

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

Editorial Board

H.-D. Holtje, H. Timmerman, J. Vacca, H. van de Waterbeemd, T. Wieland

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

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>.

© 2005 WILEY-VCH Verlag GmbH & Co.
KGaA Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

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 XV

Preface XVII

List of Contributors XIX

1 Introduction to Chemoinformatics in Drug Discovery –

A Personal View 1

Garland R. Marshall

1.1 Introduction 1

1.2 Historical Evolution 4

1.3 Known versus Unknown Targets 5

1.4 Graph Theory and Molecular Numerology 6

1.5 Pharmacophore 7

1.6 Active-Analog Approach 8

1.7 Active-Site Modeling 9

1.8 Validation of the Active-Analog Approach and Active-Site Modeling 10

1.9 PLS/CoMFA 11

1.10 Prediction of Affinity 12

1.11 Protein Structure Prediction 13

1.12 Structure-Based Drug Design 15

1.13 Real World Pharmaceutical Issues 15

1.14 Combinatorial Chemistry and High-throughput Screens 16

1.15 Diversity and Similarity 16

1.16 Prediction of ADME 17

1.17 Failures to Accurately Predict 17

1.18 Summary 18

References 19

Part I Virtual Screening 23**2 Chemoinformatics in Lead Discovery 25***Tudor I. Oprea*

- 2.1 Chemoinformatics in the Context of Pharmaceutical Research 25
- 2.2 Leads in the Drug Discovery Paradigm 27
- 2.3 Is There a Trend for High Activity Molecules? 29
- 2.4 The Concept of Leadlikeness 32
- 2.5 Conclusions 37
- References 38*

3 Computational Chemistry, Molecular Complexity and Screening Set Design 43*Michael M. Hann, Andrew R. Leach, and Darren V.S. Green*

- 3.1 Introduction 43
- 3.2 Background Concepts: the Virtual, Tangible and Real Worlds of Compounds, the “Knowledge Plot” and Target Tractability 44
- 3.3 The Construction of High Throughput Screening Sets 45
- 3.4 Compound Filters 47
- 3.5 “Leadlike” Screening Sets 48
- 3.6 Focused and Biased Set Design 54
- 3.7 Conclusion 55
- References 56*

4 Algorithmic Engines in Virtual Screening 59*Matthias Rarey, Christian Lemmen, and Hans Matter*

- 4.1 Introduction 59
- 4.2 Software Tools for Virtual Screening 61
- 4.3 Physicochemical Models in Virtual Screening 62
 - 4.3.1 Intermolecular Forces in Protein–Ligand Interactions 63
 - 4.3.2 Scoring Functions for Protein–Ligand Recognition 66
 - 4.3.3 Covering Conformational Space 67
 - 4.3.4 Scoring Structural Alignments 68
- 4.4 Algorithmic Engines in Virtual Screening 69
 - 4.4.1 Mathematical Concepts 69
 - 4.4.2 Algorithmic Concepts 76
 - 4.4.3 Descriptor Technology 81
 - 4.4.4 Global Search Algorithms 85
- 4.5 Entering the Real World: Virtual Screening Applications 89
 - 4.5.1 Practical Considerations on Virtual Screening 89
 - 4.5.2 Successful Applications of Virtual Screening 91
- 4.6 Practical Virtual Screening: Some Final Remarks 99
- References 101*

5	Strengths and Limitations of Pharmacophore-Based Virtual Screening	117
	<i>Dragos Horvath, Boryeu Mao, Rafael Gozalbes, Frédérique Barbosa, and Sherry L. Rogalski</i>	
5.1	Introduction	117
5.2	The “Pharmacophore” Concept: Pharmacophore Features	117
5.3	Pharmacophore Models: Managing Pharmacophore-related Information	118
5.4	The Main Topic of This Paper	119
5.5	The Cox2 Data Set	119
5.6	Pharmacophore Fingerprints and Similarity Searches	120
5.7	Molecular Field Analysis (MFA)-Based Pharmacophore Information	123
5.8	QSAR Models	125
5.9	Hypothesis Models	125
5.10	The Minimalist Overlay-Independent QSAR Model	126
5.11	Minimalist and Consensus Overlay-Based QSAR Models	128
5.12	Diversity Analysis of the Cox2 Compound Set	131
5.13	Do Hypothesis Models Actually Tell Us More Than Similarity Models About the Structural Reasons of Activity?	131
5.14	Why Did Hypothesis Models Fail to Unveil the Key Cox2 Site–Ligand Interactions?	134
5.15	Conclusions	136
	<i>References</i>	<i>137</i>
Part II	Hit and Lead Discovery	141
6	Enhancing Hit Quality and Diversity Within Assay Throughput Constraints	143
	<i>Iain McFadyen, Gary Walker, and Juan Alvarez</i>	
6.1	Introduction	143
6.1.1	What Makes a Good Lead Molecule?	144
6.1.2	Compound Collections – Suitability as Leads	144
6.1.3	Compound Collections – Diversity	145
6.1.4	Data Reliability	146
6.1.5	Selection Methods	149
6.1.6	Enhancing Quality and Diversity of Actives	153
6.2	Methods	154
6.2.1	Screening Library	155
6.2.2	Determination of Activity Threshold	156
6.2.3	Filtering	156
6.2.4	High-Throughput Screen Clustering Algorithm (HTSCA)	157
6.2.5	Diversity Analysis	160

6.2.6	Data Visualization	161
6.3	Results	162
6.3.1	Peptide Hydrolase	162
6.3.2	Protein Kinase	167
6.3.3	Protein–Protein Interaction	168
6.4	Discussion and Conclusion	169
	<i>References</i>	172
7	Molecular Diversity in Lead Discovery: From Quantity to Quality	175
	<i>Cullen L. Cavallaro, Dora M. Schnur, and Andrew J. Tebben</i>	
7.1	Introduction	175
7.2	Large Libraries and Collections	176
7.2.1	Methods and Examples for Large Library Diversity Calculations	177
7.3	Medium-sized/Target-class Libraries and Collections	181
7.3.1	Computational Methods for Medium- and Target-class Libraries and Collections	183
7.4	Small Focused Libraries	189
7.4.1	Computational Methods for Small and Focused Libraries	190
7.5	Summary/Conclusion	191
	<i>References</i>	192
8	<i>In Silico</i> Lead Optimization	199
	<i>Chris M.W. Ho</i>	
8.1	Introduction	199
8.2	The Rise of Computer-aided Drug Refinement	200
8.3	RACHEL Software Package	201
8.4	Extraction of Building Blocks from Corporate Databases	201
8.5	Intelligent Component Selection System	203
8.6	Development of a Component Specification Language	205
8.7	Filtration of Components Using Constraints	207
8.8	Template-driven Structure Generation	208
8.9	Scoring Functions – Methods to Estimate Ligand–Receptor Binding	209
8.10	Target Functions	212
8.11	Ligand Optimization Example	214
	<i>References</i>	219

Part III Databases and Libraries 221**9 WOMBAT: World of Molecular Bioactivity 223**

Marius Olah, Maria Mracec, Liliana Ostopovici, Ramona Rad, Alina Bora, Nicoleta Hadaruga, Ionela Olah, Magdalena Banda, Zeno Simon, Mircea Mracec, and Tudor I. Oprea

- 9.1 Introduction – Brief History of the WOMBAT Project 223
- 9.2 WOMBAT 2004.1 Overview 224
- 9.3 WOMBAT Database Structure 227
- 9.4 WOMBAT Quality Control 228
- 9.5 Uncovering Errors from Literature 231
- 9.6 Data Mining with WOMBAT 234
- 9.7 Conclusions and Future Challenges 235
- References 237*

10 Cabinet – Chemical and Biological Informatics Network 241

Vera Povolna, Scott Dixon, and David Weininger

- 10.1 Introduction 241
 - 10.1.1 Integration Efforts, WWW as Information Resource and Limitations 241
 - 10.1.2 Goals 243
- 10.2 Merits of Federation Rather than Unification 243
 - 10.2.1 The Merits of Unification 244
 - 10.2.2 The Merits of Federation 244
 - 10.2.3 Unifying Disparate Data Models is Difficult, Federating them is Easy 245
 - 10.2.4 Language is a Natural Key 246
- 10.3 HTTP is Appropriate Communication Technology 248
 - 10.3.1 HTTP is Specifically Designed for Collaborative Computing 248
 - 10.3.2 HTTP is the Dominant Communication Protocol Today 248
 - 10.3.3 HTML Provides a Universally Accessible GUI 249
 - 10.3.4 MIME “Text/Plain” and “Application/Octet-Stream” are Important Catch-alls 249
 - 10.3.5 Other MIME Types are Useful 250
 - 10.3.6 One Significant HTTP Work-around is Required 250
- 10.4 Implementation 251
 - 10.4.1 Daylight HTTP Toolkit 251
 - 10.4.2 Metaphorics’ Cabinet Library 252
- 10.5 Specific Examples of Federated Services 252
 - 10.5.1 Empath – Metabolic Pathway Chart 253
 - 10.5.2 Planet – Protein–ligand Association Network 254
 - 10.5.3 EC Book – Enzyme Commission Codebook 254
 - 10.5.4 WDI – World Drug Index 254

10.5.5	WOMBAT – World of Molecular Bioactivity	255
10.5.6	TCM (Traditional Chinese Medicines), DCM (Dictionary of Chinese Medicine), PARK (Photo ARKive) and zi4	255
10.5.7	Cabinet “Download” Service	256
10.5.8	Cabinet Usage Example	256
10.6	Deployment and Refinement	262
10.6.1	Local Deployment	264
10.6.2	Intranet Deployment	264
10.6.3	Internet Deployment	265
10.6.4	Online Deployment	266
10.7	Conclusions	266
	<i>References</i>	268

11 Structure Modification in Chemical Databases 271

Peter W. Kenny and Jens Sadowski

11.1	Introduction	271
11.2	Permute	274
11.2.1	Protonation and Formal Charges	274
11.2.2	Tautomerism	275
11.2.3	Nitrogen Configurations	276
11.2.4	Duplicate Removal	276
11.2.5	Nested Loop	276
11.2.6	Application Statistics	277
11.2.7	Impact on Docking	277
11.3	Leatherface	279
11.3.1	Protonation and Formal Charges	279
11.3.2	Tautomerism	280
11.3.3	Ionization and Tautomer Model	281
11.3.4	Relationships between Structures	282
11.3.5	Substructural Searching and Analysis	283
11.4	Concluding Remarks	283
	<i>References</i>	284

12 Rational Design of GPCR-specific Combinational Libraries Based on the Concept of Privileged Substructures 287

Nikolay P. Savchuk, Sergey E. Tkachenko, and Konstantin V. Balakin

12.1	Introduction – Combinatorial Chemistry and Rational Drug Design	287
12.2	Rational Selection of Building Blocks Based on Privileged Structural Motifs	288
12.2.1	Privileged Structures and Substructures in the Design of Pharmacologically Relevant Combinatorial Libraries	288

12.2.2	Analysis of Privileged Structural Motifs: Structure Dissection Rules	291
12.2.3	Knowledge Database	293
12.2.4	Target-specific Differences in Distribution of Molecular Fragments	295
12.2.5	Privileged versus Peripheral Retrosynthetic Fragments	296
12.2.6	Peripheral Retrosynthetic Fragments: How to Measure the Target-specific Differences?	297
12.2.7	Selection of Building Blocks	300
12.2.8	Product-based Approach: Limiting the Space of Virtual Libraries	305
12.2.9	Alternative Strategy: Property-based Approach	306
12.2.10	Kohonen Self-organizing Maps	307
12.3	Conclusions	309
	<i>References</i>	311

Part IV Chemoinformatics Applications 315

13	A Practical Strategy for Directed Compound Acquisition	317
	<i>Gerald M. Maggiora, Veerabahu Shanmugasundaram, Michael S. Lajiness, Tom N. Doman, and Martin W. Schultz</i>	
13.1	Introduction	317
13.2	A Historical Perspective	319
13.3	Practical Issues	320
13.4	Compound Acquisition Scheme	322
13.4.1	Preprocessing Compound Files	322
13.4.2	Initial Compound Selection and Diversity Assessment	325
13.4.3	Compound Reviews	327
13.4.4	Final Selection and Compound Purchase	328
13.5	Conclusions	328
13.6	Methodologies	329
13.6.1	Preprocessing Filters	329
13.6.2	Diverse Solutions (DVS)	330
13.6.3	Dfragall	330
13.6.4	Ring Analysis	331
	<i>References</i>	331
14	Efficient Strategies for Lead Optimization by Simultaneously Addressing Affinity, Selectivity and Pharmacokinetic Parameters	333
	<i>Karl-Heinz Baringhaus and Hans Matter</i>	
14.1	Introduction	333
14.2	The Origin of Lead Structures	336
14.3	Optimization for Affinity and Selectivity	338
14.3.1	Lead Optimization as a Challenge in Drug Discovery	338
14.3.2	Use and Limitation of Structure-based Design Approaches	339

14.3.3	Integration of Ligand- and Structure-based Design Concepts	340
14.3.4	The Selectivity Challenge from the Ligands' Perspective	342
14.3.5	Selectivity Approaches Considering Binding Site Topologies	344
14.4	Addressing Pharmacokinetic Problems	347
14.4.1	Prediction of Physicochemical Properties	347
14.4.2	Prediction of ADME Properties	348
14.4.3	Prediction of Toxicity	349
14.4.4	Physicochemical and ADMET Property-based Design	350
14.5	ADME/Antitarget Models for Lead Optimization	350
14.5.1	Global ADME Models for Intestinal Absorption and Protein Binding	350
14.5.2	Selected Examples to Address ADME/Toxicology Antitargets	354
14.6	Integrated Approach	357
14.6.1	Strategy and Risk Assessment	357
14.6.2	Integration	359
14.6.3	Literature and Aventis Examples on Aspects of Multidimensional Optimization	360
14.7	Conclusions	366
	<i>References</i>	367

15 Chemoinformatic Tools for Library Design and the Hit-to-Lead Process: A User's Perspective 381

Robert Alan Goodnow, Jr., Paul Gillespie, and Konrad Bleicher

15.1	The Need for Leads: The Sources of Leads and the Challenge to Find Them	381
15.2	Property Predictions	383
15.3	Prediction of Solubility	384
15.4	Druglikeness	390
15.4.1	Are There Differences between Drugs and Nondrugs?	390
15.4.2	Is the Problem Tractable within a Single Program?	391
15.4.3	Do We Have a Training Set that Will Allow Us to Address the Issue?	392
15.4.4	Approaches to the Prediction of Druglikeness	392
15.5	Frequent Hitters	394
15.6	Identification of a Lead Series	395
15.7	The Hit-to-lead Process	397
15.7.1	Prioritization of Hits	397
15.7.2	Identification of Analogs	402
15.7.3	Additional Assays	403
15.8	Leads from Libraries: General Principles, Practical Considerations	404
15.9	Druglikeness in Small-molecule Libraries	406
15.10	Data Reduction and Viewing for Virtual Library Design	407
15.11	Druglikeness	408

15.12	Complexity and Andrews' Binding Energy	408
15.13	Solubility	411
15.14	Polar Surface Area	411
15.15	Number of Rotatable Bonds	412
15.16	hERG Channel Binding	413
15.17	Chemoinformatic Analysis of the Predicted Hansch Substituent Constants of the Diversity Reagents for Design of Vector Exploration Libraries	415
15.18	Targeting Libraries by Virtual Screening	416
15.19	Combinatorial Design Based on Biostructural Information	418
15.20	Ligand-based Combinatorial Design: The RADDAR Approach	419
15.21	Virtual Screening of Small-molecule Library with Peptide-derived Pharmacophores	421
15.22	Chemoinformatic Tools and Strategies to Visualize Active Libraries	423
15.23	Visualization of Library Designs during Hit-to-lead Efforts	423
15.24	Summary and Outlook for Chemoinformatically Driven Lead Generation	425
	<i>References</i>	426
16	Application of Predictive QSAR Models to Database Mining	437
	<i>Alexander Tropsha</i>	
16.1	Introduction	437
16.2	Building Predictive QSAR Models: The Importance of Validation	438
16.3	Defining Model Applicability Domain	441
16.4	Validated QSAR Modeling as an Empirical Data-modeling Approach: Combinatorial QSAR	443
16.5	Validated QSAR Models as Virtual Screening Tools	445
16.6	Conclusions and Outlook	452
	<i>References</i>	453
17	Drug Discovery in Academia – A Case Study	457
	<i>Donald J. Abraham</i>	
17.1	Introduction	457
17.2	Linking the University with Business and Drug Discovery	457
17.2.1	Start-up Companies	457
17.2.2	Licensing	458
17.3	Research Parks	459
17.4	Conflict of Interest Issues for Academicians	459
17.5	Drug Discovery in Academia	461
17.5.1	Clinical Trials in Academia	461

XIV | *Contents*

17.6	Case Study: The Discovery and Development of Allosteric Effectors of Hemoglobin	462
17.6.1	Geduld (Patience)	463
17.6.2	Glück (Luck)	463
17.6.3	Geschick (Skill)	464
17.6.4	Geld (Money)	471
	<i>References</i>	481

Subject Index	485
----------------------	-----

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

xx | *List of Contributors*

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

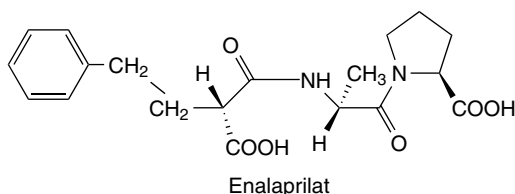
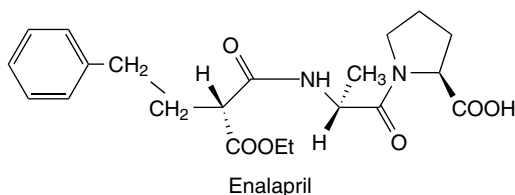
1.1

Introduction

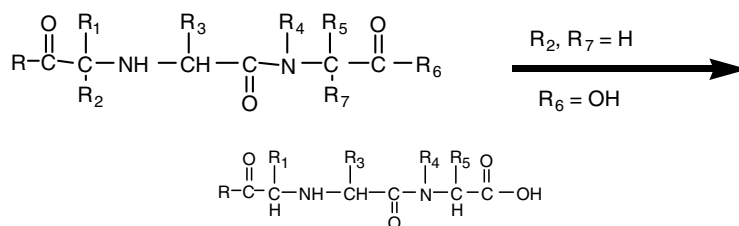
The first issue to be discussed is the definition of the topic. What is chemoinformatics and why should you care? There is no clear definition, although a consensus view appears to be emerging. “Chemoinformatics is the mixing of those information resources to transform data into information and information 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



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



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).

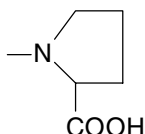
Substituent R: R is described as a lower alkoxy. The patent states that substituents are “otherwise represented by any of the variables including straight and branched chain hydrocarbon radicals from one to six carbon atoms, for example, methyl, ethyl, isopentyl, hexyl or vinyl, allyl, butenyl and the 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.

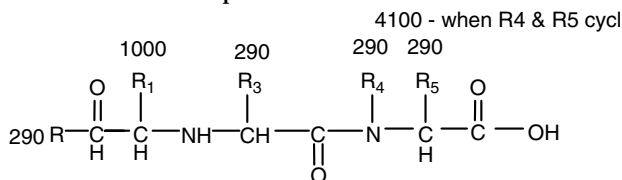
To determine the number of possible compounds when R4 and R5 are cyclized, a different approach was used. The patent states, "R4 and R5 when joined through the carbon and nitrogen atoms to which they are attached form a 4- to 6-membered ring". Preferred ring has the formula:



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



$$\begin{array}{ll}
 \text{Summation} & (290)(1000)(290)(290)(290) = 7.07 \cdot 10^{12} \quad \text{R4/R5 noncyclic} \\
 & (290)(1000)(290)(4100) = 3.44 \cdot 10^{11} \quad \text{R4/R5 cyclized}
 \end{array}$$

Sum = $7.41 \cdot 10^{12} \rightarrow 3$ chiral centers (carbons where R₁, R₃ and R₅ 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 known 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

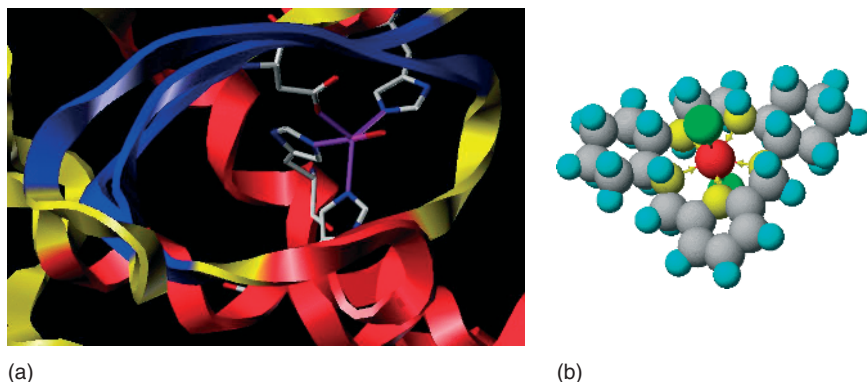


Fig. 1.1 (a) Active site of Mn superoxide dismutase (three histidine and one aspartic acid ligand to manganese) and (b) M40403, synthetic enzyme with 5 nitrogens (yellow) and two chloride (green) ligands.

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.