Chemoinformatics in Drug Discovery

Edited by
Tudor I. Oprea

WILEY-VCH

WILEY-VCH Verlag GmbH & Co. KGaA

Chemoinformatics in Drug Discovery

Edited by
Tudor I. Oprea

Methods and Principles in Medicinal Chemistry

Edited by R. Mannhold, H. Kubinyi, G. Folkers

Recently Published Volumes:

T. Lengauer (ed.)

Bioinformatics – From Genomes to Drugs
Vol. 14

2001, ISBN 3-527-29988-2

J. K. Seydel, M. Wiese

Drug-Membrane Interactions
Vol. 15

2002, ISBN 3-527-30427-4

O. Zerbe (ed.)

BioNMR in Drug Research
Vol. 16

2002, ISBN 3-527-30465-7

P. Carloni, F. Alber (eds.)

Quantum Medicinal Chemistry
Vol. 17

2003, ISBN 3-527-30456-8

H. van de Waterbeemd,
H. Lennernäs, P. Artursson (eds.)

Drug Bioavailability
Vol. 18

2003, ISBN 3-527-30438-X

H.-J. Böhm, G. Schneider (eds.)

Protein-Ligand Interactions
Vol. 19

2003, ISBN 3-527-30521-1

R. E. Babine, S. S. Abdel-Meguid (eds.)

Protein Crystallography in Drug Discovery
Vol. 20

2004, ISBN 3-527-30678-1

Th. Dingermann, D. Steinhilber,
G. Folkers (eds.)

Molecular Biology in Medicinal Chemistry
Vol. 21

2004, ISBN 3-527-30431-2

H. Kubinyi, G. Müller (eds.)

Chemogenomics in Drug Discovery

2004, ISBN 3-527-30987-X

Chemoinformatics in Drug Discovery

Edited by
Tudor I. Oprea

WILEY-VCH Verlag GmbH & Co. KGaA

Series Editors:

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

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

Prof. Dr. Gerd Folkers
Department of Applied Biosciences
ETH Zürich
Winterthurerstrasse 19
8057 Zürich
Switzerland
folkers@pharma.anbi.ethz.ch

Volume Editor:

Prof. Dr. Tudor I. Oprea
Division of Biocomputing
MSC08 4560
University of New Mexico
School of Medicine
Albuquerque, NM 87131
USA

This 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 Cataloging-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.

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

Composition Laserwords Private Ltd, Chennai, India
Printing betz-druck GmbH, Darmstadt
Bookbinding Litges & Dopf Buchbinderei GmbH, Heppenheim

ISBN-13: 978-3-527-30753-1
ISBN-10: 3-527-30753-2

Contents

A Personal Foreword

This volume brings together contributions from academic and industrial scientists who develop and apply chemoinformatics strategies and tools in drug discovery. From chemical inventory and compound registration to candidate drug nomination, chemoinformatics integrates data via computer-assisted manipulation of chemical structures. Linked with computational chemistry, physical (organic) chemistry, pharmacodynamics and pharmacokinetics, chemoinformatics provides unique capabilities in the areas of lead and drug discovery. This book aims to offer knowledge and practical insights into the use of chemoinformatics in preclinical research.

Divided in four sections, the book opens with a first-hand account from Garland Marshall, spanning four decades of chemoinformatics and pharmaceutical research and development. Part one sets the stage for virtual screening and lead discovery. Hit and lead discovery via *in silico* technologies are highlighted in part two. In part three, data collection and mining using chemical databases are discussed in the context of chemical libraries. Specific applications and examples are collected in part four, which brings together industrial and academic perspectives. The book concludes with another personal account by Don Abraham, who presents drug discovery from an academic perspective.

The progression hit identification → lead generation → lead optimization → candidate drug nomination is served by a variety of chemoinformatics tools and strategies, most of them supporting the decision-making process. Key procedures and steps, from virtual screening to *in silico* lead optimization and from compound acquisition to library design, underscore our progress in grasping the preclinical drug discovery process, its needs for novel technologies and for integrated informatics support. We now have the ability to identify novel chemotypes in a rational manner, and *in silico* methods are deep-rooted in the process of systematic discovery. Our increased knowledge in a variety of seemingly unrelated phenomena, from atomic level issues related to drug–receptor binding to bulk properties of drugs and pharmacokinetics profiling, is likely to lead us on a better path for the discovery of orally bioavailable drugs, at the same time paving the way for novel, unexpected therapeutics.

I want to acknowledge all the contributors who made this book possible. Their insights, examples and personal accounts move beyond the sometimes dry language of science, turning this volume into an interesting and fascinating book to read.

Finally, I thank Frank Weinreich and Hugo Kubinyi for their encouragement and timely pressure to prepare this book on time.

Albuquerque, January 2005 *Tudor I. Oprea*

Preface

The term "chemoinformatics" was introduced in 1998 by Dr. Frank K. Brown in the Annual Reports of Medicinal Chemistry. In his article "Chemoinformatics: What is it and How does it Impact Drug Discovery", he defines chemoinformatics as follows: "*The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization*".

In fact, Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information. Related terms of chemoinformatics are cheminformatics, chemi-informatics, chemometrics, computational chemistry, chemical informatics, and chemical information management/science.

Reflecting the above given definitions, the present volume on "Chemoinformatics in Drug Discovery"covers its most important aspects within four main sections. After an introduction to chemoinformatics in drug discovery by Garland Marshall, the first section is focused on *Virtual Screening*. T. Oprea describes the use of "Chemoinformatics in Lead Discovery" and M.M. Hann et al. deal with "Computational Chemistry, Molecular Complexity and Screening Set Design". Then, M. Rarey et al. review "Algorithmic Engines in Virtual Screening" and D. Horvath et al. review the "Strengths and Limitations of Pharmacophore-Based Virtual Screening". The next section is dedicated to *Hit and Lead Discovery* with chapters of I.J. McFadyen et al. on "Enhancing Hit Quality and Diversity Within Assay Throughput Constraints", of C.L. Cavallaro et al. on "Molecular Diversity in Lead Discovery", and of C. Ho on "In Silico Lead Optimization". Topics of the third section refer to *Databases and Libraries*. They include chapters on "WOMBAT: World of Molecular Bioactivity" by M. Olah et al., on "Cabinet – Chemical and Biological Informatics Network" by V. Povolna et al., on "Structure Modification in Chemical Databases" by P.W. Kenney and J. Sadowski, and on the "Rational Design of GPCR-specific Combinational Libraries Based on the Concept of Privileged Substructures" by N.P. Savchuk et al.

According to our intention, to provide in this series on "Methods and Principles in Medicinal Chemistry" practice-oriented monographs, the book closes with a section on *Chemoinformatics Applications*. These are exemplified by G.M. Maggiora et al. in a chapter on "A Practical Strategy for Directed Compound Acquisition", by

K.-H. Baringhaus and H. Matter on ‘‘Efficient Strategies for Lead Optimization by Simultaneously Addressing Affinity, Selectivity and Pharmacokinetic Parameters’’, by R.A. Goodnow et al. on ‘‘Chemoinformatic Tools for Library Design and the Hit-to-Lead Process’’ and by A. Tropsha on the ‘‘Application of Predictive QSAR Models to Database Mining’’. The section is concluded by a chapter of D.J. Abraham on ‘‘Drug Discovery from an Academic Perspective’’.

The series editors would like to thank Tudor Oprea for his enthusiasm to organize this volume and to work with such a fine selection of authors. We also want to express our gratitude to Frank Weinreich from Wiley-VCH for his valuable contributions to this project.

September 2004

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

List of Contributors

DONALD J. ABRAHAM
Department of Medicinal Chemistry
Virginia Commonwealth University
800 E. Leigh Street
Richmond, VA 23219-1540
USA

JUAN C. ALVAREZ
Chemical and Screening Sciences
Wyeth Research
200 Cambridge Park Drive
Cambridge, MA 02140
USA

KONSTANTIN V. BALAKIN
Chemical Diversity Labs, Inc.
Computational and Medicinal Chemistry
11575 Sorrento Valley Road
San Diego, CA 92121
USA

MAGDALENA BANDA
Romanian Academy
Institute of Chemistry
“Coriolan Dragulescu”
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

FRÉDÉRIQUE BARBOSA
Cerep S. A.
128, Rue Danton
92506 Rueil-Malmaison
France

KARL-HEINZ BARINGHAUS
Aventis Pharma Deutschland GmbH
DI&A Chemistry
Computational Chemistry
Industriepark Höchst, Bldg. G 878
65926 Frankfurt am Main
Germany

KONRAD BLEICHER
F. Hoffmann-La Roche Ltd.
PRBD-C
4070 Basel
Switzerland

ALINA BORA
Romanian Academy
Institute of Chemistry
“Coriolan Dragulescu”
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

CULLEN L. CAVALLARO
Pharmaceutical Research Institute
Bristol-Myers Squibb Co.
Province Line Road
Princeton, NJ 08540
USA

SCOTT DIXON
Metaphorics, LLC
441 Greg Avenue
Sante Fe, NM 87501
USA

THOMPSON N. DOMAN
Lilly Research Laboratories
Structural and Computational Sciences
Indianapolis, IN 46285
USA

PAUL GILLESPIE
Hoffmann-La Roche Inc.
Discovery Chemistry
340 Kingsland Street
Nutley, NJ 07110
USA

ROBERT A. GOODNOW, Jr.
Hoffmann-La Roche Inc.
New Leads Chemistry Initiative
340 Kingsland Street
Nutley, NJ 07110
USA

RAFAEL GOZALBES
Cerep S. A.
128, Rue Danton
92506 Rueil-Malmaison
France

DARREN V. S. GREEN
GlaxoSmithKline Research
and Development
Gunnels Wood Road
Stevenage, SG1 2NY
United Kingdom

NICOLETA HADARUGA
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

MICHAEL M. HANN
GlaxoSmithKline Research
and Development
Gunnels Wood Road
Stevenage, SG1 2NY
United Kingdom

CHRIS HO
Drug Design Methodologies, LLC
4355 Maryland Ave.
St. Louis, MO 63108
USA

DRAGOS HORVATH
UMR8525-CNRS
Institut de Biologie de Lille
1, rue Calmette
59027 Lille
France

PETER W. KENNY
AstraZeneca Mereside
Alderley Park
Macclesfield, SK10 4TG
United Kingdom

MICHAEL S. LAJINESS
Lilly Research Laboratories
Structural and Computational Sciences
Indianapolis, IN 46285
USA

ANDREW R. LEACH
GlaxoSmithKline Research
and Development
Gunnels Wood Road
Stevenage, SG1 2NY
United Kingdom

CHRISTIAN LEMMEN
BioSolveIT GmbH
An der Ziegelei 75
53757 Sankt Augustin
Germany

GERALD M. MAGGIORA
Dept. of Pharmacology and Toxicology
University of Arizona
College of Pharmacy
Tucson, AZ 85271
USA

BORYEU MAO
Cerep Inc.
15318 NE 95th Street
Redmond, WA 98052
USA

GARLAND R. MARSHALL
Center for Computational Biology
Washington University
School of Medicine
660 S. Euclid Ave.
St. Louis, MO 63110
USA

HANS MATTER
Aventis Pharma Deutschland GmbH
DI&A Chemistry
Computational Chemistry
Industriepark Höchst, Bldg. G 878
65926 Frankfurt am Main
Germany

IAIN MCFADYEN
Chemical and Screening Sciences
Wyeth Research
200 Cambridge Park Drive
Cambridge, MA 02140
USA

MARIA MRACEC
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

MIRCEA MRACEC
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

IONELA OLAH
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

MARIUS OLAH
Division of Biocomputing, MSC08 4560
University of New Mexico
School of Medicine
Albuquerque, NM 87131
USA

TUDOR I. OPREA
Division of Biocomputing, MSC08 4560
University of New Mexico
School of Medicine
Albuquerque, NM 87131
USA

LILIANA OSTOPOVICI
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

VERA POVOLNA
Metaphorics, LLC
441 Greg Avenue
Sante Fe, NM 87501
USA

RAMONA RAD
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

MATTHIAS RAREY
Center for Bioinformatics (ZBH)
University of Hamburg
Bundesstrasse 43
20146 Hamburg
Germany

SHERRY L. ROGALSKI
Cerep Inc. 15318 NE 95th Street
Redmond, WA 98052
USA

JENS SADOWSKI
AstraZeneca R&D Mölndal
Structural Chemistry Laboratory
SC264
43183 Mölndal
Sweden

NIKOLAY P. SAVCHUK
Chemical Diversity Labs, Inc.
Chemoinformatics
11575 Sorrento Valley Road
San Diego, CA 92121
USA

DORA M. SCHNUR
Pharmaceutical Research Institute
Bristol-Myers Squibb Co.
Province Line Road
Princeton, NJ 08540
USA

MARTIN W. SCHULTZ
Pfizer Global Research and Development
301 Henrietta Street
Kalamazoo, MI 49007
USA

VEERABAHU SHANMUGASUNDARAM
Computer-Assisted Drug Discovery
Pfizer Global Research and Development
2800 Plymouth Road
Ann Arbor, MI 48105
USA

ZENO SIMON
Romanian Academy
Institute of Chemistry
"Coriolan Dragulescu"
Bv. Mihai Viteazul No. 24
300223 Timisoara
Romania

ANDREW J. TEBBEN
Pharmaceutical Research Institute
Bristol-Myers Squibb Co.
Province Line Road
Princeton, NJ 08540
USA

SERGEY E. TKACHENKO
Chemical Diversity Labs, Inc.
Computational and Medicinal Chemistry
11575 Sorrento Valley Road
San Diego, CA 92121
USA

ALEXANDER TROPSHA
Laboratory for Molecular Modeling
University of North Carolina
Chapel Hill, NC 27599
USA

GARY WALKER
Chemical and Screening Sciences
Wyeth Research
401 N. Middletown Road
Pearl River, NY 10965
USA

DAVID WEININGER
Metaphorics, LLC
441 Greg Avenue
Sante Fe, NM 87501
USA

1
Introduction to Chemoinformatics in Drug Discovery – A Personal View

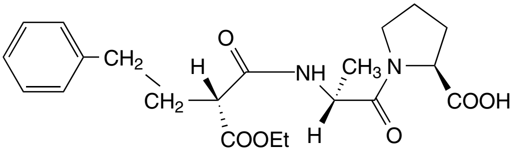
Garland R. Marshall

1.1
Introduction

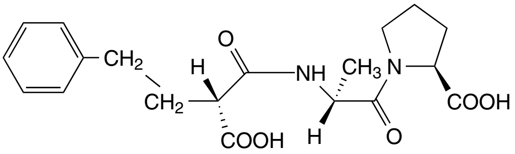
The first issue to be discussed is the definition of the topic. What is chemoinformatics and why should you care? There is no clear definition, although a consensus view appears to be emerging. "Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization" according to one view [1]. Hann and Green suggest that chemoinformatics is simply a new name for an old problem [2], a viewpoint I share. There are sufficient reviews [3–6] and even a book by Leach and Gillet [7] with the topic as their focus that there is little doubt what is meant, despite the absence of a precise definition that is generally accepted.

One aspect of a new emphasis is the sheer magnitude of chemical information that must be processed. For example, Chemical Abstracts Service adds over three-quarters of a million new compounds to its database annually, for which large amounts of physical and chemical property data are available. Some groups generate hundreds of thousands to millions of compounds on a regular basis through combinatorial chemistry that are screened for biological activity. Even more compounds are generated and screened *in silico* in the search for a magic bullet for a given disease. Either one of the two processes for generating information about chemistry has its own limitations. Experimental approaches have practical limitations despite automation; each *in vitro* bioassay utilizes a finite amount of reagents including valuable cloned and expressed receptors. Computational chemistry has to establish relevant criteria by which to select compounds of interest for synthesis and testing. The accuracy of prediction of affinities with current methodology is just now approaching sufficient accuracy to be of utility.

Let me emphasize the magnitude of the problem with a simple example. I was once asked to estimate the number of compounds covered by a typical issued patent for a drug of commercial interest. The patent that I selected to analyze was for enalapril, a prominent prodrug ACE inhibitor with a well-established commercial market. Given the parameters as outlined in the patent covering enalapril, an estimation of the total number of compounds included in the generic claim for enalaprilat, the active

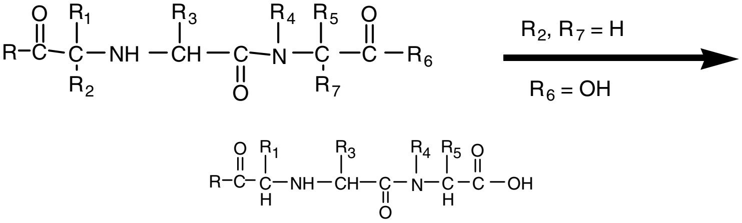


Enalapril



Enalaprilat

ingredient, was made. The following is the reference formula as described by the patent and simplified with R_6 = OH, and R_2 and R_7 = H:



Thus, one can simply enumerate the members of each class of substituent and combine them combinatorially. The following details the manner in which the number of each substituent was determined with the help of Chris Ho (Marshall and Ho, unpublished).

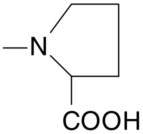
Substituent R: R is described as a lower alkoxy. The patent states that substituents are "otherwise represented by any of the variables including straight and branched chain hydrocarbon radicals from one to six carbon atoms, for example, methyl, ethyl, isopentyl, hexyl or vinyl, allyl, butenyl and the like." DBMAKER [8] was used to generate a database of compounds containing any combination of one to six carbon atoms, interspersed with occasional double and triple bonds, as well as all possible branching patterns. Constraints were employed to forbid the generation of chemically impossible constructs. Concord 3.01 [9] was used to generate and validate the chemical integrity of all compounds. 290 unique substituents were generated as a minimal estimate.

Substituent R3: This substituent is identical to substituent R, only that it is an alkyl instead of an alkoxy. Again, 290 unique substituents of six or fewer carbon atoms were generated.

Substituent R1: R1 is described as a substituted lower alkyl wherein the substituent is a phenyl group. The patent is vague with regard to where this phenyl group should reside. If the phenyl group always resides at the carbon farthest away from the main chain, then again, 290 different substituents will result. However, if the phenyl group can reside anywhere along the 1- to 6-member chain, then approximately 1000 substituents are chemically and sterically possible.

Substituents R4 & R5: These two substituents are described by the patent as being lower alkyl groups, which may be linked to form a cyclic 4- to 6-membered ring in this position. This produces two scenarios: if these groups remain unlinked, then, as before, 290 substituents are found at *each* position.

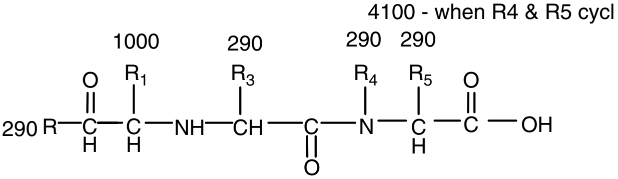
To determine the number of possible compounds when R4 and R5 are cyclized, a different approach was used. The patent states, ‘‘R4 and R5 when joined through the carbon and nitrogen atoms to which they are attached form a 4- to 6-membered ring’’. Preferred ring has the formula:



The patent is again vague in describing the generation of these cyclic systems. However, given that R4 and R5 are each 1–6 carbon alkyl groups with various branching patterns that are linked together, what results is a 4- to 6-membered ring system that may contain none, one or two side chains depending upon how R4 and R5 are connected. The overall requirement is that the total number of atoms comprising this ring system be less than or equal to 12.

To construct these ring systems, two databases were generated. The first database (‘‘ring database’’) contained three compounds – a 4-, 5- and 6-membered ring as specified by the patent. The second database (‘‘side-chain database’’) was constructed by cleaving each of the 290 alkyl compounds in half. One would assume that the first half of the alkyl chain would generate the ring, leaving the second half to dangle and form a side chain. A program DBCROSS (Ho, unpublished) was then used to join one compound from the ring database with up to two structures from the side-chain database at chemically appropriate substitution sites. Again, the overall requirement was that the number of atoms be less than or equal to 12. Approximately 4100 different cyclic systems were generated in this manner.

Total number of compounds



Summation	(290)(1000)(290)(290)(290)	=	$7.07 \cdot 10^{12}$	R4/R5 noncyclic
	(290)(1000)(290)(4100)	=	$3.44 \cdot 10^{11}$	R4/R5 cyclized

Sum $= 7.41 \cdot 10^{12} \rightarrow$ 3 chiral centers (carbons where R_1, R_3 and R_5 are attached to the backbone) in this molecule: X 8 $= 5.93 \cdot 10^{13}$ or more than 59 *trillion* compounds included in the patent.
Note: If the phenyl group of substituent R1 is limited to the position farthest from the parent chain, then the number of compounds drops to $1.72 \cdot 10^{13}$ or more than 17 *trillion* compounds included in the patent.

Actually, the number of compounds included in the patent is severalfold larger as esters of enalaprilat such as enalapril were also included. Of the 100 trillion or so compounds included in the patent, how many could be predicted to lack druglike properties (molecular weight too large? logP too high?)? How many would be predicted to be inactive on the basis of the known structure-activity data available on angiotensin-converting enzyme (ACE) inhibitors such as captopril? How many would be predicted to be inactive now that a crystal structure of a complex of ACE with an inhibitor has been published? Given the structure-activity relationships (SAR) available on the inhibitors, what could one determine regarding the active site of ACE? What novel classes of compound could be suggested on the basis of the SAR of inhibitors? On the basis of the new crystal structure of the complex? Do the most potent compounds share a set of properties that can be identified and used to optimize a novel lead structure? Can a predictive equation relating properties and affinity for the isolated enzyme be established? Can a similar equation relating properties and *in vitro* bioassay effectiveness be established? These are representative questions facing the current drug design community and one focus of chemoinformatics.

One significant tool that is employed is molecular modeling. Because I have been involved more directly with computational chemistry and molecular modeling, there is a certain bias in my perspective. This is the reason I have used "A Personal View" as part of the title. I have also chosen a historical presentation and focused largely on those contributions that significantly impacted my thinking. This approach, of course, has its own limitation, and I apologize to my colleagues for any distortions or omissions.

1.2
Historical Evolution

With the advent of computers and the ability to store and retrieve chemical information, serious efforts to compile relevant databases and construct information retrieval systems began. One of the first efforts to have a substantial long-term impact was to collect the crystal structure information for small molecules by Olga Kennard. The Cambridge Structural Database (CSD) stores crystal structures of small molecules and provides a fertile resource for geometrical data on molecular fragments for calibration of force fields and validation of results from computational chemistry [10, 11]. As protein crystallography gained momentum, the need for a common repository of macromolecular structural data led to the Protein Data Base (PDB) originally located at Brookhaven National Laboratories [12]. These efforts focused on the accumulation and organization of experimental results on the three-dimensional structure of molecules, both large and small. Todd Wipke recognized the need for a chemical information system to handle the increasing numbers of small molecules generated in industry, and thus MDL and MACCS were born.

With the advent of computers and the availability of oscilloscopes, the idea of displaying a three-dimensional structure of the screen was obvious with rotation providing depth cueing. Cyrus Levinthal and colleagues utilized the primitive computer graphics facilities at MIT to generate rotating images of proteins and nucleic acids to provide insight into the three-dimensional aspects of these structures without having to build physical models. His paper in Scientific American in 1965 was sensational and inspired others (including myself [13]) to explore computer graphics (1966/1967) as a means of coping with the 3D nature of chemistry. Physical models (Dreiding stick figures, CPK models, etc.) were useful accepted tools for medicinal chemists, but physical overlap of two or more compounds was difficult and exploration of the potential energy surface hard to correlate with a given conformation of a physical model.

As more and more chemical data accumulated with its implicit information content, a multitude of approaches began to extract useful information. Certainly, the shape and variability in geometry of molecular fragments from CSD was mined to provide fragments of functional groups for a variety of purposes. As series of compounds were tested for biological activity in a given assay, the desire to distill the essence of the chemical requirements for such activity to guide optimization was generated. Initially, the efforts focused on congeneric series as the common scaffold presumably eliminated the molecular alignment problem with the assumption that all molecules bound with a common orientation of the scaffold. This was the intellectual basis of the Hansch approach (quantitative structure-activity relationships, QSAR), in which substituent parameters from physical chemistry were used to correlate chemical properties with biological activity for a series of compounds with the same substitution pattern on the congeneric scaffold [14, 15].

1.3 Known versus Unknown Targets

Intellectually, the application of molecular modeling has dichotomized into those methods dealing with biological systems where no structural information at the atomic level is known, the unknown receptor, and those systems that have become relatively common, where a three-dimensional structure is know from crystallography or NMR spectroscopy. The Washington University group has spent most of its efforts over the last three decades focused on the common problem encountered where one has little structural information. Others, such as Peter Goodford and Tak Kuntz, have taken the lead in developing approaches to therapeutic targets where the structure of the target was available at atomic resolution. The seminal work of Goodford and colleagues [16] on designing inhibitors of the 2,3-diphosphorylglycerate (DPG) binding site on hemoglobin

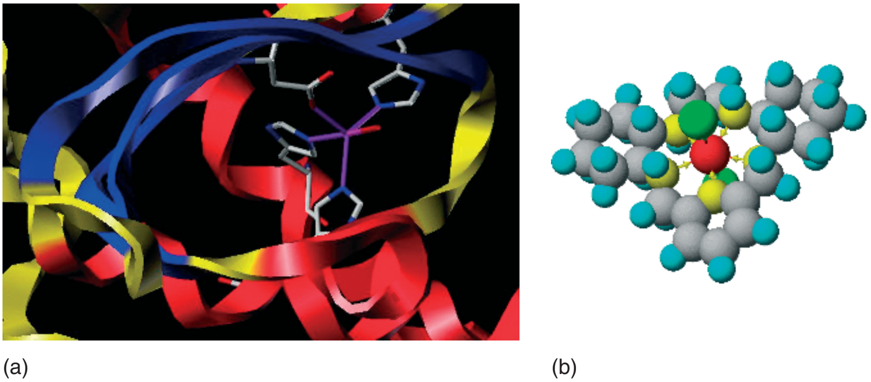


Fig. 1.1 **(a)** Active site of Mn superoxide dismutase (three histidine and one aspartic acid ligand to manganese) and **(b)** M40403, synthetic enzyme with 5 nitrogens (yellow) and two chloride (green) ligands.

for the treatment of sickle-cell disease certainly stimulated many others to obtain crystal structures of their therapeutic target. The most dramatic example of computer-aided drug design of which I am aware is the development of superoxide dismutase mimetics of below 500 molecular weight by Dennis Riley of Metaphore Pharmaceuticals. By understanding the redox chemistry of manganese superoxide reductase, Riley was able to design a totally novel pentaazacrown scaffold complexed with manganese (Figure 1.1) that catalyzes the conversion of superoxide to hydrogen peroxide at diffusion-controlled rates [17, 18]. This is the first example of a synthetic enzyme with a catalytic rate equal to or better than nature’s best. The advances in molecular biology provided the means of cloning and expressing proteins in sufficient quantities to screen a variety of conditions for crystallization. Thus, it is almost expected that a crystal structure is available for any therapeutic target of interest. Unfortunately, many therapeutic targets such as G-protein-coupled receptors are still significant challenges to structural biology.

1.4
Graph Theory and Molecular Numerology

Considerable literature developed around the ability of numerical indices derived from graph theoretical considerations to correlate with SAR data. This was a source of mystery to me for some time. A colleague, Ioan Motoc, from Romania, with experience in this arena and a very strong intellect, helped me understand the ability of various indices to be useful parameters in QSAR equations [19–21]. Ioan correlated various indices with more physically relevant (at least to me) variables such as surface area and molecular volume. Since computational time was at a premium during the early days of QSAR and such indices could be calculated with minimal computations, they played a useful role and continue to be used. As a chemist, however, I am much more comfortable with parameters such as surface area or volume.