, pp. 73–106.

5 A.C. Good, S.-J. Cho, J.S. Mason, *J. Comput.-Aided Mol. Des.* **2004**, *7–9*, 523.

6 G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894.

7 J.L. Jenkins, M. Glick, J.W. Davies, *J. Med. Chem.* **2004**, *47*, 6144.

8 P. Willett, *J. Med. Chem.* **2005**, *48*, 4183.

9 M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc.* **1995**, *117*, 7769.

10 R. Todeschini, V. Consonni, *Handbook of MolecularDescriptors*, Wiley-VCH, Weinheim, **2000**.

11 G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 359.

12 G. Moreau, P. Broto, C. Vandycke, *Eur. J. Med. Chem.* **1984**, *19*, 66.

13 H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205.

14 M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, *J. Med. Chem.* **2000**, *43*, 3233.

15 P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849.

16 R.P. Sheridan, M.D. Miller, D.J. Underwood, S.K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128.

17 R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572.

18 N. Stiefl, K. Baumann, *J. Med. Chem.* **2003**, *46*, 1390.

19 Y.C. Martin, R.D. Brown, M.G. Bures, *Combinatorial Chemistry and Molecular Diversity*, Wiley, New York, **1998**.

20 T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, *Introduction to Algorithms*, 2nd edn, MIT Press, Cambridge, MA, **2001**.

21 E.W. Dijkstra, *Numer. Math.* **1959**, *1*, 269.

22 R.W. Floyd, *Commun. ACM* **1961**, *4*, 42.

23 B.L. Bush, R.P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756.

24 MOE, Molecular Operating Environment, Chemical Computing Group, Montreal, http://www.chemcomp.com.

25 T. Clark, *J. Mol. Graph. Model.* **2004**, *22*, 519.

26 A. Bender, H.Y. Mussa, G.S. Gill, R.C. Glen, *J. Med. Chem.* **2004**, *47*, 6569.

27 J.M. Barnard, G.M. Downs, P. Willett, in *Virtual Screening of Bioactive Molecules* (Eds H.J. Böhm, G. Schneider), Wiley-VCH, Weinheim, **2000**, p. 59.

28 G. Schneider, M. Nettekoven, *J. Comb. Chem.* **2003**, *5*, 233.

29 A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391.

30 M. Stahl, M. Rarey, G. Klebe, in *Bioinformatics: From Genomes to Drugs*, Vol. 2 (Ed. T. Lengauer), Wiley-VCH, Weinheim, **2001**, p. 137.

31 U. Fechner, G. Schneider, *ChemBiochem* **2004**, *5*, 538.

32 H. Xu, D.K. Agrafiotis, *Curr. Top. Med. Chem.* **2002**, *2*, 1305.

33 U. Fechner, L. Franke, S. Renner, P. Schneider, G. Schneider, *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687.

34 P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713.

35 (a) J. Gasteiger, C. Rudolph, J. Sadowski, *Tetrahedron: Comput. Methods* **1990**, *3*, 537; (b) Molecular Networks, Erlangen, http://www.mol-net.de.

36 J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177.

37 J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.* **2004**, *2*, 3256.

38 U. Fechner, G. Schneider, *QSAR Comb. Sci.* **2004**, *23*, 19.

39 D. Horvath, B. Mao, *QSAR Comb. Sci.* **2003**, *22*, 489.

40 D. Horvath, in *Combinatorial LibraryDesign and Evaluation: Principles, Software Tools and Applications* (Eds A. Ghose, V.

Viswanadhan), Marcel Dekker, New York, **2001**, p. 429.

41 J. Paetz, in *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, Chicago, IL, **2003**, p. 14.

42 U. Fechner, J. Paetz, G. Schneider, unpublished results.

43 S. Renner, C. Schwab, J. Gasteiger, G. Schneider, unpublished results.

44 E. Perola, P.S. Charifson, *J. Med. Chem.* **2004**, *47*, 2499.

45 (a) C. Schwab, in *Handbook of Cheminformatics* (Ed. J. Gasteiger), Wiley-VCH, Weinheim, **2003**, p. 262; (b) Molecular Networks, Erlangen, http://www.mol-net.de.

46 G.W. Bemis, M.A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887.

47 (a) Y. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181; (b) Y.J. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912.

48 H.L. Morgan, *J. Chem. Soc.* **1965**, *5*, 107.

49 L. Naerum, L. Norskov-Lauritsen, P.H. Olesen, *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525.

50 G. Schneider, M.-L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487.

51 G. Schneider, O. Clement-Chomienne, L. Hilfiger, P. Schneider, S. Kirsch, H.J. Bohm, W. Neidhart, *Angew. Chem. Int. Ed.* **2000**, *39*, 4130.

52 X.Q. Lewell, B.J. Duncan, S.P. Watson, M.M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511.

53 *Derwent World Drug Index*, Derwent Information, London, Nov. 1998.

54 M. Roger-Evans, A.I. Alanine, K.H. Bleicher, D. Kube, G. Schneider, *QSAR Comb. Sci.* **2004**, *23*, 426.

55 S. Renner, T. Noeske, C.G. Parsons, P. Schneider, T. Weil, G. Schneider, *ChemBiochem* **2005**, *6*, 620.

56 G. Schneider, P. Schneider, in *Chemogenomics in Drug Discovery* (Eds H. Kubinyi, G. Müller), Wiley-VCH, Weinheim, **2004**, p. 341.

57 ASINEX, Moscow, http://www.asinex.-com.

58 H. Matter, *J. Med. Chem.* **1997**, *40*, 1219.

59 Y.C. Martin, J.L. Kofron, L.M. Traphagen, *J. Med. Chem.* **2002**, *45*, 4350.

60 S. Renner, G. Schneider, *J. Med. Chem.* **2004**, *47*, 4653.

61 O. Güner, *Pharmacophore Perception, Development and Use in Drug Design*, International University Line, La Jolla, CA, **2000**.

62 A. Palomer, F. Cabre, J. Pascual, J. Campos, M.A. Trujillo, A. Entrena, M.A. Gallo, L. Garcia, D. Mauleon, A. Espinosa, *J. Med. Chem.* **2002**, *45*, 1402.

63 S. Renner, V. Ludwig, O. Boden, U. Scheffer, M. Göbel, G. Schneider, *ChemBiochem* **2005**, *6*, 1119.

64 Specs, Delft, http://www.specs.net.

65 S. Renner, G. Schneider, *Chem. Med. Chem.* **2006**, *1*, 181.

4
Feature Trees: Theory and Applications from Large-scale Virtual Screening to Data Analysis

Matthias Rarey, Sally Hindle, Patrick Maaß, Günther Metz, Christian Rummey, and Marc Zimmermann

4.1
Introduction: from Linear to Non-linear Molecular Descriptors

Based on the ideas of Paul Ehrlich [1] and Emil Fischer [2], pharmacophores and molecular similarity became the most prominent and frequently used concepts in molecular design. In the early days of molecular design, the absence of protein structures in atomic detail was the major motivation for comparing small molecules. Owing to the tight binding of bioactive compounds to their specific receptor, the compound can act as a partial negative imprint of the active site. Molecules similar in their physico-chemical properties relevant for binding therefore have a high chance of also showing a similar binding profile with respect to proteins.

Although nearly 29 000 protein structures were available as of March 2005 (http://www.rcsb.org/pdb/) and protein structure-based design techniques are available, there are several applications making similarity-based methods a key technology in molecular design. For whole classes of pharmaceutically interesting target proteins such as GPCRs or ion channels, a protein structure with atomic resolution is still out of reach. Here, similarity-based methods are without alternative. Furthermore, molecular similarity plays an important role in target-unrelated pre- and post-processing steps such as library design and diversity analysis, prediction of ADME properties or drug/lead likeness and the analysis of screening results via clustering around active compounds. The concept of molecular similarity underlies a large variety of computational techniques for molecular design, ranging from pharmacophore elucidation via structural alignment of molecules to descriptor-based similarity searching. In this chapter, we focus on a certain descriptor technology called Feature Trees.

Descriptors are widely used for efficient retrieval of similar compounds and also for clustering and property prediction (see [3] for a recent review). The task of the descriptor is to represent a compound such that a biologically (or chemically) relevant similarity can be deduced efficiently from the comparison of two descriptors with a computer. The difficulty in developing a descriptor is, therefore, to find a good trade-off between the coverage of important physico-chemical properties

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

and efficient computability. For the latter, the format of the descriptor is of importance. Most descriptors in use today have a linear format: individual properties are calculated from the compound and stored in a vector. These properties can be expressed by numbers (e.g. molecular weight, log*P*) or by booleans (such as the absence or occurrence of a chemical fragment). Since each individual property represents the whole compound, it is necessary, when comparing two compounds, to compare the corresponding properties and calculate an overall similarity value from simple numerical equations. This *alignment-free* comparison algorithm is extremely fast, but it also bears some severe disadvantages. Studying the process of binding reveals that the relative arrangement of functional groups on the molecular surface plays a dominant role. This relative arrangement, however, is only weakly described in linear descriptors. Only on the basis of an alignment can the relative arrangement be adequately considered. Furthermore, the information concerning a certain part of a molecule is covered by several values in the vector. This makes it difficult to apply these descriptors to combinatorial sets of compounds such as combinatorial libraries or chemical fragment spaces.

At the other extreme, a three-dimensional (3D) model of the molecule itself can be considered as a descriptor. In order to compare them, the molecules have to be aligned in 3D space, which is a difficult task, mostly owing to the conformational flexibility of most compounds of interest. Such 3D alignment-based comparisons of molecules are therefore time intensive and bear the risk of missing the right alignment.

The question arises as to whether compounds can be compared based on an alignment but without the necessity to deal with conformational flexibility, i.e. whether we can develop an *alignment-based* but *conformation-independent* descriptor. The Feature Tree [4] is an attempt to achieve this goal. The descriptor is based on a reduced representation of the molecular graph as proposed earlier [5–7]. Although the comparison is more difficult than with linear descriptors owing to the necessity for calculating the alignment, the algorithms are still efficient enough to look at large data sets. With several examples, it can be shown that the descriptor preserves the global arrangement of functional groups within the molecule without depending too much on the molecular graph. In the following, we will describe the Feature Tree descriptor and the algorithms for creating and comparing them. We will then summarize several applications from virtual screening via chemical fragment space search and HTS data analysis to similarity-driven visualization of compounds.

4.2 Creating Feature Trees from Molecules

A Feature Tree represents a molecule by a tree structure. The tree should capture the major building blocks of the molecule in addition to their overall arrangement. Detailed information of less importance for protein binding such as the molecular graph should be neglected. In this way, not only is the complexity of the comparison problem reduced, but also so-called lead hopping between chemical classes with compounds sharing the same wanted biological activity is supported. In order to circumvent the problem of dealing with conformational spaces, 3D information is neglected. Since most compounds do have rings, the question arises as to why the molecule should be represented by a tree. The tree structure has a striking advantage when it comes to comparison algorithms: Whenever a single edge of the tree is removed, the tree falls into two well-defined pieces. In Section 4.3, three efficient pairwise comparison algorithms will be explained. All of them are based on this special feature of tree structures. Fortunately, although most compounds have rings, most of them are still tree-like. As long as only short cycles occur, each cycle can be considered as a building block and is then represented by a single node in the tree. However, the tree representation is inadequate for long macrocycles and large, highly bridged ring systems such as fullerenes.

In order to create a Feature Tree, the major task is the division of the molecule into building blocks. In a first step, each bond of the molecule is marked as *terminal*, *cyclic* or *acyclic*. End-standing bonds have an incident atom with only one bond which is easy to detect. In order to detect cyclic bonds, a depth-first search algorithm customized for detecting biconnected components in graphs can be applied [8]. After cutting all acyclic bonds, a first set of building blocks is defined. All atoms not contained in any ring form a building block together with the connected end-standing atoms.

The second step comprises the division of the ring systems if possible. For each atom within a ring system, the shortest cycle containing the atom is determined with another depth-first-search algorithm. After removing duplicates from the set of rings, a *cycle graph* is constructed. Each cycle forms a node in the cycle graph and two nodes are connected by an edge if the corresponding cycles share at least one atom. Within the cycle graph, the depth-first-search algorithm for detecting biconnected components can be applied again. Nodes forming a cycle in the cycle graph cannot be separated from each other and are therefore considered as a single building block. Acyclic nodes in the cycle graph are considered as individual building blocks. Note that the process described above does not reflect a division based on the smallest set of smallest rings (SSSR). In contrast to SSSR, the division process within Feature Trees is unique. This is necessary in order to guarantee the detection of identical compounds. The process of dividing cycles is shown in Fig. 4.1. It should be further noted that not every atom is assigned to a single building block. Within a ring system, an atom might be assigned to two neighboring building blocks in the case that the atom is contained in two rings. Once the building blocks are defined, the Feature Tree can be easily constructed. Each building block is a single node in the Feature Tree. Two nodes are connected by an edge if there is an atom or two adjacent atoms covered by the corresponding building blocks.

Finally, the Feature Tree nodes are marked with labels describing the shape and chemical properties of the building block. In principle, every kind of descriptor can be used as a label provided that the descriptor is additive over the building blocks. In our Feature Tree implementation, we normally work with a shape

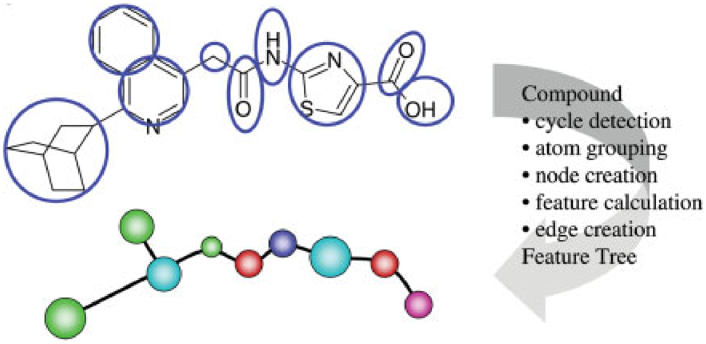


Fig. 4.1 The conversion of a molecule into a Feature Tree descriptor. The major phases are summarized on the right. A cyclic system is divided into individual rings only if this can be done uniquely. Features stored at the Feature Tree nodes are shape or chemistry related. The chemical feature (interaction potential) is color coded as follows: red, H-bond acceptor; blue, H-bond donor; green, hydrophobic.

and a chemistry descriptor. The shape descriptor that we use has two components: a ring closure count and an approximated van der Waals volume, which is the volume of the van der Waals spheres less the sphere overlap along covalent bonds. In addition, a path length descriptor can be used, which is not described here (see [9]). The purpose of the chemistry descriptor is to reflect the interaction pattern that a building block can form with a surrounding protein. In order to do so, a profile of potential interactions is derived from the building block. The FlexX interaction scheme is employed, which represents hydrogen bond donors and acceptors and three subtypes of hydrophobic interactions (for details, see [10]). Both the shape and the chemistry descriptor are obviously additive. For the comparison algorithms, a function is required for calculating a similarity value between pairs of shape and chemistry descriptors. If $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ are the descriptor vectors, we use the following equation, motivated by the idea that the minimum defines the number of features both molecules have in common:

$$c(a,b) = \begin{cases} 1 & \text{if } \sum\limits_i a_i + b_i = 0 \\ \dfrac{2\sum\limits_i \min(a_i, b_i)}{\sum\limits_i a_i + b_i} & \text{otherwise} \end{cases} \tag{1}$$

Similarities between a shape descriptor and a chemistry descriptor may be combined to calculate the final similarity $\text{sim}(m)$ for a match m of nodes representing two building blocks.

4.3 Algorithms for Pairwise Comparison of Feature Trees

Once a Feature Tree can be created from a molecule, the question arises of how to compare two Feature Trees. Using Eq. (1), we are able to compare two individual Feature Tree nodes. Owing to the additivity of the features stored at a node, we can also compare two sets of Feature Tree nodes. This is done by adding the features over all nodes within a set and applying Eq. (1) again. Obviously, we can also compare two complete Feature Trees in this way: we just add all features in the two trees and apply Eq. (1). We call such a comparison *level-0,* because no division of the tree into pieces has been performed. Level-0 comparisons closely resemble the way linear descriptors work. If we assume for a moment that all components of a linear descriptor are additive and can be computed for each building block individually (such as the volume descriptor), adding the feature values over all Feature Tree nodes will create the linear descriptor.

The quality of a descriptor can be significantly improved if features are compared locally with consideration of the molecule's topology rather than globally. In order to achieve such a local comparison, we have to compute which molecule parts should be compared with which; in other words, we have to compute an alignment (or matching) between molecule parts. We call a comparison based on such a subdivision of the molecule a *level-x* comparison. Since different matchings will result in different similarity values, determining the matching with the highest possible similarity value becomes an optimization problem.

In the computer science literature, several algorithms for matching trees can be found [11–14]. They all perform a node-to-node matching between the trees. In a Feature Tree, a node can represent building blocks of variable size from a single atom to a ring system. A node-to-node matching is therefore inappropriate. In order to match molecule parts of roughly the same size, small sets of connected nodes, so-called *subtrees,* must be matched with each other. In this way, a small chain can be matched with a ring of roughly the same size although the ring might be represented by a single node and the chain by a series of 2–3 nodes.

A match of subtrees should reflect the fact that certain parts of the molecules interact with the same sub-pocket of a protein. Let us now consider a set of matches $m_1 = (a_1, b_1)$, $m_2 = (a_2, b_2)$ and $m_3 = (a_3, b_3)$. The matched subtrees must be arranged in the molecules A and B such that the corresponding molecule parts can interact with the same respective sub-pockets simultaneously. Whether such an arrangement is possible can only be answered in 3D space. Nevertheless the topology of the trees can give us a good estimate. If in molecule A the subtrees are in the order a_1–a_2–a_3 and in molecule B in the order b_1–b_3–b_2, it is unlikely (but not impossible) that conformations for molecules A and B exist which allow the placement of all three molecule parts in the same respective sub-pockets. A matching is called *topology-maintaining* if, for all pairs m_1, m_2 of matches, the following holds: a_1 is connected to a_2 by a path containing only unmatched nodes, if and only if this is also true for b_1 and b_2.

Once a topology-maintaining matching M of subtrees has been calculated, an overall similarity value can be derived. For each match, we calculate the similarity value as already described above for a level-0 comparison. For the whole matching, a size-weighted average over all matches gives the final similarity value

$$S_m(A, B) = \frac{\frac{1}{2}\sum\limits_{m\in M} \mathrm{size}(m)\mathrm{sim}(m)}{\omega \max[\mathrm{size}(A),\, \mathrm{size}(B)] + (1-\omega)\min[\mathrm{size}(A),\, \mathrm{size}(B)]} \tag{2}$$

in which size( ) gives the number of non-hydrogen atoms covered by the matched subtrees or in the whole molecule A or B and sim( ) gives the similarity value for the individual matches. [Note that Eq. (2) has changed since the original publication in 1998. The new form has the advantage of being independent of the matching size, thus resolving optimization problems already mentioned in the 1998 publication.]

The parameter ω allows the similarity value to be tuned towards global or partial similarity. To illustrate the effect, let us assume that molecule A might be fully contained in B and $T=\mathrm{size}(A)=\mathrm{size}(B)/2$; 100% of molecule A will be matched to 50% of B with similarity value $\mathrm{sim}(m)=1$, resulting in a numerator T. If $\omega=1$, the denominator is $2T$ resulting in an overall similarity value of 0.5. If, however, $\omega=0$, the denominator is T, resulting in an overall similarity value of 1.0.

In the following, we roughly describe the three algorithms available for calculating a subtree matching of Feature Trees. We intend to present the overall idea here rather than covering every detail of the algorithm. For the latter, we refer to the original publications [4, 15].

4.3.1
Recursive Division: the Split-search Algorithm

The first algorithm developed for Feature Tree matching follows the basic principle of “divide and conquer”. The algorithm will subsequently divide the Feature Trees into smaller subtrees until the subtree size falls below a certain value. For an explanation of this algorithm, we have to define the division process clearly. By cutting a tree edge, the tree is broken into two subtrees. A *directed cut* is an edge $e=(a,b)$ together with a direction from one node a to another node b. A *split* is a pair of directed cuts belonging to two different Feature Trees. A split defines four subtrees, two in each Feature Tree. It also defines two subtree matches, the first match consists of the subtrees containing the *from*-nodes and the second consists of the subtrees containing the *to*-nodes (see Fig. 4.2 for an illustration of splits). Based on the similarity equation given above, a split can be scored by calculating the similarity values of the two implicitly defined matches.

The split-search algorithm recursively divides the two Feature Trees by introducing splits. At every stage of the recursion, all splits which provide a topology-maintaining matching will be scored. The three best-scoring splits are subsequently considered. For each of them, the split-search algorithm is recursively called twice, once for each pair of subtrees defined by the split. If the size of a subtree falls below a certain threshold, the recursion is stopped and the similarity value is returned. The split-search algorithm stores the splits which resulted in the highest similarity value and returns the similarity value to the calling function.

From an algorithmic point of view, two challenging tasks have to be resolved here. First, an efficient algorithm for scoring all splits must be developed. Second, a function is necessary to distinguish splits which result in topology-maintaining matchings from those which do not. For both tasks, appropriate solutions are given in the original publication [4].

A few aspects of the split-search algorithm are worth mentioning. First, and most importantly, the algorithm is heuristic, since not all possible splits but only the three best-scoring ones are evaluated. Second, the algorithm performs redundant calculations: a set of splits can result in the same matching independent of the order of the splits. The split-search algorithm creates the splits in a certain order and, therefore, potentially creates a single matching multiple times. Third, the split-search algorithm allows leaving nodes unmatched in the middle of the Feature Trees without penalizing them. Despite these deficiencies, the split-search algorithm is extremely fast and works well in virtual screening exercises.

4.3.2 Subsequently Growing Matchings: the Match-search Algorithm

The most problematic issue of the split-search algorithm as described above is that unmatched nodes may occur between matched nodes. This “gap”, also called an *inner-NIL match*, becomes harder to justify the larger it gets since the two matched parts of the molecule are assumed to interact with the same subpockets of an active site (see also Fig. 4.3). In the following, we will assume that inner-NIL matches are forbidden – which makes the development of an alternative matching algorithm necessary. We will first describe the new algorithm, called *match-search*, in a recursive fashion which operates on two trees *A* and *B*.

The first step of the algorithm is identical with the initial call of the split-search algorithm, namely the search for a small set of high-scoring initial splits. The match-search algorithm then iterates through the list of splits performing the following calculations for each split. The split produces two subtree-matches, one on the from-node and the other one on the to-node side. The from- and to-nodes adjacent to the respective cuts are called *head-nodes* in the following. The algorithm refines these two subtree matches independently following the same strategy. For both subtrees from *A* and *B* of a match, all subtrees smaller than a certain size limit and including the head-node are enumer-

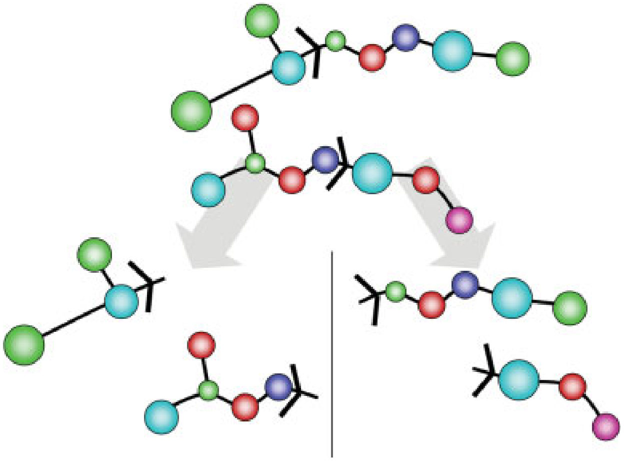


Fig. 4.2 A split is a pair of directed cuts (black wedges). A directed cut is an edge $e=(a,b)$ together with a direction from one node a to another node b. The split automatically defines a division of two trees as well as the assignment of subtrees (owing to the directionality).

ated. These subtrees are matched and scored and a small set of high scoring subtree matches are kept. For each of these subtree matches, a series of cuts are necessary to separate the newly formed subtree from the rest. As a result of this step, we have a match of two subtrees containing the head nodes and a series of new subtrees from A and also from B lying behind the matched part (see also Fig. 4.4). Every new subtree from A is now matched with every new subtree from B. For all these subtree matches, similarity values can be computed by recursively applying the match-search algorithm. Finally, the combination of matches resulting in the highest similarity value is chosen and the similarity value is returned.

An implementation of the above algorithm in this recursive fashion would have an exponential asymptotic runtime behavior. A simple observation shows that such high computing demand is unnecessary. If two trees A and B with n_A and n_B nodes, respectively, are compared, only $4(n_A–1)(n_B–1)$ different calls of

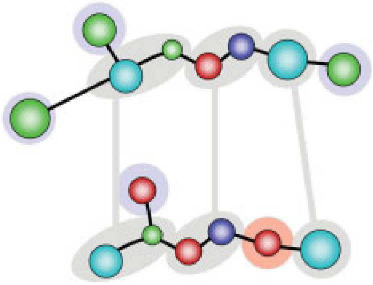


Fig. 4.3 A matching assigns subtrees of one Feature Tree to subtrees of another (gray ellipsoids). An unmatched subtree or node is either endstanding (blue) or an inner node (red). An unmatched inner node is called an inner-NIL match.

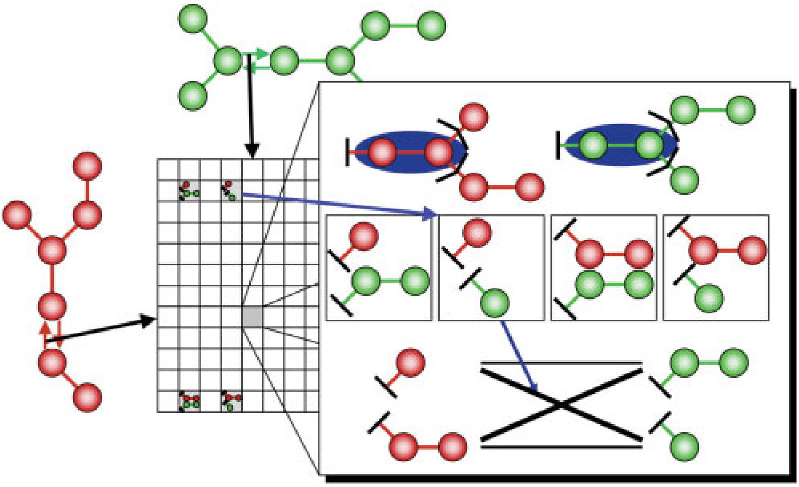


Fig. 4.4 The match search algorithm creates a matrix with one cell for each pair of directed tree edges. The cell stores the overall similarity of the two subtrees. The similarity value is calculated with a dynamic programming scheme shown on the right. First, an extension match (blue ellipsoid) is searched. Then the subtrees are cut and matched in all possible combinations. For each combination, a similarity value can be extracted from the matrix (exemplarily shown by the blue arrows). A maximum-weight bipartite matching solves the assignment of the subtrees.

the match-search algorithm can be performed, since every call starts with a certain split. We only have to cache the results from the match-search algorithm in a $2(n_A{-}1)\times 2(n_B{-}1)$ matrix and reuse them once the result is available. This technique converts the recursive algorithm into a dynamic programming scheme with polynomial runtime.

The match-search algorithm forbids inner-NIL matches and therefore produces other matchings than the split-search algorithm. Its runtime is dominated by the search of the optimal assignment of subtrees. Trees with a high node degree can cause long computing times. On typical drug-like compounds, however, the algorithm computes the matching within milliseconds.

4.3.3 Match-Search with Gaps: the Dynamic Match-search Algorithm

The match-search algorithm described above works well for similar Feature Trees of equal sizes or if one tree is fully contained in the other tree. However, as the algorithm cannot generate inner-NIL matches, variable linker regions between pharmacophoric groups cannot be modeled (see Fig. 4.5).

Therefore, a further improvement of the match-search algorithm was developed: the *dynamic match-search* algorithm. The new algorithm extends the dy-

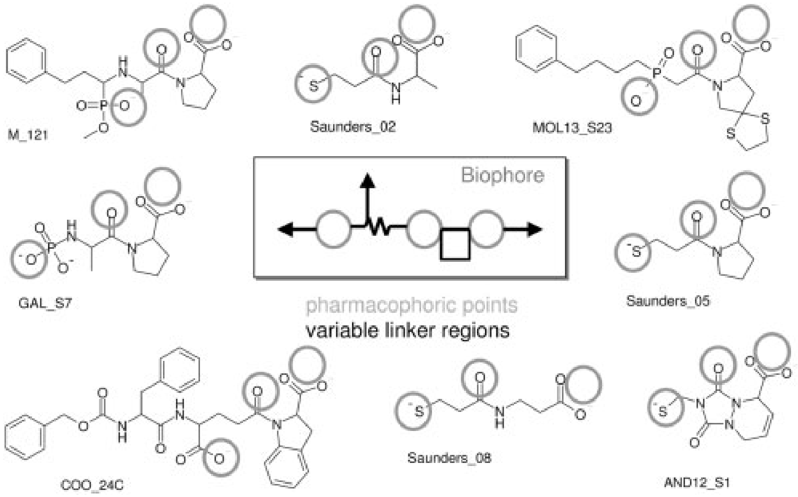


Fig. 4.5 Examples of different scaffolds for ACE inhibitors having three pharmacophoric features separated by variable linker regions.

namic programming scheme of the match-search algorithm by allowing gaps (as in sequence comparison algorithms [16]). A gap corresponds to an inner-NIL match. The penalty score of each inner-NIL match depends on the size of the skipped subtree. Instead of using a recursive procedure, the algorithm computes every cell of the dynamic programming matrix (being the cache from Section 4.3.2) in a bottom up fashion. First, the terminal nodes of both Feature Trees are compared and matched in all possible combinations by introducing a split after each node. Then, the initial matches are extended by either a match-, merge- or gap-operation. A match-operation aligns the nodes adjacent to the previous match. A merge-operation aligns subtrees neighboring the previous match ignoring the topology within the matched subtrees. This operation therefore introduces a certain degree of fuzziness into the matching process. A gap-operation allows a subtree in one of the two trees to be skipped. In each cell of the matrix, the result of the three possible operations which gives the best similarity score for the currently constructed matching will be chosen. In the case that a branching node is met, the dynamic match-search algorithm has to decide which outgoing edge of one subtree should be mapped to which outgoing edge of the second. Every possible combination has to be evaluated and the best one chosen. The decision can be made when all ingoing edges are already computed and are, therefore, present in the dynamic programming matrix. In order to handle nodes with a high number of outgoing edges, a bipartite matching procedure [17] is used to solve efficiently the problem of finding the best assignment.

The algorithm stops when the trees are completely covered by matches. After having computed the entire dynamic programming matrix, an optimal matching can be extracted using a back-tracking procedure to follow the path of the highest local match scores. This procedure will find every possible alignment between two unrooted trees. Unlike the other two algorithms, it does not rely on the heuristics of choosing an initial split. Whether the resulting matching is topology-maintaining can be easily checked in each of the three possible match-extension operations. The runtime of the dynamic match-search algorithm scales quadratic with the size of the Feature Trees (the size of the dynamic programming matrix is growing in the number of edges) and the maximum node degree (bipartite matching algorithm).

4.3.4 Building Multiple Feature Tree Models

With an algorithm for comparing Feature Trees in hand, we can now describe how to build multiple Feature Tree (MTree) models. Based on the matching calculated during the comparison of two trees, a new tree combining the information from both input Feature Trees can be created. The nodes of the new tree are the matches containing the features of the mapped subtrees of the two trees. The edges are formed following the topology of the input Feature Trees. The resulting MTree model has the same structure as a Feature Tree. Using the same matching algorithms, an MTree model can be compared to other MTree models or Feature Trees.

In order to build an MTree model from more than two Feature Trees, the dynamic match-search algorithm can be applied in a hierarchical manner. We developed two efficient heuristics for this task:

- The first strategy is to assemble a hierarchical model in a bottom-up fashion based on a cluster dendrogram of the Feature Trees. Therefore, in the first step, a cluster dendrogram of the Feature Trees is created. This is done by a single linkage clustering algorithm using the pairwise similarity scores of the Feature Trees. Starting at the bottom of the dendrogram, the molecules are pairwise combined into a model. Each model is defined by an MTree, which can be used like any other Feature Tree for comparisons. At the next hierarchy level, the models (MTree models) are also combined pairwise to create new models. For the computation of the dendrogram of n trees, $n(n–1)/2$ comparisons are required. The complete MTree model of all Feature Trees can then be computed with another n–1 comparisons.
- The second strategy is to add incrementally individual Feature Trees to the model (starting with two Feature Trees). In each step, the most similar Feature Tree to the then-current model is chosen. This strategy affords $n–i$ comparisons in the ith step and n–1 steps altogether until a single MTree model remains.

MTree models constructed in this way can be used for screening purposes. By merging the information of the underlying trees into an MTree model, virtual screening can be done by simple pairwise comparisons. Hence we do not have to compare the ranks of several virtual screening runs comparing individual query molecules with each database molecule.

Another advantage is that the matches in a model can be weighted by the local similarity of the corresponding subtrees in those matches, thereby pronouncing certain elements of the model. An exemplary application is provided in Section 4.6.

4.4 Feature Trees in Similarity Searching and Virtual Screening

When it comes to selection strategies involving large numbers of compounds, the term virtual screening (VS) has been used to describe the task of data reduction to a manageable size [3, 18–21]. The number of compounds to be compared in a reasonable timeframe influences the compromise between descriptor accuracy (2D, 3D, multiconformer 3D) and speed. Large databases of commercially available compounds have been assembled and, unlike real compound collections, such databases are easy to maintain. These searchable collections can be extended to chemically accessible virtual libraries vastly exceeding the size of any physical compound collection. Generally, fast ligand-based methods facilitate the screening of millions to billions of compounds and can be used as a prefilter for more sophisticated 3D-based algorithms. Of course, active compounds need to be at hand from either in-house data, literature search or patent information. Efforts in virtual screening then focus on finding either close structural analogs for lead optimization or diverse but biologically similar compounds to open up novel chemical routes.

The latter scenario is sometimes referred to as scaffold or lead hopping [22–25]. This is a formidable challenge for the descriptor and the similarity measure. While avoiding the chemical graph and atom type-based molecular representation, the essential features required for activity have to be retained. By definition, such a task will be prone to picking out false positives and, therefore, requires a fast search in large and diverse databases together with a tunable level of similarity.

4.4.1 Virtual Screening

The first Feature Tree publication demonstrated the ability to retrieve known actives from databases in addition to the potential to bridge different structural classes based on Feature Tree similarity [4]. A collection of 581 randomly chosen compounds from the MDDR together with 391 active compounds covering five target classes constituted the dataset for enrichment studies. Each compound served as a query and enrichment rates were compared within the different target classes. The improved retrieval rates relative to random selection were compared with results obtained using the 2D fingerprint descriptor from Daylight. Enrichment rates for Feature Trees were either lower than with Daylight fingerprints (ACE), similar (PAF, HMG) or higher (TXA2, 5HT3). The average over-

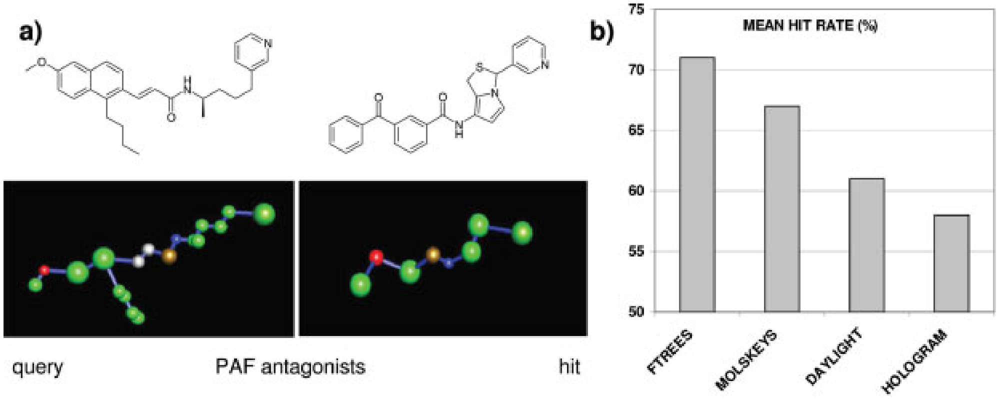


Fig. 4.6 (a) Feature Tree search result based on query (left) retrieves a high ranked (rank 5) active compound with low 2D similarity. The corresponding feature trees are also visualized. (b) Comparison of hit rates in an enrichment study (Feature Trees, MDL Keys, Daylight Fingerprints, Molecular Holograms).

lap of hits found by both techniques was about 50%. Careful examination of the results revealed the potential of Feature Trees to retrieve actives that show low 2D similarity. Figure 4.6 shows such a lead hop example for active compounds not retrieved by the 2D fingerprint-based methods.

In a publication dedicated to novel so-called affinity fingerprints, the same dataset as described above was originally used to compare a number of 2D descriptors [26]. The hit rate was defined by the number of correctly classified hits (with respect to the target class) within the 10 nearest neighbors for a given query. Overall performance was measured by the mean hit rate of all 391 active compounds taken once as a query. All descriptors showed hit rates of around 60–70% with Feature Trees slightly outperforming MACCS keys, Daylight fingerprints and molecular hash key (Fig. 4.6).

In the circumstances surrounding more true to life drug discovery projects, actives have to be retrieved from much larger virtual compound collections during virtual screening. In this case, the modeler will invariably need a strategy for selecting subsets of molecules in order to reduce the dataset to a more manageable size. This is still very much an issue when the modeler wishes to investigate molecules using slower virtual screening approaches such as docking. One such scenario involving the application of Feature Trees is presented below, where a very large dataset of molecules was to be investigated using a combination of similarity-based and structure-based screening approaches.

The target protein was CDK2 with known 3D structure. Associated with the target was a diverse set of 57 active molecules taken from the literature [27–32]. A couple of the actives can be seen in Fig. 4.7a, shown in red. The large dataset of ~1.3 million compounds was compiled from vendor catalogs (including Aldrich Rare Chemicals [33] and Chemstar, September 2002 [34]; doubly occurring entries were removed). The object of the experiment was to apply available virtual screening tools to these data in order to identify further potential CDK2 actives. Feature Trees were chosen as the similarity descriptor because of their speed and lead hopping capabilities and also because the software allows interactive parameter tuning by the user. The descriptor was applied in the first instance to identify molecules in the large dataset that showed similarity to the available actives. Although a dataset of this size can be handled quickly by Feature Trees (for example, the dynamic match-search algorithm is fast enough to search this number of molecules in about 40 min on a current single PC processor: P4, 2.4 GHz), it was too large to be processed with slower modeling tools in a reasonable amount of time. It was decided that approximately one-tenth of the large dataset (130 000 molecules) would provide a good-sized subset for the structure-based part of the experiment. Feature Trees were then applied in the second instance to select appropriate molecules to form the smaller subset. How the selection was made was of great concern and critical to the final outcome of the experiment.

In the ideal case, the smaller subset should of course contain all molecules from the large dataset most likely to be potential CDK2-active molecules. The simplest attempt to achieve this goal would be to calculate the similarity values for all dataset molecules compared with one of the active molecules using Feature Trees and select the most similar 130 000. This strategy brings with it the following problem: the subset will be very restricted in terms of diversity. The modeler often has the structure of an active molecule available but is actually seeking new actives that contain a different scaffold in order to achieve a lead hop. A superior subset should contain molecules similar to the actives, yet diverse amongst themselves. An improved attempt would take into consideration all information available from the diverse set of 57 actives. Each dataset molecule is compared with all the active molecules to give lists of the most similar molecules to each active. The 130 000 molecules most similar to any of the actives are then selected for the subset. However, we are now in danger of overfitting the data. To avoid overfitting, the available data can be split so that one part is used in a similar manner to select the subset, while the other part can be used to measure the performance of the selection method. The final procedure used in the experiment was as follows.

Ten of the known actives were picked from the 57. This can be done, for example, by picking 10 actives at random, but could also be done more strategically by taking molecules representative of clusters after the actives were clustered amongst themselves or simply by hand picking 10 of the molecules. The remaining actives were hidden in the large dataset. Each dataset molecule was compared with the 10 active molecules. To force the subset to contain molecules similar to each of the 10 actives, the most similar 1 300 molecules to each were selected first to be part of the subset (multiply occurring molecules were of course removed). The subset was then completed by taking the next most similar molecules to any of the 10 actives. The number of hidden actives present in the subset was counted. In fact, what was more important was to note the presence of different chemotypes of the hidden actives in the subset. This hopefully will reflect the diversity among the rest of the subset, which is really what is of interest to the modeler. It is easy to recognize at this stage the influence that the selection of the 10 actives will have on the contents of the subset and how the selection of the 10 actives can actually be optimized to achieve the desired effect of finding the most diverse set of hidden chemotypes at the end; this was how the experiment proceeded.

As mentioned above, the Feature Trees software allows to user to tune various parameters to influence how molecules are compared with each other. When using the dynamic match-search algorithm, it is possible to change, for example, whether the alignment of a common molecule core or the alignment of potential pharmacophoric groups at the edge of the molecules should be more important. Parallel to refining the selection of 10 actives, the parameters of the software were tuned to also achieve the desired effect of finding the most diverse set of hidden chemotypes. This was actually done following the final selection of 10 actives but, in a more thorough experiment, these two steps could be done in parallel or iterated.

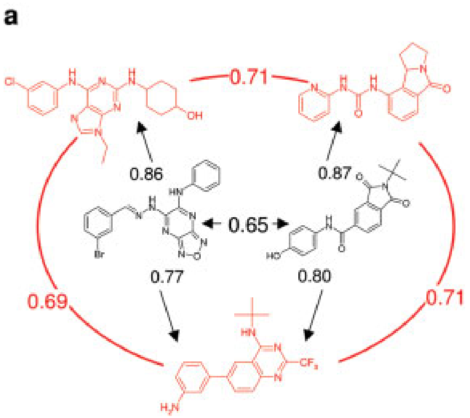
With the final selection of 10 actives and tuned parameter settings for the Feature Trees software, 40 of the remaining 47 actives were present in the final subset of 130 000 molecules. The diversity of the subset is represented in Fig. 4.7 a. Here, two dataset molecules found to be simultaneously in the 1300 most similar molecules to two actives can be seen, along with the corresponding actives themselves – the Feature Tree similarity between the molecules is also shown. A small table in Fig. 4.7 b gives some brief statistics of the dataset or subset size and the number of hidden actives found at two stages of the subset creation.

This subset was further investigated using the slower modeling methods to try to identify potential actives, known as plausible hits. An example of a molecule selected from the results of a docking experiment is shown in Fig. 4.8. This molecule had a similarity score of 0.93 to an active and is shown docked with the typical kinase inhibitor binding pattern. Both the active and the plausible hit are not drug-like from a medicinal chemistry perspective, but this example demonstrates well how the Feature Tree descriptor captures similarity between two molecules.

4.4.2 Virtual Screening Based on Multiple Query Compounds

For fuzzy virtual screening purposes, highly active molecules with different scaffolds are combined into an MTree model. By combining the information of remotely related actives into a single model, efficient database searches with molecule ensembles are possible.

Here, we provide an example of a previously published study [35] of an application of our analysis strategy to an HTS data set for DHFR of *E. coli* [36]. The goal was to find a good predictor for a correct classification of active and inactive molecules using MTree models. The data set consisted of 100 000 compounds split in two equal-sized parts: a training set and a test set both with bio-



b

stage of subset calculation	set size	no. actives	enrichment factor
start: complete dataset	1 289 408	47	-
extract 1 300 nearest neighbors to each of the ten actives and remove doubles	5902	27	125
finish: final subset	130 000	40	8

Fig. 4.7 (a) Three of the 10 actives (colored red) used to select the final subset are shown with two "unknowns" (colored black) found amongst their nearest neighbors. The Feature Tree similarity values (calculated with the dynamic match-search algorithm) are also shown. The actives and unknown molecules are diverse with respect to molecular scaffold. (b) Statistics of the dataset or subset size and the number of hidden actives found from the 47 at two stages during the creation of the 130000 molecule subset.

logical activities. For a comparison of the MTree model approach with the standard Feature Trees similarity search, we decided to use a simple kernel classification approach. Here, for an arbitrary given molecule t and for a given radius δ, let $\rho_\delta(t)$ be the *activity density* of the arbitrary molecule t:

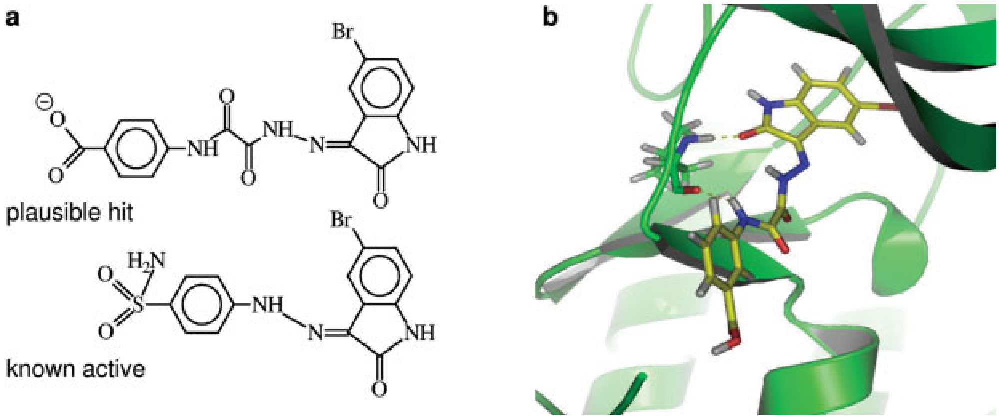


Fig. 4.8 (a) A plausible hit molecule (top) selected after completion of docking experiments on the small subset of molecules. A similar known active is shown beneath. (b) The docking pose of the plausible hit in the active site of CDK2 as predicted with the docking software FlexX – the typical binding pattern of kinase inhibitors around the hinge region of the active site is clear to see. Picture created using PyMOL: DeLano, W.L. The PyMOL molecular graphics system (2002), http://www.pymol.org.

$$\rho_\delta(t) = \frac{|U_\delta(t) \cap Act|}{|U_\delta(t)|} \tag{3}$$

Act being the set of all active molecules and $U_\delta(t)$ being the set of all molecules within the Feature Trees similarity radius delta around molecule *t*. A molecule is classified as active if the number of actives in its neighborhood is greater than the threshold ρ. The cross-validation results for the classifier on the training set supported the choice of $\delta = 0.9$ and $\rho = 0.22$. The train and test set were more diverse than expected, so we went down to $\delta = 0.8$. We reported a ranked list of the top λ molecules with the highest activity density. For the MTree model approach, all actives from the training set were clustered and the most diverse cluster centers were selected. From this selection, several MTree models were constructed and used for virtual screening of the test set. With the help of the MTree models and the classifier, we independently selected 5000 molecules for further analysis. The best MTree model retrieved more actives (20 of 119) from the test set than the classifier (15) in the selection of the best scored 5000 molecules.

4.4.3 Tagged Feature Trees

Virtual screening often benefits from an expert bias which helps focus on more desirable results, given in the form of additional information. A point in case is docking under pharmacophore constraints [37] or the concept of relative pharmacophores with a “special” internal reference point [38]. Yet another application for a directionally biased compound comparison is the selection of chemical reagents where functional attachment points are aligned and pharmacophoric features are examined relative to this point of reference by a procedure termed GaP [39]. Other concepts which try to describe combinatorial products in terms of their educts and need a special reference point are shape-based Topomers [24] and pharmacophore based OsPreys [40].

The Feature Tree comparison algorithm defines directionality by comparing two trees and selecting one alignment of the nodes from among many possibilities. It always finds the alignment with the best possible overall score. However, there might be situations where a user-defined node match is desired in order to incorporate expert knowledge in the search. With this in mind, an extension of the Feature Trees match-search algorithm has been implemented where common user-defined substructures are forced to be aligned with each other. In fact, the defined substructures are converted into special Feature Tree tag nodes, which then always match each other in the Feature Trees matching. We briefly describe three useful scenarios.

For example, ACE inhibitors display a distinct set of pharmacophore points (an acidic group, a carbonyl group and a zinc binding group [41, 42]; see also Fig. 4.5). Recently, the structure of the 3D target along with the co-crystallized ligand lisinopril (LPR) was published [43] revealing the binding mode as shown in Fig. 4.9a. As there are acidic groups at both ends of the molecule, aligning the distinct zinc bind-

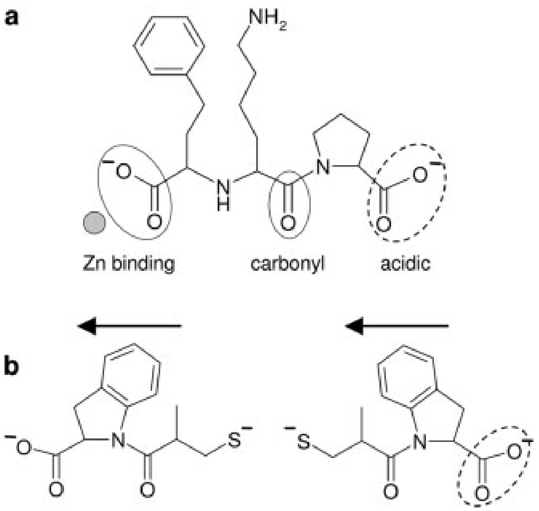


Fig. 4.9 Two ACE inhibitors. (a) Lisinopril (LPR – the co-crystallized ligand). LPR exhibits the three pharmacophore points common to ACE inhibitors: a zinc binding (carboxylate), carbonyl and acidic group. (b) A second ACE inhibitor where the zinc binding group is a thiolate group. The standard Feature Tree descriptor aligns the inhibitor to LPR back-to-front with a similarity score of 0.77, as shown on the left (the direction of the arrow indicates the direction of the alignment to LPR above). When the acidic group substructure is identified as a tagged node (dashed oval), the alignment of the two inhibitors is corrected although the similarity score falls to 0.44, as shown on the right.

ing and acidic group pharmacophore points in ACE inhibitors correctly is difficult because they obviously display very similar characteristics. This is true of LPR itself. Another ACE inhibitor is shown in Fig. 4.9 b, where thiolate replaces the carboxylate group seen in LPR at the zinc binding position. The Feature Tree descriptor chooses to align the ACE inhibitor to LPR in a back-to-front manner, matching the thiolate zinc binding group to the carboxylate group in LPR, as in Fig. 4.9. By defining a set of acidic substructures (for example, carboxylate and tetrazole) to be tagged Feature Tree nodes, the substructures can be forced to match with each other. Then, from this forced starting point (or forced initial split), the Feature Tree comparison algorithm must only find the alignment between the remaining parts of the two molecules. It is important to note that forcing a starting point in the matching nearly always results in a lower similarity score for the alignment.

In a different application scenario, tagged nodes can be used to define a known common anchor point in a set of molecules. A set of approximately 400 R-group instances from a combinatorial library (based around a quinazoline core) forms an example of such a set of molecules, where the functional attachment point forms the tagged Feature Tree node. A Feature Tree similarity

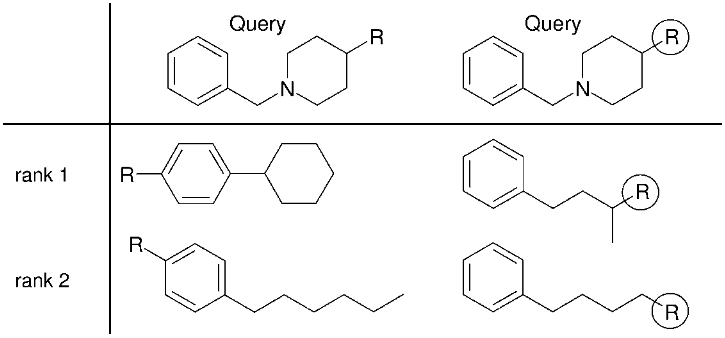


Fig. 4.10 R-group molecules or instances from a combinatorial library: the common anchor point of the instances is marked with an R. One instance was used as a query in a similarity search (top molecule). On the left, the two most similar instances in the library without consideration of the anchor point are shown. When the anchor point is identified as a tagged node (circles), the most similar instances in the library have their anchor point at equivalent positions in the molecules, as shown on the right.

search was carried out in this set using a query R-group instance, both with and without the tagged node approach, and the instances were ranked according to similarity. The query and top results can be seen in Fig. 4.10. Without a forced match between the common anchor points, the Feature Tree descriptor identifies R-group instances as most similar which align with the query back-to-front. This is because the alignment algorithm has no information about the anchor points and searches for the globally best scoring alignment. Once the tagged nodes are forced to match, the Feature Tree descriptor finds instances to be most similar under the condition that the direction of the alignment is fixed.

Screening on chemical microarrays [18, 44] – where chemical compounds are immobilized on a solid support – represents a particularly striking example of the relative pharmacophoric features. Of course, the similarity of compounds in this case depends strongly on the site of attachment. Tagged Feature Trees have been used to describe compound similarities as a function of the mode of immobilization. This approach provided a useful guidance for library design and screening data analysis in the context of affinity-based screening on chemical microarrays [45].

4.5
Searching Combinatorial Fragment Spaces with Feature Trees

So far, we have used the Feature Tree descriptor in pairwise comparisons. Most of the practical applications of molecular similarity can be reduced to the problem of comparing a pair of molecules. In some applications, however, it is extremely inefficient to do so. For combinatorial libraries, for example, we can search for the most similar compound by enumerating the library and comparing every library molecule with the query molecule. The number of library molecules is polynomial in the number of building blocks used. It would, therefore, be much more efficient to search the library in its closed form, namely on the basis of the building blocks and the synthesis rules. In the following, we will describe how such similarity searches can be performed with the Feature Tree descriptor.

The search space that we consider is a so-called *combinatorial fragment space*. It consists of a set of molecular fragments and rules defining how these fragments can be combined to build molecules. Each fragment contains one or multiple link atoms of a certain link type. The rules describe which link types are compatible with each other and which chemical modifications have to be performed if a pair of fragments are connected via a certain link-type combination. Combinatorial fragment spaces are a very powerful description. They can handle molecule libraries (fragments without links) and combinatorial libraries (every R-group has its own link type) as special cases. In its most general fashion, a combinatorial fragment space describes what can be made synthetically from a set of educts (fragments) with a fixed set of reactions (rules). A popular way of creating fragment spaces is the retrosynthetic analysis of a compound set [46]. Here, the rules are used to break compounds into fragments at chemically meaningful positions and to add the corresponding linker atoms.

4.5.1 Search Algorithm

Given a query molecule, the question arises of how to find molecules from a combinatorial fragment space most similar to the query. Obviously, in most cases an enumeration of the space is prohibitive owing to the enormous number of compounds which can theoretically be constructed. Therefore, we have to search on the level of molecular fragments and rules instead. The initial step to do so is to convert the combinatorial fragment space into the Feature Tree domain. Every fragment can be converted into a Feature Tree if we make sure that link atoms become separate nodes; so-called *link-nodes*. We can create the Feature Tree of a molecule of the space without going back to the molecular level: The link-nodes at the connected fragments are removed and a new edge is formed between the nodes adjacent to the link-nodes (see Fig. 4.11 for an example). After converting the query molecule into a Feature Tree, the remaining task is to construct a Feature Tree from the Feature Tree fragments of the space with highest possible similarity to the query.

In order to solve this task, we will again develop a dynamic programming scheme as in the case of pairwise Feature Tree comparisons. For every directed cut in the query tree and every link type in the fragment space, we would like to know the most similar fragment of the space under the assumption that the link of the fragment is matched to the atom behind the directed cut (see Fig. 4.12). For all pairs of directed cuts and link types, a list of the most similar

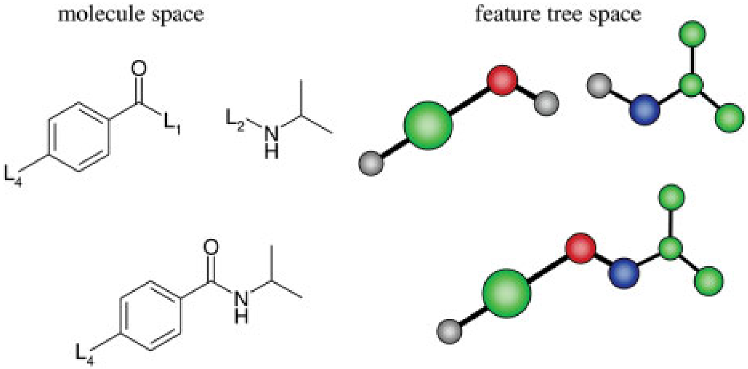


Fig. 4.11 Chemical fragment spaces consist of molecule fragments with linkers and a set of rules defining how fragments can be connected. With Feature Trees, the process of building molecules from a chemistry space can be done directly and easily on the descriptor level by combining the trees of the corresponding fragments. In this way, the challenge of searching a chemistry space without enumerating the compounds can be addressed.

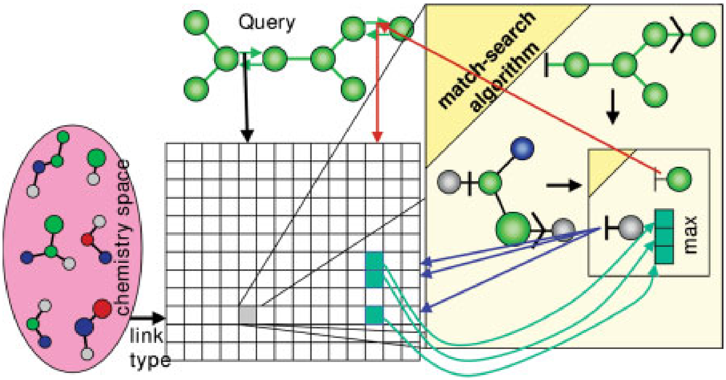


Fig. 4.12 A chemistry space can be searched for the compound most similar to a query with the following dynamic programming procedure. For every directed edge and every link type, a list of the most similar fragments is calculated with the match-search algorithm. When a second link is found in the fragment, the dynamic programming matrix can be used in order to find the highest possible similarity value (red arrow, the part of the query which has to be matched to the link; blue arrows, the compatible link types; green arrows, previously calculated similarity values for these edge–link-type combinations).

fragments is stored in a matrix called the *edge-link-similarity matrix*. If we assume for a moment that the fragment space contains only fragments with a single link, then this matrix can be computed by applying the match-search algorithm (without the initial split search) to every pair made up of a directed cut in the query and a fragment from the space. The final, most similar molecule can be found by taking every edge of the query tree together with every allowed combination of link types and combining the similarity results achieved for the antiparallel directed cuts at this edge.

In the generic case, a fragment might have multiple links. We can still compare a part of the query tree with a fragment of the fragment space using the match-search algorithm. Now, while the match-search algorithm proceeds, the following situation might occur. After selecting a match, the algorithm will cut at a link-node. Subsequently, we have to compare the link-node of type t with a certain part of the query tree starting with edge e. We cannot perform this comparison because the link-node is only a placeholder for a fragment which we can add via a compatible link-node. At this point, we can make use of the edge-link-similarity matrix. Let us assume that t' is a link type compatible to t. In the matrix under position (t', e), we can find the fragment of the space most similar to the query tree part starting at bond e and also the corresponding similarity value for that fragment. The match-search algorithm can, therefore, stop here and just reuse the results of a previous run. The algorithm can easily be extended to deal with multiple compatible link types.

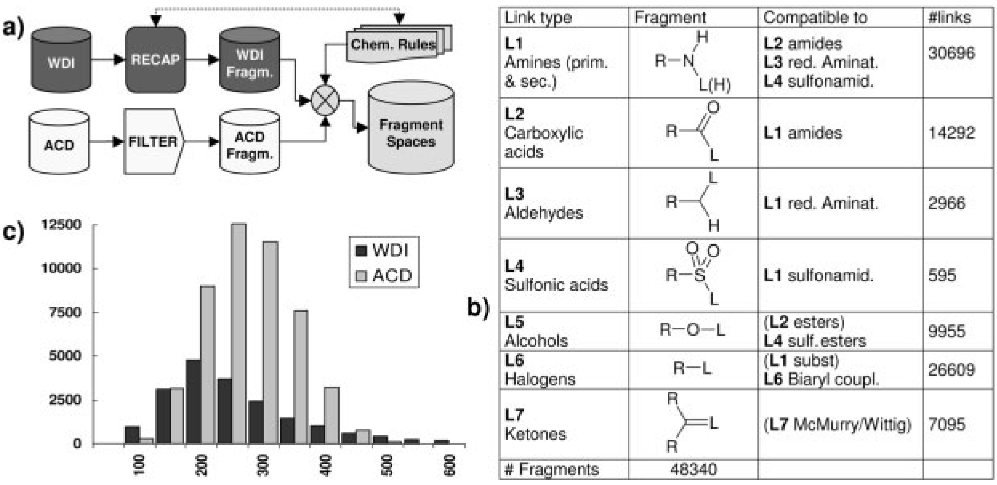
Two further phases are necessary to complete the combinatorial fragment space search algorithm. In a preprocessing phase, the order in which the edge-link-similarity matrix is computed has to be determined. The order of the link types (the rows of the matrix) does not play a role. The order of the edges, however, is of importance. Since we are reusing data from the matrix, it must be ensured that the results for small subtrees are computed first. As in the dynamic match-search algorithm, the computation is started at the terminal nodes and ordered such that the subtree size increases. In a post-processing phase, the most similar Feature Trees have to be constructed since the result of the above-described dynamic programming scheme is only the similarity value. At each entry of the edge-link-similarity matrix, a list of fragments achieving the highest similarity values is stored. Based on these lists, a recursive algorithm not described here can reconstruct all Feature Trees (and the corresponding compounds) which result in the previously calculated similarity values.

4.5.2 Set-up of Fragment Spaces

A popular way of creating fragment spaces is the retrosynthetic analysis of a compound set described in the context of Feature Trees in [9]. Here, the rules are used to break compounds into fragments at chemically meaningful positions and add the corresponding linker atoms. Fragmentation can be performed within relevant databases of drug-like chemical compounds such as the WDI, CMC or MDDR [47]. An advantage of such starting material is the focus on drug-like chemical space with given precedence for a real synthesis. An example of such a space is the WDI-derived chemical space as described in the original Feature Tree fragment space paper. RECAP [46] – like shredding of the WDI based on 11 reactions – led to 17 000 unique fragments with an average molecular weight of 200 Da. The chemically allowed rules for combinations of fragments can be based on the same (retro-)synthetic rules which were used to split the compounds in the database. However, a bias to existing compounds may exist and hinder novelty. In addition, fragments defining the virtual hits have to be "translated" to commercial reagents in order to start synthesis.

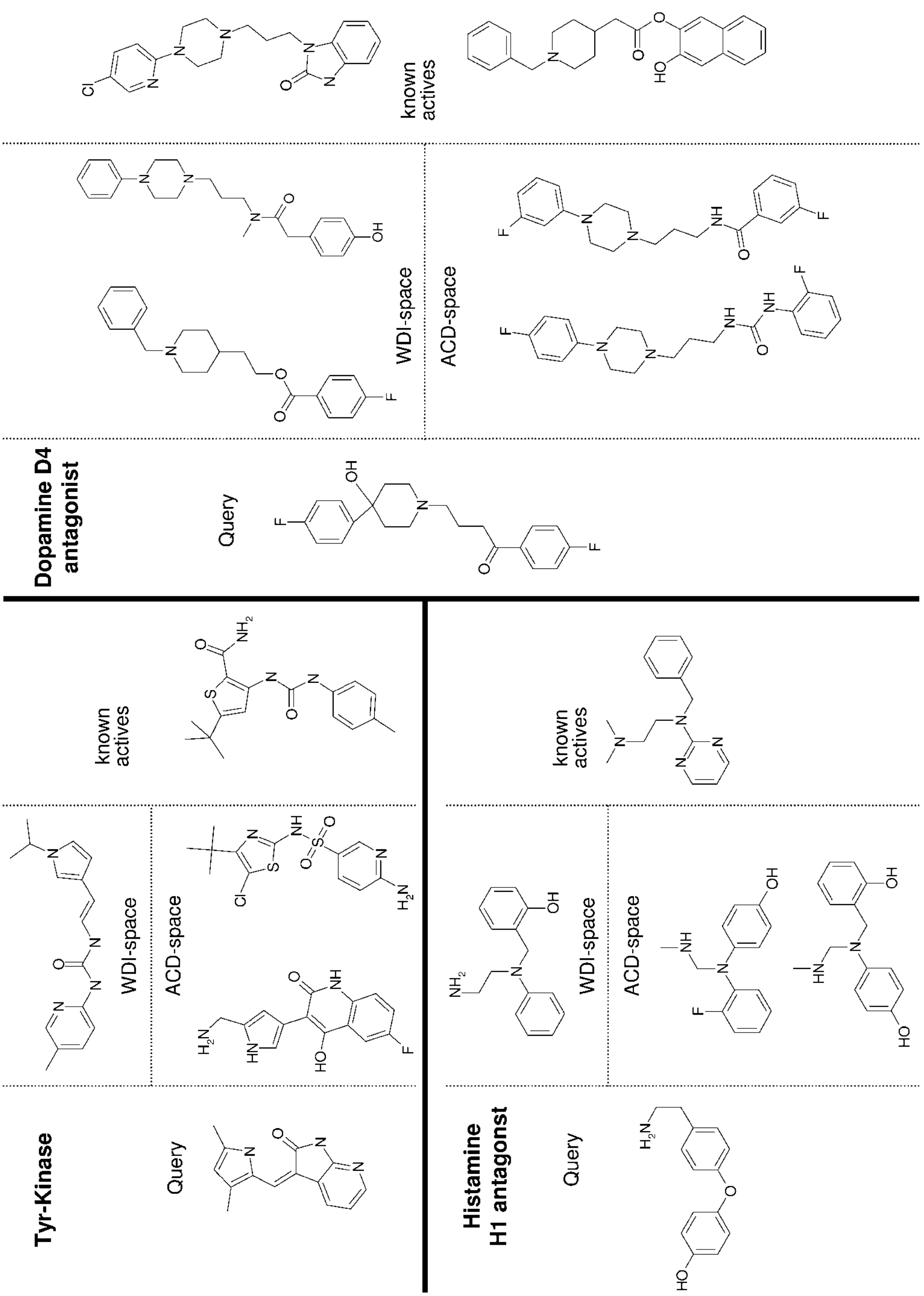
Alternatively, available chemical reagents from vendor databases together with established chemistries can guide the setup of a supply based fragment space (Fig. 4.13 a). In this way, the search space is defined by all compounds that could be made with all reagents and established chemistries "at hand". By defining the likelihood of accessibility of reagents, e.g. in stock or obtainable from a reliable supplier, versus chemicals that would still need to be ordered, hits can be classified by availability. Likewise, synthetic ease can be captured by the estimated do-ability of chemical synthesis, which can further help classify real accessibility of virtual hits for biological screening experiments. This can be a crucial aspect in discovery projects where timely supply of interesting compounds is a key factor.

Here we describe in practical terms the necessary steps to set up such a tailored chemical fragment space. The chemistries for fragment assembly are



Link type	Fragment	Compatible to	#links
L1 Amines (prim. & sec.)	H R—N L(H)	**L2** amides **L3** red. Aminat. **L4** sulfonamid.	30696
L2 Carboxylic acids	O R— L	**L1** amides	14292
L3 Aldehydes	L R— H	**L1** red. Aminat.	2966
L4 Sulfonic acids	O O R—S L	**L1** sulfonamid.	595
L5 Alcohols	R—O—L	(**L2** esters) **L4** sulf. esters	9955
L6 Halogens	R—L	(**L1** subst) **L6** Biaryl coupl.	26609
L7 Ketones	R =L R	(**L7** McMurry/Wittig)	7095
# Fragments	48340		

Fig. 4.13 (a) Workflow of a Feature Tree fragment space set-up. The underlying fragment database can be derived from retrosynthetic analysis (RECAP) of compound databases (WDI) and/or from a selection of reagents from vendor databases (ACD). In addition, chemical rules for allowed fragment combinations need to be defined. (b) The number and molecular weight distribution of unique fragments present after the retrosynthetic analysis of the WDI and the reagents collected from vendor databases. (c) Chemistry definitions and link types used in the ACD fragment space.


Dopamine D4 antagonist
Query
WDI-space
ACD-space
known actives
Tyr-Kinase
Query
WDI-space
ACD-space
known actives
Histamine H1 antagonst
Query
WDI-space
ACD-space
known actives

based on a list of common chemical reactions used in combinatorial chemistry. The definition of link-types differs somewhat from that of the initial RECAP-WDI space. We do not distinguish between amines and ureas, in fact the former can be used to build up the latter simply by including carbonic acid into the fragment space. Currently, seven functional groups and also seven reaction types are covered (Fig. 4.13 b). Depending on the query or desired hit properties, reactions such as ester formation are removed.

The source of reagents was the Available Chemicals Directory [47]. The commercially available building blocks were filtered for desired properties and undesired chemical features in order to direct the chemical space towards drug-like or lead-like molecules. A list of defined protecting groups was removed and at least one functional group related to the chemistries defined for linking fragments had to be present in a database compound. It is important to include reagents with multiple link types to allow diverse combinations of fragments beyond simple two-component products. Duplicate fragments are removed during the labeling process, e.g. identical fragments descendent from the carboxylic acid and the acid chloride thereof. In addition, one has to consider the elimination of redundant fragments that cannot be distinguished by the descriptor, such as substitution patterns of terminal aromatic ring systems. It can be advantageous to enrich the collection selectively with small linker-type fragments and to include small reagents, such as water, ammonia, hydrazine or similar, into the space, which can take on a kind of "bridging" function between larger fragments. Spiking with target-based scaffolds not necessarily covered by commercial compounds provides a means additionally to tailor the search space towards more focused libraries. As the definition of chemistries currently does not handle ring closures, a number of carefully selected heterocycles was also added to the collection. Overall, the fragment space contains approximately 50 000 fragments with a molecular weight distribution higher than the WDI-derived fragment space (Fig. 4.13 c).

4.5.3 Searching in Fragment Spaces

The Feature Tree fragment spaces screening concept was exemplified for a number of known actives across different target classes in the paper introducing the general concept [9]. Starting from a single active compound as the query structure, the ability of the search method to construct and identify so-called

Fig. 4.14 Search results for three different targets. Whereas hits close to the query could be retrieved with a high similarity value (not shown), at a lower similarity threshold so-called plausible hits were retrieved which are close to actives for the same target but chemically distinct from the query. For each target the query is shown (left) together with plausible hits in the WDI (middle top) and ACD fragment space (middle bottom) and known actives (right).

plausible hits was investigated. The plausibility of hits was judged based on the structural similarity to either the query or other known actives. In many cases, the Feature Tree descriptor was able to produce hits that were distinct from the query but structurally related to other actives from a different chemical series, hence showing its ability to jump between classes. The targets under investigation were dopamine D4 antagonists, histamine H1 antagonists, Cox-2 inhibitors, Tyr-kinase inhibitors and angiotensin II antagonists. We have taken the same examples to investigate the ability to retrieve plausible hits from the fragment space based on commercial reagents as described above. Figure 4.14 shows selected hits from searches within both fragment spaces.

As long as we perform similarity searches in small libraries of up to a few million compounds, searching for a list of the most similar compounds is a useful task. When the search space increases, however, two problems occur which make the redefinition of the similarity search problem necessary. First, in a very large fragment space, the chance is high that the most similar compound is nearly identical or even identical with the query compound. In this case, the similarity search would be useless. It becomes clear that our true aim is not to find the most similar compound but to find molecules with a certain similarity value. Only the limited search space and the imperfect search algorithms make these two problems appear the same in daily practice. To tackle this problem within the Feature Trees software, a target similarity value can be defined such that molecules with a selected degree of similarity are created. The influence of such similarity fine tuning was investigated with a set of 55 dopamine D4 antagonists [9] demonstrating the gradual morphing of compounds to different topologies as the similarity level decreases. These changes often go hand in hand with increased molecular complexity of hits and the user needs to judge carefully the appropriate parameter settings.

The second problem appears when we browse through hit lists resulting from similarity searches in large spaces. Owing to the size of the space, the compounds in the hit list might be very similar to each other. It is therefore necessary to control the diversity within the hit list. Within Feature Trees, we have several criteria at hand; for example, the number of common or different fragments between two molecules in the hit list can be limited. Alternatively, Feature Tree similarity within the hit list can be used to increase diversity.

While the number and nature of compounds can be controlled as described above, Feature Tree fragment spaces still makes it possible to generate an almost unlimited number of similar compounds. In order to benefit from the vastness of the search space, the user is generally interested in the creation of as many suggestions as possible without being overwhelmed by the sheer number of compounds. We found that even with carefully chosen parameter settings, the process of selecting molecules for synthesis and testing from a large hit list can still be a challenge. One approach to guide this process involves post-processing by clustering the hits. It is a distinct and desired characteristic of Feature Tree fragment spaces that compounds with varying topologies are generated. Therefore, the concept of molecular frameworks [48] can be very use-

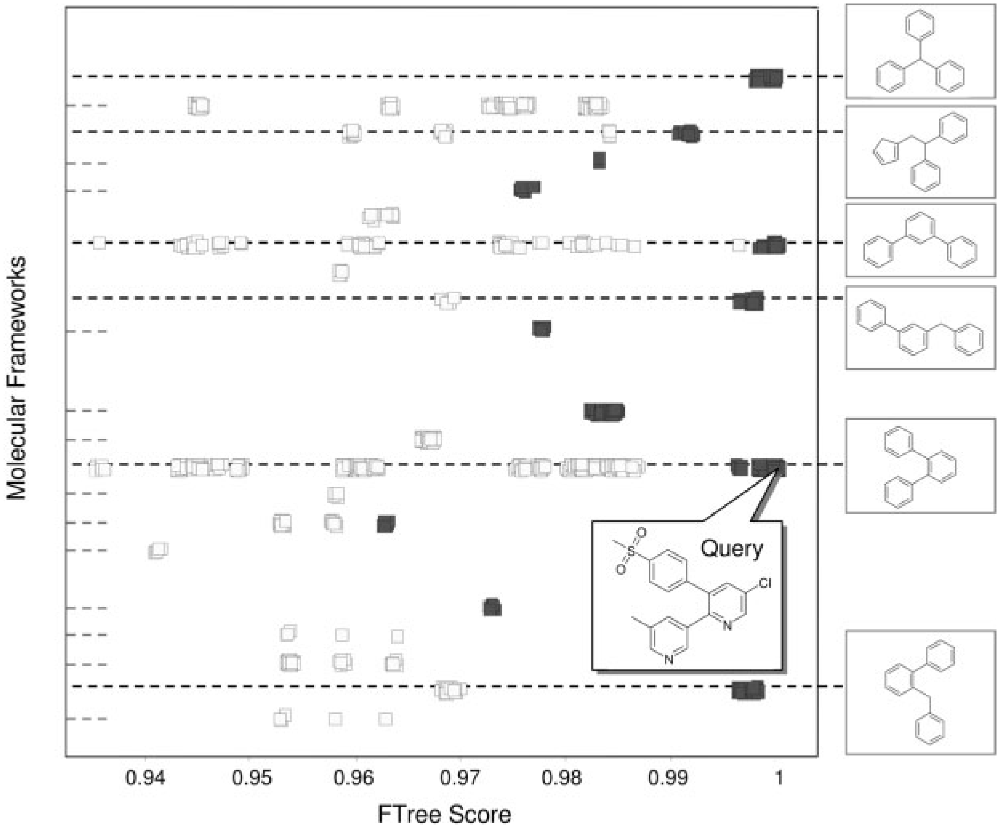


Fig. 4.15 Hit list from Feature Tree fragment space searches are grouped by a molecular framework analysis (*y*-axis) and the Feature Tree similarity score (*x*-axis). The search algorithm indeed produces compounds from variant chemical classes. Hits for further processing are then selected not only based on the overall top ranked molecules, but the top-ranked molecules within distinct frameworks (black squares).

ful to group hits that share a common chemical graph-based topology. Figure 4.15 shows an example with a Cox-2 inhibitor query. The hits belonging to different frameworks are grouped along horizontal lines with the *x*-axis indicating the similarity value. The figure shows that Feature Tree fragment spaces indeed allow exploration of divergent topologies spanning variant ranges of similarity. Within a molecular framework, assorted decorations or atom-type scaffold variations can be inspected. It has been observed that the algorithm tends to produce more complex compounds with decreasing similarity to the query [9]. Therefore, a representation as in Fig. 4.15 helps concentrate on simpler compounds. Hit compounds for further investigations can be selected not only by reflecting similarity, but also by ensuring that diverse chemical classes are covered by prototype representatives.

4.6 Multiple Feature Tree Models: Applications in HTS Data Analysis

Owing to the rapid progress in the fields of combinatorial chemistry and parallel synthesis, large sets of diverse structures are available for high-throughput screening (HTS). The computational analysis of HTS results becomes an important task in computer-aided molecular design because of the significant noise and high failure rates. The identified screening hits are compared with each other in order to generate a hypothesis about the underlying lead structure. Similarity-based methods are also used for identifying structural classes around the detected screening hits in combination with fast clustering or partitioning algorithms. After such grouping of similar hits, SAR (structure–activity relationships) models can be generated which relate the biological activity to the presence or absence of substructures or functional groups. These models can be used to prioritize molecules for further testing.

In the following section, the software tool *HTSview* is presented. It has been designed for the extraction of knowledge from HTS data by means of multiple Feature Tree alignments [15]. The targeted application is the analysis of the primary screen and on selection of molecules from the test set for the secondary screen in order to identify suitable lead structures. An interactive graphical user interface combines the concept of molecular similarity with data mining and visualization methods and aims at the identification of appropriate chemical series for optimization. All the methods rely only on similarity comparisons and measured activity data alone and work without any information on the target structure. The main idea behind HTSview is a design cycle which combines information gained from experimental and computer-based methods (see Fig. 4.16). In each iteration the results of the experiments are analyzed and then new molecules which should be tested are proposed. We normally start the process with several hundred thousand measured activities of a compound library usually containing yields from several hundred up to a thousand hits. In order to reduce the data and to extract only relevant information, the so-called activity region is computed. This includes all highly active and similar yet inactive compounds. Note that HTS data are in general of poor accuracy, containing significant systematic and statistical errors. Therefore, we have to deal with many false positives and false negatives which are likely to be found in the activity regions.

First the pairwise similarity between all considered active molecules in the HTS collection and all inactive ones are computed. The Feature Tree dynamic match-search algorithm is chosen for similarity calculations. The most interesting inactives are selected using a similarity cut-off. This usually reduces the data set to 2000–5000 compounds. For these molecules, an all-by-all similarity matrix is computed. Standard data mining methods can be applied to this matrix in order to extract relationships between the structural similarity of the molecules and their measured biological activities. In HTSview, several clustering algorithms have been integrated which allow grouping of similar molecules. Similarity cut-offs or kNN-methods such as Jarvis–Patrick clustering [49] are

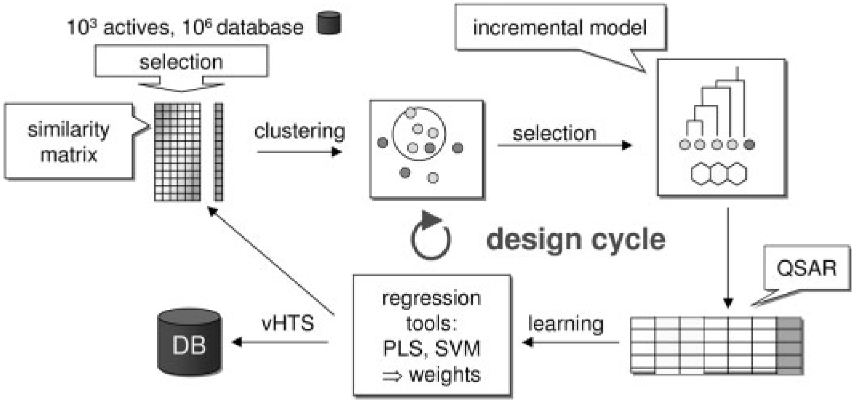


Fig. 4.16 Main workflow used by HTSview. First a similarity matrix is computed based on Feature Tree similarity, Then the Feature Trees are clustered. For selected clusters MTree models are constructed. A QSAR matrix is computed based on the MTree as an alignment template. In the next step, the activity-related weights are extracted by standard regression tools. The resulting biophores can be used for virtual screening or to redefine the similarity measure.

very efficient and can deal with huge matrices. Hierarchical clustering algorithms (e.g. single linkage or complete linkage clustering [50]) are better suited to smaller data sets and can be applied to selections resulting from the kNN methods. For each cluster showing a good activity profile, a multiple Feature Tree alignment is computed using the MTree algorithm (cf. Section 4.3.4).

The next step is to correlate biological activity with the structural information of the topological alignment in order to find the relevant motifs from an HTS experiment. We call this type of model a *biophore* model. In this context, we define a *biophore* as the fragment-based description of an ensemble of molecules with similar biological activities and structures, based on the Feature Tree descriptor. Biophore models are generated in a two-step procedure. The first step is called *fragment sampling*. A set of similar actives and inactives is aligned with the MTree of the cluster under consideration. The fragments in each match are collected in a *match list*. In the second step, the MTree is used to align all the molecules of the activity region. The similarity between each molecule from the activity region and each fragment in the match list is computed and stored in a compound-property matrix as used in classical QSAR analysis.

Using linear regression or approaches such as PLS, a weight vector can be computed based on the compound-property matrix. Owing to the linear nature of this analysis, those weights provide information about the importance of each fragment at a certain position in the alignment (i.e. a node in the MTree). Positive or negative weights in the MTree nodes now indicate fragments that are positively or negatively correlated with activity, respectively. Favorable fragments have high positive weights. The resulting model can be applied to search databases efficiently for molecules with related features using the algorithm from Section 4.3.4.

Table 4.1 Biophore models for series from different target families and therapeutic indications. For GABA A and MMP-8 some compounds were excluded since Feature Trees do not distinguish stereoisomers or aromatic substitution patterns. Here, the position of substituents had a dramatic effect on the biological activity

Family	Target	No. of compounds	Biophore model				CoMFA model	
			Sampling inactives		Sampling actives		q^2	r^2
			q^2	r^2	q^2	r^2		
Serine proteins	Factor Xa	138	0.441	0.609	0.393	0.499	0.760	0.913
Metalloproteins	ACE	68	0.609	0.778	0.699	0.831	0.630	0.882
	ACE (neutral)	68	0.644	0.799	0.647	0.795	0.630	0.882
	MMP-8*	81	0.509	0.739	0.354	0.535	0.569	0.905
	Thermolysin	61	0.321	0.664	0.312	0.624	0.636	0.941
Kinases	CDK-2	86	0.463	0.636	0.499	0.614	0.630	0.860
Ion channel	GABA A*	28	0.480	0.780	0.460	0.800	0.745	0.946

In order to validate the alignment and model building procedures, we selected internal and literature data sets representing several protein target families [51]. For most datasets, 3D-QSAR models and X-ray structures were available which allowed a detailed comparison with biophore models. The data sets encompassed inhibitors for serine proteases (factor Xa), metalloproteinases (ACE, MMP-8, thermolysin), kinase (CDK-2) and ion channels (GABA-A). For every data set, significant biophore models were obtained with reasonable or good predictivity, expressed as leave-one-out cross-validated r^2 (q^2) values (see Table 4.1). The predictivity is almost as good as for 3D-QSAR techniques, although 3D information is not used, suggesting that model building does not necessarily require 3D structural alignments.

As an application to real HTS data, we applied this method to a proprietary kinase inhibitor screen. A significant biophore model with a cross-validated r^2 value of 0.465 was obtained, which explains the important SAR features and provides relevant information for follow-up investigation. This particular model was generated using 57 hits from a purine scaffold, split into actives and inactives with activities in the nanomolar to micromolar range. The biophore model was generated using the actives only. All compounds, including the inactive compounds, were aligned on the biophore model. A linear regression analysis revealed a good correlation between the highly weighted fragments and activity. Subsequent virtual screening of a large compound collection using this biophore model leads to a significant enrichment of active compounds. Hence we were able to explain the SAR of various classes sharing comparable features important for activity for this target.

4.7 Drawing Similar Compounds in 2D Using Feature Tree Mappings

One of the most interesting properties of the Feature Tree matching algorithms is that they preserve the topology of the molecules, i.e. the relative arrangement of fragments within the molecule is considered by the matching algorithms. For drawing similar molecules in a similar way, the Feature Tree comparisons are useful as they detect similarity beyond mere substructure identity and they also provide a mapping that can be viewed as alignments of molecule fragment pairs.

A first algorithm for similar molecule drawing is given in [52]. Here, embeddings into a supertree are used to guide molecule drawings. In the following, we summarize an alternative algorithm based on Feature Trees that draws structure diagrams of similar molecules in a similar manner [53].

The basic algorithm for automated drawing of structure diagrams (SDG) goes back to Zimmerman [54] (see [55] for a review on SDG algorithms). The idea is to start the drawing with one atom and subsequently add atoms and bonds to the drawing such that the drawing is always connected and free of overlaps. A variant is to split molecules into parts called drawing units and then to proceed analogously. Such a procedure has several degrees of freedom for placing the atoms (or units) on the drawing plane. The main challenges are to exploit these degrees of freedom in order to achieve non-overlapping diagrams and the drawing of complex ring systems. Our algorithm tries to place atoms according to the current standards in chemistry (see, e.g., [56]), for example by drawing bonds according to specific angle patterns. In the following, the set of rules used is referred to as drawing rules.

Given two similar compounds, the matching created during the Feature Tree comparison can be used for laying out the two SDGs similarly. In order to derive directional constraints from a Feature Tree matching, bond paths both between matches and within matched subtrees and the relative arrangement of rings in ring systems are considered.

The algorithm used to employ this information is based on the concept of "drawing under user-defined constraints": each bond of the molecule can be annotated with a direction (north, north-east, east, etc.) describing the orientation of the drawing unit containing the bond. The algorithm (for details, see [53]) searches for a structure diagram fulfilling these directional constraints. The overall method for drawing a set of similar molecules is as follows. First, one of the molecules – the "template" – is drawn using the unconstrained drawing algorithm. If the molecule set does not imply a template molecule, it is recommended to select the molecule with the largest number of rotatable bonds. The other molecules are drawn subsequentially. Bond paths are extracted from the Feature Tree matching between the template and the currently drawn molecule. The directions for the extracted bond paths in the current molecule are taken from the template and used as directional constraints for the drawing algorithm.

In Fig. 4.17, a set of benzodiazepines is shown that were aligned using Feature Tree matchings. The bond orientations extracted from the template are shown by arrows. The algorithm has been shown to be very useful to visualize similarities beyond common subgraphs. Further test cases can be found in [53].

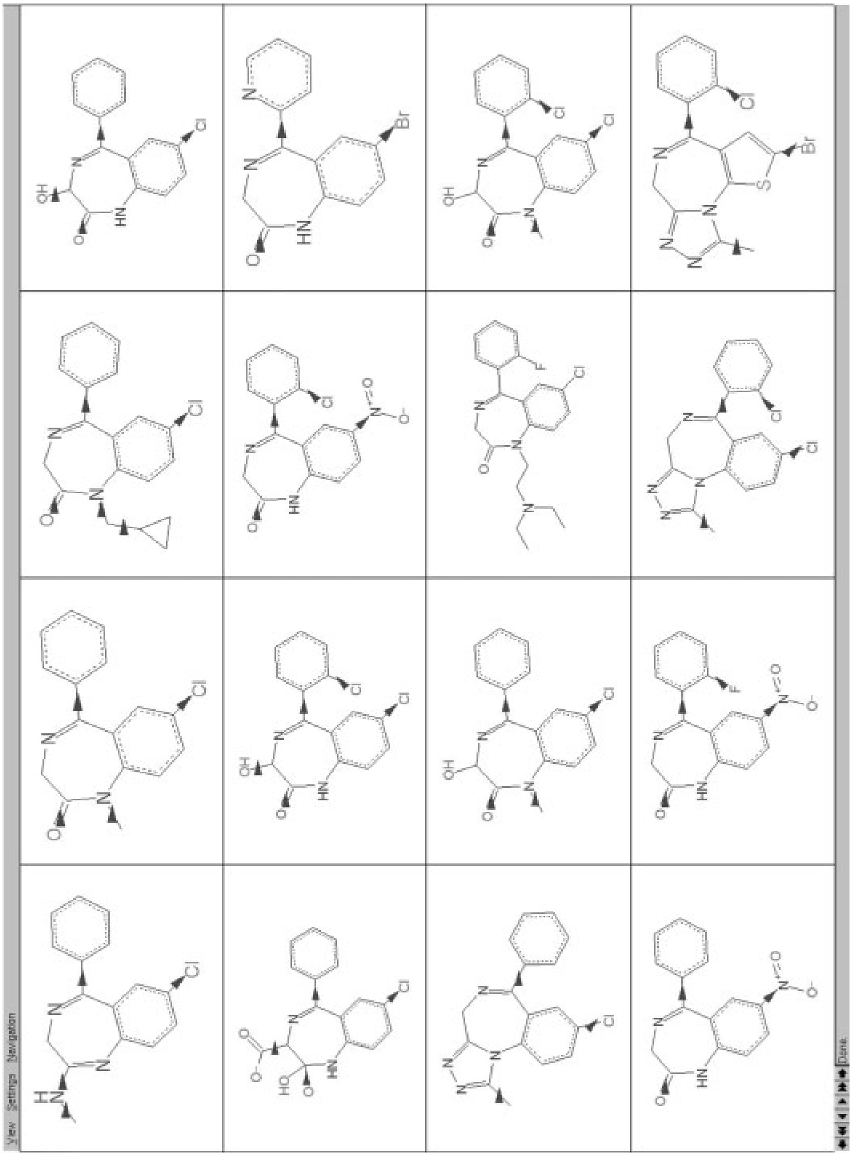


Fig. 4.17 Benzodiazepines aligned using Feature Tree matchings. The bond orientations extracted from the template compound (row 3, column 3) are shown by arrows.

4.8 Conclusion

Owing to the relevance of similarity-based techniques in molecular design, descriptor technologies have been under investigation for decades, resulting in hundreds of ways to represent molecular structure. Obviously, the question of which is the right descriptor cannot be answered in general, but generally depends on the specific application. The Feature Tree approach differs from most other descriptors in structure. The node-labeled tree structure is more closely related to the molecule than linear structures; it implies, however, more complex comparison algorithms. The descriptor combines conformation independence with alignment dependence, which makes it somehow unique. Alignment dependence is often seen as a disadvantage owing to the more time-consuming comparison and the potential bias resulting from heuristic alignment schemes. Both arguments do not hold in the case of Feature Trees, since the optimal alignment can be computed within milliseconds by employing dynamic programming techniques. Also, the alignment of structures allows for local comparison of molecular properties instead of a global comparison only. This yields not only a higher degree of accuracy, but also facilitates a variety of applications, several of which were reviewed in this chapter. The alignment dependence typically improves enrichment rates in ligand-based virtual screening. In HTS data analysis, a bridge between similarity-based and QSAR-based approaches can be built. For the analysis of fragment spaces, the local character of the Feature Tree comparison allows for a full, non-heuristic search. To our knowledge, comprehensive fragment space searching is a unique characteristic of Feature Trees. Finally, the alignment information can be used to draw compounds in such a way that similarity is expressed in their structure diagrams. Numerous other applications come to mind for which local, alignment-based comparison is advantageous such as combinatorial library design, library comparison or clustering of compounds. The further development of Feature Trees will therefore be an active field of research.

Acknowledgments

We would like to thank Dirk Tomandl (Santhera Pharmaceuticals) for stimulating discussions, molecular framework analysis and cheminformatics support. Thanks are also due to Marcus Gastreich (BioSolveIT) for his assistance in the CDK2 virtual screening experiment and for his general advice on chemistry matters. Finally we would like to thank Christian Lemmen (BioSolveIT) for many valuable suggestions on the manuscript.

References

1 Ehrlich, P. Address in pathology on chemotherapeutics: scientific principles, methods and results. *Lancet* **1913**; *ii*, 445–451.

2 Fischer, E. Einfluß der Konfiguration auf die Wirkung der Enzyme. *Berichte der Deutschen Chemiker Gesellschaft* **1894**, *27*, 2985–2993.

3 Lengauer, T., Lemmen, C., Rarey, M., Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27–34.

4 Rarey, M., Dixon, J. S. Feature Trees: a new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design* **1998**, *12*, 471–490.

5 Gillet, V. J., Downs, G. M., Holliday, J. D., Lynch, M. F., Dethlefsen, W. Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *Journal of Chemical Information and Computer Science* **1991**, *31*, 260–270.

6 Gillet, V. J., Willett, P., Bradshaw, J. Similarity searching using reduced graphs. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 338–345.

7 Barker, E. J., Gardiner, E. J., Gillet, V. J., Kitts, P., Moris, J., et al. Further development of reduced graphs for identifying bioactive compounds. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 346–356.

8 Tarjan, R. E. Depth-first search and linear graph algorithms. *SIAM Journal of the ACM* **1972**, *22*, 146–160.

9 Rarey, M., Stahl, M. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 497–520.

10 Rarey, M., Kramer, B., Lengauer, T., Klebe, G. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* **1996**, *261*, 470–489.

11 Zhang, K., Shasha, D. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing* **1989**, *18*, 1245–1262.

12 Tai, K.-C. The tree-to-tree correction problem. *Journal of the ACM* **1979**, *26*, 422–433.

13 Jiang, T., Wang, L., Zhang, K. Alignment of Trees – an alternative to Tree Edit. Department of Computer Science and Systems, Technical report, McMaster University, Department of Computer Science and Systems, 93,8, Hamilton, ON, **1993**.

14 Gupta, A., Nishimura, N. Finding largest subtrees and smallest supertrees. *Algorithmica* **1998**, *21*, 183–210.

15 Zimmermann, M., Master Thesis, Multiple Überlagerung von Wirkstoffmolekülen auf der Basis des FeaterTree Deskriptors, University of Bonn, Bonn, **1998**.

16 Waterman, M. S. *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman and Hall, New York, **1995**.

17 Ahuja, R. K., Magnati, T. L., Orlin, J. B. *Network Flows*, Prentice Hall, Englewood Cliffs, NJ, **1993**.

18 Böhm, H.-J., Schneider, G. *Virtual Screening For Bioactive Molecules*, Wiley-VCH, Weinheim, **2000**.

19 Ajay, N. J. Virtual screening in lead discovery and optimization. *Current Opinion in Drug Discovery and Development* **2004**, *7*, 396–404.

20 Walters, W. P., Stahl, M. T., Murcko, M. A. Virtual screening – an overview. *Drug Discovery Today* **1998**, *3*, 160–178.

21 Schneider, G., Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64–70.

22 Schneider, G., Lee, M. L., Stahl, M., Schneider, G. *De novo* design of molecular architectures by evolutionary assembly of drug-derived building blocks. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 487–494.

23 Jenkins, J. L., Glick, M., Davies, J. W. A 3D Similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *Journal of Medicinal Chemistry* **2004**, *47*, 6144–6159.

24 Andrews, K. M., Cramer, R. D. Toward general methods of targeted library design: topomer shape similarity searching

with diverse structures as queries. *Journal of Medicinal Chemistry* **2000**, *43*, 1723–1740.

25 Cheeseright, T., Mackey, M., Vinter, A. Peptides to non-peptides: leads from structureless virtual screening. *Drug Discovery Today: Biosilico* **2004**, *2*, 57–60.

26 Briem, H., Lessel, U. F. *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes. *Perspectives in Drug Discovery and Design* **2000**, *20*, 231–244.

27 Gray, N., Detivaud, L., Doerig, C., Meijer, L. ATP-site directed inhibitors of cyclin-dependent kinases. *Current Medicinal Chemistry* **1999**, *6*, 859–875.

28 Sielecki, T. M., Boylan, J. F., Benfield, P. A., Trainor, G. L. Cyclin-dependent kinase inhibitors: useful targets in cell cycle regulation. *Journal of Medicinal Chemistry* **2000**, *43*, 1–18.

29 Hardcastle, I. R., Golding, B. T., Griffin, R. J. Designing inhibitors of cyclin-dependent kinases. *Annual Review of Pharmacology and Toxicology* **2002**, *42*, 325–348.

30 Kim, K. S., Kimball, S. D., Misra, R. N., Rawlins, D. B., Hunt, J. T., Xiao, H. Y., Lu, S., Qian, L., Han, W. C., Shan, W., Mitt, T., Cai, Z. W., Poss, M. A., Zhu, H., Sack, J. S., Tokarski, J. S., Chang, C. Y., Pavletchi, N., Kamath, A., Humphreys, W. G., Marathe, P., Bursuker, I., Kellar, K. A., Roongta, U., Batorsky, R., Mulheron, J. G., Bol, D., Fairchild, C. R., Lee, F. Y., Webster, K. R. Discovery of aminothiazole inhibitors of cyclin-dependent kinase 2: synthesis, X-ray crystallographic analysis and biological activities. *Journal of Medicinal Chemistry* **2002**, *45*, 3905–3927.

31 Knockaert, M., Greengard, P., Meijer, L. Pharmacological inhibitors of cyclin-dependent kinases. *Trends in Pharmacological Sciences* **2002**, *23*, 417–425.

32 Steffen, A. University of Marburg, Marburg, personal communication.

33 Sigma-Aldrich Library of Rare Chemicals, **September 2002**, www.sigmaaldrich.com.

34 Chemstar Virtual Screening Database, **December 2001**, www.chemstar.ru.

35 Zimmermann, M., Tresch, A., Maass, A., Hofmann, M. Drilling into a HTS data set of E. coli dihydrofolate reductase. In *Proceedings of the 15th European Symposium on Quantative Structure-Activity Relationships* **2004**, Aki E., Yalcin J. (eds) published by Computer Aided Drug Design & Development Society, Turkey, pp. 408–409.

36 Zolli-Juran, M., Cechetto, J. D., Hartlen, R., Daigle, D. M., Brown, E. D. High throughput screening identifies novel inhibitors of *Escherichia coli* dihydrofolate reductase that are competitive with dihydrofolate. *Bioorganic and Medicinal Chemistry Letters* **2003**, *13*, 2493–2496.

37 Hindle, S. A., Rarey, M., Buning, C., Lengauer, T. Flexible docking under pharmacophore constraints. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 129–149.

38 Mason, J. S., Cheney, D. L. Ligand–receptor 3-D similarity studies using multiple 4-point pharmacophores. *Pacific Symposium on Biocomputing* **1999**, *4*, 456–467.

39 Leach, A. R., Green, D. V., Hann, M. M., Judd, D. B., Good, A. C. What are GaPs? A rational approach to monomer acquisition and selection. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 1262–1269.

40 Martin, E. J., Hoeffel, T. J. Oriented substituent pharmacophore PRopErtY space (OSPPREYS): a substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 383–403.

41 Bersuker, I. B., Bahceci, S., Boggs, J. E. Improved electron-conformational method of pharmacophore identification and bioactivity prediction. Application to angiotensin converting enzyme inhibitors. *Journal of Chemical Information and Computer Science* **2000**, *40*, 1363–1376.

42 Mayer, D., Naylor, C. B., Motoc, I., Marshall, G. R. A unique geometry consistent with structure–activity studies. *Journal of Computer-Aided Molecular Design* **1987**, *1*, 3–16.

43 Natesh, R., Schwager, S., Sturrock, E., Acharya, K. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature* **2003**, *421*, 551.

44 Vetter, D. Chemical microarrays, fragment diversity, label-free imaging by plasmon resonance – a chemical genomics approach. *Journal of Cellular Biochemistry* **2002**, *87*, 79–84.

45 Metz, G. Personal communication; www.graffinity.com.

46 Lewell, X.Q., Judd, D.B., Watson, S.P., Hann, M.M. RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Science* **1998**, *38*, 511–522.

47 *World Drug Index (WDI), Comprehensive Medicinal Chemistry (CMC), MDL Drug Data Report (MDDR), Available Chemicals Directory (ACD)*. Derwent Information, London, http://www.derwent.co.uk.

48 Bemis, G.W., Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.

49 Willett, P. Chemical similarity searching. *Journal of Chemical Information and Computer Science* **1998**, *38*, 983–996.

50 Downs, B. Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry*. Wiley-Interscience, New York, **2002**, pp. 1–40.

51 Zimmermann, M., Rarey, M., Naumann, T., Matter, H., Hessler, G. Extracting knowledge from high-throughput screening data: towards the generation of biophore models. In *Proceedings of the 14th EuroQSAR 2002 Symposium, Bournemouth, UK: Designing Drugs And Crop Protectants: Processes, Problems and Solutions*, Blackwell, Oxford, **2003**, pp. 63–67.

52 Boissonnat, J.D., Cazals, F., Flötotto, J. 2D-structure drawings of similar molecules. *Graph Drawing* **2000**, 115–126.

53 Fricker, P., Gastreich, M., Rarey, M. Automated drawing of structural molecular formulae under constraints. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1065–1078.

54 Zimmerman, B.L., PhD Thesis, Computerized-generated structural formulas with standard ring orientations; University of Pennsylvania: Philadelphia, PA, **1971**.

55 Helson, H.E. Structure diagram generation. *Reviews in Computational Chemistry*, Wiley-VCH, New York, **1999**, pp. 313–398.

56 Maehr, H. Graphic representation of configuration in two-dimensional space. Current conventions, clarifications and proposed extensions. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 884–902.

5
Concept and Applications of Pseudoreceptors

Klaus-Jürgen Schleifer

5.1
Introduction

Successful computational approaches in drug design are mostly based on experimentally determined protein structures of a molecular target or a multitude of chemical ligands with known pharmacological (*in vitro*) effects at the same receptor site. Whereas the so-called *structure-based drug design* uses the binding pocket of a protein as the lock in order to construct the best-fitting key (e.g. LUDI [1]), *ligand-based drug design* tries to extract key functions out of a ligand-based pharmacophore model for the prediction of biological effects of structural congeners (e.g. CoMFA [2] and CoMSIA [3]).

Starting in the late 1980s, a combination of these techniques was introduced and referred to as *minireceptor* or *pseudoreceptor modeling* [4–8]. A broader distribution concomitant with an increased number of publications was achieved by the commercially available pseudoreceptor modeling software package *Yak* (for technical details, see [9, 10]). The new concept allowed the construction of a three-dimensional peptidic pseudoreceptor around any single small molecule or molecular ensemble of interest (e.g. a set of superimposed ligand molecules). As a result, guided by permanent correlation of experimental and model-derived calculated free energies of binding, a host–guest-system is created, mimicking reasonably well the interaction pattern of a *real* binding site. At the same time, pseudoreceptors were constructed applying classical molecular dynamics simulations and force field minimizations [11, 12].

This chapter does not cover these classical approaches but rather focuses on basic principles, evolution and scientific applications based on specialized pseudoreceptor modeling software packages.

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

5.2 Methodology

In order to build up an atomistic pseudoreceptor model, the following basic steps have to be carried out:

- selection of a set of ligand molecules with known affinity (K_d) or activity (IC_{50} or EC_{50}) to the same receptor site
- generation of the bioactive 3D conformation for each of these ligands
- superposition of all ligands (i.e. pharmacophore construction)
- generation of vectors associated with directional interactions (i.e. H-bond extension vectors, lone-pair vectors and hydrophobicity vectors)
- vector-cluster analysis in order to characterize essential functional groups (*anchor points*) for receptor binding
- selection of appropriate binding partners for all free anchor points (e.g. amino acid residues or metal ions)
- successive retrieval of residue templates from database to saturate all vectors, docking, orientation and optimization of the ligand–pseudoreceptor complex.

Automation of these steps was accomplished by the pseudoreceptor modeling programs *Yak* [9, 10] and *PrGen* [13]. In order not to limit the use of only organic molecules (i.e. ligands and amino acid residues), the Yeti force field [14, 15] was implemented. This force field considers – in addition to organic residues – explicitly metal ions (e.g. Zn^{II}, Ca^{II} and Mg^{II}) that might be essential for the generation of metalloproteinic pseudoreceptors.

Generation of the ligand-specific interaction vectors is a fundamental basis for the ongoing saturation with amino acid residues. These may be chosen individually from a given database and placed according to the Ponder–Richards side-chain rotamer library [16]. Characterization of the vectors is simplified via a color-coded visualization of so-called hydrogen-extension vectors (HEVs; i.e. H-bond donors), lone-pair vectors (LPVs; i.e. H-bond acceptors) and hydrophobicity vectors (HPVs) (see Fig. 5.1). In the case that all molecules of interest direct equal vectors in a common direction, the tips of these vectors are ideal starting points for the placement of amino acid residues (e.g. LPV is saturated with the alcoholic hydroxyl group of a serine or threonine).

If individually positioned residues are in close contact, the program evaluates the possibility of peptide-bond formation and thus links single residues to a peptide. Furthermore, ligand-independent extension (i.e. residues without direct contact to vectors) of the pseudoreceptor can be used in order to complete the peptidic receptor site (e.g. entirely closed shell around the ligand molecules).

Subsequently, a receptor minimization is carried out by energy minimization of all residues keeping the position, orientation and conformation of the ligands unaltered. To achieve a high correlation between experimentally derived and calculated binding energies ($\Delta G^{\circ}_{\mathrm{exp}}$ vs $\Delta G^{\circ}_{\mathrm{calc}}$), *correlation coupling* may be used. This additional quantity in the energy term couples the actual root mean square deviation of $\Delta G^{\circ}_{\mathrm{calc}}$ and $\Delta G^{\circ}_{\mathrm{exp}}$ to the force field energy of the system. By a straightforward

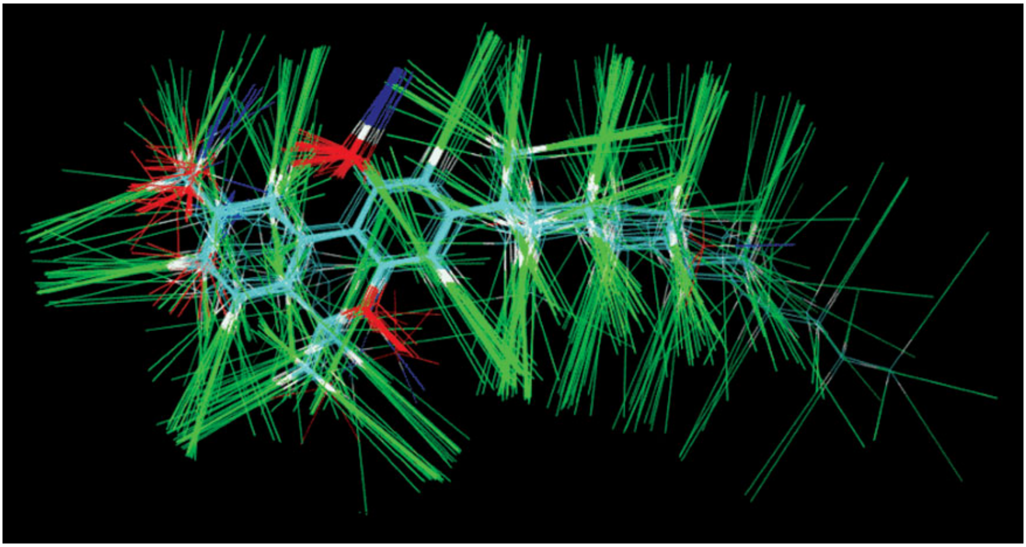


Fig. 5.1 Superimposed set of ligands with indicated hydrogen-extension vectors (HEVs; blue), lone-pair vectors (LPVs; red) and hydrophobicity vectors (HPVs; green).

minimization of this term in the course of correlation-coupled receptor minimization, an almost perfect correlation may be enforced. In the next step, the pharmacophore is allowed to relax by minimizing the ligands without constraints while the receptor remains fixed (*ligand relaxation*). In order to allow a more exhaustive search for the global minimum of the ligands, a Monte Carlo procedure may be chosen.

Internal ligand relaxation allows the removal of strain possibly imposed on the ligands by the receptor during correlation-coupled refinement but usually yields suboptimal models. Therefore, correlation-coupled receptor minimization followed by unconstrained ligand relaxation is repeated several times until a highly correlated pseudoreceptor model is obtained in the relaxed state (designated *ligand equilibration*).

To validate the equilibrated receptor, its potency to predict free energies of binding ($\Delta G^{\circ}_{\mathrm{pred}}$) is examined. Therefore, classical QSAR methods such as cross-validation via leave-*X*%-out analyses and/or prediction of activity for an external set of compounds (test set) are accomplished. Since all QSAR models are typically constructed to predict properties of new or even virtual molecules, model validation with an external test set reflects reality best (unbiased or biased random selection of training set and test set ligands is supported by the software).

The test set ligands have to be placed equally to the training set molecules into the pseudoreceptor and are minimized applying the same refinement protocol as described for the training set ligands. Finally, linear regression obtained for the training set is used to estimate free energies of binding for the test set derivatives (see Fig. 5.2).

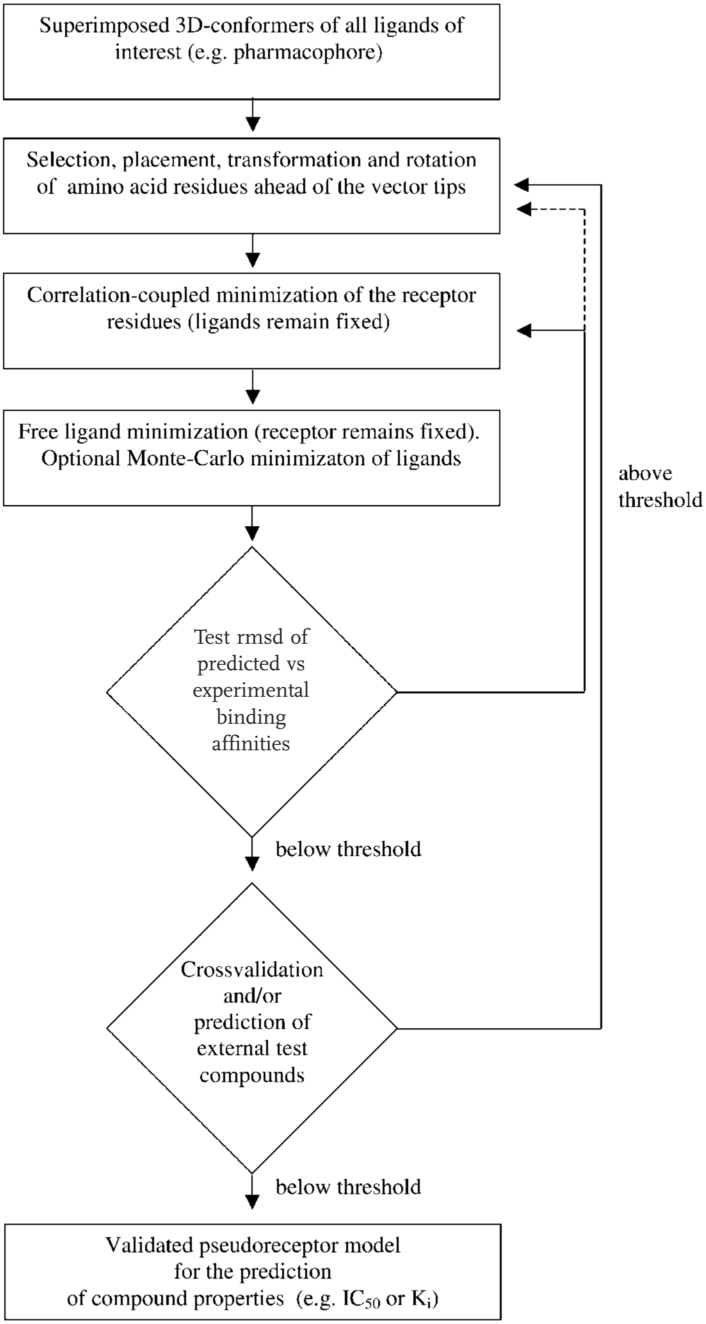

Superimposed 3D-conformers of all ligands of interest (e.g. pharmacophore)
Selection, placement, transformation and rotation of amino acid residues ahead of the vector tips
Correlation-coupled minimization of the receptor residues (ligands remain fixed)
Free ligand minimization (receptor remains fixed). Optional Monte-Carlo minimizaton of ligands
Test rmsd of predicted vs experimental binding affinities
above threshold
below threshold
Crossvalidation and/or prediction of external test compounds
below threshold
Validated pseudoreceptor model for the prediction of compound properties (e.g. IC_{50} or K_i)

Fig. 5.2 Flowchart of a typical *PrGen* pseudoreceptor model construction and validation approach.

Free energy of ligand binding is derived as follows:

$$E_{\text{binding}} \approx E_{\text{ligand-receptor}} - T\Delta S_{\text{binding}} - \Delta G_{\text{solvation,ligand}} + \Delta E_{\text{internal,ligand}} \tag{1}$$

In general, this equation is a combination of the approach of Blaney et al. [17] with the method of Still et al. [18] for estimating ligand solvation energies and a term to correct for the loss of entropy upon receptor binding following Searle and Williams [19]. The term $E_{\text{ligand–receptor}}$ corresponds to the enthalpic contribution of the ligand–receptor interaction and is determined using the directional Yeti force field [14, 15]. The term $\Delta G_{\text{solvation,ligand}}$ corresponds to the energy required to strip the solvent molecules off the ligands when binding from an aqueous environment to a hydrophobic receptor cavity. $\Delta G_{\text{solvation}}$ is calculated using the algorithm of Still et al. [18]. $T\Delta S_{\text{binding}}$ is estimated by assigning the amount of 0.7 kcal mol^{-1} to every freely rotatable single bond, excluding terminal methyl groups. The term $\Delta E_{\text{internal,ligand}}$ accounts for the potential increase in the ligand internal energy – relative to a strain-free reference conformation in aqueous solution – while bound to the receptor surrogate.

Free energy of binding is calculated according to the Gibbs–Helmholtz equation by conversion of experimental dissociation constants (K_d) at e.g. 25 °C:

$$\Delta G^{\circ}_{\text{exp}} = RT \ln K_d \approx 1.364\ (\text{kcal mol}^{-1}) \times \log K_d \tag{2}$$

Parallel to atomistic pseudoreceptors, a second strategy was embarked yielding *quasi-atomistic receptor models* by use of the program *Quasar* [20]. Instead of a shell of amino acid residues, a three-dimensional binding site surrogate, represented by a 3D (dot) surface surrounding a series of ligands at the van der Waals distance is generated. Each of these dots (called virtual particle) bears relevant atomistic properties (i.e. H-bond donor, H-bond acceptor, H-bond flip-flop particles, salt bridges, neutral and charged hydrophobic particles, virtual solvent and void space) that are visualized color-coded (Fig. 5.3).

Quasar not only considers one conformer per molecule but also represents each molecule by an ensemble of conformers in different orientations and protonation states (called fourth dimension [21]), thereby reducing the bias associated with the choice of the bioactive conformation. The fifth dimension refers to the possibility to consider an ensemble of different induced-fit models [22] and the sixth dimension allows for the simulation of local induced H-bond flip-flop and various solvation effects [23].

The multiple-conformation or multiple-orientation consideration of *Quasar* is taking into account an additional term ($\Delta E_{\text{env.adapt.,lig}}$) in Eq. (1) for the energy uptake on modifying the average receptor envelope to the individual receptor envelope:

$$\begin{aligned} E_{\text{binding}} \approx\ & E_{\text{ligand-receptor}} - T\Delta S_{\text{binding}} - \Delta G_{\text{solvation,ligand}} + \Delta E_{\text{internal,ligand}} \\ & + \Delta E_{\text{env.adapt.,ligand}} \end{aligned} \tag{3}$$

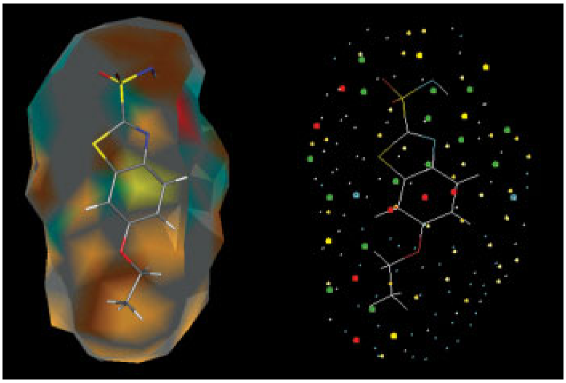


Fig. 5.3 Surface (left) and dot representation (right) of *Quasar*. Color-code indicates specific interaction sites projected to the van der Waals surface of a sulfonamide ligand, e.g. H-bond donor (green) and H-bond acceptor (yellow).

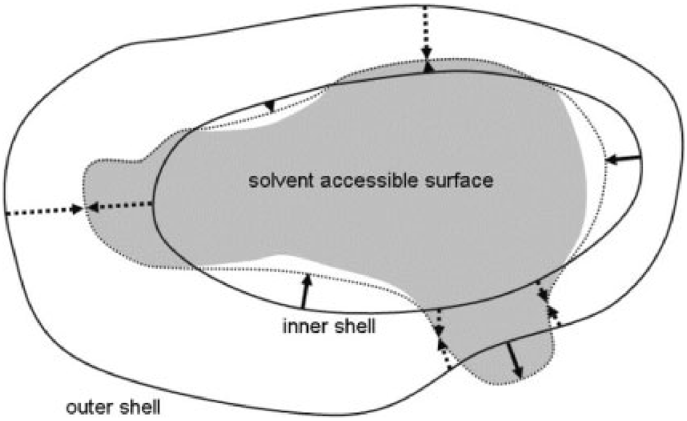


Fig. 5.4 Dual-shell representation of the receptor surrogate by *Raptor* [24]. During the steric adaptation process, the fields generated by the protein binding site on to the ligand's solvent accessible surface (SAS, gray surface) are computed by linear interpolation between inner and outer shell, if the ligand's SAS lies between those two shells (dashed arrows). For surface points located inside the inner layer, the latter may adapt only in part to the ligand topology (solid arrows; the dotted line represents the topologically adapted receptor surface).

To allow for ligand-dependent induced fit in receptor modeling, the software *Raptor* [24] uses a dual-shell representation of the binding site. The inner shell is yielded by optimizing the most potent compounds of the training set and represents the most favorable region of binding. Another compound featuring additional sterically demanding groups may experience different fields as a consequence of induced fit. Therefore, *Raptor* generates a second, outer layer, yielded by fields for altered binding behavior of compounds (Fig. 5.4).

5.3 Application of Pseudoreceptors

The pseudoreceptor modeling concept was utilized for (i) reconstruction of experimentally determined receptor sites, (ii) exploration of crucial ligand–receptor interaction sites and (iii) prediction of pharmacological activities of molecules, sometimes compared with results derived from other 3D-QSAR techniques.

One early study attempted to reconstruct the active site of the enzyme human carbonic anhydrase I (pdb code 2CAB) based on the structures of four potent sulfonamide inhibitors [9]. The central zinc ion and eight relevant amino acid residues for saturation of the vectors were extracted from the X-ray structure and placed around the four ligands. The final pseudoreceptor model derived from *Yak* was compared with the experimentally determined binding pocket geometry of the crystal structure. Superposition of both *receptors* indicates the zinc ion and the chelating histidine residues at almost identical sites (Fig. 5.5 a). Only the imidazole rings of two histidines are rotated. Most deviation is observed in the lower part of the binding site. In order to keep the putative binding geometry, the native glutamine (Gln) had to be replaced by a short-chain asparagine (Asn) in the pseudoreceptor. Thereby, essential interactions with the ligands (H-bond via the side-chain amide group) and the peptide link to the neighboring phenylalanine were enabled. Leu 131 is shifted closer to the ligands in order to increase the direct contact area.

Thr 199 and Leu 198 of the pseudoreceptor model are slightly shifted from their original positions. Nevertheless, two crucial H-bond interactions with the ligands have been conserved. It is interesting that occupation of the binding pocket with a sulfonamide ligand (pdb code 1AZM) does not dramatically change its topology. However, in analogy with the pseudoreceptor model, especially Leu 198 and Thr 199 are shifted away from the ligand (Fig. 5.5 b).

Although the orientations and calculated free binding energies of the sulfonamide ligands are not presented in the original paper, reconstruction of the receptor site via a pseudoreceptor approach was promising and stimulated further activities.

Sippl et al. [25] developed a binding site model for histamine H_3-receptor agonists based on a set of 16 histamine congeners with measured $\log(1/EC_{50})$ values between 3.7 and 9.12 (i.e. ΔG° from –5.2 to –12.9 kcal mol^{-1} at 37 °C).

The pseudoreceptor model was constructed with six amino acid residues. The imidazole ring of the ligands and the terminal basic function were saturated via

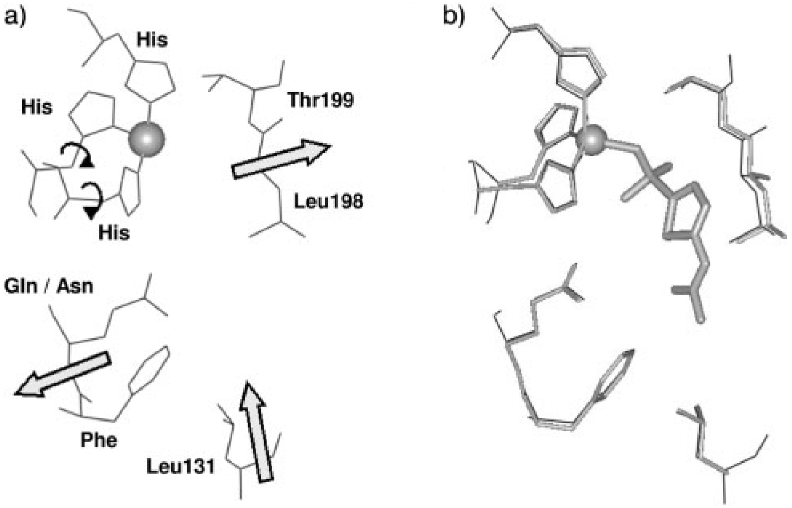


Fig. 5.5 (a) Binding site of human carbonic anhydrase I (pdb code 2CAB) with a bound zinc ion (sphere). Arrows indicate the directions of the pseudoreceptor model to rotate and shift amino acid residues. (b) Superposition of the empty (lines) and the sulfonamide occupied binding site (sticks, pdb code 1AZM) indicating the obvious shift of Leu 198 and Thr 199.

H-bond interactions (Tyr and Asn) and a charged aspartate (Asp). In addition, three suitable hydrophobic residues were chosen (Phe, Ile and Leu) to build up a hydrophobic cleft for the molecules. The correlation yielded for 12 training set molecules ($r^2=0.98$, r.m.s.d. = 0.21 kcal mol^{-1}) and four test set molecules (r.m.s.d. = 0.66 kcal mol^{-1}) indicates the high quality of the model. The topology of this virtual pseudoreceptor model was checked by means of X-ray crystallographically determined histamine complexes and interaction pattern derived from the program *GRID* [26] (Molecular Discovery, Oxford, UK). In this regard, two L-histidine-binding proteins of *Escherichia coli* (pdb code 1HSL) and *Salmonella typhimurium* (pdb code 1HBP) were investigated with respect to their imidazole ring complementarity. This conserved region bears a tyrosine and a leucine residue parallel to the aromatic imidazole, thus being in good agreement with the pseudoreceptor‘s phenylalanine and isoleucine (Fig. 5.6).

The probe-based algorithm *GRID* indicates three favorable common binding regions for an aliphatic hydroxyl group. The pseudoreceptor model occupies these regions with hydroxy (Tyr), carbonyl (Asn) and carboxylate functions (Asp) in order to saturate hydrophilic features (NH or =N) of the agonists (Fig. 5.6). Furthermore, the hydrophobic amino acid residues (Phe, Ile and Leu) are also located in the center of the grids generated with a hydrophobic methyl probe (not shown).

Schmetzer et al. [27] performed a joint *CoMFA* and *Yak* study for 31 cannabinoids acting on the CB1 receptor. Starting from a ligand-based pharmacophore model, a classical CoMFA investigation was accomplished yielding a correlation

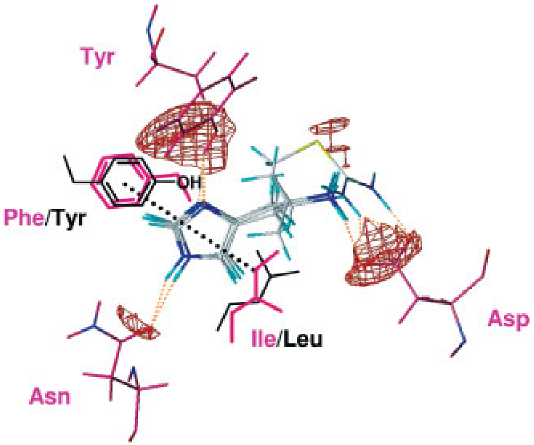


Fig. 5.6 Part of the pseudoreceptor model for histamine H_3-receptor antagonists [25] (one leucine beside the isoleucine residue is omitted for clarity). Hydrogen bonds from ligands to complementary functions of asparagine (Asn), tyrosine (Tyr) and aspartate (Asp) are located in the *GRID*-derived interaction space [26]. The model's phenylalanine and isoleucine residues parallel to the imidazole rings are positioned almost identically to detected residues (Tyr and Leu; black) of known histidine binding sites.

of $r^2 = 0.977$ (five principal components) and a leave-one-out cross-validation of $r^2_{cv} = 0.630$ ($s_{pred} = 0.792$). Subsequently, the same ligands were applied for a *Yak*-directed pseudoreceptor alignment in order to yield a pseudoreceptor-based pharmacophore model for the ligands. To compare both strategies, again *CoMFA* was accomplished for statistical analysis. It is interesting that in spite of a lack of an experimental 3D structure of the real binding site, the robustness of the new pseudoreceptor-derived pharmacophore model is significantly increased (Table 5.1).

This is inferred from better correlations (r^2 and r^2_{cv} values) and increased sensitivity associated with a reduction of principal components (PCs). Hence the r^2_{cv} value yielded for the pharmacophore alignment (PA) drops from 0.63 (five PCs)

Table 5.1 CoMFA statistics [a)] for 29 cannabinoids based on a pharmacophore (PA) and a pseudoreceptor (PrA) alignment

	r^2	s_{est}	r^2_{cv}	s_{pred}	$\Delta G_{ext.set}$	PC
PA	0.977	0.197	0.630	0.792	0.39/–0.13	5
PrA	0.985	0.161	0.788	0.603	0.70/–0.56	5

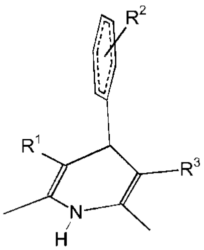
a) Standard error of estimate (s_{est}); leave-one-out cross-validated squared correlation (r^2_{cv}); standard error of prediction (s_{pred}); ΔG_{calc}–ΔG_{exp} in kcal mol^{-1} for two external test set derivatives ($\Delta G_{ext.set}$); principal components (PC).

to 0.54 (three PCs) while the pseudoreceptor-based alignment (PrA) with three PCs remains robust ($r^2_{cv}=0.64$).

The predictive powers of the two models are almost identical although, at first glance, deviation from experimental values is smaller making use of the classical pharmacophore–CoMFA approach [e.g. $\Delta\Delta G^\circ=0.39$ (PA) vs 0.70 kcal mol^{-1} (PrA)]. On closer inspection, one has to take into account that the pseudoreceptor-derived model shows a stronger internal correlation (higher r^2 and r^2_{cv} values) and therefore the differences in the prediction are statistically not relevant.

In a reversed type of approach, we used the pseudoreceptor modeling concept not to develop a tool for SAR predictions but to characterize two discrete ion channel modes on a molecular level [28]. For this purpose, a set of 13 well-characterized 1,4-dihydropyridine (DHP) derivatives with experimentally determined dissociation constants (K_d) for the voltage-gated calcium channel (VGCC) in the resting state (rs) and the open/inactivated state (is) were investigated (Table 5.2).

Table 5.2 Investigated DHP antagonists and agonists with corresponding experimentally determined (ΔG°_{exp}) free energies of binding (kcal mol^{-1}) in the resting state (rs) and the inactivated state (is) of VGCCs. Compounds **X–XIII** represent test set derivatives. Solvation energies of the ligands (ΔG_{solv}) are indicated in kcal mol^{-1}



Derivative	R^1	R^2	R^3	ΔG°_{exp} rs	ΔG°_{calc} is	G_{solv}
I	$COOCH_3$	2′-NO_2	$COOCH_3$	–10.502	–13.184	–14.198
II	$COOCH_3$	3′-CN	$COOCH_3$	–9.708	–12.108	–10.743
III	$COOCH_3$	4′-Cl	$COOCH_3$	–8.209	–8.964	–9.201
IV	NO_2	2′-OCF_2H	$COOCH_3$	–9.571	–10.474	–14.696
V	$COOCH_3$	2′-OCF_2H	NO_2	–9.264	–10.564	–14.543
VI	NO_2	2′-CF_3	H	–6.967	–7.634	–10.313
VII	H	2′-CF_3	NO_2	–7.364	–7.741	–10.330
VIII	NO_2	2′-CF_3	NO_2	–8.256	–9.110	–17.275
IX	NO_2	2′-OCF_2H	NO_2	–7.817	–8.660	–16.994
X	NO_2	2′-CF_3	$COOCH_3$	–9.704	–10.641	–14.150
XI	$COOCH_3$	2′-CF_3	NO_2	–8.803	–10.296	–13.929
XII	NO_2	2′-OCF_2H	H	–7.860	–8.277	–10.096
XIII	H	2′-OCF_2H	NO_2	–7.422	–7.783	–11.269

For the construction of the receptor envelope, only experimentally detected residues crucial for high-affinity binding or related (mainly smaller or more restricted) amino acids with identical functional features were considered.

Structural comparison of pure enantiomers demonstrated the importance of the right-hand side of DHPs for high-affinity binding in the resting state of the channel. Therefore, a threonine, which has been experimentally proven to be crucial for binding (Thr 1066), was placed as hydrogen-bond donor at this side of the pharmacophore. The NH function of the DHP ring was saturated by the carbonyl oxygen of the glycine backbone that imitates a crucial glutamine amide function (Gln 1070). A methionine (Met 1188 or Met 1491) was located axially beside and a phenylalanine on top of the substituted 4-phenyl ring. Two additional tyrosines (Tyr 1490) were arranged below the DHP ring and parallel to the 2′- and 3′-substituted phenyl ring, respectively. Correlation-coupled receptor minimization followed by free ligand relaxation obtained a satisfactory correlation of $R=0.99$ (r.m.s.d. = 0.097 kcal mol^{-1}) between experiment and calculation. To resolve the problem of multiple local minima in conformational space, a Monte Carlo search was performed to find the best adjustment of the ligands within the binding cavity.

Free energies of binding were successfully predicted for four test set ligands (r.m.s.d. = 0.532 kcal mol^{-1}) using the linear regression obtained with the training set. It should be mentioned, that two further test ligands could not be predicted accurately because of an additional charge-transfer interaction, which is not considered by classical force fields. The author reasonably explained this behavior by use of quantum mechanical calculations [28, 29].

With 19 additional nifedipine analogs, a second model was constructed supporting the above-mentioned receptor hypothesis of the resting state model (r.m.s.d. = 0.409 kcal mol^{-1} for the test set derivatives).

Selectivity of the resting state pseudoreceptor model was checked with the same ligands via prediction of their binding behavior to the inactivated channel mode.

In spite of a good correlation (r.m.s.d. = 0.115 kcal mol^{-1}) for the training set, prediction for the relevant test set molecules demonstrated the lack of any predictivity (r.m.s.d. value 2.033 vs 0.532 kcal mol^{-1}).

For the construction of a pseudoreceptor model of the open/inactivated state, ligand-derived information was considered indicating the left-hand side of DHPs to be essential for binding [28].

Based on these observations, a second hydrogen-bond donor (e.g. Lys$^+$, Arg$^+$, Ser, Gln, etc.) was placed at the left side to simulate the open/inactivated channel mode. A threonine proved to be best yielding perfect correlation for the training set (r.m.s.d. = 0.123 kcal mol^{-1}) and convincing prediction (r.m.s.d. = 0.848 kcal mol^{-1}) for the test set derivatives (Fig. 5.7 b).

Since all pseudoreceptor models are composed of the same six amino acid residues – Thr, Phe, Gly, Met, Tyr, Tyr – transition from resting to open/inactivated state could be described by one additional hydrogen-bond donor interaction (Thr) at the left-hand side of DHPs.

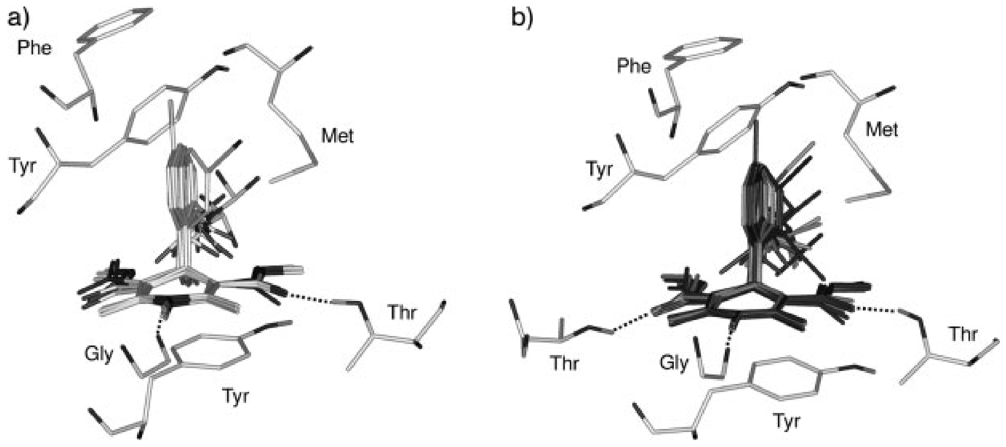


Fig. 5.7 Pseudoreceptor model of the L-type VGCC in the resting state (a) and the opened/inactivated channel mode (b) with one additional threonine (Thr) at the left side of DHPs. Training set (black) and test set structures (gray) are stabilized via three H-bonds (dashed lines). For clarity only NH and OH hydrogens are displayed.

Vedani et al. [30] used a 5D-QSAR *Quasar* approach in order to deduce novel ligands for the chemokine receptor-3 (CCR3). In this study, two receptor surrogates were built based on a total of 141 compounds comprising *N*-(alkyl)benzylpiperidines (series-1) and ureidoalkylpiperazines, aminoalkylpiperazines and amidoalkylpiperazines (series-2). In the simulations, these compounds were represented by a total of 421 conformers (4D-QSAR) while simultaneously exploring six different induced-fit scenarios (5D-QSAR): a linear induced fit scaled to 75%, four field-based modes (steric, electrostatic, H-bond, lipophilicity) and a protocol based on energy minimization. The receptor surface was constructed via the van der Waals surface generated starting from all conformers of the ligands defining the training set. Subsequently, domains or discrete points on the receptor surface were randomly populated with atomistic properties and optimized by simulating crossover events applying a genetic algorithm. For series-1 (40 training set and 10 test set compounds) and series-2 (66 training set and 25 test set derivatives) acceptable leave-8/11-out cross-validated r^2 values (0.95/0.86) and a predictive r^2 value of 0.879/0.798 for the test molecules, respectively, were obtained. Subsequently, both series were combined (the authors commented that the reason for not envisioning this first hand was of a technical nature) and a common receptor surrogate was generated. To match the volume of series-1, one torsional angle of the series-2 ligands had to be altered from gauche (–60°) to trans (180°) conformation. After 8000 crossovers (32 generations), the simulation reached the cross-validated r^2 of 0.907 (r.m.s.d. = 0.4 kcal mol^{-1}) and a predictive r^2 of 0.899 (r.m.s.d. = 0.34 kcal mol^{-1}). These quantities reflect values averaged over 250 models that, among themselves, differ in 27% of the mapped 138 properties on 342 available positions of the surface.

Based on these models and a functional-group analysis, 58 novel compounds with (i) lipophilic substituents (e.g. CH_3, Cl, CN) to increase hydrophobic inter-

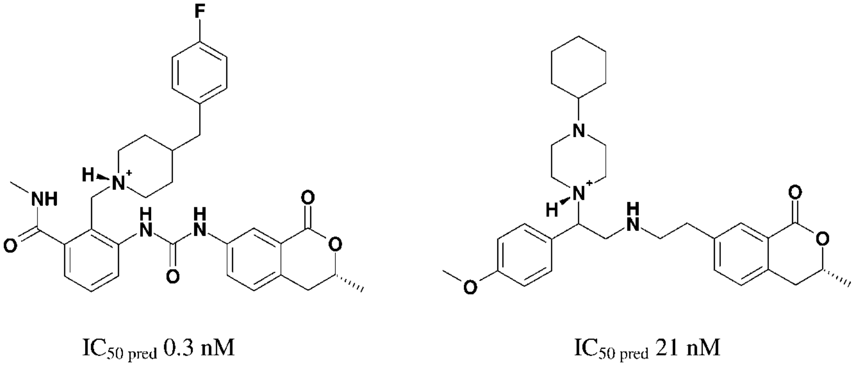


Fig. 5.8 CCR3 antagonists with predicted IC_{50} values ($IC_{50\ pred}$) derived from *Quasar* models for series-1 (left) and series-2 [30].

actions and simultaneously reduce desolvation energy or (ii) amphiphilic H-bond accepting moieties (e.g. acetate, pyridine, isocoumarin) aimed at strengthening hydrogen-bond interactions were screened *in silico*.

For series-1, 10 novel ligand molecules with calculated IC_{50} values from 0.3 to 16 nM were tested. For series-2, the 48 tested molecules yielded calculated K_i values from 21 nM to 2.5 μM. In both series, an isocoumarin ring combining H-bond-acceptor features with a delocalized, polarizable ring system yields best results (Fig. 5.8).

Unfortunately, the proposed novel structures have not been synthesized or tested so far. The fact that 11 of the proposed ligands show calculated binding affinities higher than any compound of the training set should initiate experimental inspection.

5.4 Conclusion

Pseudoreceptor models were originally constructed to yield an atomistic picture of hitherto unsolved receptor sites. By means of a specialized force field, this binding pocket was subsequently applied to derive 3D-QSAR studies (*Yak* and *PrGen*). Unrealistic consideration of one common binding site for a set of several tightly bound ligands yielded non-atomistic models (i.e. *Quasar* and *Raptor*) implicitly accounting for different binding poses of the ligands, "breathing" binding pockets and altered H-bond interaction patterns.

Use of these programs is documented in a multitude of publications trying to predict biological activity of compounds modulating the human carbonic anhydrase [9, 10], G protein-coupled receptors (GPCRs) [10, 13, 25, 27, 30, 31], varicella-zoster virus and human thymidine kinases [32], cytochrome P-450-dependent lanosterol 14α-demethylase ($P450_{14DM}$) [33], a sweet taste receptor [34], β-tubulin [35], ion channels [28, 36, 37] and others.

It will be interesting to follow whether pseudoreceptors up to the sixth dimension (6D-QSAR [23]) represent the final state or just a breakpoint to further evolutions (e.g. 7D-QSAR).

References

1 H.J. Böhm, *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

2 R.D. Cramer III, D.E. Patterson, J.D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

3 G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130–4146.

4 H. Frühbeis, R. Klein, H. Wallmeier, *Angew. Chem. Int. Ed. Engl.* **1987**, *26*, 403–418.

5 J.P. Snyder, S.N. Rao, *Chem. Des. Autom. News* **1989**, *4*, 13–15.

6 H.D. Höltje, S. Anzali, *Pharmazie* **1992**, *47*, 691–698.

7 J.P. Snyder, S.N. Rao, K.F. Koehler, R. Pellicciari, *Trends in Receptor Research*, Elsevier, Amsterdam, **1992**, 367–403.

8 J.P. Snyder, S.N. Rao, K.F. Koehler, A. Vedani, *3D QSAR in Drug Design*, ESCOM Science Publishers, Leiden, **1993**, 336–354.

9 A. Vedani, P. Zbinden, J.P. Snyder, *J. Recept. Res.* **1993**, *13*, 163–177.

10 A. Vedani, P. Zbinden, J.P. Snyder, P.A. Greenidge, *J. Am. Chem. Soc.* **1995**, *117*, 4987–4994.

11 V. Frecer, B. Ho, J.L. Ding, *Eur. J. Biochem.* **2000**, *267*, 837–852.

12 E. Gálvez-Ruano, I. Iriepa-Canalda, A. Morreale, K.B. Lipkowitz, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 57–68.

13 P. Zbinden, M. Dobler, G. Folkers, A. Vedani, *Quant. Struct.–Act. Relat.* **1998**, *17*, 122–130.

14 A. Vedani, D.W. Huhta. *J. Am. Chem. Soc.* **1990**, *112*, 4759–4767.

15 A. Vedani, D.W. Huhta. *J. Am. Chem. Soc.* **1991**, *113*, 5860–5862.

16 J.W. Ponder, F.M. Richards, *J. Mol. Biol.* **1987**, *193*, 775–791.

17 J.M. Blaney, P.K. Weiner, A. Dearing, P.A. Kollman, E.C. Jorgensen, S.J. Oatley, J.M. Burridge, J.F. Blake, *J. Am. Chem. Soc.* **1982**, *104*, 6424–6434.

18 W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

19 M.S. Searle, D.H. Williams, *J. Am. Chem. Soc.* **1992**, *114*, 10690–10697.

20 A. Vedani, M. Dobler, P. Zbinden, *J. Am. Chem. Soc.* **1998**, *120*, 4471–4477.

21 A. Vedani, D.R. McMasters, M. Dobler, *Quant. Struct.–Act. Relat.* **2000**, *19*, 149–161

22 A. Vedani, M. Dobler, *J. Med. Chem.* **2002**, *45*, 2139–2149.

23 A. Vedani, M. Dobler, M.A. Lill, *J. Med. Chem.* **2005**, *48*, 3700–3703.

24 M.A. Lill, A. Vedani, M. Dobler, *J. Med. Chem.* **2004**, *47*, 6174–6186.

25 W. Sippl, H. Stark, H.D. Höltje, *Pharmazie* **1998**, *53*, 433–437.

26 P.J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.

27 S. Schmetzer, P. Greenidge, K.A. Kovar, M. Schulze-Alexandru, G. Folkers, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 278–292.

28 K.J. Schleifer, *J. Med. Chem.* **1999**, *42*, 2204–2211.

29 K.J. Schleifer, *Pharmazie* **1999**, *11*, 804–807.

30 A. Vedani, M. Dobler, H. Dollinger, K.M. Hasselbach, F. Birke, M.A. Lill, *J. Med. Chem.* **2005**, *48*, 1515–1527.

31 J.M. Jansen, K.F. Koehler, M.H. Hedberg, A.M. Johansson, U. Hacksell, G. Nordvall, J.P. Snyder, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 812–818.

32 P.A. Greenidge, A. Merz, G. Folkers, *J. Comput.-Aided Mol. Des.* **1995**, *9*, 473–478.

33 M. Botta, F. Corelli, F. Manetti, C. Mugnaini, A, Tafi, *Pure Appl. Chem.* **2001**, *73*, 1477–1485.

34 A. Bassoli, L. Merlini, G. Morini, *Pure Appl. Chem.* **2002**, *74*, 1181–1187.

35 L. Maccari, F. Manetti, F. Corelli, M. Botta, *Farmaco* **2003**, *58*, 659–668.

36 K.J. Schleifer, E. Tot, H.-D. Höltje, *Pharmazie* **1998**, *53*, 596–602.

37 K.J. Schleifer, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 467–475.

6
Pharmacophores from Macromolecular Complexes with LigandScout

Gerhard Wolber and Robert Kosara

6.1
Introduction

Amongst others, the concept of describing pharmacon–drug interactions via pharmacophore models consisting of relevant chemical features has become a well-accepted technique, which is very appropriate for use in high-throughput virtual screening [1, 2]. There are two ways of modeling drug–target interactions: either by starting from a set of ligands that are known to bind to the same target in a comparable way (ligand-based drug design) or by investigating the geometry of the target and the bound ligand if its structure is available (structure-based drug design). While several pharmacophore approaches exist for ligand-based design [3], structure-based design is often still performed using docking algorithms. This chapter will introduce LigandScout, a program for structure-based pharmacophore modeling.

6.1.1
Structure-based Drug Design Methods

If the 3D coordinates of the target receptor obtained from X-ray or NMR structure analysis are known, the most obvious way of deriving a model for drug–receptor interaction is to investigate a molecule's complementarity with the target binding site. A commonly used structure-based drug design approach is docking small molecules into the binding site assuming the binding site to be rigid and the ligand to be flexible. There are several docking tools available: Rarey et al. developed a tool called FlexX [4–6] that is able to find the bio-active conformation of many known complexes without manual interaction. DOCK, developed by Oshiro and Kuntz [7], uses several scoring functions to evaluate conformational space and the steric interaction of the ligand with the protein. AUTODOCK, a docking tool developed by Morris et al. [8], uses a rapid grid-based method of energy evaluation operating on the AMBER [9] force field. A third tool named GOLD, developed by Jones et al. [10, 11], is based on a genetic algorithm and supports partial protein flexibility.

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

A slightly different approach was presented in the computer program LUDI by Böhm [12]. It searches for interaction centers in the protein and assembles potential new ligands by combining fragments from a three-dimensional structure library. The scoring function for selecting the fragments depends not only on chemical features, but also on the number of rotatable bonds available in the yet to be assembled ligand molecule. Structure-based design becomes even more interesting when the increasing number of known target structures is considered. Since high-throughput *in silico* screening of combinatorial libraries has become relevant, docking tools have been used for that purpose. However, the re-investigation of the binding site for each docked ligand is computationally inefficient compared to the mapping of a ligand to a pre-elucidated pharmacophore describing the binding site.

6.1.2
Why Structure-based Pharmacophores?

When describing target–ligand interactions, it is essential that the hypothetical model is (i) selective enough to filter active compounds from inactive ones and that it is (ii) general enough to find new ligands that were not known at the time of model creation. The model definition should be so general that it can be applied across several classes of compounds and still reflect the desired mode of action. It should (iii) be possible to match automatically large compound libraries in a batch process in reasonably short time. We describe here our approach to design models fulfilling these requirements by implementing a simple and robust pharmacophore definition that is transparent to the user and is capable of being adapted once additional information on a specific ligand–target interaction becomes available.

In order to prove its general applicability, pharmacophores were derived from the biggest repository of publicly available protein–ligand complexes: the Protein Data Bank (PDB) [13]. Work with the PDB brings up some problems resulting from historical growth. Whereas recently submitted complexes may conform to a high-quality standard, structures submitted decades ago may show severe geometric problems and even missing atoms and bonds [14]. We present an advanced preprocessing procedure that incorporates already published techniques [14–16] for cleaning up this information, which is a prerequisite for chemical investigation and ultimately automatic pharmacophore generation.

6.2
The Data Source: Clean-up and Interpretation of PDB Ligand Molecules

Data formats in the Brookhaven Protein Databank have become an intensively discussed topic in the last few years. The original PDB file format [17] was created in the late 1970s and maintained by the Research Collaboratory for Structural Bioinformatics [13]. In order to improve the organization of bibliographic information and meta data, the mmCif format [18], which is a subset of the STAR (Self-defining Text Archive and Retrieval) format [19], was created in 1990 and has been continuously modified since then. mmCif is based on a defined set of data items, described in a dictionary description language (DDL), and enables potential software to more easily parse and validate content and meta data. Bernstein et al. created a tool for converting PDB to mmCif [20], which has to fight inconsistencies and ambiguities in the historically grown file format. In order to address these problems, the PDB Data Uniformity Project was initiated to improve data quality by correcting obvious errors [21]. In the past few years, XML technologies [22, 23] have become a general standard for storing and converting all kinds of data and technically more robust solutions than mmCif based on XML, such as PROXIML [24], were proposed but still not accepted by the Research Collaboratory for Structural Bioinformatics.

Our tool LigandScout deals with data mining in protein complexes that have been continuously submitted for over 30 years. The file format used was mainly created to describe proteins, never focusing on ligands or their detailed description. In order to yield the best results from data gathering, a potential interpretation algorithm needs to eliminate any possible means of data tampering caused by automated conversion. This is the reason why this work has to perform interpretation starting from the slightly outdated original PDB file format, which was used to submit the largest part of all ligands complexed in proteins.

The first part of this section describes how this interpretation is done and which assumptions have been made to retrieve plausible results. The second part will then describe the algorithms used to compute bond characteristics from geometric information, because the PDB file format and its successors include no means to specify hybridization states or bond orders, which are essential for the characterization of properties of small organic ligands. Ligand bond characteristics interpretation is a prerequisite for the step to follow: the detailed description of protein–ligand interactions by pharmacophore models.

6.2.1 Topological Analysis

A first step in the interpretation of the ligands is the analysis of topological information (typically referred to as “2D information”) contained in the molecular graph leaving the positions of the atoms disregarded. These calculations regularly can be performed at a much lower computational cost than a three-dimensional analysis. Brown and Martin [25] even concluded that 2D descriptors generally contain the same information as geometric descriptors.

In the case of PDB ligands, only part of the graph information is defined in the PDB file format sepcification: bond types and atom hybridization states are missing and will later be derived from the three-dimensional arrangement of connected atoms. Connectivity information, however, is already present and can be used to seperate cyclic from non-cyclic molecule parts. Especially for planar rings, this separation is a prerequisite for geometry interpretation, because of their different geometry characteristics: neighboring bonds to an sp^3 atom have a default tetrahedral bond angle of 109.5°, which can be distinguished from an sp^2 atom with a default bond angle of 120°. A planar aromatic five-membered ring has a typical angle of 108°, even though it contains sp^2 hybridized atoms.

From a graph-theoretical point of view, the recognition of chemically relevant cycles is not a trivial problem in terms of what kind of graph cycles are chemically relevant, which is also reflected in the extended discussion and large number of papers dealing with this problem [26–30]. Lynch et al. [31] reviewed ring perception algorithms for chemical graphs and investigated which kind of ring set suffices to describe all rings in a molecule. It was concluded that there are many valid solutions for describing a ring set, as long as the description is consistent and reproducible. An enumeration of all possible *smallest sets of smallest rings* is suggested to meet this requirement.

With the aim of interpreting and finding planar rings in biologically relevant ligands, reproducibility for complex ring combinations is not essential provided that all relevant ring atoms are covered. Therefore, an efficient algorithm for finding only one smallest set of smallest rings was implemented and will be described in the next section. From the many reports on algorithms for finding the smallest set of smallest rings (SSSR), the report of Figueras describing an algorithm using breadth-first search [32] was chosen to form the basis for adapting the existing data structures of the ilib framework. The reasons for this choice are the awareness of the relevant previous reports [29, 30] in his paper, a very concrete description of the algorithm presented and its high efficiency. It was shown that, for common cases, the breadth-first approach can be 2000 times faster then the previously suggested depth-first search. For this algorithm, the following terms need to be defined:

- *path:* a list of adjacent atoms describing a "walk" through the molecular graph
- *ring closure:* a cycle in the molecule graph
- *path intersection between P_1 and P_2*: a path I that contains all elements (atoms) from P_1 and from P_2 that are part of both P_1 and P_2
- *valid ring closure:* a ring closure that has a path intersection consisting of exactly one atom.

The assembly of the ring set was implemented as follows:

1. Simultaneously extend paths from a starting atom following all neighbors in a breadth-first manner.
2. If an atom already has a path, a ring closure is found. If this closure is a valid ring closure (i.e. the intersection of the colliding paths is a one-atom path), the corresponding path will be added to the set of rings.

The implementation was adapted using reference lists instead of arrays performing the following steps:

1. Create a BFS-Queue Q containing the first atom.
2. Get the first atom a from the queue and remove it from Q.
3. For each neighboring atom n with respect to a, which is not equal to a:

- If the path of n is *null*, extend path by n and put n to the back of the queue.
- If the path belonging to n is not *null*, compute the intersection between the path of the first element of Q and the path of n. If the intersection is a single atom, compute the ring from the path of n appended by the path of the first element of Q.

4. Go to step 2.

All atoms with more than one neighbor are checked for valid ring closures using the procedure described above and are added to the set of rings if the found ring is not already a member of the ring set.

6.2.2 Geometric and Semantic Analysis

From the PDB and topological analysis, we construct a molecular graph with untyped bonds and recognized chemical rings. For correct chemical recognition, we need to determine hybridization states, from which we then try to derive correct bond types.

The first assignment step was to detect planar rings that obey the Hückel rule of $4n+2$ p-electrons and flag them as aromatic, i.e. force all the atoms into an sp^2 hybridization state. For planarity recognition, three arbitrarily chosen adjacent ring atoms form a base plane. Subsequently, for each atom of the ring, the distance to this base plane is computed, d being zero for the ideal case. In vector notation, the plane equation can be described as follows:

$$(\vec{P} - \vec{P}_{Atom1})(\vec{P} - \vec{P}_{Atom2})(\vec{P} - \vec{P}_{Atom3}) = d$$

where $\vec{P}_{Atom1}$, $\vec{P}_{Atom2}$ and $\vec{P}_{Atom3}$ form a plane and $\vec{P}$ is the point to be investigated. In component notation, the same equation can be written as determinant derived from the 3D coordinates of the three atoms *A1*, *A2* and *A3*:

$$\begin{vmatrix} x - x_{A1} & y - y_{A1} & z - z_{A1} \\ x_{A2} - x_{A1} & y_{A2} - y_{A1} & z_{A2} - z_{A1} \\ y_{A3} - x_{A1} & y_{A3} - y_{A1} & z_{A3} - z_{A1} \end{vmatrix} = d$$

As rings are rarely perfectly planar, a threshold for the in-plane-distance had to be determined: Comparing distorted pyrroles in protoporphyrin rings with sp^3 atoms in pyrrolidines reveals that a threshold of 0.65 Å is adequate for the determinant described above.

For hybridization state assignment in chains, a novel method using geometry templates is created. The reason for this new method was that all previously used methods are derived from force-field potentials and use bond angles or tetrahedral angles. Whereas these angles only consider at most two neighbors of an atom, our new method considers all neighbors and their geometric relations.

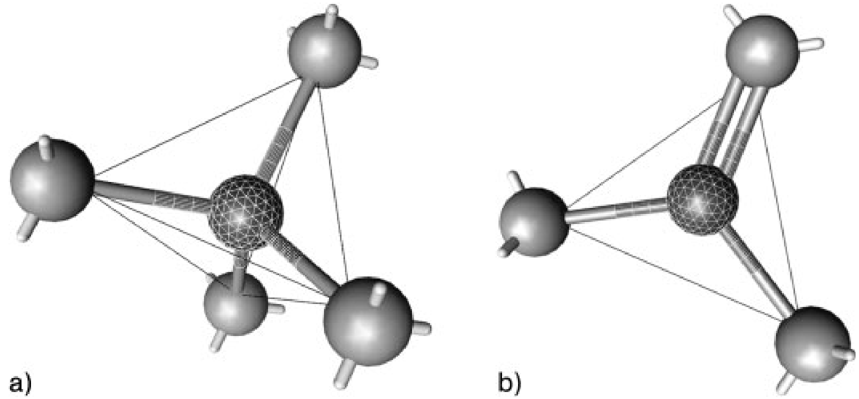


Fig. 6.1 Geometry templates for a tetrahedral sp^3 atom (a) and a triangular planar sp^2 atom (b).

For each hybridization state, a rigid geometric body is created and optimally aligned with the central atom. For sp^3, a tetrahedral geometric template is created (see Fig. 6.1 a), whereas sp^2 atoms are represented by a planar triangle (see Fig. 6.1 b) and an sp state is indicated by a linear alignment. Absolute distances are then summed and divided by the number of neighbors to determine which template fits best and determines the hybridization state. Details on this algorithm can be found in a previous paper describing LigandScout in more detail [33].

6.2.3 Double Bond Distribution

Whereas sp^3 and sp hybridization states directly indicate bond order, sp^2 atom chains show an alternating bond pattern and therefore the assignment of bond orders needs some further investigation. The assignment of commonly occurring functional groups and double bond patterns as proposed by Sayle [16] is followed by a procedure that recursively assigns the maximum number of double bonds to the remaining adjacent sp^2 atoms.

6.3 Chemical Feature-based Pharmacophores Used by LigandScout

Selecting chemical feature types for the construction of pharmacophores is the most important step towards pharmacophore creation. In early pharmacophore modeling techniques such as the active analog approach described by Marshall et al. [34], pharmacophore features could contain any fragment or atom type. Later techniques such as the software package Catalyst [35], however, use a more general way for building pharmacophore queries, for example using a single geometric entity for all negative ionizable groups. In real-life applications, these features are customized to achieve the desired filtering efficiency [36]. The discussion below will show that several arguments exist for continuing with this trend and even further extending the generalization of chemical functionalities.

6.3.1 Characteristics of Chemical Features: Specific or Comparable?

General definitions may result in models that are universal to the detriment of selectivity. Selectivity, nonetheless, is a major issue in pharmacophore validation and, therefore, feature descriptions that are too general need to be changed from reflecting universal chemical functionality to representing distinct functional groups. A common approach is to derive a model from distinct ligands in order to represent the specific mode of interaction as a chain of functional groups or exclusions thereof [36]. By restricting general chemical feature definitions in the way described above, the number of standard well-known features increases at the cost of comparability. However, only comparable pharmacophores are sufficiently universal and can represent a mode of action instead of a set of already existing ligands. Additionally, automated processing of pharmacophores becomes more transparent if chemical features stay comparable.

In order to describe the levels of universality and specificity of chemical features, a simple layer model is proposed to allow referral to these properties more easily. Table 6.1 shows a proposed classification of abstraction layers of chemical features: A lower level corresponds to higher specificity and therefore lower universality.

Below are some examples of chemical features according to the specificity of this layer model:

- *Layer 1:* A phenol group facing a parallel benzenoid system within a distance of 2–4 Å.
- *Layer 2:* A phenol group.

Table 6.1 Abstraction layers of chemical feature constraints

Layer	Classification	Universality	Specificity
4	Chemical functionality (positive ionizable area, lipophilic contact) without geometric constraint	+++	–
3	Chemical functionality (hydrogen bond donor, acceptor) with geometric constraint	++	+
2	Molecular graph descriptor (atom, bond) without geometric constraint	–	++
1	Molecular graph descriptor (atom, bond) with geometric constraint	– –	+++

Table 6.2 SMARTS patterns for chemical features (HBA-F, hydrogen bond acceptor/electrostatic fluorine interaction; HBD, hydrogen bond donor; PI, positive ionizable; NI, negative ionizable)

	Inclusion patterns	Exclusion patterns
HBA-F	{[O,S]}[#1]	c1nnnn1
	{N}[#1]	
	C{F}	
HBD	{[N,O,S;X1,X2]}	[–,–2,–3]
PI	{[NX3]}([CX4])([CX4,#1])[CX4,#1]	
	{N}=[CX3]({[N;H1,H2]})[! N]	
	N=[CX3]({[NH1]}){[NH1]}	
	{[+,+2,+3;! $(*[–,–2,–3])]}	
NI	[S,P](={O})(={O}){[OH]}	
	[S,C,P](={O}){[OH]}	
	{c}1{n}{n}{n}{n}1	
	{[–,–2,–3;! $(*[+,+2,+3])]}	

- *Layer 3:* H-bond acceptor vector including an acceptor point as well as a projected donor point; aromatic ring including a ring plane.
- *Layer 4:* H-bond acceptor without the projected point; lipophilic group.

The most frequent reason for creating features on the low universality levels 1 and 2 is that the definitions of the higher levels are not sufficient to describe the features occurring in the training set (see [36] for an example). Even if customization results in a layer 1 or layer 2 feature, there should be a possibility of including layer 3 or 4 information in order to categorize and thus increase comparability (for example, a carboxylic acid as a layer 2 feature is a subcategory of 'negative ionizable', which is a layer 4 feature).

LigandScout should serve as a basis for the comparison of feature locations and properties. Therefore, we needed to design a chemical feature set that is still universal but yet selective enough to reflect all relevant types of ligand-receptor interaction. This set is described in the next section.

6.3.2
Fully Automated Perception of Chemical Features

The chemical feature definitions described in the following sections are all categorized into hydrogen bond interactions, which are described as layer 3 features, and also into charge interactions and lipophilic interactions, which are represented as level 4 features. All current layer 4 features are represented as points with a tolerance radius forming a sphere, whereas layer 3 features are represented by vectors. The chemical feature definitions are specified in Daylight SMARTS notation [37] (listed in detail in Table 6.2) and directly imported into the program. The SMARTS syntax has been slightly extended by marking atoms used for geometry constructions by the insertion of braces, { and }, around the atom definition. If several atoms are marked, the center point between them is regarded as the reference point.

6.3.3 Vectors: Hydrogen Bonding

Hydrogen bonding occurs when covalently bound hydrogen atoms with a positive partial charge interact with another atom with a negative partial charge [38]. This typically happens when the partially positively charged hydrogen atom is positioned between partially negatively charged oxygen and nitrogen atoms, but is also found in different situations as described below. Additionally, electrostatic interactions of fluorine atoms with hydrogen bond donors were treated in a similar way as hydrogen bond acceptors, because they exhibit comparable geometric constraints.

Assuming an ideal hydrogen bond angle of 180°, the hydrogen bond is supposed to be broken when the angle difference exceeds 34°. This was derived from hydrogen bonding in water [39] and is reflected in the 146° angle shown in Fig. 6.2

For an sp^2 hybridized donor atom, the position of the added hydrogen can be used as a basis for the calculations because there is only one plausible coordinate position. If the angle formed by the two heavy atoms and the shared hydrogen exceeds 146°, the H-bond is considered to be broken. For the case of an sp^3 atom being the donor atom, the hydrogen may rotate freely and, consequently, the artificially added hydrogen position cannot be taken into account. In order to formulate a valid constraint reflecting hydrogen bond plausibility, the first adjacent atom on the donor side is added to the consideration. The angle that determines when the hydrogen bond is considered to be broken is set to a default deviation of 34° [39]. Hydrogen bond acceptors are to be regarded symmetrically.

6.3.4 Points: Lipophilic Contacts and Charge-transfer Interactions

Both lipophilic contacts and charge-transfer interactions represent layer 4 features and are implemented as 3D points with a specified tolerance.

6.3.4.1 Hydrophobic Contacts

Similarly to the concept of the Catalyst software package [35, 40], hydrophobic areas are implemented in the form of spheres located in the center of hydrophobic atom chains, branches or groups. First, a hydrophobicity scoring function pursuant to the Catalyst definition is implemented. As a next step, the algorithm checks if an ensemble of adjacent atoms is able to attain a sufficient overall hydrophobicity score. If this condition is met and a hydrophobic area in the macromolecule exists, a level 4 feature consisting of a sphere with a tolerance

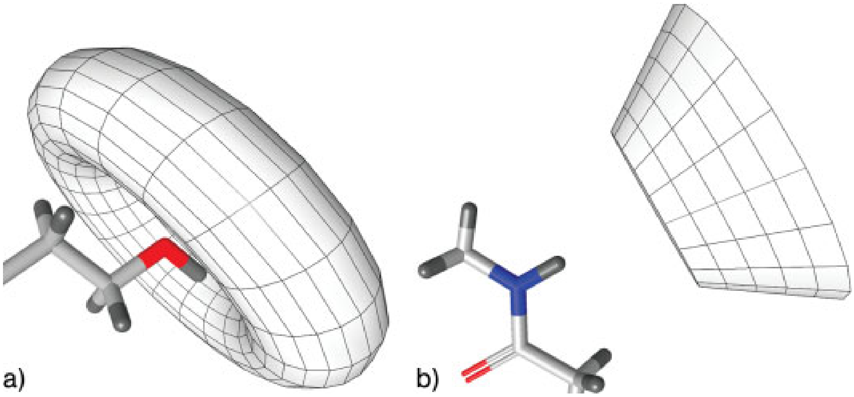


Fig. 6.2 Hydrogen bonding geometry for sp^2 (a) and sp^3 atoms (b).

radius of 1.5 Å is added to the weighed center of these atoms. A hydrophobic feature sphere is added if and only if a hydrophobic feature exists on the macromolecule side within a distance of 1–5 Å. The maximum distance was set to 5 Å because a larger gap would permit water molecules to be located in between.

6.3.4.2 **Positive and Negative Ionizable Areas**

Positive ionizable areas are represented by atoms or groups of atoms that are likely to be protonated at a physiological pH. These are summarized in Table 6.2 and include basic amines, basic secondary amidines, basic primary amidines, basic guanidines and positive charges not adjacent to a negative charge. Negative ionizable areas are atoms or groups of atoms that are likely to be deprotonated at physiological pH, including sulfonic acids, phosphonic acids, sulfinic, carboxylic or phosphinic acids, tetrazoles and negative charges which are not neutralized by adjacency to a positive charge.

A sphere with a tolerance radius of 1.5 Å represents an ionizable feature and is added for the case that a reversely charged ionizable feature can be found on the macromolecule side within a plausible distance range. This interval is set to 1.5–5.6 Å and is user-adjustable.

6.4
Overlaying Chemical Features

For the purpose of finding common chemical features occurring in different structures with bound ligands in comparable binding modes, an overlaying algorithm for several pharmacophore models is implemented: A compatibility graph of all feature point pairs is constructed regarding vector base points and projected points as independent classes of chemical features. Feature pairs are defined as being compatible if and only if their distance range lies within the defined tolerance for both tolerance spheres. From the largest compatible subset of chemical feature compatibility pairs determined by maximum clique detection [41], two distinct common feature pharmacophore models are formed and subsequently aligned in 3D space performing a single rotation. The calculation of the transformation matrix necessary for the rotation uses an analytical, efficient algorithm by Kabsch [42], which supports the alignment of point sets with specific weights. A weight of 1.0 was assigned to chemical features located on the ligand and a weight of 0.1 to excluded volume spheres; this was necessary to prioritize the alignment of chemical functionality rather than the alignment of exclusion spheres.

For sufficiently similar models, this algorithm is extremely useful for deriving "common-feature" pharmacophores. Its use, however, is restricted to the nearly identical binding pockets; conformational differences in the two compared protein pockets will lead to useless results.

6.5 3D Visualization and Interaction

In order to understand the three-dimensional structure of the ligand and its environment, and to interact with it, an interactive graphical user interface including visualization techniques was implemented in LigandScout. The visualization is implemented using OpenGL, an industry standard for 3D graphics programming, which leverages available graphics hardware to achieve fast rendering and interaction. OpenGL is available on virtually all current operating systems and thus helps to make the program largely platform-independent.

When a new macromolecule is loaded, it is first displayed in a mode showing only its backbone [43], in addition to the binding sites containing the ligands. Clicking on a binding site will take the user to a different view of just this ligand and its immediate environment (Fig. 6.3).

6.5.1 Core and Environment Visualization

When visualizing ligand–macromolecule interactions, we differentiate between the ligand "core" molecule) and its immediate environment. The environment consists of all atoms in the macromolecule residing within a distance of up to 5 Å from the ligand.

Molecule visualization adheres to the standard conventions for depicting molecules in chemistry by providing the usual display modes for molecules, using the standard colors for elements and so on. Additionally, it can optionally show context-related information in tool-tips (similar to additional information in user interfaces), and also provide means to interact with the molecule.

There are several different display styles that can be used to depict the molecule and its environment (Fig. 6.4). In the standard setup, the core is drawn using the stick mode, and the environment is rendered using lines. The differ-

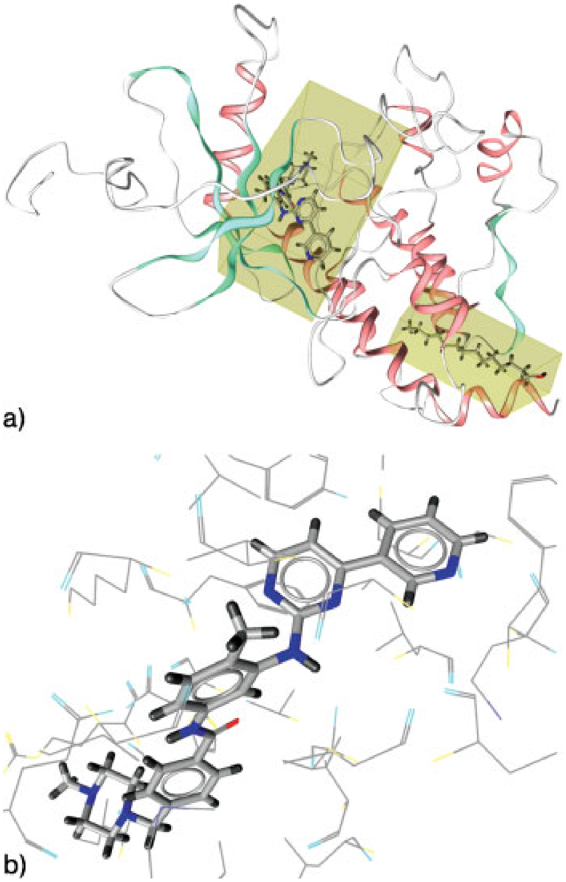


Fig. 6.3 After loading a PDB file, the macromolecule's backbone is shown, along with the binding sites (a). After selecting a binding site, the user is taken to a zoomed-in view of the ligand and its immediate environment (b).

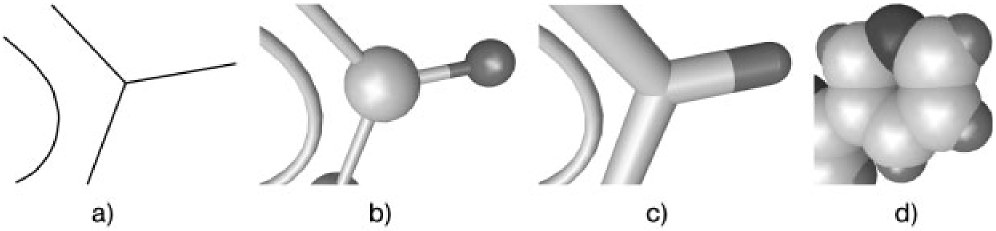


Fig. 6.4 Render styles in LigandScout: (a) lines, (b) ball and stick, (c) stick and (d) spacefilling/CPK.

ence between the two is thus clearly defined and the environment is less likely to occlude parts of the ligand.

Hydrogens can be shown or hidden and they can also be drawn as stubs (Fig. 6.1). Stubs are useful to show the possible positions of hydrogens when editing molecules or when hydrogens are of little interest.

For human perception of the depth of three-dimensional molecular structures, it is important to provide depth cues [44]. Movement is the strongest depth cue, so it is very important to render the image quickly enough to provide smooth movements of the object on the screen. Lighting is used to provide a better impression of the shapes of objects, and also additional depth. Finally, objects that are further away appear as if seen through fog, which improves the impression of three-dimensional space.

6.5.2 Pharmacophore Visualization

LigandScout pharmacophores solely consist of chemical features classified as layers 3 and 4 (Table 6.2). Visualization mainly distinguishes between point and vector features: point features (layer 4) are defined as a center with a tolerance; this group encompasses hydrophobic, positively ionizable and negatively ionizable areas, in addition to excluded volume spheres. Hydrogen donors, acceptors and donor–acceptor pairs belong to the vector features group.

Point features are rendered as spheres with different colors to differentiate them (Fig. 6.5): hydrophobic/lipophilic features are drawn in yellow and positive and negative ionizable features are drawn in red and blue, respectively. Excluded volume spheres use a dark gray color to signify their meaning.

Spheres are drawn as semi-transparent objects, with a wire frame on their surface to enhance the impression of depth and to make it easier to judge the size of the spheres in the third dimension (Fig. 6.6).

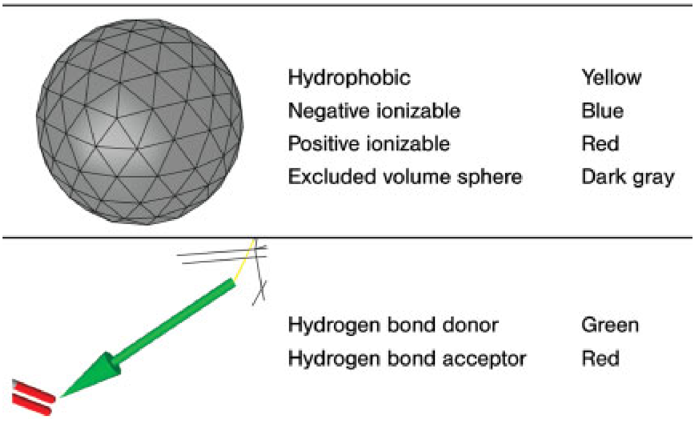


Figure 6.5 Different features in the pharmacophore visualization.

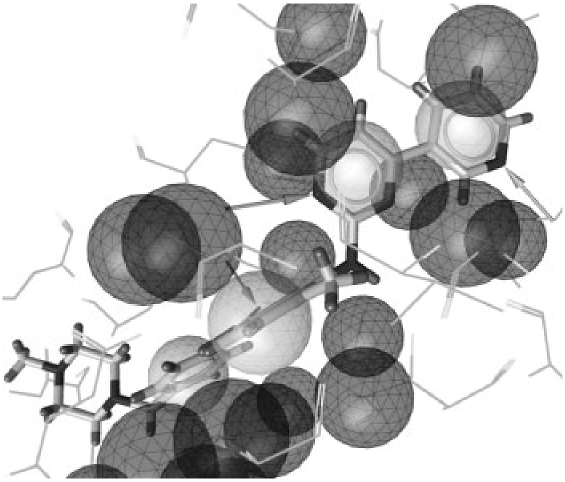


Fig. 6.6 Pharmacophore of complex 1opj, consisting of four lipophilic aromatic points (light gray spheres), three vectorized HBA (arrows) and excluded volume spheres (dark gray spheres).

Vector features are drawn as three-dimensional pointers or as a pair of two opposing pointers in the case of a donor–acceptor pair.

6.5.3 **Interaction**

Any visible object in the visualization may be selected for subsequent user interaction. The distinction between core and environment is useful to restrict the pickable objects to either group. This is especially important when selecting an entire region, because it is likely that unwanted objects in front of or behind the desired ones would otherwise be included in the selection.

Tool-tips provide additional information on atoms, bonds and chemical features and allow the user to measure angles and distances (Fig. 6.7). Tool-tips are semi-transparent, two-dimensional labels that are overlaid over the molecular image and point to the object for which they provide information.

When a new ligand is loaded, the render style of each atom or bond is determined by whether it is in the core or the environment. This can be changed for the complete ligand or environment at once or individually for each atom or bond.

The information about selected objects is also communicated to other views in LigandScout, such as a 2D depiction and a tree-viewer. In this way, the user can easily identify parts of the molecule in the different views and find additional information about them.

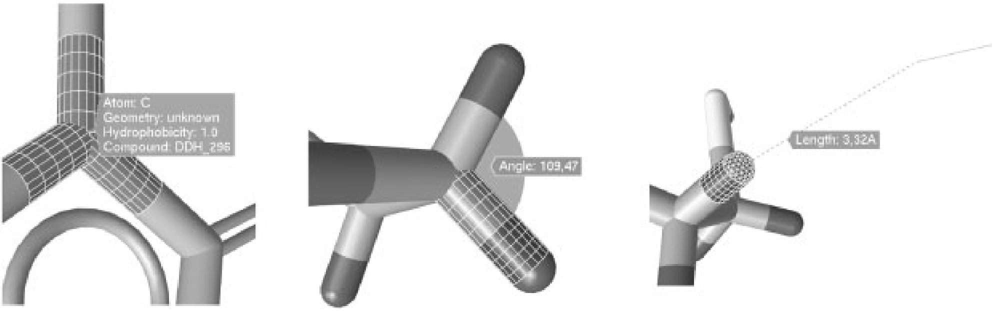


Fig. 6.7 Examples of different types of tool-tips for atoms, angles and distances.

6.6 Application Examples: Pharmacophore Generation and Screening

In order to assess the usability of pharmacophores with general chemical feature definitions for virtual screening experiments, we demonstrate two application examples that have already been published [33]. We chose Catalyst as the screening platform, because it allows one to incorporate the chemical feature definitions described in Table 6.2 directly into the program. These feature definitions contain five different types: lipophilic points (LIP), positive ionizable points (PI), negative ionizable points (NI), hydrogen bond donor vectors (HBD) and hydrogen bond acceptor vectors extended by electrostatic interactions occurring between fluorine atoms and hydrogen donors (HBA-F). Additionally, exclusion volume spheres were used as steric constraints. Pharmacophores were imported as hypotheses into Catalyst by interfacing the hypoedit tool from LigandScout.

The hypotheses were applied as screening filters to two different databases: a database consisting of all "drug-like" PDB ligands and the Maybridge 2003 database. For the "drug-like PDB ligands", a molecular weight constraint of minimum 250 and maximum 600 in combination with the "Lipinski rule of 5" (maximum 10 acceptors, maximum five donors, maximum log $P=5$) is applied [45, 46]. Owing to the lack of experimental log P values, the topological *c*log P estimation algorithm of Wildman and Crippen [47] is used as a filter. Although this rough kind of filtering may be considered problematic, the result should give an idea about the number of drug-like molecules in the PDB and the sort of enrichment that a pharmacophore is able to provide against a background of pharmaceutically relevant compounds. From the 6680 unfiltered PDB ligands with removed duplicates, 2765 conforming to the simple drug-likeness criteria remained. These were converted to Catalyst's multi-conformational format using the FAST method.

Additionally, the Maybridge compound library (Version 2003, containing 59 194 compounds) is converted into multi-conformational format and screened, analyzing the resulting hit lists for their accordance with the Lipinski drug-likeness criteria as applied to the PDB ligand database. A maximum mol weight of

Table 6.3 Screening results

Database	HRV coat protein subtype 16		ABL tyrosine kinase	
	Hits	W11-like hits	Hits	Gleevec-like hits
PDB ligands (2,765 entries)	1	1		2
2Maybrige 2003(59 194 entries)	48	0		7

600 is chosen for the two examples described below; it may need to be changed for pharmacophores targeting larger compounds.

6.6.1 HRV Coat Protein Inhibitor

Three PDB entries (1ncr, 1nd3 and 1c8m) contain pleconaril (PDB id: w11) bound to the protein hull of Rhinovirus subtype 16. From these entries, three pharmacophores were automatically created and overlaid. The final common feature pharmacophore consists of three lipophilic points, two aromatic lipophilic points, one H-bond acceptor and 22 excluded volume spheres, which characterize the protein environment of the lipophilic points. The results of the screening experiments are shown in Table 6.3. The search in the PDB ligand database yielded four hits: pleconaril (PDB id: w11), WIN61209 (PDB id: w01), WIN68934 (w02) and WIN65099 (w03), which are all reported to be Rhinovirus 16 hull protein binding agents [48]. Searching the Maybridge database yielded 67 hits; the drug-likeness criteria could further reduce the list to 47.

6.6.2 ABL Tyrosine Kinase Inhibitor

STI-571 (Gleevec) has been approved in the USA for the treatment of chronic myelogenous leukemia (CML) under the trade-name Gleevec. Crystal structures of a close analog of STI-571 revealed that STI-571 binds to the inactive form of ABL tyrosine kinase, stabilizes it and thus prevents activation [49]. This binding mode was targeted in the second application example of LigandScout. Three relevant PDB entries for this investigation were identified: 1fpu (Fig. 6.8), 1iep and 1opj. From these three entries, six pharmacophores were created from three complexes (all three records contained two different chains with a ligand each) and two different ligand molecules, STI-571 (PDB id: sti) and its variant *N*-(4-methyl-3-{[4-(3-pyridinyl)-2-pyridinyl]amino}phenyl)-3-pyridinecarboxamide (PDB id: prc) used in 1 fpu. In a straightforward approach, all six pharmacophore models were merged together into a single hypothesis using the clique detection algorithm together with Kabsch alignment described above.

The resulting pharmacophore model contains four lipophilic aromatic areas, two acceptors and eight excluded volume spheres. The same screening experi-

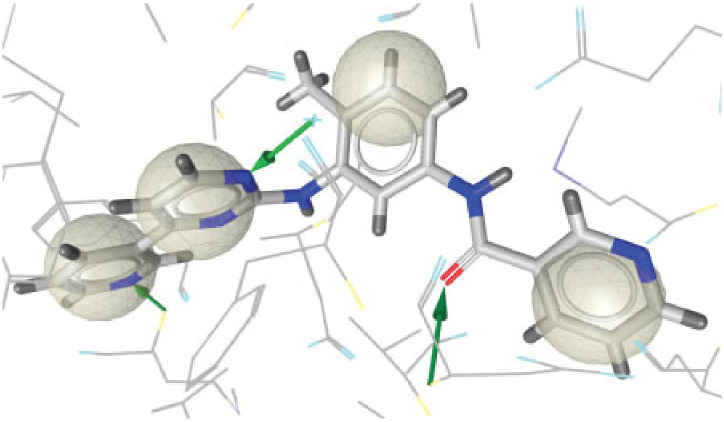


Fig. 6.8 Pharmacophore model automatically derived from PDB entry 1fpu consisting of four lipophilic aromatic points (light gray) and three hydrogen bond acceptors (green vectors).

ments as already carried out with the first example gave the results depicted in Table 6.3. The pharmacophore model is able to identify all Gleevec entries from the PDB database whereas from the Maybridge database, seven compounds were identified which might be potential lead structures for ABL tyrosine kinase inhibitors.

6.7 Conclusion

Techniques and paradigms used in computer-aided drug discovery have changed rapidly over the past few years. Docking is still the most commonly used method for structure-based drug design. Pharmacophore modeling, however, shows clear advantages regarding the computational cost for virtual screening and the understanding of the interaction between macromolecule and ligand. Ligand Scout promises to become a useful tool to make interaction information available as a transparent 3D model, which not only can be used for efficient virtual screening, but also provides means to understand intuitively the binding mode of a small-molecule ligand to a target. Overlaying for the generation of 'common feature pharmacophores' is one interesting application; another one might be to add information in the form of geometric constraints or weights to significantly improve enrichment. The need for transparency and the ability to modify automated results emphasize the need for high-quality visualization and the ability for interaction provided with LigandScout.

Acknowledgments

We thank our co-workers at Inte:Ligand, Martin Biely and Alois Dornhofer, for their work on LigandScout. We also thank Thierry Langer of the Institute of Pharmaceutical Chemistry of the University of Innsbruck and his computer-aided molecular design group for their support and for performing the screening experiments, especially Eva Krovat, Theodora Steindl, Oliver Funk, Daniela Schuster, Christian Laggner and Johannes Kirchmair.

References

1 T. Langer, R. Hoffmann, F. Bachmair, S. Begle. Chemical function based pharmacophore models as suitable filters for virtual screening. *J.Mol. Struct. (THEOCHEM) 503*, 59–72, **2000**.

2 W.P. Walters, M.T. Stahl, M.A. Murcko. Virtual screening – an overview. *Drug Discov. Today, 3*(4), 160–178, **1998**.

3 T. Langer, G. Wolber. Lead optimization: pharmacophore definition and 3d searches. *Drug Discov. Today, Technol., 1*(3), 203–207, **2004**.

4 *SYBYL 6.6/FlexX*. Tripos, St. Louis, MO.

5 M. Rarey, B. Kramer, T. Lengauer, G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol., 261*, 470–489, **1996**.

6 M. Rarey, B. Kramer, T. Lengauer, G. Klebe. Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des., 11*, 369–384, **1997**.

7 C.M. Oshiro, I.D. Kuntz. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Des., 9*, 113–130, **1995**.

8 G.M. Morris, DS. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson. Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem., 19*, 1639–1662, **1998**.

9 S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc., 106*, 765–784, **1984**.

10 G. Jones, P. Willett, R.C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol., 245*, 43–53, **1995**.

11 G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol., 267*, 727–748, **1997**.

12 H.-J. Böhm. LUDI: Rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des., 6*, 593–606, **1992**.

13 H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The protein data bank. *Nucleic Acids Res., 28*, 235–242, **2000**.

14 *Relibase: a Program for Searching Protein–Ligand Databases*. Cambridge Crystallographic Data Centre, Cambridge, **2000**.

15 M. Hendlich, F. Rippmann, G. Barnickel. BALI: automatic assignment of bond and atom types for protein ligands in the brookhaven protein databank. *J. Chem. Inf. Comput. Sci., 37*, 774–778, **1997**.

16 R. Sayle. *From Cruft to Content: Perception of Molecular Connectivity from 3D Coordinates*. Bioinformatics Group, Metaphorics LLC, Santa Fe, NM, **2001**.

17 *PDB File Format Contents Guide, Version 2.2*, **1996**, available from the PDB consortium, Piscatway, NS and La Jolla, CA (www.pdb.org)

18 J. Westbrook, P. Bourne. STAR/mmCIF: an extensive ontology for macromolecular structure and beyond bioinformatics. *Bioinformatics, 16*, 159–168, **2000**.

19 S.R. Hall, N. Spadaccini. The STAR file: detailed specification. *J. Chem. Inf. Comput. Sci.*, *34*, 505–508, **1994**.

20 H. Bernstein, F. Bernstein, P. Bourne. pdb2cif: translating PDB entries into mmCIF format. *J. Appl. Crystallogr.*, *31*, 282–295, **1998**.

21 J. Westbrook, Z. Feng, S. Jain, T.N. Bhat, N. Thanki, V. Ravichandran, G.L. Gilliland, W. Bluhm, H. Weissig, D.S. Greer, P.E. Bourne, H.M. Berman. The protein data bank: unifying the archive. *Nucleic Acids Res.*, *30*, 245–248, **2002**.

22 B. DuCharme. *XML, the Annotated Specification*. Prentice-Hall PTR, Upper Saddle River, NJ, **1999**.

23 T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler. *Extensible Markup Language (XML) 1.0*, 2nd ed., The w3c consortium, Cambridge, MA, USA, **2000**.

24 D.C. McArthur, *An Extensible XML Schema Definition for Automated Exchange of Protein Data: PROXIML (PROtein eXtensIble Markup Language)*. University of California, Santa Cruz, CA, **2001**, available from http://xml.coverpages.org/proximl.xml.

25 R.D. Brown, Y.C. Martin. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.*, *37*, 1–9, **1997**.

26 J. Gasteiger, C. Jochum. An algorithm for the perception of synthetically important rings. *J. Chem. Inf. Comput. Sci.*, *19*, 43–48, **1979**.

27 W.T. Wipke, T.M. Dyott. Use of ring assemblies in ring perception algorithm. *J. Chem. Inf. Comput. Sci.*, *15*, 140–147, **1975**.

28 J.B. Hendrickson, D.L. Grier, A.G. Toczko. Condensed structure identification and ring perception. *J. Chem. Inf. Comput. Sci.*, *24*, 195–203, **1984**.

29 R. Balducci, R.S. Pearlman. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.*, *34*, 822–831, **1994**.

30 B.T. Fan, A. Panaye, J.P. Doucet, A. Barbu. Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table. *J. Chem. Inf. Comput. Sci.*, *33*, 657–662, **1993**.

31 G.M. Downs, V.J. Gillet, J.D. Holliday, M.F. Lynch. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, *29*, 172-187, **1989**.

32 J. Figueras. Ring perception using breadth-first search. *J.Chem. Inf. Comput. Sci.*, *36*, 986–991, **1996**.

33 G. Wolber, T. Langer. Ligandscout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.*, *45*, 160 –169, **2005**.

34 G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler, D.A. Dunn. The conformational parameter in drug design: the active analog approach. In *Computer-Assisted Drug Design*, Vol. 112, E.C. Olson and R.E.Christoffersen (eds), American Chemical Society, Washington, DC, **1979**, pp. 205–226

35 J. Green, S. Kahn, H. Savoj, P. Sprague, S. Teig. Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.*, *34*, 1297–1308, **1994**.

36 E.M. Krovat, T. Langer. Non-peptide angiotensin II receptor antagonists: chemical feature based pharmacophore identification. *J. Med. Chem.*, *46*, 716–726, **2003**.

37 Smiles ARbitrary Target Specification (SMARTS). Daylight Chemical Information Systems, Mission Viejo, CA, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

38 L. Pauling. *The Nature of the Chemical Bond*, 2nd edn. Cornell University Press, Ithaca, NY, **1948**.

39 A. Khan. A liquid water model: density variation from supercooled to superheated states, prediction of H-bonds and temperature limits. *J. Phys. Chem.*, *104*, 11268–11274, **2000**.

40 *Catalyst, Version 4.10*. Accelrys, San Diego, CA, **2001**.

41 G. Wolber, T. Langer. Comb^{i}Gen: a novel software package for the rapid generation of virtual combinatorial libraries. In *Rational Approaches to Drug Design*, H.-D. Höltje and W. Sippl (eds), Prous Science, Barcelona, Spain, **2000**, pp. 390–399.

42 W. Kabsch. A solution for the best rotation to relate to sets of vectors. *Acta Crystallogr., Sect. A*, *32*, 922–923, **1976**.

43 A. Halm, L. Offen, D.W. Fellner. Visualization of complex molecular ribbon structures at interactive rates. In *8th International Conference on Information Visualization (IV)*, IEEE Computer Society Press, Los Alamitos, CA, **2004**, pp. 737–744.

44 E.B. Goldstein. *Sensation and Perception.* Wadsworth Publishing, Belmont, CA, **2001**.

45 C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.*, *23*, 3–25, **1997**.

46 C.A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods*, *44*, 235–249, **2000**.

47 S.A. Wildman, G.M. Crippen. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.*, *39*, 868–873, **1999**.

48 A.T. Hadfield, G.T. Diana, M.G Rossmann. Analysis of three structurally related antiviral compounds in complex with human rhinovirus 16. *Proc. Natl. Acad. Sci. USA*, *96*, 14730, **1999**.

49 T. Schindler, W. Bornmann, P. Pellicena, W.T. Miller, B. Clarkson, J. Kuriyan. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science, 289*, 1938, **2000**.

7
GRID-based Pharmacophore Models: Concept and Application Examples

Francesco Ortuso, Stefano Alcaro, and Thierry Langer

7.1
Introduction

The pharmacophore concept is a widely accepted and useful approach in both early drug discovery stages of hit determination and lead optimization [1]. A vast number of slightly different methods to build such models exist and some of them are discussed in this book. Usually pharmacophore models are created by collecting the most relevant structural features of biologically active compounds. When no three-dimensional structure information on the target exists most cases of the time, chemical intuition is necessary for completing the ligand-based pharmacophore generation, in ambiguous cases possibly leading to erroneous models. One of the most advanced applications of pharmacophore models is to use them as virtual screening filters of large compound database sets against a wide variety of multiple macromolecular targets [2]. This emerging technique, now considered as a new source of novel drug leads [3], is attracting more and more the attention of industrial pharmaceutical research.

Among the computational methodologies widely adopted in drug design studies, Goodford's GRID [4] program is very well accepted and trusted in the scientific community. It works by mapping the three-dimensional space around molecular targets with probes mimicking the main chemical properties of most common atom types and small moieties that are found in ligands. GRID data can be used to identify the best probe locations as map display and also 3D information for chemometric analysis [5–8]. The large availability of crystal and NMR structures of macromolecular complexes deposited in the Protein Data Bank (PDB) [9] is an excellent source for studying interactions between molecules of different nature (proteins, nucleic acids, small organic ligands).

In this chapter, we describe the results of our studies we aimed at the development of a general computational procedure to generate automatically and unbiased objective pharmacophore models using the GRID approach and starting with PDB macromolecular complexes. Within the context of structure-based pharmacophore modeling, it represents an approach that is somehow complementary to that described in Chapter 6. We have used logically combined maps

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

computed with the GRID force field in order to derive essential information on the interactions between occurring within the molecules of a PDB complexes for to generating chemical feature-based pharmacophore models. However, this approach can also be extended to cases where only one macromolecular complex partner is present, since the computation of GRID maps requires at least one binding partner. The versatility of the new computational approach has been tested benchmarked in several application examples using molecular complexes of different nature.

7.2 Theoretical Basis of the GBPM Method

The GRID-based pharmacophore model (GBPM) is created in a six-step procedure as depicted in Fig. 7.1.

The *first step* is dedicated to the PDB file pretreatment, which often contains water molecules and no hydrogen atoms. In the pretreatment, the user should fix typical problems such as missing residues, missing side-chains and wrong bond orders, especially for bound organic compounds. The GREAT and GRIN modules of the GRID software help contribute to this task and allow the preparation of the GRID mapping procedure. Assuming that the complex has two interacting molecules α and β, as in the case of protein–protein or protein–ligand complexes, the main goal of this first step is to obtain three interaction energy maps with from the $\alpha+\beta$, α and β subunits, keeping the atomic coordinates of the original PDB model (Fig. 7.1).

The *second step* performs the GRID calculation with a given probe on the three subunit models. In order to make the application of Boolean operations with the map files as easy as possible, the matrix dimension of the GRID box is exactly maintained as in the largest model, i.e. that with $\alpha+\beta$ subunits, maintaining, for both subunits, the original complex atom coordinates. The three maps obtained are named **A**, **B** and **C**, respectively (Fig. 7.1).

The *third step* is based on the GRAB procedure, implemented in GRID v. 21, performing a Boolean operation [10] between the maps **B** and **A**. The resulting map **D** has, by definition, the same matrix dimension of the original maps and reports, with negative energy values, the α–β interaction areas. According to the GRAB algorithm [10], the α components are converted into positive or zero values comparing maps **D** and **C**. The resulting map **E** reports the acceptance degree of a certain probe into the α–β binding site. Such an indication represents a first, interesting, advantage of the GBPM method, since actually no indication has been given in order to identify the right positioning and the extension of map **E**. Definitely, each point of the α–β interaction area is automatically defined with unbiased influence of the user.

The *fourth step* is dedicated to the identification of the most important interaction areas of map **E**. This task is carried out using the MINIM utility included into the GRID program. This program collects all points within a certain energy threshold, allowing the interpolation of the closest ones. The choice of an energy threshold value is a biased task *per se* but, considering a pharmacophore model as a minimum interaction descriptor built by few features, we have generally found an energy threshold about 10% higher than the global minimum value to be appropriate. This means, in most cases, about 1 kcal mol^{-1} above the global minimum energy determined. Actually, such a value allows at least one feature to be collected for each probe used. Often the above energy threshold yields too complicated pharmacophore models that can be reduced using the GRID energy as a cutting criterion.

In order to design a suitable pharmacophore model, all reported operations should be repeated using at least three different probes: the hydrophophic probe (DRY), a hydrogen bond acceptor (O) and a hydrogen bond donor (N1). This choice allows a basic characterization of most of the interaction areas; however, more sophisticated and selective models can be obtained by adding other GRID probes such as halogen or charged atoms. In the *fifth step*, the information originating from the different probe experiments are simply merged into a preliminary pharmacophore model (multiple probe features of Fig. 7.1).

The *sixth step* is dedicated to the validation of this the preliminary model and eventually its modulation in terms of number of features (i.e. its complexity). The quality of the pharmacophore model is tested as the capability to recognize selectively the original ligand present in the PDB file. Technically the evaluation step can be carried out by the Catalyst software [11], in particular using the CiTest fit module [12]. The preliminary model is imported converting the GBPM points into Catalyst features. The GRID energies are also included in the fit analysis as feature weight according to the following equation:

$$wF_{ij} = EF_i | AEF_j \tag{1}$$

where wF_{ij} is the weight for the feature i into the hypothesis j, EF_i is the GRID energy for the features i and AEF_j is the average GRID energy value for the hypothesis j. This approach allows a maximum fit value (MFV) equal to the total number of features available for the hypothesis j. Taking into account the GRID energies, several preliminary models (hypotheses) can be designed reducing the number of features. Unfortunately, owing to the high variability, i.e. extension and interaction type, of the α–β subunit interface, the number of preliminary models can not be predefined. Therefore, in order to identify the best one, all possible models are submitted to a CiTest fit.

A fit index (FI), defined as ratio between the CiTest fit and MFV, is used for the evaluation of each hypothesis and as a choice criterion for the identification of the best GBPM.

Moreover, the FI descriptor, which makes possible comparisons among models with different numbers of features, can be used to extend the evaluation step including other molecules known to interact with the same β subunit binding site. Such an eventuality was found to improve strongly the quality of the final model.

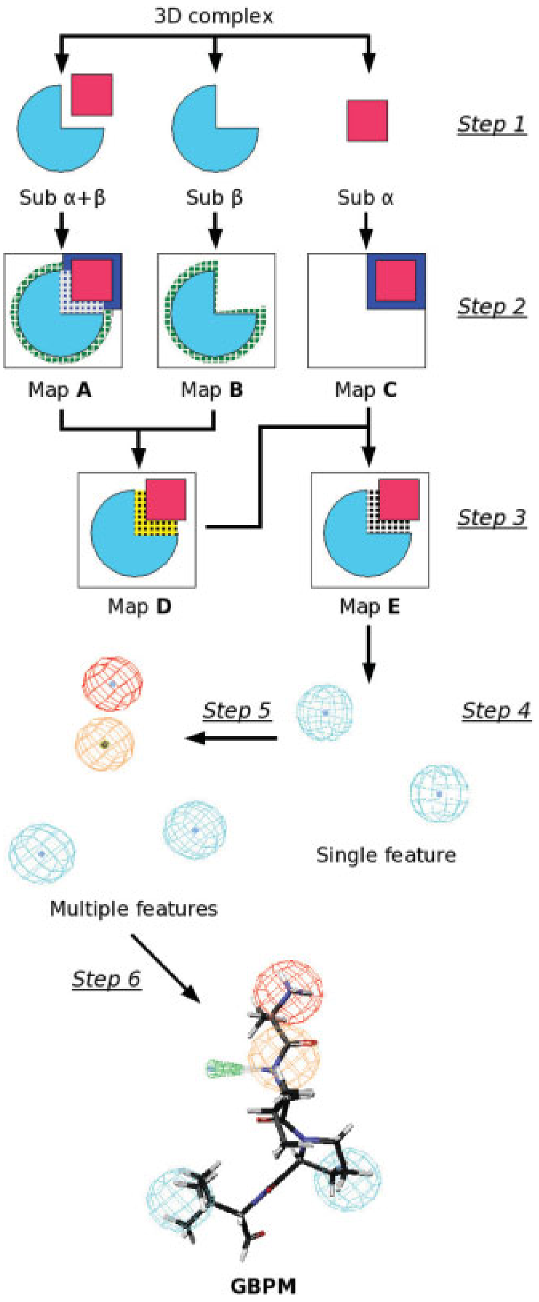


Fig. 7.1 Flow chart of the GBPM starting from a PDB complex. The bottom figure represents a generic feature-based pharmacophore model.

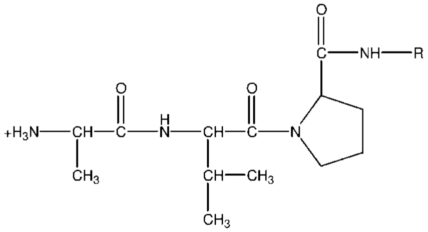
7.3 Application Examples

In this section, we describe the application of the GBPM method to different selected complexes. The resulting pharmacophore models have been tested extensively for their capacity to retrieve known ligands for the target. The application examples were selected to be representative of a certain type of molecular interaction, including protein–protein, and DNA–ligand interactions. More application examples will be described elsewhere [13].

7.3.1 Protein–Protein Interaction: XIAP

This kind of interaction can be considered as the most challenging among our application examples because it is intrinsically characterized by the lowest degree of information [14]. We demonstrate the application of the GBPM procedure to a member of a family of proteins involved in the regulation of apoptosis, the X-linked inhibitor of apoptosis (XIAP). Its third baculovirus IAP repeat domain (BIR3) recognizes compounds **1–5** as shown in Fig. 7.2.

The structure of **1** has been determined by NMR spectroscopy and is deposited in PDB entry 1G3F [15]. Compound **2** conformations were isolated from 1XB0 and 1XB1 models [16]. Five conformations of **3** were considered from the 1XB0 model. Another conformation of this peptide was extracted from the chi-



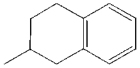
Compound		PDB data		
#	**R**	**Conformations**	**entry ID**	**Source**
1	-Ile-Ala-Gln-Lys-Ser-Glu	1	1G3F	NMR
2	-Ile-Ala	6	1XB1	X-RAY (2.70 Å)
		1	1XB0	X-RAY (2.20 Å)
3	-Ile	5		
		1	1TW6	X-RAY (1.71 Å)
4	-Ile-Ala-Gln	1		
5		1	1TFQ	NMR

Fig. 7.2 Chemical structures of compounds **1**–**5** interacting with BIR3 domain of XIAP and their PDB record.

mera 1TW6 structure [17]. From the same model, **4** was also obtained. The conformation of the peptidomimetic **5** was obtained from the 1TFQ model [18]. Finally, the Smac protein **6**, complexed with XIAP, was considered using the PDB 1G73 crystallographic model (resolution 2.00 Å). Since **6** is significantly larger than **1–5**, the recognition with XIAP is not exclusive to the BIR3 domain, but additionally involves other regions. A pharmacophore model able to describe entirely this kind of recognition is technically feasible but useless for the virtual screening of 3D databases, because it is unusual to search for compounds with these structural prerequisites of low drug likeness. Hence the GBPM was derived using the original PDB 1G3F complex in which is reported the recognition of **1**, a relatively small synthetic peptide, and the XIAP BIR3 domain. Moreover, **1** is the largest ligand among **1–5**, allowing a more exhaustive description of the interaction area.

The recognition area with caspase-9 is fairly extended (about 700 Å^2) [19], and several hydrophobic, electrostatic and hydrogen bond interactions are involved, so a simple receptor based pharmacophore model resulted relatively hard to derive. Moreover, the very small number of molecules known to recognize the XIAP BIR3 domain does not allow a rigorous classical ligand-based approach. For these reasons, GBPM represents a useful tool for the XIAP case study. Computational work followed the flow chart reported in Fig. 7.1. After the pretreatment step 1, **1** was considered as the α and the XIAP as the β subunit. The PDB 1G3F complex was used to compute GRID molecular interaction fields with O, N1 and DRY probes (maps **A**). These procedures, using the same complex box dimensions, were repeated separately on the α (maps **B**) and the β (maps **C**) subunits, maintaining, in both cases, the respective complex atom coordinates (step 2). Maps **A** and **B** were compared by the GRAB algorithm obtaining the maps **D** that were used, together with maps **C**, to obtain maps **E** (step 3). The three maps **E** were submitted to MINIM, selecting, after interpolation, those points with an interaction energy within the first kcal mol^{-1} with respect to the global minimum. This approach allowed us to obtain four features with the N1 probe, three with DRY and only one with O. The preliminary model (HYP1) was converted into the a Catalyst pharmacophore model using for each probe the most corresponding feature: for N1 the hydrogen bond donor feature (HBD) was used, for DRY the generic hydrophobic (HPB) and for O the hydrogen bond acceptor (HBA) were used as features. The weight of each feature was scaled taking into account the GRID interaction energies using Eq. (1) (Table 7.1).

The resulting HYP1 model was tested, with both rigid and flexible CiTest algorithms, using ligand **1**. Since the resulting FI index, equal to 0.02, revealed only a poor recognition of **1**, HYP1 was simplified by removing the less relevant hydrogen bond donor HDB3 and HDB4 features and rescaling the weight of the remaining ones as reported for HYP1. The evaluation of the new seven-feature model, HYP2, indicated a better fit with **1** (FI index increased to 0.57). The simplification of the pharmacophore model proceeded with the elimination of the two less relevant hydrophobic features (HPB3 and HPB4) adjusting the

Table 7.1 Preliminary 1G3F GBPM

GRID probe	Catalyst feature	GRID IE[a)]	CiTest weight
N1	HBD1	–7.31	1.53
N1	HBD2	–7.06	1.48
N1	HBD3	–6.49	1.36
N1	HBD4	–6.48	1.36
O	HBA1	–10.37	2.17
DRY	HPB1	–1.96	0.41
DRY	HPB2	–1.15	0.24
DRY	HPB3	–1.07	0.22
DRY	HPB4	–1.05	0.22

a) GRID interaction energy in kcal mol^{-1}.

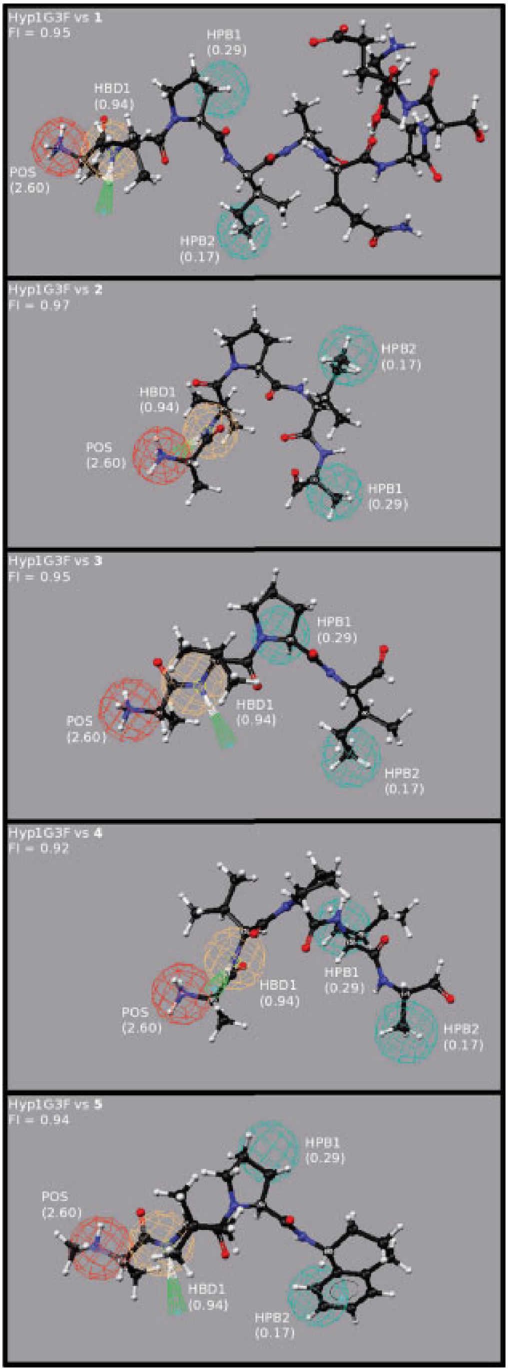
Table 7.2 HYP5 model components

GRID probe	Catalyst feature	GRID IE[a)]	CiTest weight
N^+	POS	–17.88	2.33
N1	HBD2	–7.06	0.98
O	HBA1	–10.37	1.37
DRY	HPB1	–1.96	0.15
DRY	HPB2	–1.15	0.26

a) GRID interaction energy in kcal mol^{-1}.

weight of the remaining ones. The new five-feature model, HYP3, notably increased the FI index to 0.88. In order to make a comparison between HDB and HPB components, we designed another model, HYP4, including in HYP3 all the HDB features. After rescaling of the HYP4 feature weights, an FI value of 0.49 was reached. The comparison of the FI values revealed a similar behavior of HBDs and HPBs, indicating HYP4 to be the most promising model. With the aim of improving the selective recognition capabilities of HYP4 and taking into account the presence of a positively charged N-terminus on **1–5**, we introduced the positive ionizable feature POS. This task was carried out including in the GBPM flow chart (Fig. 7.1) the GRID probe N^+. The single feature map revealed for the N^+ only one point within the first kcal mol^{-1} with respect to its global minimum. Its interaction energy, equal to –17.88 kcal mol^{-1}, was the most relevant with respect to all other probes. Interestingly, the location of this point was coincident with that of the HDB1 which showed an interaction energy of –7.31 kcal mol^{-1}. Consequently, we built a new model, HYP5, substituting HDB1 with POS and rescaling the feature weights (Table 7.2).

Surprisingly, the CiTest fit of ligand **1** with HYP5 indicated a low FI value of about 0.4. On investigating the reason for this, a different recognition pattern of **1** was found. Actually, the POS feature was exactly located on the positively charged N-terminus whereas the previous HBD1 showed a positioning far from



Hyp1G3F vs 1
FI = 0.95

Hyp1G3F vs 2
FI = 0.97

Hyp1G3F vs 3
FI = 0.95

Hyp1G3F vs 4
FI = 0.92

Hyp1G3F vs 5
FI = 0.94

Fig. 7.3 Hyp1G3F model recognition of compounds **1**–**5**. Feature weights are reported in parentheses.

such a moiety. The roto-translation introduced by the new feature, possessing a higher weight, does not allow the HBA1 interception of the hydrogen bond acceptors located on **1**. The CiTest algorithm, following the HBA1 weight, tried to superpose it on the compound oxygen atoms of **1**, leading to a lack of recognition of all other, less relevant, features. According to the interaction energies, the competition between POS and HBA1 was resolved on removing this last feature. The new model, Hyp1G3F, after the usual rescaling of the component weights, was submitted to the evaluation task and found to be superior to the previous one, with an FI value of 0.95.

Hyp1G3F was then evaluated with respect to **2–5**. With all molecules, taking into account all their experimentally determined conformations, FI values higher than 0.90 were reached, confirming the high degree of recognition of known molecules interacting with the XIAP BIR3 domain of Hyp1G3F (Fig. 7.3).

Thus the GBPM approach was successfully applied to the XIAP application example, clearly indicating the most relevant features for the BIR3 domain interaction. Such information was derived starting from only one model and was able to recognize other XIAP BIR3 domain ligands.

7.3.2 Protein–Protein Interaction: the Interleukin 8 Dimer

In the second application example for evaluating the GBPM procedure in protein–protein interactions, the interleukin 8 (IL8) dimer was analyzed. IL8 plays a relevant role in immune cell trafficking and in host defense against infection. It is known that IL8 can exist in both dimeric and monomeric forms and only the latter is able to interact productively with the CXCR receptors [20]. Therefore, the equilibrium between the dimeric and monomeric forms can be considered as an interesting target for modulating IL8 activity. In the present application example, we designed a GBPM pharmacophore model for the IL8 dimer interface which could be useful for discovering molecules modulating the equilibrium between active and inactive form of IL8.

Our case study was carried out using the NMR-derived PDB model 1IL8, which represents an average structure of the IL8 homodimer in solution [24]. To apply the GBPM approach, taking into account the general scheme reported in Fig. 7.1, the 1IL8 chain A was considered as β subunit and the chain B as the α subunit (Fig. 7.4).

The GRID probes O, N1 and DRY were used in the 1IL8 case study. The preliminary pharmacophore model (HYP1) for the IL8 dimer interface was designed selecting for each probe map **E** all points with an interaction energy within 1 kcal mol^{-1} above the global minimum. As shown in Fig. 7.5, such a model was built with 20 hydrophobic features (HPB), six hydrogen bond donors (HBD) and five hydrogen bond acceptors (HBA).

As observed in all our GBPM applications, the first 'raw' model, due to the large number of its features, was often not usable as a pharmacophore model for screening. As shown in Fig. 7.6, our approach was able to recognize cor-

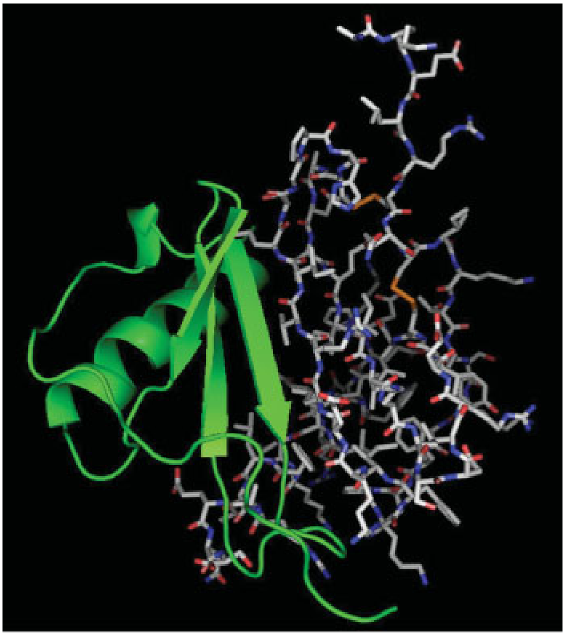


Fig. 7.4 1IL8 PDB structure. Chain A is represented in green cartoons and chain B in stick CPK notation.

Table 7.3 Hyp1IL8 components

GRID probe	Catalyst feature	GRID IE[a)]	CiTest weight
N1	HBD1	–7.01	1.33
N1	HBD2	–6.90	1.31
O	HBA1	–6.98	1.32
O	HBA2	–6.27	1.19
O	HBA3	–6.18	1.17
DRY	HPB1	–1.83	0.35
DRY	HPB2	–1.77	0.34

a) GRID interaction energy in kcal mol^{-1}.

rectly the residues responsible for dimer formation, notably those reported by Clore et al. [21]. Moreover, GBPM identified additional amino acids located at the dimer interface that could contribute to modulation of the equilibrium between the active and inactive forms of the IL8. These observation allowed us to assess positively the application of GBPM to the IL8 case study. In order to design a more suitable pharmacophore model, we reduced the total number of features of HYP1, removing the points attributed to a lower contribution to the overall interaction energy. The best of the models obtained by HYP1 simplifica-

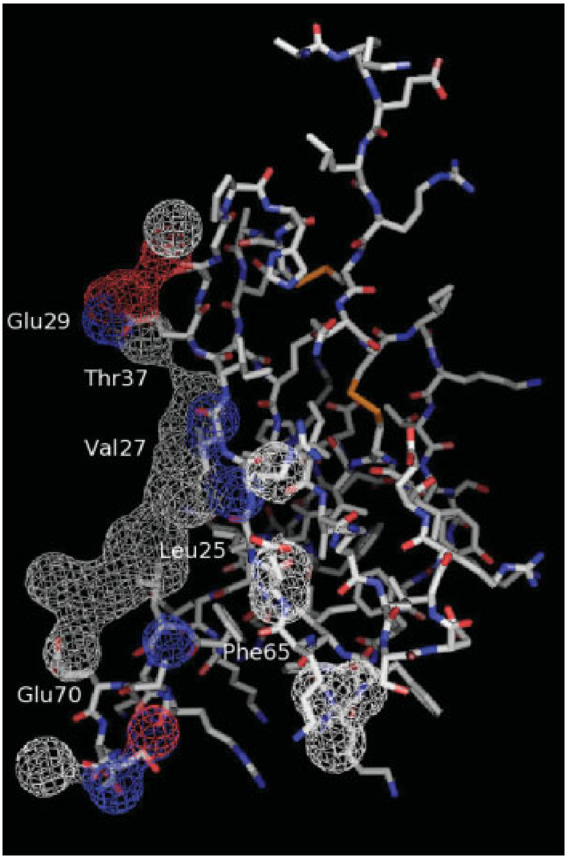


Fig. 7.5 1IL8 GBPM preliminary model. Grey meshes represent hydrophobic features, hydrogen bond donors are reported in blue and hydrogen bond acceptors in red.

tion, denoted Hyp1IL8, showed a fit index equal to 0.80 and a good recognition of the dimerization region. The Hyp1IL8 composition is reported in Table 7.3.

In Fig. 7.6, the interleukin 8 dimerization interface recognition of Hyp1IL8 is shown.

Although no ligand validation was possible in the present case study, we considered IL8 to be a good application for GBPM. Actually, no information was available about the dimer interface and our method designed a pharmacophore model able to recognize the original ligand and was therefore considered useful for virtual screening purposes.

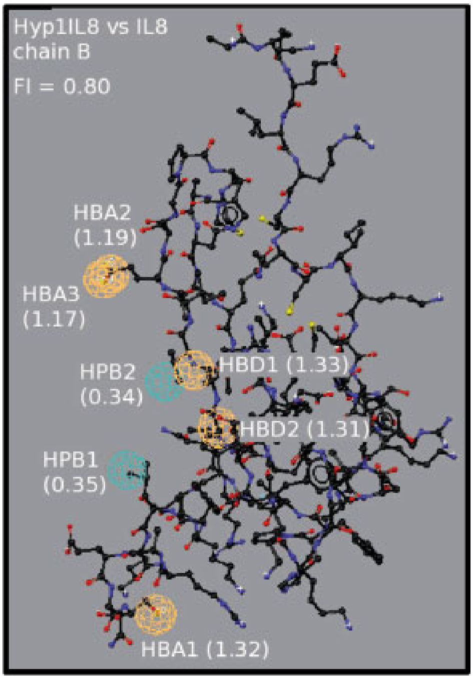


Fig. 7.6 Hyp1IL8 chain B recognition. Feature weights are reported in parentheses, hydrogen atoms have been hidden for clarity.

7.3.3
DNA–Ligand Interaction

The activity of several anticancer and antiviral drugs is due to their affinity to DNA. Many of these compounds interact with the biological target by binding to either the minor and/or the major groove. Virtual screening of 3D molecular databases with pharmacophore models derived from structural information related to this phenomenon can help to identify new molecules with exhibiting such a mechanism of action. These reasons stimulated us to evaluate the GBPM approach also for designing pharmacophore models starting from DN–ligand complexes.

The wide number of structures available in the PDB focused our attention on DNA complexes with minor groove binding compounds and therefore 18 structures solved by crystallographic methods were selected from the PDB [22]. These models reported minor groove complexes to the d(CGCGAATTCGCG)$_2$ dodecamer sequence with different binders. Taking into account their common chemical scaffold, the ligands were classified into five subclasses, A, B, C, D and E. In Table 7.4 the common scaffold-based classification, binder chemical structures, PDB codes and their crystallographic resolutions are reported.

Interestingly, all ligands reported in Table 7.4 revealed the same binding mode to DNA. In detail, **6–18** bind to the dodecamer's minor groove from the fourth pair of nucleotides until the ninth pair. This feature permits the investi-

Table 7.4 Chemical scaffold-based classification, ligand structures and PDB records of models used to test the GBPM approach to DNA–ligand case study

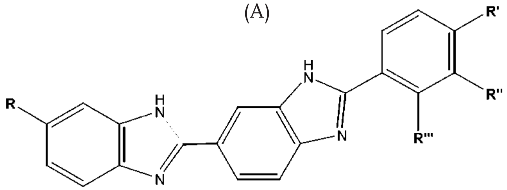
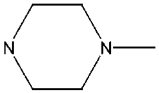
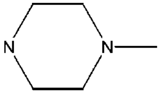
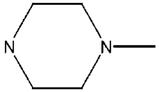
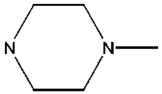
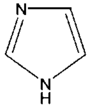
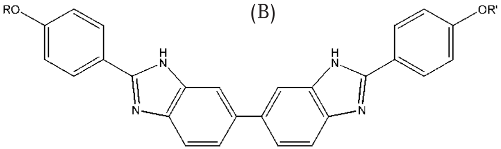
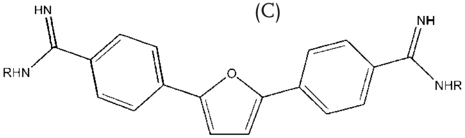
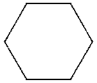
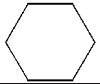
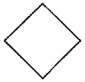
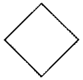
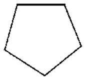
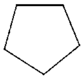
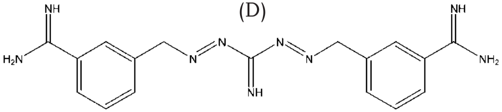
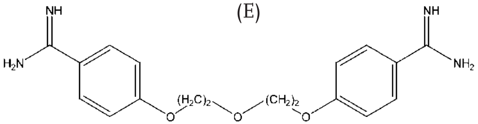
Common chemical scaffold classification	R	R′	R″	R‴	X[a)]	Y[a)]	Z[a)]
(A)  		OCH_3	I	H	**6**	442D	1.60
						448D	2.20
						444D	2.40
		H	I	H	**7**	443D	1.60
						449D	2.10
						445D	2.60
		$N(CH_3)_2$	H	H	**8**	447D	2.20
						1QV8	2.50
		$N(CH_3)_2$	H	CH_3	**9**	1QV4	2.50
		OH	H	H	**10**	109D	2.00
	H	H	–	–	**11**	453D	1.80

Table 7.4 (Continued)

Common chemical scaffold classification	R	R′	R″	R‴	X a)	Y a)	Z a)
(B) 	H	H	–	–	**11**	453D	1.80
	$(CH_2)_3N(CH_3)_2$	$(CH_2)_3N(CH_3)_2$	–	–	**12**	1FTD	2.00
(C) 	CH_2CH_3	CH_2CH_3	–	–	**13**	360D	1.85
			–	–	**14**	1FMS	1.90
			–	–	**15**	1FMQ	2.00
			–	–	**16**	1EEL	2.40
(D) 	–	–	–	–	**17**	1M6F	1.78
(E) 	–	–	–	–	**18**	166D	2.20

aX=compound reference; Y=PDB code; Z=resolution in Å.

Table 7.5 Best fitting models feature composition and FI values.

Hypothesis	PDB model	Compound	HBA [a)]	HBD [a)]	HPB [a)]	FI
Hyp442D	442D	**6**	1	1	3	0.75
Hyp453D	453D	**11**	1	3	1	0.52
Hyp360D	360D	**13**	1	2	3	0.85
Hyp1M6F	1M6F	**17**	1	2	2	0.55
Hyp166D	166D	**18**	1	1	4	0.60

a) HBA, HBD and HPB indicate hydrogen bond acceptor, donor and hydrophobic features number, respectively.

gation of two principal aspects of the GBPM approach: (i) the recognition capability with respect to molecules different from the original structure and (ii) the relationship of our pharmacophore hypotheses with the complex subunit *α*. These tasks were performed by selecting for each common scaffold subclass the PDB model with the best crystallographic resolution. Entries 442D, 453D, 360D, 1M6F and 166D were thus considered for the subclasses A, B, C, D and E, respectively, and used for building five independent pharmacophore models. These feature-based hypotheses were evaluated taking into account their FI values computed with **6**, **11**, **13**, **17** and **18**, respectively. The best fitting models were tested on the remaining ligand set.

The GRID probes C1=, N1 and O were used for building the GBPM hypotheses. We substituted DRY with C1=, which mimic the sp^2 carbon atom, because all ligands possess aromatic moieties. Moreover, the DNA minor groove, owing to the large number of polar groups, exhibits hydrophilic properties and therefore the DRY probe, which identifies those areas where water molecules are not well accepted, could indicate underestimated information. As reported in the previous examples, the first preliminary models obtained, owing to their intrinsic complexity, were not useful for virtual screening purposes. The number of features was sequentially reduced following the procedure indicated in the previous case studies. In Table 7.5 we report the best fitting model compositions, together with their validating ligand and FI.

As reported in Table 7.5, our hypotheses show different feature compositions. It was not surprising to observe that in all cases only one hydrogen bond acceptor was detected. Actually, the binding sites of **6–18**, reported in crystallographic models, show only two hydrogen bond donors represented by two guanine nucleobases located in position 9 that undergo only weak interactions with the ligands. Conversely, as indicated by the larger number of HBD, several hydrogen bond acceptors are located within the DNA minor groove, interacting strongly with acidic hydrogen atoms of **6–18**. Hydrophobic features, derived by the C1=GRID probe, were observed in all models. Only model 453D reported one of this kind of feature. Actually also in this case a second hydrophobic feature was found within the first preliminary model, but was removed because it was superposed on a more important hydrogen bond donor. In Fig. 7.7a representa-

Table 7.6 Best fitting hypotheses FI values computed on compounds **6**–**18**

Compound	PDB entry	Hypothesis				
		Hyp442D	Hyp453D	Hyp360D	Hyp1M6F	Hyp166D
6	442D	0.75	0.44	0.52	0.59	0.61
	444D	0.71	0.44	0.51	0.59	0.60
	448D	0.75	0.44	0.52	0.60	0.59
7	443D	0.68	0.43	0.52	0.52	0.59
	449D	0.67	0.43	0.52	0.52	0.59
	445D	0.65	0.43	0.52	0.52	0.59
8	447D	0.68	0.43	0.52	0.60	0.60
	1QV8	0.67	0.43	0.52	0.60	0.60
9	1QV4	0.66	0.43	0.52	0.59	0.59
10	109D	0.69	0.68	0.91	0.70	0.47
11	453D	0.64	0.52	0.52	0.45	0.60
12	1FTD	0.76	0.44	0.49	0.59	0.58
13	360D	0.80	0.68	0.85	0.61	0.60
14	1FMS	0.65	0.65	0.75	0.59	0.61
15	1FMQ	0.78	0.67	0.84	0.60	0.60
16	1EEL	0.79	0.69	0.85	0.63	0.60
17	1M6F	0.54	0.55	0.67	0.56	0.61
18	166D	0.52	0.50	0.50	0.53	0.60

tion of the pharmacophore model recognition with respect to the generating ligand is reported, including, for each hypothesis, the feature CiTest weights.

After the preliminary validation, carried out on the generating ligands, the best fitting hypotheses were evaluated with respect to **6**–**18**. For binders **6**–**8**, all crystallographic conformation was taken into account. CiTest results are reported in Table 7.6.

The ligand DNA recognition can be considered an extremely difficult case study for all pharmacophore model design methods. Actually, nucleic acid complexes, in particular with minor groove binders, are characterized by interaction between a highly hydrophilic subunit, the DNA, and a low hydrophilic or often hydrophobic subunit which is the binder. Water molecules have to be displaced from the nucleic acid, increasing the entropic contribution of binding. In order to build the complex, the total energy of the system has to be reduced by new interaction between the DNA and the binder. Such interactions can be addressed to hydrogen bonding and electrostatic terms. The first ones have to be formed by the interacting agent because the DNA minor groove possesses much more acceptor groups than donors. Also, electrostatic moieties, located on the ligand, cannot be considered with a net negative charge because this could generate repulsion forces with respect to the target structure. Positively charged agents could therefore interact with nucleic acid phosphate moieties and not with the minor groove. These aspects have clearly been indicated by the GRID analyses performed within the present case study. Actually, we changed the

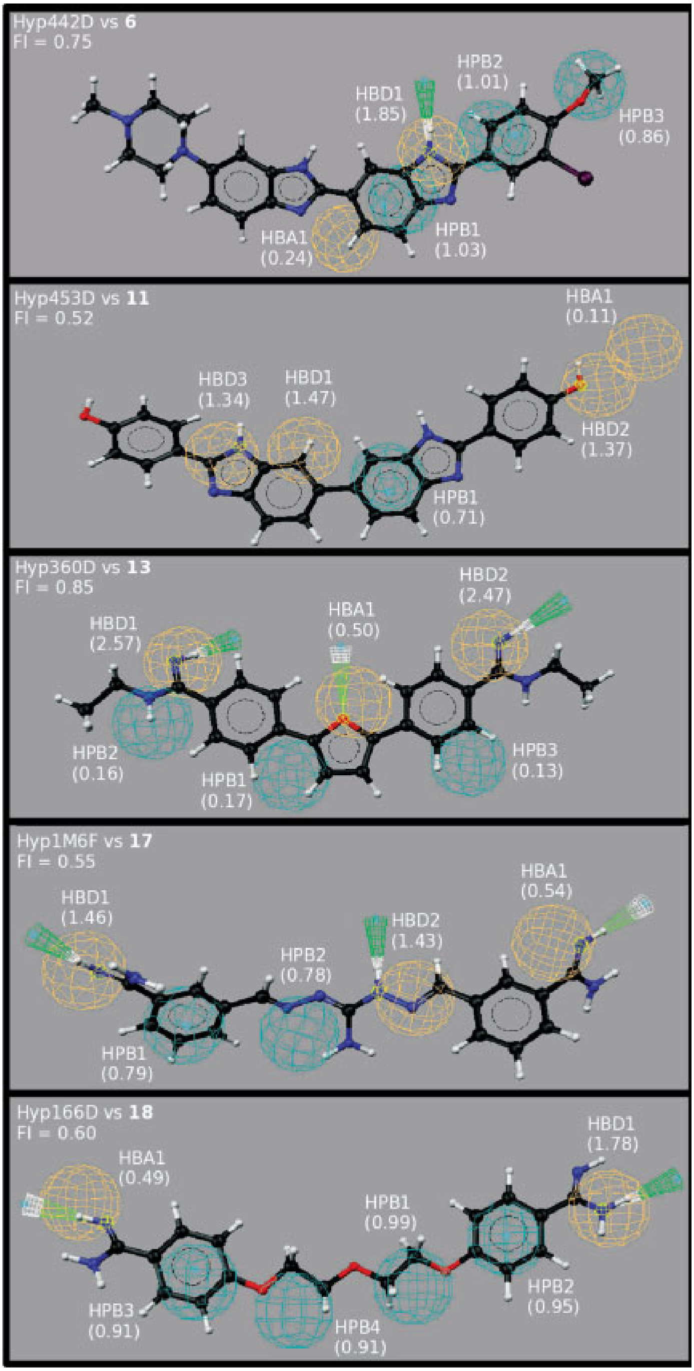


Fig. 7.7 Fit of our best pharmacophore models with respect to their generating subunit α. Feature weights are reported in parentheses.

DRY probe to C1=because the former indicated a very poor interaction with the target. The hydrogen bond network has been widely investigated by O and N1 GRID probes and GBPM highlighted the poor presence of hydrogen bond donor groups within the DNA minor groove, reporting only one HBA feature in all pharmacophore models. Conversely, HBD features were strongly suggested by the results of the analysis. The application of the GBPM approach to the present example indicates that it represents a useful computational tool to design pharmacophore models also for DNA binders. In particular, both different and common interactions of chemical scaffolds **A**–**E** have been highlighted, giving interesting information not only for virtual screening purpose but also for rational optimization of known ligands. GPBM revealed a low degree of dependence on the building complex subunit *a*, recognizing DNA binders showing a different structure with respect to the builder subunit.

7.4 Conclusions

The GBPM was developed with the aim of defining a general computational protocol for creating unbiased pharmacophore models starting from well-referenced experimental complex models such as those deposited in the PDB. The preliminary validation of the methodology in diverse molecular complexes (protein–protein, DNA–ligand, enzyme–inhibitor) will be extended further to other examples, but the results obtained so far already indicate the great versatility of the GBMP approach. All application examples described in this chapter reveal the good capabilities of this approach to identify the most relevant host–guest interaction occurring within the analyzed complexes, indepently of their nature. GBPM revealed no impediments for its application to both macromolecular and small organic compound targets. Owing to the large number of GRID probes available, a notable improvement of the generated pharmacophore models can be achieved, suggesting also new substituents for known compounds or contributing to rationalizing their structure–activity relationships. Consequently, the GBPM approach may be useful not only in the lead identification process but also in the lead optimization phase.

References

1 For general pharmacophore model applications, see: Milne G.W.A. Pharmacophore and drug discovery. In *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R. (ed.). Wiley, New York, **1998**, Vol. 3, pp. 2046–2056.

2 Langer, T., Krovat, E.M. Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. *Curr. Opin. Drug Discov. Dev.* **2003**, *3*, 370–376.

3 Alvarez, J.C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *4*, 365–370.

4 Goodford, P.J. A computational procedure for determining energetically fa-

vourable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

5 Pastor, M., Cruciani, G., Watson, K.A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure–activity relationship analysis. *J. Med. Chem.* **1997**, *40*, 4089–4102.

6 Kastenholz, M.A., Pastor, M., Cruciani, G., Haaksma, E.E.J., Fox, T. GRID/CPCA: a new computational tool to design selective ligands.*J. Med. Chem.* **2000**, *43*, 3033–3044.

7 Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.

8 Crivori, P., Cruciani, G., Carrupt, P.A., Testa, B. Predicting blood–brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204–2216.

9 http://www.rcsb.org/pdb; The Protein Databank.

10 The GRAB procedure energetically compares two generic **A** and **B** GRID maps adopting for each node of them the following algorithm: if **A**>0 and **B**>0 then **D**=0, if **A**>0 and **B**<0 then **D**=–**B**, if **A**<0 and **B**>0 then **D**=**A**, if **A**<0 and **B**<0 then **D**=(**A**–**B**).

11 Accelrys, *Catalyst, Version 4.9*. Accelrys, San Diego, CA, 2003; http://www.accelrys.com.

12 The CiTest performs the evaluation of the pharmacophore hypothesis computing a non energy-weighted fit value.

13 Ortuso, F., Langer, T., Alcaro, S. GBPM: GRID based pharmacophore model. Concept and application studies to protein-protein recognition. *Bioinformatics*, Advanced Access March 27, 2006, doi: 10.1093/bioinformatics/btl115.

14 In these cases the fit values were computed considering only the amino acids of subunits *α* and *β* interacting at the protein interface.

15 Liu, Z., Sun C., Olejniczak, E.T., Meadows, R.P., Betz, S.F., Oost, T., Herrmann, J., Wu, J.C., Fesik, S.W. Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. *Nature* **2000**, *408*, 1004–1008.

16 Shin, H., Renatus, M., Eckelman, B.P., Nunes, V.A., Sampaio, C.A., Salvesen, G.S. The BIR domain of IAP-like protein 2 is conformationally unstable: implications for caspase inhibition. *Biochem. J.* **2005**, *385*, 1–10.

17 Vucic, D., Franklin, M.C., Wallweber, H.J., Das, K., Eckelman, B.P., Shin, H., Elliott, L.O., Kadkhodayan, S., Deshayes, K., Salvesen, G.S., Fairbrother, W.J. Engineering ML-IAP to produce an extraordinarily potent caspase 9 inhibitor: implications for Smac-dependent anti-apoptotic activity of ML-IAP. *Biochem. J.* **2005**, *385*, 11–20.

18 Oost, T.K., Sun, C., Armstrong, R.C., Al-Assaad, A.S., Betz, S.F., Deckwerth, T.L., Ding, H., Elmore, S.W., Meadows, R.P., Olejniczak, E.T., Oleksijew, A., Oltersdorf, T., Rosenberg, S.H., Shoemaker, A.R., Tomaselli, K.J., Zou, H., Fesik, S.W. Discovery of potent antagonists of the antiapoptotic protein XIAP for the treatment of cancer. *J. Med. Chem.* **2004**, *47*, 4417–4426.

19 Huang, Y., Park, Y.C., Rich, R.L., Segal, D., Myszka, D.G., Wu, H. Structural basis of caspase inhibition by XIAP: differential roles of the linker versus the BIR domain. *Cell.* **2001**, *104*, 781–790.

20 Fernando, H., Chin, C., Rösgen, J., Rajarathnam, K. Dimer dissociation is essential for interleukin-8 (I-8) binding to CXCR1 receptor. *J. Biol. Chem.* **2004**, *279*, 36175–36178.

21 Clore, G.M., Appella, E., Yamada, M., Matsushima, K., Gronenborn, A.M. Three-dimensional structure of interleukin 8 in solution. *Biochemistry* **1990**, *29*, 1689–1696.

22 PDB entries selected for DNA–ligand case study. (a) 442D, 448D, 444D, 443D, 449D, 445D, 447D: Squire, C.J., Baker, L.J., Clark, G.R., Martin, R.F., White, J. Structures of *m*-iodo Hoechst–DNA complexes in crystals with reduced solvent content: implications for minor groove binder drug design. *Nucleic Acids Res.* **2000**, *28*, 1252–1258; (b) 1QV8, 1QV4: Martin, R.F., Broadhurst, S., Reum, M.E., Squire, C.J., Clark, G.R., Lobachevsky, P.N., White, J.M., Clark, C., Sy,

D., Spotheim-Maurizot, M., Kelly, D.P. In vitro studies with methylproamine: a potent new radioprotector. *Cancer Res.* **2004**, *3*, 1067–1070; (c) 109D: Czarny, A., Boykin, D.W., Wood, A.A., Nunn, C.M., Neidle, S., Zhao, M., Wilson, W.D. Analysis of van der Waals and electrostatic contributions in the interactions of minor groove binding benzimidazoles. *J. Am. Chem. Soc.* **1995**, *117*, 4716–4717; (d) 453D: Neidle, S., Mann, J., Rayner, E.J., Baron, A., Opoku-Boahen, Y., Simpson, I.J., Smith, N.J., Fox, K.R., Hartley, J.A., Kelland, L.R. Symmetric bis-benzimidazoles, a new class of sequence-selective DNA-binding molecules. *J. Chem. Soc., Chem. Commun.* **1999**, 929–930; (e) 1FTD: Mann, J., Baron, A., Opoku-Boahen, Y., Johansson, E., Parkinson, G., Kelland, L.R., Neidle, S. A new class of symmetric bisbenzimidazole-based DNA minor groove-binding agents showing antitumor activity. *J. Med. Chem.* **2001**, *44*, 138–144; (f) 360D: Guerri, A., Simpson, I.J., Neidle, S. Visualisation of extensive water ribbons and networks in a DNA minor-groove drug complex. *Nucleic Acids Res.* **1998**, *26*, 2873–2878; (g) 1FMS, 1FMQ, 1EEL: Simpson, I.J., Lee, M., Kumar, A., Boykin, D.W., Neidle S. DNA minor groove interactions and the biological activity of 2,5-bis-[4-(*N*-alkylamidino)phenyl]furans. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2593–2597; (h) 1M6F: Nguyen, B., Lee, M.P. H., Hamelberg, D., Joubert, A., Bailly, C., Brun, R., Neidle, S., Wilson, W.D. Strong binding in the dna minor groove by an aromatic diamidine with a shape that does not match the curvature of the groove. *J. Am. Chem. Soc.* **2002**, *124*, 13680–13681; (i) 166D: Nunn, C.M., Jenkins, T.C., Neidle, S. Crystal structure of gamma-oxapentamidine complexed with d(CGCGAATTCGCG)$_2$. The effects of drug structural change on DNA minor-groove recognition. *Eur. J. Biochem.* **1994**, *226*, 953–961.

8
"Hot Spot" Analysis of Protein-binding Sites as a Prerequisite for Structure-based Virtual Screening and Lead Optimization

Ruth Brenk and Gerhard Klebe

8.1
Introduction

The molecular recognition properties of a binding pocket are determined by the amino acids forming the cavity. The spatial arrangements of these amino acids and their physicochemical properties define the shape and the properties that a ligand has to complement in order to be qualified to bind to the pocket. Therefore, the structure of the binding pocket can be used to map putative interaction sites for certain functional groups such as hydrogen-bond donors and acceptors and hydrophobic features into the binding site. These favorable interactions sites are also referred to as "hot spots". Up to now, there has been no method available to translate these interaction sites directly into chemically accessible new molecules. An indirect way is to derive a pharmacophore hypothesis based on the calculated "hot spots". This pharmacophore is then subsequently used for virtual database screening or to guide docking of a pre-assembled library. In addition, "hot spots" can also be used to tailor scoring functions for a specific target in order to improve their predictive power.

In this chapter, methods to derive "hot spots" and to translate them into real molecules are introduced. The approach is illustrated by using examples of successful protein structure-based virtual screening.

8.2
Calculating "Hot Spots"

Energetically favorable interaction sites in protein binding pockets can be computed and analyzed in terms of so-called "hot spots" of binding. The archetypal program, still being widely in use is Goodford's program GRID [1]. GRID embeds the protein under consideration into a 3D grid. Subsequently, the interaction energy of a molecular probe is calculated at each grid point with respect to the surrounding protein atoms applying the distance-dependent functional form of a force field. Functional groups such as methyl groups as hydrophobic probe

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

or carbonyl oxygens as hydrogen-bond acceptor and amine nitrogens as hydrogen-bond donor probes can be used, representing the most important pharmacophoric properties. Once energy values have been assigned to the individual grid points according to the energy function, a contour surface can be calculated and visualized in terms of appropriate energy levels. If negative energy values are experienced, the areas encompassed by the contour surface represent regions favorably interacting with the selected probe (Fig. 8.1, GRID contour calculated for the binding pocket of carbonic anhydrase). One of the most prominent examples for the successful application of GRID was the development of inhibitors for sialidase [2]. Through the calculation of "hot spots", it was predicted that replacing a hydroxyl group by a basic group should result in more favorable interactions with the surrounding protein, thus improving the affinity of the weak lead structure Neu5Ac2en. In consequence, 4-guanidino-Neu5Ac2en (zanamivir) was designed and subsequently synthesized. This compound turned out to inhibit sialidase with subnanomolar affinity. It was later launched to market as a drug against influenza (Relenza).

SuperStar [3, 4] and DrugScore [5, 6] (Fig. 8.1) are two alternative, popular programs to calculate "hot spots" in binding pockets. As in GRID, the protein is embedded into a 3D grid, but different types of functional forms are used for evaluating the interactions of a probe with the surrounding binding pocket (Fig. 8.1). SuperStar, a knowledge-based approach, uses information about intermolecular interactions observed in the crystal packing of small organic molecules whereas DrugScore relies on knowledge-based potentials assembled from individual protein–ligand complexes. In SuperStar, relevant composite-crystal field environments of a particular type of functional group around a central group of interest, compiled from a large set of small molecule crystal structures, are mapped on to binding-site exposed residues. Facing such distributions to a random assembly of contacting groups, SuperStar allows one to estimate the propensity of finding a particular contacting group next to the central group under investigation. DrugScore uses the annotated information in Relibase [7] and performs histographic statistics on the contact frequencies of ligand functional groups in the neighborhood of protein functional groups. By defining an appropriate reference state (e.g. the mean distribution of all atom types), some sort of probability for the formation of ligand-to-protein contacts can be expressed. These contact preferences can be related to some type of statistical potential that provides a means of determining how favorably a particular contact can be estimated. Recently, potentials for protein–water interactions have been added to DrugScore [8]. A comprehensive analysis of crystallographically determined protein–ligand complexes with Relibase has revealed that in two-thirds of all complexes a water molecule is involved in ligand binding, e.g. by mediating a contact between protein and ligand [9]. Accordingly, tools to predict and analyze the relevance for water binding sites are crucial for the success of docking techniques or other rational approaches to structure-based drug design. In Fig. 8.2, the computed "hot spots" using the DrugScore implementation for water contacts are shown for arabinose-binding protein together with a bound sugar. The

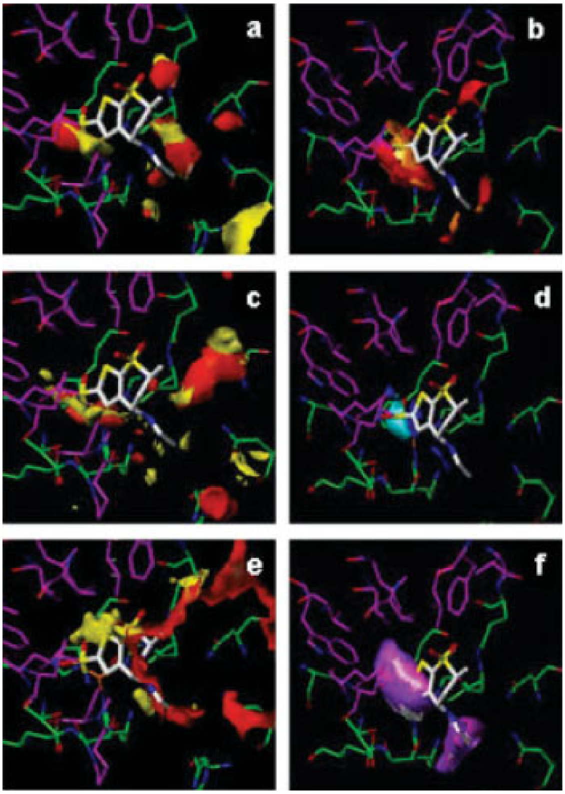


Fig. 8.1 Mapping "hot spots" of binding in the active site of human carbonic anhydrase II, for orientation the crystallographically determined binding geometry of the inhibitor dorzolamide is also shown. (a) The putative hydrogen-bond acceptor properties, calculated for a carbonyl oxygen probe with the program GRID (yellow contours indicate favorable areas) and SuperStar (red contours depict regions of enhanced propensity); (b) the isopleths indicate favorable regions as highlighted by DrugScore of an O.2 (red) and O.3 (orange) probe; (c) favorable hydrogen-bond donor sites determined for an NH amino group with GRID (yellow) and SuperStar (red); (d) similar properties indicated by DrugScore of an N.3 (blue) and N.am (cyan) probe; (e) "hot spots" for hydrophobic properties as computed by GRID for the probe DRY (yellow) and a methyl group in SuperStar (red); (f) similar properties as highlighted by DrugScore of a C.3 (violet), C.ar (magenta) and C.2 (white) probe. The various grid points were contoured at a 10% level above the global minimum in each map. Reprinted with permission from *J. Med. Chem.* **2002**, *45*, 3588–3602.

indicated "hot spots" clearly denote water positions which are actually occupied by interstitial water molecules involved in ligand binding.

A somewhat different approach has been implemented in MCSS (multiple copies simultaneous search) [10]. In this method, up to 5000 copies of a functional group are randomly distributed in the binding pocket and simultaneously minimized by a force field. The simultaneous optimization is performed in

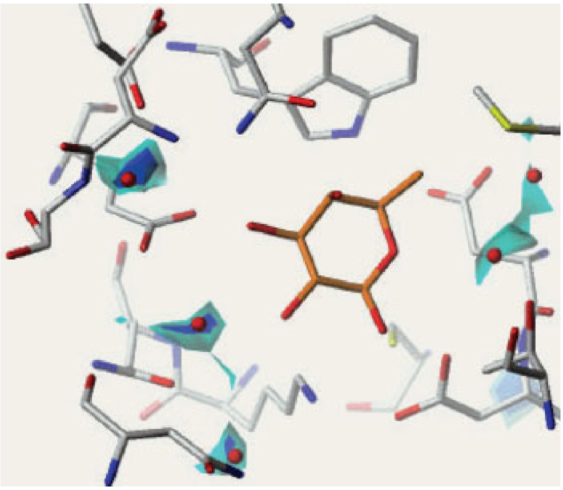


Fig. 8.2 “Hot spots” calculated for water sites with DrugScore in the binding pocket of arabinose binding protein (5abp). The DrugScore potential has been contoured at two levels (cyan/blue) and the crystallographically determined water molecules are shown as red spheres.

such a way that none of the probe fragments experiences any of the others, but all fragments encounter interactions with the protein binding site residues. The resulting clusters of probe fragments indicate the most favorable interaction sites and possible interaction geometries.

8.3 From “Hot Spots” to Molecules

Up to now, there has been no method available which directly translates the “hot spots” into the chemical structure of real molecules. An indirect way is to derive a pharmacophore hypothesis for protein structure-based virtual screening based on the calculated interaction sites. Two widely applied programs for this task are UNITY [11] and Catalyst [12]. By use of both programs, spheres can be placed into the binding pocket with the radius (considering an appropriate search tolerance) adjusted to encompass the underlying “hot spot”. The pharmacophoric property of the corresponding “hot spot”, e.g. being a hydrogen-bond donor or -acceptor or hydrophobic site is subsequently assigned to the sphere. In addition, the approximate shape of the binding pocket can be considered in terms of excluded volumes to mimic the protein environment that a ligand is not allowed to penetrate. Subsequently, the derived pharmacophore can be used for database screening. In this step, a geometric and chemical mapping between the assigned tolerance spheres and the compounds stored in a database is calculated. Catalyst operates on precalculated multiple conformations of the small-molecule candidates, whereas UNITY generates them on the fly, starting with a very fast incipient tweak algorithm to test for crude pharmacophore matching. An alternative approach to retrieve putative candidate molecules from a database is the program FeatureTrees developed by Rarey and Dixon [13]. This method is focused on the comparison of ligands and retrieves by very fast algorithms candidate molecules with topographical similarity to a given lead reference. The molecules to be compared are described by a tree of knots to which generic properties such as H-bond donor and acceptor or hydrophobic properties have been assigned.

Recently, existing docking programs have been extended to consider pharmacophore hypotheses during docking. In FlexX-Pharm [14], an extension of the original FlexX [15], the user has the option to incorporate pharmacophore features either as a pre-filter for docking, as a constraint during docking or as a post-filter to rank and evaluate docking solutions. In the pre-filter mode for docking, each ligand is checked first in order to estimate whether it can potentially satisfy the pharmacophore hypothesis. Only if this prerequisite is met is the compound actually docked into the binding pocket. If the pharmacophore is used as a constraint during docking, only poses in agreement with this hypothesis are stored. Since undesired docking poses are eliminated early on, this filter step allows new poses to emerge which otherwise would not have been explored. In the post-filter mode, the ligands are docked unconstrained, but all solutions not in agreement with the pharmacophore hypothesis are subsequently discarded from the list of putative hits. By using these different options, the computing time is significantly reduced, poses closer to the experimentally determined binding mode can be generated and better enrichment of known binders in database searches can be achieved.

PhDOCK [16, 17], an extension of DOCK 4.0 [18], also considers pharmacophores during docking. This enhanced release requires a database storing molecules in a configuration superimposed on the largest common pharmacophore. Such clusters are composed of different conformers of the same molecule but they also comprise similar molecules exhibiting the same pharmacophore (Fig. 8.3). For docking these clusters into the binding pocket, the cavity is filled with spheres. These spheres are labeled with pharmacophoric properties derived from a "hot spot" analysis. All members of one cluster are simultaneously placed in the binding pocket by matching the pharmacophore representing the cluster with spheres reflecting the protein-based pharmacophore properties [16]. Consequently, all poses obtained are scored separately for each conformer. Only the best scoring pose of each compound is stored in the final hit list. This procedure results in reduced computational efforts and better performance with respect to enrichment of known actives is achieved.

All the above-described methods share in common that the user has to select the calculated "hot spots" to be considered in the pharmacophore hypothesis. But even without establishing an explicit pharmacophore hypothesis, "hot spots" can be useful for structure-based drug design.

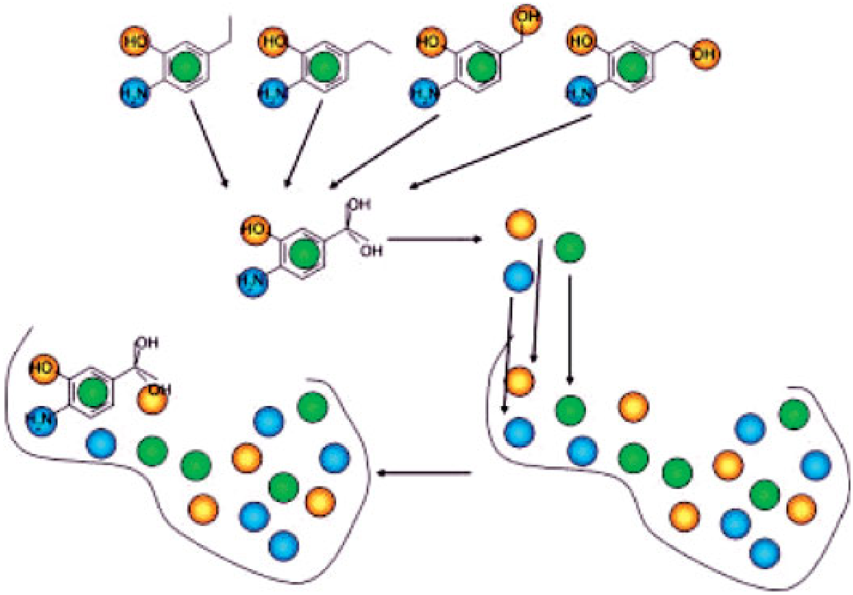


Fig. 8.3 Concept of PhDOCK. Conformers of similar molecules are aligned on the largest common pharmacophore. The pharmacophore of each cluster is consequently matched on spheres in the binding pocket labeled with pharmacophoric properties. Based on this match a transformation matrix is calculated and the molecules present in the cluster are docked into the binding site. Each molecule present in the cluster is scored and the best scoring conformer per molecule is stored in the final hit list.

One option is the direct docking of ligands on to "hot spot" grids [19]. In a feasibility study, using AutoDock [20] docking on to grids assigned by DrugScore potentials reveals more relevant ligand poses. This strategy is superior to a simple re-ranking of ligand poses using DrugScore.

If a set of compounds binding to the protein of interest is known, "hot spots" can be used to tailor a given general purpose scoring function [21]. In this approach, named AFMoC (*a*daptation of *f*ields for *mo*lecular *c*omparison), DrugScore potential fields are generated in the binding pocket. Methodologically the approach is related to the well-known comparative molecular field analysis tools CoMFA [22] and CoMSIA [23], but with the important advantage that the protein environment is explicitly considered. CoMFA and CoMSIA apply grids, assigned to the field values of a uniform probe, and use a functional form to map the ligand properties on to these grid points which corresponds either to Lennard–Jones and Coulomb potentials or uses molecular similarity indices. In contrast, AFMoC starts with a grid of pre-assigned values. These non-uniform values at the individual grid point consider the DrugScore potential values computed for various atoms according to the contact geometries with the surrounding protein environment. By use of some ligands with known binding mode and experimentally determined binding affinity, the actually placed ligand atoms introduce an affinity-based weighting of the individual DrugScore potential values [24]. The resulting interaction fields for the training set ligands are evaluated by PLS. Once such a comparative molecular field analysis has been established and a QSAR equation is derived, it can be used to predict binding affinities of novel ligands, e.g. such as result from a docking run. Accordingly, an AFMoC-derived QSAR model serves as a tailor-made scoring function that has implicitly been weighted with respect to "hot spot" data. It has been shown that the tailored scoring function achieves much better correlations between experimentally determined and computed affinities and possesses superior predictive power in estimating binding affinities compared with the original general purpose DrugScore [24].

8.4 Real-life Examples

To illustrate how pharmacophore searches can be applied successfully to the discovery of novel ligands, we chose two examples from our own recent work.

The enzyme tRNA–guanine transglycosylase (TGT) catalyzes the complete exchange of a base in tRNA [25, 26]. Upon reaction, guanine in the wobble position of tRNAs with the anticodon sequence GUN is replaced by the modified bases

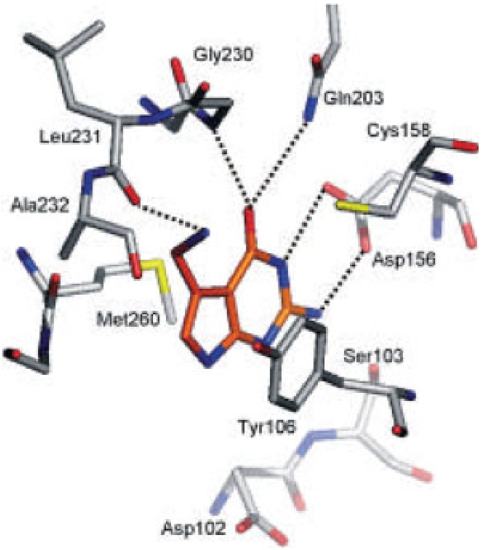


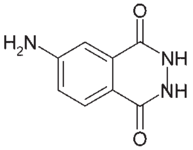
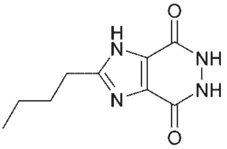
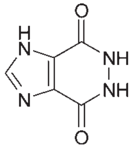
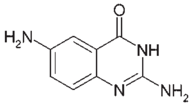
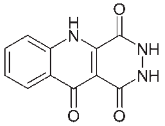
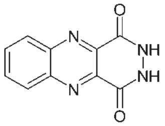
Fig. 8.4 Binding mode of preQ$_1$ complexed with *Z. mobilis* TGT. The substrate is intercalated between the hydrophobic side-chains of Tyr106 and Met260. Specific recognition occurs via hydrogen bonds towards Asp156, Gln203, Gly230 and Leu231. Reprinted with permission from *J. Mol. Biol.* **2004**, *338*, 55–75.

preQ_1. In subsequent reaction steps, preQ_1, once incorporated into tRNA, is converted to queuine. In *Shigella*, the causative agent of shigellosis, the occurrence of tRNA modified by queuine is a prerequisite for efficient translation of *virF*, a major virulence factor [27]. Mutational studies showed that *Shigella flexneri* lacking the *tgt* gene suffers drastically reduced pathogenicity compared with the wild type [28]. This prompted us to target TGT in an effort to design new antibiotics against shigellosis [29]. Since no crystal structure of *Shigella* TGT is available, the structure of the *Zymomonas mobilis* enzyme in complex with preQ_1 (Fig. 8.4) served as the platform for our studies. Sequence analysis has shown that, apart form a replacement of Tyr106 by Phe, all amino acids in the binding pocket are identical with those present in the *S. flexneri* enzyme [30]. In addition, we could show that mutating Tyr106 to Phe does not change the kinetic properties [31].

In the first design cycle, pyridazinediones, e.g. **1** (Table 8.1), were discovered using the *de novo* design program LUDI [32]. Subsequently, the crystal structure of the TGT · ligand complex was determined [29]. This analysis revealed that the discovered lead exhibits a substrate-like binding mode (Fig. 8.5 a). Based on these findings, we focused in a second design cycle on closely related analogs of the original pyridazinedione skeleton, resulting in a family of imidazole-fused pyridazinediones such as **2** and **3** [33]. To our surprise, these compounds adopt a distinct binding mode (Fig. 8.6). Upon ligand binding, the amide carbonyl group of Leu231 which in all previously determined crystal structures faced the ligand binding site, is flipped in the opposite direction towards the interior of the protein. Instead, the adjacent amide NH group of Ala232 is now exposed to the binding cavity. In addition, an interstitial water molecule (W1) is present that mediates a contact between the ligand and Ala232.

To exploit this new and unexpected binding mode for drug discovery of further inhibitors, we derived a protein-based pharmacophore for virtual screening. In a first step, "hot spots" were calculated in the cavities for both alternative binding site conformers using DrugScore [6] and SuperStar [3]. To represent the important properties of putative ligands, a hydrogen-bond donor (N.3 in DrugScore and an uncharged amino group in Superstar), hydrogen-bond acceptor (O.2 and carbonyl oxygen probe, respectively) and a hydrophobic probe (C.ar probe and aromatic C probe, respectively) were chosen. Representative results are shown in Fig. 8.5. Superimposing the crystallographically observed binding modes of the ligands on the interaction sites predicted as favorable shows that the given properties of the ligands correspond very well to the indicated "hot spots". For example, all hydrogen-bond donor groups present in **1** (Fig. 8.5 a) match convincingly with the predicted interaction sites for this type of functional group. One of the carbonyl oxygens of **2** also coincides very well with a hydrogen-bond acceptor "hot spot" (Fig. 8.5 c), whereas the second carbonyl group is placed in a region not predicted as favorable. An explanation for this finding is the fact that this group only forms hydrogen bonds to surrounding water molecules which were not considered in the "hot spot" calculations (Fig. 8.7). Interestingly, for the position of the nitrogen in ligand **2**, interacting with water molecule W1, the "hot spot" analysis predicts favorable interactions

Table 8.1 Inhibitors of *Z. mobilis* TGT

No.	Compound
1	
2	
3	
4	
6	
7	

for hydrogen-bond donors and acceptors (Fig. 8.5 b and c). This is due to the bifunctionality of the water molecule, which can act as hydrogen-bond donor and acceptor. This observation is convincingly confirmed by two inhibitors, **6** and **7**, synthesized in our lead optimization program (Table 8.1) [34]. Both compounds exhibit a binding affinity in the low micromolar range, but **6** exposes a protonated nitrogen function towards the interstitial water molecule, whereas **7** interacts with this water via its basic nitrogen, most likely exposing its acceptor properties (Fig. 8.8).

The hydrophobic probe highlights an area of the binding pocket where the bi- and tricyclic inhibitors are actually accommodated with their non-polar ring system (Fig. 8.5 d).

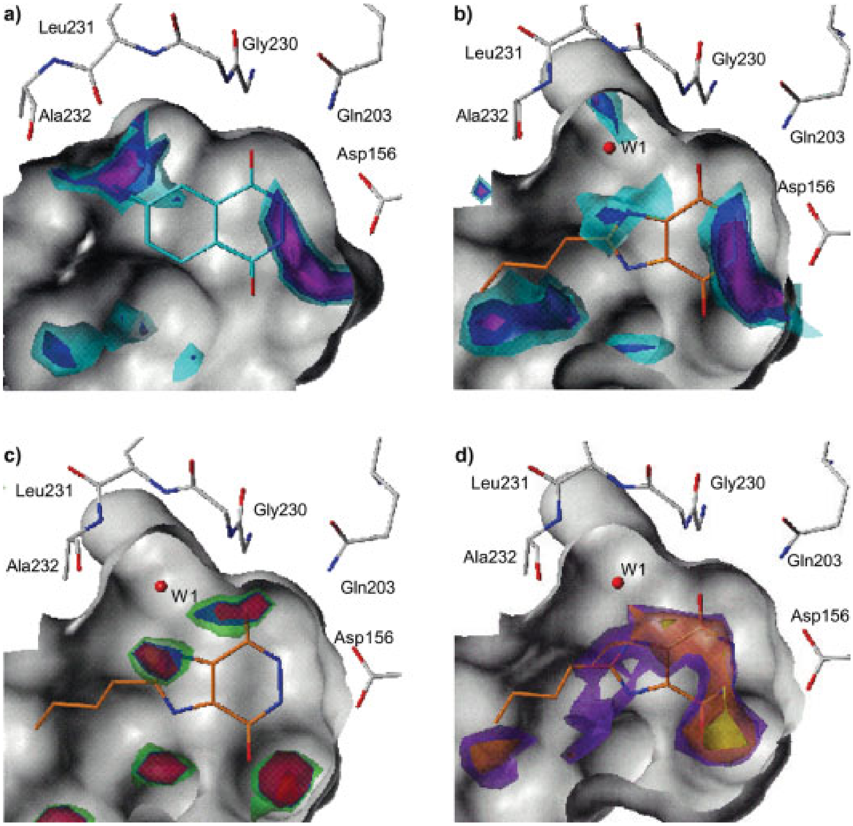


Fig. 8.5 Mapping of putative binding “hot spots” in the active-site of TGT. For orientation, the binding geometry of the inhibitor **1** (a) or **2** (b–d), studied crystallographically, is also shown. (a), (b) “Hot spots” calculated with DrugScore using a hydrogen-bond donor probe (N.3), contoured at an 80 (cyan), 84 (blue) and 88% (magenta) level with respect to the global minimum; (c) “hot spots” calculated with SuperStar using a carbonyl oxygen probe, contoured at a propensity level of 4 (green), 8 (blue) and 10 (red) (a propensity of 1 corresponds to random occurrence); (d) “hot spots” calculated with DrugScore using a hydrophobic probe (C.ar), contoured at an 89 (magenta), 91 (orange) and 95% (yellow) level. Reprinted with permission from *J. Med. Chem.* **2003**, *46*, 1133–1143.

In the following step, the calculated “hot spots” considering various probes are combined and translated into a pharmacophore hypothesis that serves as input query for a UNITY [11] search. Including all prominent spots would result in a far too complex pharmacophore model, providing only a small chance of finding ligands that would fulfil such complex criteria. However, picking too few spots could result in a very general description which would not be sufficient to describe specific TGT inhibitors. Therefore, a well-balanced hypothesis has to be established and we focused on key interactions which have been observed repeatedly in previously determined crystal structures of TGT–ligand

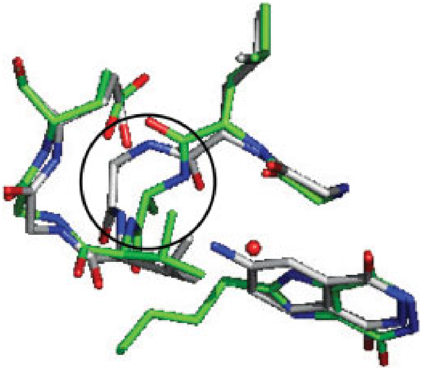


Fig. 8.6 Superposition of TGT **1** (green) and TGT·**2** (gray). The peptide bond at Leu231 to Ala232 (circle) in the structure of TGT complexed with **2** is flipped compared with that with **1**. The flip rotates the carbonyl group of Leu231 in the opposite direction from the binding site. Instead, the adjacent NH of Ala232 is now facing towards the ligand. Reprinted with permission from *J. Med. Chem.* **2003**, *46*, 1133–1143.

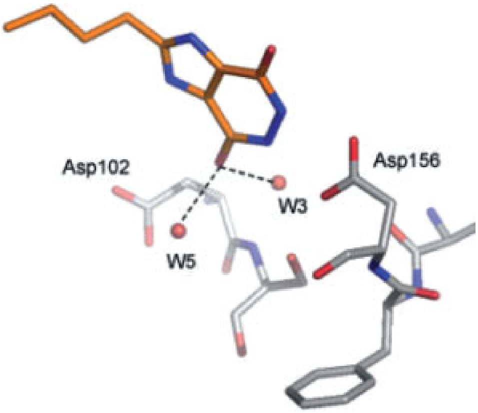


Fig. 8.7 Part of the crystal structure of TGT·**2**. One of the ligand's carbonyl groups only forms hydrogen bonds to surrounding water molecules (W3 and W5).

complexes. The finally established pharmacophore hypothesis is shown in Fig. 8.9. We hypothesized that all inhibitors require a hydrophobic core structure to intercalate between Tyr106 and Met260. Furthermore, they have to form a twinned hydrogen bond to Asp156 and, in addition, a hydrogen bond to either Gln203 or Gly230. Additionally, a hydrogen-bond donor directly interacting with Leu231 or either a hydrogen-bond donor or acceptor interacting via the interstitial water molecule W1 with the flipped peptide bond is required. The spatial

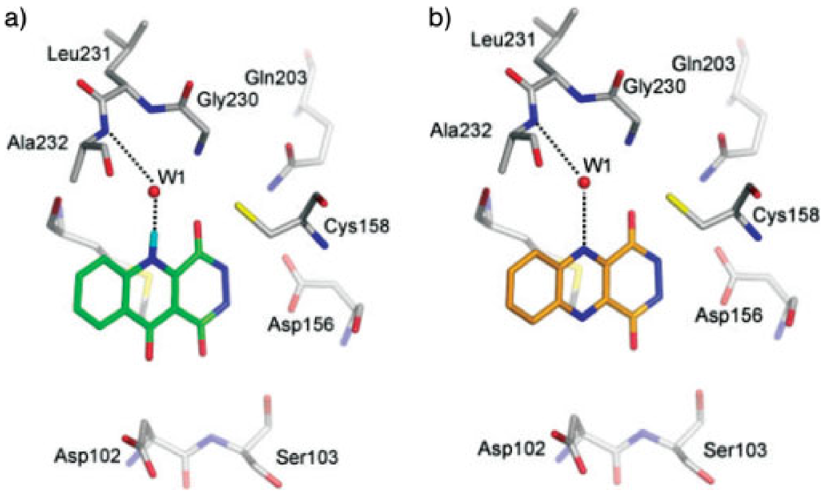


Fig. 8.8 Modeled binding mode of **6** [green (a)] and **7** [orange (b)]. Ligand **6** exposes a hydrogen-bond donor towards the water molecule W1, whereas ligand **7** interacts with this water molecule via a hydrogen-bond acceptor functionality.

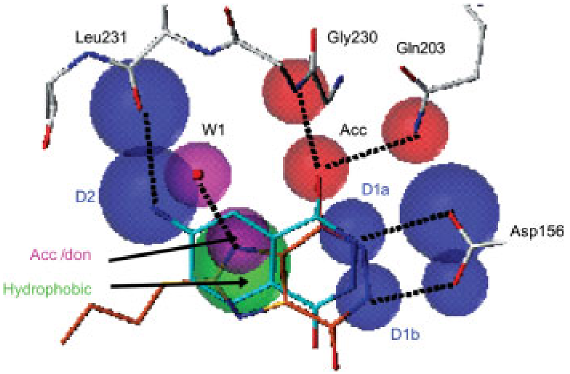


Fig. 8.9 Composite protein structure-based pharmacophore used for virtual screening. Donor features and their corresponding interaction partners in the protein are colored blue, acceptor features and their corresponding interaction partners are colored red and the acceptor/donor feature and the corresponding acceptor/donor site is colored magenta. The hydrophobic feature is colored green. The interaction to the carbonyl group of Leu231 is considered alternatively to the interaction to the water molecule W1. Reprinted with permission from *J. Med. Chem.* **2003**, *46*, 1133–1143.

tolerances for the placement of the putative pharmacophore groups were adjusted to the size of the underlying “hot spots”. To consider directionality of the hydrogen bonds, all hydrogen-bond donor and acceptor features were connected to the corresponding groups in the protein.

Table 8.2 Summary of the hierarchical filtering of a database of small molecules using the pharmacophore described in Fig. 8.9 (screening 1) and Fig. 8.12 (screening 2)

Filter step	Screening 1		Screening 2	
	No. of compounds	%	No. of compounds	%
	826952	100.00	826952	100.00
1. Rotatable bonds/MW	419737	50.76	419737	50.76
2. Requested number of hydrophobic, donor, and acceptor properties	168387	20.36	242005	29.30
3. Pharmacophore hypothesis	3309	0.40	39080	4.73
4. Excluded volumes	872	0.11	620	0.08

In the following, this composite pharmacophore was used as input for database searches. As the database of putative candidate molecules we assembled several hundred thousand compounds offered by commercial vendors. To speed up the search and to keep control over the search strategy, we followed a hierarchical protocol with increasingly complex filters (Table 8.2, Screening 1). All compounds exceeding a molecular weight of 450 Da and comprising more than seven rotatable bonds were discarded in the first step in order to focus on more drug- and lead-like compounds. Subsequently, only compounds with a minimal number of functional groups required to satisfy the pharmacophore were allowed to pass the second filter. As the third step, enhanced consideration of the above-described pharmacophore was applied. In addition to the requested presence of appropriate functional groups, it was queried whether these selected functional groups are also topographically arranged in such a way as to agree with the spatial geometry of the pharmacophore. This provides a very tight filter and only about 3300 entries could pass this step. In the next stage, the shape of the binding site was considered in terms of excluded volumes. After all filter steps, the original number of compounds was reduced to about 0.1%. The remaining compounds were docked by FlexX [15] into the TGT binding pocket in a pharmacophore-unrestrained fashion. All compounds still satisfying the pharmacophore hypothesis after docking were subsequently inspected visually. Criteria for purchasing compounds for testing were (a) the overall matching of the desired hydrogen-bonding network, (b) complementarity between ligand and protein surfaces in terms of spatial occupancy and matched contacts in hydrophobic/hydrophilic surface patches and (c) the absence of any unfavorable intermolecular interactions after minimizing the compounds in the binding pocket using the MAB force field [35, 36]. In total, nine compounds were tested (Table 8.3). All of these compounds showed at least submicromolar inhibition. Three inhibited the enzyme in the low micromolar range (**10**, **14** and **15**) and two were even submicromolar inhibitors (**8** and **9**). Most important for the following

Table 8.3 Compounds selected by virtual screening using the pharmacophore described in Fig. 8.9 and subsequently tested for TGT inhibition

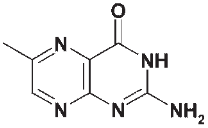
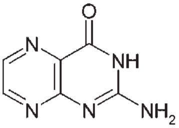
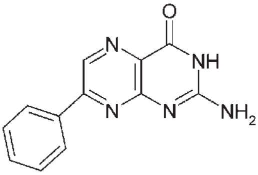
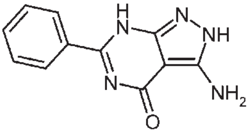
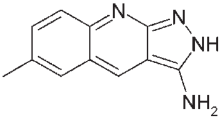
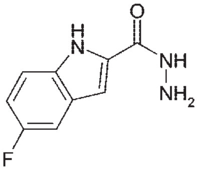
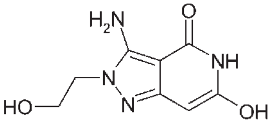
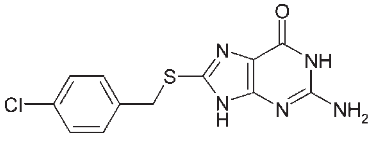
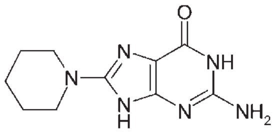
No.	Compound	K_i (μM) [a)]
8		0.25
9		0.6
10		3.8
11		249
12		156
13		72
14		8.1
15		2.7

Table 8.3 (continued)

No.	Compound	K_i (μM) [a)]
16		37

a) Owing to the elaborate determination of the K_i values, the error is assumed to be about 20–30% [29]. Data have not been corrected for competitive and uncompetitive contributions to inhibition [39].

lead optimization, our approach using a hierarchical protein-based virtual screening protocol retrieves compounds originating from different chemical classes such as pteridines, guanines, hydrazides and pyrazoles. This provides the greatest opportunities for a successful lead optimization program by structure-based chemical synthesis.

8.5
Replacement of Active-site Water Molecules

In a second attempt to explore further the recognition properties of the TGT, we focused on 4-aminoquinazolinones as initial lead structure (e.g. **4** in Table 8.1) [37–39]. Crystal structure analysis revealed that in TGT · **4** the side-chain of Asp102 is rotated towards the ligand forming a twinned hydrogen bond (Fig. 8.10). In consequence, the water molecule W3 which was present in the TGT–**1** and TGT–**2** (Fig. 8.7) complexes is displaced. In addition, the 2-amino group of **1** displaces the water molecule W5. These rearrangements attracted our attention to this part of the binding pocket. The "hot spot" maps for a hydrogen-bond donor (N.3) and an acceptor probe (O.2) are shown in Fig. 8.11 a and b. Water molecule W4 is found at the rim of the hydrogen-bond donor area contoured at a predefined level (Fig. 8.11 a). However, water molecule W3 coincides nicely with the center of a hydrogen-bond acceptor spot (Fig. 8.11 b). In addition, water molecule W5, which is in hydrogen-bond distance of W3, also matches convincingly with a hydrogen-bond acceptor spot. Based on this analysis, we concluded that the water molecule W4 is too deeply buried in the pocket and its position is not easily accessible for ligand donor functional groups. In contrast, the water molecules W3 and W5 occupy two easily accessible spots favorable for hydrogen-bond acceptor functionalities. According to these results, we formulated a modified pharmacophore hypothesis in particular addressing this part of the binding site (Fig. 8.12). In the above-described pharmacophore search, no ligands were found which, in addition to the required features, also

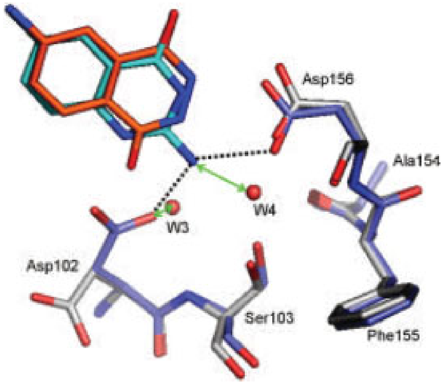


Fig. 8.10 Superposition of TGT·**1** (orange ligand, gray protein structure) and TGT·**4** (cyan ligand, blue protein structure). Upon binding of **4**, Asp102 rotates towards the ligand. In consequence, the water molecules W3 and W4 present in TGT·**1** are displaced. Reprinted from *J. Mol. Biol.* **2004**, *338*, 55–75, with permission from Elsevier.

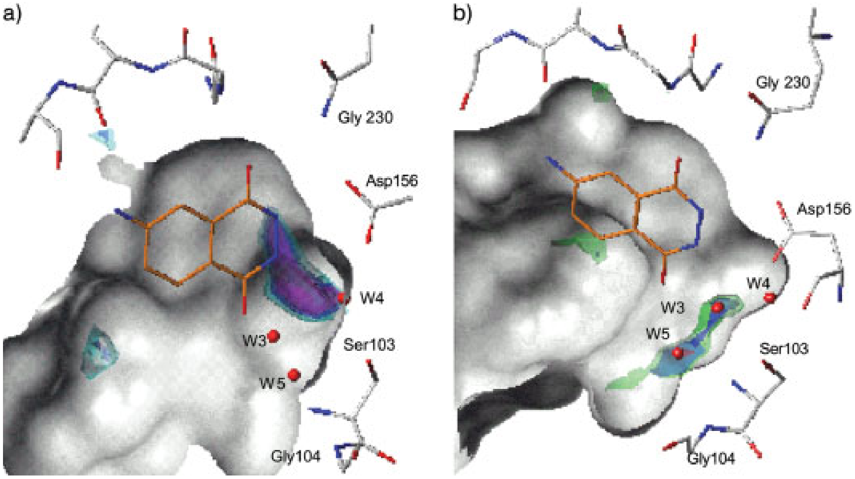


Fig. 8.11 Mapping of putative binding “hot spots” calculated with DrugScore close to the water molecules W3–W5 in the active-site of TGT. For orientation, the binding mode of inhibitor **1** is also shown. (a) “Hot spots” calculated using a hydrogen-bond donor probe (N.3), contoured at an 80 (cyan), 84 (blue) and 88% (magenta) level with respect to the global minimum. The water molecules W3–W5 do not coincide with these spots. (b) “Hot spots” calculated using a hydrogen-acceptor probe (O.2), contoured at an 80 (green), 84 (blue) and 92% (red) level with respect to the global minimum. The water molecule W3, which is displaced upon rotation of the side-chain of Asp102, and the water molecule W5 are located in a position predicted as favorable for a hydrogen-bond acceptor. Reprinted from *J. Mol. Biol.* **2004**, *338*, 55–75, with permission from Elsevier.

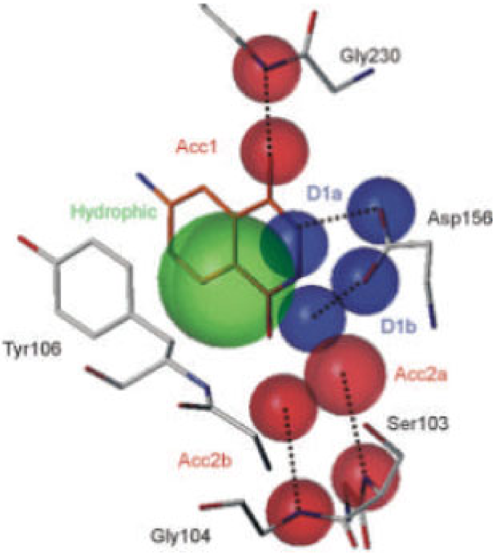


Fig. 8.12 Protein structure-based pharmacophore derived for ligands which should displace either water molecule W3 or W5. Donor features and their corresponding interaction partners in the protein are colored blue and acceptor features and their corresponding interaction partners are colored red. The hydrophobic feature is colored green. The interactions to the carboxylate group of Asp156 and to acceptor groups Acc2a and Acc2b are requested in the search query as alternative options. Reprinted from *J. Mol. Biol.* **2004**, *338*, 55–75, with permission from Elsevier.

placed an acceptor functionality into the acceptor spot areas next to W3 and W5. In order possibly to retrieve ligands addressing these two spots, it was therefore important to relax the stringent requirements defined for the previous pharmacophore hypothesis. In the updated pharmacophore, instead of a twinned hydrogen bond to Asp156, a single one was considered sufficient. In addition, a hydrogen bond to Gly230 and a hydrophobic moiety were still assumed as essential binding prerequisites. Accordingly, the novel pharmacophore hypothesis comprised two additionally defined hydrogen-bond acceptors interacting with Gly104 and Ser103. They were placed to the centers of the "hot spots" where the water molecules W3 and W5 were crystallographically observed.

Subsequently, the new pharmacophore hypothesis was used for virtual screening. The same strategy involving a series of hierarchical filters, as described above, was applied. The statistical results of this second search are listed in Table 8.2 (Screening 2). In total, about 700 molecules could pass all filter steps. After docking with FlexX and visual inspection, six compounds were selected for testing their inhibitory potency (Table 8.4). All of them inhibited TGT, at least at the submillimolar level; compounds **19–22** were the most potent compounds with inhibition constants in the two digit micromolar range. All retrieved ligands, except **17** and **20**, exhibit a low Tanimoto similarity index [40, 41] taking the original lead compounds **1** and **2** as references. Within this similarity metric, the newly discovered hits can therefore be classified as "dissimilar" or "novel". Unfortunately, up to now we have not succeeded in determining a crystal structure with e.g. **21**. Accordingly, we assume a binding mode as proposed by docking which is shown in Fig. 8.13 a. The ligand's nitro group is hosted in the area that was previously occupied by both waters W3 and W5. Compared with the TGT inhibitors studied earlier, **21** is rotated by 90° with respect to its longitudinal axis in the binding pocket (Fig. 8.13 b). As a consequence, in this orientation, the lower right-hand part of the binding pocket, adjacent to Ser103 and Gly104, is completely occupied. Through decoration with substituents in an ortho or meta position with respect to the nitro group at the central phenyl ring, regions of the binding pocket so far unexplored can now be addressed. These considerations make the latter lead a valuable candidate for further optimization.

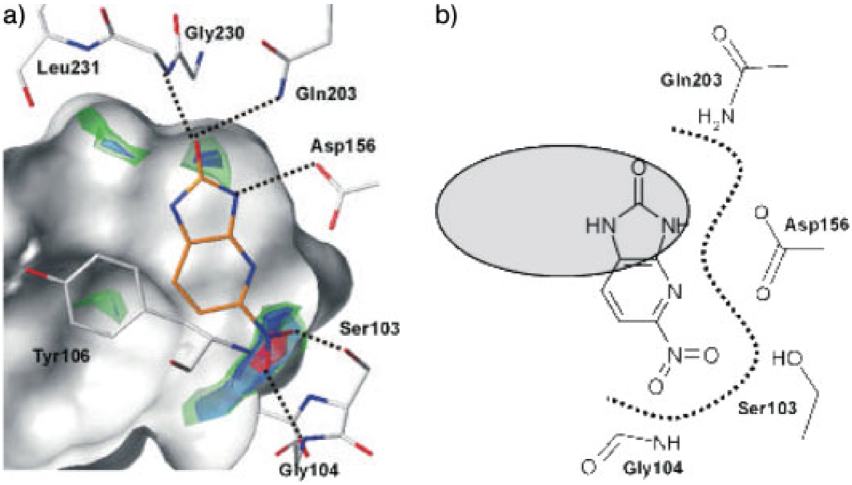
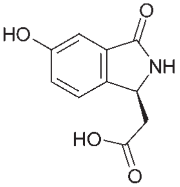
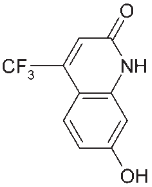
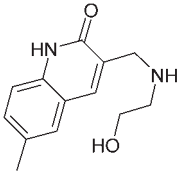
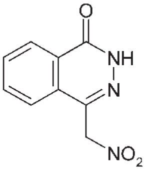
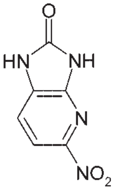
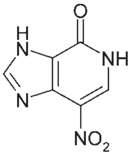


Fig. 8.13 (a) Best docking solution (rank 1) for ligand **21**. The ligand exposes its nitro group in a region predicted to be favorable for hydrogen-bond acceptors by DrugScore ("hot spots" contoured at an 80 (green), 84 (blue) and 92% (red) level). **(b)** Sketch of the binding mode of **21**. Compared with the so far known binding modes of TGT inhibitors (gray), the compound is rotated 90° with respect to its longitudinal axis. In consequence, the previously unoccupied lower right part of the binding pocket towards Ser103 and Gly104 is now addressed. Reprinted from *J. Mol. Biol.* **2004**, *338*, 55–75, with permission from Elsevier.

Table 8.4 Compounds selected by virtual screening using the pharmacophore described in Fig. 8.12 and subsequently tested for TGT inhibition.

No.	Compound	Tanimoto index to 1	Tanimoto index to 2	K_i (μM) [a)]
17		0.39	0.21	403 ± 33
18		0.27	0.22	158 ± 17
19		0.29	0.24	58 ± 15
20		0.45	0.24	31 ± 5
21		0.22	0.23	27 ± 3
22		0.19	0.29	15 ± 1

a) Owing to the elaborate determination of the K_i values, the error is assumed to be about 20–30% [29]. Data have not been corrected for competitive and uncompetitive contributions to inhibition [39].

8.6 Conclusions

Structure-based virtual screening has been established as an alternative tool to high-throughput screening for the discovery of novel leads. In particular, owing to early access of relevant crystal or NMR structures of putative drug targets, rational structure-based design methods are increasingly applied in the drug development process. Of special interest in this context is a detailed analysis of the binding pocket of the target protein. In principle, the residues exposed to the binding pocket define the shape and chemical properties of putative ligands to be accommodated by the target protein. Using different molecular probes, the binding pocket can be analyzed in terms of "hot spots" of binding. They indicate in a very generic way the required physicochemical properties, e.g. where in space a hydrogen-bond donor or acceptor facility is necessary to achieve successful binding. However, in a non-trivial step this information displayed by the ensemble of various "hot spots" (also termed protein-based pharmacophore) has to be translated into molecules that satisfy this hypothesis. This latter aspect can currently only be resolved indirectly by screening large amounts of pre-generated candidate molecules and assessing whether they are in agreement with the generic pharmacophore hypothesis. A much more efficient way would be to assemble directly candidate molecules *de novo* considering the constraining conditions of the generic pharmacophore. However, such an approach requires a better understanding of how to translate "hot spots" directly into chemistry. It is hoped that future methodological developments will provide efficient solutions to this as yet unresolved issue.

Acknowledgments

The present studies on TGT were supported by the Deutsche Forschungsgemeinschaft (Grant KL 1204/1) and the Graduiertenkolleg "Protein function at the atomic level". We are grateful for fruitful collaboration with the research group of Prof. Dr F. Diederich (ETH Zurich, Switzerland). We thank CCDC for making IsoStar and SuperStar available to us. BioSolvIT kindly provided us with FlexX and FeatureTrees and Tripos made their Sybyl modeling package available. Accelrys and MDL generously gave us access to their software products CATALYST and ACD, respectively. R.B. is supported by a fellowship of the Ernst Schering Research Foundation.

References

1 Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

2 von Itzstein, M., Wu, W.Y., Kok, G.B., Pegg, M.S., Dyason, J.C., Colman, P.M., Varghese, J.N., Ryan, D.M., Wodds, J.M., Bethell, R.C., Hotham, V.J., Cameron, J.M., Penn, C.R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418–423.

3 Verdonk, M.L., Cole, J.C., Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

4 Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V., Willett, P. SuperStar: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.* **2001**, *307*, 841–859.

5 Gohlke, H., Hendlich, M., Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.

6 Gohlke, H., Hendlich, M., Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein–ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discov. Des.* **2000**, *20*, 115–144.

7 Günther, J., Bergner, A., Hendlich, M., Klebe, G. Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.* **2003**, *326*, 621–636.

8 Günther, J., Gohlke, H., Klebe, G. unpublished results.

9 Günther, J. PhD Thesis, University of Marburg, 2004.

10 Miranker, A., Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29–34.

11 http://www.tripos.com/sciTech/inSilicoDisc/chemInfo/unity.html.

12 http://www.accelrys.com/catalyst/cat_info.html.

13 Rarey, M., Dixon, J.S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.

14 Hindle, S.A., Rarey, M., Buning, C., Lengaue, T. Flexible docking under pharmacophore type constraints. *J. Comput. Aided-Mol. Des.* **2002**, *16*, 129–149.

15 Rarey, M., Kramer, B., Lengauer, T., Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

16 Joseph-McCarthy, D., Alvarez, J.C. Automated generation of MCSS-derived pharmacophoric DOCK site points for searching multiconformation databases. *Proteins* **2003**, *51*, 189–202.

17 Joseph-McCarthy, D., Thomas, B.E.T., Belmarsh, M., Moustakas, D., Alvarez, J.C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins* **2003**, *51*, 172–188.

18 Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411–428.

19 Sotriffer, C.A., Gohlke, H., Klebe, G. Docking into knowledge-based potential fields: a comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45*, 1967–1970.

20 Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293–304.

21 Gohlke, H., Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153–4170.

22 Cramer, R.D., Patterson, D.E., Bunce, J.D. Comparative molecular-field analysis (Comfa). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

23 Klebe, G., Abraham, U., Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.

24 Silber, K., Kurz, T., Heidler, P., Klebe, G. AFMoC enhances predictivity of CoMFA: A case study with DOXP-reductoisomerase. *J. Med. Chem.* in press.

25 Romier, C., Ficner, R., Suck, D. Structural basis of base exchange by tRNA guanine transglycosylase. In *Modification and Editing of RNA*, ASM Press, Washington, DC, 1998, pp. 169–182.

26 Slany, R.K., Kersten, H. Genes, enzymes and coenzymes of queuosine biosynthesis in procaryotes. *Biochimie* **1994**, *76*, 1178–1182.

27 Durand, J.M., Dagberg, B., Uhlin, B.E., Björk, G.R. Transfer RNA modification, temperature and DNA superhelicity have a common target in the regulatory network of the virulence of *Shigella flexneri*: the expression of the virF gene. *Mol. Microbiol.* **2000**, *35*, 924–935.

28 Durand, J.M., Okada, N., Tobe, T., Watarai, M., Fukuda, I., Suzuki, T., Nakata, N., Komatsu, K., Yoshikawa, M., Sasakawa, C. vacC, a virulence-associated chromosomal locus of *Shigella flexneri*, is homologous to tgt, a gene encoding tRNA–guanine transglycosylase (TGT) of *Escherichia coli* K-12. *J. Bacteriol.* **1994**, *176*, 4627–4634.

29 Grädler, U., Gerber, H.D., Goodenough-Lashua, D. M., Garcia, G. A., Ficner, R., Reuter, K., Stubbs, M.T., Klebe, G. A new target for shigellosis: rational design and crystallographic studies of inhibitors of tRNA–guanine transglycosylase. *J. Mol. Biol.* **2001**, *306*, 455–467.

30 Romier, C., Meyer, J.E., Suck, D. Slight sequence variations of a common fold explain the substrate specificities of tRNA-guanine transglycosylases from the three kingdoms. *FEBS Lett.* **1997**, *416*, 93–98.

31 Brenk, R., Stubbs, M.T., Heine, A., Reuter, K., Klebe, G. Flexible adaptations in the structure of the tRNA modifying enzyme tRNA–guanine transglycosylase and their implications for substrate selectivity, reaction mechanism and structure-based drug design. *ChemBiochem* **2003**, *4*, 1066–1077.

32 Böhm, H.J. The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

33 Brenk, R., Naerum, L., Grädler, U., Gerber, H.D., Garcia, G.A., Reuter, K., Stubbs, M.T., Klebe, G. Virtual screening for submicromolar leads of tRNA–guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J. Med. Chem.* **2003**, *46*, 1133–1143.

34 Brenk, R., Gerber, H.-D., Kittendorf, J.D., Garcia, G.A., Reuter, K., Klebe, G. From hit to lead: *de novo* design based on virtual screening hits of inhibitors of tRNA–guanine transglycosylase, a putative target of shigellosis therapy. *Helv. Chim. Acta* **2003**, *86*, 1435–1452.

35 Gerber, P.R. Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *J. Comput. Aided-Mol. De.s* **1998**, *12*, 37–51.

36 Gerber, P.R., Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 251–268.

37 Brenk, R., Meyer, E., Reuter, K., Stubbs, M.T., Garcia, G.A., Dietrich, F., Klebe, G. Crystallographic study of inhibitors of tRNA–guanine transglycosylase suggests a new structure-based pharmacophore for virtual screening. *J. Mol. Biol.* **2004**, *338*, 55–75.

38 Meyer, E.A., Brenk, R., Castellano, R.K., Furler, M., Klebe, G., Dietrich F. *De novo* design, synthesis and *in vitro* evaluation of inhibitors for prokaryotic tRNA–guanine transglycosylase: a dramatic sulfur effect on binding affinity. *ChemBiochem* **2002**, *3*, 250–253.

39 Meyer, E.A, Donati, N., Guillot, M., Schweizer, W.B., Dietrich, F., Stengl, B., Brenk, R., Reuter, K., Klebe, G. Synthesis, Biological Evaluation, and Crystallographic Studies of Extended Guanine-Based (lin-Benzoguanine) Inhibitors for tRNA-Guanine Transglycosylase (TGT), *Helv. Chim. Acta* **2006**, *89*, 573–597.

40 Dean, P.M. *Molecular Similarity in Drug Design*, Blackie Academic and Professional, London, 1995.

41 Downs, G.M., Willett, P. Similarity Searching in databases of chemical structures. In *Reviews in Computational Chemistry*, K.B. Lipkowitz, D.B. Boyd (eds.) VCH, New York, 1996, pp. 1–66.

9
Application of Pharmacophore Fingerprints to Structure-based Design and Data Mining

Prabha Karnachi and Amit Kulkarni

9.1
Introduction

Virtual screening is being increasingly used to prioritize compounds for biological testing in pharmaceutical lead discovery programs. 2D methods such as similarity searching and 3D methods such as pharmacophore mapping and ligand docking are routinely used for virtual screening of corporate databases. Currently, structure-based methods (e.g. docking) for virtual screening of corporate libraries, external compound collections and virtual compounds are relatively slow.

Pharmacophore-based screening has become common in the field of computer-assisted drug design (CADD). The success stories achieved with traditional pharmacophore modeling have led many groups to look at ways of describing molecules in a similar way without the need for alignment or derivation of single pharmacophores. The pharmacophore represents the key elements of a protein–ligand interaction and thus a pharmacophore-based descriptor attempts to describe molecules based on their biology rather than their chemistry. The standard 2D similarity measures based around daylight fingerprints or ISIS keys group compounds are based on common chemistry. Pharmacophore-based descriptors attempt to move away from this chemistry-biased representation. Compounds similar in a pharmacophore space do not need to look similar in the chemical space.

The pharmacophore concept is based on the kind of interactions observed in molecular recognition: hydrogen bonding, charge-charge and hydrophobic interactions. A pharmacophore is a set of functional group types in a spatial arrangement that represents the interactions made in common by a set of small ligands with a protein receptor. The pharmacophore-based screening methodology can be used in cases where only the active ligands are known, without any knowledge of the X-ray structure of the protein/enzyme to which it binds and can be applied to large datasets in high-throughput screening (HTS) applications.

Pharmacophore fingerprints represent an extension of this approach whereby a basis set of pharmacophores is generated by enumerating all pharmacophoric types with the corresponding distances between them within a specified range. The 3D fingerprint for a molecule is defined as the collection of all combina-

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

tions of three pharmacophoric features (three-point) and four pharmacophoric features (four-point) in 3D space for all conformers. Each multiplet is characterized by a set of feature types and the corresponding inter-feature distances. The concept has been described previously [1–7] and applications to SAR have been explored in atom-paired descriptors [8].

9.2 Applications of 3D Pharmacophore Fingerprints

The use of 3D pharmacophore fingerprints in CADD applications can be broadly classified into the areas of design of combinatorial focused/diverse libraries, analyzing ligand–protein interactions and virtual HTS (vHTS) and protein selectivity.

9.2.1 Focused/Diverse Library Design Using Pharmacophore Fingerprints

Pickett and co-workers described pharmacophore-derived query (PDQ) as a novel methodology for diversity analysis based on three-point pharmacophores, expressed by a compound, as a descriptor [3]. The method considers both shape and property for diversity calculations and allows for conformational flexibility. The method uses 3D conformers to account for shape and important drug–receptor interactions such as hydrogen-bond donor, hydrogen-bond acceptor, acid, base, aromatic center and hydrophobe to account for property. The distance ranges (2–24 Å) covering most expected pharmacophore sizes were used. Pickett et al. considered the conformational flexibility using the ChemDBS-3D search engine. The PDQ method profiles the final structures of the library in 3D. This also makes the technique suitable for analyzing compound collections, which have not been constructed in a combinatorial sense for, e.g., corporate databases or collections of compounds available for purchase. The method can also be used as a preliminary design filter on building blocks or derivatives of the building blocks.

The DIVSEL program was developed by Pickett et al. for combinatorial reagent selection using three-point pharmacophores as the descriptor for similarity calculations [2]. The algorithm starts by selecting the compound most dissimilar to the others in the set and then iteratively selects compounds most dissimilar to those already selected. DIVSEL was used to select a set of carboxylic acids from a collection of 1100 monocarboxylic acids for an amide library, based on the pharmacophoric diversity of the products. Eleven diverse amines were selected based on pharmacophoric diversity. A virtual library of 12100 amides was constructed from the 11 amines and 1100 carboxylic acids. The DIVSEL program used the pharmacophore fingerprints for the product virtual library to select a diverse set of the carboxylic acids. The products of 90 acids with the 11 amines selected with DIVSEL covered 85% of the three-point pharmacophores represented by the entire 12100 compound virtual library.

Davies and Briant proposed a procedure for selecting reagents that exhibit most of the pharmacophores exhibited by the entire set of molecules based on their frequency in a set of combinatorial products chosen to maximize the number of different three-point pharmacophores covered [9]. Several groups have developed approaches for combinatorial reagent selection by coupling various pharmacophore diversity-based scoring functions with stochastic [10, 11] and genetic algorithms [12] using three- and four-point pharmacophore fingerprints [13, 14]. McGregor and Muskal used PharmPrint fingerprints [5] and principal component analysis for the analysis and design of virtual combinatorial libraries using common scaffolds and building blocks [6].

Mason et al. [15] described a method for measuring molecular similarity and diversity using four-point 3D multiple potential pharmacophores and a modified similarity measure for application to ligand–ligand and ligand–receptor interactions. The use of four- instead of three-point pharmacophores added to the shape information and resolution, including the ability to distinguish chirality. This method was applied in the design of combinatorial libraries for 7-transmembrane G-protein-coupled receptors around a "privileged" substructure where, in a four-point pharmacophore, one of the points was forced to be a "special" feature associated with the "privileged" substructure. This allowed for the design of libraries that optimized both the coverage of pharmacophoric shapes found in the known active ligands and exploring new diversity with the option of focusing around the "privileged" substructure.

9.2.2
Analyzing Protein–Ligand Interactions Using Pharmacophore Fingerprints

When a protein X-ray structure is available, the Design in Receptor [13, 16] (DiR Chem-X module; Oxford Molecular) method utilized both pharmacophore and shape information obtained from X-ray structures of proteins. The functional groups present in the binding sites were mapped and complementary features were placed within the binding site. A pharmacophore fingerprint based on these complementary features was then calculated. Compounds were docked into the binding site and the docked orientations were scored, based on the number of pharmacophore hypotheses that they matched. Docked orientations that have unfavorable steric contacts with the protein were rejected. Mason and Beno [13] used DiR to rank combinatorial reagents for a library based on the Ugi condensation reaction [17] for the factor Xa binding site. Using four-point pharmacophore keys, Mason et al. [15] addressed the issue of enzyme selectivity for thrombin, factor Xa and trypsin. Receptor similarity based on the number of common four-point pharmacophore keys for each ligand–receptor pair seemed to enhance the resolution for enzyme selectivity compared with three-point pharmacophore keys.

Deng et al. [18] recently described an approach to representing and analyzing 3D protein-ligand binding interactions. The Structural Interaction Fingerprint (SIFt; see also Chapter 10) method represents a ligand by the interactions it undergoes in the binding site of a protein. Using seven bits per binding-site residue to represent seven different types of interaction, interaction fingerprint translates 3D structural binding information from a protein–ligand complex into a 1D binary string. Although this requires knowledge of the binding mode of each of the ligands with a common protein, the method has been used in post-processing the output from a docking study and as a filter during a virtual chemical library screening process. The SIFts were clustered using a hierarchical clustering algorithm and Tanimoto similarity coefficient for a set of ligands for a particular protein. This allowed the ligands to be grouped into similar binding modes and the crystallographically observed binding mode was identified. The SIFt method also demonstrated a good database enrichment performance in a virtual library screen for p38 inhibitors, outperforming the scoring functions ChemScore [19] and PMF [20]. More recently, Chuaqui et al. [21] introduced interaction profiling (p-SIFt) for the analysis of protein–inhibitor complexes. They used p-SIFts as a target-specific scoring function by comparing the p-SIFts of compounds with the target-specific group of active inhibitors. This virtual screening method was applied to p38 and CDK2. The authors noted that this methodology might miss molecules with novel binding modes compared with the reference inhibitor in a virtual screening experiment. On the other hand, p-SIFts can be used to identify molecules with novel binding modes by looking for binding modes dissimilar to the known inhibitors. Based on analysis of ~90 known X-ray crystal structures of protein kinase–inhibitor complexes using SIFt, the proteins were classified into three clusters (p38, CDK2 and other ATP-like molecule-bound clusters). Further, using p-SIFts, the authors were able to show the selective differences in the interaction patterns between these clusters of kinases.

Sharing the basic premise of the above-mentioned SIFt method, Kelly and Mancera [22] developed an expanded fingerprint method for post-processing *in silico* docking and automated ligand generation data. Their method extends SIFt by representing the interactions the atomic level as opposed to the residue level and including measures of the strength of the interactions or their geometric grouping. Classifying automated ligand generation output on the basis of their binding modes with a target protein allowed for both the identification of ligands sharing similar binding modes to known active compounds and the filtering out of those ligands demonstrating unfavorable binding modes. These expanded methods were applied to the post-processing of binding poses generated in a docking study for 220 proteins and to the analysis of ligands generated by an automated ligand generation algorithm for the anthrax edema factor.

9.2.3
Virtual High-throughput Screen (vHTS) and Protein Selectivity

The Fingerprint-based Lead Identification Protocol (FLIP) [23] uses the information about the known or potential active site of a protein to data mine compound collections to select compounds that are likely to bind to the defined active site.

Underlying FLIP is the generation of interaction fingerprints that converts 3D structural binding information into a 1D binary string. Each compound in the collection is also converted to a multiconformer-based 1D binary string. The hits are identified by a similarity coefficient between the fingerprints of the active site and each molecule in the collection. FLIP uses LUDI [24, 25] technology to model protein–ligand interactions through the use of *interaction sites* based only on the protein active site information. The possible features considered are negative charge, positive charge, negative ionizable, positive ionizable, hydrogen-bond donors and projection point, hydrogen-bond acceptors and projection point, ring aromatic and projection point and hydrophobic groups. The features present in the 3D pharmacophore fingerprint uses *catFeatures* as implemented in the Catalyst software [26]. Features and inter-feature distances feature files generated by *catFeatures* are mapped to a grid to generate a 3D fingerprint file. Figure 9.1 shows an example of mapping a three-point feature to a grid.

The three features F1, F2 and F3 from a molecule form a triangle and the three sides of this triangle are *d*1 (F1–F2), *d*2 (F2–F3) and *d*3 (F1–F3) with *d*3 > *d*2 > *d*1. This information is mapped on to a grid. First F1 is placed at the origin (0,0) since it is the feature at which the minimal and maximal feature distances meet. F2 is the feature of minimal distance from feature F1 and lies on the *X*-axis (0,A). Once the coordinates of F1 and F2 have been assigned, it is easy to extrapolate the coordinates for F3 (B,C). This is represented by the pharmacophoric index *P* = *F* (F1,F2,F3,A,B,C).

In the Cerius^2 [27] modeling package users are able to select different binning schemes. In the current implementation, we have six different types of binning schemes for three-point/four-point pharmacophores: (1) uniform bins, (2) user-defined bins, (3) geometric progressions, (4) exponential progressions, (5) arithmetic progressions and (6) overlap bins. The uniform binning scheme is straightforward based on intuition and its implementation is very easy. However, this simple binning scheme suffers from some notable problems. The problem most often seen is that smaller grid intervals around the distances of interest are not permitted owing to the constraint of uniformity. Another problem often seen is that two very similar pharmacophore keys may be treated as two distinct ones and two distinct pharmacophore keys might be treated as the same pharmacophore. The former

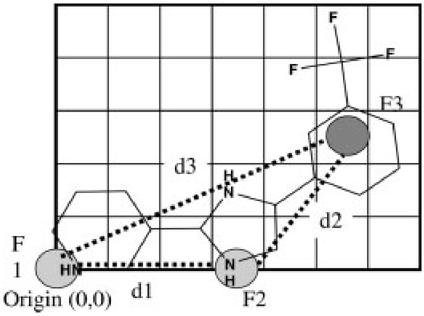


Fig. 9.1 Mapping of a three-point pharmacophore to a grid.

situation could happen when two pharmacophore keys with very similar edge distances happen to fall either side of a bin boundary. The latter situation would happen when the grid resolution is rough. Such problems suggest that the non-regular and overlap binning schemes are necessary. The non-regular binning scheme would allow more granularity around the distances of interest for any specific problems. The functionality of the overlap bin scheme would allow the elimination, or at least reduction, of edge effects in which two pharmacophores with very similar edge distances that happen to fall either side of a bin boundary are currently considered to be different.

The 3D fingerprints generated can be in either binary or non-binary form. The binary fingerprint maps the presence or absence of a particular three- or four-point pharmacophoric feature. A non-binary fingerprint, in addition to the presence or absence of a fingerprint, also keeps track of the occurrence counts of the fingerprints. By recording the total occurrence of each fingerprint over the conformations, comparisons between molecules may be more discriminating. It is possible to edit manually the interaction map generated by LUDI to remove any "noise" and edit some of the interaction site definitions. Either the edited interaction map can be used "as is" or further clustering of the interaction sites can be done. The clusters can also be edited and the user can reassign interactions points to other clusters. If the interaction sites are clustered then the cluster centers are used to represent the interaction map. Hence either all the interaction site features or the cluster center features in the active site can be used to represent the receptor interaction map. The LUDI interaction map is then converted to a Catalyst feature file. Every possible combination of three features (three-point) and four features (four-point) and their corresponding inter-feature distance is considered. This is then mapped to a grid and a 3D fingerprint file for the receptor active site is generated.

The source of the compound collection could be varied, such as corporate collection, combinatorial libraries and virtual compound collections. Using a standard SD/SMILES file for a given collection of compounds, a multi-conformer Catalyst database is constructed. The time-consuming step in the FLIP protocol is the construction of Catalyst databases. However, with the Linux OS version of Catalyst, database construction is significantly faster on a Linux cluster. The program CatFeatures, in the Catalyst environment, is used to identify the features present in the molecules. Once the features have been identified, a module within the software Cerius^2 is used to construct the fingerprints for all the molecules as described above [27]. For a given molecule and for all the conformations of the molecule, every possible combination of three features (three-point) and four features (four-point) and their corresponding inter-feature distance is considered. This is then mapped to a grid and 3D fingerprint for the molecule is generated. Figure 9.2 illustrates the steps and programs required to calculate 3D/4D fingerprints for a library or any set of molecules in an SD or smiles file format.

The 3D fingerprint for the receptor is compared against the 3D fingerprint for each molecule in the Catalyst database to select the top percentage of hits. There are two possible ways of comparing 3D fingerprints:

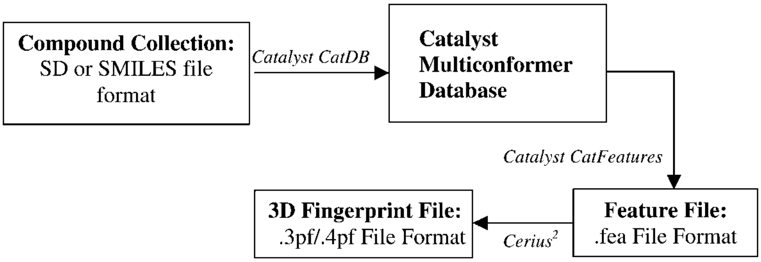

Compound Collection: SD or SMILES file format
Catalyst CatDB
Catalyst Multiconformer Database
Catalyst CatFeatures
Feature File: .fea File Format
Cerius²
3D Fingerprint File: .3pf/.4pf File Format

Fig. 9.2 3D fingerprint generation for virtual libraries.

1. Compute similarity coefficients (Tanimoto, Dice, Ochia, Hamming) between the active site fingerprint and the fingerprint for each molecule in the Catalyst database. The top *N*% of compounds can be selected by ranking the compound collection in descending order based on the similarity coefficient.
2. Another technique uses the fingerprints OnBits metric for comparing 3D fingerprints. It is based on generating a "modal fingerprint" for a set of *N* molecules, in which a bit is *on* if it is present in at least one molecule in the set. The modal 3D fingerprint of the compound collection (candidate library) is compared with the modal fingerprint of the receptor active site (reference library), reporting the number of *on* bits in each library, the number of common bits, the number of *on* bits in the candidate library not present in the reference library and the number of *on* bits in the reference library not present in the candidate library. This method allows one to list the molecules in the candidate library with *on* bits present in the reference library and to select the top *N* molecules from the candidate library with the highest number of common bits with the reference library.

The first method used a similarity metric to select the top percentage of hits and the second method does the selection based only on number of common pharmacophores between the receptor active site fingerprint and the 3D fingerprint for compounds in the virtual library. Both analysis techniques are extremely fast. One of the major advantages of FLIP technology is its throughput.

9.2.3.1 **Application of FLIP Technology**

We present the data for the application of FLIP as a virtual screening tool to identify the actives of fibroblast growth factor (Fgf) receptor (pdb code: 2FGI) [28] that were seeded in a virtual library of random molecules. The FGF receptor is a protein tyrosine kinase. The active site is the hydrophobic ATP-binding site with key hydrogen-bonding interactions. The conserved Lys514 moves significantly upon ligand binding to provide access to an adjacent pocket that can be used to design selective molecules. The active ligands for FGF are shown in Fig. 9.3.

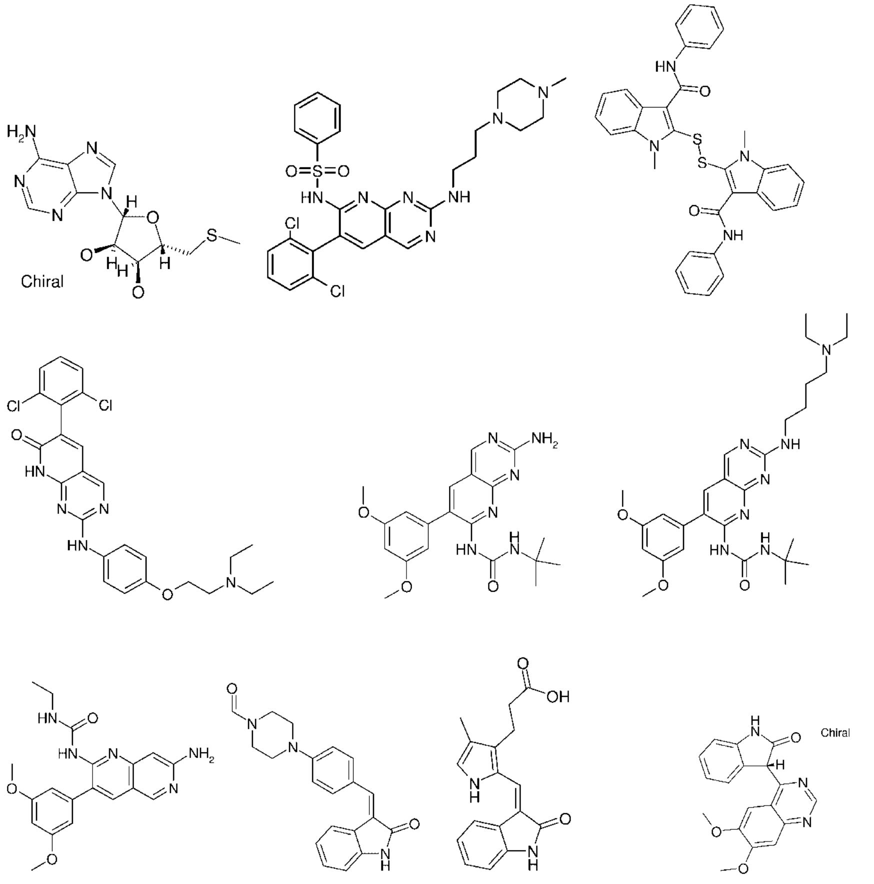


Fig. 9.3 The active ligands for FGF.

The Advailable Chemicals Directory (ACD v. 2000-1, Molecular Design, San Leandro, CA, USA) was first filtered in order to eliminate chemical reagents [29, 30], inorganic compounds and molecules with unsuitable molecular weights (<250 or >550). Out of the 80000 remaining molecules, the 1000 most diverse structures using the maxmin algorithm in Cerius^2 [27] were selected and their 3D coordinates were generated using Corina [32]. Hydrogen atoms and Gasteiger–Marsili [31] atomic charges were then added using Cerius^2 [27]. The known active ligands for FGF (Fig. 9.3) were prepared using the procedure described above starting from 2D ISIS structures. Special care was taken to retain the correct ionization states of ionizable groups (amines, amidines, carboxylic acids, etc.) at a physiological pH of 7.4. These new active structures were then individually added to 1000 random molecules previously selected to generate a virtual library.

The parameters that were varied in order to study their effects on data mining of the active molecules included (1) number of conformations of virtual com-

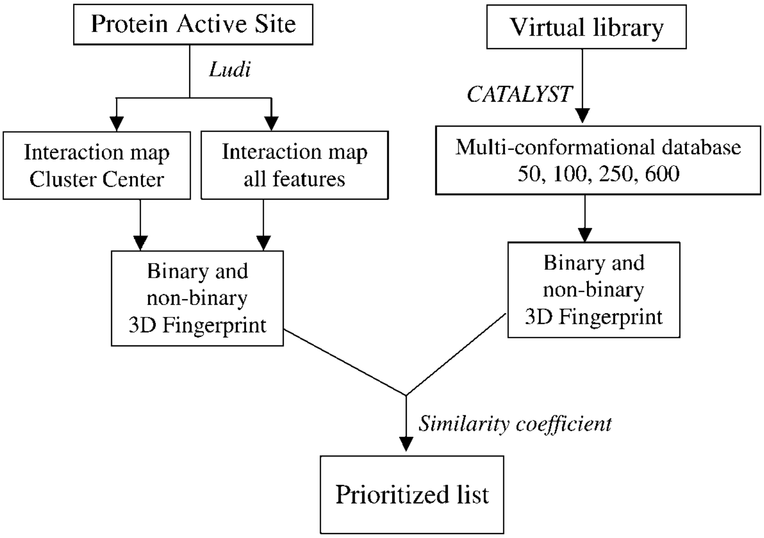

Protein Active Site
Ludi
Interaction map Cluster Center
Interaction map all features
Binary and non-binary 3D Fingerprint
Virtual library
CATALYST
Multi-conformational database 50, 100, 250, 600
Binary and non-binary 3D Fingerprint
Similarity coefficient
Prioritized list

Fig. 9.4 Protocol for varying different parameters used for FLIP analysis.

pounds, (2) cluster center feature versus all features for the interaction site model, (3) binary versus non-binary fingerprints and (4) similarity coefficients. Figure 9.4 illustrates the protocol and total schemes generated by using different combinations of the above-mentioned factors. A default grid size of 10 Å with a grid space of 2 Å and a minimum separation of 2.5 Å between the pharmacophoric features was used.

In order to study the effect of number of conformations for each molecule in the Catalyst database on the FLIP protocol, four Catalyst databases with different maximum numbers of conformations, 50, 100, 250 and 600, were generated for the virtual library. Conformations for ligands in Catalyst database format were generated using the FAST method of catConf (built-in Catalyst conformer generation engine). The binary and non-binary 3D fingerprints of the virtual library were generated. For the protein, the interaction map was generated and manually edited to remove interaction points outside the binding cavity of the ligand. The interaction site was clustered using a hierarchical clustering method. Both clustered and non-clustered interaction maps were used to generate a binary and non-binary 3D fingerprint. Similarity coefficients between the enzyme and the virtual library were calculated. The similarity coefficients calculated included Tanimoto, Ochiai, Dice and Hamming.

The Ochiai similarity coefficient with 250 maximum conformations provided the best retrieval rates of the actives from the virtual library (Fig. 9.5).

The non-binary all feature combination of the 3D fingerprint and interaction feature was able to retrieve ~90% of the actives on screening 20% of the database. The Ochiai coefficient was slightly better than Tanimoto and significantly better than Dice and Hamming in retrieving actives. All features for the enzyme

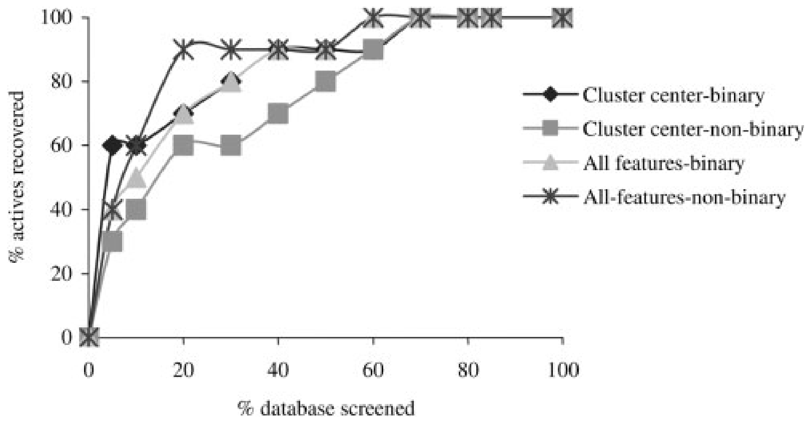


Fig. 9.5 Hit rates for FGF using the Ochiai similarity coefficient with the maximum conformers parameter set at 250.

active site in combination with the non-binary 3D fingerprint ranked actives higher than the random compounds in the virtual library. The combination of features and 3D fingerprints such as cluster center–binary and all feature–binary perform were slightly worse (70% of the actives on screening 20% of the database) than the cluster center–binary in providing enrichment. The combination of cluster center–non-binary is significantly lower in retrieving actives (60% of the actives retrieved on screening 20% of the database).

In this exercise for the FGF target on the seeded database, we retrieved 70–80% of actives by screening 20–30% of the database. The best combination for retrieving maximum actives when 5% of the database was screened was "non-binary–all feature–Ochiai similarity coefficient". The effect of conformations is dependent on the flexibility of the molecules in the database. Based on our analysis, the maximum number of conformations set to 100 was sufficient for retrieving 70–80% of the actives. The hit list is 12–16 times enriched with respect to random selection from the database on screening only 5% of the database.

This methodology can also potentially be used for addressing the issue of protein selectivity. The advantage of FLIP is that comparisons of active sites can be made without prior alignment of the active sites. The 3D fingerprints for the active sites of proteins can be compared through similarity coefficients to identify proteins that may or may not be homologous in their sequence, but may have similarities in their 3D structure. Also, by generating a similarity matrix with the protein targets as columns and molecules as rows, one can potentially address the issue of selectivity against an enzyme class.

Preliminary data on the selectivity issue are shown in Fig. 9.6. The virtual library comprised 1000 molecules from ACD (ACD v. 2000-1, Molecular Design) as described previously. Ten actives for FGF and CDK2 proteins were added to the random molecules to generate the virtual library for screening. The goal of the experiment was to isolate selectively FGF actives from the virtual library

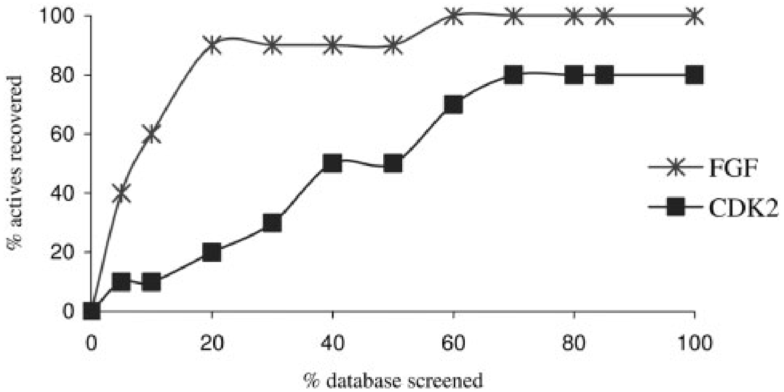


Fig. 9.6 Retrieval rates using three-point fingerprints for FGF and CDK2 actives using FGF as the target protein.

using FGF as the target protein. Using the non-binary–all feature combination along with Ochiai coefficient and 250 maximum conformers, the virtual library was prioritized and the actives for FGF and CDK2 were identified. The enrichment data are shown in Fig. 9.6. It can be seen that when FGF is the target protein, the rates of retrieval of CDK2 ligands using 3D fingerprints was similar to random selection of these actives. However, using the FGF as the target protein, 90% of the FGF actives were retrieved on screening 20% of the database. Studies are ongoing in this area to validate further some of the initial findings. Mason et al. pointed out in that four-point fingerprints were more discerning in identifying ligand–protein selectivity [15]. Using four-point fingerprints in the FLIP protocol may increase the information content in the fingerprints.

Some initial studies using a shape-based filter seemed to increase the hit rates in FLIP (unpublished data). This is not surprising as the conformers/molecules that do not fit into the active site are filtered out. This perhaps increases the “signal-to-noise” ratio. FLIP can also potentially be used as a preprocessor to a more time-consuming docking-based virtual screening. FLIP combined with a 2D descriptor-based sequential screening method has been shown to increase significantly the hit rates of actives in a virtual screening experiment [23, 33]. This seems to indicate that FLIP can be used for “scaffold” identification followed by 2D methods for exploring regions around the active scaffold(s).

9.3 Conclusion

With increasing numbers of X-ray structures being solved, the 3D information can be exploited for ligand design and optimization. There is a need for fast methods for structure-based virtual screening. The recent use of machine clusters and porting of software codes to Linux and Windows platforms has contributed to a significant speeding up of conformer database building and virtual screening using docking and pharmacophore based screens. An alternative method for structure-based virtual screening is the use of pharmacophore fingerprints. FLIP can be used for the rapid identification of leads from a database for HTS. The speed of this method (2–3 s per molecule) is an obvious advantage. It can be used on a protein target for which either a crystal or NMR structure or homology model of the protein is available. The FLIP methodology relies on commonality between fingerprints of the active site and virtual library compounds. This can be used as a pre-filter to a more time-consuming docking study.

3D fingerprints are very easy to compute and as they are computed in distance space there is no dependence on target alignment. Hence this methodology can be used for the rapid virtual screening of compounds and to compare hits for related targets. This provides us with a tool to compare both related and unrelated proteins that may have some similarities in their active site. For example, in a virtual screening experiment, a prioritized hit list for a protein active site can be compared with the prioritized hit list for the enzyme implicated in some side-effects or metabolism. Similarity matrices generated for a panel of related enzymes may potentially be used to address issues of selectivity early in the drug discovery process.

The hits identified by screening a small percentage of the database can be followed up using 2D methods and/or 3D pharmacophore-based further exploration of the database to retrieve more actives in a sequential screening fashion. The goal is to minimize the number of compounds screened and maximize the number of actives retrieved at the end of the screening rounds such that the relevant chemical space is explored with an optimal use of resources.

Acknowledgments

The authors would like to acknowledge Dr. Marvin Waldman and Dr. Chien-Tin Lin of Accelrys for exciting scientific discussions and for information on algorithms on fingerprints. The authors would also like to acknowledge Dr. Frank Brown of Johnson & Johnson Pharmaceutical Research and Development for discussions on sequential screening.

References

1 Good, A. C., Kuntz, I. D., Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373–379.

2 Pickett, S. D., Luttmann, C., Guerin, V., Laoui, A., James, E., DIVSEL and COMPLIB strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144–150.

3 Pickett, S. D., Mason, J. S., McLay, I. M., Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci* **1996**, *36*, 1214–1223.

4 Mason, J. S., Pickett, S. D., Partition-based selection. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 85–114.

5 McGregor, M. J., Muskal, S. M., Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.

6 McGregor, M. J., Muskal, S. M., Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117–125.

7 McGregor, M. J., Muskal, S. M. Pharmacophore fingerprinting in QSAR and primary library design. WO 0025106, **2000**.

8 Chen, X., Rusinko, A., Young, S. S., Recursive partitioning analysis of a large structure–activity dataset using three dimensional descriptors. *J. Chem. Inf. Comput. Sci* **1998**, *38*, 1054–1062.

9 Davies, K., Briant, C. Combinatorial chemistry library design using pharmacophore diversity. http://www.netsci.org/Science/Combichem/feature05.html.

10 Agrafiotis, D. K., Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841–851.

11 Zheng, W., Cho, S. J., Waller, C. L., Tropsha, A., Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: a novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738–746.

12 Brown, R. D., Clark, D. E., Genetic diversity: applications of evolutionary algorithms to combinatorial library design. *Expert Opin. Ther. Pat.* **1998**, *8*, 1447–1460.

13 Mason, J. S., Beno, B. R., Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. *J. Mol. Graph. Model.* **2000**, *18*, 438–451.

14 Good, A. C., Lewis, R. A., New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926–3936.

15 Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., Labaudiniere, R. F., New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing priviledge substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

16 Murray, C. M., Cato, S. J., Design of libraries to explore receptor sites. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 46–50.

17 Ugi, I., Steinbruckner, C., Isonitriles. II. Reaction of isonitriles with carbonyl compounds, amines and hydrazoic acid. *Chem. Ber.* **1961**, *94*, 734–742.

18 Deng, Z., Chuaqui, C., Singh, J., Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

19 Eldridge, M., Murray, C. W., Auton, T. A., Paolini, G. V., Lee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

20 Muegge, I., Martin, Y. C., General fast scoring function for protein–ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.

21 Chuaqui, C., Deng, Z., Singh, J., Interaction profiles of protein kinase–inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121–133.

22 Kelly, M. D., Mancera R. L., Expanded interaction fingerprint method for anlaysing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.

23 Kulkarni, A., Karnachi, P. S., 3D fingerprint-based lead identification protocol (FLIP): application to structure based design and data mining, manuscript in preparation.

24 Böhm, H. J., LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol.Des.* **1992**, *6*, 593–606.

25 Böhm, H. J., The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol.Des.* **1992**, *6*, 61–78.

26 *Catalyst 4.9*, Accelrys, San Diego, CA.

27 *Cerius^2 4.9*, Accelrys, San Diego, CA.

28 Mohammadi, M., Froum, S., Hamby, J. M., Schroeder, M. C., Panek, R. L., Lu, G. H., Eliseenkova, A. V., Green, D., Schlessinger, J., Hubbard, S. R., Crystal structure of an angiogenesis inhibitor bound to the FGF receptor tyrosine kinase domain. *EMBO J.* **1998**, *17*, 5896–5904.

29 Oprea, T., Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.

30 Walters, W. P., Stahl, M. T., Murcko, M. A, Virtual screening – an overview. *Drug Discov. Today* **1998**, *3*, 160–178.

31 Gasteiger, J., Marsili, M., Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.

32 Gasteiger, J., Rudolph, C., Sadowski, J., Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Method.* **1990**, *3*, 537–547.

33 Karnachi, P. S., Brown, F. K., Practical approaches to efficient screening information-rich screening protocol. *J. Biomol. Screen.* **2004**, *9*, 678–686.

10
SIFt: Analysis, Organization and Database Mining for Protein-Inhibitor Complexes. Application to Protein Kinase Inhibitors

Juswinder Singh, Zhan Deng, and Claudio Chuaqui

10.1
Introduction

Structure-based drug design is a critical component to lead discovery and optimization within the pharmaceutical industry. Nowadays, an increasing number of drug targets are amenable to structure determination using crystallography and NMR analysis and the past decade has witnessed an explosion in the number of three-dimensional protein–small molecule structures from both experimental and *in silico* approaches. With the recent development of high-throughput X-ray crystallography, the total number of structures will grow at an even greater rate. In parallel with the growth of experimentally determined structures, a plethora of structural information is also being generated in the rational drug discovery process. A significant challenge exists in the industry to leverage better this wealth of information.

The optimization of a small molecule at its binding site requires a detailed understanding of the intermolecular interactions within the protein–small molecule complex. Visual inspection using computer graphics is powerful at analyzing small numbers of complexes, but it becomes intractable when the number to be analyzed is very large, as is the case of the results generated from virtual library screening. The use of scoring functions to evaluate the energetics and rank libraries of virtual compounds is the primary solution to filtering large datasets.

The protein kinase family is emerging as an exciting class of targets for drug discovery [1]. Protein kinases play a pivotal role in control of cellular signaling and are involved in proliferation, differentiation and metabolism. Aberrant signaling of protein kinases has been identified in a wide range of diseases including cancer, inflammation and neurodegeneration [2, 3]. The protein kinase family exemplifies the challenges faced with the large amount of structural data being generated not only on specific drug targets, but also at the gene family level [4]. The first crystal structure of a protein kinase was solved in 1991 of the cAMP-dependent Ser/Thr protein kinase in complex with a peptide inhibitor and ATP [5]. Since then, over 120 structures of complexes have been deposited in the public databanks in complexes with small molecules bound and probably

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

a much larger number of structures are available within pharmaceutical companies.

The large amount of complex structural information requires a new method to help us analyze better the binding interactions between proteins and ligands. Ideally, such a new method should be able to facilitate the following tasks: (1) data visualization, to allow easy interpretation of the binding interactions; (2) data organization, to organize and cluster the structures in a meaningful way; (3) data analysis, to allow the comparison and profiling of the binding interactions in different structures; and (4) data mining, to help search for structures that contain key interactions or specific features. In addition, it is desirable that the method be simple and generic.

In this chapter, we describe a simple and robust approach for representing and analyzing three-dimensional protein–ligand complexes called SIFt (Structural Interaction Fingerprint) [6, 7]. We will show how this method can be applied to organizing and analyzing the structural information within the protein kinase family and also how this can be applied to virtual screening for inhibitors.

10.2
How to Generate a SIFt Fingerprint

A structural interaction fingerprint is a 1D binary representation of the interaction patterns from a 3D protein-inhibitor complex. The fingerprint representation of the interaction patterns is compact and allows for rapid clustering and analysis of massive numbers of complexes.

The first step in the construction of a SIFt interaction fingerprint is to identify a list of binding site residues that are common in all complex structures being studied (Fig. 10.1). Here, the ligand binding site is defined as the union of all the residues that are in contact with any ligand molecules in any of the structures in the group. The resulting panel of ligand binding site residues, which act as a mask covering all of the interactions occurring between the protein and the ligands, is then used as the common reference frame to construct the interaction fingerprints.

For a group of structures involving the same target protein (e.g. docking results), the ligand binding site is defined as the list of residues comprising the union of all residues involved in ligand binding over the entire library of structures. For the protein kinase–ligand complex structures, however, as the target proteins involved are different, a sequence alignment of the protein binding sites is needed which can be based on sequence and/or structural information.

After all the ligand binding site residues have been identified and all the protein–ligand intermolecular interactions have been calculated, the next step is to classify these interactions. By default, seven different types of interactions occurring at each binding residue are extracted and classified using the programs AREAIMOL [8] from the CCP4 suite [9] and the hydrogen-bonding program HBPLUS [10]. They include (1) whether or not it is in contact with the ligand;

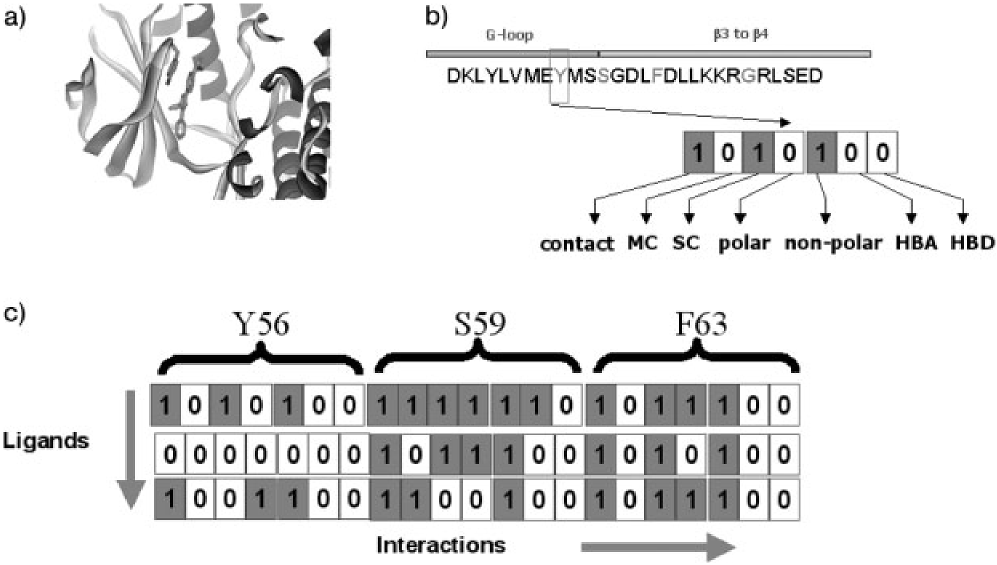


Fig. 10.1 The procedure used in the generation of the SIFt fingerprint. (a) 3D binding site of a kinase with a small molecule inhibitor bound. (b) sequence of the positions in the binding site in contact with the small molecule, together with their location in the structure of the kinase (g-loop and 3 to 4). Each binding site position is then represented by a bitstring, with each bit switched to "on" depending on whether it is involved in a contact, whether the contact is with the main chain (MC) or side-chain (SC) of the protein and if the interaction is polar, apolar, hydrogen-bond acceptor/donor. (c) Concatenation of all bitstrings for each binding site residue. This process is repeated for all ligands.

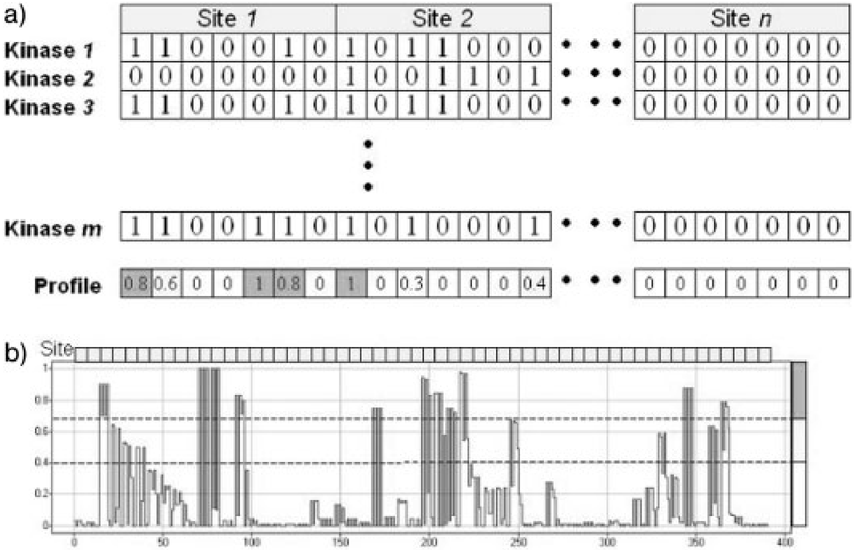
(2) whether or not any main-chain atom (MC) is involved in the contact; (3) whether or not any side-chain (SC) atom is involved in the binding; (4) whether or not a polar interaction is involved; (5) whether or not a non-polar interaction is involved; (6) whether or not the residue provides hydrogen bond acceptor(s); and (7) whether or not it provides hydrogen-bond donor(s). By doing so, each residue is represented by a seven-bit-long bitstring. The whole interaction fingerprint of the complex is finally constructed by sequentially concatenating the binding bitstring of each binding site residue together, according to ascending residue number order. Therefore, interaction fingerprints are of the same length and each bit in the fingerprint represents the presence or absence of a particular interaction at a particular binding site.

Our current implementation of SIFt uses seven bits for each binding site residue, representing seven different types of interactions. The richness of information can be improved by incorporating more bits representing other types of binding interactions, using sub-residue portions instead of the whole residue as the basic unit, taking solvent molecules into consideration or substituting the binary bits with scaled numerical data that reflect the strength and energetics of the interactions. Such an enriched SIFt provides a "higher resolution" picture of the complex. On the other hand, in situations where computational speed is a critical issue, we may construct “lower resolution” SIFts using fewer bits.

10.3 Profile-based SIFts

We have developed a profile-based approach termed p-SIFt [7] that enables us to describe the conservation of interactions between a set of protein–ligand receptor complexes. The use of profiles provides a sensitive means to compare and contrast multiple inhibitors binding to a drug target. A structural interaction fingerprint profile (p-SIFt) represents the degree to which interactions are conserved across a set of ligand–receptor complexes. The p-SIFt, $P(r)$, is derived from an array, denoted below as b, of SIFt patterns and its derivation from a set of SIFts is shown in Fig. 10.2.

The array has a length N for the total number of protein–ligand complexes and a width K of SIFt fingerprints bits. The value of each element of $P(r)$ is derived by averaging the elements in each column of the SIFt matrix, yielding a



	Site 1							Site 2								Site n						
Kinase 1	1	1	0	0	0	1	0	1	0	1	1	0	0	0	• • •	0	0	0	0	0	0	0
Kinase 2	0	0	0	0	0	0	0	1	0	0	1	1	0	1	• • •	0	0	0	0	0	0	0
Kinase 3	1	1	0	0	0	1	0	1	0	1	1	0	0	0	• • •	0	0	0	0	0	0	0
⋮																						
Kinase m	1	1	0	0	1	1	0	1	0	1	0	0	0	1	• • •	0	0	0	0	0	0	0
Profile	0.8	0.6	0	0	1	0.8	0	1	0	0.3	0	0	0	0.4	• • •	0	0	0	0	0	0	0

Fig. 10.2 The procedure used to generate a p-SIFt from a set of SIFts is illustrated in (a). The profile shown corresponds to the p-SIFt generated from the 93 kinase structures using all seven bits to compute the SIFts shown in (b). The p-SIFt is annotated with a topmost bar delineating the general kinase structural features for that portion of the fingerprint; the bar below consists of alternating blocks corresponding to each residue (site in the uniform PKA numbering scheme) in the kinase used to construct the fingerprint; the third bar consists of blocks for each bit representing the interaction features at that site.

numerical interaction frequency that varies from 0 to 1 for unobserved to fully conserved, respectively. The SIFt array, b, and resulting $P(r)$ are given by

$$b = \begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \dots & b_{1,K} \\ b_{2,1} & b_{2,2} & b_{2,3} & \dots & b_{2,K} \\ & & \vdots & & \\ b_{N,1} & b_{N,2} & b_{N,3} & \dots & b_{N,K} \end{pmatrix}$$

and

$$P(r) = (P_1 \quad P_2 \quad P_4 \quad P_K)$$

where $b_{i,r}$ is the binary bit value in the SIFt $i=1,N$ at position $r=1,K$. The values in the p-SIFt at position r are given by

$$P(r) = \sum_{i=1}^{N} b_{i,r} \Big/ N$$

Because the interaction fingerprint represents the binding mode of a ligand to a target protein, similar fingerprints imply that the corresponding ligands make similar interactions with the protein.

We used the Tanimoto coefficient [11] to measure the similarity between two SIFts, between two p-SIFts and between a SIFt and a p-SIFt. A set of SIFt patterns can be clustered using the Tanimoto similarity measure by applying standard hierarchical clustering algorithms [12, 13]. The statistical Z score was employed to measure how significant the similarity between a SIFt and a target p-SIFt (i.e. a group of structures) is above a certain background. The Z score is an indication of how many standard deviations and in what direction, an observation deviates from the background distribution's mean value.

10.4
SIFt and the Analysis of Protein Kinase – Inhibitor Complexes

As of April 2002, 93 kinase structures had been deposited in the public databanks since the first protein kinase structure was determined in 1991 by Taylor's group (5). This collection of structures covers 14 different protein kinase subfamilies and 54 unique kinase–small molecule ligands/inhibitors (Z. Deng et al., unpublished results). The catalytic domain of the protein kinases adopts a canonical fold consisting of a small N-terminal primarily β-sheet and a large C-terminal helical domain. These structures have revealed details of how ATP and substrates bind to the kinase domain and provided a valuable insight into how phosphorylation can regulate their activity [14–16]. The ATP binding site is located between the N and C-terminal domains. Relatively few substrate complexes exist and the data so far suggest a shallower binding site for substrates than for ATP. The majority of known kinase inhibitors bind to the ATP site and few inhibitors have been described that target the substrate binding site.

10.5 Canonical Protein – Small Molecule Interactions in the Kinase Family

We have described the use of SIFt to analyze the conservation of contacts between protein kinase–inhibitor complexes [6, 7]. Our analysis revealed 56 residues from the kinase catalytic domain that are involved protein–inhibitor interactions across the kinase family. These residues include (in PKA numbering): the glycine-rich loop, which is a conserved signature of the family and plays a role in binding of ATP (47–57), the hinge region, which is located between the N- and C-terminal domains and plays a role in hydrogen bonding the adenine moiety of ATP (123–125, 127, 130), b3 (70, 72), the hydrophobic pocket region (95, 104, 105, 118, 119) and the "gatekeeper" residue whose size determines inhibitor access (120) and the activation segment which is targeted by several inhibitors that stabilize the inactive conformation of the kinase.

We determined the degree of conservation across the kinase inhibitor complexes by exploring the frequency of contacts at each of the 56 positions in the 93 complexes. We classed positions using the following ranges: conserved 0.7, 0.4 intermediate <0.7, variable <0.4. Approximately 20% of the contact interactions are conserved over the 93 structures as a whole, 11% are intermediate in conservation and 69% are variable. The conserved interactions are denoted in Table 10.1 by the highlighted annotations and comprise a canonical set of interactions that are evidently fundamental for kinase binding at the ATP site. The canonical interactions are common to all inhibitors and may be used as a basic kinase-like binding filter in virtual screening.

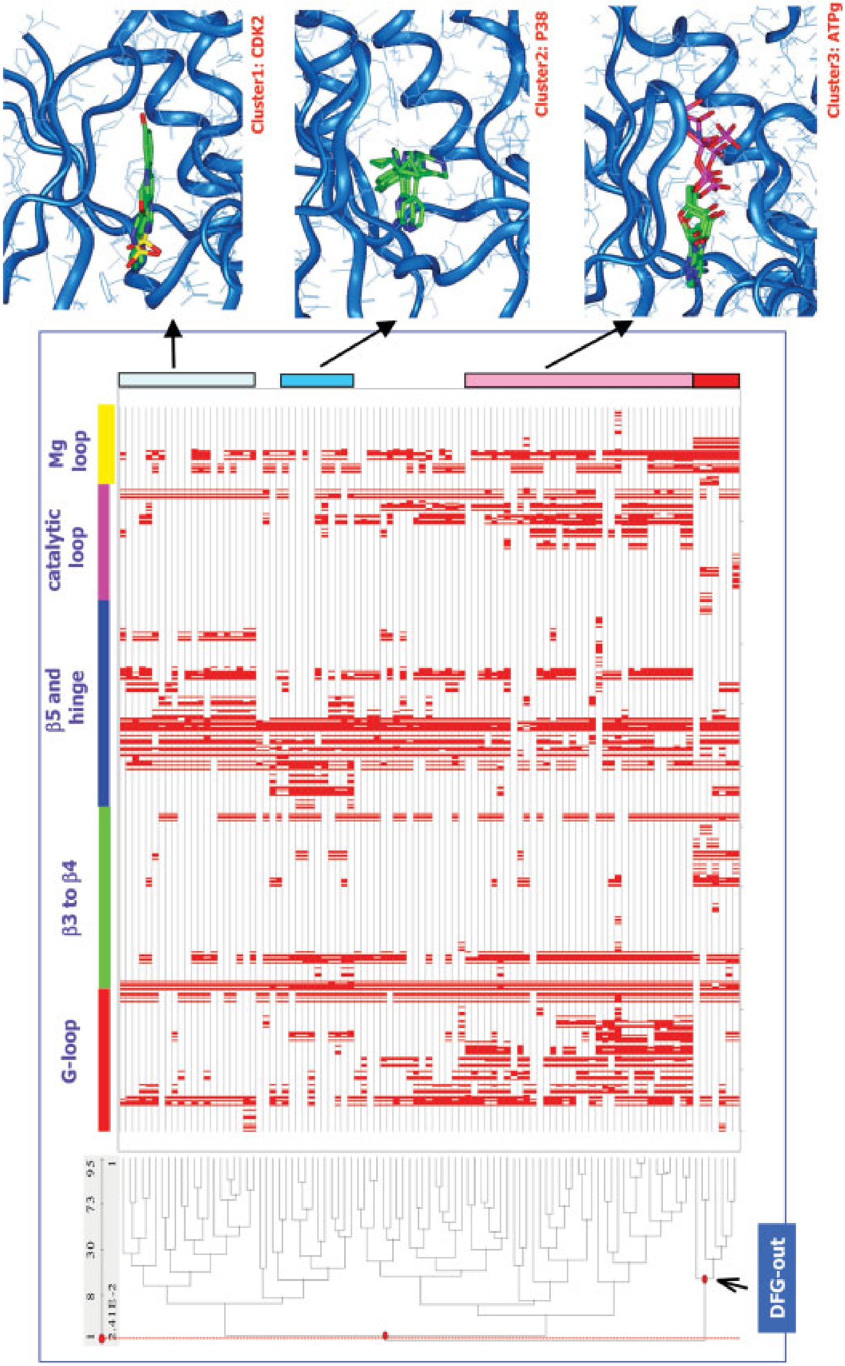
10.6 Clustering of Kinase Inhibitors Based on Interaction Fingerprints

Using SIFt, we can cluster together kinase–inhibitor complexes showing similar interaction patterns. The approach involves computing the similarity metric (Tanimoto coefficient) between all of the fingerprints and then using a hierarchical clustering algorithm to group similar interaction fingerprints together. A dendrogram derived from comparison of interaction fingerprints of the 93 protein–inhibitor complexes revealed three major clusters (Fig. 10.3), consisting of ATP and ATP analogs, p38 and CDK2 inhibitors.

The first cluster consists of p38 in complex with pyridinylimidazole inhibitors. The second cluster consists mostly of human CDK2 in complex with different compounds with diverse chemical properties. The third cluster, which does not have a clear-cut boundary, is comprised of different kinases in complex with ATP or ATP analog inhibitors denoted ATPg (GTP, AMPPNP, AMPPCP, AMP, ADP, etc.). Besides these three major clusters, about one-third of the structures are either singletons or form tiny clusters. Interestingly, the three major clusters

Table 10.1 Summary of the raw frequencies observed for contact interactions, where only residues having a frequency >0.4 for any subgroup are listed. Residues are colored according to interaction conservation: conserved ≥0.7 (italic), 0.4 intermediate <0.7 (bold), variable <0.4 (bold-italic). Highlighted cells in the columns indicate that the frequency was defined as conserved (≥0.7) for all subgroups independently. Wherever possible, information on the context of the interaction in binding ATP or inhibitors is included as an annotation

PKA No.	Raw interaction frequency					2-Structure	Interaction context
	All	ATP	CDK2	p38	Non-ATP		
49	*0.9*	*0.9*	*0.3*	0.4	*0.9*	Gly-rich Lp	ATP; hydrophobic contact with adenine
50	**0.6**	*0.9*	***0.3***	***0.2***	**0.5**	Gly-rich Lp	ATP; ribose
51	**0.5**	*0.7*	***0.3***	***0.1***	***0.4***	Gly-rich Lp	ATP; ribose
52	**0.5**	*0.9*	**0.4**	***0.0***	***0.3***	Gly-rich Lp	ATP; phosphate
53	***0.4***	*0.7*	***0.2***	***0.1***	***0.1***	Gly-rich Lp	ATP; phosphate
54	***0.3***	**0.5**	***0.2***	*0.9*	***0.2***	Gly-rich Lp	ATP; phosphate
55	***0.2***	**0.5**	***0.1***	***0.0***	***0.1***	Gly-rich Lp	ATP; phosphate
57	*1.0*	*1.0*	*0.7*	*0.8*	*1.0*	Gly-rich Lp	ATP; hydrophobic contact with adenine, ribose, phosphate
70	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	b3	ATP; hydrophobic contact with adenine
72	*0.8*	*0.9*	*0.7*	*1.0*	*0.8*	b3	ATP; phosphate
95	***0.1***	***0.0***	***0.0***	**0.6**	***0.2***	ac	Hydrophobic pocket
104	*0.7*	**0.7**	*0.7*	*0.8*	*0.8*	Lp-ac-a4	ATP; hydrophobic contact with adenine
106	***0.1***	***0.0***	***0.0***	**0.4**	***0.1***	Lp-ac-a4	Hydrophobic pocket
118	***0.2***	***0.0***	***0.0***	*1.0*	***0.3***	b5	Hydrophobic pocket
119	***0.0***	***0.0***	***0.0***	**0.7**	***0.1***	b5	Hydrophobic pocket
120	0.9	0.9	0.9	1.0	1.0	b5	Gatekeeper
121	*0.8*	*0.9*	*1.0*	*1.0*	*0.8*	b5	ATP; hydrogen bond with adenine
122	*0.7*	**0.6**	*1.0*	*1.0*	*0.8*	b5	ATP; hydrophobic contact adenine
123	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*	Hinge	ATP; hydrogen bond adenine
124	***0.3***	***0.0***	**0.6**	**0.4**	**0.5**	Hinge	ATP; adenine water-mediated interaction
125	***0.2***	***0.0***	*0.5*	*0.4*	***0.4***	Hinge	
127	**0.7**	*0.8*	*0.9*	***0.3***	**0.6**	Hinge	ATP; ribose
130	***0.3***	***0.2***	**0.5**	***0.0***	***0.4***	Hinge	ATP; ribose water-mediated interaction
168	***0.2***	**0.5**	***0.2***	***0.0***	***0.1***	Lp-b6-b7	
170	**0.6**	*0.8*	**0.4**	***0.2***	**0.4**	Lp-b6-b7	ATP; ribose
171	***0.3***	**0.4**	***0.4***	***0.0***	***0.3***	Lp-b6-b7	
173	*0.9*	*0.9*	*1.0*	***0.3***	*0.9*	Lp-b6-b7	
182	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	b8	ATP; contact with Mg-loop region
183	**0.6**	**0.5**	***0.3***	**0.4**	*0.7*	b8	ATP; hydrophobic contact with Mg-loop region
184	*0.8*	*0.9*	*0.8*	**0.7**	*0.8*	b8	ATP; contact with Mg-loop region


Cluster1: CDK2
Cluster2: P38
Cluster3: ATPg
Mg loop
catalytic loop
β5 and hinge
β3 to β4
G-loop
DFG-out
1
8
30
73
95
1
2.41E-2

represent different grouping examples of protein–ligand complexes: the first is made up of the same protein and chemically similar compounds, the second group contains the same protein but with a variety of ligands and the third cluster contains different proteins in complex with chemically similar ligands.

10.7 Profile Analysis of ATP, p38 and CDK2 Complexes

The ATP, p38 and CDK2 SIFt clusters represent a set of structures having similar conserved and variable interactions. In order to compare within and between these clusters, we developed a profile-based methodology, p-SIFt. The p-SIFt approach is analogous to profile-based techniques that have proven to be very useful in the analysis and database mining of groups of protein sequences [17] and structures [18, 19]. The sequence profile is constructed from a set of multiply aligned sequences or structures of a probe family and is used to identify distant relationships to a database of target proteins. The profile is essentially a sequence position-specific scoring matrix encoding the probability of finding any of the 20 amino acid residues at that position in the target. In the case of p-SIFt, the SIFts derived from a set of probe structures are used to derive a position-dependent profile encoding the probability that a given interaction at that position is present. The contact p-SIFts derived for the ATPg, CDK2 and P38 clusters plotted in Fig. 10.4 measure the degree of interaction conservation for each group of structures. From the p-SIFts, it is evident that CDK2 and p38 inhibitors share some common binding interactions as observed between ATP and some regions of the kinase domain while displaying marked differences in others.

The 25 members of the ATPg cluster consist of nine structures of ATP complexed with three different kinases and 16 structures of ATP analogs complexed with six kinases. The ATPg p-SIFt computed from the ATPg cluster SIFts is shown at the top of Fig. 10.4. For comparison, we also plotted the p-SIFt derived using only the nine ATP structures in the ATPg cluster. For the nine ATP complexes, 18 out of 23 contacts are classified as conserved between the kinases and the ribose, triphosphate and adenine moieties. Moreover, there are no completely variable positions. Interestingly, even for these ATP-only structures, four interactions lie in the intermediate conservation range.

A convenient way to compare directly the p-SIFts is to define a difference profile computed by the direct subtraction of one p-SIFt from another. The differ-

Fig. 10.3 (a) Hierarchical clustering of SIFts from 93 protein kinase small molecule crystal structures. On the right are the dendrogram and the corresponding distance matrix. SIFts are reorganized according to the order given by the dendrogram. Six different regions are labeled above the SIFt heat map. Three major clusters (1–3) are labeled on the left side of the heat map and also a cluster corresponding to the DFG-out conformation of the kinases. (b) Comparison of the binding modes of the three different kinase clusters.

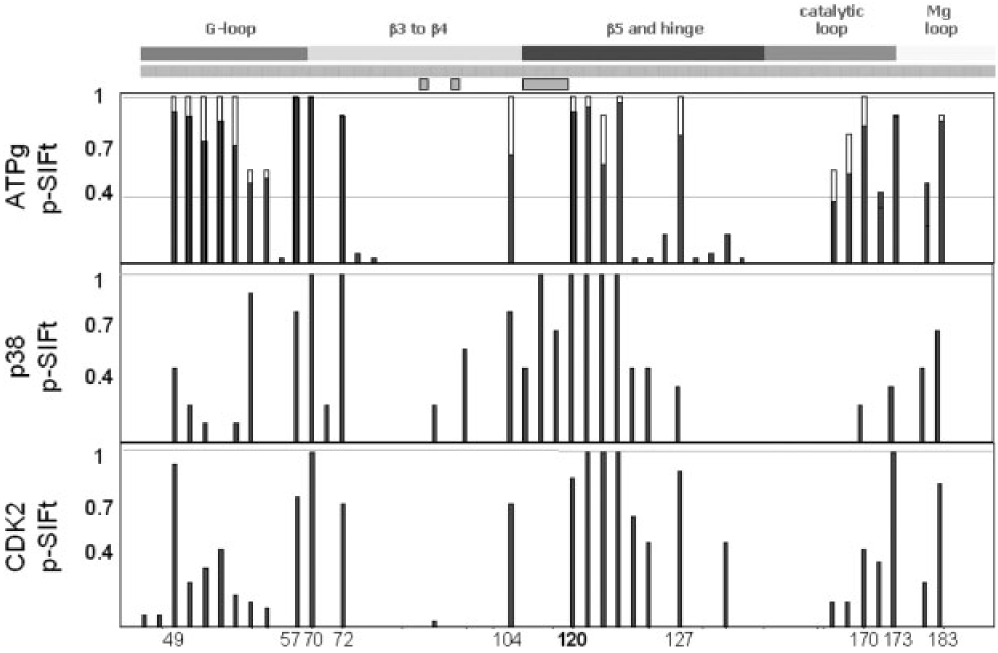


Fig. 10.4 The contact-only p-SIFts for ATPg (top), p38 (middle) and CDK2 (bottom) are plotted as a function of PKA residue numbering. The unshaded outline shown in the ATPg panel corresponds to the p-SIFt derived from the nine ATP-only structures. The increase in variability when ATP analogs are introduced is clearly visible. The blocks below the p38 p-SIFt denote residues making up the hydrophobic pocket of the kinase.

ence profiles provide an insight into how the interaction patterns observed for known kinase inhibitors differ from those detailed above for ATP. To this end, we defined difference profiles p38 – ATPg, p38 – CDK2 and CDK2 – ATPg, plotted in Fig. 10.5.

For the p38 – ATPg and p38 – CDK2 difference profiles (Fig. 10.5 a), the key distinctions are determined in part by the identity of the residue at position 120. Referred to as the “gatekeeper” residue, it controls the relative access to the hydrophobic pocket of the ATP site, a region not occupied by ATP. Bulky residues at position 120, such as Phe in CDK2, restrict access to the hydrophobic pocket, limiting the contacts available to a putative inhibitor (Fig. 10.5 b). The small Thr “gatekeeper” in p38 renders the residues making up the hydrophobic pocket accessible to small molecule inhibitors. The fact that small molecule inhibitors of p38 exploit these interactions is clearly evident from the p38 p-SIFt (Fig. 10.5 a), which indicates a set of intermediate and conserved interactions corresponding to hydrophobic pocket residues. The contrast in interaction with the hydrophobic pocket observed between p38, ATPg and CDK2 is clearly delineated by the distinct positive differences visible in the p38 – ATP and P38 – CDK2 difference profiles.

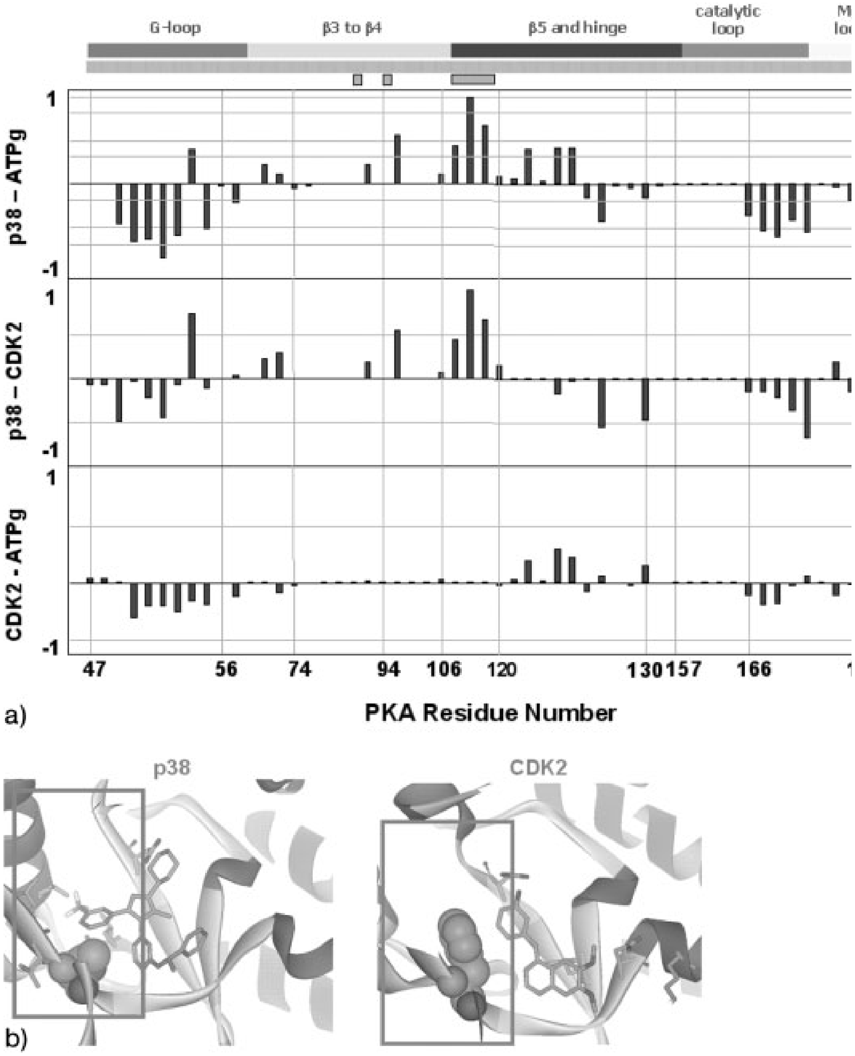


Fig. 10.5 (a) The contact-only difference profiles between p38-ATPg (top), p38-CDK2 (middle) and CDK-ATPg (bottom). The difference plots range from –1 to 1, where a value of 0 indicates that the interaction is conserved to the same degree in the two sets of structures, and a value of –1 or 1 denotes that a conserved interaction in one set of structures is not conserved in the other. (b) Binding sites for p38 and CDK2 with box highlighting the hydrophobic pocket and gatekeeper region.

In contrast, the CDK2 p-SIFt is more similar to the ATPg p-SIFt, as can be observed in the CDK2-ATP difference profile. Unlike p38, in CDK2 the Phe "gatekeeper" residue blocks access to the hydrophobic pocket. As a result, many of the residues accessible to CDK2 inhibitors will be those that also interact with ATP. In fact, of the conserved residues observed in the CDK2 p-SIFt, there are none that are not also conserved in the ATPg p-SIFt.

Unlike contacts with the hydrophobic pocket, several interactions conserved in the p38 cluster are common to CDK2, and also other non-ATP inhibitors. Finally, several interactions are conserved for ATPg and are observed with relatively low frequency for CDK2 and p38. These ATPg-specific contacts involve residues at positions 50–55, which interact with the ribose and phosphate moieties of ATP and with residues at positions 168, 170 and 171, in the vicinity of the catalytic loop.

10.8 Virtual Screening

In cases where existing structural information is available for how small molecules bind to a drug target, it would be valuable to use this information as target-based interaction constraints to discover additional leads. Our SIFt analysis identified clear interaction preferences for ATP, p38 and CDK2, clusters as well as a canonical set of conserved interactions common to all ligands bound to kinases at the ATP binding site. In this section, we will demonstrate how the p-SIFt can be applied in a VS workflow that can be tailored to a specific target without having to rely solely on the ambiguities of energy-based scoring. To this end, we tested the performance of p-SIFt-based scoring in a typical database enrichment application using p38 and CDK2 as targets.

A database containing 14 known inhibitors of p38 and 54 examples of CDK2 was spiked into a background of 1000 diverse commercially available compounds and docked against the X-ray structures of CDK2 (PDB code 1di8) and p38 (PDB code 1a9u). We then analyzed whether p-SIFT provided any advantage over popular scoring functions in enrichment of inhibitors by plotting the percentage of actives recovered as a function of the percentage of the database screened.

For p38, the enrichment obtained by applying p-SIFt scoring provided markedly superior results to those obtained using energy scoring such as Chemscore and PMF functions [7]. p-SIFt scoring performs close to the ideal enrichment curve over the first 2% of the database, meaning that 14 of the 16 known p38 actives were in the top 20 ranked ligands. On examination of the docking poses, it was discovered that for the other two inhibitors correct poses were never generated in the initial pose pool. The p-SIFt scoring method requires a pose having a correct docked binding mode to generate a high Z score, unlike energy scoring, which can generate high scores even for poses that bind incorrectly. Generating enrichments for the right reasons is a built-in advantage of the p-

SIFt scoring approach. For p38, the Hybrid Scoring scheme (p-SIFT to rule out undesirable poses followed by energy scoring function identify and rank the best pose) was found to offer no improvement in enrichment over that obtained from using p-SIFt scoring.

For CDK2, the Hybrid Scoring scheme variants performed better than the Traditional and p-SIFt schemes irrespective of what scoring function was used for pose selection and ranking. It appears that once the majority of the incorrect poses that contribute to false-positive scores have been filtered out, the differences between scoring functions visible in the results using these energy functions alone is factored out. Enrichments obtained using p-SIFt scoring are comparable to energy scoring up to 6% of the database screened and significantly better at higher levels [7].

The CDK2 p-SIFt is less selective against false poses, as evidenced by the poorer performance of p-SIFt scoring for CDK2 versus p38. Attaining database enrichments for CDK2 comparable to those obtained for p38 is a considerably more challenging task for VS. The large gatekeeper residue in CDK2 restricts the number of residues accessible in the ATP binding site. The p-SIFt for CDK2 samples less residues than p38 and conserved interactions are distributed over a relatively small spatial region. As a result, in the CDK2 there are fewer constraints on generating ligand placements and it is therefore easier to generate poses that satisfy conserved interactions in CDK2 compared with p38, where the residues of the hydrophobic pocket are accessible.

10.9 Use of p-SIFT to Enrich Selectively p38, CDK2 and ATP Complexes

The difference profiles exhibit clear regions where ATPg, CDK2 and p38 inhibitors bind to kinases in unique ways. These observations suggest that p-SIFts can be used to model the selectivity of inhibitors based on the types of interactions they are able to satisfy when binding to the kinase. In order to validate the use of p-SIFts as selectivity filters, we carried out a self-recognition experiment using the set of 93 X-ray structures as a test data set. For this purpose, p-SIFts were derived for p38, CDK2 and ATP where ~50% of the structures for each group were set aside and not used to derive the p-SIFt. For each p-SIFt, *Z* scores were then computed against all 93 kinase structures in order to assess the ability of p-SIFts to recognize members of their own group. For the p-SIFts to serve as effective molecular filters, the p38 p-SIFt needs to generate statistically significantly higher *Z* scores against the p38 cluster X-ray structures relative to the remaining structures, whereas the CDK2 and ATPg p-SIFts should perform similarly against the CDK2 and ATPg structures, respectively.

The p-SIFTs of ATPg, p38 and CDK2 p-SIFts were all successful at generating large fractions of complexes with high *Z* scores and, importantly, these were shifted relative to the counter targets (Fig. 10.6). The greatest separation in *Z* score distributions was obtained for p38, owing primarily to p-SIFt features re-

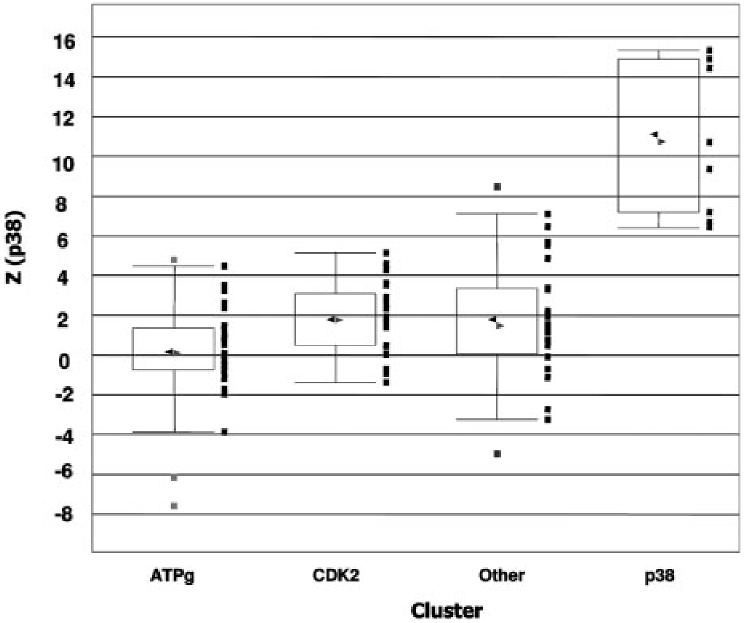


Fig. 10.6 Box plots of *Z* distributions obtained for the ATPg, p38 and CDK2, cluster subsets described in the text are shown for all kinases in the 93 X-ray structure set in (a), (b) and (c), respectively. The right and left arrows indicate the mean and the median, respectively, of the distribution; the vertical error bars delineate the upper and lower bounds of the data; the horizontal bars represent individual data points. The box outlines the upper and lower quartiles of the distribution.

flecting conserved residues in the hydrophobic pocket of the ATP binding site. The similarity between the ATPg and CDK2 p-SIFts, as discussed previously, has the consequence that 90% of the CDK2 structures overlap in *Z* score with the lowest scoring 35% of the ATPg [7]. This overlap exists primarily because the ATPg p-SIFt is in essence derived from a subset of the interactions sampled by CDK2. However, the differences between the ATP and CDK2 interaction patterns are captured in the CDK2 p-SIFt. Consequently, the highest segment in the distribution shown in Fig. 10.6 contains 19 of 20 CDK2 structures and overlaps with only two ATPg structures.

10.10 Conclusion

This chapter has introduced interaction profiling as a new and powerful approach to understand what interactions small molecules exploit in order to be competitive against ATP and, often, selective for a particular kinase. More importantly, we have shown that the information encoded in the interaction profiles can be used effectively to filter virtual libraries selectively for ligands that are inhibitors to a particular kinase. We envision that the use of SIFt should fully leverage the use of experimental information from structure-based drug design experiments into the design and optimization of virtual libraries (Fig. 10.7). This should lead to more effective focusing of chemical libraries into binding sites and may lead to improved hit rates from virtual screening.

Given the rapid growth in the number of available X-ray structures, it should be possible eventually to construct and screen against a virtual selectivity panel of interaction profiles for multiple targets in much the same way that inhibitors are routinely tested against a panel of *in vitro* kinase inhibition assays. The resulting virtual selectivity profile could be pre-computed for ligands in virtual libraries, thus providing an annotation that could be mined when selective inhibitors to any target are desired. In addition, predicted cross-reactivity to a target could be an effective starting point for lead discovery for novel targets, an approach that has been demonstrated to be fruitful for protein kinases.

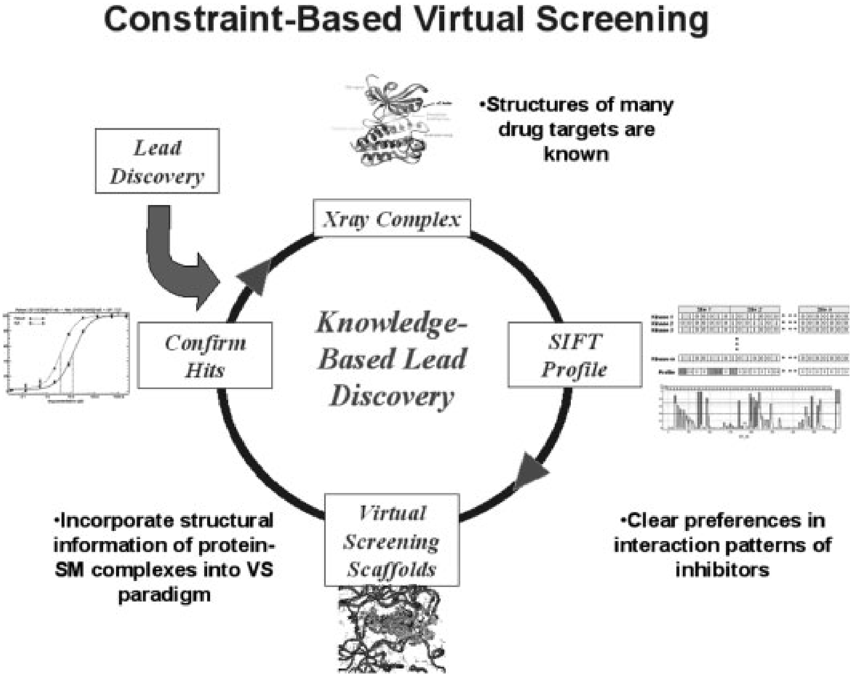

Constraint-Based Virtual Screening
•Structures of many drug targets are known
Lead Discovery
Xray Complex
Knowledge-Based Lead Discovery
Confirm Hits
SIFT Profile
•Incorporate structural information of protein-SM complexes into VS paradigm
Virtual Screening Scaffolds
•Clear preferences in interaction patterns of inhibitors

Fig. 10.7 Integration of SIFt into the structure-based drug design workflow. The experimental structure(s) of a drug target in complex with small molecules are used to generate SIFt and p-SIFt. This is used to filter a virtual chemical library, which is used to identify compounds for testing. These are confirmed as hits; their structures are determined, thus leading to further cycles of structure-based drug design.

Acknowledgments

We would like to thank Rainer Fuchs, Donovan Chin and Herman van Vlijmen for useful discussions during the course of this work.

References

1 Cohen, P. Protein kinases – the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **2002**, *1*, 309–315.

2 Blume-Jensen, P., Hunter, T. Oncogenic kinase signaling. *Nature* **2001**, *411*, 355–365.

3 Shawver, L. K., Slamon, D., Ullrich, A. Smart drugs: tyrosine kinase inhibitors in cancer therapy. *Cancer Cell* **2002**, *1*, 117–123.

4 Vieth, M., Higgs, R.E., Robertson, D.H., Shapiro, M., Gragg, E.A., Hemmerle, H. Kinomics – structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.

5 Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S., Sowadski, J.M. Crystal structure of the catalytic subunit of a cyclic adenosine monophosphate-dependent protein kinase. *Science* **1991**, *253*, 407–414.

6 Deng, Z., Chuaqui, C., Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein–ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

7 Chuaqui, C., Deng, Z., Singh, J. Interaction profiles of protein kinase inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121–133.

8 Lee, B., Richards, F.M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.

9 CCPN. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **1994**, *D50*, 760–763.

10 McDonald, I.K., Thornton, J.M. Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793.

11 Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

12 Raymond, J.W., Blankley, C.J., Willett, P. Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J. Mol. Graph. Model.* **2003**, *21*, 421–433.

13 Dubes, R., Jain, A.K. Clustering methodologies in exploratory analysis. *Adv. Comput.* **1980**, *19*, 113–228.

14 Scapin, G. Structural biology in drug design: selective protein kinase inhibitors. *Drug Discov. Today* **2002**, *7*, 601–611.

15 Johnson, L.N., Noble, M.E.M., Owen, D.J. Active and inactive protein kinases:-structural basis for regulation. *Cell* **1996**, *85*, 149–158.

16 Nolen, B., Taylor, S., Ghosh, G. Regulation of protein kinases:controlling activity through activation segment conformation. *Mol. Cell* **2004**, *15*, 661–675.

17 Luthy, R., Xenarios, I., Bucher, P. Improving the sensitivity of the sequence profile method. *Protein Sci.* **1994**, *3*, 139–146.

18 Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, *253*, 164–170.

19 Eisenberg, D., Bowie, J.U., Luthy, R., Choe, S. Three-dimensional profiles for analysing protein sequence–structure relationships. *Faraday Discuss. Chem. Soc.* **1992**, 25–34.

11
Application of Structure-based Alignment Methods for 3D QSAR Analyses

Wolfgang Sippl

11.1
Introduction

In drug design, establishing a common alignment of 3D structures of investigated molecules is an important prerequisite for several methodologies, e.g. 3D similarity analysis, prediction of biological activities and even estimation of ADME parameters [1–3]. Various methodologies and pharmacophore strategies for the superposition of small ligands have therefore been proposed in the literature and are reviewed in this book (see Chapter 2). An alignment generation procedure usually comprises two phases: superimposing the molecules and scoring of the alignments derived. Superposition techniques may either utilize information obtained from a binding site of a corresponding target protein (direct methods) or be based solely on information obtained from the ligands themselves (indirect methods). Some common assumptions, especially for the indirect methods, are that the aligned molecules interact with the same amino acids within a binding pocket and exhibit a unique binding mode. Additionally, the generated alignment ideally contains the ligands in their bioactive conformation. Superposition methods differ in how they treat flexibility and molecular representation. Molecules can be considered as flexible or rigid; alternatively, flexibility can also be modeled via a limited set of rigid conformers. The molecules to be aligned can be represented by their atoms, shape or molecular interaction fields [1, 2].

The prediction of biological activity of novel compounds based on their structure is one of the major challenges in today's drug design. A prerequisite for most approaches is the correct alignment of the molecules under study. Similarly to the alignment procedures, the prediction methods can be classified into two major categories: indirect ligand-based and direct structure-based approaches. Ligand-based methods, including traditional quantitative structure–activity relationships (QSAR) [4] and modern 3D QSAR techniques [5], are based entirely on experimental structure–activity relationships for enzyme inhibitors or receptor ligands. 3D QSAR methods are nowadays used widely in drug de-

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

sign, since they are computationally feasible and afford fast generation of models from which biological activity of newly synthesized molecules can be predicted. The basic assumption is that a suitable sampling of the molecular interaction fields around a set of aligned molecules might provide all information necessary for an understanding of their biological activities [6]. A suitable sampling is achieved by calculating interaction energies between each molecule and an appropriate probe placed at regularly spaced grid points surrounding the molecules. The resulting energies derived from simple potential functions can then be contoured in order to give a quantitative spatial description of molecular properties. If correlated with biological activity, 3D fields can be generated, which describe the contribution of a region of interest surrounding the ligands to the target properties. However, there is a major difficulty in the application of 3D QSAR methods: in order to obtain a correct model, a spatial arrangement of the ligands towards one another has to be found that is representative of the relative differences in the binding geometry at the protein binding site. The success of a molecular field analysis is therefore determined by the choice of the ligand superposition [7–9]. In most cases, the first step in a 3D QSAR study is the generation of a reliable pharmacophore model. Many alignment strategies have been reported and compared that accomplish this purpose (a detailed comparison of different methods can be found in [2]). Depending on the molecular flexibility and the structural diversity of the compounds investigated, the task of generating a unique pharmacophore can become less feasible. Despite the difficulties concerning the molecular alignment, many successful 3D QSAR case studies applying different programs have been reported in the last few years. Most CoMFA applications in drug design have been comprehensively listed and discussed in some reviews [10–13] and books [14–16].

Structure-based methods, on the other hand, incorporate information from the target protein and are able to calculate fairly accurately the position and orientation of a potential ligand in a protein binding site [17, 18]. Over the last decade, a broad range of competitive methods for scoring protein–ligand interactions has emerged [19–27]. Established approaches have been further improved, e.g. in the area of the regression-based scoring functions or methods based on first principles. In addition, well-known techniques have been applied to protein–ligand scoring by using atom–atom contact potentials to develop knowledge-based scoring functions. The major problem with modern docking programs is the inability to evaluate the free energy of binding required to score correctly different ligand–receptor complexes. The main problem in affinity prediction is that the underlying molecular interactions are highly complex and that the experimental data (both structural and biological data) are far from being perfect for computational approaches. Numerous terms have to be taken into account when trying to quantify correctly the free energy of binding [27–29]. Elaborate methods such as the free energy perturbation and the thermodynamic integration methods have been shown to be able – at least to some extent – to predict binding affinities correctly, but have the drawback of being computationally very expensive.

In order to exploit the strengths of both approaches, i.e. incorporation of protein information by docking programs and generation of predictive models for related molecules by 3D QSAR methods, we and others suggested a combination of both methods, resulting in an automated unbiased procedure [30–37]. In this context, the 3D structure of a target protein is used within a docking protocol to guide the alignment for a comparative molecular field analysis. This approach allows the generation of a kind of target-specific scoring method considering all the structure–activity data known for a related ligand data set.

This chapter focuses on computational studies which employ a combination of structure-based and 3D QSAR methods as a mean to predict the affinity of a ligand for its receptor. The comprehensive utility of this approach is exemplified by case studies published in the last few years and from our laboratory. Special emphasis will be placed on a detailed description of the combined structure–ligand-based approach and the successful application of this procedure to the design of novel drug molecules.

11.2
Why is 3D QSAR So Attractive?

The era of quantitative analysis for the correlation of molecular structures with biological data started in the 1960s with the classical equation for 2D QSAR analysis proposed by Hansch and Leo [4]. Since then, several QSAR approaches have been developed [5, 11]. The first applicable 3D QSAR method was proposed in 1988 by Cramer et al. [6]. The primary aim of 3D QSAR methods is to establish a correlation of biological activities of a series of structurally and biologically characterized compounds with the spatial fingerprints of numerous field properties of each molecule, such as steric demand, lipophilicity and electrostatic interactions. Typically, a 3D QSAR study allows the identification of the pharmacophoric arrangement of molecular features in space and provides guidelines for the design of next-generation compounds with enhanced biological potencies.

No 3D QSAR method would be applied to a dataset unless one expected that the analysis would reveal insights into useful 3D structure–activity relationships. Since the 3D properties of molecules govern biological activity, it is especially informative to see a 3D summary of how structural changes influence biological activities. Approaches that do not provide such a graphical output are often less attractive to the scientific community. An advantage of 3D over 2D QSAR methods is that they take into account 3D structures of molecules and are additionally applicable to sets of structurally diverse compounds [38]. Recent QSAR methods include 4D QSAR, where an ensemble of conformations for each ligand represents the fourth dimension [39], and 5D QSAR, which in addition considers hypotheses for changes that might occur in a conformation of a receptor due to ligand binding (induced fit) as a fifth dimension [40].

The number of 3D QSAR studies has increased exponentially over the last decade, since a variety of methods are commercially available in user-friendly, graphically guided software [6, 9, 41]. The most frequently applied methods include the comparative molecular field analysis CoMFA, the comparative molecular similarity indices analysis CoMSIA [9] and the GRID/GOLPE method (Generating Optimal Linear PLS Estimations) [41, 42]. Several reviews have been published in the last decade dealing with the underlying theory, the problems and the application of CoMFA-related approaches [11, 12, 38]. Apart from the commercial distribution, a major factor causing the ongoing enthusiasm for 3D QSAR comes from the proven ability of several of these methods to predict correctly the biological activity of novel compounds. This ability is gaining respect as scientists realize that we are far away from the hoped-for fast and accurate prediction of affinity from (the structure of) protein–ligand complexes by free-energy perturbation or empirical scoring methods [23, 28].

11.3 CoMFA and Related Methods

11.3.1 CoMFA

For many years, 3D QSAR has been used as a synonym for CoMFA [6], which was the first method that implemented in a QSAR method the concept that the biological activity of a ligand can be predicted from its three-dimensional structure. Until now, CoMFA has been probably the most often applied 3D QSAR method [12, 38]. A CoMFA study normally starts with traditional pharmacophore modeling in order to suggest a bioactive conformation of each molecule and ways to superimpose the molecules under study. The underlying idea of CoMFA is that differences in a target property, e.g. biological activity, are often closely related to equivalent changes in shapes and strengths of non-covalent interaction fields surrounding the molecules, or, stated in a different way, the steric and electrostatic fields provide all information necessary for understanding the biological properties of a set of compounds. Hence the molecules are placed in a cubic grid and the interaction energies between the molecule and a defined probe are calculated for each grid point. Normally, only two potentials, namely a steric potential in the form of a Lennard–Jones function and an electrostatic potential in form of a simple Coulomb function, are used within a CoMFA study. It is obvious that the description of molecular similarity is not a trivial task, nor is the description of the interaction process of ligands with corresponding biological targets. In the standard application of CoMFA, only enthalpic contributions of the free energy of binding are provided by the potentials used. However, many binding effects are governed by hydrophobic and entropic contributions. Therefore, one has to characterize in advance the expected main contributions of forces and whether under these conditions CoMFA will actually be able to find realistic results.

In the original CoMFA report, field values were systematically calculated for ligands at each grid point of a regularly sampled 3D grid box, that extended 4 Å beyond the dimension of all molecules in the dataset, using an sp^3 carbon atom with +1 charge as probe [6]. The grid resolution should be in a range to produce the field information that is necessary to describe variations in biological activity. On the other hand, introduction of too much irrelevant data to statistical analysis may result in a decrease in predictivity of the model. Typically, a resolution of 2 Å is utilized. Often, superior results are derived using a grid spacing of 2 Å as opposed to the more accurate 1 Å spacing [7]. In addition, the CoMFA program provides a variety of other parameters (probe atoms, charges, energy scaling, energy cutoffs, etc.) which can be adjusted by the user. This flexibility in parameter settings enables the user to fit the whole procedure as closely as possible to a problem. However, it enhances the possibility of chance correlations. Interestingly, nearly all of the successful CoMFA analyses have been done with default parameters.

11.3.2
CoMSIA

Owing to the problems associated with the functional form of the Lennard–Jones potential used in most CoMFA methods [12], Klebe et al. [9], developed a similarity indices-based CoMFA method named CoMSIA (Comparative Molecular Similarity Indices Analysis). Instead of grid-based fields, CoMSIA is based on similarity indices that are obtained by using a functional form that is adapted from the SEAL algorithm. Three different indices related to steric, electrostatic and hydrophobic potentials were used in their study of the classical Tripos steroid benchmark dataset. Models of comparable statistical quality with respect to internal cross-validation of the training set, in addition to predictivities of the test set, were derived using the CoMSIA method. The clear advantage of this method lies in the functions used to describe the molecules studied, and also the resulting contour maps. The contour maps obtained from CoMSIA are easier to interpret than those obtained by the CoMFA approach. The CoMSIA procedure also avoids cutoff values used in CoMFA to restrict potential functions by assuming unacceptably large values. Detailed descriptions of the method and its application can be found in the literature [9, 43]. Recently, the authors of CoMSIA included a novel hydrogen-bond descriptor which should overcome the problem of underestimating hydrogen bonds in CoMFA studies [43].

11.3.3
GRID/GOLPE

The GRID program [44] has been used by a number of workers as an alternative to the original CoMFA method for calculating interaction fields. An advantage of the GRID approach, apart from the large number of chemical probes available, is the use of a 6–4 potential function, which is smoother than the 6–12 form of the Lennard–Jones type, for calculating the interaction energies at the grid lattice points. Good statistical results were obtained, for example, in an analysis of glycogen phosphorylase b inhibitors by Cruciani and Watson [45]. They used the GRID force field in combination with the GOLPE program [43], which accomplishes the necessary chemometric analysis. The particularly interesting aspect of this dataset is that the X-ray structures of all protein–ligand complexes have been solved. This allowed the authors to investigate the dataset using new and different methods to develop 3D QSAR techniques further.

A further refinement of the original CoMFA technique was realized by introducing the concept of variable selection and reduction [45]. A large number of variables in the descriptor matrix (i.e. the interaction energies) represent a statistical problem in the CoMFA approach. These variables make it increasingly difficult for multivariate projection methods, such as PLS, to distinguish the useful information contained in the descriptor matrix from that of lower quality or noise. Hence approaches for separating the useful from the less useful variables were needed. The GOLPE approach was developed in order to identify which variables are meaningful for the prediction of the biological activity and to remove those with no predictivity [46]. Within this approach, fractional factorial design (FFD) is initially applied to test multiple combinations of variables [42]. For each combination, a PLS model is generated and only variables which significantly increase the predictivity are considered. Variables are then classified according to their contribution to predictivity. A further advance in GOLPE is the implementation of the smart region definition (SRD) procedure that aims to select the cluster of variables mainly responsible for activity rather than a single variable. The SRD technique seems less prone to change correlation than any single variable selection and improves the interpretability of the models [46].

11.4 Reliability of 3D QSAR Models

The quality and reliability of any 3D QSAR model is strongly dependent on the careful examination of each step within a 3D QSAR analysis. As with any QSAR method, an important point is the question of whether the biological activities of all compounds studied are of comparable quality. Preferably, biological data should be obtained in the same laboratory under identical conditions. All compounds being tested in a system must have the same mechanism (binding mode) and all inactive compounds must be shown to be truly inactive. Only *in vitro* data should be considered, since only *in vitro* experiments are able to reach a real equilibrium. All other test systems undergo time-dependent changes by multiple coupling to parallel biochemical processes (e.g. membrane permeation). Another critical point is the existence of transport phenomena and diffusion gradients underlying all biological data. One has to bear in mind that all 3D QSAR approaches were developed to describe only one interaction step in the lifetime of ligands. In all cases, where nonlinear phenomena result from drug transport and distribution, any 3D QSAR technique should be applied with caution. The biological activities of the molecules used in a CoMFA study should ideally span a range of at least three orders of magnitude. For all molecules under study, the exact 3D structure has to be reported. If no information on the exact stereochemistry of the tested compounds is given (mixtures of enantiomers or diastereomers), these compounds should be excluded from a CoMFA study.

The search for the bioactive conformation and a molecular alignment constitutes a serious problem within all 3D QSAR studies. It is one of the most important sources of wrong conclusions and errors in all 3D QSAR analysis. The risk of deriving irrelevant geometries can be reduced by considering rigid analogs. Even then, the alignment poses problems, because there are some cases of different binding modes of seemingly closely related compounds [14]. Even if the binding modes are comparable, the choice of wrong ligand conformations may still result in a 3D QSAR analysis being unreliable. Chemical feature-based pharmacophore alignments have been demonstrated to be a useful starting point for 3D QSAR studies [48–50]. Problems in the generation of conformations and the correct alignment could be avoided by deriving them from the 3D structures of ligand–protein complexes which are known from X-ray crystallography, NMR or homology modeling [35].

The final stage of a 3D QSAR analysis consists in a statistical validation in order to assess the significance of the model and hence its ability to predict biological activities of novel compounds. In most published 3D QSAR case studies, the leave-one-out (LOO) cross-validation procedure has been used for this purpose. The output of this procedure is the cross-validated q^2 and the standard deviation of error prediction (SDEP), which are commonly regarded as ultimate criteria of both the robustness and predictive ability of a model. The simplest cross-validation method is LOO, where one object at a time is removed from the dataset and predicted by the model generated. A more robust and reliable method is the leave-several-out cross-validation. For example, in the leave-20%-out cross-validation, five groups of approximately the same size are generated. Thus, 80% of the compounds are randomly selected for the generation of a model, which is then used to predict the remaining compounds. This operation must be repeated numerous times in order to obtain reliable statistical results. The leave-20%-out or also the more demanding leave-50%-out cross-validation results are much better indicators for the robustness and the predictive ability of a 3D QSAR model than the usually used LOO procedure [47, 51]. LOO often yields too optimistic models, which fail when predicting real test set molecules.

Despite the known limitations of the LOO procedure, it is still uncommon to test 3D QSAR models for their ability to predict correctly the biological activities of compounds not included in the training set. Still, many workers claim that their models, showing high LOO q^2 values, have high predictive ability in the absence of external validation (for a detailed discussion on this problem, see [51–55]). Contrary such expectations, it has been shown in several studies that a correlation between the LOO cross-validated q^2 value for the training set and the correlation coefficient r^2 between the predicted and observed activities for the test set does not exist [52, 54]. Therefore, it is highly recommended to use demanding cross-validation procedures and external test sets to validate further an established 3D QSAR model.

11.5 Structure-based Alignments Within 3D QSAR

The combination of ligand-based and structure-based approaches is an attractive strategy for ligands for which the binding site is known but the exact binding mode has not been determined experimentally. This has been demonstrated by a variety of approaches developed within the last decade. One of the earliest approaches published in this field was the VALIDATE program by Hoad et al. [56]. The method uses 12 physico-chemical and energetic parameters, including the electrostatic and steric interaction energy between a receptor protein and ligands computed with the AMBER force field, to correlate these descriptors with biological activities. The method has been validated on 51 diverse protein–ligand X-ray structures. The ligands ranged in size from 24 to 1512 atoms and spanned a pK_i range from 2.47 to 14.0. The best fit equation, using PLS analysis, yielded $r^2=0.85$ with a standard error of 1.0 log units and a cross-validated $r^2=0.78$. This QSAR was found to be predictive for at least two of three test sets of enzyme inhibitor complexes: 14 structurally diverse crystalline complexes (predictive $r^2=0.81$), 13 HIV protease inhibitors (predictive $r^2=0.57$) and 11 thermolysin inhibitors (predictive $r^2=0.72$). VALIDATE has also been successfully applied to the design of non-peptidic HIV-1 protease inhibitors [57].

Another approach which utilizes the intermolecular interaction energy between the receptor and its ligand is the COMBINE approach developed by Ortiz et al. [58]. It employs a unique method that partitions the interaction energy between receptor and ligand fragments and subjects them to a statistical analysis. This is suggested to enhance contributions from mechanistically important interaction terms and to tune out noise due to inaccuracies in the potential energy functions and molecular models. For a set of 26 phospholipase A2 inhibitors, the direct correlation between interaction energies, computed using the CFF91 DISCOVER force field and percentage enzyme inhibition, was very low, $r=0.212$. However, with the COMBINE approach, employing PLS fitting and the GOLPE variable selection procedure, good correlations with the percentage inhibition rate were observed ($q^2_{\mathrm{LOO}}=0.82$). Predictive models were also obtained for a variety of other biological targets and their ligands: Acetylcholinesterase (AChE) inhibitors ($n=35$, $q^2_{\mathrm{LOO}}=0.76$) [59], factor-Xa inhibitors ($n=133$, $q^2_{\mathrm{LOO}}=0.61$) [60], periplasmic oligopeptide binding component (OppA) ligands ($n=28$, $q^2_{\mathrm{LOO}}=0.73$) [61], neuraminidase inhibitors ($n=39$, $q^2_{\mathrm{LOO}}=0.78$) [62], cyclooygenase-2 inhibitors ($n=58$, $q^2_{\mathrm{LOO}}=0.64$ [63] and cytochrome P450 1A2 ligands ($n=12$, $q^2_{\mathrm{LOO}}=0.74$) [37].

Recent approaches that primarily employ the combination of structure-based alignment strategies and comparative molecular field analysis to predict ligand affinity have included studies of ligand binding to enzymes and receptor X-ray structures, and also protein homology models. Marshall's group was one of the first to apply this technique. They examined the binding of 59 HIV-1 protease inhibitors from different structural classes [64]. The availability of X-ray crystallographic data for at least one representative from each class bound to HIV-1 protease provided information regarding not only the active conformation of each inhibitor but also, via superposition of protease backbones, the relative positions of each ligand with respect to one another in the active site of the enzyme. The molecules were aligned and served as templates on which additional congeners were field-fit minimized. The predictivity of the derived models was subsequently evaluated using test set molecules, for which X-ray structural information was available.

Tropsha's group used the crystal structures of the three AChE inhibitors – tacrine, edrophonium and decamethonium – as a template on which other structurally analogous AChE inhibitors were superimposed. In order to obtain quantitative relationship between the structure and biological activities of the inhibitors, CoMFA in combination with a variable-selection method {cross-validated r^2 guided region selection (q^2-GRS) routine [65]} was carried out. Using the structure-based alignment of 60 AChE inhibitors and CoMFA/q^2-GRS yielded a highly predictive QSAR model with a q^2 of 0.73 [65]. Whereas in the last two studies manually derived protein-based alignments were used as input for a 3D QSAR analysis, several case studies have recently been published in which an automated docking procedure was applied for structure-based alignment generation. Whereas in the two applications of Marshall's and Trophsa's groups manually derived protein-based alignments were used as input for 3D QSAR analysis, several studies have been published recently in which an automated docking procedure has been used for the structure-based alignment generation.

Mügge and Podlogary generated a series of CoMFA models from docking-based and atom-based alignments for biphenylcarboxylic acid matrix–metalloproteinase-2 (MMP-3) inhibitors [66]. The underlying statistics of these approaches were assessed in order to determine whether a docking approach can be employed as an automated alignment tool for the development of 3D QSAR models. The docking-based alignment provided by a DOCK/PMF scoring protocol yielded statistically significant, cross-validated CoMFA models. Field fit minimization was successfully applied to refine the docking-based alignments. The statistically best CoMFA model has been created by the ligand-based alignment that has been found, however, to be inconsistent with the stromelysin crystal structure. The refined docking-based alignment resulted in a final alignment that is consistent with the crystal structure and only slightly statistically inferior to the ligand-based aligned CoMFA model.

Constantino et al. used the combination of a docking-based alignment and 3D QSAR analysis to build a predictive model for 46 poly(ADP-ribose)polymerase (PARP) inhibitors [67]. Representative PARP inhibitors were docked into the crystallographic structure of the catalytic domain of PARP by using the AutoDock 2.4 program. The docking studies provided an alignment scheme that was instrumental in superimposing all the remaining inhibitors. Based on this alignment, a 3D QSAR model was established using the RECEPTOR module within Cerius^2 [68]. The resulting statistical analysis yielded a predictive model able to explain much of the variance of the 46-compound data set ($n=46$, $q^2_{\mathrm{LOO}}=0.74$).

Matter et al. examined a series of 138 inhibitors of the blood coagulation enzyme factor Xa using CoMFA and CoMSIA [69]. To rationalize biological affinity and to provide guidelines for further design, all compounds were docked into the factor Xa binding site. Those docking studies were based on X-ray structures of factor Xa in complex with literature-known inhibitors. The docking results were validated by four X-ray crystal structures of representative ligands in factor Xa. The 3D-QSAR models based on a superposition rule derived from these docking studies were validated using conventional and cross-validated q^2 values. This led to consistent and highly predictive 3D-QSAR models with which were found to correspond to experimentally determined factor Xa binding site topology in terms of steric, electrostatic and hydrophobic complementarity ($n=138$, $q^2_{\mathrm{LOO}}=0.75$). The same strategy was successfully applied to a data set of MMP-8 matrix–metalloproteinase inhibitors ($n=90$, $q^2_{\mathrm{LOO}}=0.57$) [70].

Tervo et al. examined the binding of 92 catechol-*O*-methyltransferase inhibitors (COMT) [71]. They used a combination of the FlexX molecular docking method with a GRID/GOLPE 3D QSAR to analyze possible interactions between COMT and its inhibitors and to initiate the design of new inhibitors. The GRID/GOLPE models were made by using bioactive conformations from docking experiments, which yielded a q^2 value of 0.64. The docking results, the COMT X-ray structure and the 3D QSAR models were found to be in good agreement with each other. Interest was also focused on how well the calculated FlexX total energy scores correlated with the experimental biological activity. FlexX total energy scores for the 92 compounds were correlated with the corresponding pIC_{50} values, resulting in an r^2 value of 0.30, indicating the problem of scoring functions.

In a study by the same group, structure-based alignment techniques for 3D QSAR were analyzed and compared with traditional atom-based approaches. A set of 113 flexible cyclic urea HIV-1 protease inhibitors was used to generate CoMFA and CoMSIA models [72]. Inhibitors that were aligned automatically with GOLD were in agreement with information obtained from existing X-ray structures. Both the protein- and the ligand-based alignment strategy produced statistically significant CoMFA and CoMSIA models (best q^2 value of 0.65 and best predictive r^2 value of 0.75), whereas the GOLD-based alignment gave more robust models for predicting the activities of an external inhibitor set.

Some groups have applied the docking-based alignment strategy to develop 3D QSAR models for nuclear hormone receptor ligands. During the last decade, several X-ray structures of nuclear hormone receptors in complex with hormones, agonists and antagonists have been resolved and used for structure-based drug design [73]. In general, automated docking programs were shown to be successful in docking ligands to this receptor class [34, 74, 75]. Therefore, it was appealing to use structure-based 3D QSAR approaches also for this class of targets. Predictive and robust receptor-based 3D QSAR models have been reported for estrogen receptor agonists ($n=30$, $q^2_{\mathrm{LOO}}=0.90$, $q^2_{\mathrm{L50\%O}}=0.82$ [34] and $n=36$, $q^2_{\mathrm{LOO}}=0.63$ [76]), and also for androgen receptor ligand ($n=67$, $q^2_{\mathrm{LOO}}=0.66$ [77] and $n=25$, $q^2_{\mathrm{LOO}}=0.78$ [78]).

Moro et al. used a homology model of the A3 adenosine receptor to generate a target-based alignment [79]. Docking-based structure superimposition was used to perform a 3D QSAR analysis using the CoMFA program. A correlation coefficient q^2 of 0.84 was obtained for a set of 106 A3 receptor ligands. Both steric and electrostatic contour plots, obtained from the CoMFA analysis, were found to be in agreement with the hypothetical binding site achieved by molecular docking. Following the reported computational approach, 17 new ligands were designed, synthesized and tested. Consistently, the predicted K_i values were very close to the experimental values.

The nearly exponential growth of the Protein Data Bank in the last few years has resulted in a huge number of 3D structures of interesting target proteins which can be analyzed by means of structure-based drug design methods. It has also been shown on numerous high-resolution protein–ligand structures that docking methods are nowadays able to predict fairly accurate the position of ligands in the corresponding binding sites [25]. Therefore, it is not surprising that an increasing number of structure-based 3D QSAR models have now been published. A combination of docking and comparative molecular field analysis has been successfully applied to enzyme inhibitors of the following pharmaceutically relevant targets: non-nucleoside HIV-1 reverse transcriptase inhibitors ($n=29$, $q^2_{\mathrm{LOO}}=0.72$) [80], Raf-1 kinase inhibitors ($n=91$, $q^2_{\mathrm{LOO}}=0.53$) [81], aldose reductase inhibitors ($n=45$, $q^2_{\mathrm{LOO}}=0.56$) [82], cyclooxygenase-2 inhibitors ($n=88$, $q^2_{\mathrm{LOO}}=0.84$) [83], HIV-1 reverse transcriptase inhibitors ($n=70$, $q^2_{\mathrm{LOO}}=0.84$) [84], EGFR kinase inhibitors ($n=96$, $q^2_{\mathrm{LOO}}=0.64$) [85], *Yersinia* protein tyrosine phosphatase YopH inhibitors ($n=34$, $q^2_{\mathrm{LOO}}=0.83$) [86], HIV-1 integrase inhibitors ($n=66$, $q^2_{\mathrm{LOO}}=0.72$) [87], HIV-1 reverse transcriptase inhibitors ($n=50$, $q^2_{\mathrm{LOO}}=0.78$) [88], dihydrofolate reductase inhibitors ($n=240$, $q^2_{\mathrm{L10\%O}}=0.65$) [89] and type-B monoamine oxidase inhbitors ($n=130$, $q^2_{\mathrm{L10\%O}}=0.73$) [90].

We have successfully applied the combination of structure-based 3D QSAR to several drug design projects [34, 35, 91–95]. Our goal was the prediction of biological activities and prioritizing synthesis for proposed compounds *a priori*. To show exemplarily the potential of the combined approach, a case study is presented here, in which the structure-based 3D QSAR was used for the design of novel AChE inhibitors [33, 96]. AChE is an enzyme that hydrolyzes the neurotransmitter acetylcholine (ACh) at cholinergic synapses with a turnover rate superior to most other known enzymes [97]. Recent research interest regarding this enzyme is due not only to this high catalytic efficiency, but also to the broad implications of AChE inhibition for human health, agrochemistry and chemical agents. For example, Alzheimer's disease (AD) is associated with low *in vivo* levels of ACh, hence AChE has been the focus of many drug discovery projects aimed at maintaining available ACh via mild or reversible inhibitors such as tacrine and donepezil [98, 99]. AD is the most common cause of dementia in the elderly population. The cholinergic hypothesis of AD has provided the rationale for the current major therapeutic approach to AD. However, to date, all longer term studies have shown that clinical efficacy declines as a result of either a loss of drug efficacy or the relentless progression of the disease. Hence interest in the discovery of novel AChE inhibitors continues since the current AChE inhibitors lack perfection.

The availability of the AChE crystal structures of various species in its un/complexed form provides a solid basis for the structure-based design of novel AChE inhibitors [100]. Within AChE, two principal binding sites can be found. The catalytically active site is located at the bottom/base of a deep gorge in the enzyme. The ACh catalysis reaction is accomplished by a collective interaction of a catalytic triad consisting of Ser203, Glu334 and His447 and nearby residues (e.g. the choline binding site: Trp86) [101].

AChE also has a peripheral anionic site (PAS) located near the enzyme surface at the entrance of the active site gorge. In the PAS, the Trp286 residue plays a very important role in ligand binding that affects enzymatic activity through a combination of steric blockade of ligands moving through the gorge and allosteric alteration of the catalytic triad conformation and efficiency [102]. The gorge itself is a narrow hydrophobic channel with a length of 20 Å, connecting the PAS site to the active site [103]. An acyl binding pocket consisting of Gly122, Trp236, Phe295, Phe297 and Phe338 residues is responsible for the interaction with the acetyl group [104]. Early inhibition research was mainly focused on ligands binding in the active site (e.g. tacrine, amiridine). Recent efforts have focused on finding novel ligands that bind to both sites in order to search for more potent reversible inhibitors (e.g. TAK-147, E2020 [99]), selectively favoring the inhibition of AChE rather than the related butyrylcholinesterase (BChE). In this context, we focused on the search for novel potent and selective AChE inhibitors [33, 96]. The starting point of our AChE project was the finding that the morpholine derivative minaprine showed weak inhibition of AChE [105]. Starting with the lead structure minaprine and the available X-ray structures of AChE (uncomplexed and complexed with different inhibitors [102–108]), a variety of minaprine derivatives were developed [109].

A detailed inspection of the available AChE inhibitor X-ray structures yielded relevant information concerning the orientation of the inhibitors within the binding pocket. AChE shows a nearly identical three-dimensional structure in all known X-ray structures. The active site is located 20 Å from the protein surface at the bottom of a deep and narrow gorge. The only major conformational difference between the four complexes is the orientation of Phe330, a residue located in the middle of the gorge. Depending on the co-crystallized inhibitor, this aromatic residue adopts a different conformation. However, the positions of the four inhibitors in the binding pocket are different, indicating that more than one clearly defined binding region exists. In order to find a reliable docking/scoring strategy for this target, the known crystal structures of AChE–inhibitor complexes were taken as a positive control. AutoDock in combination with a force-field refinement yielded good results when docking the AChE inhibitors [33, 96]. The program has been used in many docking studies and shows good accuracy. However, it should be remembered that it is not the fastest method [19]. AutoDock uses a Lamarckian genetic algorithm to explore the binding possibilities of a ligand in a binding pocket. The interaction energy of ligand and protein is evaluated using atom affinity potentials calculated on a grid similar to that described by Reynolds et al. [44]. The minimized uncomplexed AChE was used as input structure for the docking simulations. During the docking procedure, all ligand atoms were considered flexible, while protein atoms were kept fixed. The 100 resulting complexes were clustered with an r.m.s.d. tolerance of 0.7 Å. In a second step, low-energy complexes were re-ranked according to the interaction energy calculated with a more detailed energetic model based on the YETI force field [110, 111]. For this second step, the 20 top-ranked complexes of the AutoDock output were selected. The protein structure was kept fixed during the minimization, whereas the ligand was allowed to change its conformation and position in the binding pocket. Applying this minimization, the ligand conformation is relaxed into a neighboring local energy minimum.

AChE stands out as a target to which it is particularly hard to dock. Looking closer at the individual docking modes of the known inhibitors, it is apparent that many of the suggested binding modes are in fact wrong. The AChE binding cavity is large, accommodates many water molecules and more than one defined binding region in the pocket has been identified. The ligand–protein interactions observed in the crystal structure of the AChE complex used for docking consist mainly of van der Waals and hydrophobic interactions, with only one positively charged ligand atom involved in electrostatic interactions. No direct hydrogen bonds between the ligand and the protein have been observed, only water-bridged hydrogen bonds. Minaprine derivatives are fairly symmetrical molecules, with aromatic rings involved in π–π interactions with the protein at both ends of the molecule. π–π stacking is not modeled by the force field employed in AutoDock. It has been observed that hydrogen bonds are particularly important for obtaining correct docking modes in most docking programs [18, 20]. The symmetry of the molecules, the lack of π–π interactions modeled and falsely predicted hydrogen bonds result in a large number of improbable docking poses. To obtain more consistently correct docking results for this particular target, one would need to include specific water molecules during the docking run or post-docking filters to select the correct docking pose. In the present study, the program GRID was used to compare the generated docking poses with the molecular interaction fields. GRID interaction fields were calculated for the binding pocket using the trimethylammonium, methyl, carbonyl, amide and DRY probe. Taking the derived interaction fields as filters along with the complexes generated by the AutoDock procedure, we were able to reproduce the conformation of all co-crystallized inhibitors. As expected, the closest agreement with the experimental data was observed when the protein structure extracted

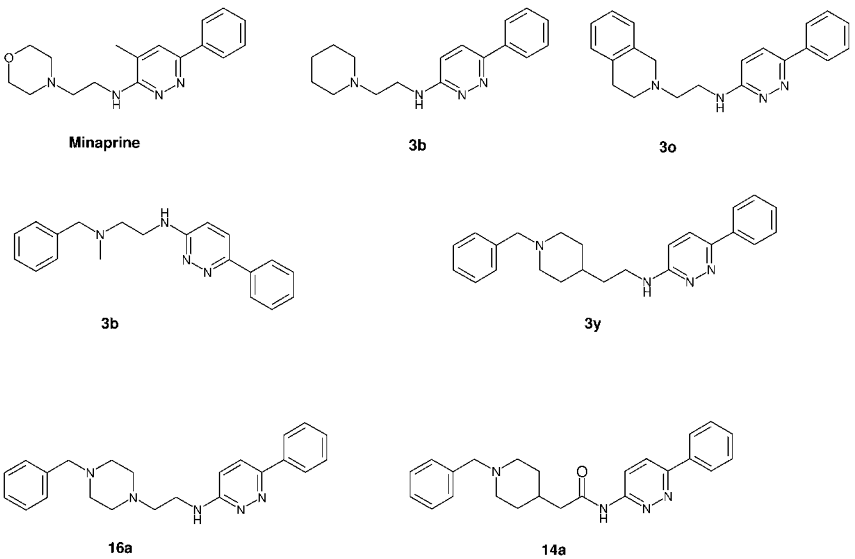


Fig. 11.1 Examples of AChE inhibitors from the training set.

from the corresponding AChE–inhibitor complex was taken as a target for the subsequent docking procedure. All r.m.s.d. values were below 1.2 Å (normally a value below 2.5 Å indicates successful docking).

The ability to predict accurately the binding conformation of tacrine, decamethonium, edrophonium and huperzine gave confidence that we could use our model to evaluate the binding conformation of aminopyridazine compounds (Fig. 11.1).

Since the aminopyridazine derivatives have a comparable size to decamethonium and it is likely that they interact in a similar way with the binding site, we took the protein structure from the AChE-decamethonium complex for further docking.

Figure 11.2 shows the predicted position of an aminopyridazine in comparison with the position of decamethonium observed in the corresponding crystal structure. The hydrophobic parts of the aminopyridazine inhibitors interact with various aromatic residues in the binding pocket. The benzyl ring of the inhibitor displays classical π–π stacking with the aromatic ring of Trp84, thereby occupying the binding site for quaternary ligands. The charged nitrogen of the piperidine moiety makes a cation–π interaction with Phe330 and electrostatic interactions with Tyr121. No direct hydrogen bonds were observed between polar groups of the inhibitor and the binding site. A similar binding orientation within the binding pocket was observed for all other inhibitors (Fig. 11.3).

In the next step, we focused on a possible correlation between the derived AutoDock scoring values and the YETI protein–ligand interaction energies and the observed biological data. In an attempt to correlate the scoring values and

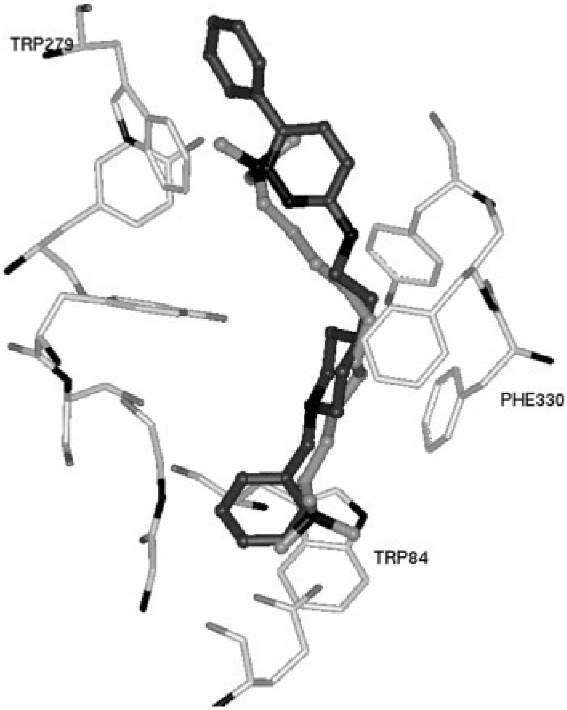


Fig. 11.2 Comparison between the predicted position of the aminopyridazine **3y** (dark gray) and the X-ray structure of the AChE–decamethonium (gray) complex.

the calculated interaction energies with the experimental data, it turned out that no correlation could be detected in the case of the scoring values and only a moderate correlation was observed for the 42 training set molecules. This observation is in agreement with many docking studies in the literature [23, 112], and represents a general problem with scoring functions.

Therefore, the force-field refined docking poses were subsequently extracted from the protein environment and were taken as input for a GRID/GOLPE analysis. Applying the variable-selection strategy incorporated within GOLPE, we obtained a significant 3D QSAR model. The significance was tested by applying a variety of validation procedures. The LOO analysis yielded a correlation coefficient with a cross-validated q^2_{LOO} of 0.94 for the water probe and 0.92 for the methyl probe. In addition, we analyzed the reliability of the model by applying leave-20%-out and leave-50%-out cross-validation (100 runs). Both models are also robust, indicated by high correlation coefficients of $q^2=0.91$ (water probe, SDEP=0.41) and 0.90 (methyl probe, SDEP=0.44) obtained by using the leave-50%-out cross-validation procedure. The statistical results gave confidence that the derived model could also be used for the prediction of novel compounds.

To get an impression of which parts of the AChE inhibitors are correlated with variation in activity, we analyzed the PLS coefficient plots (obtained by

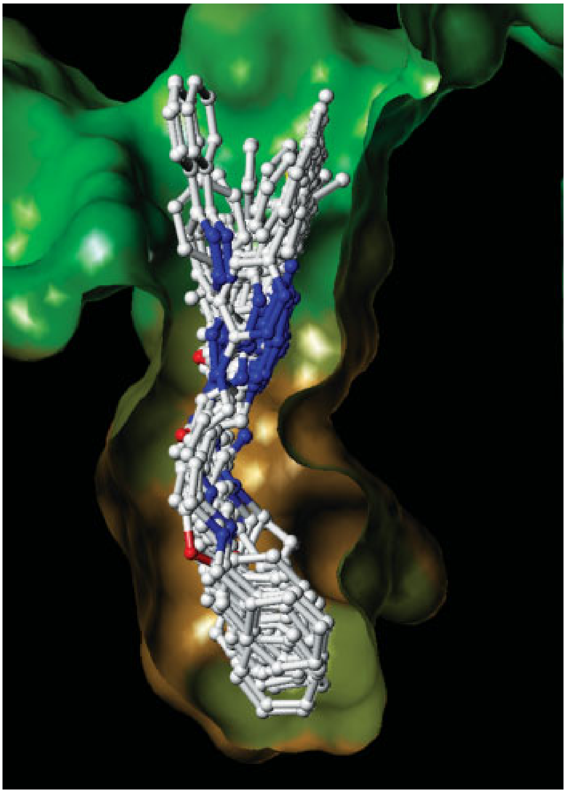


Fig. 11.3 Receptor-based alignment of all investigated inhibitors as obtained by the docking analysis. The solvent-accessible surface of the binding pocket is displayed.

using the water and the methyl probe) and compared them with amino acid residues of the binding pocket. The plots indicate those lattice points where a particular property significantly contributes and thus explains the variation in biological activity data (Fig. 11.4).

The plot obtained with the methyl probe indicated that close to the arylpyridazine moiety, a region with positive coefficients exists (region A in Fig. 11.4). The coefficients were superimposed with the original GRID field obtained for compound **4j** with the methyl probe. The interaction energies in region A are positive, therefore the decrease in activity is due to a steric overlap within this region. Hence it should be possible to obtain active inhibitors by reducing the ring size compared with compound **4j** (which is shown in Fig. 11.4 together with the PLS coefficient maps). For that reason, several molecules containing hydrophobic groups were proposed (Table 11.1).

A second interesting field was observed located above the arylpyridazine moiety in the model obtained using the water probe. Here a region exists where

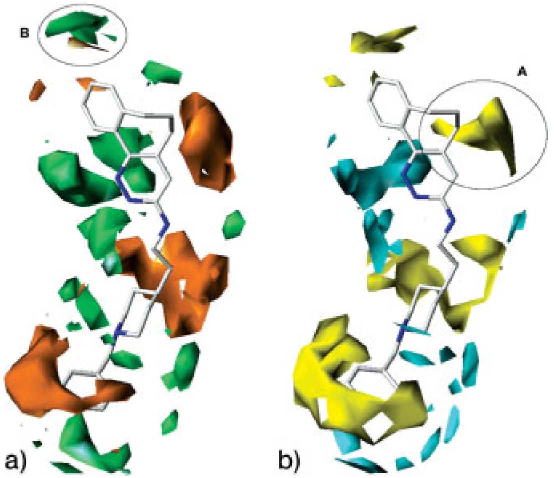
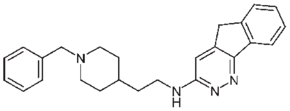
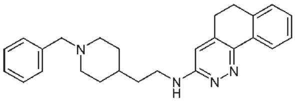
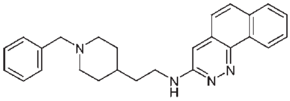
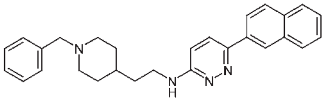
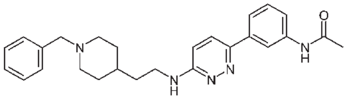
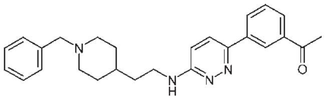
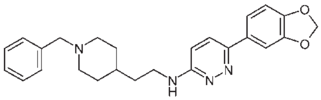


Fig. 11.4 PLS coefficient maps obtained using the water probe (a) and the methyl probe (b). Green and cyan fields are contoured at –0.003 and yellow and orange fields at +0.003 (compound **4j** is shown for comparison).

polar interactions increase activity (region B in Fig. 11.4). After analysis of the entrance of the gorge (the interaction site for the arylpyridazine system), we rationalized the design of compounds bearing polar groups. In the calculated AChE–aminopyridazine complexes we observed two polar amino acid residues (Asn280 and Asp285) located at the entrance of the gorge which could serve as an additional binding site for the substituted arylpyridazine system. To test this hypothesis, several inhibitors possessing polar groups with hydrogen-bond donor and acceptor properties were synthesized and tested. The designed inhibitors were docked into the binding pocket by applying the developed procedure and their biological activities were predicted using the PLS models. In Table 11.1 the predicted and experimentally determined inhibitor activities are listed for the novel compounds. In general, excellent agreement between the predicted and experimentally determined values was observed, indicated by the low $SDEP_{ext}$ values of 0.44 (water model) and 0.40 (methyl model). The reduction in the size of the aminopyridazine ring system resulted in highly potent inhibitors **4g**–**4i**. The molecules of the second series of designed inhibitors containing polar groups were also accurately predicted. The gain in activity compared with the non-substituted compound **3y** (Fig. 11.1) is moderate, indicating that the potential interaction with the two polar residues at the entrance does not play an important role. Since the two residues are located at the entrance of the binding pocket, it may be possible that these residues undergo stronger interactions with water molecules than with the protein side-chains.

In addition, traditional ligand-based 3D models were created in order to compare the results from structure-based and atom-based alignment techniques.

Table 11.1 Designed compounds predicted using the GRID/GOLPE model

Compound	Structure	Observed [a)]	Predicted [b)]	Predicted [c)]
4g		8.00	7.00	7.20
4h		7.41	7.62	7.66
4i		7.66	7.48	7.56
6g		7.24	6.90	6.77
6h		7.24	7.05	7.11
6i		7.27	7.25	7.2
6j		7.14	6.88	6.92

a) Inhibitory activity measured on the AChE of *Torpedo californica* [12].

b) Predicted activity using the water probe model.

c) Predicted activity using the methyl probe model.

The ligand-based alignments were carried out using the superposition program FlexS [2], and compound **4j** as rigid template (in the conformation obtained by the docking) on the one site and the Multifit routine within the SYBYL software and the same template on the other site (Fig. 11.5).

Both 3D QSAR models were generated in the same way (i.e. water probe, SRD/FFD variable selection, statistical validation) as described for the docking-based models. The significance was tested by applying cross-validation procedures. The LOO analysis yielded a correlation coefficient with a cross-validated q^2_{LOO} of 0.77 for the Multifit and an q^2_{LOO} of 0.63 for the FlexS alignment. In addition, we analyzed the reliability of the models by applying the leave-50%-out cross-validation procedure. Both models are also robust, indicated by the correlation coefficients obtained of $q^2 = 0.71$ (Multifit) and 0.57 (FlexS). However, compared with the receptor-based 3D QSAR models, both ligand-based models showed larger deviations when predicting the activities of the six designed inhibitors. The external SDEP values (0.86 and 0.81, respectively) are twice the SDEP value of the docking-based 3D QSAR model (0.44), indicating the reliability of the receptor-based approach.

In conclusion, using the receptor-based 3D QSAR strategy we were able to design potent novel AChE inhibitors which seem to interact simultaneously with the cation–π subpocket of the catalytic site and the peripheral site of the enzyme. Further support for our docking study came from the crystal structure of AChE in complex with donepezil [112]. Like our most potent inhibitors, donepezil also contains a benzylpiperidine moiety which shows a similar position and orientation in the published crystal structure to the corresponding group in our docking results. The comparison of both AChE–inhibitor complexes revealed that both kinds of inhibitors adopt comparable conformations in the narrow binding pocket. As we predicted for our aminopyridazine inhibitors, donepezil makes no direct hydrogen bond to any amino acid residue of the binding pocket. Only water-bridged hydrogen bonds have been detected for donepezil, as proposed for the described aminopyridazine compounds (Fig. 11.6).

11.6 Conclusion

In this chapter, case studies have been reviewed where a combination of 3D QSAR and receptor-based alignments led to predictive and meaningful models. Apart from the good predictive ability, the models derived are also able to indicate which interaction sites in the binding pocket might be responsible for the variance in biological activities. In the last decade, structure-based methods have become major tools in drug design, including lead finding and optimization. It has also been shown that structure-based methods are nowadays able to predict fairly accurately the position of a ligand in its binding site. Apart from the accurate prediction of experimental data, modern docking methods have become more and more efficient. Meanwhile, docking programs are being developed that accomplish docking of highly flexible ligands in a few seconds or minutes on modern PCs. The major problem is still the prediction of the binding affinity, probably limited by the approximation used in today's scoring and force field methods [24–27]. The application of 3D QSAR methods may facilitate the prediction of binding affinities if one has a series of compounds which bind in a similar way to a target protein. Up to now, the imprecise nature of docking and scoring has made blind virtual screening of large number of compounds without any information about true actives or known experimental complex structures a risky exercise. It has been shown by Jacobsson et al. that limited experimental information and proper multivariate statistical treatment of the scoring data dramatically increase the value of these kinds of computations [113]. They

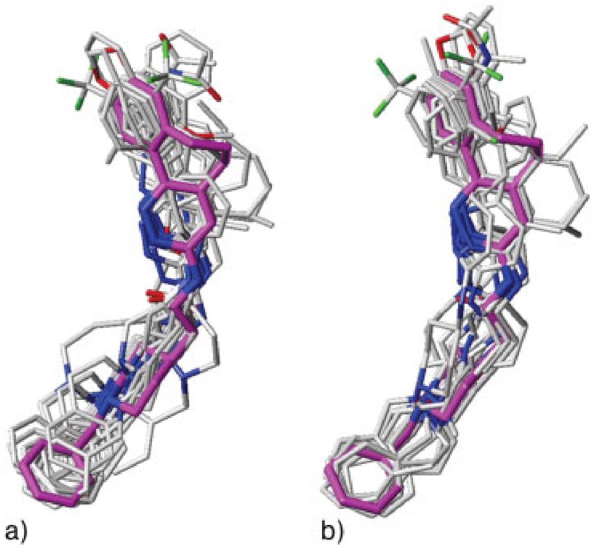


Fig. 11.5 Ligand-based alignments obtained applying the FlexS (a) and SYBYL Multifit (b) program. Compound **4j** is colored magenta.

generated scoring matrices for known actives and potential inactives for four different targets, using docking followed by scoring with seven different scoring functions. Based on these matrices, multivariate classifiers were generated and evaluated with external test sets and compared with classical consensus scoring and single scoring functions. It was found that proper multivariate analysis of scoring data is very rewarding in terms of recall of known actives and enrichment of true actives in the set of predicted actives. Another interesting strategy which might overcome the problem of neglecting the protein information in a 3D QSAR analysis has been described by Gohlke and Klebe [114].

Since a multivariate QSAR analysis considers only the information that applies to the considered data set, advantages are offered in comparison with more elaborate methods. These methods have to consider all influences on ligand binding and must calculate the corresponding amounts correctly. Thus, a multivariate QSAR analysis is able to provide a kind of scoring function valid for a particular data set. Since the reported combined strategy is able to predict biological affinity rapidly, the method can be applied to large ligand series. So long as methods are not developed that are able to solve the affinity prediction problem, structure-based 3D QSAR will remain an exciting strategy for drug design studies. Lastly, with the advent of sophisticated protein homology modeling and available protein crystal structure information, it is hoped that 3D QSAR models may provide insight into which intramolecular interactions are important and consistent with a proposed binding mode.

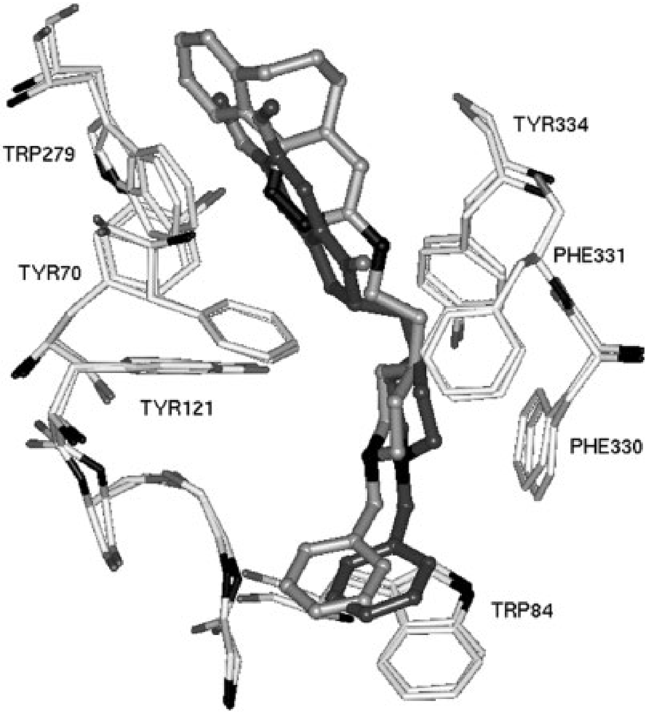


Fig. 11.6 Comparison of the experimentally determined AChE–donepezil complex and the predicted position of the aminopyridazine **4j**. The aminopyridazine (light gray) shows a similar conformation and interaction pattern to donepezil (dark gray) in the corresponding X-ray structure.

Acknowledgments

The author wishes to thank Jean-Marie Contreras, Camille-G. Wermuth (Prestwick Chemical, Illkirch, France), Dirk Classen-Houben and Hans-Dieter Höltje (Heinrich-Heine-Universität Düsseldorf, Germany) for their cooperation and support of this project. Birgit Schlegel (Heinrich-Heine-Universität Düsseldorf, Germany) is acknowledged for critical review of the manuscript.

References

1 Miller, M.D., Sheridan, R.P., Kearsley, S.K. SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.

2 Lemmen, C., Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215–232.

3 Jain, A.N. Ligand-based Structural Hypotheses for Virtual Screening. *J. Med Chem.* **2004**, *47*, 947–61.

4 Hansch, C., Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC, **1995**.

5 Kubinyi, H. QSAR and 3D QSAR in Drug Design. *Drug Discov. Today* **1997**, *2*, 457–467.

6 Cramer, R.D., III, Patterson, D.E., Bunce, J.D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

7 Folkers, G., Merz, A., Rognan, D. CoMFA: Scope and Limitations. In *3D QSAR in Drug Design. Theory, Methods and Applications*, Kubinyi, H. (ed.), ESCOM Science Publishers, Leiden, **1993**.

8 Klebe, G., Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70–80.

9 Klebe, G., Abraham, U., Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.

10 Martin, Y.C. 3D QSAR: Current State, Scope and Limitations. *Perspect. Drug Discov. Des.* **1998**, *12*, 3–23.

11 Norinder, U. Recent Progress in CoMFA Methodology and Related Techniques. *Perspect. Drug Discov. Des.* **1998**, *12*, 15–39.

12 Kim, K.H., Greco, G., Novellino, E. A Critical Review of Recent CoMFA Applications. *Perspect. Drug Discov. Des.* **1998**, *12*, 257–315.

13 Podlogar, B.L., Ferguson, D.M. QSAR and CoMFA: a Perspective on the Practical Application to Drug Discovery. *Drug Des Discov.* **2000**, *1*, 4–12.

14 Kubinyi, H. (ed.). *3D QSAR in Drug Design. Theory, Methods and Applications*, ESCOM Science Publisher, Leiden, **1993**.

15 Kubinyi, H., Folkers, G., Martin, Y.C. (eds). *3D QSAR in Drug Design. Ligand–Protein Interactions and Molecular Similarity.* Kluwer/ESCOM, Dordrecht, **1998.**

16 Kubinyi, H., Folkers, G., Martin, Y.C. (eds). *3D QSAR in Drug Design. Recent Advances.* Kluwer/ESCOM, Dordrecht, **1998.**

17 Kuntz, J.D. Structure-based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078–1082.

18 Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *11*, 935–949.

19 Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J. Distributed Automatic Docking of Flexible Ligands to Proteins. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

20 Verdonk, M.L., Cole, J.C., Hartshorn, M., Murray, C., Taylor, R.D. Improved Protein–Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609–623.

21 Meng, E., Shoichet, B.K., Kuntz. I.D. Automated Docking with Grid-based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505–524.

22 Kontoyianni, M., McClellan, L.M., Sokol, G.S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.

23 Tame, J.R.H. Scoring Functions: A View from the Bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99–108.

24 Böhm, H.J. Prediction of Binding Constants of Protein-ligands: a Fast Method for the Prioritization of Hits Obtained from De-novo Design or 3D Database Search Programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.

25 Wang, R., Lu, Y., Fang, X., Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800

Protein–Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.

26 Perola, E., Walters, W. P., Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins* **2004**, *56*, 235–249.

27 Gohlke, H., Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small Molecule Ligands to Macromolecular Receptors. *Angew. Chem. Int. Ed.* **2002**, *41*, 2644–2676.

28 Masukawa, K. M., Kollman, P. A., Kuntz, I. D. Investigation of Neuraminidase-Substrate Recognition Using Molecular Dynamics and Free Energy Calculations. *J. Med. Chem.* **2003**, *46*, 5628–5637.

29 Huang, D., Caflisch, A. Efficient Evaluation of Binding Free Energy Using Continuum Electrostatics Solvation. *J. Med. Chem.* **2004**, *47*, 5791–5797.

30 Waller, C. L., Oprea, T. I., Giolitti, A., Marshall, G. R. Three-dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.

31 Cho, S. J, Garsia, M. L., Bier, J., Tropsha, A. Structure-based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1996**, *39*, 5064–5071.

32 Vaz, R. J., McLEan, L. R., Pelton, J. T. Evaluation of Proposed Modes of Binding of (2*S*)-2-[4-[[(3*S*)-1-acetimidoyl-3-pyrrolidinyl]oxyl]phenyl]-3-(7-amidino-2-naphthyl)propan oic Acid Hydrochloride and Some Analogs to Factor Xa Using a Comparative Molecular Field Analysis. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 99–110.

33 Sippl, W., Contreras, J. M., Rival, Y. M., Wermuth, C. G. Comparative Molecular Field Analysis of Aminopyridazine Acetylcholinesterase Inhibitors. In *Proceedings of the 12th European Symposium on QSAR – Molecular Modelling and Predicting of Bioactivity*, Gundertofte, K. (ed.), Plenum Press, Copenhagen, **1998.**

34 Sippl, W. Receptor-based 3D Quantitative Structure–Activity Relationships of Estrogen Receptor Ligands. *J. Comput.-Aided. Mol. Des.*, **2000**, *14*, 559–572.

35 Sippl, W. Binding Affinity Prediction of Novel Estrogen Receptor Ligands Using Receptor-based 3D QSAR Methods. *Bioorg. Med. Chem.* **2002**, *10*, 3741–55.

36 Ortiz, A. R., Pisabarro, M. T., Gago, F., Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.

37 Lozano, J. J., Pastor, M., Cruciani, G., Gaedt, K., Centeno, N. B., Gago, F., Sanz, F. 3D QSAR Methods on the Basis of Ligand–Receptor Complexes. Application of Combine and GRID/GOLPE Methodologies to a Series of CYP1A2 Inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *13*, 341–353.

38 Akamatsu, M. Current State and Perspectives of 3D QSAR. *Curr. Top. Med. Chem.* **2002**, *2*, 1381–1394.

39 Duca, S., Hopfinger, A. J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367–1387.

40 Vedani, A., Dobler, M. 5D-QSAR: the Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45*, 2139–2149.

41 Baroni, M., Constantino, G., Cruciani, G., Riganelli, D, Valigli, R., Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): an Advanced Chemometric Tool for Handling 3D QSAR Problems. *Quant. Struct.–Act. Relat.* **1993**, *12*, 9–20.

42 Cruciani, G., Crivori, P., Carrupt, P.-A., Testa B. Molecular Fields in Quantitative Structure–Permeation Relationships. *J. Mol. Struct.* **2000**, *503*, 17–30.

43 Böhm, M., Klebe, G. Development of New Hydrogen-bond Descriptors and their Application to Comparative Molecular Field Analyses. *J. Med. Chem.* **2002**, *45*, 1585–1597.

44 Reynolds, C. A. Wade, R. C., Goodford, P. J. Identifying Targets for Bioreductive Agents: Using GRID to Predict Selective Binding Regions of Proteins. *J. Mol. Graph.* **1989**, *7*, 103–108.

45 Cruciani, G., Watson, K. Comparative Molecular Field Analysis Using GRID Force Field and GOLPE Variable Selection Methods in a Study of Inhibitors of

Glycogen Phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589–2601.

46 Pastor, M., Cruciani, G., Clementi, S. Smart Region Definition: a New Way to Improve the Predictive Ability and Interpretability of Three-dimensional Quantitative Structure–Activity Relationships. *J. Med. Chem.* **1997**, *40*, 1455–1464.

47 Oprea, T. I., Garcia, A. E. Three-dimensional Quantitative Structure–Activity Relationships of Steroid Aromatase Inhibitors. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186–200.

48 R. D. Hoffmann, R. D., Langer, T. Use of the CATALYST Program as a New Alignment Tool for 3D QSAR. In *QSAR and Molecular Modelling: Concepts,Computational Tools and Biological Applications*, Sanz, F., Giraldo, J., Manaut F. (eds), PROUS Science Publishers, Barcelona, 1995, pp. 466–469.

49 Langer, T., Hoffmann, R. D. On the Use of Chemical Function Based Alignment Generation as Input Tool for 3D-QSAR, *J. Chem. Inf. Comput. Sci.*, 1998, *38*, 325–330.

50 Langer, T., Ecker, M., Hoffmann, R. D., Chiba, P., Ecker, G. F. Lead identification for modulations of multidrug resistance based on in silico screening with a pharmacophoric feature model. *Arch. Pharm.* **2004**, *337*, 317–327.

51 Golbraikh, A., Trophsa, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

52 Kubinyi, H., Hamprecht, F.A., Mietzner, T. Three-dimensional Quantitative Similarity–Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553–2564.

53 Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y., Lee, K.-H., Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–53.

54 Norinder, U. Single and Domain Made Variable Selection in 3D QSAR Applications. *J. Chemom.* **1996**, *10*, 95–105.

55 Doweyko, A. M. 3D QSAR Illusions, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587–596.

56 Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M., Marshall, G. R. VALIDATE: a New Method for the Receptor-based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959–3969.

57 Di Santo, R., Costi, R., Artico, M., Massa, S., Ragno, R., Marshall, G. R., La Colla, P. Design, Synthesis and QSAR Studies on *N*-Aryl Heteroarylisopropanolamines, a New Class of Non-peptidic HIV-1 Protease Inhibitors. *Bioorg. Med. Chem.* **2002**, *10*, 2511–2526.

58 Ortiz, A. R., Pisabarro, M. T., Gago, F., Wade, R. C. Prediction of drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.

59 Martin-Santamaria, S., Munoz-Muriedas, J., Luque, F. J., Gago, F. Modulation of Binding Strength in Several Classes of Active Site Inhibitors of Acetylcholinesterase Studied by Comparative Binding Energy Analysis. *J. Med. Chem.* **2004**, *47*, 4471–4482.

60 Murcia, M., Ortiz, A. R. Virtual Screening with Flexible Docking and COMBINE-based Models. Application to a Series of Factor Xa Inhibitors. *J. Med. Chem.* **2004**, *47*, 805–820.

61 Wang, T., Wade, R. C. Comparative Binding Energy (COMBINE) Analysis of OppA–Peptide Complexes to Relate Structure to Binding Thermodynamics. *J. Med. Chem.* **2002**, *45*, 4828–4837.

62 Wang, T., Wade, R. C. Comparative Binding Energy (COMBINE) Analysis of Influenza Neuraminidase–Inhibitor Complexes. *J. Med. Chem.* **2001**, *44*, 961–971.

63 Kim, H. J., Chae, C. H., Yi, K. Y., Park, K. L., Yoo, S. E. Computational Studies of COX-2 Inhibitors: 3D QSAR and Docking. *Bioorg. Med. Chem.* **2004**, *12*, 1629–1641.

64 Waller, C. L., Oprea, T. I., Giolitti, A., Marshall, G. R. Three-dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.

65 Cho, S. J, Garsia, M. L., Bier, J., Tropsha, A. Structure-based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1996**, *39*, 5064–5071.

66 Mügge, I., Podlogary, B. L. 3D Quantitative Structure–Activity Relationships of Biphenyl Carboxylic Acid MMP-3 Inhibitors: Exploring Automated Docking as Alignment Method. *Quant. Struct.-Act. Relat.* **2001**, *20*, 215–223.

67 Costantino, G., Macchiarulo, A., Camaioni, E., Pellicciari, R. Modeling of Poly(-ADP-ribose)polymerase (PARP) Inhibitors. Docking of Ligands and Quantitative Structure–Activity Relationship Analysis. *J. Med. Chem.* **2001**, *44*, 3786–3794.

68 *Cerius^2*. Accelrys, San Diego, CA.

69 Matter, H., Defossa, E., Heinelt, U., Blohm, P.-M., Schneider, D., Müller, A., Herok, S., Schreuder, H., Liesum, A., Brachvogel, V., Lönze, P., Walser, A., Al-Obeidi, F., Wildgoose, P. Design and Quantitative Structure–Activity Relationship of 3-Amidinobenzyl-1*H*-indole-2-carboxamides as Potent, Nonchiral and Selective Inhibitors of Blood Coagulation Factor Xa. *J. Med. Chem.* **2002**, *45*, 2749–2769.

70 Matter, H. Schudok, M., Schwab, W., Thorwart, W., Barbier, D., Billen, G., Haase, B., Neises, B., Weithmann, K. U., Wollmann, T. Tetrahydroisoquinoline-3-carboxylate Based Matrix–Metalloproteinase Inhibitors: Design, Synthesis and Structure–Activity Relationship. *Bioorg. Med. Chem.* **2002**, *10*, 3529–3544.

71 Tervo, A. J., Nyroenen, T. H., Ronkko, T., Poso, A. A Structure–Activity Relationship Study of Catechol-*O*-methyltransferase Inhibitors Combining Molecular Docking and 3D QSAR Methods. *J. Comput.-Aided Mol. Des.* **2003**, 17, 797–810.

72 Tervo, A. J., Nyroenen, T. H., Ronkko, T., Poso, A. Comparing the Quality and Predictiveness between 3D QSAR Models Obtained from Manual and Automated Alignment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 807–816.

73 Egea, P. F., Klahoz, B. P., Moras, D. Ligand–Protein Interactions in Nuclear Receptors of Hormones. *FEBS Lett.* **2000**, *476*, 62–67.

74 Bissantz, C., Folkers, G., Rognan, D. Protein-based Virtual Screening of Chemical Databases: 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

75 Chen, Y. Z., Zhi, D. G. Ligand–Protein Inverse Docking and its Potential Use in the Computer Search of Protein Targets of a Small Molecule. *Proteins* **2001**, *43*, 217–226.

76 Wolohan, P., Reichert, D. E. CoMFA and Docking Study of Novel Estrogen Receptor Subtype Selective Ligands. *J. Comput.-Aided Mol Des.* **2003**, *17*, 313–328.

77 Soderholm, A. A., Lehtovuori, P. T., Nyronen, T. H. Three-dimensional Structure–Activity Relationships of Nonsteroidal Ligands in Complex with Androgen Receptor Ligand-binding Domain. *J. Med. Chem.* **2005**, *48*, 917–925.

78 Ai, N., DeLisle, R. K., Yu, S. J., Welsh, W. J. Computational Models for Predicting the Binding Affinities of Ligands for the Wild-type Androgen Receptor and a Mutated Variant Associated with Human Prostate Cancer. *Chem. Res. Toxicol.* **2003**, *16*, 1652–1660.

79 Moro, S., Braiuca, P., Deflorian, F., Ferrari, C., Pastorin, G., Cacciari, B., Baraldi, P. G., Varani, K., Borea, P. A., Spalluto, G. Combined Target-based and Ligand-based Drug Design Approach as a Tool to Define a Novel 3D Pharmacophore Model of Human A3 Adenosine Receptor Antagonists: Pyrazolo[4,3-*e*]1,2,4-triazolo[1,5-*c*]pyrimidine Derivatives as a Key Study. *J. Med. Chem.* **2005**, *48*, 152–162.

80 Medina-Franco., J. L., Rodrýguez-Morales, S., Juarez-Gordiano, C. A. Hernandez-Campos, A., Castillo. R. Docking-based CoMFA and CoMSIA Studies of Non-nucleoside Reverse Transcriptase Inhibitors of the Pyridinone Derivative Type. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 345–360.

81 Thaimattam, R., Daga, P., Rajjak, S. A., Banerjee, R., Iqbal, J. 3D QSAR CoMFA, CoMSIA Studies on Substituted Ureas as Raf-1 Kinase Inhibitors and its Confirmation with Structure-based Studies. *Bioorg. Med. Chem.* **2004**, *12*, 6415–6425.

82 Sun, W. S., Park, Y. S., Yoo, J., Park, K. D., Kim, S. H., Kim, J. H., Park, H. J. Rational Design of an Indolebutanoic Acid Derivative as a Novel Aldose Reductase Inhibitor Based on Docking and 3D

QSAR Studies of Phenethylamine Derivatives. *J. Med. Chem.* **2003**, *46*, 5619–5627.

83 Kim, H.J., Chae, C.H, Yi, K.Y., Park, K.L., Yoo, S.E. Computational Studies of COX-2 Inhibitors: 3D-QSAR and Docking. *Bioorg. Med. Chem.* **2004**, *12*, 1629–1641.

84 Ragno, R., Artico, M., De Martino, G., La Regina, G., Coluccia, A., Di Pasquali, A., Silvestri, R. Docking and 3D QSAR Studies on Indolyl Aryl Sulfones. Binding Mode Exploration at the HIV-1 Reverse Transcriptase Non-nucleoside Binding Site and Design of Highly Active *N*-(2-Hydroxyethyl)carboxamide and *N*-(2-Hydroxyethyl)carbohydrazide Derivatives. *J. Med. Chem.* **2005**, *48*, 213–23.

85 Assefa, H., Kamath, S., Buolamwini, J.K. 3D-QSAR and Docking Studies on 4-Anilinoquinazoline and 4-Anilinoquinoline Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Inhibitors. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 475–493.

86 Hu, X., Stebbins, C.E. Molecular Docking and 3D QSAR Studies of Yersinia Protein Tyrosine Phosphatase YopH Inhibitors. *Bioorg. Med. Chem.* **2005**, *13*, 1101–1109.

87 Kuo, C.L., Assefa, H., Kamath, S., Brzozowski, Z., Slawinski, J., Saczewski, F., Buolamwini, J.K., Neamati, N. Application of CoMFA and CoMSIA 3D QSAR and Docking Studies in Optimization of Mercaptobenzenesulfonamides as HIV-1 Integrase Inhibitors. *J. Med. Chem.* **2004**, *47*, 385–399.

88 Zhou, Z., Madura, J.D. 3D QSAR Analysis of HIV-1 RT Nonnucleoside Inhibitors, TIBO Derivatives Based on Docking Conformation and Alignment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2167–2178.

89 Sutherland, J., Weaver, D.F. Three-dimensional Quantitative Structure–Activity and Structure–Selectivity Relationships of Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309–331.

90 Carrieri, A., Carotti, A. Barreca, M.L., Altomare, C. Binding Models of Reversible Inhibitors to Type-B Monoamine Oxidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 769–778.

91 Sippl, W., Höltje, H.-D. Structure-based 3D-QSAR – Merging the Accuracy of Structure-based Alignments with the Computational Efficiency of Ligand-based Methods. *J. Mol. Struct. (Theochem)* **2000**, *503*, 31–50.

92 Cinone, N., Höltje, H.-D., Carotti, A. Development of a Unique 3D Interaction Model of Endogenous and Synthetic Peripheral Benzodiazepine Receptor Ligands. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 753–768.

93 Hammer, S., Spika, L., Sippl, W., Jessen, G., Kleuser, B., Höltje, H.-D., Schäfer-Korting, M. Glucocorticoid Receptor Interactions with Glucocorticoids: Evaluation by Molecular Modeling and Functional Analysis of Glucocorticoid Receptor Mutants. *Steroids* **2003**, *68*, 329–339.

94 Classen-Houben, D., Sippl, W., Höltje, H.-D. Molecular Modeling on Ligand–Receptor Complexes of Protein–Tyrosine–Phosphatase 1B. In *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, Ford, M., Livingstone, D., Dearden, J., Van de Waterbeemd, H. (eds), Blackwell, Bournemouth, **2002**.

95 Broer, B.M., Gurrath, M., Höltje, H.-D. Molecular modeling studies on the ORL1-receptor and ORL1-agonists. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 739–754.

96 Sippl, W., Contreras, J.M., Parrot, I., Rival, Y., Wermuth, C.G. Structure-based 3D QSAR and Design of Novel Acetylcholinesterase Inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395–410.

97 Massoulie, J., Pezzementi, L., Bon, S., Krejci, E., Velette, F.M. Molecular and Cellular Biology of the Cholinesterases. *Prog. Neurobiol.* **1993**, *41*, 31–39.

98 Crismon, M.L. Tacrine: First Drug Approved for Alzheimer's Disease. *Ann. Pharmacother.* **1994**, *28*, 744–751.

99 Barner, E.L., Gray, S.L. Donepezil in Alzheimer Disease. *Ann. Pharmacother.* **1998**, *32*, 70–77.

100 Barril, X., Orozco, M., Luque, F.J. Towards improved Acetylcholinesterase Inhibitors: a Structural and Computational Approach. *Mini-Rev. Med. Chem.* **2001**, *1*, 255–266.

101 Rachinsky, T.L, Camp, S., Li, Y., Ekstrom, J., Newton, M., Taylor, P. Molecular Cloning of Mouse Acetylcholinesterase: Tissue Distribution of Alternatively Spliced mRNA Species. *Neuron* **1990**, *5*, 317–327.

102 Bourne, Y., Taylor, P., Radic, Z., Marchot, P. Structural Insights into Ligand Interactions at the Acetylcholinesterase Peripheral Anionic Site. *EMBO J.* **2003**, *22*, 1–12.

103 Sussman, J.L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L., Silman, I. Atomic Structure of Acetylcholinesterase from *Torpedo californica*: a Prototypic Acetylcholine-binding Protein. *Science* **1991**, *253*, 872–879.

104 Harel, M., Quinn, D.M., Nair, H.K., Silman, I., Sussman, J.L. The X-ray Structure of a Transition State Analog Complex Reveals the Molecular Origins of the Catalytic Power and Substrate Specificity of Acetylcholinesterase. *J. Am. Chem. Soc.* **1996**, *118*, 2340–2346.

105 Wermuth, C.-G., Schlewer, G., Bourguignon, J.-J., Maghioros, G., Bouchet, M.J., Moire, C., Kan, J.P., Worms, P., Biziere, K. 3-Aminopyridazine Derivatives with Atypical Antidepressant Serotonergic and Dopaminergic Activities. *J. Med. Chem.* **1989**, *32*, 528–537.

106 Sussman, J.L., Harel, M., Silman, I. Three-dimensional structure of acetylcholinesterase and of its complexes with anticholinesterase drugs. *Chem. Biol. Interact.* **1993**, *87*, 187–197.

107 Harel, M., Sussman, J.L. Quaternary Ligand Binding to Aromatic Residues in the Active Site Gorge of Acetylcholinesterase. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 9031–9035.

108 Raves, M.L., Harel, M., Pang, Y.P., Silman, I., Kozikowski, A.P., Sussman, J.L. Structure of Acetylcholinesterase Complexed with the Nootropic Alkaloid Huperzine A. *Nat. Struct. Biol.* **1997**, *4*, 57–63.

109 Contreras, J.M., Rival, Y., Chayer, S., Bourguignon, J.J., Wermuth, C.G. Aminopyridazines as Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1999**, *42*, 730–741.

110 Vedani, A., Huhta, D.W. A New Force Field for Modeling Metalloproteins. *J. Am. Chem. Soc.* **1990**, *112*, 269–280.

111 Vedani, A., Dunitz, J.D. Lone-pair Directionality of H-bond Potential Functions for Molecular Mechanics Calculations: the Inhibition of Human Carbonic Anhydrase II by Sulfonamides. *J. Am. Chem. Soc.* **1985**, *107*, 7653–7658.

112 Kryger, G., Silman, I., Sussman, J.L. Structure of Acetylcholinesterase Complexed with E2020 (Aricept): Implications for the Design of New Anti-Alzheimer Drugs. *Struct. Fold. Des.* **1999**, *15*, 297–307.

113 Jacobsson, M., Liden, P., Stjernschantz, E., Bostrom, H., Norinder, U. Improving Structure-based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **2003**, *46*, 5781–5789.

114 Gohlke, H., Klebe, G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-based Pair Potentials to a Particular Protein *J. Med. Chem.* **2002**, *45*, 4153–4170.

Part III
Pharmacophores for Hit Identification and Lead Profiling: Applications and Validation

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

12
Application of Pharmacophore Models in Medicinal Chemistry

Fabrizio Manetti, Maurizio Botta, and Andrea Tafi

12.1
Introduction

Many tools and protocols are now available for the scientist involved in the drug discovery process. When the structure of the macromolecular target (usually termed the receptor) is unknown, the ligand-based drug design approach can be applied for different purposes. As an example, given a set of compounds acting through the same mechanism of action (that is, able to bind to the same site of a receptor), one can investigate the chemical features responsible for the activity and summarize them in terms of pharmacophoric models. However, a problem that sometimes arises in pharmacophore-based approaches is the need to take into account possible adverse steric interactions between inactive compounds in a dataset and the target protein counterpart. The most common situation encountered in the literature is connected with the mining of large databases. In this case, the most likely outcome of queries based on relatively simple pharmacophore hypotheses (that contain three or four features) would be very large hit lists of several hundreds of compounds, difficult to evaluate critically. Addition of excluded volume spheres to pharmacophores or ligand-forbidden zones to constrain the models is consequently expected to reduce the number of retrieved hits considerably.

Based on the above considerations, we report in this chapter two case studies where pharmacophore generation and handling plays a pivotal role in finding new hits. In the first example, a classical computational strategy consisting of pharmacophore building, pharmacophore validation, database mining, hit identification and hit optimization is described, aiming at the identification of potent antagonists of the α_1 adrenergic receptor. Additionally, we also report how this original pharmacophore model for α_1 adrenoceptor antagonists evolved towards α_{1d} subtype selectivity. In the second example, in contrast, the rationalization of the antifungal activities of azole compounds is exploited to discuss the importance and utility of adding excluded volume spheres (representing regions of the space forbidden to the ligands) to a pharmacophore.

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann

ISBN: 3-527-31250-1

12.2 Building Pharmacophore Models Able to Account for the Molecular Features Required to Target the α_1 Adrenergic Receptor (α_1-AR) and its Subtypes

12.2.1 A Pharmacophore Model for α_1-AR Antagonists

In recent years, the search for new selective α_1-AR antagonists has increased, owing to their importance in the treatment of hypertension and of benign prostatic hyperplasia (BPH). In fact, α_1-AR blockers have been employed in the treatment of BPH for more than two decades, owing to the significant improvements in lower urinary tract symptoms (LUTS) and flow-rates in patients with bladder outflow (urinary) obstruction [1, 2]. In this context, we first synthesized a new class of arylpiperazine-pyridazinone derivatives [belonging to the same structural class as compounds (**1**) and (**2**); see Table 12.1], found to be active as α_1-AR antagonists. In a second step, the goal was to gain further insight into the structural factors responsible for α_1 affinity, in order to design new ligands with increased selectivity for the α_1 receptor.

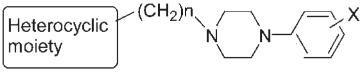
A ligand-based pharmacophore building and development method was applied to rationalize the relationships between the structure of the pyridazinone derivatives and their affinity for the α_1 adrenergic receptor. This pharmacophore model was then used as a three-dimensional query to perform a search into databases of known structures [3] with the aim of finding new hits as a starting point for structural optimization in order to improve the selectivity profile with respect to other G protein coupled receptors (GPCR), such as the α_2 adrenergic and the serotoninergic 5-HT$_{1A}$ receptors.

12.2.1.1 Pharmacophore Building

Our primary interest being the search for new high-affinity ligands, we decided to build a pharmacophore model for antagonists of the α_1-AR and in the first step did not consider any selectivity criteria between the α_1-AR subtypes and between α_1-AR and other GPCR.

The training set for pharmacophore development was chosen according to the Catalyst [4] guidelines (further information on the rules for picking training set compounds can be found at the Accelrys web site [5]). Fourteen molecules were selected from our own class of piperazine-pyridazinone derivatives (with affinity values spanning over about 2.5 orders of magnitude, between 0.60 nM, found for compound **1**, and 180 nM, found for compound **2**). Additional compounds were selected among α_1-AR antagonists reported in the literature. A large number of compounds were found with a biological behavior appropriate for our purposes. However, to ensure the highest homogeneity in the biological data with respect to those of the pyridazinone derivatives (comparable pharmacological protocols used to evaluate α_1-AR affinity of such compounds), two constraints were applied to select training set compounds from the literature:

Table 12.1 Structure and affinity of representative arylpiperazines discussed in the text.



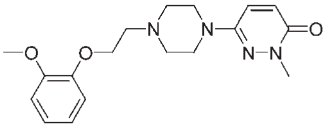
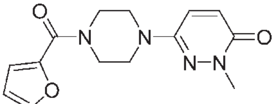
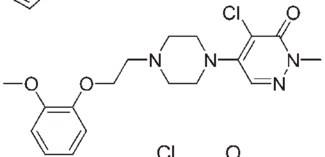
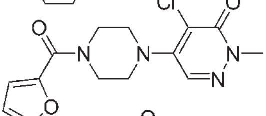
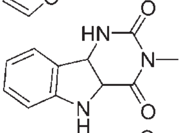
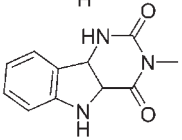
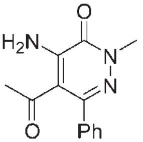
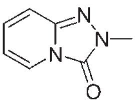
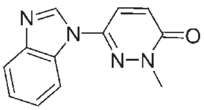
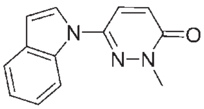
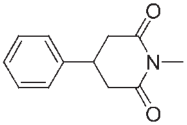
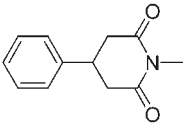
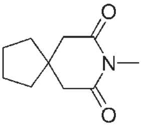
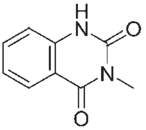
Compound	Heterocyclic moiety	*n*	X	K_i (nM)[a]	Ref.
1		4	2-OMe	α_1 0.60 (1.8)	3
2		4	2-OMe	α_1 180 (88)	3
3		7	2-OMe	α_1 1.4 (9.7)	3
4		7	2-O*i*Pr	α_1 0.052	10, 11
5		2	2-OMe	α_1 0.21 (0.11) α_{1d} 0.36 (0.33)	3, 18
6		2	4-cyclohexyl	α_{1d} 2000 (1810)	18
7		1	2-OMe	α_1 2396 (1200)	9b, 3
8, trazodone		3	3-Cl	α_1 281 (220)	3, 14
9		4	2-OMe	α_1 1.1 (1.3)[b]	16, 17
10		4	2-OMe	α_1 0.9 (0.2)[c]	16, 17

Table 12.1 (continued)

Compound	Heterocyclic moiety	*n*	X	K_i (nM) [a)]	Ref.
11		4	2-OMe	α_{1d} 16 (24)	24
12		2	2-OMe	α_{1d} 1318 (1096)	24
13		4	2,5-diCl	α_{1d} 0.67 (0.33)	19
14		3	2-OCH$_2$CF$_3$	α_{1d} 2.0 (3.1)	21

a) Estimated and predicted affinity values calculated by Catalyst for the training set and test set, respectively, are given in parentheses.

b) The affinity of this compound towards 5-HT$_{1A}$ receptor was 315 nM, with a 5-HT$_{1A}$/α_1-AR ratio of 286.

c) The affinity of this compound towards 5-HT$_{1A}$ receptor was 253 nM, with a 5-HT$_{1A}$/α_1-AR ratio of 281.

(i) the antagonist activity on α_1-AR was tested on rat cortex homogenates; and (ii) inhibition constants (K_i) were calculated according to the Cheng–Prusoff equation [6]: $K_i = IC_{50}/[1+([L]/K_d)]$, where IC_{50} is the concentration of the tested compound that produced a 50% inhibition of specific [^{3}H]prazosin binding to α_1-AR, [L] is the ligand concentration and K_d is its dissociation constant. K_d of labeled prazosin binding to α_1-AR from rat cortex membranes was 0.24 nM.

As a result, 10 additional compounds were added to obtain the final training set constituted by a total of 24 structures. Biological data associated with these compounds, expressed as K_i, ranged between 0.21 nM, found for compound **5**, and 2396 nM, found for compound **7** (Table 12.1). We assumed that all these compounds were acting through the same mechanism of action at the same binding site.

With no experimental data at hand (X-ray crystallographic atomic coordinates, NMR structure data, etc.) describing the biologically relevant conformations of the selected compounds, conformation ensembles of each compound were generated with the program MacroModel [7], within a range of 20 kcal mol^{-1} above the calculated global minimum energy conformation. At the time of this work, the generation of conformers with third-party software was clearly indicated since within the Catalyst 4.5-Confirm module, conformational models of piperazine derivatives were misleading: owing to the limitations of the built-in poling algorithm, the correct identification of a diequatorial substituent arrangement within the 1,4-disubstituted piperazine ring was not feasible [8]. However, it should be mentioned that in recent Confirm versions this problem has been resolved.

Compounds and their conformational models were imported to Catalyst and subjected to the HypoGen routine to build chemical feature-based pharmacophore models using "hydrogen bond acceptor lipid" (HBA), "hydrogen bond donor" (HBD), "positive ionizable" (PI), "ring aromatic" (RA) and "hydrophobic" (HY) as possible features. Two additional constraints were set: (i) because of the molecular flexibility and functional complexity of the training set, only pharmacophores containing five features should be considered; and (ii) the program was forced to include a positive ionizable feature in the composition of hypotheses, on the basis of the literature reporting a basic atom (usually a nitrogen) as a critical structural determinant for α_1-AR antagonistic activity.

12.2.1.2 **Pharmacophore Analysis**

The pharmacophore generation routine provides the user with a series of parameters that allow for a preliminary evaluation of the statistical significance of the pharmacophoric hypotheses generated. Using the hypothesis cost function analysis, based on the simplicity of the models and on their capacity for predicting the affinity of the molecule with a small deviation from the experimentally determined value, we could determine that the SAR signal within this training set was strong. The reliability of these models was further confirmed with the correlation coefficient and the root mean square deviation of affinity data ($r=0.92$ and r.m.s.d. = 0.89, respectively, for the first-ranked hypothesis). Using the top-scoring pharmacophore, all but one compound in this training set showed a deviation between experimental and predicted affinity values of <3. The sole exception of a 7-fold difference was found for compound **3** (Table 12.1), indicating an impressive and reliable ability of the pharmacophore model to estimate affinities of the training set compounds toward α_1-AR.

The most active compound in the training set (**5**) was able to map all the pharmacophoric elements of the model (Figure 12.1). The *o*-methoxyphenyl substituent matched the hydrophobic moiety of the model constituted by both HY1 and HY2, whereas the most basic nitrogen atom of the piperazine ring mapped the positively ionizable group, PI. The terminal condensed heterocyclic system satisfied both the hydrogen-bond acceptor feature with one of its carbonyl groups and with its distal phenyl ring corresponding to the hydrophobic region HY3. In such an orientation, the estimated affinity of 0.11 nM was in very good agreement with the experimental value (0.21 nM). The strong biological effect of this molecule suggested that it possesses many or all of the molecular features required for affinity and that, moreover, the pharmacophore model correctly estimated all the functions necessary for the major interactions between antagonists and receptor.

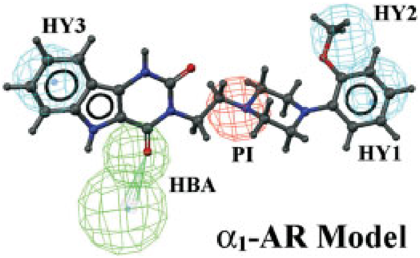


Fig. 12.1 Compound **5** (with the highest α_1-AR affinity among the training set compounds) superposed on the pharmacophoric model for α_1-AR antagonists.

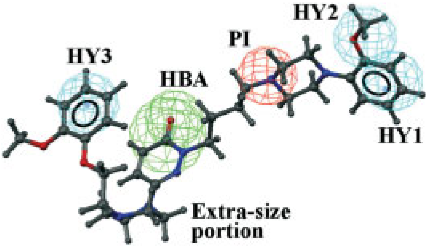


Fig. 12.2 Matching between compound **1** and the pharmacophoric model for α_1-AR antagonists. In this orientation, the extra-size portion of the molecule is represented by the piperazine ring linked to the pyridazinone moiety.

Superposition of a representative example of the pyridazinone-piperazine derivatives (compound **1**, the most active derivative belonging to the first generation of pyridazinones from our research group) on the model showed an orientation for the arylpiperazine portion similar to that found for the best antagonist (Figure 12.2).

In fact, the HY1–HY2–PI system of features was occupied by the ortho-substituted phenylpiperazine moiety. Differently, in order to present a binding pattern similar to that for compound **5** (BA, HY3), this class of compounds was characterized by an “extra-size” portion of the molecule compared with the best antagonist and other small compounds. In a first orientation, this extra-size was represented by the methoxyphenoxyethyl or furoyl portion of the molecule, whereas in a second orientation, the extra-size moiety was the piperazinyl ring bound to the pyridazinone moiety. The fact that small and very active ligands have been found to overlap the pharmacophoric model perfectly suggested that the extra-size portion of the pyridazinone derivatives is as a chemical feature unnecessary for affinity, probably representing a molecular portion able to contact particular regions of the receptor assigned to modulate the biological properties of these compounds.

Taken together, these considerations suggested that a high-affinity α_1-AR antagonist should be characterized by the following three-dimensional structural properties: (i) a substituted phenyl ring (preferably at the ortho position) is required to interact with the HY1–HY2 hydrophobic regions of the model, (ii) a basic, positively ionizable nitrogen atom is necessary to interact with a carboxylate group of an aspartic residue on the third transmembrane helix of the receptor, (iii) a polar group, part of the terminal heterocyclic moiety, corresponding to the hydrogen-bond acceptor feature of the model and (iv) an additional hydrophobic region of the model (HY3) accommodating (portions of) the terminal (heterocyclic) moieties. These results are in good agreement with the generally accepted statement, originally described through the DeMarinis pharmacophore model, that an α_1-AR antagonist interacts with the corresponding receptor with a three-pocket binding pattern [9]: the most important structural element is a protonated nitrogen atom, while the remaining molecular portions of the molecule fit two binding pockets almost symmetrically located with respect to the positively ionizable group.

12.2.1.3 **Validation of the Pharmacophore Model**

In order to assess the statistical significance of the model, a randomization trial procedure (called *catScramble*) derived from the Fischer method was performed. Results after scrambling affinity values showed that there is at least a 95% chance that the best model reports a true correlation between structural and biological data.

In addition to the above-mentioned validation test, an independent test set of compounds (11 pyridazinone derivatives and six molecules collected from the literature) was used to assess further the validity and predictive power of the pharmacophore hypothesis. Test compounds were selected based on the same criteria under which the training set molecules were chosen with special attention that affinity values had been derived from the same biological assay (ability to displace radiolabeled prazosin from α_1-AR on rat cerebral cortex). This additional validation step confirmed the high predictive power of the pharmacophore [3]. A strong point of this model is that it has accommodated much of the structure–affinity relationships of α_1-AR antagonists belonging to the arylpiperazine class. The importance of the substitution in the ortho position has emerged, in addition to the influence of the terminal heterocyclic moiety in defining the α_1-AR antagonist activity. Finally, the distance between the arylpiperazine and the terminal fragment were demonstrated to be crucial for affinity. Taken together, all these findings suggested slight modifications of parent compounds such as **1**–**3**. Variations of the arylpiperazine scaffold (i.e. the insertion on the phenyl ring of substituents larger than a methoxy group, such as an ethoxy or isopropyloxy moiety, better filling the HY1 region of the pharmacophore model), associated with the presence of structural features profitable for affinity towards α_1-AR (such as the furoyl moiety as the terminal molecular portion and a heptyl spacer) led to the discovery of ligands with affinity in the very low nanomolar range. As an example of the utility of our first model, compound **4** bearing an isopropyloxyphenylpiperazine and a furoylpyridazinone as the terminal heterocyclic moiety was later found to exhibit an even lower affinity toward α_1-AR (0.052 nM, see Table 12.1) [10, 11].

12.2.1.4 **Hit Search Through Database Mining**

Having available a reliable and robust pharmacophore model [3], it was used as a three-dimensional query to filter databases of known structures (such as the NCI, Maybridge and MiniBioByte collections, provided by Accelrys along with Catalyst) and possibly to find new and original structural motifs able to fulfil the functional and spatial constraints imposed by the model itself. The database search was also meant to assess further the validity of the model. Among the 486 compounds extracted from the databases (about 0.27% of the total), the pharmacophore model identified compounds known to have a relevant α_1-AR affinity and characterized by different scaffolds. Examples such as carpipramine (a classical 6–7–6 tricyclic antidepressant) [12] and ergotamine [13] have been reported as α_1 adrenoceptor blockers. Trazodone [14] (**8**, Table 12.1) and its open analog etoperidone [15], characterized by anti-α_1-adrenergic activity expressed as the ability to displace tritiated prazosin from α_1-AR in the rat cortical membranes, were also picked up by the pharmacophore search procedure. Although the affinity of trazodone towards α_1-AR was not of great interest – low micromolar range (0.28 μM) – its structure is worthy of further investigation. The model predicted the affinity of such a compound (0.22 μM) in good agreement with the experimental value, suggesting that the poor affinity was due to partial mapping of the *m*-chlorophenyl ring to the HY1–HY2 hydrophobic part of the pharmacophore, together with an incomplete fit to the terminal HY3 feature. In other words, the structure of trazodone, although possessing the chemical features to map all the pharmacophoric regions, suffers from the presence of a reduced length spacer between the phenylpiperazine moiety and the terminal heterocycle. On the basis of these results, in an effort to improve the adrenoceptor binding properties of trazodone and taking into account the suggestions of the SAR data for α_1-AR antagonists (i.e. the role of the ortho substitution on the phenyl ring bound to the piperazine ring, the influence of the alkyl spacer length and the existence of an extra-size molecular portion in compounds similar to arylpiperazinylalkylpyridazinones), the structure of trazodone was modified to meet the three-dimensional structural requirements imposed by our pharmacophore model for α_1-AR antagonists. The distance between the phenylpiperazine ring and the terminal moiety was increased either by inserting a pyridazinone ring directly linked to the terminal moiety or by lengthening the alkyl spacer up to a seven-membered chain. The terminal system of trazodone was simplified in an imidazole, benzimidazole and indole nucleus, while the substituents and substitution pattern on the phenyl ring attached to the piperazine were varied in several ways (i.e. *o*-methoxy and *o*-chloro instead of the *m*-chloro substituent of trazodone). As a result of this structural optimization [16, 17], we obtained derivatives with affinity values for α_1-AR comparable to those of the parent compounds. While these findings partially disappointed our expectations of finding compounds with improved affinity, the biological profile of the new compounds was noteworthy. In fact, two of the new arylpiperazines (**9** and **10**, Table 12.1) were characterized by 5-HT_{1A}/α_1 affinity ratios of 286 and 139, respectively. This increased selectivity ratio highlighted the importance of the new arylpiperazines, since the development of novel selective and potent α_1-AR antagonists bearing those chemical features is still a difficult task. In fact, *o*-methoxy- or *o*-chlorophenylpiperazinyl derivatives with an alkyl spacer of three or four methylene units, characterized by appreciable affinity for α_1-AR, are usually good ligands also for the 5-HT_{1A} serotoninergic receptor and therefore hindering α_1 selectivity.

12.2.2 Towards a Pharmacophore Model for the α_{1d}-AR Subtype

12.2.2.1 A Preliminary Model

The five-feature pharmacophore described above was compared with theoretical three-dimensional structure models of the complexes between pyridazinone inhibitors and the three α_1-AR subtypes derived from molecular dynamics simulations [9b]. However, only a qualitative comparison could made between these complexes and the pharmacophore model, since no structural information (geometric constraints between the structural elements of the complexes) was available. In agreement with both the DeMarinis model [9a] and the model we developed for α_1-AR antagonists [3], each of the complexes showed a three-subsite binding motif accommodating the inhibitors: (i) a hydrophobic pocket, corresponding to the HY1–HY2 system of the pharmacophore, to accommodate the substituted phenyl ring attached to the piperazine; (ii) an aspartic acid of the third transmembrane domain, to accommodate the most basic nitrogen atom of the piperazine ring through a hydrogen-bond interaction; and (iii) a polar binding pocket, to accommodate, in turn, the pyridazinone or the isoxazolopyridazinone moiety of the inhibitor. In particular, the complex representing the structure of the α_{1d}-AR showed a hydrogen-bond interaction between the carbonyl group of the pyridazinone ring of the inhibitors and an arginine residue in the seventh transmembrane domain, comparable to the interaction of the pyridazinone carbonyl group with the hydrogen-bond acceptor feature of the pharmacophoric model.

These considerations suggested that our model incorporates the most relevant structural features of the inhibitors of the α_{1d} adrenoceptor subtype.

To test this hypothesis, we investigated a new series of phenylpiperazines, bearing a pyrimido[5,4-*d*]indolo group as the terminal heterocyclic moiety like compound **5**, showing preferential affinity for the α_{1d}-AR subtype. A selection of 16 compounds, with affinity values spanning over 3.5 orders of magnitude, was used to generate a set of 10 five-feature pharmacophore hypotheses [18] that we compared with the previously described five-feature pharmacophoric

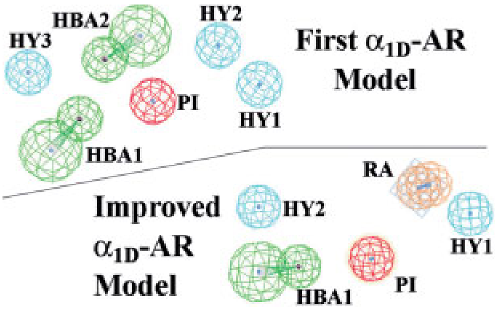


Fig. 12.3 Comparison between the first six-feature pharmacophoric model for α_{1d}-AR antagonists and the improved five-feature pharmacophoric model for the same α_1-AR subtype antagonists.

model for α_1-AR antagonists. A hierarchical cluster analysis of the pharmacophores showed that models 1–3 shared the same features at very similar spatial locations, suggesting that they could be considered as equivalent to each other. On the other hand, the fourth model showed a different feature composition and belonged to a different cluster. However, it was noted that for many of the compounds used to generate the models, the same conformer of each compound matched either the first or the fourth hypothesis with the highest fit value. This evidence, combined with the fact that models 1 and 4 shared four out of five pairs of features located almost at the same positions in the three-dimensional space, led us to merge the two models to obtain a new six-feature pharmacophore model (Figure 12.3, upper left corner).

The major difference between this model and the original pharmacophore for α_1-AR antagonists was the presence of a second hydrogen-bond acceptor feature, while the remaining features underwent only a slight variation of their spatial locations. The ‘Regress Hypothesis’ routine of Catalyst was then subsequently used to allow the model to estimate and predict the affinity of compounds of (and external to) the training set. A preliminary evaluation of the predictive properties of the new model showed that affinity values for several pyrimidoindoles were estimated in very good agreement with the experimental data. A correlation coefficient of 0.91 highlighted the power of the model in estimating affinity values of the whole training set of 16 compounds. Moreover, the model was also able to predict well the affinity of α_{1d}-AR inhibitors taken from the literature, such as SNAP 8719 (predicted 6.2 nM; determined 1.6 nM) and BMY 7378 (predicted 1.9 nM; determined 6.3 nM), whereas the affinities of smaller compounds, such as SKF 104856 (predicted 86 000 nM; determined 1.6 nM) and discretamine (predicted 14 000 nM; determined 25 nM), were underestimated. The main reason for the poor predictions was the inability of the smaller compounds to attain all the pharmacophore features as a consequence of their overall reduced size.

Although this model was characterized by excellent statistical parameters and was able to predict affinity data of a large test set also containing well-known α_{1d}-AR inhibitors (such as BMY 7378, SNAP 8719 and additional compounds from the literature), it suffered from the inability to rationalize variations in affinity due to different substituents at the para position of the phenyl ring attached to the piperazine nucleus of the pyrimido[5,4-*d*]indole. Moreover, from a recent report based on molecular docking and dynamics simulations on imido derivatives [19], two suggestions arose that could be used for building a second-generation pharmacophore model for α_{1d}-AR antagonists. In particular, the authors stated that two hydrogen bonds involving both the carbonyl groups of the glutarimido moiety were unlikely to occur at the same time, in disagreement with our pharmacophoric model for α_{1d}-AR inhibitors (see Figure 12.3). In addition, one of the most relevant interactions resulting from the molecular dynamics calculations was found to be an aromatic contact between the 2,5-dichlorophenyl substituent and a Phe residue of the α_{1d}-AR model. It is described in our pharmacophore as a more generic hydrophobic interaction involving the substituted phenyl ring attached to the piperazine nucleus. All these considerations prompted us to retrace the steps performed to build the pharmacophore model for α_{1d}-AR inhibitors and check for any possibility of improving it.

It should be mentioned that preliminary calculations to generate the pharmacophore model for α_{1d}-AR antagonists [18] produced models that should be reanalyzed in this context, because they were in partial agreement with what was reported in the literature [19]. In particular, (i) the six-feature pharmacophore for α_{1d}-AR inhibitors was obtained by merging two complementary five-feature models, one of them containing only one hydrogen-bond acceptor group, and (ii) a similar pharmacophore model with an equal number of chemical features (i.e. five) and similar physicochemical properties (i.e. hydrogen-bond acceptor, hydrophobic, a positive ionizable group and similar distribution in the three-dimensional space) was also obtained by using the HipHop routine implemented in Catalyst (not shown). However, it was characterized by the presence of a ring aromatic feature (belonging to an RA–HY pair of features and mapping the phenyl ring linked to the piperazine nucleus and its ortho substituent) instead of

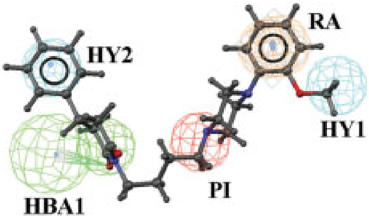


Fig. 12.4 Superposition pathway of the glutarimido derivative **11** into the new pharmacophoric hypothesis for α_{1d}-AR antagonists.

one of the hydrophobics constituting the HY1–HY2 system of our final model for α_{1d}-AR antagonists. Taken together, these data highlighted that pharmacophore models for α_{1d}-AR antagonists, characterized by only one hydrogen-bond acceptor group (corresponding to a carbonyl moiety of a pyridinedione ring) and by an aromatic interaction (involving the phenyl ring attached to the piperazine nucleus), were already hypothesized during our previous calculations. However, they have not been investigated further because of the lower statistical quality in comparison with the other pharmacophore models found during the same calculation runs.

12.2.2.2 **An Improved (Simplified) Model**

Based on the above-mentioned considerations, we planned to use the same training set (16 compounds) from which the first pharmacophore for α_{1d}-AR antagonists was derived, enlarged with five compounds belonging to a new class of glutarimido derivatives such as **11**, to re-run the hypothesis generation (HypoGen) routine within Catalyst. The new training set consisted of 21 compounds with affinity data for α_{1d}-AR spanning over five orders of magnitude (from 0.36 nM found for compound **5** to 2000 nM found for compound **6**). Only five-feature models were requested. The first-ranked pharmacophore showed a correlation of 0.94 between actual (experimental) and estimated affinity and was constituted by two hydrophobic regions (HY1 and HY2), a hydrogen-bond acceptor group (HBA), an aromatic ring (RA) and a positively ionizable feature (PI). A comparison between the new model and the previous pharmacophore for α_{1d}-AR antagonists (Figure 12.3) showed several common features. In particular, the HY1, PI, HBA1 and HY3 regions of the previous model corresponded to identical features of the new model (HY1, PI, HBA1 and HY2, respectively), with almost the same 3D constraints. An aromatic ring was found in the new model instead of the HY2 feature of the previous pharmacophore. Finally, HBA2 of the previous model disappeared and was not replaced by any other feature in the new model, decreasing the total number of features from six to five. It was important to note that the improved model for α_{1d}-AR inhibitors was very similar to the original pharmacophoric model for α_1-AR inhibitors, supporting the hypothesis that the last model incorporated the main features that a compound should possess to show a good affinity toward α_{1d} adrenoceptors.

Figure 12.4 summarizes the superposition pattern of the glutarimido derivatives on the new pharmacophore model for α_{1d}-AR antagonists. These compounds show the *o*-methoxyphenyl group matching the HY1–RA system, whereas the most basic nitrogen atom of the piperazine ring maps to the positively ionizable feature (PI) of the model. The remaining two features are mapped by the 4-substituted piperidinedione system that represents the terminal heterocyclic portion of this family of compounds. In particular, while the hydrogen-bond acceptor group (HBA) is matched by one of the carbonyl moieties, the terminal hydrophobic region (HY2) accommodates the phenyl ring. It is important to note that the glutarimido ring was found in the sofa conformation with the C3 atom pointing out of the plane defined by the remaining atoms. Moreover, the phenyl group at position 3 assumes the so-called equatorial parallel conformation in the 4-phenyl derivatives. This is different from the 4-phenyl-4-methyl derivatives where an equatorial perpendicular conformation is generally found, in agreement with data reported for phenylglutaric anhydride derivatives by Altona and co-workers [20]. The pharmacophore was able to account for the variation of affinity resulting from different alkyl spacers. In fact, when the polymethylene chain was reduced from five, four or three carbon atoms to an ethyl spacer, the affinity underwent a marked decrease. This was an unexpected result, if one compares affinity data of **12** and **5**, sharing the *o*-methoxyphenylpiperazinylalkyl scaffold and a six-membered terminal heterocyclic ring bearing two carbonyl groups at positions 2 and 6. Inspection of the superposition pattern of these compounds reveals good complementarity between the planar tricyclic moiety of **5** and the HBA–HY2 system, whereas the non-planar phenylglutarimide group of **12** showed little matching with both of the two terminal features of the model. This finding suggests that the shape of the terminal moiety could be crucial in influencing the affinity of these compounds for α_{1d}-AR (see below). On the other hand, when the distance between the arylpiperazine moiety and the terminal heterocyclic group increases, profitable interactions between the phenylglutarimide and the model are found (as an example, compound **11** shows an affinity of 16 nM).

Replacement of the *o*-methoxy group at the phenylpiperazino moiety with a chlorine atom led to comparable or slightly lower affinity, in agreement with many other experimental data reported in the literature. This result is validated by the model, showing that both the methoxy and chlorine groups interacted in a similar way with the HY1 pharmacophoric feature.

To check the reliability of the model and to assess its predictive power, a validation step was performed by predicting affinity data for a large test set of compounds collected from three different sources. (i) Glutarimido derivatives belonging to the same class of compounds used to build the training set were evaluated against the pharmacophore model to predicting their affinity values. The results obtained were in good agreement with experimental data, supporting the pharmacophore hypothesis. (ii) The new model overcomes the major limitation of the previous model (inability to evaluate affinity of compounds bearing para substituents on the phenyl ring attached to the piperazine nucleus). In fact, owing to a slightly different location of the features (with respect to the original pharmacophore model) mapped by the phenylpiperazinyl system (RA and HY1), the new pharmacophore is able to account for the influence of para substituents on affinity. As an example, derivatives of **6** bearing an isopropyl, *tert*-butyl, *sec*-butyl and butyl group as the X substituent, were predicted to exhibit an affinity of 37, 17, 30 and 40 nM, respectively. This is in good agreement with experimental data (35, 20, 63 and 251 nM, respectively) [18]. In these cases, whereas the basic nitrogen and the terminal tricyclic moiety showed a precise superposition to PI, HBA and HY2, respectively, only a partial fit was found for both RA and HY1 by the phenyl ring and its para substituent, respectively. (iii) Some imido derivatives such as **13**, recently reported in the literature [19], structurally similar to the compounds described in this paper, were also used as part of the test set to check the robustness of the pharmacophore model. Here also the importance of the alkyl spacer in determining affinity data was evidenced and represented by the pharmacophore model. Compounds with an ethyl spacer were predicted to be significantly less active than compounds with propyl and butyl spacers. Calculated (predicted) affinities for such compounds were comparable to the experimental data (Table 12.1), confirming the high predictivity of the model. The power of the pharmacophore model to predict correctly classes of compounds different from those constituting the training set was further assessed by calculating the affinity value for several arylpiperazine derivatives bearing a substituted uracil moiety as the terminal heterocyclic group, such as **14** [21]. The predicted affinity for five compounds showed by an error factor (ratio between calculated and actual affinity values or vice versa) between 1.5 and 3.0 (as an example, the affinity of **14** was calculated to be 3.1 nM, in agreement with the experimental value of 2.0 nM). Although the affinity values of these literature compounds and our derivatives were the result of different experimental measurements (displacement of [^{3}H]prazosin from cloned human α_{1d}-AR in membranes from CHO cells versus displacement of [^{125}I]BE 2254 from cloned human α_{1d}-AR in membranes from HEK293 cells, respectively), the good predictive power demonstrated by the pharmacophore model for the literature compounds further supported the model itself.

In summary, the new pharmacophore model for α_{1d}-AR antagonist was able to account for the major structure–activity relationships of substituted phenyl piperazinylalkyl derivatives with variable heterocycles as the terminal portions. The model clearly highlighted that the distance between the arylpiperazine moiety and the terminal heterocyclic fragment was a critical element to influence the affinity. The high affinity of compounds belonging to the class of compound **5** were also rationalized in terms of the shape of the terminal moiety. Planar systems (as in **5**) allowed for good interactions with all the pharmacophoric features, even when the spacer was short (two carbon atoms), whereas compounds with a non-planar phenylglutarimido moiety required at least a propyl spacer to interact profitably with the pharmacophore regions. However, as previously hypothesized independently by ourselves and other research groups on the basis of SAR analyses, in addition to the shape of the terminal heterocyclic moiety, the size and length of the alkyl spacer simultaneously play an important role in determining affinity of a compound for α_{1d}-AR. As an example, many of the above-mentioned imido derivatives [19], characterized by the spiro moiety of BMY 7378 as the terminal fragment bound to an ethyl spacer, showed affinity values for α_{1d}-AR comparable to **5** and higher than those found for compounds such as **12**. This finding evidenced that a combination of the spiro-glutarimido moiety with an ethyl spacer was more profitable for affinity than a phenylglutarimido moiety linked to the same spacer, suggesting that the smaller size of compound **5** in combination with a short chain was preferred for affinity. On the other hand, on lengthening the spacer, the affinity values increased for compounds belonging to the same class as compound **12**, whereas no significant improvement (or slight decrease) of affinity was found for imido derivatives [19].

The different spatial position of the HY1–RA features (relative to the HY1–HY2 features of the first model for α_{1d}-AR antagonists) allowed a better prediction of affinity values for compounds bearing a para substituent at the phenyl ring attached to the piperazine nucleus.

In the recent literature, the three-dimensional theoretical models determined by molecular dynamics calculations of the complexes between compounds structurally related to BMY 7378 and the three α_1-AR subtypes were described [19]. A qualitative comparison between their structures and the new pharmacophore model for α_{1d}-AR antagonists showed several common features. An expected, strong electrostatic interaction between the basic nitrogen atom of the ligand and an aspartate positioned in the third transmembrane domain (Asp176) {sequence numbering is referred to the α_{1d}-adrenergic receptor (572 residues) deposited at the Human Protein Reference Database [22]} has been found. This interaction was also present in the pharmacophore model (PI feature). An additional common structural feature represented by a hydrophobic pocket to accommodate the substituted arylpiperazine moiety of inhibitors is present. This cavity corresponds to both the HY1 and RA features of our new model. In this region, one of the most profitable interactions was represented by an aromatic contact between the dichlorophenyl ring of the of the arylpiperazine moiety of inhibitors with Phe364 of the receptor (the two phenyl rings were oriented in such a way to allow for a possible T tilted interaction). This interaction is perfectly mimicked by the RA feature of the model mapping the phenyl substituent at the piperazine ring of our compounds. Moreover, hydrophobic residues such as Val177 could correspond to the HY1 feature of the pharmacophore, interacting with the methoxy or chloro group at the ortho position of the phenyl ring bound to the piperazine nucleus. An additional binding pocket was described, allowing only for one hydrogen-bond contact between one of the carbonyl groups of the dione moiety of the inhibitors and Trp172 or/and Tyr392. The same interaction was also found in our new pharmacophore model, one of the carbonyl groups of the glutarimido moiety of compound analogs of **12** mapping to the hydrogen-bond acceptor. Finally, a series of amino acids surrounded the terminal heterocyclic moiety of the inhibitors. In particular, the side-chains of Glu157 and Lys385 are in contact (van der Waals interactions) with the edge of the ligands, similarly to the pharmacophore model, where the HY3 feature is mapped by the phenyl ring attached to the glutarimido moiety. Another theoretical model of the interactions between the α_{1d}-AR and one of its inhibitors (discretamine), reported by Carrieri et al. [23], is in good agreement with our new pharmacophore model. This model shows a hydrophobic (aromatic) contact between the inhibitors and Phe364, a salt bridge involving Asp176 and, finally, a weak hydrogen bond (or a polar interaction) with the receptor. These elements are reported as being the critical features for high affinity to α_{1d}-AR.

All these facts lead to the conclusion that our new five-feature pharmacophore model should be a good abstract representation of the most important structural elements that a compound should possess for high α_{1d} adrenoceptor affinity. Similarity between the HBA and PI features of the pharmacophore model with parts of the theoretical receptor could be also considered as an additional validation of our model.

In conclusion, we have reported how, using a stepwise approach, a generic pharmacophore model for α_1-AR antagonists has been evolved into a model able to rationalize the SAR of arylpiperazine derivatives for the α_{1d}-AR subtype. Using a classical computational protocol, the next step of this work will be the use of this pharmacophoric model as a three-dimensional search query to mine databases of virtual and/or commercial compounds, aiming at finding new hits with (improved) affinity toward α_{1d}-AR. As a possible extension, pharmacophore models could also be built for the remaining two α_1-AR subtypes (namely α_{1a}-AR and α_{1b}-AR) and compared with the model that we obtained for α_{1d}-AR antagonist, with the purpose of understanding the reason(s) for α_1-AR subtype selectivity.

12.3 Use of Excluded Volume Features in the Rationalization of the Activity Data of Azole Antifungal Agents

As already stated in the Introduction, a problem that sometimes arises in pharmacophore approaches is the need to take into account possible adverse steric interactions between inactive compounds in a dataset and the target protein counterpart. In these situations, the definition of ligand-forbidden zones by means of the addition of excluded volume spheres to a pharmacophore is nowadays considered a reasonable and effective improvement.

12.3.1 Excluded Volume Spheres in Structure-based and Ligand-based Pharmacophore Studies

Further to an early paper by Greenidge *et al.* [25], excluded volume features have been included, as a rule, in the course of the generation of protein-based pharmacophores, either to describe the shape of the binding cavity [26] or to eliminate compounds giving rise to steric conflicts with key residues of the target [27, 28]. Greenidge and co-workers exploited for the first time the knowledge of a receptor X-ray structure (the one of the complexes of rat thyroid hormone receptor with few ligands) to investigate the utility of augmenting a structure-based pharmacophore with hundreds of ligand-inaccessible volumes, used to define the demarcation of the active site [25]. Using this kind of refinement, the Maybridge database was efficiently searched (both in a timely manner and also considering specificity aspects). The number of false positives obtained was reduced by a factor of 2–5, with the clear effect of pruning and focusing the hit list. A different strategy was adopted – to reach the same goal – in a recent pharmacophore study on influenza virus neuraminidase (NA) inhibitors by Steindl and Langer [28]. In this case, the sole information about a few regions forbidden to the ligands has in fact been incorporated in the software Catalyst, during the refinement of structure-based pharmacophore hypotheses, with the aim of introducing enhanced steric selectivity. Nevertheless, the very few added excluded volume spheres (five), combined with the creation of a set of multiple hypotheses linked by defined connection strategies, have enabled the number of hits extracted from the Derwent World Drug Index to be sensibly reduced. Finally, in a dissimilar approach aimed at developing a "dynamic" receptor-based pharmacophore model for HIV-1 integrase inhibitors, Carlson et al. [27] manually introduced three excluded volume spheres with radii of 1.5 Å, centered on the side-chains of three essential residues of the protein. In this case, however, excluded volume spheres were not so used as an active site surface for volume constraint purposes, but rather as a tool suitable to eliminate some unreasonable modes of alignment of one or more of the compounds under investigation.

The issue of including information about ligand–receptor steric clashes has recently been taken into consideration also in the case of ligand-based pharmacophore studies. Accurate prediction of the activities of less active compounds in a dataset, with pharmacophore features in common with other active ones, in fact, has sometimes been reported to be a problem, where inactivity is probably due to steric clashes with the target. With regard to this subject, an interesting paper appeared in 2004, in which the construction of a ligand-based pharmacophore model for inhibitors of *Plasmodium falciparum* cyclin-dependent protein kinases (CDKs) and its positive validation by the discovery of different classes of novel inhibitors is described [29]. A Catalyst pharmacophore model composed by four features has been developed, without any structural analysis of the known X-ray CDK2 structure. The pharmacophore has been then used as a template to search an in-house database that has resulted in the discovery of 16 new potent inhibitors. Notably, the inhibitory activity of some of the new derivatives has been reported to be predicted "exceptionally well" by the model. In fact, and acknowledged in the paper, the activities of 80% of the test set compounds (measured as a minimum inhibitory concentration), and also those of 100% of the database-retrieved compounds, are slightly overestimated by the model, which does not contain any excluded volume.

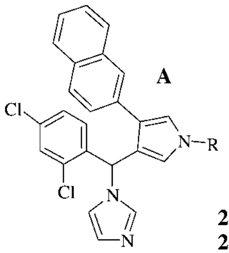
In a paper published in 2000 by Norinder [30], Catalyst was used for the first time to build a common feature pharmacophore hypothesis for HIV-1 protease inhibitors, which was then refined using in-house software (HypoOpt), after having added to it some hundreds of excluded volume spheres. These were actually derived from the X-ray structure of an inhibitor complexed to the enzyme. The aim of the approach was to obtain a computational model with some improved predictive power with respect to the corresponding hypothesis derived without receptor information.

The possibility of adding automatically excluded volume spheres to ligand-based pharmacophores, thus accounting for steric hindrance problems, has recently been tackled by Accelrys by the implementation of the HypoRefine algorithm into the software Catalyst [31]. This has been found to yield improved regression coefficients and potentially more predictive models, by the strategic placement of excluded volume spheres that can approximate steric repulsive interactions, during a simulated annealing optimization of quantitative pharmacophore models. The original quantitative HypoGen algorithm, in fact, has occasionally been reported to have difficulties in generating hypotheses that correlate well, if the steric properties of the data set make a large contribution to the activities. In such a case, inactive compounds may well have their activities overpredicted. One of the first ligand-based pharmacophore researches was published in 2004 in the *Journal of Chemical Information and Computer Science* (in the same issue that included the paper by Steindl and Langer on NA inhibitors [28]), in which the Catalyst HypoRefine module was successfully exploited to construct a pharmacophore hypothesis of some matrix metalloproteinase-1 inhibitors [32]. In that paper, Tsai and Lin state that a "top pharmacophore hypothesis" has been constructed by applying Catalyst HypoRefine (assisted by mainly a CoMSIA 3D-QSAR model), which can be used to represent the binding mode of inhibitors inside the enzyme active site. Together with the soundness of their results, the authors present some of the difficulties encountered in applying Catalyst to the construction of the model, highlighting especially (as already noted by Greenidge et al. [26]) the issue of generating reliable conformational models for all the compounds and the difficulty of properly choosing some representative features, among the several default ones provided by the program.

12.3.2 Issues Inherent in the Rational Design of Azole Antifungal Agents

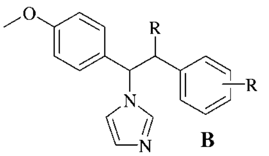
We have been involved since 1996 in computational studies of antifungal azoles [33, 34] that inhibit *Candida albicans*'s cytochrome P-450-dependent enzyme lanosterol 14α-demethylase (CA-CYP51) and our efforts have produced remarkable results, the soundness of which has been confirmed by other workers [35]. In spite of these attainments, however, our work has not been able to satisfy our expectations thoroughly for a long time. The outer predictivity of a ligand-based pharmacophore model (HYPO1) that we proposed in 2002 using the Catalyst HypoGen algorithm [34], in fact, was recently denied ($r^2_{\text{pred}}=0.19$) when – in advance of the computational studies subjected here to close examination – HYPO1 was applied to predict the activity of some newly synthesized analogs (**15 c–j**, **16 a**, **b**, **17 d–f**, **18 a**, **19 a–c**, **20 a**, **b**, **21 a** and **22 a–c** in Chart 12.1 and Table 12.2).

In fact, HYPO1's fault was not surprising if one relates its framework (HYPO1 was composed by one *ad hoc* defined aromatic-nitrogen-with-lone-pair vectorial feature to simulate the coordination interaction of azole inhibitors with the iron atom of the enzyme protoporphyrin system an three aromatic rings [34]) to the chemical structure and antifungal activity of the new compounds. The activity values of these molecules, in fact, were revealed to be strongly dependent on the presence of a hydrophobic substituent (possibly aliphatic) on



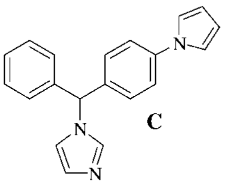
A

22a: R=H
22b: R=CH_3
22c: R=C_2H_5



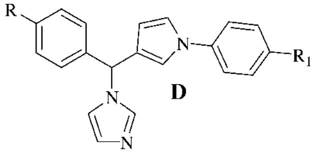
B

23a: R=2,4-Cl_2	**23g**: R=2-(1-pyrrolyl)
23b: R=4-NH_2	**23h**: R=4-NO_2
23c: R=3-(1-pyrrolyl)	**23i**: R=3-Cl
23d: R=2-Cl	**23j**: R=2-F
23e: R=3-F	**23k**: R=4-F
23f: R=2-NO_2	



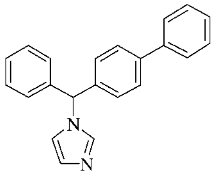
C

24a

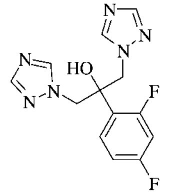


D

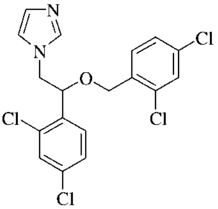
25a: R=H,	R_1=H	**25g**: R=Cl	R_1=F
25b: R=CH_3	R_1=H	**25h**: R=H	R_1=F
25c: R=Cl	R_1=Cl	**25i**: R=F	R_1=H
25d: R=F	R_1=Cl	**25j**: R=CH_3	R_1=Cl
25e: R=H	R_1=Cl	**25k**: R=F	R_1=F
25f: R=CH_3	R_1=F		



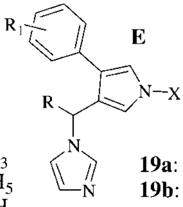
Bifonazole



Fluconazole



Miconazole



E

15a: R=Ph	R_1=4-Cl	X=H
15b: R=Ph	R_1=4-Cl	X=CH_3
15c: R=Ph	R_1=4-Cl	X=C_2H_5
15d: R=Ph	R_1=4-Cl	X=C_3H_7
15e: R=Ph	R_1=4-Cl	X=CH_2-c-C_3H_5
15f: R=Ph	R_1=4-Cl	X=CH=CH_2
15g: R=Ph	R_1=4-Cl	X=CH_2-CH=CH_2
15h: R=Ph	R_1=4-Cl	X=CH_2-CH=C$(CH_3)_2$
15i: R=Ph	R_1=4Cl	X=CH=CH-$COOCH_3$
15j: R=Ph	R_1=4-Cl	X=Ph
16a: R=Ph	R_1=2,4-Cl	X=CH_3
16b: R=Ph	R_1=2,4-Cl	X=C_2H_5
17a: R=2,4-Cl_2-Ph	R_1=4-Cl	X=H
17b: R=2,4-Cl_2-Ph	R_1=4-Cl	X=CH_3
17c: R=2,4-Cl_2-Ph	R_1=4-Cl	X=C_2H_5
17d: R=2,4-Cl_2-Ph	R_1=4-Cl	X=C_3H_7
17e: R=2,4-Cl_2-Ph	R_1=4-Cl	X=CH_2-CH=CH_2
17f: R=2,4-Cl_2-Ph	R_1=4-Cl	X=CH_2-CH$(OCH_3)_2$
18a: R=2,4-Cl_2-Ph	R_1=2-Cl	X=CH_3

19a: R=2,4-Cl_2-Ph	R_1=2,4-Cl_2	X=CH_3
19b: R=2,4-Cl_2-Ph	R_1=2,4-Cl_2	X=C_2H_5
19c: R=2,4-Cl_2-Ph	R_1=2,4-Cl_2	X=CH_2-CH=CH_2
20a: R=4-CH_3-Ph	R_1=3,4-Cl_2	X=CH_3
20b: R=4-CH_3-Ph	R_1=3,4-Cl_2	X=C_2H_5
21a: R=2,4-Cl_2-Ph	R_1=4-(1-pyrrolyl)	X=CH_3
26a: R=1-naphthyl	R_1=H	X=H
26b: R=2-naphthyl	R_1=H	X=H
26c: R=2-naphthyl	R_1=H	X=CH_3
26d: R=1-naphthyl	R_1=H	X=CH_3
26e: R=2,4-Cl_2-Ph	R_1=4-Ph	X=H
26f: R=2,4-Cl_2-Ph	R_1=4-CF_3	X=H
26g: R=2,4-Cl_2-Ph	R_1=4-CN	X=H
26h: R=2,4-Cl_2-Ph	R_1=4-NO_2	X=H
26i: R=2,4-Cl_2-Ph	R_1=4-NH_2	X=H
26j: R=2,4-Cl_2-Ph	R_1=4-(1-pyrrolyl)	X=H
26k: R=2,4-Cl_2-Ph	R_1=4-OH	X=H
26l: R=2,4-Cl_2-Ph	R_1=4-SCH_3	X=H

Chart 12.1 Structures of the azoles discussed in the text.

Table 12.2 Experimental and estimated anti-*Candida* activity values for all the azoles discussed in the text (μmol mL^{-1}).

Compound	$MIC_{cpd}/MIC_{bifonazole}$			Error	
	Experimental	Estimated (MOD1)	Estimated (MOD3)	MOD1	MOD3
15a [a)]	0.66	1.1	0.86	1.7	1.3
15b [a)]	0.025	0.053	0.13	2.1	5.1
15c	0.11	0.018	0.016	–5.8	–6.7
15d [a)]	0.023	0.014	0.0064	–1.7	–3.6
15e [a)]	0.025	0.087	0.052	3.5	2.1
15f	0.031	0.041	0.26	1.3	8.3
15g [a)]	0.019	0.020	0.0076	1.0	–2.5
15h [a)]	0.043	0.031	0.063	–1.4	1.5
15i [a)]	0.34	0.11	0.26	–3.1	–1.3
15j [a)]	1.5	0.56	1.0	–2.7	–1.5
16a	0.11	0.046	0.13	–2.3	1.2
16b	0.33	0.014	0.33	–24	1.0
17a	0.92	1.0	0.58	1.1	–1.6
17b	0.28	0.034	0.60	–8.3	2.1
17c	0.31	0.016	0.063	–19	–4.9
17d	1.70	0.014	0.13	–120	–13
17e	0.15	0.011	0.045	–14	–3.3
17f [a)]	0.48	0.80	0.54	1.7	1.1
18a [a)]	0.36	0.80	0.28	2.2	–1.3
19a	2.9	0.039	0.28	–73	–10
19b	0.23	0.0090	0.33	–25	1.5
19c	0.12	0.013	0.23	–9.3	1.9
20a [a)]	0.19	0.054	0.13	–3.6	–1.5
20b	0.47	0.0098	0.19	–48	–2.5
21a	0.62	0.74	3.3	1.2	5.4
22a [a)]	0.21	1.4	0.74	6.5	3.5
22b	0.94	0.79	0.46	–1.2	–2.0
22c	1.1	0.65	0.39	–1.7	–2.8
23a [a)]	1.0	1.3	0.62	1.3	–1.6
23b	1.4	9.8	26	6.9	19
23c	45	1.2	4.3	–38	–10
23d	63	2.9	23	–22	–2.8
23e	51	6.3	6.6	–8.1	–7.7
23f [a)]	97	4.1	23	–24	–4.1
23g	7.8	1.0	11	–7.6	1.4
23h	17	8.4	30	–2.0	1.8
23i	44	1.9	32	–23	–1.4
23j [a)]	49	7.7	22	–6.3	–2.2
23k	23	8.0	22	–2.8	–1.0
24a [a)]	6.8	5.8	3.9	–1.2	–1.7
25a	2.3	1.4	21	–1.7	9.2
25b [a)]	1.2	1.4	1.0	1.2	–1.2
25c	7.3	1.1	0.62	–6.4	–12

Table 12.2 (continued)

Compound	$MIC_{cpd}/MIC_{bifonazole}$			Error	
	Experimental	Estimated (MOD1)	Estimated (MOD3)	MOD1	MOD3
25d	2.8	1.2	0.63	−2.4	−4.4
25e [a)]	0.70	1.1	0.61	1.6	−1.1
25f	3.1	1.5	0.61	−2.1	−5.1
25g	2.7	1.7	0.83	−1.6	−3.3
25i	1.8	1.4	22	−1.3	12
25h [a)]	4.1	1.5	3.0	−2.7	−1.4
25j	1.1	1.2	0.68	1.1	−1.7
25k	1.4	1.6	3.7	1.2	2.7
26a	1.7	1.6	6.6	−1.1	3.8
26b	3.8	1.8	2.5	−2.2	−1.6
26c	63	1.1	15	−58	−4.0
26d	6.0	1.1	11	−5.5	1.7
26e	4.3	2.2	3.0	−1.9	−1.4
26f	2.2	0.86	0.52	−2.6	−4.3
26g	4.3	3.7	13	−1.2	3.0
26h	4.5	3.4	19	−1.3	4.2
26i [a)]	28	6.5	17	−4.4	−1.7
26j	1.3	0.82	0.18	−1.6	−7.1
26k [a)]	26	6.1	21	−4.2	−1.2
26l	2.8	0.74	0.92	−3.8	−3.1
Bifonazole [a)]	1.0	4.7	4.0	−4.7	4.0
Fluconazole [a)]	0.069	1.5	0.59	21	8.6
Miconazole [a)]	0.14	0.60	0.27	4.4	2.0

a) Training set compounds.

the nitrogen of the pyrrole ring, while HYPO1 (characterized by three "ring aromatic" features plus one vector "coordination bond" feature) could not recognize this obviousness, as it could not match such a substituent by any feature.

This being the case, a new pharmacophore modeling project was started [36] using newly and previously synthesized azoles belonging to five different structural classes (see Chart 12.1), using Catalyst, with the purpose of bringing our previous model up to date. During this new study, the most tricky points of our previous computational protocol have been carefully revised, attempting to hit at last the "predictivity" target, essential for the future rational design of possibly active compounds. First, the definition of the coordination bond feature (UNA: Unsubstituted Nitrogen Aromatic) has been updated. Differently than in HYPO1, UNA has been defined as a sphere; X-ray data have demonstrated, in fact, that the binding of azoles to CYP51 enzymes diverges to some extent from any fixed distance and angle [37–39]. Notably, similar arguments have recently been taken into account to develop a modified version of the docking program DOCK suitable for CYP enzymes [40]. Second, Catalyst's HypoRefine module has been included in the new computational protocol, allowing hypotheses with excluded volume spheres to be generated and thus accounting for steric hindrance problems. Finally, minimum inhibitory concentration mean values (MIC), instead of the previously used MIC_{90} values [33, 34], were preferred in this study to express the anti-*Candida* activities of the compounds studied (Table 12.2). The MIC values of the whole dataset (63 derivatives plus fluconazole, miconazole and bifonazole as the reference compounds), which cover four orders of magnitude, have been normalized and formulated as MIC compound/MIC bifonazole.

The conformational models of alternative stereoisomers of all the compounds (see above for details on the conformational search within Catalyst), to be used for pharmacophore generation, were automatically generated by means of Catalyst. The following chemical features were taken into account to build the pharmacophoric hypotheses: UNA, hydrogen-bond acceptors (HBA), hydrogen-bond donors (HBD), aromatic rings (RA) and hydrophobic groups (HY). Owing to both the molecules' flexibility and functional complexity, the hypothesis generator was constrained to report hypotheses with at least four features and to include UNA in each pharmacophore, to satisfy the key interaction between azole inhibitors and the enzyme.

In a first attempt, two pharmacophore models were generated from two different training sets. The first hypothesis (MOD1, shown in Figure 12.5) was derived from a general training set (see Table 12.2), conceived to maximize the information content of the whole set of studied compounds, which included molecules whose activities represented all the orders of magnitude covered by each distinct structural group. Further to this approach, the coordination interaction plus one aromatic ring and two hydrophobic features were recognized by Catalyst to have pharmacophoric relevance. The regression line based on MOD1 of experimental versus estimated/predicted MIC exhibited a correlation coefficient $r^2 = 0.84$ for the training set (r.m.s.d. = 1.23).

Comparison between estimated and experimentally measured MIC values of the compounds (see Table 12.2) showed, in the worst case, a 24-fold difference and in most cases was less than a 2-fold difference, indicating a reliable ability of MOD1 to estimate affinities within the training set. The execution of the

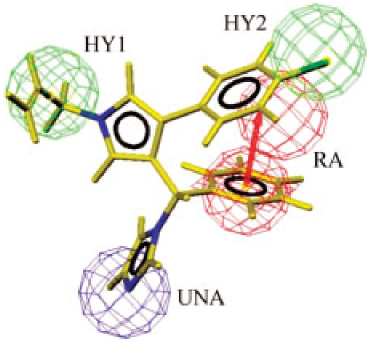


Fig. 12.5 Compound **15g** (in yellow), the most active of the whole set, mapped on to MOD1. Pharmacophore features are color coded: blue for the unsubstituted aromatic nitrogen (UNA), red for aromatic ring (RA), green for hydrophobic (HY1 and HY2).

same analysis on the test set revealed, however, a less convincing scenario as a fairly high correlation coefficient ($r^2=0.73$) was not coupled with homogeneous distribution of the corresponding marks around the regression line (result not shown). The MIC values of several compounds were predicted with very high accuracy, whereas the inhibitory potency of some derivatives was much overestimated (up to two orders of magnitude) and only one compound was underestimated (**23b**).

The prediction of the test set by MOD1 was regarded as unsatisfactory and presumed to be the after-effect of a possible lack of quantitative correlation between MIC activity data of some inhibitors and their binding interaction into the enzyme active site (MIC has been found to depend also on permeability and metabolic factors) [41]. With the aim of verifying the validity of such a guess, a reduced subset of 35 compounds out of 66 was selected for a second pharmacophore generation experiment. The majority of the molecules from the most sampled structural class (E in Chart 12.1) – covering four MIC orders of magnitude – were in fact steadily overestimated by MOD1 and consequently only the six most active molecules from this class (one order of magnitude) were considered in this generation. A second set of pharmacophore hypotheses was thus built, considering a training set and a test set of 20 and 15 compounds, respectively, from which the pharmacophore MOD2 was selected. It showed a relevant increase in the correlation for both the training set ($r^2=0.94$, r.m.s.d.$=0.84$) and test set ($r^2=0.90$); however, as was to be expected owing to the rule whereby the training set had been selected, the correlation coefficient decreased dramatically when all the compounds were included into an exhaustive (46 compounds) test set ($r^2=0.52$). MOD2 is displayed in Figure 12.6 superposed on **15g**.

Here, the coordination interaction plus three aromatic rings (RA1, RA2 and RA3) and two excluded volume spheres (EV1 and EV2) were found by Catalyst

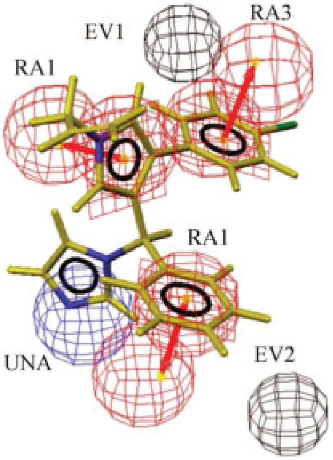


Fig. 12.6 Compound **15g** (in yellow) mapped on to MOD2. Pharmacophore features are color coded: blue for the unsubstituted aromatic nitrogen (UNA), red for aromatic ring (RA1, RA2 and RA3). In black are shown the excluded volume spheres (EV1 and EV2).

to enhance the pharmacophore relevance. The inadequate prediction of the exhaustive test set by MOD2, however, clearly demonstrated its overall unreliability.

A critical comparison was then performed between this hypothesis and MOD1 to rationalize the not entirely satisfactory results obtained up to then. Two structural aspects of MOD2 appeared in fact to be really interesting and deserved further consideration. The first unfavorable peculiarity was the presence of aromatic ring features only (RA1, RA2 and RA3) to express the attractive interactions of azoles with the amino acids of the active site of CYP51. This result was clearly a computational oversight largely induced by the exclusion from the training set of many less active compounds from the most sampled structural class. As a consequence of that choice, in fact, Catalyst was not able to recognize the relevance of the presence/absence of a pyrrole substituent (HY1 in MOD1), peculiar to the compounds of that class. The second remarkable (favorably) indication provided by MOD2 was the presence of two excluded volume spheres (EV1 and EV2). Excluded volume spheres had in fact been chased since the beginning of this study and their appearance supported a guess that the average overprediction of the 42 compound test set by MOD1 was mainly due to the lack of excluded volume spheres in that pharmacophore. Actually, neither MOD1 nor MOD2 displayed polar features besides UNA; moreover, the possibility that steric interactions might play a relevant role in the binding of azoles into the active site of CYP51 had already been assessed both in our previous research [34], based on the lipophilicity of our compounds, and, more generally, by other workers [42]. The analysis just described clarified at a molecular level the reasons for the low reliability of MOD2 with respect to MOD1 and suggested a possible adjustment to be made to correct our computational protocol. Catalyst's default SPACING parameter – that is, the minimum distance between actual features locations – was hypothesized as the critical control parameter that did not work properly for our set of compounds. A further and final three-dimensional pharmacophore model (MOD3) was consequently generated, from the first general training set, after having decreased the spacing value to 1.0 Å. MOD3 showed an interesting increase in the correlation for both the training set (24 compounds, $r^2=0.93$, r.m.s.d.$=0.80$) and the test set (42 compounds, $r^2=0.73$) with respect to MOD1. Comparison between estimated and experimental MIC values gave, in the worst case (fluconazole), an 8.5-fold difference and in most cases was less than a 2-fold difference (see Table 12.2). Interestingly, while the regression line based on MOD3 (shown in Figure 12.7) exhibited a correlation coefficient for the test set equal to that given by MOD1, the distribution around the regression line was definitely more homogeneous than that arising from MOD1, indicating a better capacity of MOD3 to predict the activities of the test set.

The MIC values of all the test set compounds were in fact predicted within the measured order of magnitude, with the exception of three compounds whose antifungal activity was underestimated by factors of 13, 19 and 12.

The coordination interaction, one aromatic ring, two hydrophobic features and two excluded volume spheres make up the final model (shown in Figure

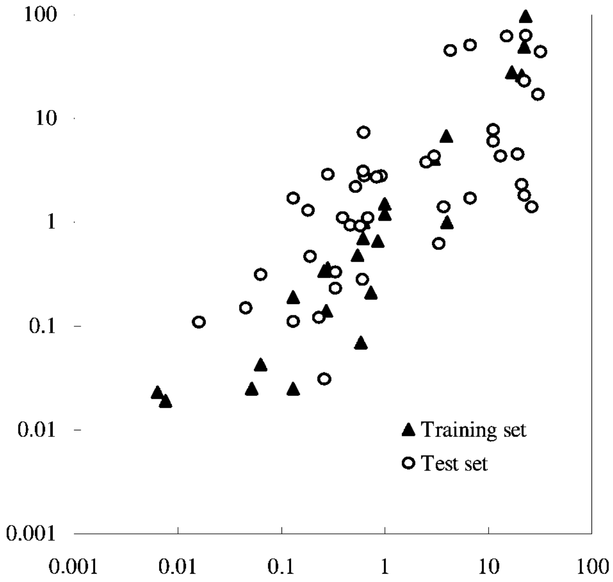


Fig. 12.7 Regression line for MOD3. Experimental versus estimated (or predicted) anti-*Candida* activity is reported for each member of the training set and of the test set.

12.8). The pharmacophore seems to be more definite, accurate, flexible and realistic with respect to the Catalyst model that we proposed in our previous studies [34], as aromatic π–π stacking interactions appear no longer to be the sole interactions able to modulate the activities of different antifungal agents. Some key interactions, and also excluded volume spheres, further to the coordination bond of azole antifungals with the demethylase enzyme, are now highlighted. It is questionable whether the excluded volume spheres properly represent the surrounding atoms in the binding pocket of CA-CYP51 or, more simply, regions randomly selected by the HypoRefine algorithm in the aligned inactive molecules far away from the active that could not contain any topology. Nevertheless, volume spheres have helped to improve the predictivity of MOD3 as they specify spherical spaces in the proximity of the pharmacophore that could not contain any atoms or bonds and this is a constraint preventing an advantageous matching of conformations of the less active compounds on to the pharmacophore.

12.4 Conclusion

Both examples reported above suggest that the pharmacophoric approach provided by Catalyst could represent a useful and efficient tool available to modelers working in the field of medicinal chemistry. However, it is necessary to em-

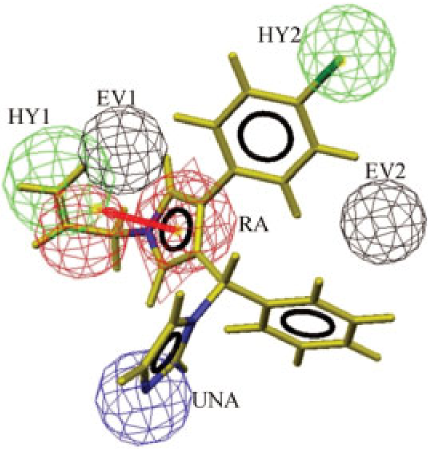


Fig. 12.8 Compound **15g** (in yellow) mapped on to MOD3 (the final pharmacophore). Pharmacophore features are color coded: blue for the unsubstituted aromatic nitrogen (UNA), red for aromatic ring (RA), green for hydrophobic (HY1 and HY2). In black are shown the excluded volume spheres (EV1 and EV2).

phasize that two prerequisites are required of the user to guarantee both full control during the steps of model generation and critical evaluation of the pharmacophoric hypothesis generated: (i) knowledge, as high as possible, of the software routines and of the parameters that they involve and (ii) chemical good sense that the user should have and apply to the analysis of results. In detail, while a good knowledge of the parameters settable within pop-up menus of the software graphical interface is usually enough to run appropriate calculations, sometimes higher control of the software is needed in order to adjust (i.e. enable, disable or set) several "hidden" variables (referred to as the Catalyst parameters) to try to solve a specific problem. As an example, if the *confAnalysis.AxialEquatorialRatio* parameters were available at the time the first case study was approached, the conformational search routine of the program would have been sufficiently efficient to treat simple compounds such as arylpiperazinyl derivatives. On the other hand, the expertise of the user in the field of chemistry and medicinal chemistry is an essential condition to analyze output data. Two examples follow to support this conclusion. The first concerns the directionality properties of hydrogen bonds. In our experience, we sometimes found that the hydrogen bond acceptor vector depicted by Catalyst was a prolongation of a C=O group, in disagreement with the angular constraint required for hydrogen-bond contacts. An additional example is represented by distorted rings (having a very low probability of existing on the basis of a Boltzmann analysis) resulting from the fitting of a compound into a pharmacophore hypothesis.

Taking into account these considerations (and others that lie outside the scope of this chapter), it is undeniable that at the moment several types of software exist (including Catalyst), based on a ligand-based drug design approach and aimed at rationalizing in an efficient and fast way biological data, proposing useful suggestions to improve the contacts between ligands and the corresponding receptor counterpart and allowing one to identify – by database searches – new hit compounds to be structurally optimized.

References

1 (a) Chang, D. J., Chang, T. K., Yamanishi, S. S., Salazar, F. H. R., Kosaka, A. H., Khare, R., Bhakta, S., Jasper, J. R., Shieh, I.-S., Lesnick, J. D., Ford, A. P. D. W., Daniels, D. V., Eglen, R. M., Clarke, D. E., Bach, C., Chan, H. W. Molecular cloning, genomic characterization and expression of novel human α_{1A}-adrenoceptor isoforms. *FEBS Lett.* **1998**, *422*, 279–283; (b) Daniels, D. V., Gevel, J. R., Jasper, J. R., Kava, M. S., Lesnick, J. D., Meloy, T. D., Stepan, G., Williams, T. J., Clarke, D. E., Chang, D. J., Ford, A. P. D. W. Human cloned α_{1A}-adrenoceptor isoforms display α_{1L}-adrenoceptor pharmacology in functional studies. *Eur. J. Pharmacol.* **1999**, *370*, 337–343.

2 (a) de Reijke, T. M., Klarskov, P. Comparative efficacy of two α-adrenoreceptor antagonists, doxazosin and alfuzosin, in patients with lower urinary tract symptoms from benign prostatic enlargement. *BJU Int.* **2004**, *6*, 757–762; (b) Michelotti, G. A., Schwinn, D. A. Mechanistic insights into the role of α_1-adrenergic receptors in lower urinary tract symptoms. *Curr. Urol. Rep.* **2004**, *4*, 258–266; (c) Lowe, F. C. Role of the newer α_1-adrenergic receptor antagonists in the treatment of benign prostatic hyperplasia-related lower urinary tract symptoms. *Clin. Ther.* **2004**, *11*, 1701–1713.

3 Barbaro, R., Betti, L., Botta, M., Corelli, F., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G., Corsano, S. Synthesis, biological evaluation and pharmacophore generation of new pyridazinone derivatives with affinity toward α_1- and α_2-adrenoceptors. *J. Med. Chem.* **2001**, *44*, 2118–2132.

4 *Catalyst (Version 4.9)*, Accelrys, San Diego, CA.

5 Accelrys, http://www.accelrys.com/support/life/catalyst/hypogen.html.

6 Cheng, Y. C., Prusoff, W. H. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (IC_{50}) of an enzymatic reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

7 Mohamadi, F., Richards, N. G. J., Guida, W. C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., Still, W. C. Macromodel – an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440–467.

8 Langgard, M., Bjornholm, B., Gundertofte, K. Pharmacophore modeling by automated methods: possibilities and limitations. In *Pharmacophore Perception, Development and Use in Drug Design*, Guner, O. F. (ed.), IUL Biotechnology Series, International University Line, La Jolla, CA, 2000, pp. 239–251.

9 (a) De Marinis, R. M., Wise, M., Hieble, J. P., Ruffolo R. R., Jr. Structure–activity relationships for alpha-1 adrenergic receptor agonists and antagonists. In *The Alpha-1 Adrenergic Receptor*, Ruffolo, R. R., Jr (ed.), Humana Press, Clifton, NJ, 1987, pp. 211–265; (b) Montesano, F., Barlocco, D., Dal Piaz, V., Leonardi, A., Poggesi, E., Fanelli, F., De Benedetti, P. G. Isoxazolo-[3,4-*d*]-pyridazin-7-(6*H*)-ones and their corresponding 4,5-disubstituted-3-(2*H*)-pyridazinone analogues as new substrates for α_1-adrenoceptor selective antagonists: synthesis, modeling and binding studies. *Bioorg. Med. Chem.* **1998**, *6*, 925–935.

10 Betti, L., Floridi, M., Giannaccini, G., Manetti, F., Strappaghetti, G., Tafi, A., Botta, M. α_1-Adrenoceptor antagonists. 5. Pyridazinone-arylpiperazines. Probing the influence on affinity and selectivity of both *ortho*-alkoxy groups at the arylpiperazine moiety and cyclic substituents at the pyridazinone Nucleus. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 171–173.

11 Betti, L., Corelli, F., Floridi, M., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G., Botta, M. α_1-Adrenoceptor antagonists. 6. Structural optimization of pyridazinone-arylpiperazines. Study of the influence on affinity and selectivity of cyclic substituents at the pyridazinone ring and alkoxy groups at the arylpiperazine moiety. *J. Med. Chem.* **2003**, *46*, 3555–3558.

12 Setoguchi, M., Sakamori, M., Takehara, S., Fukuda, T. Effects of iminodibenzyl antipsychotic drugs on cerebral dopamine and alpha-adrenergic receptors. *Eur. J. Pharmacol.* **1985**, *112*, 313–322.

13 (a) Roquebert, J., Grenie, B. Alpha 2-adrenergic agonist and alpha 1-adrenergic antagonist activity of ergotamine and dihydroergotamine in rats. *Arch. Int. Pharmacodyn. Ther.* **1986**, *284*, 30–37; (b) Bonuso, S., Di Stasio, E., Marano, E., Covelli, V., Testa, N., Tetto, A., Buscaino G.A. The antimigraine effect of ergotamine: a role for alpha-adrenergic blockade? *Acta Neurol. (Napoli)* **1994**, *16*, 1–10.

14 Giannangeli, M., Cazzolla, N., Luparini, M.R., Magnani, M., Mabilia, M., Picconi, G., Tomaselli, M., Baiocchi, L. Effect of modification of the alkylpiperazine moiety of trazodone on 5-HT$_{2A}$ and α_1 receptor binding activity. *J. Med. Chem.* **1999**, *42*, 336–345.

15 Costa, A., Martignoni, E., Blandini, F., Petraglia, F., Genazzani, A.R., Nappi, G. Effects of etoperidone on sympathetic and pituitary–adrenal responses to diverse stressors in humans. *Clin. Neuropharmacol.* **1993**, *16*, 127–138.

16 Betti, L., Botta, M., Corelli, F., Floridi, M., Fossa, P., Giannaccini, G., Manetti, F., Strappaghetti, G., Corsano, S. α_1-Adrenoceptor antagonists. Rational design, synthesis and biological evaluation of new trazodone-like compounds. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 437–440.

17 Betti, L., Botta, M., Corelli, F., Floridi, M., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G., Tafi, A., Corsano, S. α_1-Adrenoceptor antagonists. 4. Pharmacophore-based design, synthesis and biological evaluation of new imidazo-, benzimidazo-, and indoloarylpiperazine derivatives. *J. Med. Chem.* **2002**, *45*, 3603–3611.

18 Romeo, G., Materia, L., Manetti, F., Cagnotto, A., Mennini, T., Nicoletti, F., Botta, M., Russo, F., Minneman, K.P. New pyrimido[5,4-*b*]indoles as ligands for α_1-adrenoceptor subtypes. *J. Med. Chem.* **2003**, *46*, 2877–2894.

19 Leonardi, A., Barlocco, D., Montesano, F., Cignarella, G., Motta, G., Testa, R., Poggesi, E., Seeber, M., De Benedetti, P.G., Fanelli, F. Synthesis, screening and molecular modeling of new potent and selective antagonists at the α_{1d} adrenergic receptor. *J. Med. Chem.* **2004**, *47*, 1900–1918.

20 (a) Koer, F.J., Altona, C. Conformation of non-aromatic ring compounds – 84. Dipole moments and 100 MHz Fourier-transform proton magnetic resonance spectra of some substituted 3-phenylglutaric anhydrides, a conformational analysis. *Recl. J. R. Neth. Chem. Soc.* **1974**, *93/5*, 147–151; (b) Koer, F.J., Faber, D.H., Altona, C. Conformation of non-aromatic ring compounds – 89. The conformation of 3-(2,6-dichlorophenyl)glutaric anhydride studied by dynamic proton magnetic resonance and empirical force field calculations. *Recl. J. R. Neth. Chem. Soc.* **1974**, *93/12*, 307–311; (c) Koer, F.J., Altona, C. Conformation of non-aromatic ring compounds – 92. Carbon-13 nuclear magnetic resonance spectra of some 3-arylglutaric anhydrides. *Recl. J. R. Neth. Chem. Soc.* **1975**, *94/6*, 127–131.

21 Lopez, F.J., Arias, L., Chan, R., Clarke, D.E., Elworthy, T.R., Ford, A.P.D.W., Guzman, A., Jaime-Figueroa, S., Jasper, J.R., Morgans, D.J., Jr., Padilla, F., Perez-Medrano, A., Quintero, C., Romero, M., Sandoval, L., Smith, S.A., Williams, T.J., Blue, D.R. Synthesis, pharmacology and pharmakokinetics of 3-(4-aryl-piperazin-1-ylalkyl)-uracils as uroselective α_{1A}-antagonists, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1873–1878.

22 Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far,

R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., Pandey, A. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Res.* **2003**, *10*, 2363–2371; http://www.hprd.org/protein/00079.

23 Carrieri, A., Centeno, N.B., Rodrigo, J., Sanz, F., Carotti, A. Theoretical evidence of a salt bridge disruption as the initiating process for the α_{1d}-adrenergic receptor activation: a molecular dynamics and docking study. *Proteins* **2001**, *43*, 382–394.

24 Romeo, G. Personal communication.

25 Greenidge, P.A., Carlsson, B., Bladh, L.-G., Gillner, M. Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor. *J. Med. Chem.* **1998**, *41*, 2503–2512.

26 Greenidge, P.A., Merette, S.A.M., Beck, R., Dodson, G., Goodwin, C.A., Scully, M.F., Spencer, J., Weiser, J., Deadman, J.J. Generation of ligand conformations in continuum solvent consistent with protein active site topology: application to thrombin. *J. Med. Chem.* **2003**, *46*, 1293–1305.

27 Carlson, H.A., Masukawa, K.M., Rubins, K., Bushman, F.D., Jorgensen, W.J., Lins, R.D., Briggs, J.M., McCammon, J.A. Developing a pharmacophore model for HIV-1 integrase. *J. Med. Chem.* **2000**, *43*, 2100–2114.

28 Steindl, T., Langer, T. Influenza virus neuramidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1849–1856.

29 Bhattacharjee, A.K., Geyer, J.A., Woodard, C.L., Kathcart, A.K., Nichols, D.A., Prigge, S.T., Li, Z., Mott, B.T., Waters, N.C. A three-dimensional *in silico* pharmacophore model for inhibition of *Plasmodium falciparum* cyclic-dependent kinases and discovery of different classes of novel Pfmrk species inhibitors. *J. Med. Chem.* **2004**, *47*, 5418–5426.

30 Norinder, U. Refinement of Catalyst hypotheses using simplex algorithm. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 545–557.

31 Accelrys, http://www.accelrys.com/catalyst/hyporefine.html.

32 Tsai, K.-C., Lin, T.-H. A ligand-based molecular modeling study on some matrix metalloproteinase-1 inhibitors using several 3D QSAR techniques. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1857–1871.

33 Tafi, A., Anastassopoulou, J., Theophanides, T., Botta, M., Corelli, F., Massa, S., Artico, M., Costi, R., Di Santo, R., Ragno, R. Molecular modeling of azole antifungal agents active against *Candida albicans*. 1. A comparative molecular field analysis study. *J. Med. Chem.* **1996**, *39*, 1227–1235.

34 Tafi, A., Costi, R., Botta, M., Di Santo, R., Corelli, F., Massa, S., Ciacci, A., Manetti, F., Artico, M. Antifungal Agents 10. New derivatives of 1-[(aryl)[4-aryl-1*H*-pyrrol-3-yl]methyl]-1*H*-imidazole, synthesis, anti-*Candida* activity, and quantitative structure–activity relationship studies. *J. Med. Chem.* **2002**, *45*, 2720–2732.

35 Liu, J., Pan, D., Tseng, Y., Hopfinger, A.J. 4D-QSAR analysis of a series of antifungal P450 inhibitors and 3D-pharmacophore comparisons as a function of alignment. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2170–2179.

36 Di Santo, R., Tafi, A., Costi, R., Botta, M., Artico, M., Corelli, F., Forte, M., Caporuscio, F., Angiolella, L., Palamara, A.T. Antifungal agents. 11. *N*-substituted derivatives of 1-[(aryl)[4-aryl–1*H*-pyrrol–3-yl]methyl]–1*H*-imidazole: synthesis, anti-*Candida* activity, and QSAR studies. *J. Med. Chem.* **2005**, *48*, 5140–5153.

37 Poulos, T.L., Howard, A.J. Crystal structures of methyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450$_{\mathrm{cam}}$. *Biochemistry* **1987**, *26*, 8165–8174.

38 Yano, J.K., Koo, L.S., Shuller, D.J., Li, H., Ortiz De Montellano, P. R., Poulos, T.L. Crystal structure of a thermophilic cytochrome P450 from the archaeon *Sul-*

folobus solfataricus. *J. Biol. Chem.* **2000**, *275*, 31086–31092.

39 Podust, L. M., Poulos, T. L., Waterman, M. R. Crystal structure of cytochrome P450 14*α*-sterol demethylase (CYP51) from *Mycobacterium tubercolosis* in complex with azole inhibitors. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 3068–3073.

40 Verras, A., Kuntz, I. D., Ortiz De Montellano, P. R. Computer-assisted design of selective imidazole inhibitors for cytochrome P450 enzymes. *J. Med. Chem.* **2004**, *47*, 3572–3579.

41 Upadhayaya, R. S., Sinha, N., Jain, S., Kishore, N., Chandra, R., Arora, S. K. Optically active antifungal azoles: synthesis and antifungal activity of (2*R*,3*S*)-2-(2,4-difluorophenyl)-3-(5-[2-[4-arylpiperazin-1-yl]ethyl]tetrazol-2-yl/1-yl)-1-[1,2,4]-triazol-1-yl-butan-2-ol. *Bioorg. Med. Chem.* **2004**, *12*, 2225–2238.

42 (a) Sharma, P., Rane, N., Gurram, V. K. Synthesis and QSAR studies of pyrimido[4,5-*d*]pyrimidine-2,5-dione derivatives as potential antimicrobial agents. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4185–4190; (b) Gollapudy, R., Ajmani, S., Kulkarni, A. Modeling and interaction of *Aspergillus fumigatus* lanosterol 14-*α* demethylase ‘A’ with azole antifungals. *Bioorg. Med. Chem.* **2004**, *12*, 2937–2950; (c) Ji, H., Zhang, W., Zhang, M., Kudo, M., Aoyama, Y., Yoshida, Y., Sheng, C., Song, Y., Yang, S., Zhou, Y., Lü, J., Zhu, J., Structure-based *de novo* design, synthesis and biological evaluation of non-azole inhibitors specific for lanosterol 14*α*-demethylase of fungi. *J. Med. Chem.* **2003**, *46*, 474–485; (d) Wiktorowicz, W., Markuszewski, M., Krysinski, J., Kaliszan, R. Quantitative structure–activity relationship study of a series of imidazole derivatives as potential new antifungal drugs. *Acta Pol. Pharm.* **2002**, *59*, 295–306.

13
GPCR *Anti-target* Modeling: Pharmacophore Models to Avoid GPCR-mediated Side-effects

Thomas Klabunde

13.1 Introduction: GPCRs as Anti-targets

In recent years, the term *anti-target* has been adopted within the drug discovery community. Several enzymes, receptors or channels were identified as the molecular basis for several severe side-effects observed for development candidates (or even for marketed drugs) and were therefore termed *anti-targets*. *Anti-target*-mediated side-effects can put the further development of promising clinical candidates at risk, hence several pharmaceutical companies have not only started to implement appropriate *in vitro* assays in the early phase of the drug discovery chain, but in addition, structural information on these *anti-targets*, their ligands

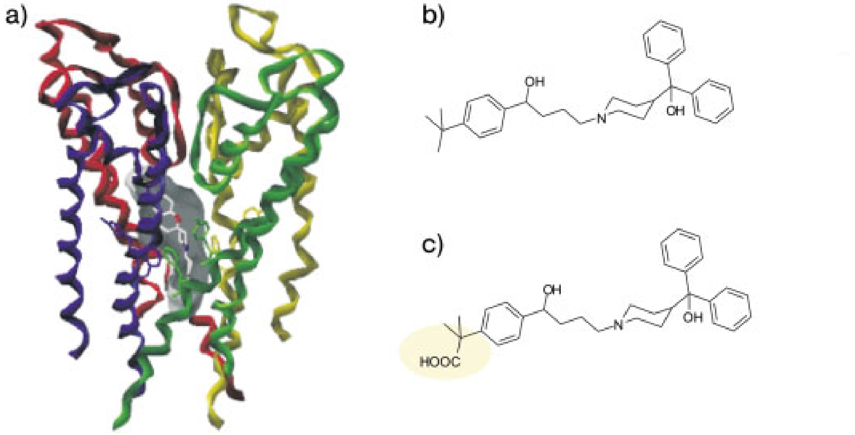


Fig. 13.1 The potassium hERG channel as *anti-target* within drug discovery. (a) Homology model of hERG channel with compound MK-499 bound [2]. (b) Histamine H1 antagonist terfenadine, which was found to be a strong blocker of the hERG channel inducing the long QT syndrome (hERG IC_{50}=56 nM). (c) Fexofenadine, a close analog of terfenadine, reveals no hERG channel affinity and is successfully marketed for seasonal rhinitis (trade name Allegra).

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

Table 13.1 Biogenic amine binding GPCR *anti-targets* and side-effects mediated by high-affinity antagonists of the respective receptor

Receptor	Expected side-effects
Adrenergic α_{1a}	Orthostatic hypotension, dizziness and fainting spells
Dopaminergic D2	Extrapyramidal syndrome (EPS), tardive dyskinesia
Serotonin 5-HT_{2C}	Weight gain, obesity
Muscarinic M1	Attention deficits, hallucinations, memory deficits

and structure-activity relationships is compiled and *in silico* tools are developed for *anti-target* modeling. These computational tools can guide the chemical optimization of novel lead series towards clinical candidates lacking *anti-target*-mediated side-effects. The probably best known example of an *anti-target* is the K^+ channel encoded by the human ether-a-go-go related gene (hERG) (Fig. 13.1 a) [1].

The hERG K^+ channel plays a crucial role in normal action potential repolarization in the heart. Within recent years, several non-cardiac drugs have been found to inhibit the hERG K^+ channel, resulting in a drug-induced long QT syndrome and sudden cardiac death. Also terfenadine (Fig. 13.1 b), a drug released for the treatment of seasonal rhinitis, was found to inhibit the hERG channel with an IC_{50} of 56 nM [2]. It caused significant QT prolongation and had to be withdrawn from the market. Interestingly, fexofenadine, a close analog of terfenadine, does not inhibit the hERG channel and is free of any cardiac-related side-effects (Fig. 13.1 c) [2]. Nowadays, pharmacophore [3], structure-based [4–6], 3D QSAR [3] and neural network models [7] have been developed for the hERG channel to support the chemical optimization of novel drug candidates towards molecules having no hERG-mediated cardiac side-effects.

Historically, the discovery of drugs acting at G-protein coupled receptors (GPCRs) has been extremely successful with 50% of all recently launched drugs targeting against GPCRs [8]. GPCRs form a large protein family that plays an important role in many physiological and patho-physiological processes. Especially the subfamily of biogenic amine-binding GPCRs has provided excellent drug targets (given in parentheses) for the treatment of numerous diseases (Table 13.1): schizophrenia (mixed D2/D1/5-HT_2 antagonists), psychosis (mixed D2/5-HT_{2A} antagonists), depression (5-HT_1 agonists), migraine (5-HT_1 agonists), allergies (H1 antagonists), asthma (β2 agonists, M1 antagonists), ulcers (H2 antagonist) or hypertension (α1 antagonist, β1 antagonist). Although representing excellent therapeutic targets, the central role that many of the biogenic amine-binding GPCRs play in cell signaling also poses a risk for new drug candidates which reveal side-affinities towards these receptor sites: These candidates have the potential to interfere with the physiological signaling process and to cause undesired effects in preclinical or clinical studies. For example, the α_{1A} adrenergic receptor modulates the relaxation of the vascular muscle tone and is therefore important for blood pressure regulation. It has been suggested as an *anti-target* that mediates cardiovascular side-effects of many drug candidates causing orthostatic hypotension, dizziness and fainting spells [9, 10]. Furthermore, in order to obtain a clean clinical profile for novel development candidates, strong molecular interactions with dopamine and serotonin receptors (such as 5-HT$_{2A}$ and D2 receptors) representing the molecular targets for many anti-psychotics (e.g. olanzapine and risperidone) need to be avoided. Table 13.1 lists some GPCR *anti-targets* and the potential side-effects mediated by high-affinity antagonists of these receptors.

13.2
In Silico Tools for GPCR Anti-target Modeling

In order to monitor affinity profiles of new drug candidates and to predict undesired GPCR-mediated side-effects, we have established a panel of biogenic amine receptor binding assays. Profiling of several hundred compounds within this panel showed that several lead compounds entering the chemical optimization phase reveal affinities towards several members of the biogenic amine *anti-target* panel. Reliable *in silico* tools to identify compounds with strong GPCR *anti-target* affinity and computational models to guide the chemical optimization towards compounds having a more favorable GPCR affinity profile thus appear to be of great value for the design and development of new drug candidates.

The challenge in the generation of these pharmacophores for *anti-target* modeling is the requirement that these models need to describe the receptor interaction points not only for a single chemical series but for several different compound classes. In addition, these *cross-chemotype* pharmacophores need to capture sufficient pharmacophoric points to describe all relevant receptor–ligand interactions. Three-dimensional pharmacophore models rationalizing the affinity of several different chemical series have recently been described for the α_{1A}, the 5-HT$_{2A}$ and the D2 receptors [11]. This chapter is focused on the *anti-target* pharmacophore models of the α_{1A} adrenergic receptor, the prototype of a GPCR *anti-target*. Using the α_{1A} receptor as an example, the generation of validation of *cross-chemotype* pharmacophore models and first applications of these *anti-target* models will be described. It will be shown how these *anti-target* pharmacophore models are capable of rationalizing the strong *anti-target* affinity of novel lead series and how they can guide the chemical optimization towards development candidates with a superior safety index.

13.3
GPCR Anti-target Pharmacophore Modeling: the α_{1a} Adrenergic Receptor

Pharmacophore models for the α_{1A} adrenergic receptor have also been described by others [12, 13]. Barbaro et al. used a series of pyridazinone derivatives based

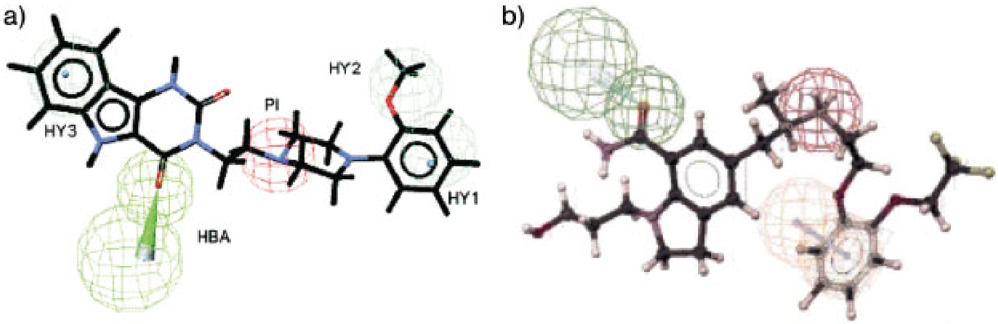


Fig. 13.2 α_{1A} adrenergic receptor pharmacophore models. (a) By Barbaro et al. using a series of pyridazinone derivatives and biological data from rat receptor [12]. Reprinted with permission from [12]. Copyright 2001, American Chemical Society. (B) By Bremner et al. derived from a diverse set of 38 compounds [13, 14]. Reprinted from [14]. Copyright 2000, with permission from Elsevier.

on biological data on the rat receptor as a training set for pharmacophore generation [12]. The model appears to be well suited for the quantitative prediction of the biological activity of the training set molecules and chemically closely related series (Fig. 13.2 a). However, it does not represent a *cross-chemotype* model suitable for mapping a diverse set of different α_{1A} chemical series. The model generated by Bremner and co-workers, on the other hand, was derived from a diverse set of 38 compounds [13, 14]. However, it comprises only three pharmacophoric features and is therefore quite generic and cannot be expected to be very specific for the α_{1A} receptor (Fig. 13.2 b). As both available models appeared to be unsuitable for the purpose of *anti-target* modeling, *cross-chemotype* pharmacophore models for the human α_{1A} adrenergic receptor have been generated [11].

13.3.1 Generation of Cross-chemotype Pharmacophore Models

The common-features hypothesis generation module of Catalyst 4.7 [15] (termed *HipHop*) was used for the generation of two *cross-chemotype* 3D pharmacophores describing α_{1A} antagonists. The common-features hypothesis generation module is designed specifically for finding chemical features shared by a set of compounds belonging to different chemical classes. It provides the compounds' relative alignments with the hypothesis expressing these common features. The training set used for the generation of the α_{1a} adrenergic pharmacophore model was extracted from the Aureus database, a structure–activity database for GPCR ligands compiled and maintained by Aureus Pharma [16]. The database covers all biological data published on GPCRs and provides chemical structural information, references to the original publication or patent and detailed information on the experimental conditions (e.g. assay type, cell line or radioligand used).

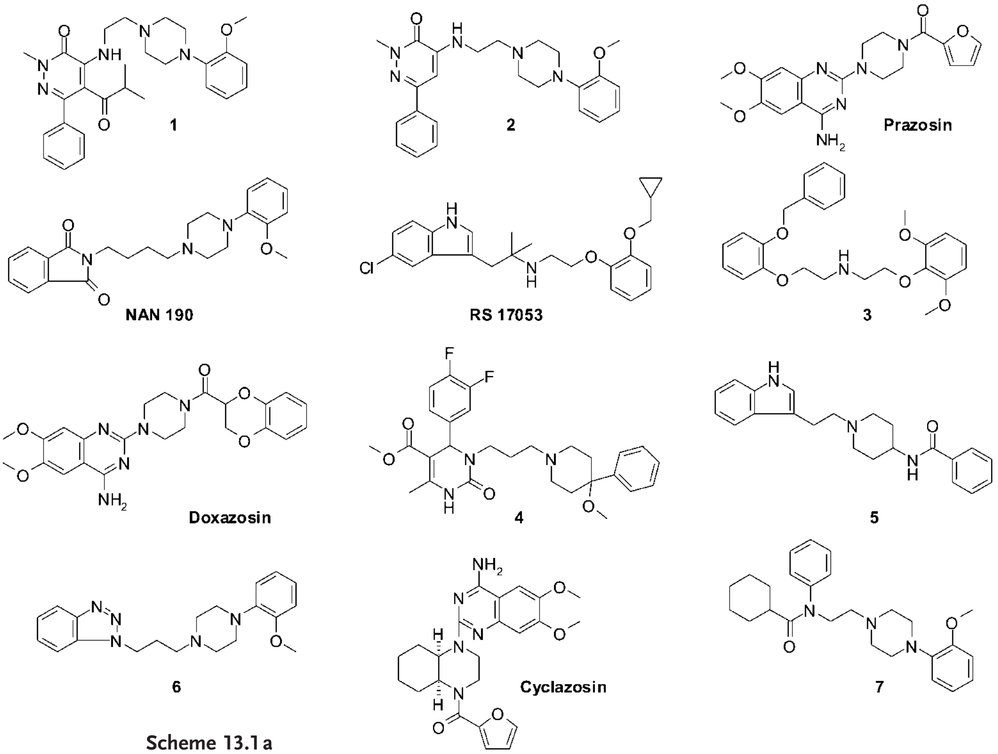
Table 13.2 Training set molecules for α_{1A} adrenergic receptor and their affinities measured in a radioligand displacement assay

Class	Compound	K_i (nM) [11]
I	**1**	0.2
I	**2**	0.2
I	Prazosin	0.3
I	NAN 190	0.4
I	RS 17053	0.5
I	**3**	0.5
I	Doxazosin	0.8
I	**4**	1.0
I	**5**	2.8
I	**6**	4.6
I	Cyclazosin	12.3
I	**7**	27.1
I	**8**	44
I	**9**	72.4
II	YM 617	0.04
II	WB 4104	0.1
II	ARC 239	0.4
II	BE 2254	0.4
II	Spiperone	25.1
II	**10**	28.2

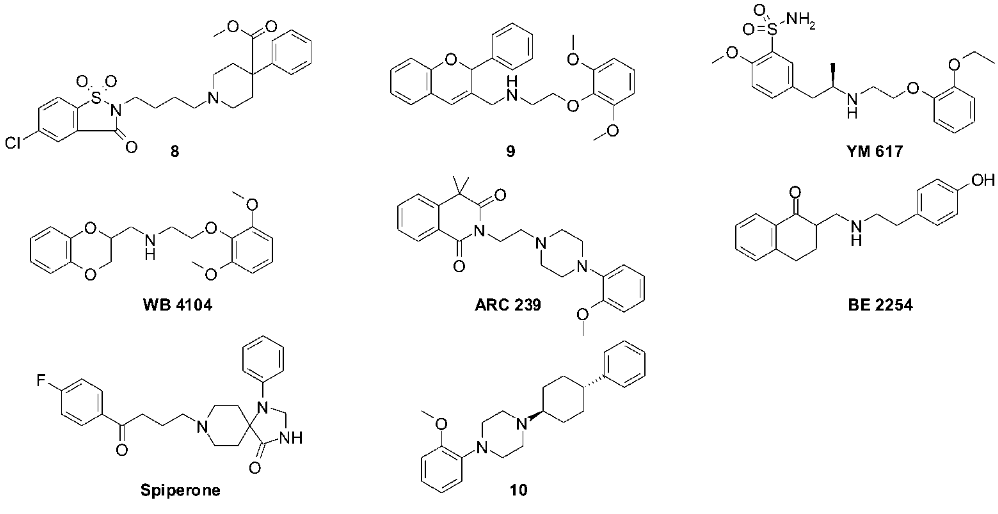
Adrenergic α_{1A} receptor antagonists with K_i values <100 nM tested against the recombinant human wild-type receptor were extracted from the Aureus database. The structural analysis of the compounds reveals that they can be grouped into two classes, probably binding overlapping but not identical binding sites within the receptor. Thus two diverse training sets covering chemotype examples of both classes were selected: (i) class II antagonists are represented by six compounds revealing two aromatic rings and a positively ionizable group positioned two to four bond lengths from the aromatic features; and (ii) 14 representatives of class I antagonists, revealing a positively ionizable group which is separated from the first aromatic ring by two to three bond lengths and by six to seven bond lengths to the second aromatic ring. Table 13.2 and Scheme 13.1 show the chemical structures of both sets of compounds used for pharmacophore generation together with the reported binding affinities.

13.3.2 Description of Cross-chemotype Pharmacophore Models

The two common-feature pharmacophores are depicted in Fig. 13.3 a and b showing the mapping on to a reference molecule representative for both classes of high affinity α_{1A} antagonists [11]. The models describe the key chemical features required for binding of structurally diverse ligands to this adrenergic re-



Scheme 13.1 a



Scheme 13.1 b

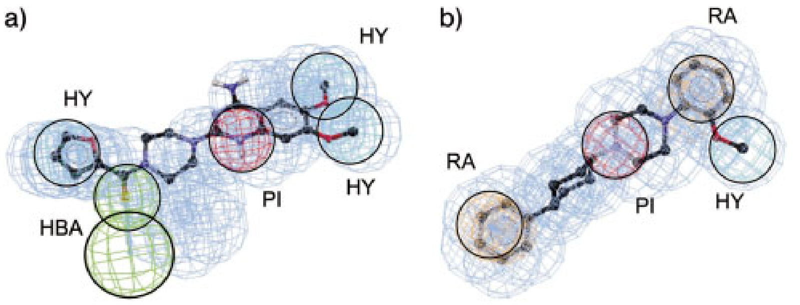


Fig. 13.3 Common-feature pharmacophores of α_{1a} adrenergic receptor antagonists [11]. On to each pharmacophore the reference has been mapped. (a) Class I pharmacophore model aligned to prazosin; (b) class II pharmacophore model aligned to compound **10**. Pharmacophoric features are red for positively ionizable (PI), green for hydrogen bond acceptors (HBA), light blue for hydrophobic (HY) and orange for ring aromatic (RA). Shape restraints are shown in light blue.

ceptor subtype: The class I pharmacophore (Fig. 13.3 a) represents a five-point pharmacophore, which is composed of three hydrophobic moieties connected through a positively ionizable group (matched by the N2 group of the quinazoline ring) and a hydrogen-bond acceptor group (mapped by the amide group of the shown compound prazosin). The class II pharmacophore (Fig. 13.3b) describes the four main pharmacophoric points of the smaller class of α_{1A} ligands lacking the hydrogen-bond acceptor group: two ring aromatic features, one hydrophobic and one positively ionizable feature.

The similarity of the right part of both pharmacophores (class I, positively ionizable, hydrophobic, hydrophobic; class II, positively ionizable, hydrophobic, ring aromatic) indicates that the "head" groups of class I and class II ligands mapping this part of the pharmacophore interact with the same site at the adrenergic receptor (see Section 13.3.4). However, both pharmacophores also reflect the differences between the two different classes of α_{1A} receptor antagonists found in the left part of the molecules as shown in Scheme 13.1. Class I ligands appear to share an acceptor group and a second hydrophobic group separated from the central positive charge by 9.5 Å (5–6 bond lengths). The shorter class II ligands, however, reveal an aromatic group connected by only 7.2 Å (2–4 bond lengths) to the positively charged nitrogen atom.

13.3.3 Validation of Anti-target Pharmacophore Models

13.3.3.1 Virtual Screening: Hit Rates and Yields

The purpose of the *anti-target* pharmacophores is to recognize and rationalize *anti-target* side-affinities within chemotypes different from those used in the training set. Hence it appears crucial to validate the pharmacophore hypotheses using external test set molecules, which have not been used for pharmacophore

Table 13.3 Hit rates and yields from virtual screen of MDDR test set database using both α_{1A} pharmacophores as selection filter [11]

Class	No. of virtual hits	No. of identified α_{1A} antagonists	Hit rate (%) [a)]	Yield (%) [b)]
I	82	26	32	52
II	146	42	29	84
I or II	168	45	27	90

a) Hit rate = (number of true actives in hit list)/(number of compounds in hit list) × 100.

b) Yield = (number of true actives in hit list)/(number of true actives in full database) × 100.

model generation. The predictive power of both α_{1A} pharmacophore models was therefore evaluated by virtual screening using a test database of 50 known α_{1A} antagonists embedded in a database consisting of 1000 drug-like molecules (active and inactive sets were taken from the MDL MDDR database) [11]. To mark the predictive power hit rates [hit rate = (number of α_{1A} antagonists in hit list)/(total number of compounds in hit list) × 100] and yields [yield = (number of α_{1A} antagonists in hit list)/(number of α_{1A} antagonists in full database) × 100] were calculated. The results are presented in Table 13.3. For both pharmacophores a hit rate of approximately 30% was obtained, which is six times higher than a random selection. In addition, with a yield of 84%, the class II pharmacophore was able to identify most of the α_{1A} antagonists within the test set, still revealing excellent specificity as reflected by a good hit rate.

The virtual screening suggests that especially the less stringent class II four-point pharmacophore is suitable to recognize most of the known α_{1A} antagonists and to provide mappings of compounds having significant α_{1A} affinity. Taken together, both pharmacophores are able to recognize 90% of the α_{1A} antagonists embedded in the test data set.

13.3.3.2 **Virtual Screening: Fit Values and Enrichment Factors**

In many cases, the performance of a pharmacophore-based virtual screen can be improved when the quality of the mapping to the respective pharmacophore is considered. Fit values of all test set molecules on to both pharmacophores can be calculated and the compounds can be sorted using the fit value of their mappings. The resulting enrichment graphs are shown Fig. 13.4 for both α_{1A} pharmacophores.

Both enrichment curves show a steep beginning, almost parallel to the ideal curve (black line). The flattening of the curves towards the right can be explained by the fact that some α_{1A} compounds of the MDDR database cannot be mapped by the pharmacophore and thus obtain fit values of 0. The steepness of the enrichment curve on the left, however, reflects, that among the top-ranked

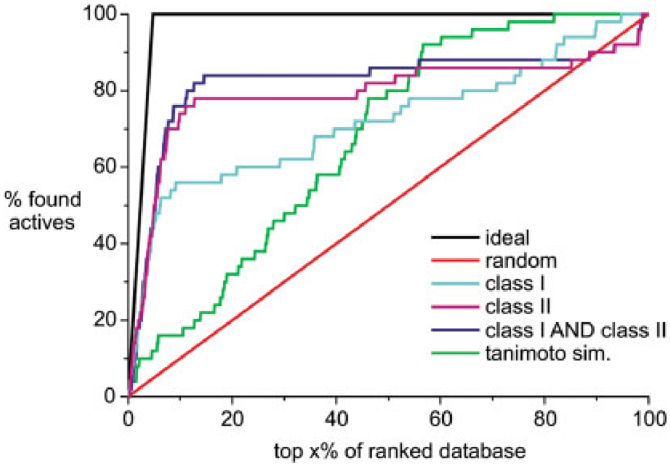


Fig. 13.4 Enrichment graph for virtual screening of α_{1A} antagonists embedded into a random MDDR library comprising 1048 compounds [11]. The curve shows the relative ranking of the 50 α_{1A} antagonists. Database compounds are ranked along the *x*-axis based on the fit value calculated for the mapping on the respective pharmacophore. Cyan, class I pharmacophore; magenta, class II pharmacophore; blue, sum of class I and class II; green, maximum Tanimoto similarity to reference compounds in training set; red, random; black, ideal. The enrichment by the class II pharmacophore (magenta) at a yield of 50% is 10-fold better than by a random selection.

compounds of the database (e.g. 1–10%), a high percentage of true α_{1A} ligands can be identified by these *anti-target* pharmacophores (e.g. among top 10 scored virtual hits using class II pharmacophore six are α_{1A} antagonists, indicating a hit rate of 60% among the top 1% of the virtual hits). Hit values, yields and enrichment factors (hit rate found versus random selection) are listed for both pharmacophores for the top 5% scorers in Table 13.4: (i) both pharmacophores

Table 13.4 Hit rates, yields and enrichment factors for the top 5% scorers from virtual screen of MDDR test set database using both α_{1A} pharmacophores as selection filters. All compounds were scored based on their fit value on to the respective pharmacophore

	No. of α_{1A} antagonists among top 5% of database	Hit rate (%) [a)]	Yield (%) [b)]	Enrichment factor
I	22	42	44	8.8
II	25	48	50	10

a) Hit rate = (number of true actives in hit list)/(number of compounds in hit list) × 100.

b) Yield = (number of true actives in hit list)/(number of true actives in full database) × 100.

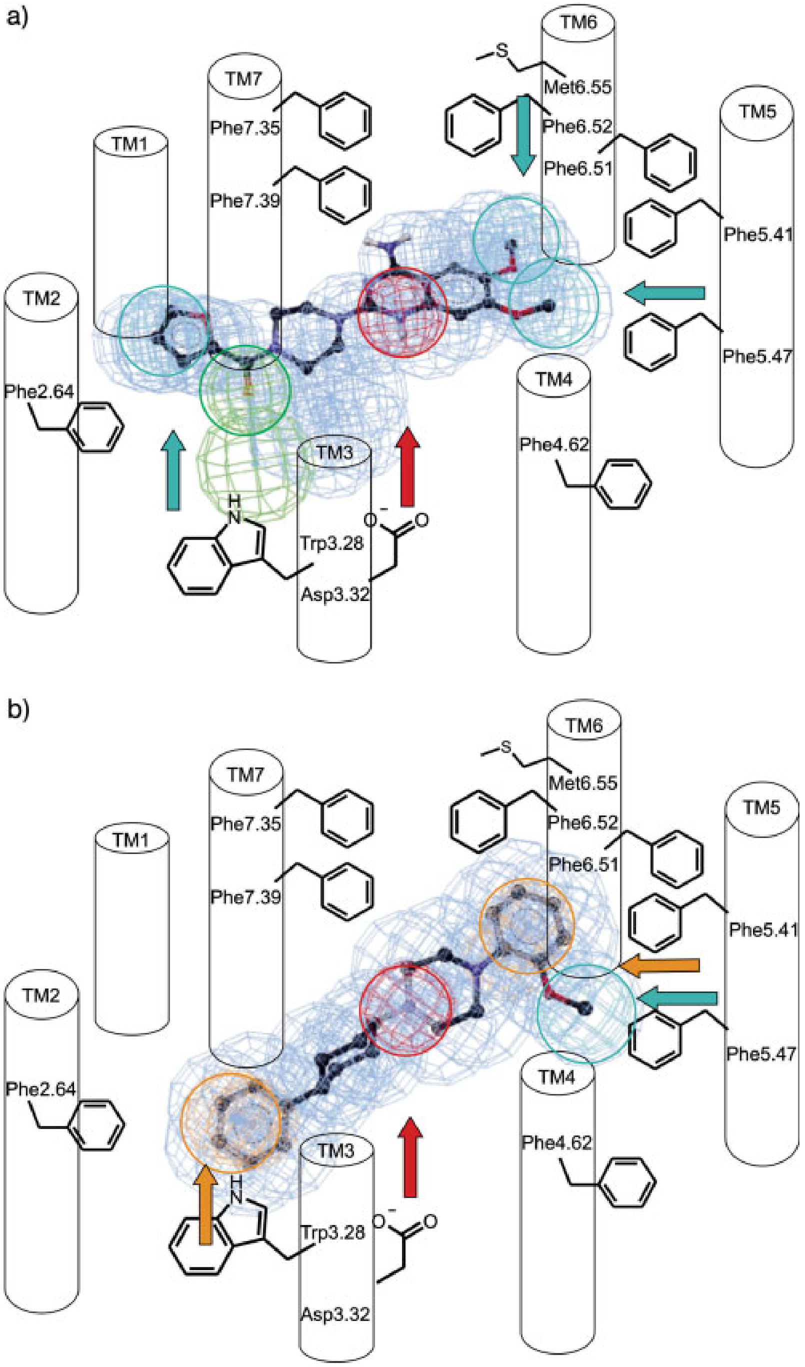
provide an excellent yield with 44 and 50% of the α_{1A} antagonists being found among the top 5% scorers of the ranked database, respectively; (ii) enrichment factors between 9 and 10 were obtained comparing the hit rate of the pharmacophore-based selection with a random selection. The excellent hit rate and yield generated cannot be explained by the structural similarity of MDDR test set molecules to the Aureus training set molecules. Figure 13.4 reveals that the yields and hit rates obtained by ranking the database compounds based on their maximal Tanimoto similarity (calculated based on 2D fingerprints: green curve) to one of the six Aureus class II training set molecules is not significantly higher than a random selection.

The excellent performance in terms of yield and enrichment factor of both pharmacophores suggests that both pharmacophores can also be useful filters for virtual screening to identify α_{1A} antagonists within large compound repositories. Indeed, both pharmacophores have been successfully applied in a virtual screening approach combining pharmacophore-based and homology model-based virtual screening [17]. Using this combined approach, novel α_{1A} antagonists exhibiting nanomolar affinity could be identified from the corporate compound collection.

13.3.4 Mapping of Pharmacophore Models into Receptor Site

Numerous site-directed mutagenesis studies have provided a conclusive picture for the molecular interactions between receptor-activating biogenic amines (e.g. serotonin, epinephrine, dopamine) and their receptors [18–22]: a highly conserved aspartate residue in transmembrane helix TM3 (Asp 3.32 according to the Ballosteros-Weinstein nomenclature) [23] conserved serine residues in TM5 (e.g. Ser 5.42 and Ser 5.46 for α_{1A}) and also hydrophobic phenylalanine residues from TM6 have been identified as being important for agonist binding. In addition, through mutational studies and comparative affinity determinations based on ligand binding, the essential amino acids involved in antagonist recognition could be identified for the α_{1A} receptor [22, 24, 25]. According to these studies,

Figure 13.5 Topographical interaction model of α_{1A} adrenergic receptor generated based on public site-directed mutagenesis. Both pharmacophore models have been mapped into the topographical model of the receptor. The model reveals putative receptor interaction sites for most of the pharmacophoric features observed within each antagonist class. (a) Class I pharmacophore with prazosin as reference compound; (b) class II pharmacophore with compound **10** as reference. Pharmacophoric features are red for positively ionizable (PI), green for hydrogen-bond acceptors (HBA), light blue for hydrophobic (HY) and orange for ring aromatic (RA). Shape restraints are shown in light blue. Arrows indicate putative molecular interactions between the pharmacophoric points and receptor sites. Color codes indicate the type of molecular interaction: light blue, hydrophobic; red, salt bridge to negative ionizable group from receptor; orange, aromatic stacking interaction.


a)
TM1
TM2
TM3
TM4
TM5
TM6
TM7
Phe7.35
Phe7.39
Met6.55
Phe6.52
Phe6.51
Phe5.41
Phe5.47
Phe2.64
Phe4.62
Trp3.28
Asp3.32
H
N
S
O
O
-
b)
TM1
TM2
TM3
TM4
TM5
TM6
TM7
Phe7.35
Phe7.39
Met6.55
Phe6.52
Phe6.51
Phe5.41
Phe5.47
Phe2.64
Phe4.62
Trp3.28
Asp3.32
H
N
S
O
O
-

the binding pocket of the prototype biogenic amine receptor antagonist stretches from the agonist binding site formed by TM3, TM5 and TM6 – interacting with the antagonist's "head" group – towards the transmembrane helices TM1, TM2 and TM7, which have been suggested to harbor the lipophilic "tail" moiety of several antagonists. Based on these experimental data, a topographical interaction model for the α_{1A} receptor has been generated as shown in Fig. 13.5 [11].

The two pharmacophore models were mapped into the topographical interaction model to indicate the putative interaction points of each pharmacophoric feature with its receptor: (i) the positive ionizable pharmacophoric feature is thought to be anchored via a salt bridge to the conserved aspartate residue in TM3; (ii) the hydrophobic and aromatic features of the "head" moieties are harbored within hydrophobic micro domains formed by aromatic and aliphatic side-chains of TM4, TM5 and TM6; the "floor" of this hydrophobic micro domain is formed by several conserved aromatic amino acids (Phe 6.44, Trp 6.48, Phe 5.47), which are conserved among the family of biogenic amine GPCRs; and (iii) the hydrophobic or ring aromatic feature observed within the "tail" moiety of almost all α_{1A} antagonists is likely to be directed towards aromatic and hydrophobic residues within TM helices TM3 (Trp 3.28) and TM2 (Phe 2.64).

13.3.5 Guidance of Chemical Optimization to Avoid GPCR-mediated Side-effects

Our company has established a panel of biogenic amine receptor binding assays to monitor affinity profiles of novel drug candidates. So far, several hundred compounds coming from GPCR directed libraries have been profiled against this panel, revealing that approximately 14% of the profiled compounds have moderate α_{1A} affinity in the submicromolar range and 3.5% of all tested compounds reveal strong α_{1A} binding with affinities <100 nM. These experimental results show once more the need to optimize the side-affinity profile of several compounds to support the further development of these drug candidates.

The main application of the generated *anti-target* pharmacophore hypotheses is to rationalize these experimental findings by providing pharmacophore mappings. Recognition of the key chemical features that are responsible for the side-affinities of a chemical series could then provide guidance for the chemical optimization of the series towards compounds having a more favorable side-affinity profile. Most importantly, 80% of all experimentally identified α_{1A} binders could be mapped on to the class II α_{1A} pharmacophore fulfilling all four pharmacophore points. Mapping of one of these compounds on to the α_{1A} class II pharmacophore is shown in Fig. 13.6. The mapping directly suggests the chemical features that are mediating the strong affinity towards this subtype of the adrenergic receptor: these are the positive charge of the piperazine moiety, the ortho-substituted phenyl ring at the 4-position of the piperazine and the aromatic ring of the phenoxy side-chain. The mapping thus provides direct guid-

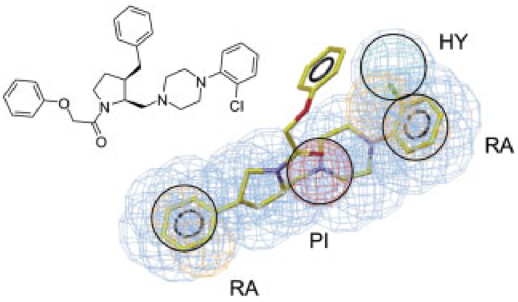


Figure 13.6 Pharmacophore mapping of high-affinity α_{1A} binder mapped on to class II adrenergic α_{1A} pharmacophore model. All pharmacophoric points are mapped. The alignment suggests that removal of the chlorine substituent within the 4-phenyl-piperidine will reduce the unfavorable site affinity on α_{1A}.

ance for the chemical optimization of the respective series to avoid the undesired α_{1A} affinity (e.g. removal of the chlorine substituent within the 4-phenyl piperazine compound series).

13.4 Conclusion

As demonstrated in this chapter, *cross-chemotype* pharmacophore models can be generated from training sets covering chemically diverse ligands. These models describe the key pharmacophoric features required for receptor binding. When site-directed mutagenesis data are available, the models can be mapped into the receptor recognition site linking each pharmacophoric point to its interaction site within the receptor. When applied as filters within virtual screening, these 3D pharmacophores offer acceptable levels of predictivity as revealed by good yields and high enrichment factors. Furthermore, such *in silico* tools can be directly applied to guide the chemical optimization of novel GPCR drug candidates towards clinical candidates with less α_{1A}-mediated side-effects (e.g. orthostatic hypotension, dizziness and fainting spells).

References

1 Keating, M.T., Sanguinetti, M.C. Molecular and cellular mechanisms of cardiac arrhythmias. *Cell* **2001**, *104*, 569–580.

2 Pearlstein, R.A., Vaz, R.J., Kang, J., Chen, X.L., Preobrazhenskaya, M., Shchekotikhin, A.E., Korolev, A.M., Lysenkova, L.N., Miroshnikova, O.V., Hendrix, J., Rampe, D. Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829–1835.

3 Cavalli, A., Poluzzi, E., De, P.F., Recanatini, M. Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K^+ channel blockers. *J. Med. Chem.* **2002**, *45*, 3844–3853.

4 Pearlstein, R., Vaz, R., Rampe, D. Understanding the structure–activity relationship of the human ether-a-go-go-related gene cardiac K^+ channel. A model for bad behavior. *J. Med. Chem.* **2003**, *46*, 2017–2022.

5 Mitcheson, J.S., Chen, J., Sanguinetti, M.C. Trapping of a methanesulfonanilide by closure of the HERG potassium channel activation gate. *J. Gen. Physiol* **2000**, *115*, 229–240.

6 Mitcheson, J.S., Chen, J., Lin, M., Culberson, C., Sanguinetti, M.C. A structural basis for drug-induced long QT syndrome. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12329–12333.

7 Roche, O., Trube, G., Zuegge, J., Pflimlin, P., Alanine, A., Schneider, G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBiochem* **2002**, *3*, 455–459.

8 Klabunde, T., Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBiochem.* **2002**, *3*, 928–944.

9 Kehne, J.H., Baron, B.M., Carr, A.A., Chaney, S.F., Elands, J., Feldman, D.J., Frank, R.A., van Giersbergen, P.L., McCloskey, T.C., Johnson, M.P., McCarty, D.R., Poirot, M., Senyah, Y., Siegel, B.W., Widmaier, C. Preclinical characterization of the potential of the putative atypical antipsychotic MDL 100,907 as a potent 5-HT2A antagonist with a favorable CNS safety profile. *J. Pharmacol. Exp. Ther.* **1996**, *277*, 968–981.

10 Peroutka, S.J., U'Prichard, D.C., Greenberg, D.A., Snyder, S.H. Neuroleptic drug interactions with norepinephrine alpha receptor binding sites in rat brain. *Neuropharmacology* **1977**, *16*, 549–556.

11 Klabunde, T., Evers, A. GPCR anti-target modeling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side-effects. *ChemBiochem* **2005**, *6*, 876–889.

12 Barbaro, R., Betti, L., Botta, M., Corelli, F., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G., Corsano, S. Synthesis, biological evaluation and pharmacophore generation of new pyridazinone derivatives with affinity toward alpha(1)- and alpha(2)-adrenoceptors. *J. Med. Chem.* **2001**, *44*, 2118–2132.

13 Bremner, J.B., Griffith, R., Coban, B. Ligand design for alpha(1) adrenoceptors. *Curr. Med. Chem.* **2001**, *8*, 607–620.

14 Bremner, J.B., Coban, B., Griffith, R., Groenewoud, K.M., Yates, B.F. Ligand design for alpha1 adrenoceptor subtype selective antagonists. *Bioorg. Med. Chem.* **2000**, *8*, 201–214.

15 *Catalyst, Version 4.7*. Accelrys, San Diego, CA, **2004**.

16 Aureus Pharma, www.aureus-pharma.com, **2004**.

17 Evers, A., Klabunde, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1a receptor. *J. Med. Chem.* **2005**, *48*, 1088–1097.

18 Chambers, J.J., Nichols, D.E. A homology-based model of the human 5-HT2A receptor derived from an *in silico* activated G-protein coupled receptor. *J. Comput.-Aided Mol. Des* **2002**, *16*, 511–520.

19 Hwa, J., Graham, R.M., Perez, D.M. Identification of critical determinants of alpha 1-adrenergic receptor subtype selective agonist binding. *J. Biol. Chem.* **1995**, *270*, 23189–23195.

20 Ishiguro, M. Ligand-binding modes in cationic biogenic amine receptors. *ChemBiochem*. **2004**, *5*, 1210–1219.

21 Pollock, N.J., Manelli, A.M., Hutchins, C.W., Steffey, M.E., MacKenzie, R.G., Frail, D.E. Serine mutations in transmembrane V of the dopamine D1 receptor affect ligand interactions and receptor activation. *J. Biol. Chem*. **1992**, *267*, 17780–17786.

22 Shi, L., Javitch, J.A. The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol*. **2002**, *42*, 437–467.

23 Ballosteros, J.A., Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations of G protein-coupled receptors. *Methods Neurosci*. **1995**, *25*, 366–428.

24 Hamaguchi, N., True, T.A., Saussy, D.L., Jr, Jeffs, P.W. Phenylalanine in the second membrane-spanning domain of alpha 1A-adrenergic receptor determines subtype selectivity of dihydropyridine antagonists. *Biochemistry* **1996**, *35*, 14312–14317.

25 Hamaguchi, N., True, T.A., Goetz, A.S., Stouffer, M.J., Lybrand, T.P., Jeffs, P.W. Alpha 1-adrenergic receptor subtype determinants for 4-piperidyloxazole antagonists. *Biochemistry* **1998**, *37*, 5730–5737.

14
Pharmacophores for Human ADME/Tox-related Proteins

Cheng Chang and Sean Ekins

14.1
Introduction

The last decade has witnessed an enormous increase in the number of compounds flowing through the drug discovery and development pipeline in the pharmaceutical industry, primarily owing to the advent of combinatorial chemistry and high-throughput screening (HTS) (Rodrigues 1997). These new technologies may have increased the chances of finding new lead compounds beyond traditional medicinal chemistry methods. However, expensive phase II and phase III clinical trial failures related to unsatisfactory ADME/Tox properties have also increased. In order to improve the rate of success in the more costly downstream stages of drug development, ADME/Tox evaluations have been shifted into the very early part of the discovery process. New technologies such as *in vitro* HTS and *in silico* approaches have been developed to meet the new challenges of large compound numbers and shortened cycle times that are characteristic of this phase of drug discovery. As the most cost-effective method, *in silico* screening has the additional advantage of being able to reduce significantly the experimental effort in the screening phase of drug discovery (Boobis et al. 2002). *In silico* approaches include three-dimensional quantitative structure–activity relationship (3D-QSAR) and pharmacophore modeling, which can be used directly as database searching methods. Recently, there have been several reviews focused on different aspects of computational ADME/Tox and more recently one of these (Ekins and Swaan 2004) has reviewed *in silico* approaches to modeling the specific proteins involved in determining ADME/Tox properties. The primary aim of this chapter is to review briefly some of the recent applications of pharmacophore technologies in drug discovery ADME/Tox studies. For other QSAR methods the reader is referred to several useful recent reviews (Barratt and Rodford 2001; Wessel and Mente 2001; Butina et al. 2002; Greene 2002; van de Waterbeemd and Gifford 2003).

Absorption into the bloodstream is the first step for drugs to reach their targets followed by distribution to tissues. The drugs are then metabolized into more readily excreted forms. All of these aspects are significantly mediated or

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

influenced by transporters, enzymes, ion channels and receptors. This complex interplay of different proteins coordinates to absorb nutrients and protect against the accumulation and toxic compounds. The potential overlap or competition of multiple compounds with affinity for the same protein raises the potential for possible drug–drug interactions (DDI), which could result in either reduced drug efficiency or increased drug toxicity owing to extended bioavailability. Methods to predict these types of interactions effectively are highly desirable and in recent years we have seen the focus of much research on computational methods. The interest in computational models based on *in vitro* data for predicting potential drug interactions via these multiple proteins (Ekins and Swaan 2004) follows the computational assessment of properties such as absorption (Palm et al. 1996, 1998; Wessel et al. 1998; Clark 1999; Kelder et al. 1999; Norinder et al. 1999; Oprea and Gottfries 1999; Stenberg et al. 1999; Egan et al. 2000; Ekins et al. 2001 b; Raevsky et al. 2001; Stenberg et al. 2001; Zhao et al. 2001; Niwa 2003), which now occurs much earlier in drug discovery than perhaps a decade ago or less (Ekins et al. 2000 c; Ekins and Rose 2002). These recently developed different computational approaches are undergoing validation yet they represent a means to improve the productivity of the drug discovery process. Owing to their highly parallel nature, computational methods are also the fastest and most cost-effective method for indication of possible toxic consequences caused by interfering with the above multiple proteins (Ekins et al. 2000 b), providing molecular insight and suggesting new hypotheses for rapid testing *in vitro*. This ability to screen large numbers of molecules computationally parallels the increase in throughput of *in vitro* assays for drug discovery over the past decade. Both *in vitro* and computational approaches can be used in tandem in an iterative manner to improve the developed models.

A pharmacophore is the representation of the spatial arrangement of structural features that are required for a certain biological activity. Pharmacophore development theory and applications have been explained in detail elsewhere (Guner 2000; Kurogi and Guner 2001; Guner 2002; Guner et al. 2004). Three widely used pharmacophore perception programs, Catalyst, GASP and DISCO, have been thoroughly described and compared by Patel et al. (2002) and the interested reader is referred to their paper for further details of the methods. The ultimate goal of *in silico* studies in ADME/Tox is to predict the disposition behavior of drugs in the whole body by incorporating all kinetic processes into one global model. However, currently only a very limited number of preliminary models at the protein level have been conducted (Yamashita and Hashida 2004). We are starting to observe a more “systems-based” approach to ADME/Tox as various databases on the interactions of small molecules with proteins are combined with multiple QSAR models and other ADME/Tox tools (Ekins et al. 2005 a, c, d). However, most published studies describe modeling of the individual protein targets related to a single ADME/Tox property, and these models will be the primary focus of this chapter.

The key proteins that have been modeled with such pharmacophore methods include the major cytochrome P450 (CYP) enzymes, UDP-glucuronosyltransferase (UGT), P-glycoprotein (P-gp), breast cancer resistance protein (BCRP), peptide transporter (PepT1), apical sodium-dependent bile acid transporter (ASBT), sodium taurocholate-transporting polypeptide (NTCP), nucleoside transporter, organic cation transporter (OCT), multiple nuclear hormone receptors including the pregnane X receptor (PXR) and human ether-a-go-go (hERG) potassium channel. We describe below some of the pharmacophore modeling efforts to date in more detail.

14.2 Cytochrome P450

Drug metabolism via the liver is the primary elimination mechanism for the majority of drugs and xenobiotics in humans. Cytochrome P450s are the most important enzymes (phase I) whereas UDP-glucuronosyl transferases are the key (phase II) enzymes for clearance of drugs by the liver (Williams et al. 2004a). Without a crystal or 3D structure for many of these proteins, the prediction of whether a molecule binds to them rests with our limited knowledge of the specificity and selectivity of the binding sites derived from *in vivo* and *in vitro* data. Various *in vitro* systems are now widely used to study metabolism (VandenBranden et al. 1998; Ekins et al. 1999e) and characterize the potential for DDI mediated by P450 enzymes (Ekins et al. 1997, 1998c, 2000c; Ekins and Wrighton 1999; Margolis et al. 2000; Gao et al. 2002). P450s have affinity for structurally diverse hydrophobic molecules in human and others species used as pharmacological (Mankowski et al. 2000) or toxicological models (Mankowski et al. 1999) and represent the most extensively studied family of drug-metabolizing enzymes. As early as the 1960s, mathematical quantitative structure–activity relationship (QSAR) models were used to assist the understanding of drug metabolism and to improve the design of new drugs (Hansch et al. 1968). In the 1990s the availability of more complex and graphically intensive software tools enabled computational pharmacophore-type models describing key molecular features of ligands for human CYP1A2 (Fuhr et al. 1993) and CYP2C9 (Jones et al. 1996). In this software, the key molecular features are translated into spheres, points or a mesh on to which molecules themselves can be mapped in 3D space.

Recent research has described and compared many pharmacophores that have been generated for P450s (Ekins et al. 2001a; de Groot and Ekins 2002), including pharmacophores derived from manual alignments of molecules (de Groot et al. 1999a, b), providing insight into the important features for interaction of ligands and the proteins. The *in vitro* data training sets have varied in size, although pharmacophores generally consist of less than 50 molecules and the size of the corresponding test sets has also varied accordingly. The human P450s which have received most of the focus of computational approaches to date include CYP2B6, CYP2C9, CYP2D6, CYP3A4, CYP3A5 and CYP3A7 (Ekins et al. 1997, 1998c, 1999a–d, 2000a, 2003b; Ekins and Wrighton 2001; Snyder et al. 2002). Models have also been built after analysis of the literature and using novel data including recombinant-derived kinetic values for CYP2B6 (Ekins et al. 1999 c; Wang and Halpert 2002), CYP2C9 (Jones et al. 1996; Afzelius et al. 2001) and CYP2D6 (Snyder et al. 2002), although these efforts have ventured beyond strictly using pharmacophores in some cases. In particular, the characterization of CYP2B6 with *in vitro* methods resulted in the first pharmacophore and 3D-QSAR models published for substrates of this enzyme, suggesting at least three hydrophobic interactions and a hydrogen-bond acceptor are important features for binding (Ekins et al. 1999 c). The combined *in vitro* and computational approaches for CYP2B6 also increased awareness of the potential of this enzyme to bind similar ligands to CYP3A4. Complex *in vitro* enzyme kinetics were also first reported for CYP2B6 in these studies, which are indicative of multiple molecules binding simultaneously to the enzyme (Ekins et al. 1997, 1998 b, c, 1999 c; Ekins and Wrighton 1999), a characteristic at that time previously only widely observed with CYP3A4. The CYP3A family of enzymes are the most important in terms of human drug metabolism (Wrighton et al. 2000) because they have a broad substrate specificity. Computational pharmacophores for CYP3A4 have therefore been derived for substrates (Ekins et al. 1999 d) and inhibitors (Ekins et al. 1999 b, 2003 a, b) using kinetic constants K_{m}, $K_{\mathrm{i}\ (\mathrm{apparent})}$ and IC_{50} data (Ekins et al. 2001 a).

The computational pharmacophore approach has also been used to provide the first example of a model for the important features of molecules which increase their own metabolism (autoactivators) via CYP3A4. This autoactivator pharmacophore for CYP3A4 possessed three hydrophobic features and one hydrogen-bond acceptor (Ekins et al. 1999 d), corresponding with residues identified using site-directed mutagenesis studies (Harlow and Halpert 1997; He et al. 1997; Domanski et al. 1998, 2000; Xue et al. 2001). This autoactivator model provided some ideas of the mechanism behind the unusual kinetic behavior of this enzyme by possibly binding a different site in the protein (Ekins et al. 1998 a). Recently, heteroactivation pharmacophores have been generated for CYP3A4 and CYP2C9 (Egnell et al. 2003, 2004) where in these cases a molecule acts to increase the metabolism of a second different molecule.

The structures of the membrane bound CYPs were unknown until the relatively recent crystallization of the rabbit and human CYP2C forms (Cosme and Johnson 2000; Williams et al. 2000, 2003) and the human CYP3A4 (Williams et al. 2004 b; Yano et al. 2004). The X-ray structures for CYP2C9 and CYP3A4 have to some extent confirmed some of the complex binding characteristics of CYPs seen *in vitro* with CYP3A4 and CYP2B6, which were extensively modeled computationally using pharmacophores. The continued generation of *in vitro* datasets using recombinant CYPs presents the opportunity for further pharmacophore modeling studies to understand binding site features. A recent study used a series of quinidine and quinine analogs as inhibitors of human CYP2D6 (Hutzler et al. 2003) to assess whether the ionic interaction of the basic nitrogen represented the most important feature for binding. We have used these data for 27 molecules with an IC_{50} range 0.01–64.1 μM to build a Catalyst pharmaco-

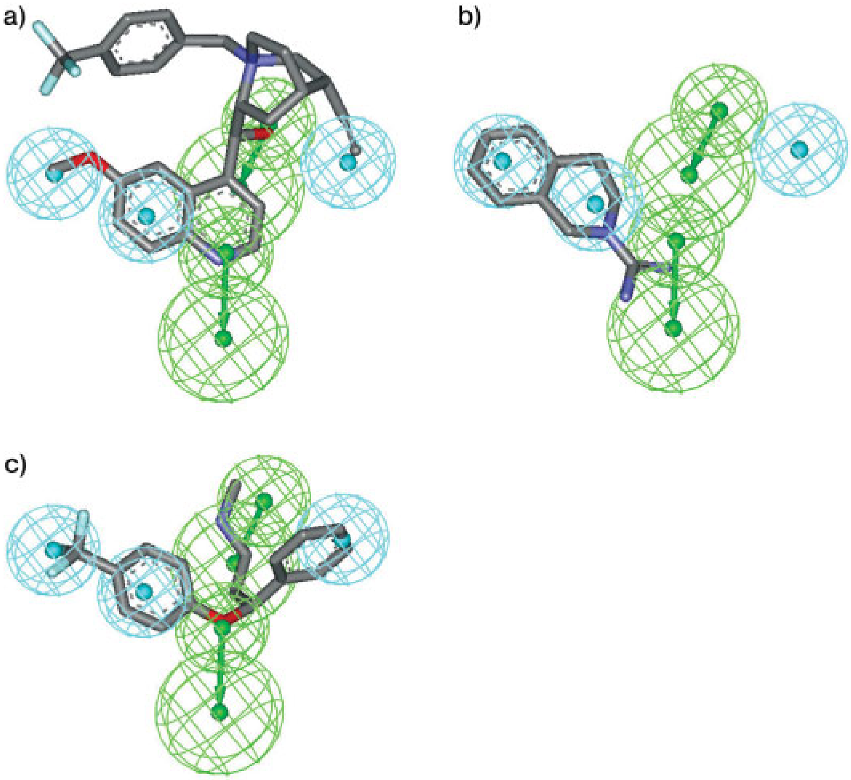


Fig. 14.1 A CYP2D6 inhibitor HypoGen pharmacophore model derived from quinidine and quinine analogs (Hutzler et al., 2003) showing (**a**) the mapping of a training set compound, (**b**) the substrate debrisoquine (Lightfoot et al., 2000) and (**c**) the substrate (Margolis et al., 2000) and inhibitor (Ekins et al., 1999 a) *S*-fluoxetine. The pharmacophore includes hydrogen-bond acceptors (green) and hydrophobes (cyan).

phore for this structurally similar series. Four pharmacophore features are derived from the quinidine backbone (three hydrophobes and one hydrogen-bond acceptor) and only one feature, a hydrogen-bond acceptor, reflects the R group modifications (Fig. 14.1 a). Interestingly no feature is mapped to the substituents on the tertiary quinuclidine nitrogen atom (Fig. 14.1 A).

The model statistics ($r=0.90$; total cost=124.8 was considerably lower than the null hypothesis total cost=170.3) indicate that this may be a useful model for understanding the orientation and mapping of the various analogs. Other known substrates and inhibitors of this enzyme were fitted to the model including debrisoquine (Fig. 14.1 b) and *S*-fluoxetine (Fig. 14.1 c), demonstrating in both cases the mapping to some, but not all, features. There was similarity in this pharmacophore to those derived previously for CYP2D6 inhibitors as the hydrophobic features were dominant, (Ekins et al. 1999 a), although this previous study contained a hydrogen-bond acceptor and hydrogen-bond donor in both pharmacophores presented. A CYP2D6 substrate Catalyst pharmacophore contained a positive ionizable feature, hydrogen-bond acceptor and two hydrophobic features (Snyder et al. 2002). These published pharmacophores may all have some general degree of overlap but they identify different interactions in the binding site that can be influenced and are probably dependent on the existence of a basic nitrogen in the molecule.

14.3 UDP-glucuronosyltransferase

Phase II metabolic processes including the glucuronidation of small lipophilic molecules are important for the clearance of drugs, endobiotics and xenobiotics in all mammalian species (Tukey and Strassburg 2000). A recent study described the glucuronidation of simple 4-substituted phenols by the human recombinant UGT1A6 and UGT1A9 isoenzymes (Ethell et al. 2002). A genetic algorithm using a range of molecular surface and atomic descriptors was used in one of the first attempts to predict the K_m for these isoenzymes, Pharmacophore development in this case was not successful. A different group has used Catalyst pharmacophores for many of these enzymes employing either HipHop or HypoGen with a custom glucuronidation feature. In this way it was possible to derive pharmacophores for UDPGT 1A4 (Smith et al. 2003b), UDPGT 1A1 (Sorich et al. 2002; Smith et al. 2003a) and others (Sorich et al. 2004). At present, the datasets from which the models were constructed are still relatively limited in terms of structural diversity compared with the P450 models, therefore the general applicability of these pharmacophore models may be restricted until more *in vitro* data are available.

14.4 P-glycoprotein (P-gp)

The efflux transporter P-gp is a large ATP-dependent membrane-bound protein expressed at the plasma membrane interface of many organs with their environment. This transporter acts as a barrier, limiting exposure to a diverse range of structurally and functionally unrelated substrates such as paclitaxel. Overexpression of P-gp in malignant cells has also been associated with multidrug resistance (MDR) in some cancers by transporting anticancer drugs such as gleevec out of these cells (Wandell et al. 1999; Ekins and Swaan 2004). P-gp is mainly expressed in the canalicular domain of hepatocytes, brush border of proximal tubule cells and capillary endothelial cells in the central nervous system (CNS). These locations for the transporter expression result in reduced oral drug absorption and enhanced renal and biliary excretion of substrate drugs. Being an obvious target for improved bioavailability of drugs, P-gp has been intensively studied, with many experimental results available in the literature. To account for the observed broad substrate specificities for P-gp, the presence of multiple drug binding sites has also been proposed by many groups (Ayesh et al. 1996; Dey et al. 1997; Scala et al. 1997; Shapiro and Ling 1997; Shapiro et al. 1999).

Computational pharmacophores have been generated to predict the inhibition of P-gp from *in vitro* data for several cell systems, including structurally diverse inhibitors of digoxin transport in Caco-2 cells, vinblastine and calcein accumulation in P-gp expressing LLC-PK1 (L-MDR1) cells and vinblastine binding in vesicles derived from CEM/VLB100 cells (Ekins et al. 2002 c, d). A pharmacophore constructed with 27 inhibitors of digoxin transport by P-gp appears to be one of the most useful models described by the authors and was predictive for molecules known to inhibit digoxin transport and vinblastine binding (Ekins et al. 2002 d). Most of these P-gp pharmacophore models correctly rank-order the data from the other probes, indicating partial overlap for the binding sites probed by digoxin and vinblastine. By merging all P-gp inhibitor pharmacophores, common areas of identical chemical features such as hydrophobes, hydrogen-bond acceptors and ring aromatic features were apparent (Ekins et al. 2002 c). A common features HipHop alignment of the P-gp substrates verapamil and digoxin produced a pharmacophore to which vinblastine partially aligned and that also indicated affinity for a similar or identical binding site(s) within P-gp, overlapping with the P-gp inhibitor pharmacophores (Ekins et al. 2002 c).

Further pharmacophore-based approaches (Pajeva and Wiese 2002) using GASP alignments to vinblastine and to Rhodamine 123 have been carried out to elucidate the verapamil binding site (Pajeva and Wiese 2002). Those studies revealed similar pharmacophore requirements as proposed above. Multiple hydrophobic and hydrogen-bonding interactions were described (Pajeva and Wiese 2002). Superimposition of a small number of P-gp ligands with SYBYL and MOLCAD was undertaken by Garrigues et al. (2002) to generate two pharmacophores. The resulting models were validated with two additional compounds. Based on the transport profile of nine glucocorticoid compounds, further P-gp pharmacophore and QSAR models were generated (Yates et al. 2003). This model is very similar to the P-gp inhibition model based on vinblastine, both having four hydrophobes and two hydrogen-bond acceptors. This P-gp transport model also has two hydrogen-bond donor site features, which could be useful in suggesting interacting amino acids within P-gp. When comparing the P-gp transport model with the previous inhibition model based on digoxin, some interesting observations can be made. Both models comprise hydrophobic and hydrogen-bond acceptor features. Owing to the similarity of glucocorticoid training set compounds, four hydrophobic features representing the steroid ring system were all identified in the transport pharmacophore model. These molecules are a very homologous set of compounds, which is in contrast to the molecules in the inhibition model and this may be illustrated by the multiple hydrogen-bond acceptor features in the transport model whereas the inhibition model has only one. Satisfying the transport model would render a compound susceptible to P-gp, but not fitting in the model does not necessarily exclude the candidate from P-gp transport. It is therefore necessary to filter through different pharmacophore models for P-gp substrate and inhibitor screening to have a more comprehensive assessment of the candidate molecules in question. A recent pharmacophore based search of the Derwent World Drug Index identified 28 P-gp inhibitors of diverse structures (Langer et al. 2004). Our own searches of databases of known P-gp substrates and non-substrates with the previously published catalyst inhibitor pharmacophores suggest that they may also be useful for identifying drugs in larger databases that had not previously been identified as P-gp substrates or inhibitors (Chang et al. 2005 a). Molecular modeling studies therefore have the potential to enhance our understanding of complex transporters such as P-gp for which we currently do not have a crystal structure, and such studies can also be applied more globally to other transporters, as the following sections demonstrate.

14.5
Human Peptide Transporter 1

In an effort to try to predict ionized molecules that are actively transported from the intestine into the circulation, a pharmacophore model of the rat peptide transporter was produced with literature K_m data (Ekins et al. 2000 d). This model suggested that two ionizable groups, a hydrophobe and a hydrogen-bond acceptor, are important features for the transporter. This preliminary work has been expanded to identify novel substrates for the human intestinal small peptide carrier (hPEPT1), which is a proton-coupled, low-affinity, high-capacity oligopeptide transport system with broad substrate specificity known to transport a range of di- and tripeptides, β-lactam antibiotics and angiotensin-converting enzyme inhibitors. The well-characterized and relatively high-affinity ligands Gly-Sar, bestatin and enalapril were used to generate a common features HipHop pharmacophore. This consisted of two hydrophobic features, a hydrogen-bond donor and acceptor and a negative ionizable feature (Ekins et al. 2005 b), representing different features and positioning compared with the rat peptide transporter substrate pharmacophore. This hPEPT1 HipHop pharmacophore was then used to search the CMC database of over 8000 drug-like molecules and retrieved 145 virtual hits mapping to the features. The highest scoring compounds within this set were selected and tested in a stably transfected CHO-hPepT1 cell model. The antidiabetic repaglinide and HMG-CoA reductase inhibitor fluvastatin were found to inhibit hPEPT1 with sub-millimolar potency (Ekins et al. 2005 b) (IC_{50} 178 ± 1.0 and 337 ± 4 μM, respectively). The pharmacophore was also able to identify known hPEPT1 substrates and inhibitors in a further database mining exercise using over 500 commonly prescribed drugs. This further validated the model and demonstrated the potential of combining computational and *in vitro* approaches to determine the affinity of compounds for hPEPT1. These computational and *in vitro* efforts provide some insights into key molecular interactions and have indicated that there may be many more drugs that are substrates for hPEPT1 than previously described.

14.6 Apical Sodium-dependent Bile Acid Transporter (ASBT)

The ASBT is a high-efficacy, high-capacity bile acid transporter located on intestinal epithelial cells. It provides an additional intestinal target for improving drug absorption. In 1970, Lack and colleagues proposed the characteristics of the ASBT binding site as having a hydrophobic pocket with three components: a recognition site for interaction with steroid nucleus, an anionic site for electrostatic interaction with co-transported sodium and a cationic site for interaction with the negatively charged side-chain (Lack et al. 1970). A Catalyst pharmacophore model developed later by Baringhaus and colleagues based on a training set of 17 chemically diverse inhibitors of ASBT gave a more accurate description of the essential features for ASBT affinity, namely one hydrogen-bond donor represented by the 7- or 12-hydroxyl group, one hydrogen-bond acceptor made up by the negatively charged side-chain and three hydrophobic features partially fulfilled by ring D and the 21-methyl group (Baringhaus et al. 1999). Since the 3-hydroxyl does not map to any essential features, the authors suggested substituting 3-hydroxyl groups in poorly absorbed drugs to increase bioavailability. To date we are not aware of any additional pharmacophores for ASBT.

14.7 Sodium Taurocholate-transporting Polypeptide (NTCP)

NTCP is responsible for the transport of bile acids into the liver. Similarly to ASBT in some respects, it is also a potential target for the improvement of absorption of poorly absorbable drugs. A successful case is the improvement of antitumor agent cisplatin. Utilizing NTCP, the novel cisplatin–bile acid the derivatives Bamet-UD2 and Bamet-R2 are more efficiently transported into the liver (Briz et al. 2002) than cisplatin, while toxic accumulation in other tissues is significantly lowered. Since the inhibition of NTCP contributes to cholestasis, Kim et al. (1999) evaluated 33 molecules in NTCP taurocholate uptake inhibition. Based on eight taurocholate uptake inhibitors with IC_{50} values ranging from 1 to 264 μM, a Catalyst pharmacophore model was generated to reveal two hydrophobic features and two hydrogen-bond donor features with a training set correlation of $r=0.97$ (Ekins et al. 2002e). This model has yet to be further tested and expanded with additional molecules, but may provide some in sight into features required for interaction with this transporter.

14.8 Nucleoside Transporters

Nucleoside transporters transport both naturally occurring nucleosides and synthetic nucleoside analogs that are used as anticancer drugs (e.g. cladribine, fludarabine and gemcitabine) and antiviral drugs (e.g. cytarabine and zalcitabine).

These transporters can be classified into two broad categories based on sodium dependence: the sodium-dependent concentrative transporters (CNTs) and sodium-independent equilibrative transporters (ENTs). Different subtypes of each category exist. CNT1 (pyrimidine specific) is localized primarily in intestine and kidney epithelia, whereas CNT2 (purine specific) and CNT3 (broad specificity) have more broad distributions (Gray et al. 2004). ENT1, ENT2 and ENT3 all have wide specificities and are widely distributed in mammalian tissues (Baldwin et al. 2004). CNTs mainly exist in the apical membrane, whereas ENTs are located in the basolateral membrane of epithelial cells, suggesting coordination between CNTs and ENTs in transepithelial nucleoside transport (Casado et al. 2002; Baldwin et al. 2004). A recent study has compared the relative and unified structural requirements of nucleosides for high-affinity interaction with CNT1, CNT2 and ENT1 (Chang et al. 2004). Unique pharmacophore models were developed for each transporter based on inhibition data from a series of uridine and adenosine analogs, including a variety of anticancer and antiviral drugs (Wang and Giacomini 1999; Patil et al. 2000; Ekins and Swaan 2004). Besides obvious similarities (two hydrophobic centers and one hydrogen-bond acceptor on the pentose ring) among the three individual transport pharmacophores, subtle differences set the individual transporters apart. hENT1 inhibitors require the presence of both a hydrogen-bond acceptor and donor feature near 3′-C while CNT1 and CNT2 inhibitors need hydrogen-bond acceptors on 3′-OH and the 2-position of the pyrimidine ring, CNT2 inhibitors have an extra requirement of a hydrogen-bond acceptor on 5′-OH. With more pharmacophore requirements, CNT2 is the most selective in inhibition while ENT1 has the broadest inhibitor specificity owing to its simple pharmacophore features. These models illustrated the common feature requirements for nucleoside transporter inhibition and identified distinctive feature requirements for each transporter subtype. This therefore represents a model for future design of high-affinity nucleoside analog anticancer and antiviral drugs.

14.9
Organic Cation Transporter 1 and 2

The OCTs influence the plasma concentration of many cationic drugs. They are typically found in barrier epithelia, including the kidney, liver and intestine, where they influence drug bioavailability, excretion and toxicity. To date, three subtypes of polyspecific cation transporters named OCT1, OCT2 and OCT3 have been cloned from different species (Grundemann et al. 1994, 1997; Gorboulev et al. 1997). The human organic cation transporter hOCT1 is therefore important in the elimination of many cationic drugs (Koepsell 1998). By assessing the extent of inhibition *in vitro* of [^{3}H]TEA uptake in HeLa cells stably expressing hOCT1 using 22 diverse molecules, a pharmacophore was produced consisting of three hydrophobic features and a positive ionizable feature (Bednarczyk et al. 2003). For a small series of eight phenylpyridinium and quinolinium analogs the correlation for observed and predicted IC_{50} values was low and the predicted values were within the range of similar compounds in the training set. A more recent study illustrated the binding requirement of both human OCT2 and rabbit OCT2 (Suhre et al. 2005) and used molecules that discriminate between the orthologs to probe the qualitative differences between molecules with high affinity to the transporters with a Catalyst HipHop alignment. An alignment of the selective inhibitors for both transporters indicated distinctive differences for recognition, manifested in features and angles recognized for each transporter. Even though the features on these pharmacophores were similar, the approach was able to identify a difference in the orientation of the hydrogen bonding features ($>37°$). This could infer variability in the disposition of critical amino acids for interaction with inhibitors within both transporters and these models may be useful alongside experimental data when deriving protein-based models for these transporters.

14.10 Organic Anion-transporting Polypeptides (OATPs)

The OATPs are key membrane-bound transporters expressed in many organs including intestine, liver, lung, choroid plexus, blood–brain barrier and other organs (Tamai et al. 2000). This family of transporters is capable of mediating sodium-independent transport of a diverse array of molecules that are mostly anions, in addition to organic cations, steroid conjugates, organic anions and xenobiotics (Bossuyt et al. 1996a; Hagenbuch and Meier 2004) by coupling uptake with the efflux of bicarbonate (Satlin et al. 1997) or glutathione (Li et al. 1998). The OATPs share some substrate overlapping specificity with other promiscuous efflux transporters such as P-gp and MRP2, indicative perhaps of some degree of coordination. The involvement of OATPs in the hepatic uptake of drugs implies a potential for drug–drug interactions (Kim 2003), as exemplified by the interaction between cerivastatin and cyclosporin A (Shitara et al. 2003) and also cerivastatin, gemfibrozil and its glucuronide metabolite (Shitara et al. 2004). Thirty-six mammalian OATPs have been identified, but only a few of these have been characterized in any detail. An alignment of 18 inhibitors of the rat Oatp1a5 using GASP identified a hydrogen-bond donor and negatively charged regions at opposite ends of a planar hydrophobic region (Yarim et al. 2005). Currently, 11 human OATPs have been identified and only recently have pharmacophore models been generated for human OATP1B1 (Chang et al. 2005b) and rat Oatp1a1. These pharmacophore models were validated using external test sets of compounds. All of these models comprised hydrophobes and hydrogen-bond acceptors. The pharmacophores for these transporters are differentiated by the exact number and position of these pharmacophore features.

Using the substrates described in the literature (Table 14.1), sufficient data were also available with oocytes expressing OATP1A2 to enable a preliminary pharmacophore for this transporter. The OATP1A2-oocyte data set consisted of

Table 14.1 Literature K_m data for human OATP1A2 derived from oocytes expressing this transporter

Molecule	K_m (µM)	Reference
BametR2	23.8	Briz et al. (2002)
Bamet ud2	14.1	Briz et al. (2002)
BSP	20	Kullak-Ublick et al. (1995)
Cholate	93	Kullak-Ublick et al. (1995)
DHEAS	6.6	Kullak-Ublick et al. (1998)
Deltorphin II	330	Gao et al. (2000)
DPDPE	202	Gao et al. (2000)
Estrone-3-sulfate	59	Bossuyt et al. (1996 b)
Ouabain	5500	Bossuyt et al. (1996 b)
N-Methylquinine	5.1	van Montfoort et al. (1999)
N-Methylquinidine	25.6	van Montfoort et al. (1999)
Taurocholate	60	Kullak-Ublick et al. (1995)
Tauroursodeoxycholate	19	Kullak-Ublick et al. (1995)
Thyroxine	3	Fujiwara et al. (2001)
Triidothyronine	2.7	Fujiwara et al. (2001)

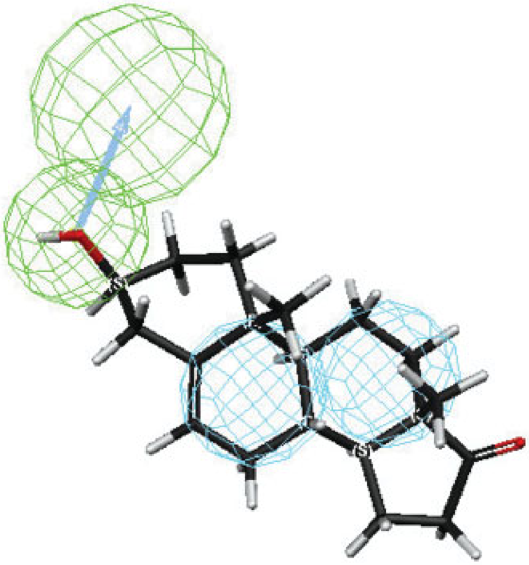


Fig. 14.2 DHEAS mapping to the OATP1A2 pharmacophore derived from data in Table 14.1. The pharmacophore includes hydrogen-bond acceptors (green) and hydrophobes (cyan).

Table 14.2 Summary of the molecules tested with human OATP1A2 and rat Oatp1a1 [rat data summarized from the literature previously (Chang et al., 2005 b)]

Molecule	Mean human K_m (µM)	Mean rat K_m (µM)	Log mean human K_m (µM)	Log mean rat K_m (µM)
BSP	20	2.52	1.30	0.40
Cholate	93	54	1.97	1.73
DHEAS	6.60	5.00	0.82	0.70
Estrone-3-sulfate	59.00	8.25	1.77	0.92
Fexofenadine	6.40	32	0.81	1.51
Taurocholate	60	38.8	1.78	1.59
Ouabain	5500	2350	3.74	3.37
Deltorphin	330	137	2.52	2.14
Tauroursodeoxycholate	19	13	1.28	1.11
DPDPE	202	48	2.31	1.68

15 molecules (K_m range 2.7–5500 µM), resulting in a Catalyst pharmacophore containing one hydrogen-bond acceptor and two hydrophobes (Fig. 14.2, correlation for training set $r = 0.89$). With only two hydrophobic features and one hydrogen-bond acceptor feature, human OATP1A2 (Fig. 14.2) appears non-selective compared with the other two pharmacophores generated to date (Chang et al. 2005 b). However, there is a good correlation between the *in vitro* data for the same 10 substrates shared with rat Oatp1a1 (Table 14.2, $r^2 = 0.74$), which is slightly higher than for eight substrates shared between human OATP1B1 and rat Oatp1a1 ($r^2 = 0.64$) (Chang et al. 2005 b). This suggests some degree of overlap but not identity between the substrate specificity for these three rat and human transporters. The OATP pharmacophore models could therefore help provide future insight into possible drug–drug interactions with these transporters and will be validated with additional compounds in the future.

14.11 Breast Cancer Resistance Protein (BRCP)

The BCRP is an ABC transporter similar to P-gp whose expression results in resistance to anticancer therapeutics and may limit intestinal absorption of drugs. However, there have been limited studies to elucidate the selectivities of drugs for P-gp and BCRP (Brooks et al. 2004). We used a published dataset of seven topoisomerase inhibitors (Maliepaard et al. 2001) to construct a HipHop model for BCRP. We then mapped the potent tyrosine kinase inhibitor Gleevec to this pharmacophore as this compound has been suggested experimentally in conflicting studies as both a substrate and inhibitor of BCRP (Burger et al. 2004; Houghton et al. 2004).

Gleevec clearly maps to the hydrophobic and hydrogen-bond donor features in this pharmacophore (Fig. 14.3 a). We also used the published P-gp substrate pharmacophore to analyze qualitatively whether gleevec maps well to these pharmacophore features also (Fig. 14.3 b). From this mapping, it seems that gleevec fits well to this pharmacophore and also that for BCRP. Recently, several studies have suggested that Gleevec binds P-gp with a K_i value of 18.3 μM, using a calcein–AM efflux assay and indicated that it is both a substrate and a modulator of human P-gp, suggestive of possible drug interactions via P-gp (Hamada et al. 2003; Mahon et al. 2003). It is important, therefore, to design future potent tyrosine kinase inhibitors without the propensity to bind to either of these transporters and pharmacophore models may be useful in this process.

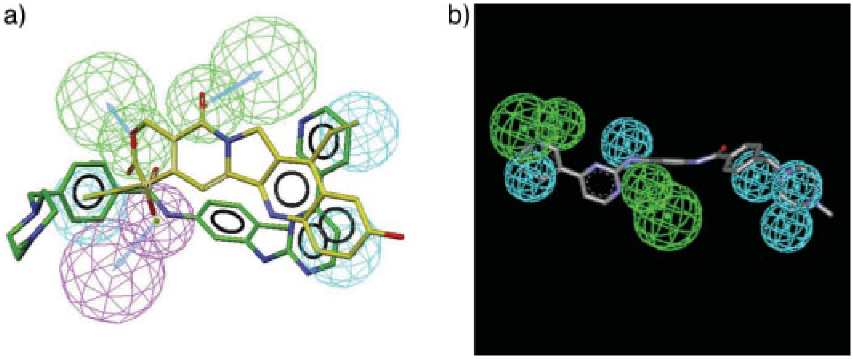
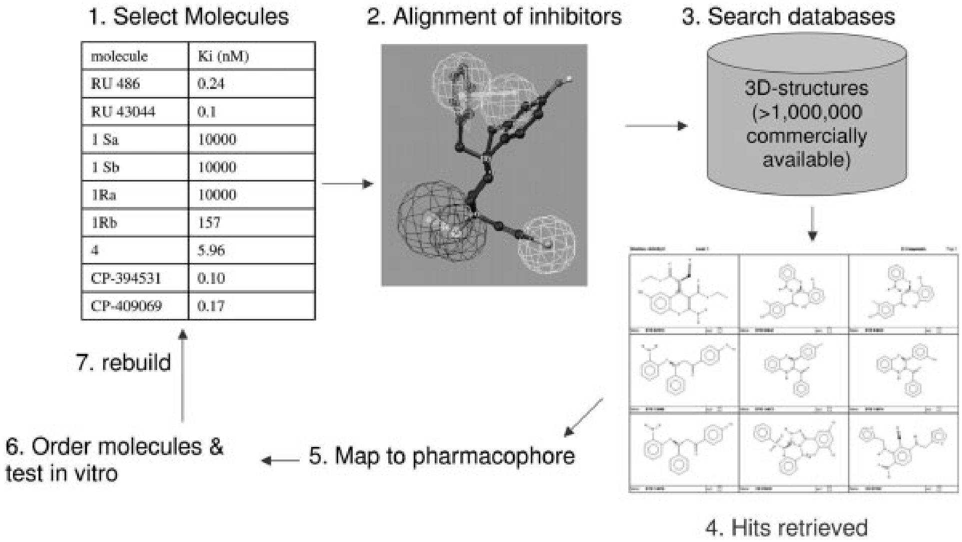


Fig. 14.3 **A**. A preliminary HipHop pharmacophore for the breast cancer resistance protein derived for seven inhibitors (Maliepaard et al., 2001) showing SN-38 (yellow) and gleevec (green). **B**. Gleevec mapped to the previously published HipHop pharmacophore for P-gp substrates (Ekins et al., 2002c). The pharmacophores include hydrogen-bond acceptors (green), hydrogen-bond donors (purple) and hydrophobes (cyan).

14.12 The Nuclear Hormone Receptors

The regulation of metabolic and transport proteins occurs via complex nuclear hormone receptor mediated pathways (Ekins et al. 2002 e) among the pregnane X-receptor (PXR), constitutive androstane receptor (CAR), glucocorticoid receptor (GR), aryl hydrocarbon receptor (AHR) and probably many other receptors. Some of these receptors have been modeled computationally with pharmacophores (Ekins and Erickson 2002; Ekins and Schuetz 2002; Ekins et al. 2002 e; Mankowski and Ekins 2003, Schuster and Langer 2005) and they may bind many pharmaceutical and environmentally relevant molecules. PXR is a transcriptional regulator of CYP3A4 (Bertilsson et al. 1998; Blumberg et al. 1998; Kliewer et al. 1998), CYP2C9, CYP2B6, P-gp and numerous other proteins and is activated by structurally diverse molecules. *In vitro* EC_{50} data for 12 molecules were used to generate a PXR pharmacophore model that defined key features of



molecule	Ki (nM)
RU 486	0.24
RU 43044	0.1
1 Sa	10000
1 Sb	10000
1Ra	10000
1Rb	157
4	5.96
CP-394531	0.10
CP-409069	0.17

1. Select Molecules
2. Alignment of inhibitors
3. Search databases
3D-structures (>1,000,000 commercially available)
4. Hits retrieved
5. Map to pharmacophore
6. Order molecules & test in vitro
7. rebuild

Fig. 14.4 Pharmacophore-based database searching for drug discovery. This example is based on a glucocorticoid receptor dataset (Morgan et al., 2002).

ligands binding to PXR (Ekins and Erickson 2002). It implicated at least four hydrophobic features and a hydrogen-bonding feature that should be avoided in future drug candidate molecules. The pharmacophore with the ligand SR12813 was fitted in the human PXR ligand binding site and compared with the orientations of the crystallized molecule once the X-ray structure was available. The pharmacophore was further tested by predicting 28 PXR ligands which had available *in vitro* data. This pharmacophore has since been used to analyze the binding of imidazole analogs to PXR, which result in increased apoA1 and HDL-C in rats and mice (Bachmann et al. 2004). A second orphan nuclear receptor, CAR, has approximately 40% identity with PXR in the ligand-binding domain and also regulates CYP2B6, CYP3A4 and other proteins. An alignment of clotrimazole, androstanol and 5β-pregnane-3,20-dione yielded a pharmacophore for human CAR, with three hydrophobic features and one hydrogen-bond acceptor (Ekins et al. 2002a, e). This planar model indicated that CAR is a less promiscuous receptor than PXR because it accommodates less flexibility in the ligands binding to it. The GR has also been implicated in the induction of CYP3A4 (Pascussi et al. 2000a, b, 2001; El-Sankary et al. 2002; Usui et al. 2003). A paper by Morgan et al. (2002) on the discovery of non-steroidal human GR antagonists provided a nine-molecule data set from which a pharmacophore model was constructed with an observed versus predicted correlation of 0.94 for the training set (Mankowski and Ekins 2003) (Fig. 14.4). This pharmacophore may be useful for predicting potential non-steroidal GR ligands as CYP3A4 inducers or for searching databases for other non-steroidal GR ligands. Computational models have also been generated for the AHR which regulates CYP1A1, CYP1B1 and several phase II enzymes. These models seem to confirm the key importance of planarity, molecular length, along with other parameters such as the frontier orbital HOMO and lipophilicity (Lewis et al. 2002). A pharmacophore derived with four nanomolar ligands, indirubin, indigo, ITE and TCDD (Adachi et al. 2001; Song et al. 2002), suggested at least two possible planar alignments, one of which included a hydrogen-bond acceptor (Mankowski and Ekins 2003). Both of these pharmacophores for the AHR possessed multiple key hydrophobic features. To date, the number of molecules available for modeling these receptors has been limited and this has restricted the size of the training sets and hence the predictive capability.

14.13 Human Ether-a-go-go Related Gene

Undesirable drug interactions may also occur via binding to ion channel proteins such as the hERG potassium channel. Cisapride, terfenadine, astemizole, sertindole and grepafloxacin were all drugs withdrawn from the market owing to cardiovascular toxicity associated with alteration of the action potential via this channel. It is understood that drugs or their metabolites may block this channel, thereby prolonging the QT interval (the period between the start of ventricular depolarization and repolarization) and in some cases this leads to the potentially life-threatening ventricular arrhythmia. Utilizing *in vitro* IC_{50} data generated with cDNA-expressed hERG channels, the first published computational pharmacophore model was derived with 15 molecules (Ekins et al. 2002b). This pharmacophore contained four hydrophobes and one positive ionizable feature (Ekins et al. 2002b) which was in agreement with a published homology model, as the hydrophobic features coincide with the F656 and Y652 residues thought to be involved in π–π stacking with aromatic residues of hERG inhibitors (Mitcheson et al. 2000). The hERG pharmacophore model produced predictions for a 22-molecule test set with a correlation r^2 of 0.83. Other sets of ligands for different therapeutic targets (5-HT_{2A} receptor anatagonists and phosphodiesterase-4 inhibitors) possess similar pharmacophores for their hERG inhibition capability (Ekins 2004), which is apparent when the models are aligned (Ekins and Swaan 2004). A second group published a pharmacophore using a larger data set of 31 inhibitors (Cavalli et al. 2002) and defined three aromatic hydrophobic features and a central nitrogen. A third group using 28 molecules including sertindole analogs produced a model which showed similar key features (Pearlstein et al. 2003). All of these pharmacophores have similar molecular features but with differences in their exact positioning, which might suggest either some flexibility in this channel or multiple binding sites at different points which is in common with other promiscuous proteins (Ekins 2004). Ultimately, these models have been useful for describing interactions in the channel and enabling predictions for experimental verification.

14.14 Conclusion

The computational pharmacophore work discussed here has considered several targets implicated in ADME/Tox: the major P450 enzymes, UDP-glucuronosyl transferase, various transporters, the nuclear hormone receptors and the potassium channel hERG. There are many other human proteins that are also relevant to ADME/Tox where additional pharmacophore modeling studies might be useful. These include but are not limited to the organic anion transporter, vitamin transporter, multidrug resistance protein, flavin-containing monooxygenase, epoxide hydrolase, sulfotransferase and glutathione *S*-transferase. In some of these cases other types of QSAR models or homology models have been applied but pharmacophores may also be useful.

Computational pharmacophore approaches have been used to describe the features that ligands possess that ultimately relate to key interactions for recognition within the binding sites of these proteins. The computational models suggest that in addition to key hydrogen-bonding features present in most models, there are multiple hydrophobic interactions shared by all of the proteins involved in undesirable drug interactions (Ekins 2004). This may go some way to explain the general promiscuity of these proteins, which relates to their overall protective ability against a wide array of molecule structures. Multiple pharmacophores may also be required for each protein to predict affinity adequately for different classes of molecules. These models may also be merged to show qualitative commonalities between different datasets. Data for the same protein generated in different *in vitro* systems can also be successfully combined to result in what we have termed a meta-pharmacophore approach (Chang et al. 2005 b). Ultimately these pharmacophores could certainly be used to rapidly search databases to identify and remove undesirable molecules or suggest molecules that could be used as novel experimental probes for the protein in question. To date there have been few instances where such ADME/Tox-related pharmacophores have been used for database searches (Langer et al. 2004; Ekins et al. 2005 b) (Fig. 14.4) and they represent a cherry-picking approach to filter compound vendor databases further.

These ADME/Tox pharmacophore models can also be combined in a multidimensional approach to assess drug interactions and possible toxicity alongside other computational models for the therapeutic target (Ekins et al. 2002 a; Shimada et al. 2002; Young et al. 2002; Ekins 2003). This combined modeling approach could be used as a decision criterion for molecule selection (Ekins et al. 2004). The integration of the models described above with computational technologies such as *de novo* growth, docking algorithms and virtual library screening will assist in improving the rate of discovery of bioactive molecules (Shimada et al. 2002) with optimal biopharmaceutical properties. Actively combining the insights from *in vitro* and computational models will result in a holistic or systems biology-based understanding of the many proteins and their interactions in regulation, transport and metabolic pathways (Ekins et al. 2002 a, 2005 a, c, d). We suggest therefore that the application of computational methods such as pharmacophores alongside experimental methods may go some way to improving the efficiency of early ADME/Tox molecule profiling and also later stage lead optimization. This will hopefully lead to an improved success rate for drug discovery and provide a useful method to eliminate late stage failures due to predictable drug interactions that would normally be identified during clinical trials.

The computational pharmacophore approach sets educational challenges as drug discovery scientists should be trained to use actively a combination of *in vitro* and computational models for multiple proteins (Ekins and Swaan 2004). We have seen some progress over the last decade in the use of computational approaches in ADME/Tox and pharmacophores in particular have been broadly applied. In the future, if these models are to be more widely applied they will need to become more accessible to other scientists outside computational chemistry groups. At present there are few if any freely available pharmacophore technologies and perhaps we will see academic groups challenge the *status quo* that has dominated this field for over a decade. One could imagine that standards for open source pharmacophore model generation, sharing and application would greatly facilitate even broader application of pharmacophore models in this and other fields. We look forward to such future developments for expanding the scope of pharmacophore applications.

Acknowledgments

We gratefully acknowledge the contributions and encouragement of our many colleagues and collaborators in modeling these proteins recently, including Dr Peter W. Swaan (University of Maryland), Dr Stephen H. Wright (University of Arizona), Dr K. Sandy Pang (University of Toronto), Dr Jim Dalton (Ohio State University) and Dr Jash Unadkat (University of Washington). We are also grateful for the considerable efforts of Mr John Ohrn (Accelrys, San Diego, CA) in making Catalyst available to us. This chapter is dedicated to Mr James H. Wikel on the occasion of his retirement and who initially actively supported S.E. in using pharmacophores for ADME/Tox research.

References

Adachi J, Mori Y, Matsui S, Takigami H, Fujino J, Kitagawa H, Miller CAI, Kato T, Saeki K, Matsuda T (**2001**) Indirubin and indigo are potent aryl hydrocarbon receptor ligands in human urine. *J Biol Chem 276*, 31475–31478.

Afzelius L, Zamora I, Ridderstrom M, Andersson TB, Karlen A, Masimirembwa CM (**2001**) Competitive CYP2C9 inhibitors: enzyme inhibition studies, protein homology modeling and three dimensional quantitative structure activity relationship analysis. *Mol Pharmacol 59*, 909–919.

Ayesh S, Shao Y-M, Stein WD (**1996**) Co-operative, competitive and non-competitive interactions between modulators of P-glycoprotein. *Biochim Biophys Acta 1316*, 8–18.

Bachmann K, Patel H, Batayneh Z, Slama J, White D, Posey J, Ekins S, Gold D, Sambucetti L (**2004**) PXR and the regulation of apoA1 and HDL-cholesterol in rodents. *Pharmacol Res 50*, 237–246.

Baldwin SA, Beal PR, Yao SY, King AE, Cass CE, Young JD (**2004**) The equilibrative nucleoside transporter family, SLC29. *Pflugers Arch 447*, 735–743.

Baringhaus KH, Matter H, Stengelin S, Kramer W (**1999**) Substrate specificity of the ileal and the hepatic Na(+)/bile acid cotransporters of the rabbit. II. A reliable 3D QSAR pharmacophore model for the ileal Na(+)/bile acid cotransporter. *J Lipid Res 40*, 2158–2168.

Barratt MD, Rodford RA (**2001**) The computational prediction of toxicity. *Curr Opin Chem Biol 5*, 383–388.

Bednarczyk D, Ekins S, Wikel JH, Wright SH (**2003**) Influence of molecular structure of substrate binding to the human organic cation transporter, hOCT1. *Mol Pharmacol 63*, 489–498.

Bertilsson G, Heidrich J, Svensson K, Asman M, Jendeberg L, Sydow-Backman M, Ohlsson R, Postlind H, Blomquist P, Berkenstam A (**1998**) Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc Natl Acad Sci USA 95*, 12208–12213.

Blumberg B, Sabbagh W, Jr, Juguilon H, Bolado J, Jr, van Meter CM, Ong ES, Evans RM (**1998**) SXR, a novel steroid and xenobiotic-sensing nuclear receptor. *Genes Dev 12*, 3195–3205.

Boobis A, Gundert-Remy U, Kremers P, Macheras P, Pelkonen O (**2002**) *In silico* prediction of ADME and pharmacokinetics. Report of an expert meeting organised by COST B15. *Eur J Pharm Sci 17*, 183–193.

Bossuyt X, Muller M, Hagenbuch B, Meier PJ (**1996a**) Polyspecific drug and steroid clearance by an organic anion transporter of mammalian liver. *J Pharmacol Exp Ther 276*, 891–896.

Bossuyt X, Muller M, Meier PJ (**1996b**) Multispecific amphipathic substrate transport by an organic anion transporter of human liver. *J Hepatol 25*, 733–738.

Briz O, Serrano MA, Rebollo N, Hagenbuch B, Meier PJ, Koepsell H, Marin JJ (**2002**) Carriers involved in targeting the cytostatic bile acid–cisplatin derivatives *cis*-diammine-chloro-cholylglycinate-platinum(II) and *cis*-diammine-bisursodeoxycholate-platinum(II) toward liver cells. *Mol Pharmacol 61*, 853–860.

Brooks TA, Kennedy DR, Gruol DJ, Ojima I, Baer MR, Bernacki RJ (**2004**) Structure–activity analysis of taxane-based broad-spectrum multidrug resistance modulators. *Anticancer Res 24*, 409–415.

Burger H, van Tol H, Boersma AW, Brok M, Wiemer EA, Stoter G, Nooter K (**2004**) Imatinib mesylate (STI571) is a substrate for the breast cancer resistance protein (BCRP)/ABCG2 drug pump. *Blood 104*, 2940–2942.

Butina D, Segall MD, Frankcombe K (**2002**) Predicting ADME properties *in silico*: methods and models. *Drug Discov Today 7*, S83–S88.

Casado FJ, Lostao MP, Aymerich I, Larrayoz IM, Duflot S, Rodriguez-Mulero S, Pastor-Anglada M (**2002**) Nucleoside transporters in absorptive epithelia. *J Physiol Biochem 58*, 207–216.

Cavalli A, Poluzzi E, De Ponti F, Recanatini M (**2002**) Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K^+ channel blockers. *J Med Chem 45*, 3844–3853.

Chang C, Swaan PW, Ngo LY, Lum PY, Patil SD, Unadkat JD (**2004**) Molecular requirements of the human nucleoside transporters hCNT1, hCNT2 and hENT1. *Mol Pharmacol 65*, 558–570.

Chang C, Ekins S, Swaan P (**2005**) Application of P-gp pharmacophore models in database screening. In: *229th ACS National Meeting*, San Diego, p. 509.

Chang C, Pang KS, Swaan PW, Ekins S (**2005**) Comparative pharmacophore modeling of organic anion transporting polypeptides:a meta-analysis of rat Oatp1a1 and OATP1B1. *J Pharmacol Exp Ther 31*, 533–541.

Clark DE (**1999**) Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J Pharm Sci 88*, 807–814.

Cosme J, Johnson EF (**2000**) Engineering microsomal cytochrome P450 2C5 to be a soluble, monomeric enzyme. Mutations that alter aggregation, phospholipid dependence of catalysis and membrane binding. *J Biol Chem 275*, 2545–2553.

de Groot MJ, Ekins S (**2002**) Pharmacophore modeling of cytochromes P450. *Adv Drug Del Rev 54*, 367–383.

de Groot MJ, Ackland MJ, Horne VA, Alex AA, Jones BC (**1999a**) A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed *N*-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J Med Chem 42*, 4062–4070.

de Groot MJ, Ackland MJ, Horne VA, Alex AA, Jones BC (**1999b**) Novel approach to predicting P450-mediated drug metabolism: development of a combined protein and pharmacophore model for CYP2D6. *J Med Chem 42*, 1515–1524.

Dey S, Ramachandra M, Pastan I, Gottesman MM, Ambudkar S (**1997**) Evidence for two nonidentical drug-interaction sites in the human P-glycoprotein. *Proc Natl Acad Sci USA 94*, 10594–10599.

Domanski TL, Liu J, Harlow GR, Halpert JR (**1998**) Analysis of four residues within substrate recognition site 4 of human cytochrome P450 3A4: role in steroid hydroxylase activity and *α*-napthoflavone stimulation. *Arch Biochem Biophys 350*, 223–232.

Domanski TL, He Y-A, Harlow GR, Halpert JR (**2000**) Dual role of human cytochrome P450 3A4 residue Phe-304 in substrate specificity and cooperativity. *J Pharmacol Exp Ther 293*, 585–591.

Egan WJ, Merz KMJ, Baldwin JJ (**2000**) Prediction of drug absorption using multivariate statistics. *J Med Chem 43*, 3867–3877.

Egnell AC, Eriksson C, Albertson N, Houston B, Boyer S (**2003**) Generation and evaluation of a CYP2C9 heteroactivation pharmacophore. *J Pharmacol Exp Ther 307*, 878–887.

Egnell AC, Houston JB, Boyer CS (**2005**) Predictive models of CYP3A4 heteroactivation: *in vitro–in vivo* scaling and pharmacophore modeling. *J Pharmacol Exp Ther 312*, 926–937.

Ekins S (**2003**) *In silico* approaches to predicting metabolism, toxicology and beyond. *Biochem Soc Trans 31*, 611–614.

Ekins S (**2004**) Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discov Today 9*, 276–285.

Ekins S, Erickson JA (**2002**) A pharmacophore for human pregnane-X-receptor ligands. *Drug Metab Dispos 30*, 96–99.

Ekins S, Rose JP (**2002**) *In silico* ADME/TOX: the state of the art. *J Mol Graph 20*, 305–309.

Ekins S, Schuetz E (**2002**) The PXR crystal structure: the end of the beginning. *Trends Pharmacol Sci. 23*, 49–50.

Ekins S, Swaan PW (**2004**) Computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. *Rev Comput Chem 20*, 333–415.

Ekins S, Wrighton SA (**1999**) The role of CYP2B6 in human xenobiotic metabolism. *Drug Metab Rev 31*, 719–754.

Ekins S, Wrighton SA (**2001**) Application of *in silico* approaches to predicting drug–drug interactions: a commentary. *J Pharmacol Toxicol Methods 44*, 1–5.

Ekins S, VandenBranden M, Ring BJ, Wrighton SA (**1997**) Examination of purported probes of human CYP2B6. *Pharmacogenetics 7*, 165–179.

Ekins S, Ring BJ, Binkley SN, Hall SD, Wrighton SA (**1998a**) Autoactivation and activation of cytochrome P450s. *Int J Clin Pharmacol Ther 36*, 642–651.

Ekins S, VandenBranden M, Ring BJ, Gillespie JS, Yang TJ, Gelboin HV, Wrighton SA (**1998c**) Further characterization of the expression and catalytic activity of human CYP2B6. *J Pharmacol Exp Ther 286*, 1253–1259.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, Wrighton SA (**1999a**) Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics 9*, 477–489.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, Wrighton SA (**1999b**) Three and four dimensional-quantitative structure activity relationship analyses of CYP3A4 inhibitors. *J Pharmacol Exp Ther 290*, 429–438.

Ekins S, Bravi G, Ring BJ, Gillespie TA, Gillespie JS, VandenBranden M, Wrighton SA, Wikel JH (**1999c**) Three dimensional-quantitative structure activity relationship

(3D-QSAR) analyses of substrates for CYP2B6. *J Pharmacol Exp Ther 288*, 21–29.

Ekins S, Bravi G, Wikel JH, Wrighton SA (**1999 d**) Three dimensional quantitative structure activty relationship (3D-QSAR) analysis of CYP3A4 substrates. *J Pharmacol Exp Ther Exp Thera 291*, 424–433.

Ekins S, Maenpaa J, Wrighton SA (**1999 e**) *In vitro* metabolism: subcellular fractions. In *Handbook of Drug Metabolism*, Woolf TF (ed.), Marcel Dekker, New York, pp. 363–399.

Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, Wrighton SA (**2000 a**) Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab Dispos 28*, 994–1002.

Ekins S, Ring BJ, Bravi G, Wikel JH, Wrighton SA (**2000 b**) Predicting drug–drug interactions *in silico* using pharmacophores: a paradigm for the next millennium. In *Pharmacophore Perception, Development and Use in Drug Design*, Guner OF (ed.), IUL, San Diego, CA, pp. 269–299.

Ekins S, Ring BJ, Grace J, McRobie-Belle DJ, Wrighton SA (**2000 c**) Present and future *in vitro* approaches for drug metabolism. *J Pharmacol Toxicol Methods 44*, 313–324.

Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH (**2000 d**) Progress in predicting human ADME parameters *in silico*. *J Pharmacol Toxicol Methods 44*, 251–272.

Ekins S, de Groot M, Jones JP (**2001 a**) Pharmacophore and three dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab Dispos 29*, 936–944.

Ekins S, Durst GL, Stratford RE, Thorner DA, Lewis R, Loncharich RJ, Wikel JH (**2001 b**) Three dimensional quantitative structure permeability relationship analysis for a series of inhibitors of rhinovirus replication. *J Chem Inf Comput Sci 41*, 1578–1586.

Ekins S, Boulanger B, Swaan PW, Hupcey MAZ (**2002 a**) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J Comput Aided Mol Des 16*, 381–401.

Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA (**2002 b**) Three dimensional quantitative structure activity relationship for the inhibition of the hERG (human ether-a-go-go related gene) potassium channel. *J Pharmacol Exp Ther 301*, 427–434.

Ekins S, Kim RB, Leake BF, Dantzig AH, Schuetz E, Lan LB, Yasuda K, Shepard RL, Winter MA, Schuetz JD, Wikel JH, Wrighton SA (**2002 c**) Application of three dimensional quantitative structure–activity relationships of P-glycoprotein inhibitors and substrates. *Mol Pharmacol 61*, 974–981.

Ekins S, Kirillov E, Rakhmatulin EA, Nikolskaya T (**2005 c**) A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism. *Drug Metab Dispos 33*, 474–481.

Ekins S, Kim RB, Leake BF, Dantzig AH, Schuetz E, Lan LB, Yasuda K, Shepard RL, Winter MA, Schuetz JD, Wikel JH, Wrighton SA (**2002 d**) Three dimensional quantitative structure–activity relationships of inhibitors of P-glycoprotein. *Mol Pharmacol 61*, 964–973.

Ekins S, Mirny L, Schuetz EG (**2002 e**) A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRa and LXRb. *Pharmacol Res 19*, 1788–1800.

Ekins S, Berbaum J, Harrison RK (**2003 a**) Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab Dispos 31*, 1077–1080.

Ekins S, Stresser DM, Williams JA (**2003 b**) *In vitro* and pharmacophore insights into CYP3A enzymes. *Trends Pharmacol Sci 24*, 191–196.

Ekins S, Berbaum J, Harrison RK, Zecher M, Yuan J, Ischchenko AV, Berezin K, Chubukov V, Lawson D, Hupcey MAZ (**2004**) Applying computational and in vitro approaches to lead selection. In *Workshop on Pharmaceutical Profiling in Drug Discovery for Lead Selection*, Borchardt RT, Kerns EM, Lipinski CA, Thakker DR, Wang B. (eds.), AAPS, Arlington, VA, pp 361–389.

Ekins S, Andreyev S, Ryabov A, Kirilov E, Rakhmatulin EA, Bugrim A, Nikolskaya T (**2005 a**) Computational prediction of human drug metabolism. *Expert Opin Drug Metab Toxicol 1*, 303–323.

Ekins S, Johnston JS, Bahadduri P, D'Souzza VM, Ray A, Chang C, Swaan PW

(**2005b**) *In vitro* and pharmacophore based discovery of novel hPEPT1 inhibitors. *Pharm Res 22*, 512–517.

Ekins S, Nikolsky Y, Nikolskaya T (**2005d**) Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol Sci 26*, 202–209.

El-Sankary W, Bombail V, Gibson GG, Plant N (**2002**) Glucocorticoid-mediated induction of CYP3A4 is decreased by disruption of a protein: DNA interaction distinct from the pregnane X receptor response element. *Drug Metab Dispos 30*, 1029–1034.

Ethell BT, Ekins S, Wang J, Burchell B (**2002**) Quantitative structure activity relationships for the glucuronidation of simple phenols by expressed human UGT1A6 and UGT1A9. *Drug Metab Dispos 30*, 734–738.

Fuhr U, Strobl G, Manaut F, Anders E-M, Sorgel F, Lopez-de-Brinas E, Chu DTW, Pernet AG, Mahr G, Sanz F, Staib AH (**1993**) Quinolone antibacterial agents: relationship between structure and *in vitro* inhibition of human cytochrome P450 isoform CYP1A2. *Mol Pharmacol 43*, 191–199.

Fujiwara K, Adachi H, Nishio T, Unno M, Tokui T, Okabe M, Onogawa T, Suzuki T, Asano N, Tanemoto M, Seki M, Shiiba K, Suzuki M, Kondo Y, Nunoki K, Shimosegawa T, Iinuma K, Ito S, Matsuno S, Abe T (**2001**) Identification of thyroid hormone transporters in humans: different molecules are involved in a tissue-specific manner. *Endocrinology 142*, 2005–2012.

Gao B, Hagenbuch B, Kullak-Ublick GA, Benke D, Aguzzi A, Meier PJ (**2000**) Organic anion-transporting polypeptides mediate transport of opioid peptides across blood–brain barrier. *J Pharmacol Exp Ther 294*, 73–79.

Gao F, Johnson DL, Ekins S, Janiszewski J, Kelly KG, Meyer RD, West M (**2002**) Optimizing higher throughput methods to assess drug–drug interactions for CYP1A2, CYP2C9, CYP2C19, CYP2D6, rCYP2D6 and CYP3A4 *in vitro* using a single point IC50. *J Biomol Screen 7*, 373–382.

Garrigues A, Loiseau N, Delaforge M, Ferte J, Garrigos M, Andre F, Orlowski S (**2002**) Characterization of two pharmacophores on the multidrug transporter P-glycoprotein. *Mol Pharmacol 62*, 1288–1298.

Gorboulev V, Ulzheimer JC, Akhoundova A, Ulzheimer-Teuber I, Karbach U, Quester S, Baumann C, Lang F, Busch AE, Koepsell H (**1997**) Cloning and characterization of two human polyspecific organic cation transporters. *DNA Cell Biol 16*, 871–881.

Gray JH, Owen RP, Giacomini KM (**2004**) The concentrative nucleoside transporter family, SLC28. *Pflugers Arch 447*, 728–734.

Greene N (**2002**) Computer systems for the prediction of toxicity: an update. *Adv Drug Deliv Rev 54*, 417–431.

Grundemann D, Gorboulev V, Gambaryan S, Veyhl M, Koepsell H (**1994**) Drug excretion mediated by a new prototype of polyspecific transporter. *Nature 372*, 549–552.

Grundemann D, Babin-Ebell J, Martel F, Ording N, Schmidt A, Schomig E (**1997**) Primary structure and functional expression of the apical organic cation transporter from kidney epithelial LLC-PK1 cells. *J Biol Chem 272*, 10408–10413.

Guner OF (**2000**) *Pharmacophore, Perception, Development and Use in Drug Design*. University International Line, San Diego, CA.

Guner OF (**2002**) History and evolution of the pharmacophore concept in computer-aided drug design. *Curr Top Med Chem 2*, 1269–1277.

Guner O, Clement O, Kurogi Y (**2004**) Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. *Curr Med Chem 11*, 2991–3005.

Hagenbuch B, Meier PJ (**2004**) Organic anion transporting polypeptides of the OATP/SLC21 family: phylogenetic classification as OATP/SLCO superfamily, new nomenclature and molecular/functional properties. *Pflugers Arch 447*, 653–665.

Hamada A, Miyano H, Watanabe H, Saito H (**2003**) Interaction of imatinib mesilate with human P-glycoprotein. *J Pharmacol Exp Ther 307*, 824–828.

Hansch C, Lien EJ, Helmer F (**1968**) Structure–activity correlations in the metabolism of drugs. *Arch Biochem Biophys 128*, 319–330.

Harlow GR, Halpert JR (**1997**) Alanine-scanning mutagenesis of a putative substrate recognition site in human cytochrome P4503A4. *J Biol Chem 272*, 5396–5402.

He YA, He YQ, Szklarz GD, Halpert JR (**1997**) Identification of three key residues in substrate recognition site 5 of human cytochrome P450 3A4 by cassette and site-directed mutagenesis. *Biochemistry 36*, 8831–8839.

Houghton PJ, Germain GS, Harwood FC, Schuetz JD, Stewart CF, Buchdunger E, Traxler P (**2004**) Imatinib mesylate is a potent inhibitor of the ABCG2 (BCRP) transporter and reverses resistance to topotecan and SN-38 *in vitro*. *Cancer Res 64*, 2333–2337.

Hutzler JM, Walker GS, Wienkers LC (**2003**) Inhibition of cytochrome P450 2D6: structure-activity studies using a series of quinidine and quinine analogs. *Chem Res Toxicol 16*, 450–459.

Jones JP, He M, Trager WF, Rettie AE (**1996**) Three-dimensional quantitative structure-activity relationship for inhibitors of cytochrome P4502C9. *Drug Metab Dispos 24*, 1–6.

Kelder J, Grootenhuis PDJ, Bayada DM, Delbressine LPC, Ploeman J-P (**1999**) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm Res 16*, 1514–1519.

Kim RB (**2003**) Organic anion-transporting polypeptide (OATP) transporter family and drug disposition. *Eur J Clin Invest 33*, 1–5.

Kim RB, Leake B, Cvetkovic M, Roden MM, Nadeau J, Walubo A, Wilkinson GR (**1999**) Modulation by drugs of human hepatic sodium-dependent bile acid transporter (sodium taurocholate cotransporting polypeptide) activity. *J Pharmacol Exp Ther 291*, 1204–1209.

Kliewer SA, Moore JT, Wade L, Staudinger JL, Watson MA, Jones SA, McKee DD, Oliver BB, Willson TM, Zetterstrom RH, Perlmann T, Lehmann JM (**1998**) An orphan nuclear receptor activated by pregnanes defines a novel steroid signalling pathway. *Cell 92*, 73–82.

Koepsell H (**1998**) Organic cation transporters in intestine, kidney, liver and brain. *Annu Rev Physiol 60*, 243–266.

Kullak-Ublick GA, Hagenbuch B, Stieger B, Schteingart CD, Hofmann AF, Wolkoff AW, Meier PJ (**1995**) Molecular and functional characterization of an organic anion transporting polypeptide cloned from human liver. *Gastroenterology 109*, 1274–1282.

Kullak-Ublick GA, Fisch T, Oswald M, Hagenbuch B, Meier PJ, Beuers U, Paumgartner G (**1998**) Dehydroepiandrosterone sulfate (DHEAS): identification of a carrier protein in human liver and brain. *FEBS Lett 424*, 173–176.

Kurogi Y, Guner OF (**2001**) Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr Med Chem 8*, 1035–1055.

Lack L, Walker JT, Singletary GD (**1970**) Ileal bile salt transport: *in vivo* studies of effect of substrate ionization on activity. *Am J Physiol 219*, 487–490.

Langer T, Eder M, Hoffmann RD, Chiba P, Ecker GF (**2004**) Lead identification for modulators of multidrug resistance based on *in silico* screening with a pharmacophoric feature model. *Arch Pharm (Weinheim) 337*, 317–327.

Lewis DFV, Jacobs MN, Dickins M, Lake BG (**2002**) Quantitative structure–activity relationships for inducers of cytochromes P450 and nuclear receptor ligands involved in P450 regulation within the CYP1, CYP2, CYP3 and CYP4 families. *Toxicology 176*, 51–57.

Li L, Lee TK, Meier PJ, Ballatori N (**1998**) Identification of glutathione as a driving force and leukotriene C4 as a substrate for oatp1, the hepatic sinusoidal organic solute transporter. *J Biol Chem 273*, 16184–16191.

Lightfoot T, Ellis SW, Mahling J, Ackland MJ, Blaney FE, Bijloo GJ, de Groot MJ, Vermeulen NPE, Blackburn GM, Lennard MS, Tucker GT (**2000**) Regioselective hydroxylation of debrisoquine by cytochrome P450 2D6: implications for active site modeling. *Xenobiotica 30*, 219–233.

Mahon FX, Belloc F, Lagarde V, Chollet C, Moreau-Gaudry F, Reiffers J, Goldman JM, Melo JV (**2003**) MDR1 gene overexpression confers resistance to imatinib mesylate in leukemia cell line models. *Blood 101*, 2368–2373.

Maliepaard M, van Gastelen MA, Tohgo A, Hausheer FH, van Waardenburg RC, de Jong LA, Pluim D, Beijnen JH, Schellens JH (**2001**) Circumvention of breast cancer resistance protein (BCRP)-mediated resistance to camptothecins *in vitro* using non substrate drugs or the BCRP inhibitor GF120918. *Clin Cancer Res 7*, 935–941.

Mankowski DC, Ekins S (**2003**) Prediction of human drug metabolizing enzyme induction. *Curr Drug Metab 4*, 381–391.

Mankowski DC, Laddison KJ, Christopherson PA, Ekins S, Tweedie DJ, Lawton MP (**1999**) Molecular cloning, expression and characterization of CYP2D17 from cynomolgus monkey liver. *Arch Biochem Biophys 372*, 189–196.

Mankowski DC, Lawton MP, Ekins S (**2000**) Characterization of transgenic mouse strains using six human hepatic cytochrome P450 probe substrates. *Xenobiotica 30*, 745–754.

Margolis JM, O'Donnell JP, Mankowski DC, Ekins S, Obach RS (**2000**) (*R*)-, (*S*)- and racemic fluoxetine *N*-demethylation by human cytochrome P450 enzymes. *Drug Metab Dispos 28*, 1187–1191.

Mitcheson JS, Chen J, Lin M, Culberson C, Sanguinetti MC (**2000**) A structural basis for the drug-induced log QT syndrome. *Proc Natl Acad Sci USA 97*, 12329–12333.

Morgan BP, Swick AG, Hargrove DM, LaFlemme JA, Moynihan MS, Carroll RS, Martin KA, Lee E, Decosta D, Bordner J (**2002**) Discovery of potent, nonsteroidal and highly selective glucocorticoid receptor antagonists. *J Med Chem 45*, 2417–2424.

Niwa T (**2003**) Using general regression and probalistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J Chem Inf Comput Sci 43*, 113–119.

Norinder U, Osterberg T, Artursson P (**1999**) Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parameterization and PLS statistics. *Eur J Pharm Sci 8*, 49–56.

Oprea TI, Gottfries J (**1999**) Toward a minimalistic modeling of oral drug absorption. *J Mol Graph Model 17*, 261–274.

Pajeva IK, Wiese M (**2002**) Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis). *J Med Chem.*

Palm K, Luthman K, Ungell A-L, Strandlund G, Artursson P (**1996**) Correlation of drug absorption with molecular surface properties. *J Pharm Sci 85*, 32–39.

Palm K, Luthman K, Ungell A-L, Strandlund G, Beigi F, Lundahl P, Artursson P (**1998**) Evaluation of dynamic polar molecular surface area as a predictor of drug absorption: comparison with other computational and experimental predictors. *J Med Chem 41*, 5382–5392.

Pascussi JM, Drocourt L, Fabre JM, Maurel P, Vilarem MJ (**2000 a**) Dexamethasone induces pregnane X receptor and retinoid X receptor-alpha expression in human hepatocytes: synergistic increase of CYP3A4 induction by pregnane X receptor activators. *Mol Pharmacol 58*, 361–372.

Pascussi JM, Gerbal-Chaloin S, Fabre JM, Maurel P, Vilarem MJ (**2000 b**) Dexamethasone enhances constitutive androstane receptor expression in human hepatocytes: consequences on cytochrome P450 gene regulation. *Mol Pharmacol 58*, 1441–1450.

Pascussi JM, Drocourt L, Gerbal-Chaloin S, Fabre JM, Maurel P, Vilarem MJ (**2001**) Dual effect of dexamethasone on CYP3A4 gene expression in human hepatocytes. Sequential role of glucocorticoid receptor and pregnane X receptor. *Eur J Biochem 268*, 6346–6358.

Patel Y, Gillet VJ, Bravi G, Leach AR (**2002**) A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J Comput-Aided Mol Des 16*, 653–681.

Patil SD, Ngo LY, Unadkat JD (**2000**) Structure–inhibitory profiles of nucleosides for the human intestinal N1 and N2 Na^+-nucleoside transporters. *Cancer Chemother Pharmacol 46*, 394–402.

Pearlstein RA, Vaz RJ, Kang J, Chen X-L, Preobrazhenskaya M, Shchekotikhin AE, Korolev AM, Lysenkova LN, Miroshnikova OV, Hendrix J, Rampe D (**2003**) Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg Med Chem 13*, 1829–1835.

Raevsky OA, Schaper K-J, Artursson P, McFarland JW (**2001**) A novel approach for prediction of intestinal absorption of drugs in humans based on hydrogen bond desccriptors and structural similarity. *Quant Struct–Act Relat 20*, 402–413.

Rodrigues AD (**1997**) Preclinical drug metabolism in the age of high-throughput screening: an industrial perspective. *Pharmacol Res 14,* 1504–1510.

Satlin LM, Amin V, Wolkoff AW (**1997**) Organic anion transporting polypeptide mediates organic anion/HCO_3^- exchange. *J Biol Chem 272,* 26340–26345.

Scala S, Akhmed N, Rao US, Paull K, Lan L-B, Dickstein B, Lee J-S, Elgemeie GH, Stein WD, Bates SE (**1997**) P-glycoprotein substrates and antagonists cluster into two distinct groups. *Mol Pharmacol 51,* 1024–1033.

Schuster D, Langer T (**2005**) The identification of ligand features essential for PXR activation by pharmacophore modeling. *J Chem Inf Model 45,* 431–439.

Shapiro AB, Ling V (**1997**) Positively cooperative sites for drug transport by P-glycoprotein with distinct drug specificities. *Eur J Biochem 250,* 130–137.

Shapiro AB, Fox K, Lam P, Ling V (**1999**) Stimulation of P-glycoprotein-mediated drug transport by prazosin and progersterone. Evidence for a third site. *Eur J Biochem 259,* 841–850.

Shimada J, Ekins S, Elkin C, Shaknovich EI, Wery J-P (**2002**) Integrating computer-based *de novo* drug design and multidimensional filtering for desirable drugs. *Targets 1,* 196–205.

Shitara Y, Itoh T, Sato H, Li AP, Sugiyama Y (**2003**) Inhibition of transporter-mediated hepatic uptake as a mechanism for drug–drug interaction between cerivastatin and cyclosporin A. *J Pharmacol Exp Ther 304,* 610–616.

Shitara Y, Hirano M, Sato H, Sugiyama Y (**2004**) Gemfibrozil and its glucuronide inhibit the OATP2(OATP1B1: SLC21A6)-mediated hepatic uptake and CYP2C8-mediated metabolism of cerivastatin – analysis of the mechanism of the clinically relevant drug–drug interaction between cerivastatin and gemfibrozil. *J Pharmacol Exp Ther 311,* 228–236.

Smith PA, Sorich MJ, McKinnon RA, Miners JO (**2003a**) *In silico* insights: chemical and structural characteristics associated with uridine diphosphate–glucuronosyltransferase substrate selectivity. *Clin Exp Pharmacol Physiol 30,* 836–840.

Smith PA, Sorich MJ, McKinnon RA, Miners JO (**2003b**) Pharmacophore and quantitative structure–activity relationship modeling: complementary approaches for the rationalization and prediction of UDP–glucuronosyltransferase 1A4 substrate selectivity. *J Med Chem 46,* 1617–1626.

Snyder R, Sangar R, Wang J, Ekins S (**2002**) Three dimensional quantitative structure activity relationship for CYP2D6 substrates. *Quant Struct-Act Relat 21,* 357–368.

Song J, Clagett-Dame M, Peterson RE, Hahn ME, Westler WM, Sicinski RR, DeLuca HF (**2002**) A ligand for the aryl hydrocarbon receptor isolated from lung. *Proc Natl Acad Sci USA 99,* 14694–14699.

Sorich M, Smith PA, McKinnon RA, Miners JO (**2002**) Pharmacophore and quantitative structure activity relationship modeling of UDP–glucuronosyltransferase 1A1 (UGT1A1) substrates. *Pharmacogenetics 12,* 635–645.

Sorich MJ, Miners JO, McKinnon RA, Smith PA (**2004**) Multiple pharmacophores for the investigation of human UDP–glucuronosyltransferase isoform substrate selectivity. *Mol Pharmacol 65,* 301–308.

Stenberg P, Luthman K, Ellens H, Lee CP, Smith PL, Lago A, Elliot JD, Artursson P (**1999**) Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharmacol Res 16,* 1520–1526.

Stenberg P, Norinder U, Luthman K, Artursson P (**2001**) Experimental and computational screening models for the prediction of intestinal drug absorption. *J Med Chem 44,* 1927–1937.

Suhre WM, Ekins S, Chang C, Swaan PW, Wright SH (**2005**) Molecular determinants of substrate/inhibitor binding to the human and rabbit renal organic cation transporters, hOCT2 and rbOCT2. *Mol Pharmacol 67,* 1067–1077.

Tamai I, Nezu J, Uchino H, Sai Y, Oku A, Shimane M, Tsuji A (**2000**) Molecular identification and characterization of novel members of the human organic anion transporter (OATP) family. *Biochem Biophys Res Commun 273,* 251–260.

Tukey RH, Strassburg CP (**2000**) Human UDP–glucuronosyltransferases: metabolism, expression and disease. *Annu Rev Pharmacol Toxicol 40,* 581–616.

Usui T, Saitoh Y, Komada F (**2003**) Induction of CYP3As in HepG2 cells by several drugs. Association between induction of

CYP3A4 and expression of glucocorticoid receptor. *Biol Pharm Bull 26*, 510–517.

van de Waterbeemd H, Gifford E (**2003**) ADMET *in silico* modeling: towards prediction paradise? *Nat Rev Drug Discov 2*, 192–204.

VandenBranden M, Wrighton SA, Ekins S, Gillespie JS, Binkley SN, Ring BJ, Gadberry MG, Mullins DC, Strom SC, Jensen CB (**1998**) Alterations in the catalytic activities of drug-metabolizing enzymes in cultures of human liver slices. *Drug Metab Dispos 26*, 1063–1068.

van Montfoort JE, Hagenbuch B, Fattinger KE, Muller M, Groothuis GM, Meijer DK, Meier PJ (**1999**) Polyspecific organic anion transporting polypeptides mediate hepatic uptake of amphipathic type II organic cations. *J Pharmacol Exp Ther 291*, 147–152.

Wandell C, Kim RB, Kajiji S, Guengerich FP, Wilkinson GR, Wood AJJ (**1999**) P-glycoprotein and cytochrome P-450 3A inhibition: dissociation of inhibitory potencies. *Cancer Res 59*, 3944–3948.

Wang J, Giacomini KM (**1999**) Characterization of a bioengineered chimeric Na^+-nucleoside transporter. *Mol Pharmacol 55*, 234–240.

Wang Q, Halpert JR (**2002**) Combined three-dimensional quantitative structure–activity relationship analysis of cytochrome P450 2B6 substrates and protein homology modeling. *Drug Metab Dispos 30*, 86–95.

Wessel MD, Mente S (**2001**) ADME by computer. *Annu Rep Med Chem 36*, 257–266.

Wessel MD, Jurs PC, Tolan JW, Muskal SM (**1998**) Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci 38*, 726–735.

Williams JA, Hyland R, Jones BC, Smith DA, Hurst S, Goosen TC, Peterkin V, Koup JR, Ball SE (**2004 a**) Drug–drug interactions for UDP–glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (auci/auc) ratios. *Drug Metab Dispos 32*, 1201–1208.

Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE (**2000**) Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. *Mol Cell 5*, 121–131.

Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovic D, Jhoti H (**2003**) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature 424*, 464–468.

Williams PA, Cosme J, Vinkovic DM, Ward A, Angove HC, Day PJ, Vonrhein C, Tickle IJ, Jhoti H (**2004 b**) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science 305*, 683–686.

Wrighton SA, Schuetz EG, Thummel KE, Shen DD, Korzekwa KR, Watkins PB (**2000**) The human CYP3A subfamily: practical considerations. *Drug Metab Revs 32*, 339–361.

Xue L, Wang HF, Wang Q, Szklarz GD, Domanski TL, Halpert JR, Correia MA (**2001**) Influence of P450 3A4 SRS-2 residues on cooperativity and/or regioselectivity of aflatoxin B_1 oxidation. *Chem Res Toxicol 14*, 483–491.

Yamashita F, Hashida M (**2004**) *In silico* approaches for predicting ADME properties of drugs. *Drug Metab Pharmacokinet 19*, 327–338.

Yano JK, Wester MR, Schoch GA, Griffin KJ, Stout CD, Johnson EF (**2004**) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-Å resolution. *J Biol Chem 279*, 38091–38094.

Yarim M, Moro S, Huber R, Meier PJ, Kaseda C, Kashima T, Hagenbuch B, Folkers G (**2005**) Application of QSAR analysis to organic anion transporting polypeptide 1a5 (Oatp1a5) substrates. *Bioorg Med Chem 13*, 463–471.

Yates CR, Chang C, Kearbey JD, Yasuda K, Schuetz EG, Miller DD, Dalton JT, Swaan PW (**2003**) Structural determinants of P-glycoprotein-mediated transport of glucocorticoids. *Pharmacol Res 20*, 1794–1803.

Young SS, Ekins S, Lambert C (**2002**) So many targets, so many compounds, but so few resources. *Curr Drug Discov* December, 17–22.

Zhao YH, Le J, Abraham MH, Hersey A, Eddershaw PJ, Luscombe CN, Butina D, Beck G, Sherborne B, Cooper I, Platts JA, Boutina D (**2001**) Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. *J Pharmacol Sci 90*, 749–784.

15
Are You Sure You Have a Good Model?

Nicolas Triballeau, Hugues-Olivier Bertrand, and Francine Acher

15.1
Introduction

The objective of this last chapter is to ask a series of questions related to pharmacophore model validation and provide a few keys to answer them. Obviously the master question that comes up when one reaches the validation step in pharmacophore investigation is "*Do I have a good model?*". Addressing such a question requires one first to get back briefly to the very definition of pharmacophore. Indeed, the goodness of a given model first depends on how close the generated model is to the ideal pharmacophore and therefore depends on the definition one chooses to adopt. Box 1 reports several of them.

Box 1. Examples of pharmacophore definitions

Definition 1 (Paul Ehrlich, 1909): A pharmacophore is a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) biological activity [1].

Definition 2 (Peter Gund, 1977): A pharmacophore is an arrangement of molecular features or fragments forming a necessary but not sufficient condition for biological activity ([2] quoted in [3]).

Definition 3 (IUPAC recommendations, 1998): A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response [4].

Definition 4 (IUPAC recommendations, 1997): Pharmacophore generation is a procedure to extract the most important common structural features relevant for a given biological activity from a series of molecules with a similar mechanism of action [5].

Interestingly, there is a fairly good consensus between definitions, including the first one that was formulated when Paul Ehrlich first coined the term "pharmacophore" about a century ago. However, apart from definition 4, which evokes the knowledge of a lead series, they are all very theoretical, making the pharmacophore a conceptual template, an idea for the mind, more than an entity endowed with reality. Indeed, in the real world, one is limited by two material as-

Pharmacophores and Pharmacophore Searches. Edited by T. Langer and R. D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1

pects: (1) the quality of the available dataset and (2) the method envisaged to perceive the pharmacophore.

In fact, most (if not all) datasets are incomplete and multiple pharmacophore solutions can be consistent with the definition. Van Drie insists on the fact that some datasets are easy to work with and some are hard [6, 7]. Regarding the method being used to search the pharmacophore, we are first limited by the oversimplification that had to be performed to describe drug–target interactions for computer programs. To name a few, they assume that similar compounds will bind in the same way, they assume that all H-bond interactions have the same strength and they assume that water molecules have the same behavior upon binding of every ligand. Although sufficient in many cases, such postulates have numerous counterexamples [8]. Second, we should be aware either of inherent errors of algorithms used to generate pharmacophore or, alternatively, of the bias introduced by a manual construction. In fact, experts will often call the resulting solutions "pharmacophore models" or "hypotheses" to affirm their difference from the ideal pharmacophore that is capable of accurately predicting the activity of any envisaged compound.

In summary, pharmacophore model validation is the building of a body of evidence by first relying on validation methods adapted to the search algorithm (and its limitations) and second by resorting to approaches that will incorporate external data and therefore account for some inherent imperfections of the dataset. The first section of this chapter will report some of the most used validation methods and which are related to one or both of these aspects.

Many case studies have been described for pharmacophore models. The second section will report some success stories in which pharmacophore models have been used. The validation methods to which the authors have to resort will be particularly emphasized. Undeniably, the ultimate validation method for a model is to demonstrate its practical usefulness for drug discovery!

In the third and last section, we will report the construction of a new pharmacophore model for metabotropic glutamate receptor subtype 4 (mGlu4R) agonists. In spite of the fact that the available dataset is particularly difficult for pharmacophore modeling, we will show that the validation methods used allowed us to have great confidence in the hypotheses generated.

15.2 Validation Methods: Different Answers Brought to Different Questions

15.2.1 Software-related Validation Methods

15.2.1.1 Ligand-based Pharmacophore Research

Most automated pharmacophore model generators are equipped with an internal scoring method to rank different hypotheses (here termed "ranking function") and provide the expert with the most relevant solutions. Indeed, real-life datasets often have inherent ambiguities and different models (sometimes several tens of them) can be output by a given search algorithm. Knowing how the software searches and estimates acceptable models is the first prerequisite to understand the solutions, better comprehend their limitations and adapt validation accordingly. The design features of the most commonly used programs are summarized in Box 2.

Box 2. Ligand-based automated pharmacophore model generators

Catalyst HipHop [9]

Objective: Find common feature configurations amongst a set of active molecules.
Algorithm: HipHop uses a pruned exhaustive search method. Starting with simple two-feature pharmacophores, the program tries to add one extra common feature at a time until no larger common pharmacophore configuration exists [10]. Combinations that cannot be completed to reach a minimum number of features are not further explored.
Ranking function: Its internal scoring function depends on the displacement in the alignment of the input molecules and the uniqueness of the pharmacophore (e.g. other things being equal, a positive ionizable feature is more unique than a hydrophobe feature).

Catalyst HypoGen [9]

Objective: Retain models with features that better explain the differences between activities.
Algorithm: In a constructive phase, HypoGen uses a simplified version of HipHop to catalogue all common pharmacophores among the most active compounds. This pharmacophore space is then reduced to solutions that match less than half of the least actives. In a last stage, 3D-QSAR models are built [via a linear regression between the geometric fit and the log(activity) of the compounds] and optimized by simulated annealing [11].
Ranking function: HypoGen tries to minimize a function that describes the cost of the model in number of bits (Occam's razor principle: all other things being equal, the simplest model is the best). The cost function is a weighted sum of the error cost (divergence of the estimated activities from the actual values), the configuration cost (entropy cost that depends on the size of the search space) and weight cost (divergence of the feature weights from their theoretical contribution to the activity –2.0).

DISCO [12]

Objective: Find 3D alignments of the pharmacophore features in different molecules.
Algorithm: DISCO uses the Bron-Kerbosch method to detect the maximum cliques in a given graph [13]. Here the graph is made of nodes representing ensembles of matching features between molecules. Two nodes are connected by a graph edge if the two feature ensembles can be aligned simultaneously.
Ranking function: DISCO outputs all possible solutions. It is up to the user to decide their relevance further [14].

GASP [15]

Objective: Determine the correspondence between functional groups in different molecules and the alignment of these groups in a common geometry.
Algorithm: GASP uses a genetic algorithm that iteratively optimizes a population of chromosomes according to the fitness function described here below [16]. Each chromosome encodes angles of rotations about flexible bonds and mappings between features.

Ranking function: The fitness function is the weighted sum of three terms: number and similarity of overlaid elements, common volume of all molecules and internal van der Waals energy of each molecule.

Phase [17]

Algorithm: Phase uses a partitioning algorithm in which pharmacophore configurations are placed in multi-dimensional boxes. Each box represents a common pharmacophore only if it contains a sufficient number of active ligands. The resulting alignment can be used to build 3D-QSAR models. To our knowledge, this algorithm has not been further described in the literature nor validated.
Ranking function: The scoring function is made of the weighted contributions of three aspects: quality of the alignment, similarity (common volume/total volume) and selectivity (rarity).

If one compares Catalyst-HipHop with Catalyst-HypoGen, for instance, it is common to obtain very different models. Since HipHop searches common features amongst active molecules, the output models may have retained features that HypoGen may have discarded as not being relevant to explain activity. The example that we will detail at the end of this chapter is a good illustration of this. It is worth mentioning that this is due to the input dataset and that Catalyst provides solutions to circumvent this.

In general, the ranking function is only adapted to the search algorithm and, although the contribution of their different factors may be tweaked by the user, it is hard to gauge the validity of the proposed models with the resulting output values. In fact, the objective of ranking functions is more to compare different solutions together than to give insight into the validity of the models. Some authors, however, have relied solely on them to select a pharmacophore model and successfully used it in their project. The following example is an illustration of this point.

Example (Dayam et al., 2005)
Dayam et al. used four *β*-diketo acid HIV-1 integrase inhibitors to build several HipHop hypotheses [18]. Despite some noticeable differences from the resolved protein–ligand complexes, the first-ranking hypothesis could be used as a filter prior to docking in the binding site. It is worth noting that the retained model contained all the features that one can determine from the 3D structure, although not in the correct orientation. At the end of their screening campaign, 48 out of the 110 virtual hits exhibited IC_{50} below 100 μM (hit rate: ca. 43%).

Catalyst-HypoGen's cost function is an exception and is commonly used as a first step for validation purposes. The so-called "cost analysis" provides an answer to the question, "*How strong is the activity signal given the input parameters and dataset?*". In fact, the program will always output two reference hypotheses to give insight into the significance of the results. The fixed cost is the calculated cost for the ideal model that fits the data perfectly. It is therefore the lowest possible cost for the specified parameters and dataset. In contrast, the null cost is the cost of the hypothesis which predicts all activities at the average of the input values. Obviously, the wider the difference between the fixed and null scores, the greater is the significance of the results. As a rough rule of thumb, a 40–60 bits difference between the cost of an output hypothesis and the cost of the null hypothesis leads to a predictive correlation probability of 75–90% [19]. Therefore, if the fixed – null cost difference is below 40–60 bits, finding a predictive model will probably be difficult (but not impossible).

15.2.1.2 Protein Structure-based Pharmacophore Research

The other commonly used approach to building pharmacophore models is to exploit the information provided by protein and protein–ligand complexes the structures of which have been resolved. Many programs are available using simple approaches such as MOE [20] and DS ViewerPro [21] that allow one to build pharmacophores manually into more sophisticated ones, such as LigandScout [22, 23] for automatic pharmacophore perception from 3D complexes. Other methods, such as structure-based focusing (SBF) [24] and MUSIC (which account for protein flexibility while building a hypothesis [25, 26]), can be used even if the protein binding site is empty by combining accessible interaction sites determined with geometric or energetic criteria.

15.2.1.3 Critical Remarks Regarding Structure-based Pharmacophore Models

Starting solely from the structural analysis of a few protein–ligand complexes, many studies have been published claiming to extract the “pharmacophore”. In fact, although protein–ligand complexes can clear up some ambiguities about a pharmacophore, strictly, the extracted information cannot be termed “pharmacophore” if it is not followed by a proper structure–activity relationship analysis. Because such complexes only display active molecules bound to the target site, they do not bring any information regarding the requirement of each interaction to the activity. Worse, it may be impossible to extract a distinctive pharmacophore with the structure-based method: In the case of receptor antagonists or enzyme inhibitors (i.e. most cases of medicinal applications), different ligands can exhibit unique interaction patterns. Provided that they protrude in a part of the binding site that the natural agonist (or substrate) needs to access, a competition can occur (Fig. 15.1 illustrates this point). In the case of agonists, however, the pharmacophore is more likely to be univocal since the ligands have to induce a more specific conformational modification on the receptor in order to trigger the activation. In summary, direct structural analyses of protein–ligand complexes are rarely capable on their own of providing an acceptable pharmacophore model without being properly validated.

Whether the plausible models are solutions obtained by a ligand-based or protein-based approach, the first validation screen is certainly the expert himself or herself, who will discard unsatisfactory solutions before further analysis.

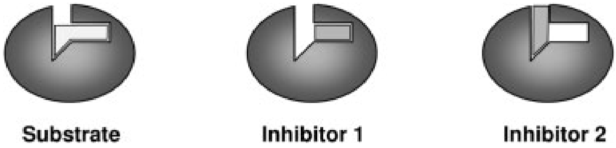


Fig. 15.1 Schematic representation of the binding mode of a substrate in an enzyme to be inhibited (left). Inhibitor 1 (center) and 2 (right) are both competitive inhibitors but their interaction pattern is different, making pharmacophore investigation tricky.

15.2.2 Visual Inspection

Do I like the proposed hypothesis? Although very subjective, this approach tends to counter-balance the inherent imperfections of automated methods. Programs such as DISCO and HipHop will provide multiple output models among which the expert has to choose the most acceptable. The eye is often the first and best critic for the pharmacophore hypotheses.

Thus, the nature of chemical features might differ between two hypotheses and the most specific is generally favored. For instance, models with more features will be favored as well as models with directional features (H-bond donor/acceptors, ring aromatics).

Models with less frequent features can also be more interesting. Obviously acidic/basic groups are less common than hydrophobic moieties amongst known drugs. A possible illustration of this point is to analyze the composition of a database containing “drug-like” molecules in term of chemical features. Figure 15.2 reports the number of compounds from the Derwent World Drug Index (WDI) 2003 database that map the most common chemical features. Interestingly, several hydrophobic features are easily found amongst drug-like molecules whereas ionizable groups are more seldom, especially if several are required. In contrast, it is fairly common to find up to four (and more) hydrophobic groups in drug-like molecules: More than 25% of Derwent WDI contains more than three hydrophobic features. Hence ionizable groups will certainly bring more selectivity if the model is to be used for database mining.

One may also favor models with spread-out features over models with features clustered on a specific part of the ligands. This may occur when the input molecules are large and feature rich, such as peptides and peptide mimics.

One may also prefer the models that align similar molecules (e.g. with the same scaffold) in the same orientation to be in line with one of the fundamental assumptions of pharmacophore modeling, which states that similar compounds have analogous binding modes. Despite rare counterexamples, drug discoveries in modern history have more than backed up this assumption.

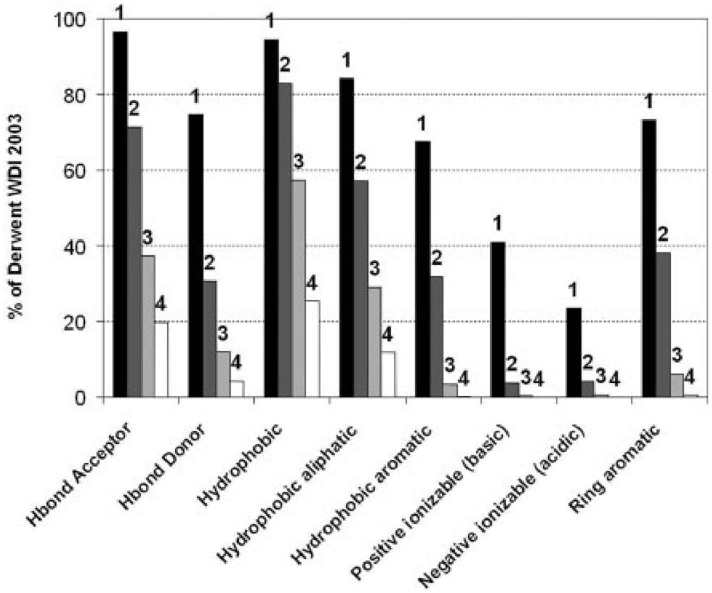


Fig. 15.2 Percentage of the Thompson Derwent WDI database (2003) that maps common Catalyst features. Only entries with molecular weights between 100 and 600 were considered (that is, a total of 51 726 entries). Colors show the fraction of the database that contain at least one (black), two (dark grey), three (light grey) and four (white) features.

15.2.3 Consistency with Structure – Activity Relationships

As a second prerequisite, a valid pharmacophore model has to provide insight into the structure–activity relationships (SARs) or at least explain them. At this stage, the question is, "*Can the hypothesis interpret a SAR?*" Such analysis will judge without appeal a poor dataset or an over-simplistic pharmacophore searching method.

15.2.3.1 Some Limitations of Computer Programs

Computer programs are often limited by the means that they have to explain the inactivity of some chemical series members. By design, a molecule is declared active if at least one low-energy conformation can map all chemical features of the model. In other words, a molecule is inactive only because it cannot exhibit the complete interaction pattern required by the binding site. In reality, many other reasons can be put forward to explain the inactivity of a compound. Martin evokes the following [14]:

1. Despite its ability to satisfy the pharmacophore model, a compound can contain groups that sterically prevent interaction.

2. It can contain other groups that are unfavorable to activity (e.g. an acidic moiety can disrupt the favorable interaction when it is located in the vicinity of an acidic residue of the binding site).
3. It is less soluble than its bioactive concentration.

It is worth noting that a new Catalyst module (Catalyst-HypoRefine [9]) is capable of accounting for false positives due to unfavorable steric contacts (point 1) by automatically placing exclusion volumes on strategic points around either HipHop or HypoGen models. To our knowledge, electrostatic or hydrophobic repulsions (point 2) are not yet handled by any pharmacophore searching programs. Field-based 3D-QSAR models (CoMFA [27], CoMSIA [28], etc.), on the other hand, have long been used to highlight such unfavorable zones (both steric and electrostatic), but most require the expert to provide a somewhat unrealistic alignment of active and inactive molecules.

15.2.3.2 **Retained Chemical Features**

If the input datasets on the other side do not bring any information about the influence (presence or absence) of a given chemical group on the binding affinity, one cannot expect the program to have a correct SAR interpretation on the missing information. In general, datasets are selected by experts in order to be as instructive as possible, but some pieces of information may simply not yet be available. Therefore, from a qualitative point of view, models that incorporate chemical features which are known to be significant to the activity should be favored. The use of a “test set” can help underline missing SAR features (see Section 15.2.4.3).

15.2.3.3 **Spatial Arrangement**

The space arrangement of the important feature is also to be considered. Indeed, many SAR studies include length modifications in scaffold side-chains to estimate the influence of flexibility and steric tolerance of the binding site (see, for example, [29, 30]). Chirality of the model is another spatial criterion that is necessary if two enantiomers of different activity level are to be discriminated. At least four points are necessary, but not sufficient to assure enantio-selectivity. The use of directional features (H-bond acceptor/donor) or shape criteria can help in obtaining chiral models.

Example of pharmacophore model with key spatial arrangement (Jullian et al., 1999; Bessis et al., 1999)

Good pharmacophore models are generally capable of predicting the bioactive conformation of the bound ligands. Jullian, Bessis and co-workers, for instance, have demonstrated that despite a more stable folded conformation in aqueous buffer, L-glutamate adopts an extended conformation when bound to metabotropic glutamate receptors (mGluR) [31, 32]. One year later, the publication of the crystal structure of L-glutamate bound to the ligand binding domain of mGlu1R [33] confirmed this key assertion for drug design.

15.2.3.4 3D-QSAR Pharmacophore Models

Some pharmacophore searching programs are designed to provide 3D-QSAR models that are capable of predicting the activity from a quantitative point of view. Apex-3D [34], Catalyst HypoGen [9] and Phase [17] are examples. Consequently, such a model should have correct statistics and abide by the common QSAR validation approaches. Box 3 describes some of them very briefly as they have been reviewed elsewhere (see [35–41]).

Box 3. Statistical measures of the quality of linear 3D-QSAR pharmacophore models

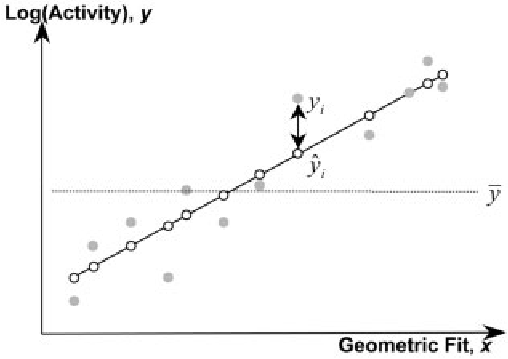
For a linear model, the estimated activity $\hat{y}_i$ (in log units) of a molecule (open circle on the graph) is a function of its geometric fit on the pharmacophore x_i: $\hat{y}_i = a \cdot x_i + b$. The n observed activities (y_i, gray circles) are compared with the predictions $\hat{y}_i$ with the three following sums:

Total sum of squares: $$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares: $$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares: $$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

One can show that $TSS = ESS + RSS$



where $\bar{y}$ is the arithmetic mean over all y_i. If the model is fitted with the least-squares technique, $\bar{y}$ is also the arithmetic mean over all $\hat{y}_i$.

The *Pearson's correlation coefficient* (r) gives an insight into how well a linear model fits the dataset; in other words, "*how much of the activity can be explained by a linear model?*". The closer r^2 is to 1, the better the model is. An r^2 of 0.89, for instance, means that 89% of the variance is explained by the linear model.

$$r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{3.1}$$

The *standard deviation* of error of prediction (SD) gives an insight into the accuracy of the prediction. If one considers the errors being normally distributed, more than 68% of estimations are performed with an error below SD.

$$SD = \left(\frac{1}{n-2} \cdot RSS\right)^{1/2} \tag{3.2}$$

Similarly, the *root mean square* (RMS) of errors gives an insight into the errors of the prediction. In Catalyst-HypoGen, the RMS is scaled according to the uncertainty of each activity measure (Unc):

$$RMS = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\log(Unc_i)}\right)^{1/2} \tag{3.3}$$

The *F statistic* illustrates the overall statistical significance. By comparing the calculated value with tabulated values for 1 and $n-2$ degrees of freedom (linear model) at a given level of confidence, one can assert that the model is or is not significant at this level:

$$F = \frac{ESS}{RSS/(n-2)} \tag{3.4}$$

Catalyst-HypoGen proposes another interesting validation method by analyzing the statistical significance of the SAR described by a model. Based on the Fisher randomization test, the observed activities are scrambled and randomly redistributed to the compounds of the training set. If some models generated from this new scrambled dataset exhibit a lower cost than the reference model, the significance of the original model is weakened. Of course, this "scrambling" work needs to be performed a certain number of times to reach a given significance level. Thus, if a total of N_{run} HypoGen runs are performed (reference and randomised runs) and n_{lc} models of lower cost than the reference are numbered, the significance is given by

$$Significance = \left(1 - \frac{1 + n_{lc}}{N_{run}}\right) \times 100$$

For instance, if a significance level of 95% is to be reached, at least 19 randomization runs need to be performed. If one cannot reach an acceptable level of significance, a new dataset and different searching parameters must be used.

15.2.4
External Data to Back Up a Pharmacophore Model

15.2.4.1 Biophysical Data

Returning to the definitions of pharmacophore (Box 1), one of the most direct methods to validate a pharmacophore model is to show its consistency with the interaction pattern that known ligands exhibit with the targeted protein site. The Protein Data Bank (PDB, http://www.rcsb.org/pdb/) is the *de facto* repository for macromolecular structures resolved by NMR or diffraction methods [42]. The structures of many protein–ligand complexes have been resolved and their atomic coordinates can be downloaded from the PDB web portal for further analysis.

Example of a structure-based pharmacophore model (Brenk et al., 2003)

Brenk et al. have successfully used structure-based pharmacophore models for ^{t}RNA guanine transglycosylase (TGT) inhibitors [43]. Starting from two complexes resolved by X-ray diffraction, three different hypotheses were derived depending on the presence or absence of a structural water molecule. The resulting pharmacophore model envisaged was an ensemble of these three hypotheses constructed with SYBYL [44]. The authors took particular care to validate it by showing that it explained structure–activity relationships. The inclusion of the model in a virtual screening project as a filter prior to a docking step allowed them successfully to enrich a composite database of ca. 800 000 molecules. Out of the nine selected molecules, seven were inhibited by TGT in the micromolar range and two in the sub-micromolar range.

Albeit very prudent, Ghose and Wendoloski suggested the use of a biophysical data source to determine the conformation of the free ligand as another means to validate a pharmacophore model [45]. After all, when one has no other data to start with, why would one discard it?

15.2.4.2 Other Published Pharmacophore Models

Although probably not sufficient alone, another commonly used approach to bring further credit to a model is to compare the resulting hypothesis with a previously reported pharmacophore hypothesis. Thus, Barbaro et al. [30] correlated their α_1-adrenergic receptor pharmacophore model with the work of De Marinis et al. [46]. Laggner et al. [47] as well reported similarities between their σ_1 receptor pharmacophore hypothesis and the model published by Glennon et al. in 1994 [29]. At that time, a thorough analysis of SAR data allowed the authors to propose a simple pharmacophore model containing three chemical features: two hydrophobic sites A and B and an amino moiety (Fig. 15.3 a). Using a more modern approach with Catalyst-HypoGen as a pharmacophore search engine, Laggner et al. described a more sophisticated model which agrees remarkably with Glennon's initial hypothesis (Fig. 15.3 b).

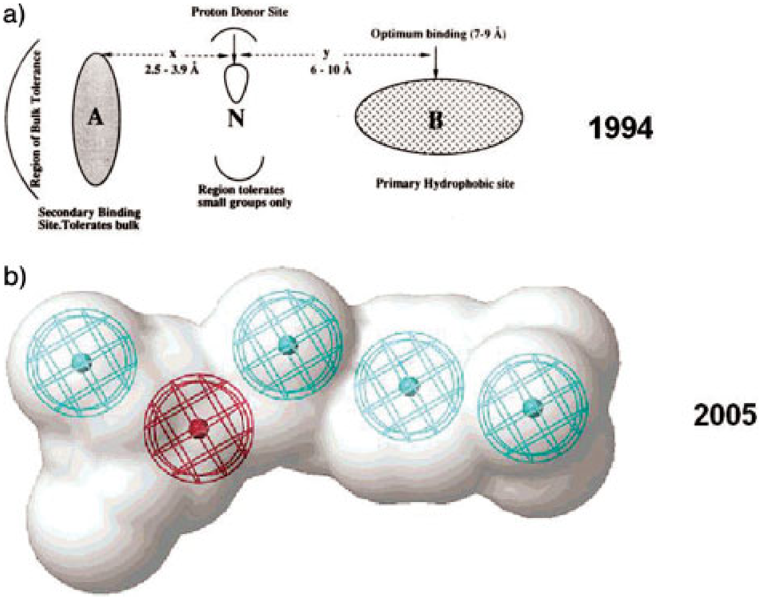


Fig. 15.3 Sigma-1 receptor pharmacophore hypotheses reported by Glennon et al. [29] (a) and Laggner et al. [47] (b; the red and cyan spheres represent positive ionizable and hydrophobic features respectively). The volume colored white accounts for the shape a high-affinity ligand. Reprinted with permission from [29] and [47]. Copyright 1994 and 2005, American Chemical Society.

15.2.4.3 The "Test Set" Approach and the Kubinyi Paradox

Since pharmacophore model design is often used when no three-dimensional information is available for the protein target, the "test set" approach is probably the most commonly used validation method, especially with 3D-QSAR pharmacophore models.

The idea is to take apart some molecules of known activity from the dataset set to confront later the generated model with different compounds to those of the "training set". In theory, this method brings an answer to the question, "*Can we extrapolate the predictions of the model to other different molecules?*" and it is certainly legitimate that one asks this question if the model is to be further exploited. For 3D-QSAR models, a statistical metric (often termed r^2_{pred}) similar to r^2 [see Eq. (3.3)] can be calculated.

The problem is how to define "different" in the above question. If the molecules of the training set are too similar to those of the test set, this aspect of the validation is not very challenging for the model and, conversely, if they are too different from each other, there is a risk of depriving the training set of some key SAR information. Therefore, molecules kept for the test set should be neither too similar nor too different from the training set compounds.

To circumvent this issue, "cross-validation" methods have been proposed to evaluate the internal predictivity of the model by discarding one or several compounds at a time from the training set and to compute an average correlation coefficient (q^2) with the subsequently derived models. Although long used by the QSAR community, Kubinyi et al. showed in a relatively recent survey that r^2_{pred} and q^2 are not correlated [48]. In other words, internal predictivity cannot guarantee extrapolability. This "Kubinyi paradox" is the strongest caveat when one relies on the "test set" approach for pharmacophore validation.

Furthermore, real datasets are often redundant and/or too small to rely on this idea and/or come from disparate *in vitro* assays. Once again, the know-how of the expert is key when it comes to deciding which molecule is to be taken in the training set.

A different but related approach consists in mining databases containing some known active compounds.

15.2.5 Database Mining

Many strategies have been proposed to enrich a set of molecules with active compounds by virtual screening: similarity search [49], docking-scoring [50], QSAR [51] and, of course, pharmacophore models. This section reviews some methods that can be applied to any of those approaches.

If a model is to be used as a query to search for active molecules in a database, a common validation method is to demonstrate its performance on a database for which the pharmacological activity of each compound is known (or at least flagged as active or inactive). Most often, such databases are made artificially for this purpose. Thus, after gathering a set of active compounds, one would seed them in a larger database of randomly selected (and supposedly inactive) molecules, the idea being to mimic some HTS results. The model is finally evaluated according to its ability to search the database for the actives and perform better than a random search (enrichment).

Although frequently used, this method has recently been called into question regarding the choice of the decoys (e.g. [52, 53]). Indeed, if selected randomly, compounds to confound the query are often inactive for obvious reasons that do not require a sophisticated model to discard them (the molecular weight is generally a good discriminator, for instance). In a real HTS run, molecules are often congeneric as they come from parallel synthesis and, consequently, are structurally more similar to one another. It would therefore be more reasonable to select decoys according to their similarity to the active molecules. In other words, such inactive compounds are more likely to produce a stronger interfering noise, making the search for activity signals more challenging for the query model. The problem is that one needs to be sure of the inactivity of the decoys as they are more likely to be actives than the randomly selected molecules. Such a piece of information is not always easy to acquire.

Several authors have successfully use pharmacologically indexed databases such as the Thomson Derwent WDI [54] and the KEGG COMPOUND [55] database in which several thousand molecules are recorded according to their known biological activities [47]. Screening of such databases allows one to validate the model for database mining by assessing its ability to retrieve active molecules (true positives) and discard inactive ones. However, it is worth mentioning that the false negative rate can be artificially high since a pharmacophore model is valid for a defined binding site. Hence one cannot expect it to retrieve molecules that bind to a different pocket of the same target.

15.2.5.1 Some Metrics to Assess Screening Performances

A plethora of metrics have been proposed to quantify the performance of a model upon database mining. Not all have been used in the particular case of pharmacophore models, but they are all applicable in the virtual screening context. Some have even been coined in a very different area to virtual screening. Matthews' correlation coefficient, for instance, was first proposed to evaluate the accuracy of the secondary structure predictions for T4 phage lysozyme. Since the idea is to be able to discriminate between good and bad predictions, Frimurer et al. [56] and later Goldblum [57] have used it in the context of drug discovery. Box 4 reports some metrics the optimization of which helps in selecting the best performing hypothesis.

Box 4. Metrics for performance assessment in virtual screening

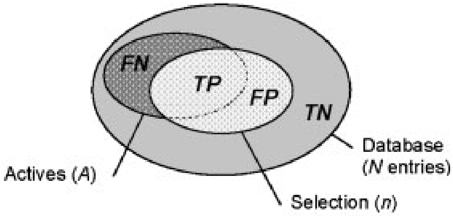
In order to facilitate the comparison between the metrics, the equations have been transcribed according to the notation summarized in the diagram shown. In a database containing a total of *N* entries among which *A* molecules are active on the investigated target, the virtual screening protocol selects *n* compounds as being actives.

TP: true positives

TN: true negatives

FN: false negatives

FP: false positives



Sensitivity [53, 58]:

$$Se = \frac{TP}{A} = \frac{TP}{TP + FN} \qquad (4.1)$$

Specificity [53, 58]:

$$Sp = \frac{TN}{N-A} = \frac{TN}{TN+FP} \tag{4.2}$$

Yield of actives [59, 60]:

$$Ya = \frac{TP}{n} \tag{4.3}$$

Enrichment [60–62]:

$$E = \frac{TP/n}{A/N} \tag{4.4}$$

Statistical significance [63]:

$$S = \sum_{k=TP}^{A} \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} \tag{4.5}$$

Balanced labeling performance [41, 64]:

$$\ell_{bal} = \frac{1}{2} \cdot Se + \frac{1}{2} \cdot Sp \tag{4.6}$$

Accuracy [60, 65, 66]:

$$Acc = \frac{TP+TN}{N} = \frac{A}{N} Se + \left(1 - \frac{A}{N}\right) \cdot Sp \tag{4.7}$$

Ford's M [67]:

$$M = \omega \cdot Se + (1-\omega) \cdot Sp \tag{4.8}$$

where ω is an adjustable weighting coefficient.

Discrimination ratio [68]:

$$DR = \frac{TP/A}{TN/(N-A)} = \frac{Se}{Sp} \tag{4.9}$$

Information content [69]:

$$I = TP \log\left(\frac{TP}{FP}\right) + FN \log\left(\frac{FN}{TN}\right) \tag{4.10}$$

"Matthews" correlation coefficient [56, 70]:

$$C = \frac{TP \cdot TN - FN \cdot FP}{((TN + FN)(TN + FP)(TP + FN)(TP + FP))^{1/2}} \tag{4.11}$$

"Goodness of hit list" [59]:

$$GH = \left(\frac{3}{4} \cdot Ya + \frac{1}{4} \cdot Se\right) \cdot Sp \tag{4.12}$$

Analysis of efficiency [71]:

$$AE = \frac{1}{2} \cdot (Se + Sp) \cdot \left(1 - \frac{U_s}{U_{total}}\right) \tag{4.13}$$

where U_s and U_{total} are the number of compounds of unknown activity selected and the total number of compounds of unknown activity in the database.

Each of these metrics gives a different insight into the performance of a screening workflow in retrieving active molecules or discarding inactive ones. Jacobsson et al., for example, used accuracy, sensitivity (which they termed the "recall'), yield of actives (termed "precision of the active class') and enrichment to characterize the performance of different structure-based virtual screening workflows [60].

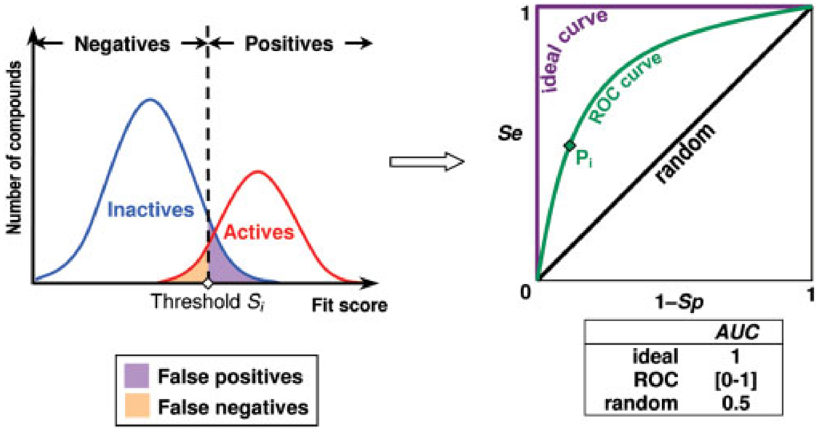
Among all the equations, the yield of actives [Eq. (4.3)] and the enrichment [Eq. (4.4)] are probably the most often used: the first is the hit rate one that would have if the *n* selected molecules were tested (again) and the second indicates how many times the virtual screening workflow performs better than a random selection in retrieving active compounds.

More directly related to pharmacophore model validation and hit list assessment, Güner and Henry designed the *GH* score [Eq. (4.12)] [59]. The different weighting coefficients are adapted to favor the high value of actives (*Ya*) over sensitivity (*Se*) because databases may contain compounds that bind the target in a different site (or interaction pattern) and which, obviously, cannot be identified by the pharmacophore model. The true meaning of several of the above metrics can be difficult to comprehend, but interestingly, many rely on two simple values: sensitivity [*Se*, Eq. (4.1)] and specificity [*Sp*, Eq. (4.2)]. For instance, one would notice that the "balanced labeling performance", the "accuracy", Ford's *M* and the "analysis of efficiency" are all linear combinations of sensitivity and specificity (if $\omega = A/N = \frac{1}{2}$ and $U_s = 0$, they even are equal).

In fact, sensitivity and specificity are the main characteristic features of any test which is to be used to categorize two populations. *Se* gives an insight into the ability of the model to select truly active molecules and *Sp*, in contrast, is the goodness in discarding inactive compounds. These two terms have the property to evolve in opposite directions when the number *n* of selected molecules changes. Indeed if most of the database is selected ($n \approx N$), most compounds are taken as actives and a minority of the truly active molecules will be lost. This maximizes *Se* ($Se \approx 1$). However, specificity will be minimized as most of the inactive molecules will be in the selection. The situation is reversed in the case where n is very small ($n \approx 0$). Consequently, one cannot optimize both *Se* and *Sp* at the same time and a trade-off is to be determined. If one can choose to rely on one of the above metrics to find an optimum, we have recently advocated the use of a simpler graphical technique which has been adopted as a gold standard in many other research areas: the receiver operating characteristic (ROC) curve method [53].

15.2.5.2 **The ROC Curve Approach**

Receiver operating characteristic (ROC) curves basically report the evolution of *Se* as a function of (1 – *Sp*) when n changes. In signal detection theory, *Se* is the perceived signal (here the activity) and (1 – *Sp*) is related to the detected background "noise" emitted by inactive molecules. The objective here is to answer the question: "*Considering available SAR data, how good is the model in discriminating active compounds from inactive ones?*". In other words, the ROC curve approach provides an answer to one of the key questions in virtual screening.



	AUC
ideal	1
ROC	[0-1]
random	0.5

Fig. 15.4 Performance assessment with ROC curves. The theoretical distributions for active (red curve) and inactive compounds (blue) as a function of their fit score on the pharmacophore (left). In most cases, these distributions overlap, leading to false predictions (colored areas). Upon threshold modification, proportions of such erroneous classifications change dramatically. Hence to any selection threshold S_i corresponds a unique point P_i (1 – Sp_i, Se_i) on the ROC graph and *vice versa*. The relative position of the ROC curve with respect to the 45° diagonal (random fit score distribution) and the ideal plot (when the distributions do not overlap) gives an insight into the overall accuracy of the computer test. Calculating the area under the ROC curve (*AUC*) is a practical way to quantify it.

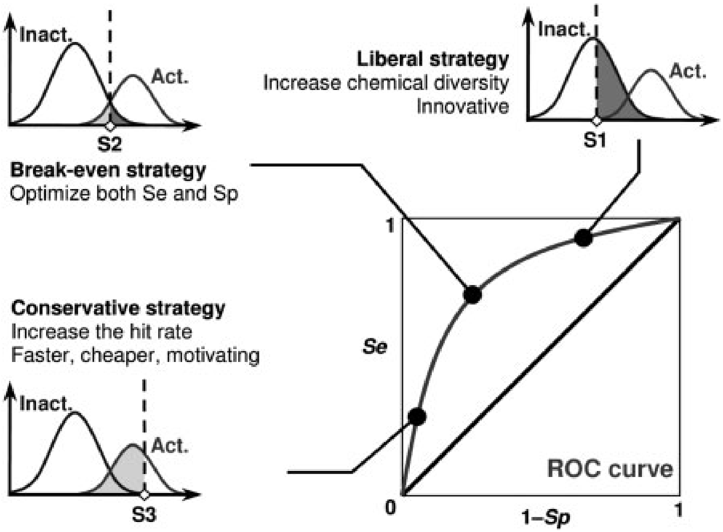


Fig. 15.5 Decision making from ROC curves. Different selection thresholds (S1–S3) correspond to different points on the ROC curve and allow one to tune S according to different strategies in drug discovery and different stages of R&D.

Most pharmacophore screening packages will provide a geometric fit score according to the best alignment of each molecule with the pharmacophore query. This allows one to rank molecules according to their fit scores and to define the selection as the *n* best fitting molecules. Figure 15.4 illustrates how to plot a ROC curve from the distributions of both active and inactive compounds as a function of the fit score.

Instead of searching for the single point on the graph (*Se*, *Sp*) that maximizes one of the above equations (Box 4), the area under the ROC curve (*AUC*) objectively characterizes the overall performance of the model by considering all possible thresholds. No particular statistics or mathematical equations are required. In addition, the second advantage of the ROC approach is that it lets the user decide where to set the threshold between the selected compounds (those which are worth testing further) and the discarded ones (those which are likely to be inactive). In fact, the selection threshold might evolve according to practical needs in addition to the progress in the drug discovery process (Fig. 15.5). For example, early hit finding may favor sensitivity over specificity to allow more structural variety amongst selected compounds. In contrast, during lead optimization, scaffold “hopping” is less a priority and a more demanding selection threshold (favoring specificity) may be chosen (conservative strategy, point S3 in Fig. 15.5).

A few recommendations regarding the set of molecules to be used to exploit the ROC approach fully can be given as follows. First, as the objective is to account for all available SAR data, the more molecules are included, the better the analysis is. The aim, however, is to balance both active and inactive populations in order not to favor one aspect of the analysis over the other (ability to select active compounds and ability to discard inactive compounds). The second recommendation if one has to choose the molecules to be considered is to favor chemical diversity amongst the active molecules and to select the inactive compounds which are more similar to the actives. In this way, the model is truly challenged for its discriminatory abilities. Finally, a mixture of training set and test set molecules is preferable since, as highlighted by the Kubinyi paradox (see Section 15.2.4.3), both sets can account for different SAR aspects.

15.3 A Successful Application: the Ultimate Validation Proof

This section will underline the importance of some of the above-described validation methods depending on the planned use for the pharmacophore model to be determined. It will be illustrated by several case studies. Successful applications are undoubtedly the best way to validate a pharmacophore hypothesis: they bring a straightforward answer to the practical question: *Was the model useful?*

Several articles and reviews have reported impressive lists of possible pharmacophore model uses to advocate for pharmacophore research [3, 6, 45, 61]. In fact, there are three main domains of applications to exploit pharmacophore models:

1. databases mining for active molecules (virtual screening);
2. guiding medicinal/computational chemistry in the design of new ligands;
3. activity prediction.

Depending on the envisaged use, some of the above-described validation methods are more suitable than others. Of course, any pharmacophore should be neat and convey SAR properly, but some approaches are more tailored for a particular application than others. The following three sections will treat each case in turn to illustrate this point.

15.3.1 Validation of Pharmacophore Models for Virtual Screening

Pharmacophores are known for their speed in database screening, especially when compared with the “high-throughput docking” approach. Indeed, by design, a pharmacophore is a more simplistic object than a protein. For example, no steric hindrance has to be evaluated when a molecule is to be fitted on a four-feature pharmacophore. Consequently, pharmacophore models are often used for virtual screening purposes either as they are or as a filter prior to a more time-consuming docking step. The examples reported below show different approaches.

15.3.1.1 Which Validation Method Should One Insist On?

In addition to reporting SAR data, pharmacophore models used for virtual screening have to give evidence of their ability to discriminate active from inactive molecules. Indeed, the goal of virtual screening being to enrich a set of molecules with active molecules, a simple yes/no decision has to be made regarding further evaluation of a given molecule. Consequently, the ROC curve approach is particularly suited to validate the model by measuring the *AUC*. Different strategies can be envisaged: if the pharmacophore model is used as a simple filter prior to a more time-consuming activity prediction (which is believed to capture finer aspects of the binding), sensitivity is generally favored over specificity to prevent the loss of active candidates. Conversely, if the pharmacophore model is the ultimate decision maker, a more specific selection threshold may be set.

If new scaffolds are to be found, a second important feature of such models is their completeness. It allows different structural solutions to fulfil the interaction pattern required by the pharmacophore. Accuracy in the activity prediction, on the other hand, is not paramount.

Example of successful use of a pharmacophore model for virtual screening (Bhattacharjee et al., 2004)

In this case study, Bhattacharjee et al. have used a 3D-QSAR pharmacophore model derived with Catalyst-HypoGen to identify potential *Plasmodium falciparum* cyclin-dependent kinase (Pfmrk) inhibitors [95]. Using a training set of 15 structurally diverse inhibitors with activities spanning the 0.13–1100 μM range, the best model (two H-bond acceptors, one ring aromatic and one aliphatic hydrophobic) exhibited excellent statistical parameters: correlation coefficient $r=0.9$, $RMS=0.8$ (log unit) and correlation with a test set of 15 other inhibitors led to $r_{pred}=0.7$. A cost analysis and a structural validation based on ortholog proteins were also performed to confirm the model further. A virtual screening campaign performed on their in-house collection (ca. 290000 molecules) allowed the author to retrieve 16 compounds with predicted activities below 52 μM. All 16 molecules were actual Pfmrk inhibitors *in vitro* with activities below the 100 μM threshold.

Example of use of a pharmacophore model as a filter in a complete virtual screening workflow (Steindl et al., 2005)

Their purpose being to identify new human rhinovirus coat protein inhibitors, Steindl et al. recently reported a successful virtual screening workflow with multiple hierarchical filtering steps [96]. The structure-based pharmacophore model that they used as the first filter was extracted from a PDB complex and expressed as a set of Catalyst features coupled with a shape query. Following thorough analysis of different known complexes, the built model was validated for database mining applications, i.e. according to its selectivity for known inhibitors when compared with the Derwent WDI database. Using the sensitivity/specificity paradigm, this first filter exhibited $Se=0.5$ and $Sp=0.95$. Ten compounds were retrieved from a chemical provider catalogue with this model and further analyzed via docking (LigandFit), scoring (LigScore2) and finally PCA-based clustering. The six best performing molecules on the overall protocol were all active on the viral coat protein in the micromolar range.

Example of a poorly validated model (Sirois et al., 2004)

The emergence of highly infectious agents such as the SARS-associated corona virus (SARS-CoV) has given rise to urgent research for new anti-infectious. With the aim of identifying the inhibitors of the SARS-CoV main proteinase, Sirois et al. exploited the recently published X-ray structure of this putative target [97]. A structure-based pharmacophore model was built using the software MOE to allow the screening of an impressive 3.6 million compound database. Since no SAR data were available, the authors could not carry out any validation of the proposed model, their hypothesis relying exclusively on the 3D structure of a unique inhibitor candidate docked in the binding site. Moreover, even though a list of the 500 best fitting compounds is reported in their paper, no *in vitro* assay results were provided. In their defense, the SARS-CoV is a particularly difficult case since a P3 facility is required to manipulate this highly pathogenic agent. Clones of the viral main proteinase were therefore not easily accessible at the time they submitted their paper.

15.3.2
Validation of Pharmacophore Models to Guide Medicinal and Computational Chemistry

As we have seen, pharmacophore models are compilations of SAR data. Consequently, they can be used by both medicinal and computational chemists to guide them in their research.

Assuming that better fits of molecules on the pharmacophore model will improve the activity, chemists can exploit the generated hypothesis for library design and lead optimization. In this objective, pharmacophore fingerprints facilitate similarity calculations (see [72, 73] as examples).

Additionally, computational chemists often use the resulting output alignment of the molecules as input for 3D-QSAR modeling. As already stated, most field-based 3D-QSAR approaches (such as CoMFA) need a pre-aligned set of molecules and the pharmacophore method is certainly one of the best ways to obtain an objective alignment of the compounds. Klabunde et al., for instance, have recently reported the use of a pharmacophore model of human liver glycogen phosphorylase inhibitors together with 3D information from inhibitor–enzyme complexes to derive a predictive CoMFA model [98].

Apart from expressing SAR, there is no validation method that is particularly recommended for this use. Of course, the selectivity of the pharmacophore will definitely facilitate library design and a possible way to assess it is to screen a database of molecules flagged as active and inactive. In this respect, this is rather similar to the virtual screening usage and the ROC curve approach could be used to assess the ability to discriminate between active and inactive compounds. However, in contrast to this use, the completeness of the model is not an issue here. In practice, lead optimization is less focused towards scaffold "hopping" than the virtual screening process. Consequently, if the model is tuned for a specific molecular series (e.g. series enumerated by a combinatorial approach), designing a library in the same series does not require more complete pharmacophore models.

Example of design of new α_1-adrenoceptor antagonists (Betti et al., 2002) following pharmacophore investigation (Barbaro et al., 2001)

In order to rationalize the design of α_1-adrenoceptor antagonists, Barbaro et al. generated a pharmacophore model for a series of pyridazone derivatives [74]. A set of 24 molecules with activity values spanning the range 0.21–2396 nM was used as input for Catalyst-HypoGen. Their best model was validated via cost analysis (fixed cost, 101; model cost range, 113–133.3; null cost, 155), statistical parameters for structure–activity relationship ($r = 0.92$; $RMS = 0.89$ log of activity; Fisher randomization test, 95% of significance) and visual inspection of the mapping of the molecules on the model. In particular, the influence of the length of a polymethylene chain and the importance of ortho substituents could be explained by the retained model. Finally, the model was validated by database mining to assess the performance of the model in retrieving known α_1-adrenocepter antagonists with different scaffolds and show good concordance with third-party models of the receptor. Betti et al. then exploited the model to guide them in the design of a novel series of α_1-antagonists [99]. Starting from the structure of a virtual hit, they managed to synthesize new compounds with remarkable α_1 affinity and selectivity: $K_i \approx 1$ nM, α_1/α_2 affinity ratio >280).

15.3.3 Validation of Pharmacophore Models for Activity Prediction

Activity prediction can be performed on a wide range of targets. In most cases, the model estimates the affinity for a protein target which is (sometimes hypothetically) linked to the treatment of a particular disease. Some models, however, have been proposed to predict molecular affinity for protein sites relevant for pharmacokinetics (e.g. P-glycoproteins), metabolism [cytochromes P_{450} (CYP)] or toxicology (hERG channel) assessments. Thus, following the trend for earlier predictions of ADMET properties using *in vitro* assays on key proteins [75], *in silico* techniques can be used to weed out molecules that are likely to exhibit poor ADMET properties [76]. Norinder has recently reviewed this subject [77] and shown that pharmacophore models are often used for this purpose.

15.3.3.1 Which Validation Method Should One Insist On?

When compared with the first application (virtual screening), pharmacophore models for activity prediction require more sophisticated models to capture more subtle effects (e.g. orientation of directional features, different weights to account for different contributions of each feature). Akin to other QSAR approaches, the models are often biased towards one or several particular congeneric series, namely those used in the training set. Consequently, although it may reinforce the confidence one may have in the generated model, performances in database mining are less important than accuracy in the activity predictions. Moreover, one can expect that, like other QSAR methods [78], such models may perform better on molecules similar to those of the training set.

Example of affinity prediction of 5-HT$_7$ antagonists (López-Rodríguez et al., 2000 and 2003)

In a short paper published in 2000, López-Rodríguez et al. reported the first pharmacophoric hypothesis for 5-HT$_7$ antagonists [79]. At that time, 30 compounds of different structures were used to derive Catalyst-HypoGen models. The best hypothesis was validated first by the goodness of SAR correlation (measures by r^2: 0.921) and then by the synthesis of a series of naphtholactams and naphthosultams. Only those that fitted the pharmacophore exhibited pK_i values above 6.5. In 2003, a more refined model was reported with a selection of 38 diverse antagonists with activities spanning five orders of magnitude [100]. The best model exhibited a poor correlation coefficient (0.74). However, after the removal of the non-selective ligands, the remaining 24 antagonists allowed a far better model to be obtained, showing excellent validation results: cost analysis, SAR correlation (r=0.91 and prediction within 1 log unit of the experimental pK_i). Database screening studies further suggested the synthesis of 34 new molecules to assess the predictive power of the retained model. Compounds that met the requirements of the pharmacophore were all actives and SAR modulations could be rationalized with the model. This definitely validated their model for further research.

Example of CYP heteroactivation prediction with pharmacophore models (Egnell et al., 2003)

Egnell et al published a series of two papers to report pharmacophore models capable of predicting CYP2C9 [80] and CYP3A4 [81] heteroactivation. Despite high structural variability, this group managed to build two suitable models capable of predicting clinical drug interactions. Accounting for structure-heteroactivation of these two isoforms, the retained models showed good internal predictivity and correlation between external test sets and observed *in vitro* positive cooperativity.

15.4
Case Study: a New Pharmacophore Model for mGlu4R Agonists

15.4.1
Metabotropic Glutamate Receptors as Potential Therapeutic Targets

The design of glutamate receptor agonists and antagonists has been the subject of many drug discovery programs as glutamate plays an essential role in both physiological and pathological processes in the CNS and its receptors are therefore important targets [82–85]. Glutamate (L-Glu) is the major excitatory neurotransmitter in the brain. It activates two classes of receptors: the ionotropic glutamate receptors (iGluRs) and the metabotropic glutamate receptors (mGluRs). The iGluRs are channel-gated receptors which mediate fast excitatory synaptic transmission [86] whereas mGluRs are G-protein coupled receptors (GPCRs) that modulate synaptic transmission [87]. The iGluRs are of three types named from their selective agonist: *N*-methyl-D-aspartic acid (NMDA), 2-amino-3-(3-hydroxy-5-methylisoxazol-4-yl)propionic acid (AMPA) and kainic acid (KA). Eight mGluRs were identified and classified in three groups according to their sequence similarity, transduction pathway and pharmacological profile. Group I includes mGlu1 and mGlu5 receptors which activate phospholipase C (PLC) and group II (mGlu2, mGlu3) and group III (mGlu4, mGlu6, mGlu7 and mGlu8) receptors inhibit AMP cyclase (AC).

Initial efforts were devoted to NMDA antagonists. As there were no 3D structures or models of the glutamate binding site available, pharmacophore models were generated and used for the design of new compounds [88, 89]. However because of the essential role of iGluR and the difficult fine tuning, all competitive drugs displayed severe side-effects and their development was discontinued. Interest was then focused on mGluRs, the other class of glutamate receptors. As mGluRs modulate synaptic transmission, they were expected to be better drug targets. Indeed, several mGluR agonists/antagonists proved to be successful and one was even taken to advanced phase II clinical assays [84]. Before the disclosure of homology models [90, 91] and the crystal structure [33] of the mGluR binding domain, pharmacophore models allowed the glutamate bioactive conformation and agonist selective features to be defined [31, 32]. Today, new mGluR ligands may be discovered through two complementary approaches: structure-based virtual screening [53] and ligand-based screening. Here, we report on the generation of a new pharmacophore model of mGlu4 receptor and its use in screening for new agonists.

15.4.2
Pharmacology of Metabotropic Glutamate Receptor Subtype 4 (mGlu4)

Known competitive mGlu4 ligands are all glutamate analogs of the linear, cyclic or phenylglycine type [87]. Agonists that were selected for the “training set” and “ROC set” to generate and validate the pharmacophore model, are displayed in Fig. 15.6.

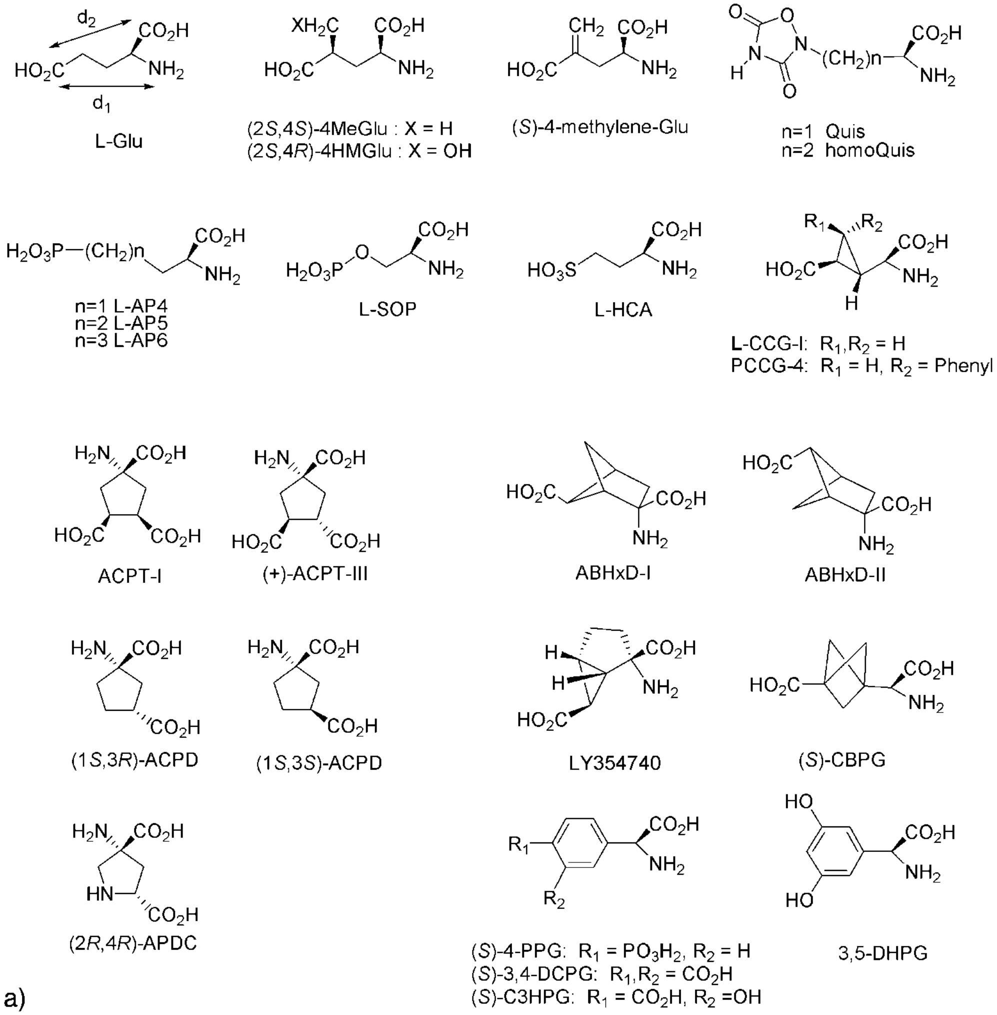


Fig. 15.6 (a) Chemical structure of mGlu4 agonists and other inactive compounds. Distances d_1 and d_2 are indicated on the L-glutamate structure (top left). Linear glutamate analogs are grouped at the top. The left bottom section displays cyclopentyl derivatives and the bottom right section shows rigid ligands (above) and phenylglycines (below).

They were chosen so that the largest range of activities is exhibited. Activities are expressed as the ratio between the agonist EC_{50} and the glutamate EC_{50} measured in the same assay as defined previously [32]. L-AP4 and most mGlu4 ligands hold an additional acidic function compared to glutamate that infer group III selectivity [e.g. L-AP4, L-SOP, ACPT, (*S*)-4-PPG, (*S*)-3,4-DCPG] (Fig. 15.6). The distal acidic groups of agonists were shown to interact with a

Name	Activity	Training set	ROC set
L-AP4	0.03		A
L-SOP	0.11	×	A
(*S*)-4-PPG	0.16	×	A
D-AP4	0.25	×	
ACPT-I	0.45	×	A
(2*S*)-CCG-I	0.52	×	A
(+)-ACPT-III	0.55	×	A
(*S*)-3,4-DCPG	0.88	×	A
L-Glu	1	×	A
ABHxD-I	2.3	×	A
L-HCA	4.9	×	A
L-AP5	8.4	×	
PCCG-4	10	×	
(2*S*,4*R*)-4HMGlu	25	×	A
L-AP6	35	×	
ABHxD-II	>100	×	
(*S*)-C3HPG	>100	×	
(*S*)-CBPG	>100	×	I
LY354740	>100	×	I
(S)-4methylene-Glu	>100	×	I
(2*S*,4*S*)-4MeGlu	>100		I
3,5-DHPG	>100		I
Quis	>100		I
homoQuis	>100		I
(1*S*,3*R*)-ACPD	>100		I
(1*S*,3*S*)-ACPD	>100		I
(2*R*,4*R*)-APDC	>100		I

b)

Fig. 15.6 (b) Activities of the mGlu4 agonists reported with L-glutamate (L-Glu) as a reference. Molecules used in the training set for Catalyst-HypoGen are flagged with an "×". Those taken to plot the ROC curves are flagged either as actives (A) and inactives (I).

set of lysines and arginines in a homology model of mGlu4 binding domain [91] (Fig. 15.7 b).

The flexibility of the side-chain of these residues allows their basic function to interact with the acidic groups of agonists with variable length. In the previous pharmacophore model, we did not include agonists that hold their distal acidic groups at longer d_1, d_2 distances (Fig. 15.6) from the α-amino acid moiety than glutamate. The reason for this was that we did not know which of the proximal or distal ionizable functions should be superimposed. It is now established that all amino acid mGlu ligands bind that moiety similarly to a common set of residues [92] (Fig. 15.7 b). Hence it is now possible to include in the model agonists such as (*S*)-4-PPG and (*S*)-3,4-DCPG that are characterized by longer d_1, d_2 distances.

With the previous pharmacophore model of mGlu4 receptor, we wished to determine the glutamate bioactive conformation, as the tertiary structure of the binding site was unknown, and to define selective features in comparison with mGlu1R and mGlu2R pharmacophore models. With the present model, we also intended to explain why some closely structurally related ligands bind selectively to other mGlu subtypes and not to mGlu4. These are found among the selected inactive compounds (ACPD, APDC, LY354740, CBPG, C3HPG) (Fig. 15.6).

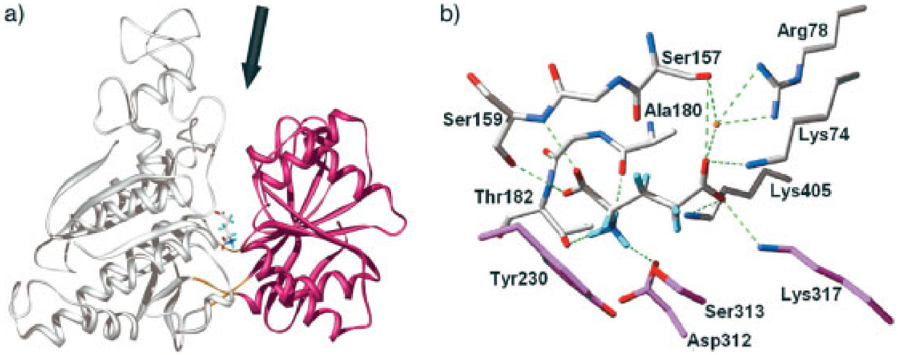


Fig. 15.7 (a) Homology model of the ligand binding domain of the mGlu4 receptor. This domain adopts a bilobate fold (shown as a ribbon, lobe 1 in white and lobe 2 in magenta) separated by a flexible hinge region (orange strands). The L-glutamate is displayed in the center in stick mode as it is trapped by the closure of the domain upon receptor activation. (b) Interaction pattern obtained with L-glutamate after docking by molecular dynamics in a model the mGlu4 receptor [top view along the black arrow in (a)]. Only residues that are in the vicinity of the ligand (center with hydrogens in cyan) are displayed. Residues are shown with colored carbons either in white (lobe 1) or magenta (lobe 2). Hydrogen bonds are shown with dashed green lines and a putatively "structural" water molecule is represented by a yellow sphere.

15.4.3
Training Set Elaboration

In this case study, the training set elaboration is utterly important because the mGlu4 agonists dataset is amongst the difficult ones (poor activities, inaccurate EC_{50} values for the least active compounds, poor structural variability, noticeable flexibility of the receptor, etc.). We report here how the training set was built to be used by Catalyst-HypoGen.

For HypoGen, activity values are specified with an uncertainty factor to account for biological variability. Hence this pharmacophore search engine will consider activity brackets for each compound instead of a sharp discrete value. This influences the "constructive phase" during which the pharmacophore space is generated with the most active compounds. Indeed, the set of "most active" molecules is defined as the subset of molecules for which the activity bracket overlaps the activity bracket of most active compound (see the "activity window", Fig. 15.8).

Although L-AP4 is the most potent agonist, this molecule was discarded from the training set in order to allow more actives in the construction phase of HypoGen. Hence L-SOP was the lead compound. Importantly, it shares many pharmacophoric configurations with L-AP4, therefore such essential structural information was provided to the program. The second most active is (*S*)-4-PPG, the first representative of the phenylglycines. This again was done deliberately because HypoGen generates the pharmacophoric space as the intersection of all accessible pharmacophoric configurations of the two most active compounds. The pharma-

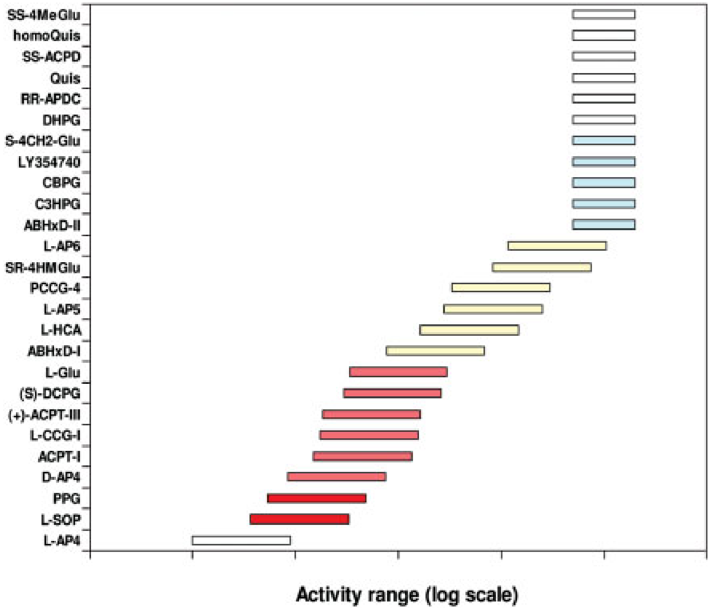


Fig. 15.8 Activity brackets used for our pharmacophore models. Compounds used in the training set have colored bars: red and light red for the "most active" set (constructive phase), yellow for moderately active molecules and cyan for the "least active" set (subtractive phase).

cophoric space is then reduced to the models that have a minimum of four features required to be found in the remaining "most active" compounds.

After the "constructive phase", the retained pharmacophore models are screened according to the mapping of the "least active" molecules. A molecule is taken as "least active" if its activity values differ by more than 3.5 log units from the activity of the most active compound (here L-SOP). For this particular dataset, the default value was too large and was therefore reduced to 2.9 to allow more molecules in the "least active" set. Only pharmacophore models that roughly discriminate between the most actives and the least actives will survive to this "subtractive phase".

15.4.4
Strategy for Perceiving the Pharmacophore

The default Catalyst features were edited to account for some important information regarding the interaction pattern exhibited by the best agonists in the orthosteric site of the mGlu4 receptor. Directed mutagenesis, resolved complex structures and homology models revealed the importance of the proximal part of the binding pocket. Indeed, this area is particularly structured, almost rigid, as many protruding side-chains are maintained by other residues via H-bonds. One of our main hypotheses in this search for new agonists is to keep the primary ammonium moiety. In fact, 3D models have shown that the three hydrogens of this moiety participate in three different H-bonds with the surrounding residues of the proximal zone, namely Ala180, Thr182 and Asp312, and consequently three polar hydrogens are required to satisfy these interactions (Fig. 15.7 b). Moreover, an ionic bridge with Asp312 is observed together with a cation-π-interaction with Tyr230, making the positive charge of the ammonium moiety a second constraint in the design of mGluR agonists. Consequently, the Positive Ionizable feature was required for the output pharmacophore models. Its presence had to be forced otherwise HypoGen would not have retained it as not being essential to discriminate between actives and inactives. This is due to our imperfect dataset in which even the least active molecules exhibit a primary ammonium.

The distal part of the binding site is flexible with residues featuring longer side-chains (Lys, Arg). It is this relative flexibility that allows molecules such as phenylglycines to bind in spite of their longer proximal–distal distance. We therefore edited the default HB Acceptor feature to relax the position of the acceptor atoms. This was simply done by deleting the foot point of the feature, therefore retaining only the projected point. Trials with the complete vectorized feature and with the foot point (location of H-bond acceptor atom, non directional) were also performed.

The Negative Ionizable feature was augmented to map on *N*-acylamides (O=C–NH–C=O) motif such as in Quis (Fig. 15.6 a).

Last, the default minimum feature distance was reduced to 230 pm (2.3 Å) owing to the relatively small size of the agonists.

15.4.5 Four Criteria to Validate our Pharmacophore Model

Our objective is to build a model to be used for virtual screening of commercial databases to identify novel mGlu4 receptor agonists.

Our validation criteria to retain output models are defined as follows:

How did we like the generated hypotheses?

1. Knowing the key role of the proximal part of the binding pocket, the hypotheses that aligned the amino acid moieties of the most actives in a similar way were retained. In particular, small and flexible agonists (such as L-AP4 or L-Glu), bulky agonists [such as ACPT-I and (+)-ACPT-III] and phenylglycines [(*S*)-4-PPG and (*S*)–3,4-DCPG]) should have similar mappings for their amino acid moieties. Should some models meet this criterion, they will represent a clear improvement compared with those previously generated in which phenylglycines could not be taken into account.

How well known SAR was provided and interpreted by the program?

2. Regarding the statistics reporting the SAR data, we were rather lenient since accurate activity prediction was not our objective. Hence, models exhibiting a

root mean square error [*RMS*, see Eq. (3.3)] below 1.5 log activity and a correlation coefficient above 0.8 were accepted.

How do the models comply with external biophysical data?

3. The three-dimensional structure of the ligand-binding domain constitutes external information that can be exploited to bring further credit to the generated model. In this respect, hypotheses that could be rationalized from the target structure were favored.

Considering known SAR, how good are the models for virtual screening applications?

4. The final decision was made according to the ROC curves plotted using a representative set of molecules to evaluate the performance for virtual screening (measured by the *AUC*) and to set a selection threshold. Models capable of exhibiting an *AUC* above 0.85 were retained. In order to be consistent with our previously reported virtual screening work resorting to a docking-scoring approach [53], the same set of 21 molecules was used. Figure 15.6 b recalls them and specifies those which were taken as actives (A) and those considered as inactives (I).

15.4.6 Results of Our Pharmacophore Model Research with Catalyst-HypoGen and HypoRefine

Different parameter sets were tested in a first row in order to restrict our pharmacophore search in a smaller space.

In an initial run, two different H-bond acceptor features (foot point and projected points; see Section 15.4.4) were selected in order to evaluate which of these two definitions was retained to build the best models. As expected, the H-bond acceptor feature defined as a simple projected point constraint was clearly favored. Even if more complex models allowing variable tolerances and variable weights were required, the foot point feature was never as good as the projected point (data not shown). Consequently, the H-bond acceptor feature defined as a foot point was discarded from further pharmacophore model research.

Table 15.1 reports the validation criteria that are met by the models obtained with the projected acceptor feature. Following the Occam razor principle implemented in HypoGen's cost function, we first tried to build simple models. The standard method uses a simulated annealing algorithm to optimize the retained models. During this process, features are added, moved or removed and eventually changed into a different feature in order to reduce the total cost of a given model. Unfortunately, the dataset did not allow HypoGen to produce simple models, the best hypothesis having an *RMS* of 1.63 (run 1). Consequently, different refinement methods were tested with various combinations: variable tolerances allow HypoGen to change the tolerance of the location constraints of each retained feature; similarly, variable weights allow the program to assign different weighting coefficients to each feature; Finally, the new Catalyst-HypoR-

Table 15.1 Summary of our pharmacophore search using various refinement methods[a)].

Run	Algorithm	Refinement method	Model 1	Model 2	Model 3
1	HypoGen	Standard	④	①③	①③④
2		Var. weights	②	④	③
3		Var. tolerances	④	④	①④
4		Var. tolerances and weights	②④	Null	④
5	HypoRefine	Standard	②	②③④	①④
6		Var. weights	①③④	①③④	③④
7		Var. tolerances	②③	③	④
8		Var. tolerances and weights	①②③④	Null	②

[a)] Variable weights and/or tolerances between features ("Var. weights" and "Var. tolerances" in the table, respectively) and the new HypoRefine algorithm to add excluded volumes were tested in various combinations. Cell shading indicates whether the three (cost-wise) best models meet (gray) or do not meet (white) criteria numbers ① (compound mappings), ② (*RMS* <1.5 and correlation coefficient $r>0.8$), ③ (from target rationalization) and ④ (ROC curve *AUC* >0.85).

efine algorithm [9] was tested to allow excluded volumes to be added to the models. With this last method, the "least active" molecules are not used to identify the most discriminating models but to find candidate locations to place excluded volumes during the optimization phase.

Interestingly, the generated models have lower costs with the new HypoRefine algorithm than with the standard HypoGen method (data not shown). We can conclude that Catalyst can better explain the differences in activities between the molecules of our dataset by resorting to excluded volumes than by simply using feature mappings. This observation backs up our previous statement regarding high structural similarities between the most active and the least active compounds of this tricky dataset.

Only a synergetic combination of different refinement methods could produce a model capable of meeting our four criteria (model 1, run 8; $RMS=0.76$, $r=0.96$, $AUC=0.87$). Rare are the cases for which Catalyst needs to resort to such extreme methods to obtain good models. This, again, underlines the difficulties with the mGlu4 dataset.

In addition to this pharmacophore hypothesis, although it met only three of the four criteria, model 1 from run 6 was retained. Surprisingly, despite criterion number 2 not being satisfied ($RMS=1.62$, $r=0.79$), this model exhibits a remarkable ability to discriminate between active and inactive compounds as assessed by the ROC curve, $AUC=0.95$. In contrast, model 1 from run 8 has good statistics ($RMS=0.76$, $r=0.96$) but a lower *AUC* of 0.87. This illustrates that a good model for activity prediction may not be the best for virtual screening applications. Let us analyze these two pharmacophore hypotheses further.

15.4.7 Description of the Two Retained Pharmacophore Models

15.4.7.1 Hypothesis 1 (Catalyst-HypoRefine with Variable Weights)

This model includes four projected points of H-bond acceptor features, one positive ionizable volume (as required) and one excluded volume. Figure 15.9 shows the mappings of three representative agonists: a small and flexible molecule (L-AP4, the most potent agonist known to date), the cyclic glutamate mimic ACPT-I and a phenylglycine, (*S*)-4-PPG.

First, the retained bioactive conformation corresponds to the extended conformation that was originally suggested by Jullian, Bessis and co-workers from the first pharmacophore models of mGlu receptor agonists [31, 32] and the later observed conformation in the resolved L-Glu/mGlu1 receptor complex by Kunishima et al. [33].

Interestingly, the excluded volume lies in the forbidden area represented by lobe 1 of the ligand binding domain (EV1, Fig. 15.9). This is in line with the commonly accepted mechanism of binding, which states that agonists must

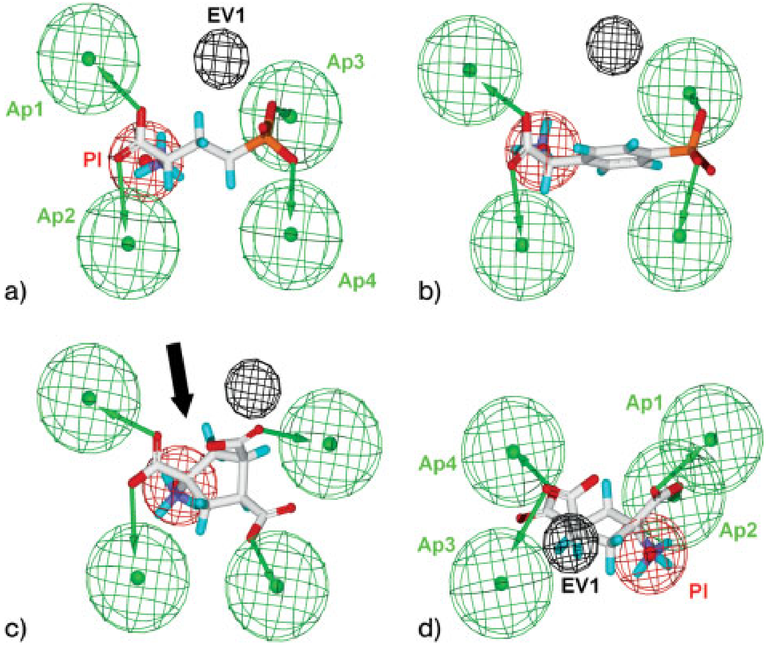


Fig. 15.9 Mapping of three agonists on hypothesis 1 [(a) L-AP4, (b) (*S*)-4-PPG and (c) ACPT-I, all top views, and (d) ACPT-I, side view along the black arrow in (c)]. The green spheres represent the tolerances surrounding projected points of H-bond acceptor features (Ap1–4) and the green vectors show which atom of the compound map to each of those feature; the red sphere locates a positive ionizable feature (basic group, PI) and the black volume is the sterically forbidden area (excluded volume, EV1).

first bind to the first lobe to allow an equilibrium displacement from an open-inactive conformation to a closed-activated state of the receptor.

As far as the projected points of H-bond acceptor features are concerned, it is not reasonable to propose a correspondence with some of the key residues identified by mutagenesis studies. Indeed, there are no agonists which contain a rigidified structure capable of constraining the proximal and distal acidic moieties in a specific orientation. In the case of L-AP4 for instance, the acidic functions (carboxylic and phosphonic) are free to rotate about the Cα–CO_2H and Cγ–PO_3H_2 bonds, therefore not providing any explicit direction for the acceptor features. It is worth noting, however, that the program perceived the importance of such interactions for the activity.

15.4.7.2 Hypothesis 2 (Catalyst-HypoRefine with Variable Weights and Tolerances)

Similarly to hypothesis 1, this second hypothesis comprises four projected points of H-bond acceptor features, one positive ionizable, but this time two excluded volumes were retained during the optimization phase (Fig. 15.10).

It is noticeable that allowing tolerances to vary led HypoRefine to reduce tolerance spheres significantly. These more stringent location constraints reduce the number of alternate mappings, therefore allowing a model with good statistics

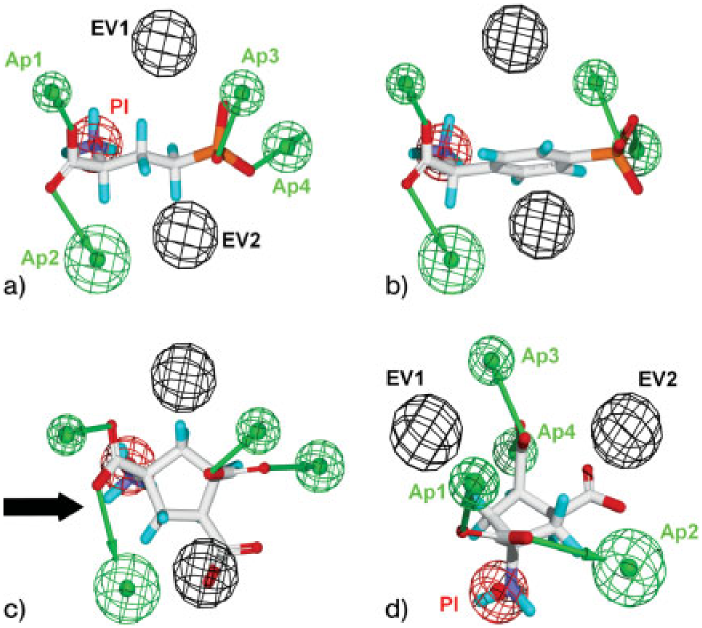


Fig. 15.10 Mapping of three agonists [(a) L-AP4, (b) (*S*)-4-PPG, (c) ACPT-I top views and (d) ACPT-I, front view along black arrow] on hypothesis 2. The color legend is as in Fig. 15.9.

to be obtained as mentioned above. Once again, the known extended conformation was chosen for small and flexible agonists such as L-Glu and L-AP4 (Fig. 15.10a). More remarkable is the second forbidden area that is added to hypothesis 2. Whereas a first excluded volume mimics the lobe 1 of the ligand binding domain, the second clearly overlaps with lobe 2, forcing the compounds to map in a narrow groove flanked by these two forbidden regions. Knowing that, by design, excluded volumes are added by Catalyst-HypoRefine to discriminate compounds with different activities better, it is noteworthy that the program was capable of perceiving such sterically forbidden zones by reporting solely on the ligands structures. In fact, it seems that the activation of the receptor depends on the closing angle of the ligand binding domain. Hence, using mGlu8 receptor as a model, we studied the mutation of the tyrosine-227 from lobe 2 (conserved in all subtypes and corresponding to Tyr230 in mGlu4; Fig. 15.7b) into alanine [93]. The results clearly suggest that this residue participates in the antagonist activity of α-methyl-AP4, whereas its "nor" derivative (L-AP4; Fig. 15.6a) is a full agonist: steric hindrance between the methyl group and the side-chain of Tyr230 might prevent the receptor from fully closing and reaching the conformation required for proper activation.

15.4.7.3 Comparison of the Two Retained Hypotheses

Although rather similar in their composition, hypotheses 1 and 2 would probably show different performances depending on their use. The preliminary statistics are clearly in favor of hypothesis 2 (see Fig. 15.11, top), making this model a good candidate for activity prediction. A more thorough validation would be required if it were to be used for this purpose. In particular, further assessment with compounds external to the training set would be necessary.

Our goal here is different as it is to identify novel mGlu4 agonists by virtual screening. Consequently, the ROC curve assessment was the core of the validation process (Fig. 15.11, bottom). Our evaluation set being rather small (only 21 representative molecules), we first evaluated the odds of obtaining better *AUC*s at random. To do that, the observed activity values were scrambled and randomly redistributed to each of the molecules 99 times. The 99 corresponding ROC curves were then plotted to measure 99 *AUC*s. Given that none of those *AUC* values could exceed the reference *AUC*s (0.95 for hypothesis 1 and 0.87 for hypothesis 2), Fisher's test indicates that the statistical significance of our models reaches 99%.

In a second row, when the *AUC*s of the two retained pharmacophore models are compared, hypothesis 1 appears to perform better than hypothesis 2. Not only is its *AUC* larger when compared with hypothesis 2, but also its ROC curve shape is more interesting. Hence, if sensitivity is to be maximized to reduce the false negative rate and increase the chances of finding novel leads ($Se=1$), it is possible to increase the specificity to 0.73 with hypothesis 1 (point S_1, Fig. 15.11), whereas the highest possible specificity value for $Se=1$ is only 0.36 with hypothesis 2 (point S_2, Fig. 15.11). In other words, if we analyze the

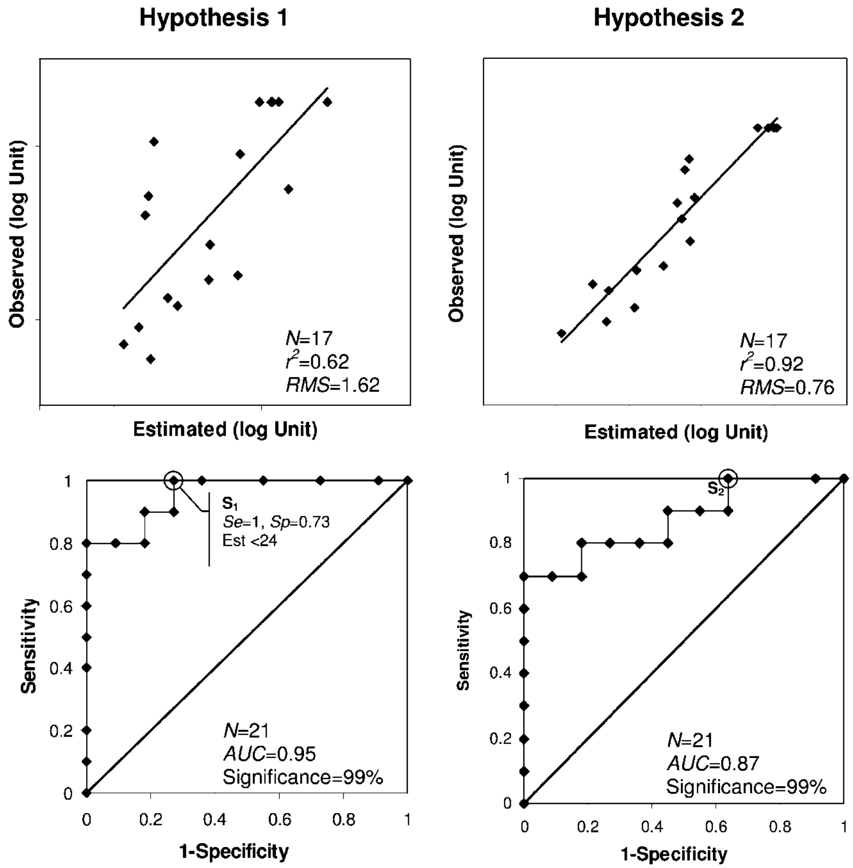


Fig. 15.11 Statistical performances (above) and ROC curves (below) of hypothesis 1 (left) and hypothesis 2 (right). The 17 compounds used for the statistical assessment are those from the training set whereas the 21 compounds used to plot the ROC curves were taken from a representative set of actives and inactive agonists (see Fig. 15.6b and text). Each straight line shows the ROC curve of a random classification and can be used as a reference.

binary classification between active and inactive, only three molecules out of 21 are misclassified with hypothesis 1 (namely 4-methylene-Glu, Quis and homo-Quis, which are declared as actives). In contrast, hypothesis 2 over-predicts seven compounds out of the 21 (i.e. ca. 63% of the inactives are found amongst the selected molecules).

In conclusion, hypothesis 1 was selected because it satisfies three of our four validation criteria and seems particularly appropriate for virtual screening (ROC curve validation). The selection threshold was set according to point S_1 (Fig. 15.11), corresponding to an activity threshold of 24 (i.e. estimated activity below 24 times the activity of L-Glu).

15.4.8
Further Validation: Virtual Screening of the CAP Database

The CAP (Chemicals Available for Purchase) database is a compilation of commercially available compounds from all major vendors. This electronic catalogue is maintained by Accelrys Inc. in several formats [94] and contains more than 1.6 million unique entries.

The virtual screening of the CAP database using hypothesis 1 as a query was organized in several rounds. The first round was focused on unprotected *α*-amino acids of low molecular weight (below 300 g mol^{-1}) because this subset is expected to contain more hits with possibilities of optimizing them. Some results of this first virtual screening round are reported in Table 15.2.

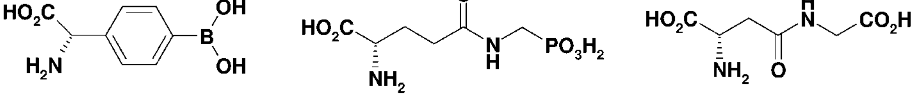
An initial screen with hypothesis 1 yielded 251 amino acids of molecular weight below 300 g mol^{-1} and among which 156 were predicted to be active on mGlu4R (estimated activity <24). A quick visual inspection of these 156 virtual hits has revealed 28 known and commercially available mGlu4 agonists (e.g. L-Glu, L-AP4, *α*-methyl-AP4, *β*-methyl-AP4, PCCG-4). To our knowledge the 128 remaining hits have not been tested on this receptor. Figure 15.12 reports some interesting structures.

The phenylboronic derivative **1** is particularly interesting as it might contract a dative bond with one of the distal lysine residues. This could noticeably reduce the k_{off} constant upon binding to mGlu4 and provide a new pharmacological tool to study this receptor (especially if the lysine residue involved is in lobe 2, i.e. Lys317). Other series of molecules that we have selected for *in vitro* tests

Table 15.2 Virtual screening of the CAP database using successive queries[a)].

Query	Query description	Number of compounds	Maximum enrichment
0	None	1619612	1
1	Unprotected *α*-amino acid	788	2055
2	Query 1 and hypothesis 1	251	6452
3	Query 2 and estimated activity <24	156	10382

[a)] For each step, the number of virtual hits is reported together with the maximum possible enrichment [enrichment calculated with Eq. (4.4) in the case where all selected compounds are active].



1 **2** **3**

Fig. 15.12 Structures of three virtual hits identified with hypothesis 1 in the CAP database.

include γ-glutamyl and γ-aspartyl derivatives such as **2** and **3**. Indeed, many new chemical derivatives could be envisaged synthetically by simply coupling primary amines to the distal acidic function of L-Glu or alternatively l-Asp. *In vitro* assays performed with these virtual hits will be reported elsewhere together with non-amino acid hits identified from other screening rounds.

15.5 Conclusion

The success of the pharmacophore approach in drug discovery no longer needs to be demonstrated. This book is an additional testimony to this. Successful pharmacophore investigations, however, have unavoidably passed a key step prior to application: validation. What we have tried to show throughout this chapter is that this critical stage depends on three practical considerations.

First, the available dataset is the foremost limitation on pharmacophore research. Ideally, it should contain compounds of various structures and activities with evidence that they bind to the same binding site. Potent, feature-poor and conformationally constrained compounds greatly help in reducing the number of solutions. In reality, however, rare are the cases that meet all these criteria together. Nevertheless, this does not mean that pharmacophore research is unthinkable for difficult cases.

Second, the approach used to perceive the pharmacophore will add further limitations. We have discussed the pitfalls of ligand-based and structure-based pharmacophore models and the different approaches used by automatic pharmacophore generation programs. Simplifications are inevitable and one cannot pretend to capture fully the complexity of a biological phenomenon with a set of 5–6 features. The validation process is therefore focused on showing that, for a given dataset, the generated model describes some of its important aspects. In particular, showing that the model explains SAR data is critical.

Third, the validation is often adapted to the final objective in the search for a pharmacophore. Whether its is to be used as a query for virtual screening, to predict accurately the activity of a series of molecule or to serve as a guide for drug design, some validation approaches can be emphasized over others. So, if the title of this chapter were to be rephrased, we could have asked: *Are you sure you have the right model … for the envisaged application?*

Last, we have searched for a new pharmacophore model of metabotropic glutamate receptor subtype 4 agonists. The retained hypothesis, generated via the new Catalyst-HypoRefine algorithm, showed significant advantages over the previously reported models by including both glutamate-like agonists and phenylglycines. It was validated for virtual screening applications by the ROC curve method and indexed database screening. Finally, the model was used as a query approach to search for new mGlu4 receptor agonists amongst a compilation of chemical vendors' catalogues. *In vitro* results of the retained virtual hits will be reported elsewhere.

Acknowledgments

We would like to thank Dr. Rémy Hoffmann (Accelrys SARL, France) for his support in the generation of the new mGlu4 pharmacophore model and his insightful remarks on the manuscript. This work was supported by the Agence Nationale de la Recherche Technique and Accelrys, Inc. (CIFRE grant No. 738/2002).

References

1 P. Ehrlich, *Ber. Dtsch. Chem. Ges.* **1909**, *42*, 17–47.

2 P. Gund, in *Progress in Molecular and Subcellular Biology*, F. E. Hahn (ed.). Springer, Berlin, **1977**, pp. 117–143.

3 J.S. Mason, A.C. Good, E.J. Martin, *Curr. Pharm. Des.* **2001**, *7*, 567–597.

4 C.G. Wermuth, C.R. Ganellin, P. Lindberg, L.A. Mitscher, *Pure Appl. Chem.* **1998**, *70*, 1129–1143.

5 H. van de Waterbeend, R.E. Carter, G. Grassy, H. Kubinyi, Y.C. Martin, M.S. Tute, P. Willett, *Pure Appl. Chem.* **1997**, *69*, 1137–1152.

6 J.H. Van Drie, in *Pharmacophore Perception, Development and Use in Drug design*, O.F. Güner (ed.), International University Line, La Jolla, CA, **2000**, pp. 50–68.

7 J.H. Van Drie, in *Computational Medicinal Chemistry for Drug Discovery*, P. Bultinck, H. DeWinter, W. Langenaeker and J. P. Tollenaere (eds.). Marcel Dekker, New York, **2004**, pp. 437–460.

8 S.J. Teague, *Nat. Rev. Drug Discov.* **2003**, *2*, 527–541.

9 Accelrys. *Catalyst*®. Accelrys, San Diego, CA; www.accelrys.com.

10 D. Barnum, J. Greene, A. Smellie, P. Sprague, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563–571.

11 H. Li, J. Sutter, R. Hoffmann, in *Pharmacophore Perception, Development and Use in Drug Design*, O.F. Güner (ed.). La Jolla CA, **2000**, pp. 171–189.

12 Tripos. *DISCO (DIStance COmparisons)*, available as DISCOtech^{TM} from Tripos, Inc., St. Louis, MO, with additional implemented functionalities; www.tripos.com.

13 Y.C. Martin, E. Dahaner, J. De Lazzer, I. Lico, P.A. Pavlik, *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83–102.

14 Y.C. Martin, in *Pharmacophore Perception, Development and Use in Drug Design*, O. F. Güner (ed). La Jolla, CA, **2000**, pp. 49–68.

15 Tripos. *GASP^{TM} (Genetic Algorithm Similarity Program)*. Tripos, Inc., St. Louis, MO; www.tripos.com.

16 G. Jones, P. Willett, R.C. Glen, *J. Comp.-Aided Mol. Des.* **1995**, *9*, 532–549.

17 Schrödinger. *Phase*. Schrödinger, LLC, New York; www.schrodinger.com.

18 R. Dayam, T. Sanchez, O. Clement, R. Shoemaker, S. Sei, N. Neamati, *J. Med. Chem.* **2005**, *48*, 111–120.

19 Internal randomization studies performed by Accelrys, Inc., San Diego, CA.

20 Computer Consulting Group. *MOE^{TM} (Molecular Operating Environment)*. Computer Consulting Group, Montreal.

21 Accelrys. *DS ViewerPro*. Accelrys, Inc., San Diego, CA.

22 G. Wolber, T. Langer, *J. Chem. Inf. Model.* **2005**, *45*, 160–169.

23 Inte:Ligand. *LigandScout*. Inte:Ligand, GmbH, Maria Enzerdorf, Austria; www.inteligand.com.

24 C.M. Venkatachalam, P. Kirchhoff, M. Waldman, in *Pharmacophore Perception, Development and Use in Drug Design*, O. F. Güner (ed.). La Jolla, CA, **2000**, pp. 339–350.

25 H.A. Carlson, K.M. Masukawa, K. Rubins, F.D. Bushman, W.L. Jorgensen, R.D. Lins, J.M. Briggs, J.A. McCammon, *J. Med. Chem.* **2000**, *43*, 2100–2114.

26 H.A. Carlson, K.M. Masukawa, J.A. McCammon, *J. Phys. Chem. A* **1999**, *103*, 10213–10219.

27 R.D. Cramer III, D.E. Patterson, J.D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

28 G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130–4146.

29 R.A. Glennon, S.Y. Ablordeppey, A.M. Ismaie, M.B. El-Ashmawy, J.B. Fischer, K.B. Howie, *J. Med. Chem.* **1994**, *37*, 1214–1219.

30 R. Barbaro, L. Betti, M. Botta, F. Corelli, G. Giannaccini, L. Maccari, F. Manetti, G. Strappaghetti, S. Corsano, *Bioorg. Med. Chem.* **2002**, *10*, 361–369.

31 N. Jullian, I. Brabet, J.-P. Pin, F.C. Acher, *J. Med. Chem.* **1999**, *42*, 1564–1555.

32 A.-S. Bessis, N. Jullian, E. Coudert, J.-P. Pin, F. Acher, *Neuropharmacology* **1999**, *38*, 1543–1551.

33 N. Kunishima, Y. Shimada, Y. Tsuji, T. Sato, M. Yamamoto, T. Kumasaka, S. Nakanishi, H. Jingami, K. Morikawa, *Nature* **2000**, *407*, 971.

34 *Apex-3D.* Originally developed by V. Golender, B. Vesterman and E. Vorpagel, this software is no longer distributed.

35 D.J. Livingstone, *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. Oxford University Press, New York, **1995**.

36 H. Kubinyi, *Drug Discov. Today* **1997**, *2*, 538–546.

37 H. Kubinyi, *Drug Discov. Today* **1997**, *2*, 457–467.

38 D.M. Hawkins, S.C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.

39 D.M. Hawkins, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

40 A.M. Doweyko, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587–596.

41 R. Guha, P.C. Jurs, *J. Chem. Inf. Model.* **2005**, *45*, 65–73.

42 H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

43 R. Brenk, L. Naerum, U. Grädler, H.-D. Gerber, G.A. Garcia, K. Reuter, M.T. Stubbs, G. Klebe, *J. Med. Chem.* **2003**, *46*, 1133–1143.

44 Tripos. *SYBYL*®. Tripos, Inc., St. Louis, MO.

45 A.K. Ghose, J.J. Wendoloski, *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 253–271.

46 R.M. De Marinis, M. Wise, J.P. Hieble, R.R. Ruffolo Jr., R.R. Ruffolo Jr. in *The Alpha-1 Adrenergic Receptors*, R.R. Ruffolo Jr. (ed). Humana Press, Clifton, NJ, **1987**, pp. 211–265.

47 C. Laggner, C. Schieferer, B. Fiechtner, G. Poles, R.D. Hoffmann, H. Glossmann, T. Langer, F.F. Moebius, *J. Med. Chem.* **2005**, *48*, 4754–4764.

48 H. Kubinyi, F. A. Hamprecht, T. Mietzner, *J. Med. Chem.* **1998**, *41*, 2553–2564.

49 P. Willett, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

50 R. Wang, Y. Lu, X. Fang, S. Wang, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.

51 A.K. Debnath, *Mini Rev. Med. Chem.* **2001**, *1*, 187–195.

52 M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T.M. Mooij, C.W. Murray, R.D. Taylor, P. Watson, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.

53 N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, *J. Med. Chem.* **2005**, *48*, 2534–2547.

54 Thompson Derwent. *Thompson Derwent World Drug Index*; scientific.thomson.com/products/wdi.

55 KEGG. *KEGG Kyoto Encyclopedia of Genes and Genomes*; www.kegg.com.

56 T.M. Frimurer, R. Bywater, L. Nrum, L.N. Lauritsen, S. Brunak, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315–1324.

57 A. Goldblum, personal communication, **2005**.

58 S. Yoon, W.J. Welsh, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.

59 O.F. Güner, D.R. Henry, in *Pharmacophore Perception, Development and Use in Drug Design*, O. F. Güner (ed.). La Jolla CA, **2000**, pp. 193–212.

60 M. Jacobsson, P. Lidén, E. Stjernschantz, H. Boström, U. Norinder, *J. Med. Chem.* **2003**, *46*, 5781–5789.

61 E.A. Hecker, C. Duraiswami, T.A. Andrea, D.J. Diller, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1204–1211.

62 D.J. Diller, R. Li, *J. Med. Chem.* **2003**, *46*, 4638–4647.

63 D.J. Diller, J. Kenneth M. Merz, *Proteins: Struct. Funct. Genet.* **2001**, *43*, 113–124.

64 J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, B. Scholkopf, *Bioinformatics* **2003**, *19*, 764–771.

65 A.J. Shepherd, D. Gorse, J.M. Thornton, *Protein Sci.* **1999**, *8*, 1045–1055.

66 H. Gao, C. Williams, P. Labute, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164–168.

67 M.G. Ford, presented at the 2nd CMTPI, Thessaloniki, Greece, **2003**.

68 E.K. Bradley, J.L. Miller, E. Saiah, P.D.J. Grootenhuis, *J. Med. Chem.* **2003**, *46*, 4360–4364.

69 E.K. Bradley, P. Beroza, J.E. Penzotti, P.D.J. Grootenhuis, D.C. Spellmeyer, J.L. Miller, *J. Med. Chem.* **2000**, *43*, 2770–2774.

70 B. Matthews, *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

71 E. Martineau, A.M. Aman, X. Kong, in *Accelrysworld*. Accelrys, Inc., San Diego, CA, **2004**.

72 J.S. Mason, D.L. Cheney, *Pac. Symp. Biocomput.* **2000**, 576–587.

73 J.E. Eksterowicz, E. Evensen, C. Lemmen, G.P. Brady, J.K. Lanctot, E.K. Bradley, E. Saiah, L.A. Robinson, P.D.J. Grootenhuis, J.M. Blaney, *J. Mol. Graph. Model.* **2002**, *20*, 469–477.

74 R. Barbaro, L. Betti, M. Botta, F. Corelli, G. Giannaccini, L. Maccari, F. Manetti, G. Strappaghetti, S. Corsano, *J. Med. Chem.* **2001**, *44*, 2118–2132.

75 G. Zlokarnik, P.D.J.Grootenhuis, J.B.Watson, *Drug Discov. Today* **2005**, *10*, 1443–1450.

76 H. Yu, A. Adedoyin, *Drug Discov. Today* **2003**, *8*, 852–861.

77 U. Norinder, *SAR QSAR Environ. Res.* **2005**, *16*, 1–11.

78 R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

79 M.L. López-Rodríguez, E. Porras, B. Benhamú, J.A. Ramos, M.J. Morcillo, J.L. Lavandera, *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1097–1100.

80 A.-C. Egnell, C. Eriksson, N. Albertson, B. Houston, S. Boyer, *J. Pharmacol. Exp. Ther.* **2003**, *307*, 878–887.

81 A.-C. Egnell, J.B. Houston, C.S. Boyer, *J. Pharmacol. Exp. Ther.* **2005**, *312*, 926–937.

82 J.A. Kemp, R.M. McKernan, *Nat. Neurosci.* **2002**, *5*, 1039–1042.

83 M.J. Marino, O. Valenti, P. J. Conn, *Drugs Aging* **2003**, *20*, 377–397.

84 G. Marek, *Curr. Opin. Pharmacol.* **2004**, *4*, 18–22.

85 C.J. Swanson, M. Bures, M.P. Johnson, A.-M. Linden, J.A. Monn, D.D. Schoepp, *Nat. Rev. Drug Discov.* **2005**, *4*, 131–144.

86 R. Dingledine, K. Borges, D. Bowie, S.F. Traynelis, *Pharmacol. Rev.* **1999**, *51*, 7–61.

87 J.-P. Pin, F. Acher, *Curr. Drug Targets* **2002**, 297–317.

88 D.F. Ortwine, T.C. Malone, C.F. Bigge, J.T. Drummond, C. Humblet, G. Johnson, G.W. Pinter, *J. Med. Chem.* **1992**, *35*, 1345–1370.

89 J.P. Whitten, B.L. Harrison, H.J. Weintraub, I.A. McDonald, *J. Med. Chem.* **1992**, *35*, 1509–1514.

90 A.-S. Bessis, H.-O. Bertrand, T. Galvez, C.D. Colle, J.-P. Pin, F. Acher, *Protein Sci.* **2000**, *9*, 2200–2209.

91 H.-O. Bertrand, A.-S. Bessis, J.-P. Pin, F. Acher, *J. Med. Chem.* **2002**, *45*, 3171–3183.

92 F.C. Acher, H.-O. Bertrand, *Biopolymers* **2005**, *80*, 357–366.

93 A.-S. Bessis, P. Rondard, F. Gaven, I. Brabet, N. Triballeau, L. Prézeau, F. Acher, J.-P. Pin, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 11097–11102.

94 Accelrys. *CAP*. Available from Accelrys, Inc., San Diego, CA, in ADE, AEI and Catalyst formats; www.accelrys.com

95 A.K. Bhattacharjee, J.A. Geyer, C.L. Woodard, A.K. Kathcart, D.A. Nichols, S.T. Prigge, Z. Li, B.T. Mott, N.C. Waters, *J. Med. Chem.* **2004**, *47*, 5418–5426.

96 T.M. Steindl, C.E. Crump, F.G. Hayden, T. Langer, *J. Med. Chem.* **2005**, *48*, 6250–6260.

97 S. Sirois, D.-Q. Wei, Q. Du, K.-C. Chou, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1111-1122.

98 T. Klabunde, K.V. Wendt, D. Kadereit, V. Brachvogel, H.-J. Burger, A.W. Herling, N.G. Oikonomakos, M.N. Kosmopoulou, D. Schmoll, E. Sarubbi, E. von Roedern, K. Schörafinger, E. Defossa, *J. Med. Chem.* *2005*, *48*, 6178–6193.

99 L. Betti, M. Botta, F. Corelli, M. Floridi, G. Giannaccini, L. Marccari, F. Manetti, G. Strappaghetti, A. Tafi, S. Corsano, *J. Med. Chem.* **2002**, *45*, 3603–3611.

100 M.L. López-Rodríguez, F. Porras, M.J. Morcillo, B. Benhamú, L.J. Soto, J.L. Lavandera, J.A. Ramos, M. Olivella, M. Campillo, L. Pardo, *J. Med. Chem.* **2003**, *46*, 5628–5650.

Subject Index

Pharmacophores and Pharmacophore Searches.
Edited by Thierry Langer and Remy D. Hoffmann
Copyright © 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-31250-1