

# Molecular Modeling and Prediction of Bioactivity

Edited by

**Klaus Gundertofte**

*H. Lundbeck A/S  
Valby, Denmark*

and

**Flemming Steen Jørgensen**

*Royal Danish School of Pharmacy  
Copenhagen, Denmark*

**KLUWER ACADEMIC / PLENUM PUBLISHERS**  
New York, Boston, Dordrecht, London, Moscow

Library of Congress Cataloging-in-Publication Data

Molecular modeling and prediction of bioactivity / edited by Klaus Gundertofte and Flemming Steen Jørgensen.

p. cm.

"Proceedings of the 12th European Symposium on Quantitative Structure-Activity Relationships ... held August 23-28, 1998, in Copenhagen, Denmark"--CIP copyright p. Includes bibliographical references and index.

ISBN 0-306-46217-6

1. QSAR (Biochemistry)--Congresses. 2. Drugs--Design--Congresses. 3. Biomolecules--Computer simulation--Congresses. I. Gundertofte, Klaus. II. Jørgensen, Flemming S. III. European Symposium on Quantitative Structure-Activity Relationships (12th : 1998 : Copenhagen, Denmark)

RM301.42 .M64 1999  
615'.19--dc21

99-044859

Proceedings of the 12th European Symposium on Quantitative Structure-Activity Relationships: Molecular Modeling and Prediction of Bioactivity, held August 23-28, 1998, in Copenhagen, Denmark

ISBN 0-306-46217-6

©2000 Kluwer Academic/Plenum Publishers, New York  
233 Spring Street, New York, N.Y. 10013

<http://www.wkap.nl>

10 9 8 7 6 5 4 3 2 1

A C.I.P. record for this book is available from the Library of Congress

All rights reserved

No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the Publisher

Printed in the United States of America

## PREFACE

The 12<sup>th</sup> European Symposium on *Quantitative Structure-Activity Relationships* was held in Radisson SAS Falconer Center in Copenhagen August 23–28, 1998.

Three hundred and forty participants attended the meeting from 34 countries, representing leading scientists from the industry and academic institutions. The Symposium featured invited lectures, oral reports, and more than 150 posters remained on show for the whole week and helped to stimulate the scientific discussions during the meeting. This volume is based on these contributions.

Major companies screen hundreds of thousands of compounds in their effort to find new leads. Quantitative Structure-Activity Relationships (QSAR), Molecular Modeling, Similarity Assessment, and Information Technologies play an increasing role in the scientific community working in the area of structure-activity relationships. Predictions of affinity, efficacy, selectivity, bioavailability, or metabolism are important in advising medicinal chemists to prepare or not to prepare a given compound. Diversity assessment is used to extend the information content in focused as well as in diverse libraries from combinatorial chemistry methods.

The rapid identification of genes will be accompanied by the equally rapid availability of new targets, which will further increase the need for efficient information systems. Such systems are needed too, in order to facilitate rapid exchange of information within drug discovery teams. Data mining tools are being developed to help the researchers digest the wealth of information produced in high-throughput organic synthesis and screening.

With this volume we have tried to illustrate the many facets of the QSAR discipline having evolved since the first important contributions from Corwin Hansch, the founder of QSAR, to whom one of the major sessions was dedicated on the occasion of his 80th birthday.

The volume covers very important topics in the challenging process from lead finding to drug candidates. Focus has been set on the potential usefulness of methods for design of lead discovery libraries, lead optimization, computational chemistry methods for the calculation of energetics of protein-ligand interaction, and computer simulations of biological activities. Important topics include new developments in chemometrics and rational molecular design, as well as different aspects of structure representation, knowledge-based approaches to structure identification, and information handling.

The contributions contained herein reside at the very cutting edge of the discipline and add to the further proliferation of the basic ideas of QSAR.

Klaus Gundertofte  
Flemming Steen Jørgensen

## CONTENTS

### Section I: Overview

Strategies for Molecular Design Beyond the Millennium .....	3
James P. Snyder and Forrest D. Snyder	

### Section II: New Developments and Applications of Multivariate QSAR

Multivariate Design and Modelling in QSAR, Combinatorial Chemistry, and Bioinformatics .....	27
Svante Wold, Michael Sjöström, Per M. Andersson, Anna Linusson, Maria Edman, Torbjörn Lundstedt, Bo Nordén, Maria Sandberg, and Lise-Lott Uppgård	
QSAR Study of PAH Carcinogenic Activities: Test of a General Model for Molecular Similarity Analysis .....	47
William C. Herndon, Hung-Ta Chen, Yumei Zhang, and Gabrielle Rum	
Comparative Molecular Field Analysis of Aminopyridazine Acetylcholinesterase Inhibitors .....	53
Wolfgang Sippl, Jean-Marie Contreras, Yveline Rival, and Camille G. Wermuth	
The Influence of Structure Representation on QSAR Modelling .....	59
Marjana Novič, Matevž Pompe, and Jure Zupan	
The Constrained Principal Property (CPP) Space in QSAR—Directional and Non-Directional Modelling Approaches .....	65
Lennart Eriksson, Patrik Andersson, Erik Johansson, Mats Tysklind, Maria Sandberg, and Svante Wold	

### Section III: The Future of 3D-QSAR

Handling Information from 3D Grid Maps for QSAR Studies .....	73
Gabriele Cruciani, Manuel Pastor, and Sergio Clementi	
Gaussian-Based Approaches to Protein-Structure Similarity .....	83
Jordi Mestres, Douglas C. Rohrer, and Gerald M. Maggiora	
Molecular Field-Derived Descriptors for the Multivariate Modeling of Pharmacokinetic Data .....	89
Wolfgang Guba and Gabriele Cruciani	

Validating Novel QSAR Descriptors for Use in Diversity Analysis.....	95
Robert D. Clark, Michael Brusati, Robert Jilek, Trevor Heritage, and Richard D. Cramer	

#### **Section IV: Prediction of Ligand-Protein Binding**

Structural and Energetic Aspects of Protein-Ligand Binding in Drug Design.....	103
Gerhard Klebe, Markus Böhm, Frank Dullweber, Ulrich Grädler, Holger Gohlke, and Manfred Hendlich	
Use of MD-Derived Shape Descriptors as a Novel Way to Predict the <i>in Vivo</i> Activity of Flexible Molecules: The Case of New Immunosuppressive Peptides .....	111
Abdelaziz Yasri, Michel Kaczorek, Roger Lahana, Gérard Grassy, and Roland Buelow	
A View on Affinity and Selectivity of Nonpeptidic Matrix Metalloproteinase Inhibitors from the Perspective of Ligands and Target .....	123
Hans Matter and Wilfried Schwab	
On the Use of SCRF Methods in Drug Design Studies .....	129
Modesto Orozco, Carles Colominas, Xavier Barril, and F. Javier Luque	
3D-QSAR Study of 1,4-Dihydropyridines Reveals Distinct Molecular Requirements of Their Binding Site in the Resting and the Inactivated State of Voltage-Gated Calcium Channels .....	135
Klaus-Jürgen Schleifer, Edith Tot, and Hans-Dieter Höltje	
Pharmacophore Development for the interaction of Cytochrome P450 1A2 with Its Substrates and Inhibitors.....	141
Elena López-de-Briñas, Juan J. Lozano, Nuria B. Centeno, Jordi Segura, Marisa González, Rafael de la Torre, and Ferran Sanz	

#### **Section V: Computational Aspects of Molecular Diversity and Combinatorial Libraries**

Analysis of Large, High-Throughput Screening Data Using Recursive Partitioning.....	149
S. Stanley Young and Jerome Sacks	
3D Structure Descriptors for Biological Activity .....	157
Johann Gasteiger, Sandra Handschuh, Markus C. Hemmer, Thomas Kleinöder, Christof H. Schwab, Andreas Teckentrup, Jens Sadowski, and Markus Wagener	
Fragment-Based Screening of Ligand Databases .....	169
Christian Lemmen and Thomas Lengauer	
The Computer Simulation of High Throughput Screening of Bioactive Molecules .....	175
Frank R. Burden and David A. Winkler	

#### **Section VI: Affinity and Efficacy Models of G-Protein Coupled Receptors**

5-HT <sub>1A</sub> Receptors Mapping by Conformational Analysis (2D NOESY/MM) and "THREE WAY MODELLING" (HASL, CoMFA, PARM) .....	183
Maria Santagati, Arthur Doweyko, Andrea Santagati, Maria Modica, Salvatore Guccione, Chen Hongming, Gloria Uccello Barretta, and Federica Balzano	

Design and Activity Estimation of a New Class of Analgesics.....	195
Slavomir Filipek and Danuta Pawlak	
Unified Pharmacophoric Model for Cannabinoids and Aminoalkylindoles.....	201
Joong-Youn Shim, Elizabeth R. Collantes, William J. Welsh, and Allyn C. Howlett	
Chemometric Detection of Binding Sites of 7TM Receptors.....	207
Monica Clementi, Sara Clementi, Sergio Clementi, Gabriele Cruciani, Manuel Pastor, and Jonas E. Nilsson	

### **Section VII: New Methods in Drug Discovery**

SpecMat: Spectra as Molecular Descriptors for the Prediction of Biological Activity.....	215
R. Bursi and V.J. van Geerestein	
Hydrogen Bond Contributions to Properties and Activities of Chemicals and Drugs .....	221
Oleg A. Raevsky, Klaus J. Schaper, Han van de Waterbeemd, and James W. McFarland	

### **Section VIII: Modeling of Membrane Penetration**

Predicting Peptide Absorption .....	231
Lene H. Krarup, Anders Berglund, Maria Sandberg, Inge Thøger Christensen, Lars Hovgaard, and Sven Frokjaer	
Physicochemical High Throughput Screening (pC-HTS): Determination of Membrane Permeability, Partitioning and Solubility .....	237
Manfred Kansy, Krystyna Kratzat, Isabelle Parrilla, Frank Senner, and Björn Wagner	
Understanding and Estimating Membrane/Water Partition Coefficients: Approaches to Derive Quantitative Structure Property Relationships.....	245
Wouter H. J. Vaes, Eñaut Urrestarazu Ramos, Henk J. M. Verhaar, Christopher J. Cramer, and Joop L. M. Hermens	
Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure.....	249
M. D. Wessel, P. C. Jurs, J. W. Tolan, and S. M. Muskal	

### **Section IX: Poster Presentations**

#### **Poster Session I: New Developments and Applications of Multivariate QSAR**

Free-Wilson-Type QSAR Analyses Using Linear and Nonlinear Regression Techniques...	261
Klaus-Jürgen Schaper	
QSAR Studies of Picodendrins and Related Terpenoids—Structural Differences between Antagonist Binding Sites on GABA Receptors of Insects and Mammals ..	263
Miki Akamatsu, Yoshihisa Ozoe, Taizo Higata, Izumi Ikeda, Kazuo Mochida, Kazuo Koike, Taichi Ohmoto, Tamotsu Nikaido, and Tamio Ueno	
Molecular Lipophilicity Descriptors: A Multivariate Analysis .....	265
Raimund Mannhold and Gabriele Cruciani	

World Wide Web-Based Calculation of Substituent Parameters for QSAR Studies .....	267
Peter Ertl	
COMBINE and Free-Wilson QSAR Analysis of Nuclear Receptor-DNA Binding .....	269
Sanja Tomic, Lennart Nilsson, and Rebecca C. Wade	
QSAR Model Validation .....	271
Erik Johansson, Lennart Eriksson, Maria Sandberg, and Svante Wold	
QSPR Prediction of Henry's Law Constant: Improved Correlation with New Parameters ..	273
John C. Dearden, Shazia A. Ahmed, Mark T. D. Cronin, and Janeth A. Sharra	
QSAR of a Series of Carnitine Acetyl Transferase (CAT) Substrates .....	275
G. Gallo, M. Mabilia, M. Santaniello, M. O. Tinti, and P. Chiodi	
"Classical" and Quantum Mechanical Descriptors for Phenolic Inhibition of Bacterial Growth .....	277
S. Shapiro and D. Turner	
Hydrogen Bond Acceptor and Donor Factors, $C_a$ and $C_d$ : New QSAR Descriptors .....	280
James W. McFarland, Oleg A. Raevsky, and Wendell W. Wilkerson	
Development and Validation of a Novel Variable Selection Technique with Application to QSAR Studies .....	282
Chris L. Waller and Mary P. Bradley	
QSAR Studies of Environmental Estrogens .....	284
M. G. B. Drew, N. R. Price, and H. J. Wood	
Quantitative Structure-Activity Relationship of Antimutagenic Benzalacetones and Related Compounds .....	286
Chisako Yamagami, Noriko Motohashi, and Miki Akamatsu	
Multivariate Regression Excels Neural Networks, Genetic Algorithm and Partial Least-Squares in QSAR Modeling .....	288
Bono Lučić and Nenad Trinajstić	
Structure-Activity Relationships of Nitrofurans Derivatives with Antibacterial Activity .....	290
José Ricardo Pires, Astréa Giesbrecht, Suely L. Gomes, and Antonia T. do-Amaral	
QSAR Approach for the Selection of Congeneric Compounds with Similar Toxicological Modes of Action .....	292
Paola Gramatica, Federica Consolaro, Marco Vighi, Roberto Todeschini, Antonio Finizio, and Michael Faust	
Strategies for Selection of Test Compounds in Structure-Affinity Modelling of Active Carbon Adsorption Performance: A Multivariate Approach .....	293
L.-G. Hammarström, I. Fängmark, P. G. Jönsson, P. R. Norman, A. L. Ness, S. L. McFarlane, and N. M. Osmond	
Design and QSAR of Dihydropyrazolo[4,3-c]Quinolinones as PDE4 Inhibitors .....	295
M. López, V. Segarra, M. I. Crespo, J. Gràcia, T. Doménech, J. Beleta, H. Ryder, and J. M. Palacios	
QSAR Based on Biological Microcalorimetry: On the Study of the Interaction between Hydrazides and <i>Escherichia coli</i> and <i>Saccharomyces cerevisiae</i> .....	297
Maria Luiza Cruzera Montanari, Anthony Beezer, and Carlos Alberto Montanari	
Cinnoline Analogs of Quinolones: Structural Consequences of the N Atom Introduction in the Position 2 .....	299
Marek L. Główka, Dariusz Martynowski, Andrzej Olczak, and Alina Staszewska	

Joint Continuum Regression for Analysis of Multiple Responses .....	301
Martyn G. Ford, David W. Salt, and Jon Malpass	
Putative Pharmacophores for Flexible Pyrethroid Insecticides .....	303
Martyn G. Ford, Neil E. Hoare, Brian D. Hudson, Thomas G. Nevell, and John A. Wyatt	
Predicting Maximum Bioactivity of Dihydrofolate Reductase Inhibitors .....	305
Matevž Pompe, Marjana Novič, Jure Zupan, and Marjan Veber	
Evaluation of Carcinogenicity of the Elements by Using Nonlinear Mapping .....	307
Alexander A. Ivanov	

## Poster Session II: The Future of 3D-QSAR

Partition Coefficients of Binary Mixtures of Chemicals: Possibility for the QSAR Analysis .....	311
Miloň Tichý, Marián Rucki, Václav B. Dohalský, and Ladislav Feltl	
A CoMFA Study on Antileishmaniasis Bisamidines .....	314
Carlos Alberto Montanari	
Antileishmanial Chalcones: Statistical Design and 3D-QSAR Analysis .....	316
Simon F. Nielsen, S. Brøgger Christensen, A. Kharazmi, and T. Liljefors	
Chemical Function Based Alignment Generation for 3D QSAR of Highly Flexible Platelet Aggregation Inhibitors .....	318
Rémy D. Hoffmann, Thierry Langer, Peter Lukavsky, and Michael Winger	
3D QSAR on Mutagenic Heterocyclic Amines That are Substrates of Cytochrome P450 1A2 .....	321
Juan J. Lozano, Manuel Pastor, Federico Gago, Gabriele Cruciani, Nuria B. Centeno, and Ferran Sanz	
Application of 4D-QSAR Analysis to a Set of Prostaglandin, PGF <sub>2</sub> α, Analogs .....	323
C. Duraiswami, P. J. Madhav, and A. J. Hopfinger	
Determination of the Cholecalciferol-Lipid Complex Using a Combination of Comparative Modelling and NMR Spectroscopy .....	325
Mariagrazia Sarpietro, Mario Marino, Antonio Cambria, Gloria Uccello Barretta, Federica Balzano, and Salvatore Guccione	
Comparative Binding Energy (COMBINE) Analysis on a Series of Glycogen Phosphorylase Inhibitors: Comparison with GRID/GOLPE Models .....	329
Manuel Pastor, Federico Gago, and Gabriele Cruciani	
EVA QSAR: Development of Models with Enhanced Predictivity (EVA_GA) .....	331
David B. Turner and Peter Willett	
3D-QSAR, GRID Descriptors and Chemometric Tools in the Development of Selective Antagonists of Muscarinic Receptor .....	334
Paola Gratteri, Gabriele Cruciani, Serena Scapecchi, M. Novella Romanelli, and Fabrizio Melani	
Small Cyclic Peptide SAR Study Using APEX-3D System: Somatostatin Receptor Type 2 (SSTR2) Specific Pharmacophores .....	336
Larisa Golender, Rakefet Rosenfeld, and Erich R. Vorpagel	



3D Quantitative Structure-Activity Relationship (CoMFA) Study of Heterocyclic Arylpiperazine Derivatives with 5-HT <sub>1A</sub> Activity .....	338
Ildikó Magdó, István Laszlovszky, Tibor Ács, and György Domány	
Molecular Similarity Analysis and 3D-QSAR of Neonicotinoid Insecticides .....	340
Masayuki Sukekawa and Akira Nakayama	
3D-SAR Studies on a Series of Sulfonate Dyes as Protection Agents against $\beta$ -amyloid Induced <i>in Vitro</i> Neurotoxicity .....	342
M. G. Cima, G. Gallo, M. Mabilia, M. O. Tinti, M. Castorina, C. Pisano, and E. Tassoni	
A New Molecular Structure Representation: Spectral Weighted Molecular (SWM) Signals and Spectral Weighted Invariant Molecular (SWIM) Descriptors .....	344
Roberto Todeschini, Viviana Consonni, David Galvagni, and Paola Gramatica	
3D QSAR of Prolyl 4-Hydroxylase Inhibitors .....	345
K.-H. Baringhaus, V. Guenzler-Pukall, G. Schubert, and K. Weidmann	
Aromatase Inhibitors: Comparison between a CoMFA Model and the Enzyme Active Site .....	347
Andrea Cavalli, Maurizio Recanatini, Giovanni Greco, and Ettore Novellino	
Imidazoline Receptor Ligands—Molecular Modeling and 3D-QSAR CoMFA .....	349
C. Marot, N. Baurin, J. Y. Mérour, G. Guillaumet, P. Renard, and L. Morin-Allory	

### Poster Session III: Prediction of Ligand-Protein Binding

Reversible Inhibition of MAO-A and B by Diazoheterocyclic Compounds: Development of QSAR/CoMFA Models .....	353
Cosimo D. Altomare, Antonio Carrieri, Saverio Cellamare, Luciana Summo, Angelo Carotti, Pierre-Alain Carrupt, and Bernard Testa	
Modelling of the 5-HT <sub>2A</sub> Receptor and Its Ligand Complexes .....	355
Estrella Lozoya, Maria Isabel Loza, and Ferran Sanz	
Towards the Understanding of Species Selectivity and Resistance of Antimalarial DHFR Inhibitors .....	357
Thomas Lemcke, Inge Thøger Christensen, and Flemming Steen Jørgensen	
Modeling of Suramin-TNF $\alpha$ Interactions .....	359
Carola Marani Toro, Massimo Mabilia, Francesca Mancini, Marilena Giannangeli, and Claudio Milanese	
De Novo Design of Inhibitors of Protein Tyrosine Kinase pp60 <sup>c-src</sup> .....	361
T. Langer, M. A. König, G. Schischkow, and S. Guccione	
Elucidation of Active Conformations of Drugs Using Conformer Sampling by Molecular Dynamics Calculations and Molecular Overlay .....	363
Shuichi Hirono and Kazuhiko Iwase	
Differences in Agonist Binding Pattern for the GABA <sub>A</sub> and the AMPA Receptors Illustrated by High-Level <i>ab Initio</i> Calculations .....	365
Lena Tagmose, Lene Merete Hansen, Per-Ola Norrby, and Tommy Liljefors	
Stabilization of the Ammonium-Carboxylate Ion-Pair by an Aromatic Ring .....	367
Tommy Liljefors and Per-Ola Norrby	

Structural Requirements for Binding to Cannabinoid Receptors .....	369
Maria Fichera, Alfredo Bianchi, Gabriele Cruciani, and Giuseppe Musumarra	
Design, Synthesis, and Testing of Novel Inhibitors of Cell Adhesion .....	371
David T. Manallack, John G. Montana, Paul V. Murphy, Rod E. Hubbard, and Richard J. K. Taylor	
Conformational Analysis and Pharmacophore Identification of Potential Drugs for Osteoporosis.....	373
Jan Høst, Inge Thøger Christensen, and Flemming Steen Jørgensen	
Molecular Modelling Study of DNA Adducts of BBR3464: A New Phase I Clinical Agent.....	375
G. De Cillis, E. Fioravanzo, M. Mabilia, J. Cox, and N. Farrell	
Prediction of Activity for a Set of Flavonoids against HIV-1 Integrase.....	377
Jarmo Huuskonen, Heikki Vuorela, and Raimo Hiltunen	
Structure-Based Discovery of Inhibitors of an Essential Purine Salvage Enzyme in <i>Tritrichomonas foetus</i> .....	380
Ronald M. A. Knegtel, John R. Somoza, A. Geoffrey Skillman Jr., Narsimha Mungala, Connie M. Oshiro, Solomon Mpoke, Shinichi Katakura, Robert J. Fletterick, Irwin D. Kuntz, and Ching C. Wang	
A 3D-Pharmacophore Model for Dopamine D <sup>1</sup> Receptor Antagonists .....	382
Jonas Boström, Klaus Gundertofte, and Tommy Liljefors	
Molecular Modeling and Structure-Based Design of Direct Calcineurin Inhibitors .....	384
Xinjun J. Hou, John H. Tatlock, M. Angelica Linton, Charles R. Kissinger, Laura A. Pelletier, Richard E. Showalter, Anna Tempczyk, and J. Ernest Villafranca	
Conformational Flexibility and Receptor Interaction .....	386
Lambert H. M. Janssen	
Investigating the Mimetic Potential of $\beta$ -Turn Mimetics .....	388
Susanne Winiwarter, Anders Hallberg, and Anders Karlén	
Conformational Aspects of the Interaction of New 2,4-Dihydroxyacetophenone Derivatives with Leukotriene Receptors.....	390
Miroslav Kuchař, Antonín Jandera, Vojtěch Kmoníček, Bohumila Brůnová, and Bohdan Schneider	
Conformational Studies of Poly(Methylidene Malonate 2.1.2).....	393
Eric Vangrevelinghe, Pascal Breton, Nicole Bru, and Luc Morin-Allory	
A Peptidic Binding Site Model for PDE 4 Inhibitors .....	395
E. E. Polymeropoulos and N. Höfgen	
Molecular Dynamics Simulations of the Binding of GnRH to a Model GnRH Receptor.....	397
A.M. ter Laak, R. Kühne, G. Krause, E. E. Polymeropoulos, B. Kutscher, and E. Günther	
Analysis of Affinities of Penicillins for a Class C $\beta$ -Lactamase by Molecular Dynamics Simulations .....	399
Keiichi Tsuchida, Noriyuki Yamaotsu, and Shuichi Hirono	
Theoretical Approaches for Rational Design of Proteins .....	401
Jiří Damborský	

Amisulpride, Sultopride, and Sulpiride: Comparison of Conformational and Physico-Chemical Properties .....	404
Audrey Blomme, Laurence Conraux, Philippe Poirier, Anne Olivier, Jean-Jacques Koenig, Mireille Sevrin, François Durant, and Pascal George	
Entropic Trapping: Its Possible Role in Biochemical Systems .....	406
Adolf Miklavc and Darko Kocjan	
Structural Requirements to Obtain Potent CAXX Mimic p21-Ras-Farnesyltransferase Inhibitors.....	408
A. Laoui	
Hydrogen-Bonding Hotspots as an Aid for Site-Directed Drug Design .....	410
James E. J. Mills and Philip M. Dean	
Superposition of Flexible Ligands to Predict Positions of Receptor Hydrogen-Bonding Atoms .....	412
James E. J. Mills and Philip M. Dean	
Comparative Molecular Field Analysis of Multidrug Resistance Modifiers .....	414
Ilza K. Pajeva and Michael Wiese	
Pharmacophore Model of Endothelin Antagonists .....	416
Mitsuo Takahashi, Kuniya Sakurai, Seji Niwa, and Seiji Oono	
The Electron-Topological Method (ETM): Its Further Development and Use in the Problems of SAR Study.....	418
Nathaly M. Shvets and Anatholy S. Dimoglo	

#### **Poster Session IV: Computational Aspects of Molecular Diversity and Combinatorial Libraries**

MOLDIVS—A New Program for Molecular Similarity and Diversity Calculations.....	423
Vadim A. Gerasimenko, Sergei V. Trepalin, and Oleg A. Raevsky	
Easy Does It: Reducing Complexity in Ligand-Protein Docking .....	425
Djamal Bouzida, Daniel K. Gehlhaar, and Paul A. Rejto	
Study of the Molecular Similarity among Three HIV Reverse Transcriptase Inhibitors in Order to Validate GAGS, a Genetic Algorithm for Graph Similarity Search.....	427
Nathalie Meurice, Gerald M. Maggiora, and Daniel P. Vercauteren	
A Decision Tree Learning Approach for the Classification and Analysis of High-Throughput Screening Data.....	429
Michael F. M. Engels, Hans De Winter, and Jan P. Tollenaere	

#### **Poster Session V: Affinity and Efficacy Models of G-Protein Coupled Receptors**

Application of PARM to Constructing and Comparing 5-HT <sub>1A</sub> and $\alpha_1$ Receptor Models....	433
Maria Santagati, Hongming Chen, Andrea Santagati, Maria Modica, Salvatore Guccione, Gloria Uccello Barretta, and Federica Balzano	
A Novel Computational Method for Predicting the Transmembranal Structure of G-Protein Coupled Anaphylatoxin Receptors, C5AR and C3AR.....	440
Naomi Siew, Anwar Rayan, Wilfried Bautsch, and Amiram Goldblum	
Receptor-Based Molecular Diversity: Analysis of HIV Protease Inhibitors .....	442
Tim D. J. Perkins, Nasfim Haque, and Philip M. Dean	

Application of Self-Organizing Neural Networks with Active Neurons for QSAR Studies .....	444
Vasyl V. Kovalishyn, Igor V. Tetko, Alexander I. Luik, Alexey G. Ivakhnenko, and David J. Livingstone	
Application of Artificial Neural Networks in QSAR of a New Model of Phenylpiperazine Derivatives with Affinity for 5-HT <sub>1A</sub> and $\alpha_1$ Receptors: A Comparison of ANN Models .....	446
María L. López-Rodríguez, M. Luisa Rosado, M. José Morcillo, Esther Fernandez, and Klaus-Jürgen Schaper	
Atypical Antipsychotics: Modelling and QSAR .....	448
Benjamin G. Tehan, Margaret G. Wong, Graeme J. Cross, and Edward J. Lloyd	

## Poster Session VI: New Methods in Drug Discovery

Genetic Algorithms: Results Too Good To Be True? .....	453
M. G. B. Drew, J. A. Lumley, N. R. Price, and R. W. Watkins	
Property Patches in GPCRs: A Multivariate Study .....	455
Per Källblad and Philip M. Dean	
A Stochastic Method for the Positioning of Protons in X-Ray Structures of Biomolecules .....	458
M. Glick and Amiram Goldblum	
Molecular Field Topology Analysis (MFTA) as the Basis for Molecular Design .....	460
Eugene V. Radchenko, Vladimir A. Palyulin, and Nikolai S. Zefirov	
Rank Distance Clustering—A New Method for the Analysis of Embedded Activity Data ..	462
John Wood and Valerie S. Rose	
The Application of Machine Learning Algorithms to Detect Chemical Properties Responsible for Carcinogenicity .....	464
C. Helma, E. Gottmann, S. Kramer, and B. Pfahringer	
Study of Geometrical/Electronic Structures—Carcinogenic Potency Relationship with Counterpropagation Neural Networks .....	466
Marjan Vračko	
Combining Molecular Modelling with the Use of Artificial Neural Networks as an Approach to Predicting Substituent Constants and Bioactivity .....	468
Igor I. Baskin, Svetlana V. Keschtova, Vladimir A. Palyulin, and Nikolai S. Zefirov	
Application of Neural Networks for Calculating Partition Coefficient Based on Atom-Type Electrotopological State Indices .....	470
Jarmo J. Huuskonen and Igor V. Tetko	
Variable Selection in the Cascade-Correlation Learning Architecture .....	472
Igor V. Tetko, Vasyl V. Kovalishyn, Alexander I. Luik, Tamara N. Kasheva, Alessandro E. P. Villa, and David J. Livingstone	
Chemical Fingerprints Containing Biological and Other Non-Structural Data .....	474
Fergus Lippi, David Salt, Martyn Ford, and John Bradshaw	
Rodent Tumor Profiles Induced by 536 Chemical Carcinogens: An Information Intense Analysis .....	476
R. Benigni, A. Pino, and A. Giuliani	

Comparison of Several Ligands for the 5-HT <sub>1D</sub> Receptor Using the Kohonen Self-Organizing-Maps Technique .....	478
Joachim Petit and Daniel P. Vercauteren	
Binding Energy Studies on the Interaction between Berenil Derivatives and Thrombin and the B-DNA Dodecamer D(CGCGAATTCGCG) <sub>2</sub> .....	480
Júlio C. D. Lopes, Ramon K. da Rocha, Andrelly M. José, and Carlos A. Montanari	
A Comparison of <i>ab Initio</i> , Semi-Empirical, and Molecular Mechanics Approaches to Compute Molecular Geometries and Electrostatic Descriptors of Heteroatomic Ring Fragments Observed in Drug Molecules.....	482
G. Longfils, F. Ooms, J. Wouters, A. Olivier, M. Sevrin, P. George, and F. Durant	
Elaboration of an Interaction Model between Zolpidem and the $\omega_1$ Modulatory Site of GABA <sub>A</sub> Receptor Using Site-Directed Mutagenesis.....	484
A. Olivier, S. Renard, Y. Even, F. Besnard, D. Graham, M. Sevrin, and P. George	
 <b>Poster Session VII: Modeling of Membrane Penetration</b>	
SLIPPER—A New Program for Water Solubility, Lipophilicity, and Permeability Prediction .....	489
O. A. Raevsky, E. P. Trepalina, and S. V. Trepalin	
Correlation of Intestinal Drug Permeability in Humans ( <i>in Vivo</i> ) with Experimentally and Theoretically Derived Parameters.....	491
Anders Karlén, Susanne Winiwarter, Nicholas Bonham, Hans Lennernäs, and Anders Hallberg	
A Critical Appraisal of logP Calculation Procedures Using Experimental Octanol-Water and Cyclohexane-Water Partition Coefficients and HPLC Capacity Factors for a Series of Indole Containing Derivatives of 1,3,4-Thiadiazole and 1,2,4-Triazole ....	493
Athanasia Varvaresou, Anna Tsantili-Kakoulidou, and Theodora Siatra-Papastaikoudi	
Determination of Accurate Thermodynamics of Binding for Proteinase-Inhibitor Interactions.....	495
Frank Dullweber, Franz W. Sevenich, and Gerhard Klebe	
 <b>Author Index</b> .....	 497
<b>Subject Index</b> .....	501

# **Section I**

## **Overview**

## STRATEGIES FOR MOLECULAR DESIGN BEYOND THE MILLENNIUM

James P. Snyder and Forrest D. Snyder

Department of Chemistry, Emory University  
1515 Pierce Drive, Atlanta, GA 30322  
e-mail: snyder@euch4e.chem.emory.edu

### INTRODUCTION

When asked to open the 12<sup>th</sup> European Symposium on QSAR with some projections into the years ahead, I was immediately drawn to the words of Niels Bohr who changed the face of science so many years ago.

“Predictions are difficult, especially about the future.”

Bohr, of course, was awarded the Nobel Prize in 1922 for work on the quantum model of atomic structure; work performed in the city of our gathering, Copenhagen, Denmark. The complementary fields of molecular modeling and QSAR are amply summarized elsewhere.<sup>1</sup> Rather than attempt a comprehensive survey, I decided to tell a few stories as representative of current developments that may have a strong influence in the field for the decade ahead. Thus, four themes will be touched in the paragraphs to follow: 1) Receptor structure – molecular detail; 2) Molecular design and re-design; 3) Bioavailability and other imponderables; 4) The human factor. To test Bohr’s proposition, at the end of each theme, a set of near-future predictions will be ventured.

### RECEPTOR STRUCTURE – MOLECULAR DETAIL

At the present time there are four experimental methods that provide atomic resolution for molecules of biological interest: X-ray crystallography,<sup>2</sup>

neutron diffraction,<sup>3</sup> nuclear magnetic resonance spectroscopy<sup>4</sup> and high resolution electron microscopy, also referred to as electron crystallography.<sup>5</sup> The latter differs from X-ray spectroscopy by deconvoluting electron diffraction rather than X-ray diffraction patterns. Complementary methodologies for protein structure that depend on knowledge of the structure of a related protein are homology modeling and threading.<sup>6</sup> While the three-dimensional structures of more than 7600 soluble proteins, protein-nucleotide aggregates and protein-ligand complexes are known,<sup>7</sup> the X-ray crystal structures of only ten different types of membrane bound proteins have been solved to date (Table 1).

**Table 1.** X-ray crystal structures of proteins with a membrane embedded domain <sup>a,b</sup>

Protein	R, <sup>c</sup> Å	Year of publication
Bacteriorhodopsin <sup>8</sup>	2.5	1997
Bacterial photoreaction centers <sup>9</sup>	2.2-3.1	1984, 1986, 1993, 1994, 1996
Light harvesting complexes <sup>10</sup>	2.5	1995, 1996
Photosystem I <sup>11</sup>	4.0	1996
Porins <sup>12</sup>	1.8-3.1	1991, 1992, 1994, 1995, 1997, 1998
Alpha-hemolysin <sup>13</sup>	1.9	1996
Prostaglandin synthase-I <sup>14</sup>	3.5	1993
Prostaglandin synthase-II <sup>15</sup>	2.5-3.0	1996
Cytochrome c oxidase <sup>16</sup>	2.8	1995
Cytochrome bc <sub>1</sub> complex <sup>17</sup>	2.8-3.0	1996, 1997, 1998

<sup>a</sup> Table adapted from P. C. Preusch, J. C. Norvell, J. C. Cassatt, M. Cassman, *Int. Union Cryst. Newsletter* **1998**, 6, 19; <sup>b</sup> Literature citations in REFERENCES, <sup>c</sup> Structure resolution.

Each of these crystal structures provides exquisite detail. An illustrative example is the cytochrome c oxidase complex (CcO) located at the terminus of the electron transport chain in the oxidative phosphorylation pathway. The structure reveals the domains of the enzyme within the mitochondrial inner membrane as well as those projecting on both sides of it. The location of both hemes and the two copper sites (Cu<sub>A</sub> and Cu<sub>B</sub>) provides a clear spatial picture of the relay of electrons from the external and mobile cytochrome *c* to the first metal center (Cu<sub>A</sub>), which passes them to the heme iron of cytochrome *a*.



Finally, the electrons are delivered to the third metal center containing a closely associated iron-heme (cytochrome  $a_3$ ) and a ligated copper atom. It is here that  $O_2$  is converted to water with concomitant priming of the proton pump responsible for production of ATP. Among many other things, the structure resolved a long standing problem as to precisely how many copper atoms occupy the  $Cu_A$  site; two.

This level of molecular detail is eagerly sought for proteins that form unique membrane spanning structures arising from multiple passage across the bilayer. Examples<sup>18</sup> include the 24-strand sodium channel  $\alpha$ -subunit, a 14-strand anion transport protein and the 12-strand a-factor and the dopamine transport protein. The structure in each case is believed to consist of membrane-embedded  $\alpha$ -helices. By contrast, the 16-strand *E. coli*. transport protein, *PhoE*, which employs  $\beta$ -sheets as membrane spanners. At present, the somewhat less complex 7-transmembrane G-protein coupled receptors that transmit the messages of numerous polypeptide hormones and other small molecules such as acetylcholine, dopamine and serotonin are of prime interest.

### Electron Crystallography - The Tubulin Dimer

The question posed here is whether high-resolution electron microscopy can provide 7-TM GPCR structure in the near future. Generally, one thinks of EM as a tool for observing small whole organisms in great detail: insect eyes, blood cells, bacteria and viruses to name a few.<sup>19</sup> During the past decade or so, however, a number of developments have converged to increase the resolution of EM to below 5 Å. Small well-ordered molecular crystals can yield structures to 1-2 Å resolution.<sup>20</sup> A spectacular example is the structure of the inorganic solid  $Ti_{11}Se_4$  which has been solved to an accuracy of 0.02 Å resolution.<sup>21</sup> At this level of accuracy, the technique is justifiably referred to as electron crystallography (EC). While many large biomolecular aggregates have been solved in the at 10-40 Å range, the structures of three proteins have been obtained at < 4 Å resolution: bacteriorhodopsin (3.5 Å),<sup>22</sup> spinach light-harvesting complex (3.4 Å)<sup>23</sup> and the  $\alpha,\beta$  tubulin dimer (3.7 Å).<sup>24</sup> The first two, bR and LHC respectively, are membrane-bound proteins. EC would appear to be a natural technique for the latter as it requires the preparation of 2-D crystals for which extended lipid layers are eminently suitable. The third soluble protein, the primary constituent of microtubules, is three times larger than bR and four times larger than LHC. Determination of the tubulin dimer structure including molecules of bound GDP and GTP is a landmark for both biology and electron crystallography.

Apart from the raw size of the  $\alpha,\beta$  tubulin dimer, another aspect of the structure justifies discussion. The 2-D crystal used in the EC analysis was

stabilized by taxol, a marketed drug that arrests a variety of cancers presumably by blocking the depolymerization of microtubules during cell division.<sup>25</sup> The Nature report that describes the dimer structure includes the small X-ray structure of a taxol surrogate, taxotere, docked in the taxol binding site. Unfortunately, the electron density of the ligand is insufficient to define the conformation of the three taxol side chains. As part of a collaboration with the Berkeley EC group, we have assembled nearly two dozen empirically viable conformations of taxol derived from pharmacophore mapping, 2-D NOE NMR analysis and the small molecule X-ray crystallographic literature. These were individually fitted to the partial electron density in the taxol-tubulin EC structure and ranked for goodness of fit.<sup>26</sup> Only one of the conformers matches the density, a molecular shape distinct from previous proposals for the bioactive conformation of taxol.<sup>27</sup> An important lesson from this study is the possibility for determining binding site ligand conformation in favorable cases by combining the results of a high resolution EC protein-ligand structure with those from small molecule modeling. Were electron crystallography to be successful in solving 7-TM GPCR structure at 3-4 Å resolution, a similar synergy between structure determination and modeling can be anticipated.

## SAR by NMR

A separate but tantalizing recent development in spectroscopy is SAR by NMR, a creation of the Abbott NMR group.<sup>28</sup> In principle, the technique is simple. A library of small molecules is presented to a protein. Both the location of the binding site and the corresponding  $K_D$  is sampled by <sup>15</sup>N NMR. The ability to treat compounds binding in the low potency  $\mu\text{M}$ -mM range is a highlight of the method. Once a pair of suitable molecules are located in contiguous sites, linkers are introduced synthetically. Discovery of nonpeptide inhibitors in the low nM range for stromelysin,<sup>28c</sup> a matrix metalloproteinase, and the FK506 binding protein has been achieved in this manner.<sup>28b</sup> The NMR-based approach has its counterparts in the area of purely computational *de novo* design. MCSS/HOOK,<sup>29</sup> LUDI<sup>30</sup> and Agouron's approach whimsically labeled "virtual SAR by NMR"<sup>31</sup> all operate by docking small molecules in a protein binding site, ranking them with a free-energy scoring function, connecting them with appropriate spacers and reevaluating the composite structures for improved binding affinity. While the Agouron workers have succeeded in mimicking the Abbott results entirely within the computer, the *de novo* approaches have yet to make a substantial impact on the drug candidate pipeline.

## Predictions

- 2-D Crystals of proteins in planar lipid films will become routinely accessible.<sup>32</sup> Electron crystallography will employ novel 2-D crystal preparations to provide an increasing number of membrane-bound protein structures, including 7TM GPCRs.
- Electron crystallography in combination with small molecule conformational analysis will provide ligand conformation for membrane-bound proteins.
- SAR by NMR will become a widely used technique for protein-bound ligand conformer analysis and ligand design.

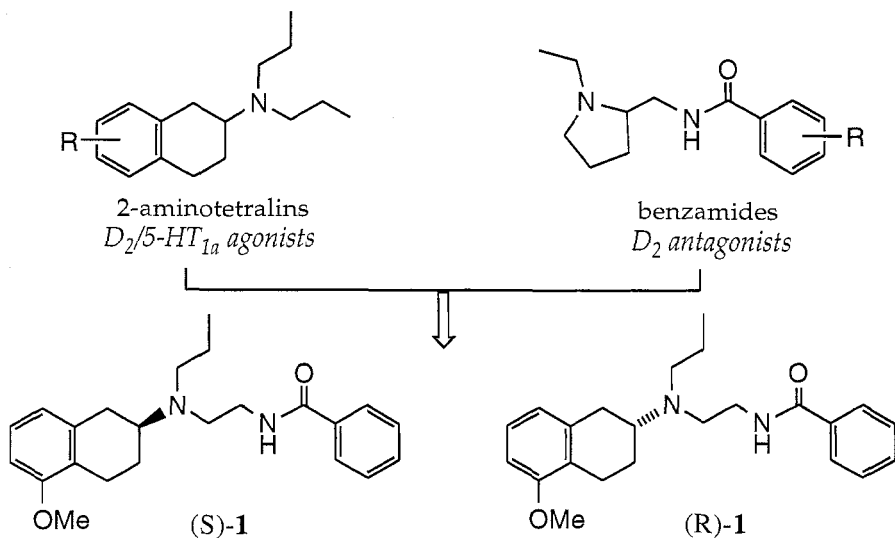
## MOLECULAR DESIGN AND RE-DESIGN

Sequences for numerous G-protein coupled receptors are now known, as is the influence of an impressive amount of point mutation data on ligand binding.<sup>33</sup> Many molecular models of the GPCRs have been constructed by homology with bR, a protein uncoupled to a G-protein. Justification follows from the bR 7-TM motif and knowledge that mammalian opsins, true members of the GPCR family, may form an evolutionary link between bR and the ligand-binding GPCRs.<sup>34</sup> Independently, the SAR of chiral small-molecule drug leads has stimulated the development of pharmacophores that include both weak and potent ligands.

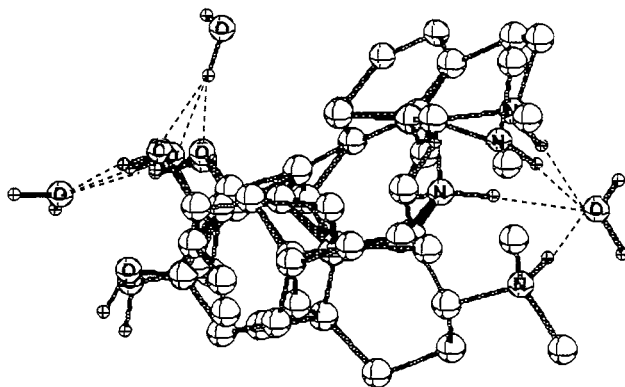
One approach to understanding drug action at structurally ill-defined macro-molecular receptors combines the features of modeled proteins and pharmacophores. The unified methodology provides novel design opportunities by borrowing the strengths of each of the latter. To my knowledge this concept was first presented by the Uppsala group.<sup>35</sup> In the following, two separate stories are intertwined to illustrate a pathway from GPCR sequence to semi-quantitative structure-based design.

### Mixed Dopamine Antagonists and Serotonin Agonists

The first thread in the weave takes its inspiration from studies by the Groningen group.<sup>36</sup> The just printed Ph.D. thesis of Evert Homan explores drug remedies for schizophrenia by focusing on atypical antipsychotic agents.<sup>37</sup> In particular, attempts to prepare mixed dopamine D<sub>2</sub> receptor antagonists and serotonin 5-HT<sub>1a</sub> agonists sprung from hybrids of substituted benzamides (D<sub>2</sub> antagonists) and 2-aminotetralins (5-HT<sub>1a</sub> agonists). Enantiomers (R)-1 and (S)-1, among others, were shown to exhibit the relevant biology.

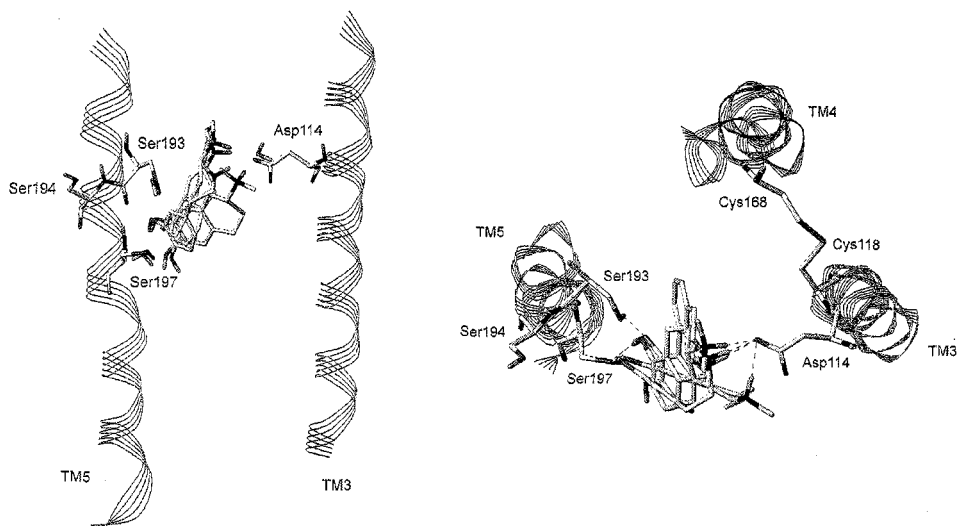


Using Macromodel<sup>38</sup> and APOLLO<sup>39</sup> software and a carefully selected set of active compounds, Homan developed independent pharmacophores for the D and 5-HT receptor subtypes (Figure 1). The unexceptional pharmacophores are complemented by the placement of water molecules at sites where the protein ligand side chain atoms of the putative biological receptor would interact with individual bound ligands.



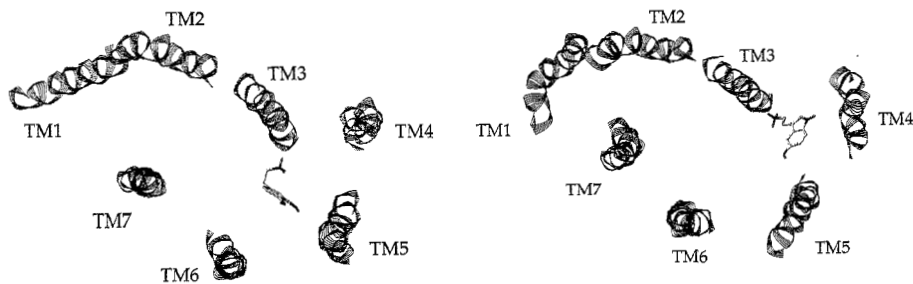
**Figure 1.** Superposition of several dopamine agonists in their pharmacophore derived dopamine  $D_2$  receptor binding conformations. The water molecules mimic putative amino acid residues from the receptor capable of forming hydrogen bonds with the ligands.

In a second modeling exercise, helices for the two 7TM receptors were constructed by sequence alignment and homology with bR and subsequently rhodopsin by means of Sybyl software.<sup>40</sup> These were then docked around the pharmacophores by employing the conserved residues in both receptors as anchor points. For example, the conserved Asp114 located on TM3 in the D<sub>2</sub> receptor was positioned to replace the pharmacophore water molecule coordinated to the aromatic OH groups. Similarly, TM5 was positioned to permit Ser193 and Ser197 to replace the remaining pharmacophore receptor site waters as shown in Figure 2.



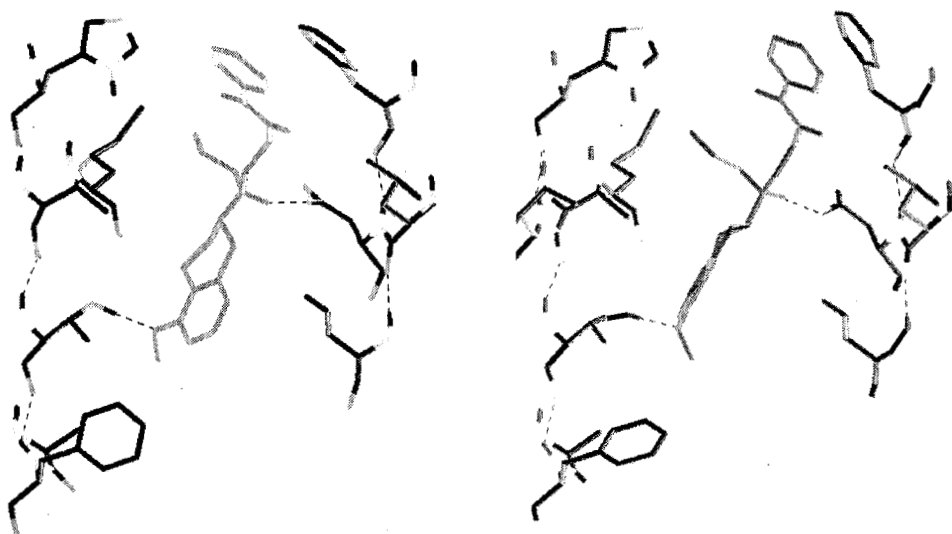
**Figure 2.** Illustration of the stepwise construction of the dopamine D<sub>2</sub> receptor model. The diagram at left shows the positioning of TM3 and TM5 helices with the aid of the pharmacophore water molecules. The diagram at right offers a top-to-bottom view of the relative positions of TM3, TM4 and TM5. The TM4 location was guided by the formation of a disulfide bridge between Cys118 in TM3 and Cys168 in TM4. TM domain backbones are displayed as line ribbons.

A consistent build-up procedure led to the D<sub>2</sub> and 5-HT<sub>1a</sub> 7TM models illustrated in Figure 3. While details of synthesis, biotesting and modeling can be found in the original Groningen publications,<sup>41</sup> it's clear that the receptor ligand complexes derived by the hybrid procedure are substantially different from the bR model, but similar to the Herzyk-Hubbard rhodopsin model.<sup>42</sup>



**Figure 3.** Topological arrangements of the TM domains of the final 7TM models of the dopamine D<sub>2</sub> (left) and serotonin 5-HT<sub>1a</sub> (right) receptors. Backbones of the TM domains are displayed as line ribbons.

Additional ligands including (R)-1 and (S)-1 were docked into the 7TM receptor. The entire binding pocket including ligands and interacting receptor side chains was subsequently extracted and transferred to the PrGen software for optimization of the individual ligand-receptor interactions.<sup>43</sup> Final 5-HT<sub>1a</sub> binding site minireceptor models are illustrated in Figure 4. Both enantiomers enjoy identical hydrophobic and hydrogen-bonding interactions with the receptor side chains, a result achieved by the molecules' adoption of diastereomeric conformations near the stereogenic carbon. The modeling outcome is consistent with the observation that both compounds are nearly equipotent agonists at this receptor subtype.



**Figure 4.** (S)-1 and (R)-1 in the optimized 5-HT<sub>1a</sub> minireceptor binding site model.

The same mirror image molecules at the modeled D<sub>2</sub> receptor provide a qualitatively different picture. The (S)-1 agonist participates in four clear-cut hydrogen bonds and a series of hydrophobic contacts (Figure 5). By contrast, the (R)-1 antagonist differs by failing to present a hydrogen bond from its 5-methoxy group on the left side of the diagram. Is this configurationally and conformationally determined difference responsible for the transition from agonist to antagonist in 1? It would be difficult to judge unless the binding site were coupled dynamically to a molecular-based signal transducing mechanism. Nevertheless, the Groningen modeling exercise is remarkably faithful to the types of variations in nonbonded ligand-receptor interactions expected to be responsible for stabilization of receptor conformations representing active and inactive 7TM forms.

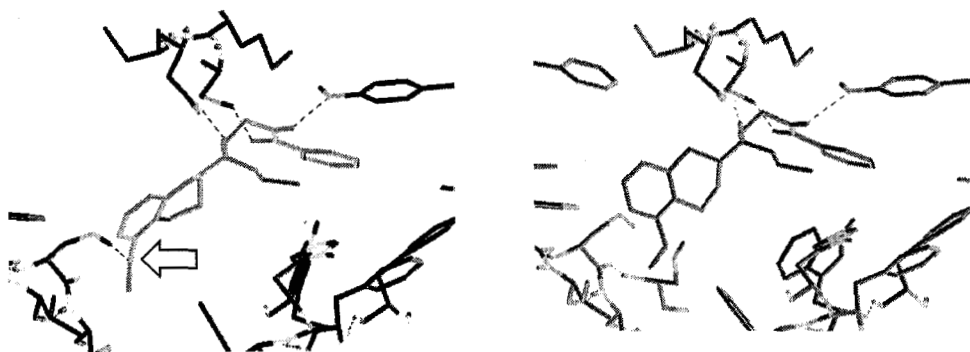


Figure 5. (S)-1 and (R)-1 in the optimized D<sub>2</sub> minireceptor binding site model. The bold arrow at left indicates the additional hydrogen-bond established by the S-enantiomer.

The minireceptors depicted in Figures 4 and 5 are suitable for exploitation by methods germane to structure-based design, namely 3-D database searching and *de novo* design. While these lead-seeking activities were not pursued in the Groningen study, we shift targets to show how refined minireceptors could have served this purpose here and can do so in other therapeutic areas.

### Vasopressin Antagonists

The second thread in the weave was stimulated by work at Emory University. The peptide hormone arginine vasopressin (AVP) operates in the central nervous system, the cardiovascular bed and the kidney. In the latter organ AVP serves to regulate water balance by causing GPCR-activated synthesis of cAMP, the deposition of aquaporins (water channels) in the cell membrane and the subsequent reabsorption of water on its way to the urinary

tract. Blockade of  $V_2$  receptors may prove useful in treating disorders characterized by excess renal absorption of water. Congestive heart failure, liver cirrhosis and CNS injuries are among them.

Accordingly, a  $V_2$  receptor pharmacophore was developed and augmented by constructing the corresponding PrGen optimized antagonist minireceptor without resorting to a preliminary 7TM model. In turn, the minireceptor was further refined to provide a semiquantitative correlation of empirical and calculated binding free energies.<sup>44</sup> The training set  $K_i$ 's span seven orders of magnitude (from low mM to sub nM) corresponding to a  $\Delta\Delta G_{\text{bind}}$  range of 6.5 Kcal/mol ( $R = 0.99$ ,  $\text{rms} = -0.41$  Kcal/mol). So far, the 3-D QSAR model has been utilized in two ways. First, a close collaboration between synthetic chemists and computational chemists has led to the intuitive and interactive conception of several novel series of analogs. Each candidate for synthesis has been subjected to a full conformational analysis, conformer screening and  $K_i$  prediction by the model. A set of candidate antagonists with a predicted  $K_i \geq 10^{*-8}$  were synthesized and challenged by three separate in vitro bioassays. Although the work is still preliminary, more than 50% of the 22 compounds tested proved to be strong  $V_2$  antagonists at low nM concentrations.<sup>45</sup> Further work is underway to demonstrate selectivity and to incorporate favorable ADME (absorption, distribution, metabolism, elimination) properties.

Second, the  $V_2$  minireceptor has been subjected to a flexible 3-D search of the Chapman Hall Database of natural products by means of the Tripos Unity software. Of the 83,000 compounds sampled in this database, forty-five simultaneously matched the pharmacophore spatial characteristics and the minireceptor occupied space.<sup>40,46</sup> The next phase of the project will subject the best candidates to the  $K_i$  prediction protocol to select further structures for synthesis and assay. We expect the project to iterate several times and to incorporate combinatorial library steps before a selective, bioavailable development candidate is designated for toxicity screening.

## Generalization

The dopamine/serotonin and vasopressin ligand vignettes illustrate a general problem and a powerful solution when one is confronted with a molecular design challenge for a structurally undetermined receptor protein target. The problem, of course, is the lack of 3-D atomic coordinates for the protein. The solution is either to combine a rough 7TM GPCR model with a pharmacophore or to construct an *ad hoc* minireceptor around the pharmacophore. In either case, the optimized ligand-based binding pocket offers the potential to generate a predictive  $K_i / \Delta G_{\text{bind}}$  correlation. With both



the latter and a binding site model, the tools of structure-based design can now be employed in what formerly was a receptor mapping context. To be sure, a largely empirical combinatorial library approach can generate novel leads and a useful SAR.<sup>47</sup> Some research centers are gambling that the same combinatorial methods will provide refined development candidates without intervention of the modeling/QSAR/design steps. In this context, the computational chemist's priorities are naturally shifted entirely to the task of virtual library design. Only time will tell if such "combinatorial" optimism is warranted.

## Predictions

- Complex pharmacophores will be developed routinely by expert systems utilizing genetic algorithms and neural networks.
- Problem oriented but structurally diverse 3-D databases will be scanned and sorted for leads and backups by employing highly accurate docking methods and much improved  $K_i / \Delta G_{\text{bind}}$  scoring functions. *De novo* design technology will mature.
- Computers and robots will be linked to analyze SAR, develop hypotheses and synthesize/screen iteratively on massively parallel computer chips. The first lead-finding step, but not subsequent steps in drug discovery, will be fully automated.
- The Sea's natural products will succeed in supplying novel and therapeutically useful molecular structures far beyond previous yields from the forests and soil sample microorganisms.<sup>48</sup>

## DRUG ORAL ACTIVITY

Bioavailability can be defined as the dissemination of a drug from its site of administration into the systemic circulation. For effective oral delivery the agent must be absorbed across the GI tract's small intestine, traverse the portal vein and endure the liver's 'first pass' metabolism. Only then does it enter the bloodstream.<sup>49</sup> The drug discovery and refinement methods described above are focused almost entirely on compound potency once the drug arrives at its site of action. Much needed are early predictors of absorption, distribution, metabolism and elimination (i.e. ADME), the vital pharmacokinetic factors that govern movement of drug from application site to action site. One very recent attempt to devise a broadly applicable guideline during the lead generation phase is the "Rule of 5".<sup>50</sup> Developed by Pfizer researchers, the measure suggests that poor absorption of a drug is more likely when its structure is characterized by i) MW > 500, ii) log P > 5, iii) more than 5 H-bond donors expressed as the sum of NHs and OHs, and iv) more than 10 H-bond

acceptors expressed as the sum of Ns and Os. The data supporting this simple analysis was taken from 2200 compounds in the World Drug Index, the "USAN/INN" collection. Since each of the substances had survived Phase I testing and were scheduled for Phase II evaluation, it was assumed that they possess desirable oral properties. Statistical analysis of the collection scored by the Rule of 5 demonstrated that less than 10% of the compounds show a combination of any two of the four parameters outside the desirable ranges. With the exception of substrates for bio-transformers, the Pfizer group recommended the following to their colleagues: "Any designed or purchased compound that shows two undesirable parameters be struck from the priority list for synthesis to assure downstream solubility and bioavailability." To be sure, compounds that pass this test do not necessarily show acceptable bioavailability. The purpose of the rule is to eliminate weak candidates from a larger collection of potential leads and backups. In this way the prospects for oral activity through enhanced solubility and permeability are improved simultaneous with potency increases designed to achieve the same goal.

While the Rule of 5, if applied judiciously, is certain to be of value, the need for protocols to make specific and accurate predictions of aqueous solubility, permeability and ADME factors is still great. Lipophilicity predictions as measured by log P, though not perfect, are highly developed.<sup>51</sup> A number of schemes for estimating aqueous solubility have been devised, but none in the open literature appear to treat complex drug structure accurately.<sup>52</sup> In the present meeting a number of promising schemes based both on descriptor derivation and physical chemical principles offer possibilities for addressing some of the key issues: solubility,<sup>53,54,55</sup> permeability,<sup>53,54,55,56</sup> intestinal absorption,<sup>57,58</sup> oral bioavailability.<sup>59</sup> Only application in a vigorous program of molecular design, synthesis and bioassay can elicit a judgment on the predictability and durability of the evolving methods.

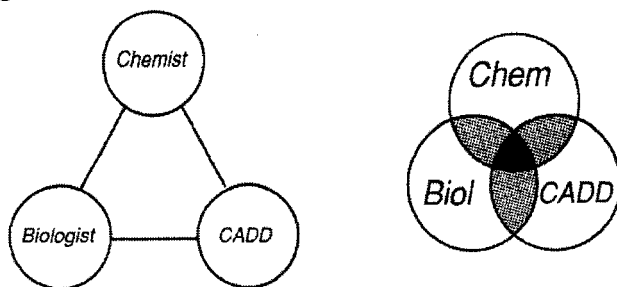
## Predictions

- Reliable methods for estimating drug absorption and permeability (e.g. as measured by CaCo-2 cells) will appear shortly. The current limitation is insufficient data.
- A combination of computers, synthesis robots, high capacity screening and design feedback loops should furnish potent lead compounds with optimal bioavailability qualities. Thus, auto-combinatorial methods will expand beyond potency screening.
- Metabolism and toxicity are more difficult, though modest progress has been made.<sup>60</sup> In the near future, experiments focused on specific lead

compounds and lead series will continue to be a necessity. The *next human generation* will enjoy useful correlations and accurate predictors.

## THE HUMAN FACTOR

Eight years ago I wrote of the need for a tight couple among chemists, biologists and computational scientists in order to create a seamless interdisciplinary interface and to heighten the chances for discovery of new therapeutic agents.



It was concluded that "At the level CADD groups are presently integrated throughout industry, there is little chance they will make a fundamental impact on drug discovery in the short term." However, a note of conditional optimism was sounded. "If management and synthetic chemists with decision-making responsibility commit to a true, collaborative integration of CADD into the research process, the current peripheral emphasis can be redirected with potential major consequences for the drug industry."<sup>61</sup>

The results have been spotty. To be sure, compounds reaching development can be identified as having their roots in collaborative encounters.<sup>62</sup> However, in spite of the fact that the great majority of pharmaceutical firms maintain a CADD group, "major consequences" have yet to materialize. Part of the reason, of course, is that computational models, like all models, are born with flaws and wide-ranging assumptions. Imaginative and effective use requires a deep knowledge of all aspects of the chemistry and biology of a project, superior judgement and persistence. Individual CADD practitioners can be faulted for the former. Anecdotes from industry suggest that persistence, follow-through and the necessary iteration are still hampered to a large degree by skepticism from experimentalists concerning the potential of modeling-based molecular design. Such skepticism combined with weak project management is, of course, self-fulfilling. In some quarters, modeling groups have consequently been diverted from the molecular design function and refocused on the fabrication of virtual combinatorial libraries.<sup>63</sup> Simultaneously, a cottage industry providing libraries-for-sale has sprung up. The new companies, many supporting the larger pharmaceutical firms with

full development and clinical resources, likewise employ computational chemists. Although it is still too early to tell, it may be here that CADD researchers prove to be a major driving force in the discovery effort.

## Predictions

- Given the natural tension between components of human behavior that regulate competition on the one hand and sharing on the other, and the lack of full-fledged management efforts to channel it, not much change in multidisciplinary molecular design collaboration can be expected in the short term.
- *Possible exceptions* The Scandinavian countries, small well-managed biotech start-ups, exceptionally well-coordinated units in large pharma and the emerging combinatorial library industry.
- Introduction of individual interactive audio & visual communication across computer networks may introduce new variables into the sharing process.

## CONCLUSIONS

In spite of the world economies' present and uncertain struggle with global capitalism, Europe's tentative feints toward unification and the lingering annoyance of Y2K, the twenty-first century ought to be anticipated with optimism. Our technical future appears very bright, indeed. Deconvolution of the human genome will provide uncountable opportunities for drug therapy, immune system regulation and "quality of life" experimentation. Discrete genes will provide protein sequences, which can be expected, in turn, to rapidly yield 3-D structures for both soluble and membrane-embedded entities. Thus, the number of health-related targets will increase as will information-rich intervention strategies. Tools of the QSAR and pharmaceutical trades will be exquisitely sharpened to permit accurate predictions of structure, potency, efficacy, selectivity, resistance, bioavailability and, ultimately, metabolism and side-effects sometime during the coming century.

One is reminded of "Ancient Man", an impressive late-eighteenth century painting by the British painter-poet, William Blake. Created at a moment of emergence for modern science, the work depicts ancient man "compelled to live the restrained life of reason as opposed to the free life of imagination. The colossal figure holds the compass down onto the black emptiness below him, perhaps symbolizing the imposition of order on chaos."<sup>64</sup> Clearly, in the twenty-first century the imposition of control over

biological and other events will require the exercise of both reason and imagination.

## ACKNOWLEDGEMENTS

I'm particularly grateful to Dr. Evert Homan and Professors Håkan Wikström and Cor Grol (University of Groningen, The Netherlands) for permission to discuss their mixed dopamine antagonist and serotonin agonist work prior to publication. Professor Marek Glówka (Technical University, Lodz) graciously pointed out the wealth of data found in Table 1, while Dr. Peter Preusch (NIGMS, NIH) generously provided access to its literature.

## REFERENCES

1. a) 3D QSAR in Drug Design, H. Kubinyi, ed., ESCOM, Leiden, 1993; b) 3D QSAR in Drug Design, H. Kubinyi, G. Folkers, Y. C. Martin, eds., ESCOM, Leiden, Vol. 2 & 3, 1998; c) 3D QSAR in Drug Design: Recent Advances. Perspectives in Drug Discovery (PD3), H. Kubinyi, G. Folkers, Y. C. Martin, eds, Kluwer/ESCOM, Vols 12-14, January 1998.
2. a) J. R. Helliwell, and M. Helliwell, X-Ray crystallography in structural chemistry and molecular biology. *Chem. Commun.* **1996**, 1595-1602; b) D. Ringe, G. A. Petsko. A consumer's guide to protein crystallography. *Protein Eng. Des.* **1996**, 205-229; c) A. D. Robertson, K. P. Murphy, Protein structure and the energetics of protein stability. *Chem. Rev.* **1997**, *97*, 1251-1267.
3. a) T. Hauss, Cold neutron diffraction in biology. *PSI-Proc.* **1997**, *97*, 205-218; b) J. R Helliwell. Neutron Laue diffraction does it faster. *Nat. Struct. Biol.* **1997**, *4*, 874-876.
4. a) K.H. Gardner, L. E. Kay, The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 357-406; b) G. M. Clore, A. M. Gronenborn, New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5891-5898; c) G. M. Clore, A. M. Gronenborn, Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol.* **1998**, *16*, 22-34.
5. W. Mertin, Electron crystallography - Now a handy method. *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 46- 47.
6. a) C. Sander, R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9*, 56-68; b) M. Sippl, H. Flöckner, Threading thrills and threats. *Structure* **1996**, *4*, 15-19; c) H. Flöckner, M. Braxenthaler, P. Lackner, M. Jaritz, M. Ortner, M. Sippl, Progress in fold recognition. *Proteins* **1995**, *23*, 376-386; d) A. Tramontano, Homology modeling with low sequence identity. *METHODS: A companion to Methods in Enzymology* **1998**, *14*, in press.
7. Protein data bank (PDB, <http://pdb.pdb.bnl.gov/>) as of 9/13/98.
8. a) E. Pebay-Peyroula, G. Rummel, J. P. Rosenbusch, E. M. Landau, X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipidic cubic phases. *Science (Washington D.C.)* **1997**, *277*, 1676-1681; b) cf. F. Hucho, X-ray structure of bacteriorhodopsin at 2.5 Å resolution. *Angew. Chem. Int. Edn. Engl.* **1998**, *37*, 1518-1519.

9. a) J. Deisenhofer, H. Michel, The photosynthetic reaction center from the purple bacterium *Rhodospseudomonas viridis*. *Angew. Chem. Int. Edn. Engl.* **1989**, *28*, 829-847; b) R. Huber, A structural basis of light energy and electron transfer in biology. *Angew. Chem. Int. Edn. Engl.* **1989**, *28*, 848-869; c) J. P. Allen, G. Feher, T. O. Yeates, H. Komiya, D. C. Rees, Structure of the reaction center from *Rhodobacter sphaeroides* R-26: The cofactors. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 5730-5734; H. L. Axelrod, G. Feher, J. P. Allen, A. J. Chirino, M. W. Day, B. T. Hsu, D. C. Rees, Crystallization and X-ray structure determination of cytochrome c2 from *Rhodobacter sphaeroides* in three crystal forms. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *D50*, 596-602; d) M. Schiffer, R. Norris, Structure and function of the photosynthetic reaction center of *Rhodobacter sphaeroides*. *Photosynth. React. Cent.* **1993**, *1*, 1-12; C.-H. Chang, O. El-Kabbani, D. Tiede, J. Norris, M. Schiffer, Structure of the membrane-bound protein photosynthetic reaction center from *Rhodobacter sphaeroides*. *Biochemistry* **1991**, *30*, 5352-5360; O. El-Kabbani, C. H. Chang, D. Tiede, J. Norris, M. Schiffer, Comparison of reaction centers from *Rhodobacter sphaeroides* and *Rhodospseudomonas viridis*: overall architecture and protein-pigment interactions. *Biochemistry* **1991**, *30*, 5361-5369; M. H. B. Stowell, T. M. McPhillips, D. C. Rees, S. M. Soltis, E. Abresch, G. Feher, Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron-proton transfer. *Science* **1997**, *276*, 812-816; cf. M. Huber, Light-induced structural changes in the primary processes of photosynthesis: Watching an enzyme in action. *Angew. Chem. Int. Edn. Engl.* **1998**, *37*, 1073-1075.
10. a) G. McDermott, M. Prince, A. A. Freer, A. M. Hawthornthwaite-Lawless, M. Z. Papiz, R. J. Cogdell, N. W. Isaac, Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature (London)* **1995**, *374*, 517-21; b) X. Hu, K. Schulten, J. Koepke, H. Michel, Structure and dynamics of light-harvesting complexes II of *rhodospirillum molischianum*. Book of Abstracts, 212th ACS National Meeting, Orlando, FL, August 25-29 **1996**, PHYS-037.
11. a) P. Fromme, H. T. Witt, W.-D. Schubert, O. Klukas, W. Saenger, N. Krauss, Structure of photosystem I at 4.5 Å resolution: a short review including evolutionary aspects. *Biochim. Biophys. Acta* **1996**, *1275*, 76-83; b) W.-D. Schubert, O. Klukas, N. Krauss, W. Saenger, P. Fromme, H. T. Witt, Photosystem I of *Synechococcus elongatus* at 4 Å resolution: comprehensive structure analysis. *J. Mol. Biol.* **1997**, *272*, 741-769.
12. a) M. S. Weiss, U. Abele, J. Weckesser, W. Welte, E. Schiltz, G. E. Schulz, Molecular architecture and electrostatic properties of a bacterial porin. *Science (Washington, D. C.)* **1991**, *254*, 1627-1630; M. S. Weiss, G. E. Schulz, Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* **1992**, *227*, 493-509; b) S. W. Cowan, T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R. A. Pauptit, J. N. Jansonius, J. P. Rosenbusch, Crystal structures explain functional properties of two *E. coli* porins. *Nature (London)* **1992**, *358*, 727-733; c) T. Schirmer, T. A. Keller, Y.-F. Wang, J. P. Rosenbusch, Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. *Science (Washington, D. C.)* **1995**, *267*, 512-514; d) B. Schmid, M. Kroemer, G. E. Schulz, Expression of porin from *Rhodospseudomonas blastica* in *Escherichia coli* inclusion bodies and folding into exact native structure. *FEBS Lett.* **1996**, *381*, 111-114; e) J. E. W. Meyer, M. Hofnung, G. E. Schulz, Structure of maltoporin from *Salmonella typhimurium* ligated with a nitrophenyl-maltotrioxide. *J. Mol. Biol.* **1997**, *266*, 761-775.
13. L. Song, M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, J. E. Gouaux, Structure of staphylococcal  $\alpha$ -hemolysin, a heptameric transmembrane pore. *Science (Washington, D. C.)* **1996**, *274*, 1859-1866.
14. a) D. Picot, P. J. Loll, R. M. Garavito, The x-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. *Nature (London)* **1994**, *367*, 243-249; b) R. M. Garavito, D. Picot, P. J. Loll, The x-ray structures of complexes of cyclooxygenase-1 and inhibitors. *Med. Chem. Res.* **1995**, *5*, 375-383.
15. R. G. Kurumbail, A. M. Stevens, J. K. Gierse, J. J. McDonald, R. A. Stegeman, J. Y. Pak, D. Gildehaus, J. M. Miyashiro, T. D. Penning, K. Seibert, P. C. Isakson, W. C. Stallings,

- Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature (London)* **1996**, *384*, 644-648; *ibid.* **1997**, *385*, 555.
16. a) T. Tomitake, H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, S. Yoshikawa, Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 Å. *Science (Washington, D. C.)* **1995**, *269*, 1069-1074; b) S. Iwata, C. Ostermeier, B. Ludwig, H. Michel, Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. *Nature (London)* **1995**, *376*, 660-9.
  17. a) D. Xia, C.-A. Yu, H. Kim, J.-Z. Xia, A. M. Kachurin, L. Zhang, Li; L. Yu, J. Deisenhofer, Crystal structure of the cytochrome bc<sub>1</sub> complex from bovine heart mitochondria. *Science (Washington, D. C.)* **1997**, *277*, 60-66; b) A. R. Crofts, E. A. Berry, Structure and function of the cytochrome bc<sub>1</sub> complex of mitochondria and photosynthetic bacteria. *Curr. Opin. Struct. Biol.* **1998**, *8*, 501-509; c) S. Iwata, J. W. Lee, K. Okada, J. K. Lee, M. Iwata, B. Rasmussen, T. A. Link, S. Ramaswamy, B. K. Jap, Complete structure of the 11-subunit bovine mitochondrial cytochrome bc<sub>1</sub> complex. *Science (Washington, D. C.)* **1998**, *281*, 64-71; d) <http://arc-gen1.life.uiuc.edu/bc-complex-site/>
  18. For a beautiful pictorial description of membrane-spanning proteins and a discussion of their function, see R. H. Garrett, C. M. Grisham, *Biochemistry: Molecular Aspects of Cell Biology* (supplement), Saunders College Publishing, Harcourt Brace College Pub., New York, pp 1130-1152 and pp 1191-1198, **1995**.
  19. a) For an endless collection of electron microscopy images, examine the links on the Microscopy Society of America site, <http://www.msa.microscopy.com/>; b) EM reference books, <http://www.med.yale.edu/celling/examine.html>; c) For some Emory University images: <http://euch3i.chem.emory.edu/~nmr/apk/>
  20. a) D. L. Dorset, *Structural Electron Crystallography*, Plenum, New York, **1995**; b) J. C. H. Spence, *Experimental High-Resolution Electron Microscopy*, 2nd ed.; Oxford Univ. Press, Oxford, **1988**; c) V. A. Drits, *Electron Diffraction and High-Resolution Electron Microscopy of Mineral Structures* (translator: Bella B. Smoliar), Springer, Berlin, **1987**.
  21. T. E. Weirich, R. Ramlau, A. Simon, S. Hovmöller, X. Zou, A crystal structure determined with 0.02 Å accuracy by electron microscopy. *Nature* **1996**, *382*, 144-146.
  22. a) R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, K. H. Downing, A model for the structure of bacteriorhodopsin based on high resolution electron cryomicroscopy. *J. Mol. Biol.* **1990**, *213*, 899-929. b) N. Grigorieff, T. A. Ceska, K. H. Downing, J. M. Baldwin, R. Henderson, Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **1996**, *259*, 393-421; c) Y. Kimura, D. G. Vassilyev, A. Miyazawa, A. Kidera, M. Matsushima, K. Mistuoka, K. Murata, T. Hirai, Y. Fujiyoshi, High resolution structure of bacteriorhodopsin determined by electron crystallography. *Photochem. Photobiol.* **1997**, *66*, 764-767.
  23. a) W. Kühlbrandt, D. N. Wang, Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. *Nature (London)* **1991**, *350*, 130-134; b) W. Kühlbrandt, D. N. Wang, Y. Fujiyoshi, Atomic model of plant light-harvesting complex by electron crystallography. *Nature (London)* **1994**, *367*, 614-621.
  24. a) E. Nogales, S. G. Wolf, K. H. Downing, Structure of the αβ tubulin dimer by electron crystallography. *Nature (London)* **1998**, *391*, 199-203; b) E. Nogales, S. G. Wolf, K. H. Downing, Visualizing the secondary structure of tubulin; Three-dimensional map at 4 Å. *J. Struct. Biol.* **1997**, *118*, 119-127.
  25. a) M. A. Jordon, L. Wilson, Microtubule polymerization dynamics, mitotic block and cell death by paclitaxel at low concentrations. In "Taxane Anticancer Agents," Eds. G. I. Georg, T. T. Chen, I. Ojima, D. M. Vyas, ACS Symposium Series 583, ACS, Washington, D.C., **1995**, 138-153; b) K. C. Nicolaou, K. C., W.-M. Dai, R. K. Guy, Chemistry and biology of taxol. *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 15-44.
  26. J. P. Snyder, E. Nogales, K. Downing, unpublished work.
  27. a) J. Dubois, D. Guénard, F. Guéritte-Voegelein, N. Guedira, P. Potier, B. Gillet, J.-C. Beloeil, Conformation of taxotere® and analogs determined by NMR spectroscopy and molecular modeling studies. *Tetrahedron* **1993**, *49*, 6533-6544; b) H. J. Williams, A. I. Scott, R. A.

- Dieden, C. S. Swindell, L. E. Chirlian, M. M. Francl, J. M. Heerding, N. E. Krauss, NMR and molecular modeling study of the conformations of taxol and of its side chain methylester in aqueous and non-aqueous solution. *Tetrahedron* **1993**, *49*, 6545-6560; c) D. G. Vander Velde, G. I. Georg, G. L. Grunewald, C. W. Gunn, L. A. Mitscher, "Hydrophobic collapse" of taxol and taxotere solution conformations. *J. Am. Chem. Soc.* **1993**, *113*, 11650-11651; d) L. G. Paloma, R. K. Guy, W. Wrasidlo, K. C. Nicolaou, Conformation of a water-soluble derivative of taxol in water by 2D-NMR spectroscopy. *Chemistry and Biology* **1994**, *1*, 107-112.
28. a) S. B. Shuker, P. J. Hajduk, R. P. Meadows, S. W. Fesik, Discovering high-affinity ligands for proteins: SAR by NMR. *Science (Washington D. C.)* **1996**, *274*, 1531-1534; b) P. J. Hajduk, R. P. Meadows, S. W. Fesik, Discovering high-affinity ligands for proteins. *Science (Washington D. C.)* **1997**, *278*, 497-499; c) P. J. Hajduk, G. Sheppard, D. G. Nettlesheim, E. T. Olejniczak, S. B. Shuker, R. P. Meadows, D. H. Steinman, G. M. Carrera, Jr., P. A. Marcotte, J. Severin, K. Walter, H. Smith, E. Gubbins, R. Simmer, T. F. Holzman, D. W. Morgan, S. K. Davidsen, J. B. Summers, S. W. Fesik, Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.* **1997**, *119*, 5818-5827.
  29. a) A. Caflisch, M. Karplus, Computational combinatorial chemistry for de novo ligand design: Review and assessment. *Persp. Drug Disc. Des. (PD3)*, K. Müller, ed, **1995**, *3*, 51-84; b) A. Caflisch, Computational combinatorial ligand design: Application to human  $\alpha$ -thrombin. *J. Comp.-Aided Molec. Des.* **1996**, *10*, 372-396.
  30. H.-J. Böhm, Site-directed structure generation by fragment-joining. *Persp. Drug Disc. Des. (PD3)*, K. Müller, ed, **1995**, *3*, 21-33.
  31. P. W. Rose, B. A. Luty, T. J. Marrone, Virtual "SAR by NMR", Abstracts, 12<sup>th</sup> European Symposium on QSAR Relationships: Molecular Modelling and Prediction of Bioactivity, O.23, Copenhagen, Denmark, August 23-28, **1998**.
  32. a) A. Brisson, A. Olofsson, P. Ringler, M. Schmutz, S. Stoylova, Two-dimensional crystallization of proteins on planar lipid films and structure determination by electron crystallography. *Biol. Cell* **1994**, *80*, 221-228; b) C. Ostermeier, H. Michel, Crystallization of membrane proteins. *Curr. Opin. Struct. Biol.* **1997**, *7*, 697-701.
  33. a) M. D. Houslay, G-protein linked receptors: a family probed by molecular cloning and mutagenesis procedures. *Clin. Endocrinol.* **1992**, *36*, 525-534; b) C. D. Strader, T. M. Fong, M. R. Tota, D. Underwood, Structure and function of G protein-coupled receptors. *Annu. Rev. Biochem.* **1994**, *63*, 101-132; c) M. R. Tota, T. M. Fong, C. D. Strader, The use of fluorescent ligands to explore the ligand binding site of G protein coupled receptors. *Alfred Benzon Symp.* **1996**, *39*, 162-170; d) T. W. Schwartz, Surprisingly elusive binding sites for non-peptide ligands in 7TM receptors. Abstract L.8, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, **1998**.
  34. a) J. Findlay, E. Eliopoulos, Three-dimensional modeling of G-protein-linked receptors. *Trends Pharmacol. Sci.* **1990**, *11*, 492-499; b) P. A. Hargrave, J. H. McDowell, Rhodopsin and phototransduction: a model system for G protein-linked receptors. *FASEB J.* **1992**, *6*, 2323-2331; c) L. Pardo, J. A. Ballesteros, R. Osman, H. Weinstein. On the use of the transmembrane domain of bacteriorhodopsin as a template for modeling the three-dimensional structure of guanine nucleotide-binding regulatory protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 4009-4012.
  35. a) G. Nordvall, U. Hacksell, Binding-site modeling of the muscarinic m1 receptor: A combination of homology-based and indirect approaches. *J. Med. Chem.* **1993**, *36*, 967-976; b) Å. Malmberg, G. Nordvall, A. M. Johansson, N. Mohell, U. Hacksell. Molecular basis for the binding of 2-aminotetralins to human dopamine D<sub>2a</sub> and D<sub>3</sub> receptors. *Molec. Pharmacol.* **1994**, *46*, 299-312; c) C. J. Grol, J. M. Jansen. The high affinity melatonin binding site probed with conformationally restricted ligands. II. Homology modeling of the receptor. *Bioorg. Med. Chem.* **1996**, *4*, 1333-1339; d) J. M. Jansen, K. F. Koehler, M. H.



- Hedberg, A. M. Johansson, U. Hacksell, G. Nordvall, J. P. Snyder. Molecular design using the minireceptor concept. *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 812-816.
36. E. J. Homan, H. V. Wikström and C. J. Grol.
  37. E. J. Homan, The medicinal chemistry of 2-aminotetralin-derived benzamides. A novel class of potential atypical antipsychotic agents. Ph.D. Thesis, University of Groningen, **1998**.
  38. MacroModel: <http://www.columbia.edu/cu/chemistry/mmod/mmod.html>
  39. a) J. P. Snyder, S.N. Rao, K.F. Koehler, A. Vedani and R. Pellicciari, APOLLO pharmacophores and the pseudoreceptor concept, In "Trends in QSAR and Molecular Modeling 92," Ed. C.G. Wermuth, ESCOM, Leiden, pp 44-51, **1993**; b) J. P. Snyder, S.N. Rao, K.F. Koehler and A. Vedani, Minireceptors and Pseudoreceptors. In "3D QSAR in Drug Design, Theory, Methods and Applications," Ed. H. Kubinyi, ESCOM, Leiden . pp 336-354, **1993**.
  40. Sybyl 6.4/Unity 3.0: <http://www.tripos.com/>
  41. a) E. J. Homan, S. Copping, L. Elfström, T. Van Der Veen, J.-P. Hallema, N. Mohell, L. Unelius, Johansson, H. V. Wikström, C. J. Grol, 2-Aminotetralin-derived substituted benzamides with mixed dopamine D2, D3 and serotonin 5-HT1A receptor binding properties. A novel class of potential atypical antipsychotic agents. *Bioorg. Med. Chem.* **1998**, in press; b) E. J. Homan, S. Copping, L. Unelius, D. M. Jackson, H. V. Wikström, C. J. Grol. Synthesis and pharmacology of the enantiomers of the potential atypical antipsychotic agents 5-Ome-BPAT and 5-Ome-(2,6-di-Ome)-BPAT. *Bioorg. Med. Chem.*, submitted.
  42. P. Herzyk, R. E. Hubbard. Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys. J.* **1995**, *69*, 2419-2442.
  43. a) A. Vedani, P. Zbinden, J. P. Snyder, P. Greenidge, Pseudoreceptor modeling: The construction of three-dimensional receptor surrogates. *J. Am. Chem. Soc.* **1995**, *117*, 4987-4994; b) P. Zbinden, M. Dobler, G. Folkers, A. Vedani, PrGen: Pseudoreceptor modeling using receptor-mediated ligand alignment and pharmacophore equilibration. *Quant. Struct.-Act. Relat.* **1998**, *17*, 122-130.
  44. X. Xia, M. Wang, J. P. Snyder, in preparation.
  45. H. Venkatesan, M. Davis, M. Wang, D. Liotta, J. P. Snyder, D. Eaton, N. Albaldawi (Emory University), unpublished.
  46. J. H. Nettles, J. P. Snyder, unpublished.
  47. a) A. W. Czarnik, J. A. Ellman, eds, Special Issue on Combinatorial Chemistry. *Acc. Chem. Res.* **1996**, *29*, 112-170; b) "Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery." E. W. Chaiken, K. D. Janda, eds., American Chemical Society, Washington, DC, **1996**; c) J. Szostak, ed, Combinatorial Chemistry. *Chem Rev.* **1997**, *97*, 347-510.
  48. a) M. Wataru, Chemical, biochemical and biotechnological importance of marine natural products. *Bio. Ind.* **1998**, *15*, 7-12; b) D. J. Faulkner. Marine natural products. *Nat. Prod. Rep.* **1998**, *15*, 113-158; c) R. Riguera, Isolating bioactive compounds from marine organisms. *J. Mar. Biotechnol.* **1997**, *5*, 187-193; d) G. R. Dietzman, The marine environment as a discovery resource. *High Throughput Screening* **1997**, 99-144; e) V. S. Bernan, M. Greenstein, W. M. Maiese, Marine microorganisms as a source of new natural products. *Adv. Appl. Microbiol.* **1997**, *43*, 57-90; f) A. M. Rouhi, Seeking drugs in natural products. *Chem. Eng. News* **1997**, April 7, 14-29; g) P. R. Jensen, W. Fenical, Marine bacterial diversity as a resource for novel microbial products. *J. Ind. Microbiol. Biotechnol.* **1996**, *17*, 346-351; h) P. J. Scheuer, Marine natural products. Diversity in molecular structure and bioactivity. *Adv. Exp. Med. Biol.* **1996**, *391*(Natural Toxins 2), 1-8; i) A. M. Rouhi, Supply issues complicate trek of chemicals from sea to market. *Chem. Eng. News* **1995**, April 7, 42-44.
  49. F. M. Belpaire, M. G. Bogaert, The fate of xenobiotics in living organisms. In "The Practice of Medicinal Chemistry," Ed. C. G. Wermuth, Academic Press, New York, pp 593-614, **1996**.
  50. 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.* **1997**, *23*, 3-25.

51. a) A. Leo, C. Hansch, D. Elkins, Partition coefficients and their uses. *Chem. Rev.* **1971**, *71*, 525-616; A. Leo, Calculating log P oct from structures. *Chem. Rev.* **1993**, *93*, 1281-1306; b) A. K. Ghose, G. M. Crippen, Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565-77; c) N. Bodor, Z. Gabanyi, C. K. Wong, A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783-3786; N. Bodor, P. Buchwald, Molecular size based approach to estimate partition properties for organic solutes. *J. Phys. Chem. B* **1997**, *101*, 3404-3412; d) G. E. Kellogg, G. S. Joshi, D. J. Abraham, New tools for modeling and understanding hydrophobicity and hydrophobic interactions. *Med. Chem. Res.* **1991**, *1*, 444-53; D. J. Abraham, G. E. Kellogg, The effect of physical organic properties on hydrophobic fields (HINT). *J. Comput.-Aided Mol. Des.* **1994**, *8*, 41-49; e) G. Klopman, S. Wang, A Computer Automated Structure Evaluation (CASE) approach to calculation of partition coefficients. *J. Comput. Chem.* **1991**, *12*, 1025-1032; G. Klopman, J.-Y. Li, S. Wang, M. Dimayuga, Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752-781; f) I. Moriguchi, S. Hirono, Q. Liu, Y. Nakagome, Y. Matsushita, Simple method of calculating octanol/water partition coefficients. *Chem. Pharm. Bull.* **1992**, *40*, 127-130; g) H. Van De Waterbeemd, R. Mannhold, Lipophilicity descriptors for structure-property correlation studies: overview of experimental and theoretical methods and a benchmark of log P calculations. *Methods Princ. Med. Chem.* **1996**, (Lipophilicity in Drug Action and Toxicology), 401-418; H. van de Waterbeemd, R. Mannhold, Programs and methods for calculation of log P values. *Quant. Struct.-Act. Relat.* **1996**, *15*, 410-412; h) R. Mannhold, R. F. Rekker, C. Sonntag, A. M. Ter Laak, K. Dross, E. E. Polymeropoulos, Comparative evaluation of the predictive power of calculation procedures for molecular lipophilicity. *J. Pharm. Sci.* **1995**, *84*, 1410-1419; i) W. M. Meylan, P. H. Howard, Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83-92; LogKow, log P estimation program: <http://esc.syrres.com/~esc1/estsoft.htm>; j) K. Takacs-Novak, Computerized log P prediction using fragment methods. *Acta Pharm. Hung.* **1998**, *68*, 39-48.
52. a) cf. Ref. 50 and citations therein for a review of methods up through 1995; b) WSKow, water solubility estimation program: <http://esc.syrres.com/~esc1/estsoft.htm>
53. O. A. Raevsky, K. J. Schaper, H. van de Waterbeemd, J. W. McFarland, Hydrogen bond contributions to properties of chemicals and drugs. Abstract O.22, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
54. M. Kansy, K. Kratzat, I. Parrilla, F. Senner, B. Wagner, Physicochemical high throughput screening (PC-HTS): In the determination of membrane permeability, partitioning and solubility. Abstract O.25, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
55. O. A. Raevsky, E. P. Trepalina, S. V. Trepalin. Slipper – A new program for water solubility, lipophilicity and permeability prediction. Abstract P.151, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
56. A. Karlén, S. Winiwarter, N. Bonham, H. Lennernäs, A. Hallberg, Correlation of intestinal drug permeability in humans (*in vivo*) with experimentally and theoretically derived parameters. Abstract P.152, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
57. P. C. Jurs, M. D. Wessel, Prediction of human intestinal absorption of drug compounds from molecular structure. Abstract O.27, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
58. L. H. Krarup, A. Berglund, M. Sandberg, I. T. Christensen, L. Hovgaard, S. Frøkjær. Predicting peptide absorption. Abstract O.24, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.

59. W. Guba, G. Cruciani, The use of molecular field-derived descriptors for the multivariate modeling of pharmacokinetic data. Abstract O.6, 12<sup>th</sup> European Symposium on QSAR, Molecular Modeling and Prediction of Bioactivity, Copenhagen, Denmark, August 23-28, 1998.
60. a) C. Hansch, D. Hoekman, A. Leo, L. Zhang, P. Li, The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicol. Lett.* **1995**, *79*, 45-53; b) N. Bodor, P. Buchwald, M. -J. Huang, Computer-assisted design of new drugs based on retrometabolic concepts. *SAR QSAR Environ. Res.* **1998**, *8*, 41-92; c) C. A. Marchant, R. D. Combes, Artificial intelligence: the use of computer methods in the prediction of metabolism and toxicity. *Bioact. Compd. Des.* **1996**, 153-162.
61. J. P. Snyder, Computer-assisted drug design. Part I. Conditions in the 1980s. *Med. Res. Rev.* **1991**, *11*, 641-662.
62. D. B. Boyd, Progress in rational design of therapeutically interesting compounds. In "Rational Molecular Design in Drug Research," T. Liljefors, F. S. Jorgensen, P. Krosgaard-Larsen, eds, Munksgaard, Copenhagen, pp 15-23, **1998**.
63. a) E. J. Martin, D. C. Spellmeyer, R. E. Critchlow, Jr., J. M. Blaney, Does combinatorial chemistry obviate computer-aided drug design? *Rev. Comp. Chem.*, K. B. Lipkowitz, D. B. Boyd, eds, V. 10, pp 75-100, **1997**; b) J. H. Van Drie, M. S. Lajiness, Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274-283; c) W. P. Walters, M. T. Stahl, M. A. Murcko, Virtual screening - an overview. *Drug Discovery Today* **1998**, *3*, 160-178.
64. L. W. Fine, "Chemistry Decoded," Oxford University Press, p xvi, **1976**.

**Section II**  
**New Developments and**  
**Applications of**  
**Multivariate QSAR**

# MULTIVARIATE DESIGN AND MODELLING IN QSAR, COMBINATORIAL CHEMISTRY, AND BIOINFORMATICS

Svante Wold,<sup>a</sup> Michael Sjöström,<sup>a</sup> Per M. Andersson,<sup>a</sup> Anna Linusson,<sup>a</sup> Maria Edman,<sup>a</sup> Torbjörn Lundstedt,<sup>b</sup> Bo Nordén,<sup>c</sup> Maria Sandberg,<sup>a</sup> and Lise-Lott Uppgård<sup>a</sup>

<sup>a</sup>Research Group for Chemometrics, Department of Organic Chemistry, Institute of Chemistry, Umeå University, SE-904 87 Umeå, Sweden, [www.chem.umu.se/dep/ok/research/chemometrics](http://www.chem.umu.se/dep/ok/research/chemometrics)

<sup>b</sup>Structure Property Optimization Center (SPOC), Pharmacia & Upjohn AB, SE-751 82 Uppsala, Sweden

<sup>c</sup>Medicinal Chemistry, Astra Hässle AB, SE-431 83 Mölndal, Sweden

## Abstract

The last decade has witnessed much progress in how to characterize and describe chemical structure, how to synthesize large sets of compounds, how to make simple and fast *in-vitro* assays, and how to determine the structure (sequence) of our genetic material. The possible consequences of this progress for drug design are great and exciting, but also bewilderingly complicated.

Fortunately, the last decade has also seen progress in how to investigate and model complicated systems, of which relationships between chemical structure and biological activity provide typical examples. These relationships are central in drug design and some related areas, notably combinatorial chemistry and bioinformatics.

The essential steps in the investigation of complicated systems include the following:

1. The appropriate quantitative parameterization of its parts (here the varying parts of the chemical structures / biopolymer sequences).
2. The appropriate measurements of the interesting properties of the system (here the "biological effects").
3. Selecting a representative set of molecules (or other systems) to investigate and make the following measurements.
4. The analysis of the resulting data.
5. The interpretation of the results.

The use of multivariate characterization, design, and modelling in these steps will be discussed in relation to drug design, combinatorial chemistry (which compounds to make and test, and how to deal with the biological test results), and bioinformatics (how to parameterize and analyze biopolymer sequences).

## 1. Introduction

Much of chemistry, molecular biology, and drug design, are centered around the relationships between chemical structure and measured properties of compounds and polymers, such as viscosity, acidity, solubility, toxicity, enzyme binding, and membrane penetration. For any set of compounds, these relationships are by necessity complicated, particularly when the properties are of biological nature. To investigate and utilize such complicated relationships, henceforth abbreviated SAR for structure-activity relationships, and QSAR for *quantitative* SAR, we need a description of the variation in chemical structure of relevant compounds and biological targets, good measures of the biological properties, and, of course, an ability to synthesize compounds of interest. In addition, we need reasonable ways to construct and express the relationships, *i.e.*, mathematical or other models, as well as ways to select the compounds to be investigated so that the resulting QSAR indeed is informative and useful for the stated purposes. In the present context, these purposes typically are the conceptual understanding of the SAR, and the ability to propose new compounds with improved property profiles.

Here we discuss the two latter parts of the SAR/QSAR problem, *i.e.*, reasonable ways to model the relationships, and how to select compounds to make the models as "good" as possible. The second is often called the problem of statistical experimental design, which in the present context we call statistical molecular design, SMD.

### 1.1 Recent Progress in Relevant Areas

In the last decades, we have made great progress in several areas of relevance for the SAR problem. The advances include improvements in our ability to determine the structures of substrates and receptors in any reaction occurring in living systems, as well as the quantitative description, parameterization, of these structures. Also the actual synthesis of interesting molecules has been simplified and partly automated, leading to the creation of large ensembles of compounds, libraries, being routinely synthesized in so-called combinatorial chemistry. Finally, a field of great interest in the present context is the determination of the structure (sequence) of the genetic material of both humans and various other organisms of interest, *e.g.*, viruses, bacteria, and parasites. Also here the last few years have seen an enormous acceleration of technology and ensuing results, and today many millions of sequence elements (amino acids or base pairs) are determined per day in laboratories all over the world.

### 1.2 Some Nagging Difficulties

These advances undoubtedly are ground for a great enthusiasm and optimism. But, interestingly, these advances are also causing great difficulties due to the huge amounts of resulting quantitative data, the "data explosion". These difficulties are similar to those in other fields of science and technology, exemplified by process engineering (multitudes of process variables measured at ever increasing frequencies), geography (satellite images), and astronomy (several types of spectra of huge numbers of stars and galaxies). For science, these vast amounts of data present great problems since all theory and most tools for analyzing data were developed for a situation when the data were few and arrived at a comfortable pace of, say, less than one number an hour. Consequently we continue to think of one molecule or process sensor or galaxy at a time, and pretend that our deep understanding in some miraculous way will be able to cope with the large numbers of events and items that we have not considered.

### 1.3 A Possible Approach

Besides organizing data in data bases, we need proper tools to get some kind of "control" of these data masses and utilize their potential information. The only tools of any generality that substantially can contribute to this objective are those of (computer based) modelling and data analysis, coupled with the proper selection of items (here molecules) to constitute the basis for the analysis. The latter selection problem is called sampling if the items already exist, and experimental design if the "items" do not (yet) exist.

If an appropriate selection of items is made and a proper model is developed, this model may cover a large chunk of the data mass. Hence, with a few well selected loosely coupled models, the whole data mass may be brought under "control".

We shall below discuss this approach and its consequences in the areas of QSAR, combinatorial chemistry, and bioinformatics.

## 2. Investigation of Complicated Systems (Modelling)

The more complicated the studied system is, the more approximate are, by necessity, the models used in the study. This because we are unable to construct "exact" models for any system more complicated than that of three particles, exemplified by  $\text{He}^+$  and  $\text{H}_2^+$ . Hence, for any molecular system of interest in the present context, with over a thousand electrons and atomic nuclei, models are highly approximate. This is so regardless if the models are derived from quantum or molecular mechanics, or if they are "empirical" linear models based on measured data. Consequently, there are deviations between the model and the observed values and the models need to have an element of statistics.

Another interesting property of complicated systems is their multivariate nature. Consider a typical organic compound with 20 to 50 atoms of type C, H, N, O, S, and P. This may also be a short peptide or a short DNA or RNA sequence. As chemists we like to think of compounds in terms of "atom groups", such as rings, chains, functional groups, "substituents", amino acids, and nucleic bases. Each such group is characterized by at least 5 properties; lipophilicity, polarity, polarizability, hydrogen bonding, and size. The latter may need sub-properties such as width and depth to be adequately described. Consequently, the investigation of a structural "family" by means of varying the structure of this "mother compound" corresponds to the variation of up to 50 -70 "factors". The modelling of resulting measurements made on this structural family must therefore also cope with a multitude of possible "factors"; the modelling must be multivariate.

### 2.1 Parameterization

One of the first problems to solve in the present context is the parameterization of the items investigated, here molecules and polymers. This parameterization must of course be consistent with chemical and biological theory. However, since this theory is highly incomplete with respect to SAR/QSAR, we must take recourse also to measured data as the basis for parameterization. Traditionally, the QSAR field has used single parameters derived from measurements on model systems, for instance  $\sigma$ ,  $\pi$ , MR, and  $E_S$  [1]. For more complicated "atomic groups", it is very difficult to find measurement systems that result in "clean" parameters, and instead some kind of multivariate parameterization is easier. Thus, multiple measurements and calculations are made on compounds of interest, and then "compressed" by means of principal component analysis (PCA) or a similar multivariate analysis to give some kind of descriptor "scales". Examples of this approach are the amino acid "principal properties" of Hellberg *et. al.* [2-5]. Fauchère *et. al.* have published a similar approach [6]. Carlson, Lundstedt, *et. al.* [7-11], and Eriksson *et. al.* [12-15] have

published numerous examples of this approach with application specific "scales" for, *e.g.*, amines, ketones, and halogenated aliphatic hydrocarbons. Martin, Blaney, *et. al.* [16] have applied this approach in the combinatorial chemistry of peptoids.

Other approaches to structure parameterization include the use of molecular modelling (CoMFA, GRID, *etc.*), "topological" indices, fragment descriptors, simulated spectra, and more. We do not here have time or space to discuss the merits of various kinds of parameterization, but just point out that there is no general agreement of how to adequately describe the structural variation in SAR/QSAR problems.

However when the parameterization is done, the result is an array of numbers, "structure descriptors", for each compound included in the investigation. We denote the array of the *i*:th compound by  $x_i$ . In CoMFA [17] and GRID [18-20], these arrays may have more than a hundred thousand elements, while in a simple Hansch model they may have two or three elements.

## 2.2 Specification and Measurement of the Biological "Activity"

Any model needs a "compass" to indicate which events or items that are "better" and which are "worse" with respect to the stated objectives of the investigation. Here, this compass is constituted by the values of the biological properties of the investigated compounds, the so called responses,  $Y$ . These responses have to be *relevant*, *i.e.*, indeed give information about the stated objective, for instance anti-inflammatory activity or calcium channel inhibition. The responses should also be fairly precise so one can recognize the effect of a change of structure as clearly as possible.

The importance of a relevant and fairly precise  $Y$  matrix is so evident that we often do not even think about this point. However, in combinatorial chemistry, somewhat discussed below, the immense possible size of the data set with hundreds of thousands of compounds, prohibits the measurement of a relevant  $Y$ -matrix, and instead fast and crude so called HTS measurements are made (HTS = high throughput screening) [21]. The resulting low information content of the response matrix,  $Y$ , makes the success of this approach highly uncertain. Only the selection of a much smaller subset of compounds makes it possible to measure a "good"  $Y$ . This will be further discussed below.

## 2.3 Compound Selection (Sampling or Statistical Experimental Design)

The second necessary step in any modelling is the selection of the set of items, molecules, on which the model is to be "calibrated". This set is usually called the "training set". In SAR/QSAR this is a neglected issue, with resulting melancholically poor models and serious difficulties for the interpretation and use of the resulting models. This will be discussed in more detail below, illustrated by some examples.

## 2.4 The Mathematical Form of the Model

The purpose of SAR/QSAR modelling is to find the relationship between chemical structure and biological activity. We can hypothesize that there is a fundamental "truth" which relates the "real structure" expressed as a  $N \times K$  matrix  $Z$  to the  $N \times M$  biological activity matrix,  $Y$ , for the  $N$  compounds under investigation. This "truth" is expressed as:

$$Y = F(Z) + \epsilon$$

Here the residuals,  $\epsilon$ , express the error of measurement in  $Y$ .



However, we have little knowledge about the real form of the function  $F$ , and hence instead use a serial expansion of it, usually a polynomial, here denoted by 'Polyn'. Also, we do not know exactly how to express the structure as  $Z$ . We therefore use a simplified version,  $X$ , which reflects our present "belief" about  $Z$ . Usually we do not know the relative importance of the different "factors" in  $X$ . Hence we also introduce a parameter vector,  $\beta$ , the values of which can be changed to make the model "fit" the data. The use of a serial expansion instead of  $F$ , and of  $X$  instead of  $Z$  introduces further "errors",  $\delta$ , giving our model:

$$Y = \text{Polyn}(X, \beta) + \delta + \varepsilon$$

## 2.5 Estimating the Model From Data, and Interpreting the Results

In a given investigation we have now decided (a) which biological responses to measure, (b) which class of compounds to investigate, (c) how to express the structural variation, and (d) the general form of their relationship. We then select the compounds to synthesize (or get our hands on them in some other way) and then subject the compounds to the biological testing. After this is done, we have data constituting an  $N \times K$  "structure" matrix,  $X$ , plus an  $N \times M$  "activity" matrix,  $Y$ . Then a phase of data analysis follows, where the model is "fitted" to the data by finding optimal values of the parameters in the vector  $\beta$ . However, this phase involves much more than that, including the appropriate transformation of the data to make them suitable for the analysis, the search for outliers and other heterogeneities in the data that would make the resulting model misleading, the investigation of the "noise" which is a combination of  $\delta$  and  $\varepsilon$  (see above), the estimation of the uncertainties of the parameters, and often, the prediction of  $Y$  for new hypothetical compounds with the structure descriptors  $X_{\text{pred}}$ .

Provided that the data set has been well selected and measured, and that the modelling and estimation have been done properly, the resulting model can finally be interpreted, *i.e.*, related to our theory of chemistry and biology. This is perhaps the most important part of the modelling, but will not be much discussed here, where we are mainly concerned with the prerequisites for a good and useful model, *i.e.*, relevant data.

## 3. Some Examples

Below we show a few examples chosen to illustrate some aspects of modelling, notably the selection of a relevant set of compounds, statistical molecular design, SMD, and multivariate analysis.

### 3.1 A "QSAR"

In any issue of medicinal chemistry, molecular biology, or bio-organic chemistry journals, or in almost any book in one of these subjects, one finds data sets similar to the one shown in Table 1 below. The present example was published some time ago, but the reference is not given to avoid possible embarrassment. The objective was to develop an anti-inflammatory compound with the general structure Z-Phen1-D-Phen2. Here D symbolizes a constant connecting chain, and Z is a constant pharmacophore. A number of different compounds ( $N=12$ ) were made with different substituents in the two phenyl rings (see Table 1).

An *in vivo* test of the decrease of the volume of an animal joint for a given dose was measured as "activity". High values correspond to "good" activity. Quantum chemical

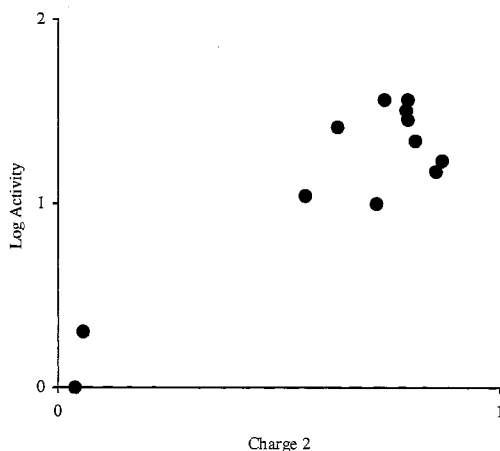
calculations were used to estimate the charge excess in the two phenyl rings, and the conclusion was that the charge on ring 2 (column 4 in Table 1) was a good predictor of the (logarithmic) activity.

Inspection of Table 1 shows a typical "L-design" where first the substituents on ring 1 are changed, then the ones on ring 2 are changed, and finally a few compounds are made where some changes are made in both rings. "L-design" stands for the resulting configuration in an abstract space in the shape of an "L". This is also often called a "COST" design for Changing One Site at a Time.

**Table 1.** Substituents on phenyl rings 1 and 2, calculated charge on phenyl ring 2, and logarithmic activity of N=12 compounds Z-Phen1-D-Phen2.

No	Phen1	Phen2	Charge 2	Log Activity
1	H	H	0.635	1.415
2	4-Me	H	0.040	0.000
3	5-Me	H	0.559	1.041
4	6-Me	H	0.056	0.301
5	H	2-Cl	0.809	1.342
6	H	3-Cl	0.856	1.176
7	H	4-Cl	0.792	1.462
8	H	2,4-Cl <sub>2</sub>	0.740	1.568
9	H	3,4-Cl <sub>2</sub>	0.723	1.000
10	H	4-Me	0.870	1.230
11	5-F	4-Cl	0.791	1.568
12	5-Me	4-Cl	0.790	1.505

Plotting the "model" of log activity vs charge 2 gives Figure 1. Although the model has an apparently "significant"  $R^2$  of 0.84 and a Y-residual SD of 0.22, the plot shows that there are actually only two clusters, only two degrees of freedom. With the typical error of measurement of  $\pm 0.3$  log units, there are actually only two points in this plot.



**Figure 1.** Y = log activity (vertical) plotted against charge in ring 2 (horizontal axis).

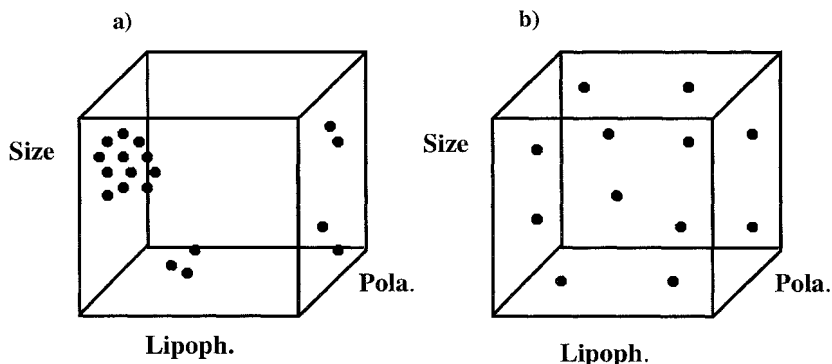
Hence, this data set gave little information about the posed question. The reason is the uninformative selection of compounds according to the "COSTly L-design". Due to the small resulting degrees of freedom, the conclusions are at best doubtful.

#### 4. Statistical Molecular Design - SMD

The selection of a set of compounds corresponds to the selection of a set of points in a multidimensional space where the number of axes equals the number of factors varied in the investigation. In example 1 above there are three substituent sites on each ring (no. 4,5,6 and 2,3,4 respectively) that are to be varied. In each we can put a large or small substituent, which is lipophilic or not, *etc.* Restricting ourselves to five factors per site – size, lipophilicity, polarity, polarizability, and hydrogen bonding -- we can see the selection of compounds for a linear model to be equivalent to the variation of 30 factors (3 + 3 sites times 5 factors). Each of these factors has a smallest and largest possible value, and hence we can see this problem as one of putting points in a rectangular 30-dimensional box.

In the initial phase of an investigation, linear models and corresponding linear designs are normally used since this allows the screening of many positions and factors. Once the dominating positions and factors are identified, one may use more detailed models where interactions (synergisms / antagonisms) between positions, curvature (quadratic terms), *etc.*, may be of interest and therefore a corresponding quadratic design is then needed.

Without a formal design protocol, one usually ends up with a selection similar to that shown in Figure 2a. This was the case in the first example where clustering is seen in the XY plot, Figure 1. Instead one should use an objective selection tool. These selections efficiently cover the structural space, and hence provide the maximal degrees of freedom for the data analysis and interpretation.



**Figure 2.** a) and b) shows the distribution of compounds resulting from a lack of SMD (left) and from the use of SMD (right).

This results in selections shown in Figure 2b. Although the boxes in Figure 2 have only three axes, one can mathematically construct and work with higher dimensional boxes. With 30 factors, one would need at least 35 compounds to get information about the 5 factors in the 6 substituent sites. If we have prior knowledge about the problem, we may be able to reduce the number of factors, stating, for instance, that only lipophilicity is important in all 6 positions, size in positions 4 and 6 on ring 1, polarity only in positions

2,3, and 4 on ring 2, *etc.* If this reduces the number of factors from 30 to 15, the number of compounds needed in an initial design is reduced to 20.

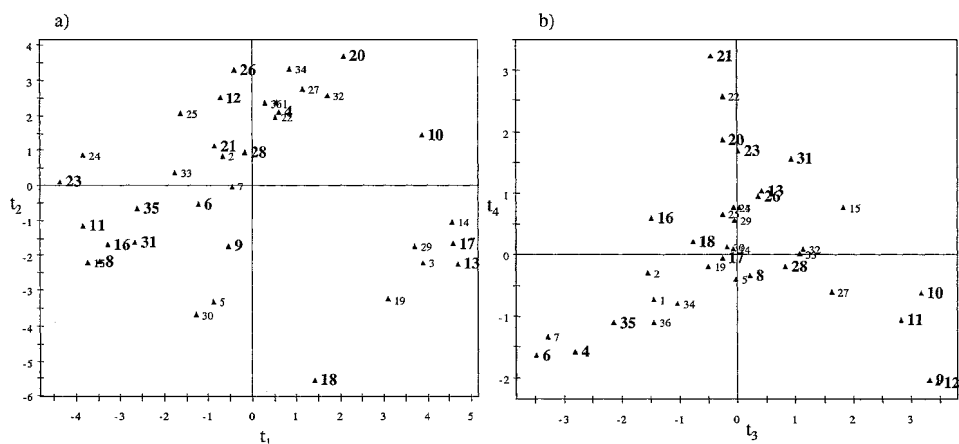
A difficulty with design of compounds is that the things that are changed – structural features – are not the same as the factors in the design and the model. Rather, the change of a substituent at a given site corresponds to the change of possibly five to seven factors. Hence, the design is first constructed in terms of these structural factors, and thereafter one identifies substituents or fragments with the correct profile of the factors. With the use of D-optimal design, this is accomplished by having a list of available substituents at each varied position together with their values of the pertinent "factors" (size, lipophilicity, *etc.*). The D-optimal selection procedure then searches for a combination of substituents at the different sites that gives the best coverage of the multidimensional factor space.

This use of statistical experimental design for the selection of informative set of compounds, we call statistical molecular design, SMD. Typical design types used in SMD include D-optimal [22] designs with center points and space-filling designs [23].

Statistical design goes back to Hansch and Craig [24] who showed how to select one substituent to investigate both lipophilicity and polarity ("pi-sigma plots"), and Hansch and Unger [25] who looked for clusters in the structure descriptor space and then selected one compound from each cluster. This was followed by Austel who introduced formal design in the QSAR area [26], and Hellberg *et. al.*, who developed multivariate design based on a combination of PCA and design [2,3]. The latter will be used in example 2 below.

#### 4.1 A Better "QSAR"

In the second example we show the use of SMD in the investigation of the toxicity of non-ionic technical surfactants recently published by Lindgren *et. al.* [27, 28]. Here N=36 surfactants were characterized by K=19 descriptors, *e.g.*, logP, MW, the "Griffin" and "Davis" hydro-lipophilicity balances, and the length of the alcohol part. These 19 descriptors are correlated and cannot be independently manipulated. Therefore, a PCA (see below) was made of the 36 x 19 X-matrix to find the underlying "latent factors". This PCA gave A=4 component model, *i.e.*, indicating 4 "latent factors". These are shown in Figure 3 a and b.



**Figure 3.** The first four PC scores ( $t_n$ ) of the N=36 surfactants times 19 descriptors X-matrix. X was mean centered and column-wise scaled to unit variance before the PCA. Bold-faced numbers indicate training set members selected by the D-optimal design for testing and Quantitative Structure-Property Relationship (QSPR) PLS model development. Left a):  $t_1$  vs  $t_2$ . Right b):  $t_3$  vs  $t_4$ .

### 4.1.1. Toxicity of the Surfactants

The aquatic toxicity of the selected  $N=18$  surfactants was measured towards two freshwater animal species, the fairy shrimp, *Thamnocephalus platyurus* and the rotifer *Brachionus calyciflorus*. The activities are defined as the logarithm (base ten) of the  $LC_{50}$  values, *i.e.* the lethal concentration at 50 % mortality after 24 hours. A large  $\log LC_{50}$  value, close to 2.0, corresponds to low toxicity.

### 4.1.2. Selection of a Representative Training Set of Surfactants

The scores of PCA of a matrix  $\mathbf{X}$  provide an optimal summary of all the variables (columns) in  $\mathbf{X}$ . Hence, these scores ( $t_a$ ) can be used as design variables for the selection of "spanning rows" of  $\mathbf{X}$ , *i.e.*, for the selection of a set of compounds that well represents the structural variation expressed by  $\mathbf{X}$ .

To allow a model whose results are (almost) interpretable in terms of the original 19 descriptors, it was decided to select  $N=18$  compounds for the training set. A D-optimal design in the four components scores (Figure 3 a and b) give the selected  $n_{\text{train}} = 18$  compounds.

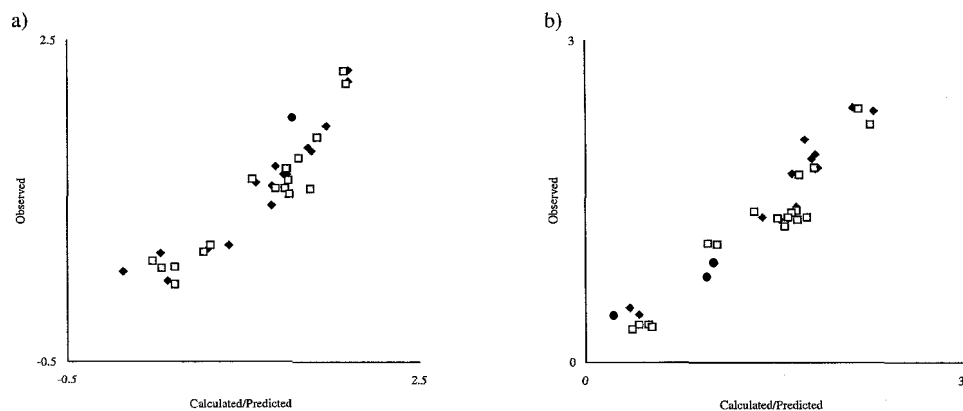
### 4.1.3. The Analysis of the Data

A PLS model (see below) was developed for the  $N=18$  observations, comprising  $K=19$  descriptor variables ( $\mathbf{X}$ ) and two activity values (toxicity),  $\mathbf{Y}$ . The model has  $A=2$  significant components according to cross-validation (CV). It explained  $R^2 = 89.3\%$  of the  $\mathbf{Y}$ -variation, and can predict  $Q^2 = 80.3\%$  of this variation according to the CV.

The important structure descriptor variables in this model are the hydrophobicity ( $\log P$ ), the number of atoms in the hydrophobic part ( $C$ ), the hydrophilic-lipophilic balance according to Davis, and the critical micelle concentration (CMC).

### 4.1.4. Prediction of the Remaining Compounds

In Figure 4 we see the predicted and observed values of all the surfactants, both the 18 training set compounds and the 18 in the prediction set. Both sets are seen to be well distributed over both axes, and the prediction set compounds are well predicted.



**Figure 4.** Observed versus predicted and calculated values for  $y = \log LC_{50}$  of the  $N=18 + 18$  training (filled diamonds) and prediction set surfactants (open squares). a) *Thamnocephalus platyurus* and b) *Brachionus calyciflorus*.

#### 4.1.5. Conclusion of the Surfactant Example

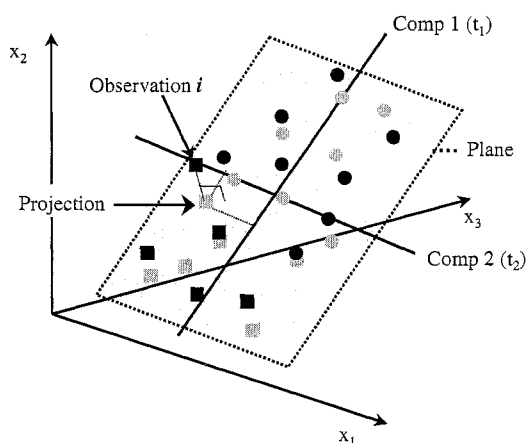
The excellent predictions of the remaining  $n=18$  surfactants from their  $K=19$  structure variable values ( $x_k$ ) demonstrates the possibility for constructing predictive QSAR / QSPR models. The selection of the model training set according to a design makes the results interpretable and the model having predictive power over the whole structural domain of the given 36 compounds.

### 5. Multivariate Analysis by Means of Projections

In the previous example (surfactants) the structure descriptor matrix  $\mathbf{X}$  of dimension  $36 \times 19$  was compressed to a  $(36 \times 2)$  dimensional matrix,  $\mathbf{T}$ . This was done to have an adequate representation of the compounds for the selection of a training set, *i.e.*, the statistical molecular design (SMD). The compression was made using a method of multivariate projection, the so called principal component analysis (PCA), further discussed below. These projections can be understood geometrically in terms of a  $K$ -dimensional space where each object (row of  $\mathbf{X}$ ) is represented as a point, and hence the  $N \times K$  data table is a swarm of  $N$  points.

By means of perturbation theory it can be shown that as long as there is some degree of similarity between the objects – corresponding to the rows in the data table,  $\mathbf{X}$  – then the data swarm can be well approximated by a low dimensional plane or hyper-plane in this space. And the greater the degree of similarity, the fewer dimensions (components, latent factors) are needed for this hyper-plane to have a given faithfulness of approximation [29].

In the present context we use two variants of multivariate projections, namely principal component analysis (PCA) and projections to latent structures using partial least squares (PLS). The former, PCA, projects a matrix  $\mathbf{X}$  to a matrix  $\mathbf{T}$  in an optimal way, *i.e.*, makes  $\mathbf{T}$  summarize  $\mathbf{X}$  as well as possible according to the least squares criterion. The latter, PLS, is used when besides the data matrix  $\mathbf{X}$ , there is also a response matrix  $\mathbf{Y}$ . PLS then makes a projection of  $\mathbf{X}$  to  $\mathbf{T}$  with two objectives, namely that (a)  $\mathbf{T}$  provides a good summary (not quite optimal) of  $\mathbf{X}$ , and (b) that  $\mathbf{T}$  is well correlated with the response matrix  $\mathbf{Y}$ .



**Figure 5.** Multidimensional space where each object is a point, and a plane gives a good approximation of the data (the  $N$  object points in  $\mathbf{X}$ ). Each object is projected onto the plane, giving the coordinate values = score values ( $t_1$  and  $t_2$ ), which when plotted, gives a picture of  $\mathbf{X}$ .

With both PCA and PLS, the resulting "score matrix"  $\mathbf{T}$  is a linear combination of the original  $\mathbf{X}$ -variables. The number of columns of  $\mathbf{T}$  ( $A$ ) is small, usually two to four, and they are orthogonal, *i.e.*, completely independent.

PCA is useful to compress a matrix of structure descriptors to a few "principal properties", PP's – the columns of  $\mathbf{T}$  [2]. These PP's can then be used as the basis of a statistical molecular design (SMD), *i.e.*, for the selection of a minimal set of compounds that well represent the total set of molecules of a given investigation.

## 5.1 Principal Component Analysis (PCA)

The principles of PCA are very simple. Pertinent reviews are given by Jackson [30] and Wold *et. al.* [31]. The  $N$  row vectors of the  $N \times K$  data matrix  $\mathbf{X}$  (*e.g.*,  $K$  descriptors of  $N$  compounds) are represented as a swarm of points in a  $K$ -dimensional space. The axes of this space are usually normalized to the same length ( $1/N$ , *i.e.*, unit variance of each variable). This is accomplished by dividing each column in  $\mathbf{X}$  by its standard deviation. Also, the data are usually centered before the analysis, *i.e.*, the mean value is subtracted from each column.

Due to correlations between the  $K$  variables (columns of  $\mathbf{X}$ ) the point swarm is not round, but rather looks like an elongated pancake. And the more similar the objects (here compounds) are, the more closely the data lie to this elongated pancake, an  $A$ -dimensional hyper-plane (Figure 5).

Algebraically, this corresponds to the modelling of the (centered and scaled)  $N \times K$  matrix  $\mathbf{X}$  by the product of an  $N \times A$  matrix  $\mathbf{T}$  and an  $A \times K$  matrix  $\mathbf{P}'$  plus an  $N \times K$  residual matrix,  $\mathbf{E}$ .

$$\mathbf{X} = \mathbf{T} \mathbf{P}' + \mathbf{E}$$

The score matrix,  $\mathbf{T}$ , optimally summarizes the information about the objects (compounds), and are hence often called the matrix of principal properties, PP's. Analogously, the loading matrix,  $\mathbf{P}$ , summarizes the information about the variables. Objects (index  $i$ ) that are similar will have similar values of the row vectors  $\mathbf{t}_i'$ , and objects that are dissimilar will have dissimilar values of these row vectors. Hence these row vectors can be used to select a set of "diverse" compounds as those with as dissimilar row vectors,  $\mathbf{t}_i'$ , as possible. This is the basis of SMD based on principal properties (PP's). Analogously, variables (index  $k$ ) with similar values of their loading vectors,  $\mathbf{p}_k$ , will have a similar information, they are strongly correlated. *Vice versa*, variables with dissimilar loading vectors are dissimilar, have different information content.

We shall here use this property of the  $\mathbf{T}$  matrix of summarizing  $\mathbf{X}$  to select "diverse" sets of compounds that provide an optimally "diverse" (spanning) information for a given objective. Interestingly, this means that the library size in combinatorial chemistry can be reduced to a few hundreds of compounds without loss of structural information. Hence, a much deeper and broader biological testing can be made making the total resulting information about the combination of structure and activity vastly superior to that of a large library that is crudely tested by HTS.

## 5.2 A Combinatorial Chemistry Application

This example is presented as a small but fairly realistic illustration of a reasonable approach to solve the "combinatorial curse of testing", *i.e.*, the inability to make an adequate biological testing of a large combinatorial library of compounds. The recourse to a HTS (high throughput screening) testing of all compounds in a large library has many

serious problems, the most serious in our view being the very low information content in the resulting test data about the "real" clinical activity, toxicity, bio-availability, uptake properties, *etc.* Hence, a selection of compounds based on their HTS results is highly risky in that it is based on very limited information.

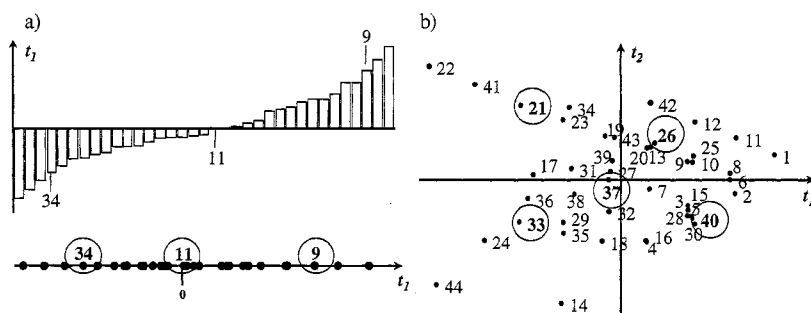
To get around the "combinatorial curse of testing", we recommend the obvious approach to make and test only a small set of selected compounds which adequately represents the structural variation of the whole potential library. By basing the selection on small sets of representative building blocks, one arrives at surprisingly small numbers of compounds needed to be made and tested. Hence, this small set of compounds can be tested much broader and deeper, thus providing a much more reliable biological basis of data for the following step of compound selection. This approach has been presented in several recent papers [16, 32-35], and much of the present example is taken from ref. [35].

Consider a combinatorial library consisting of the products of the reaction between a primary aliphatic amine and an aromatic aldehyde. And let us assume that we have access to building block libraries of  $n_1 = 35$  primary amines and  $n_2 = 44$  aromatic aldehydes. The full combinatorial library would comprise  $35 \times 44 = 1540$  products. We can now ask whether all these really are needed. And can we really test them?

We shall use SMD (statistical molecular design) to select a small but representative set of amines (with 3 members) and a second small but representative set of aldehydes (with 5 members). Finally, we shall combine the two sets to a small library with only  $n_{\text{final}} = 9$  compounds. This is small enough to allow an extensive biological testing of all its members.

This approach involves a number of steps, namely (1) characterizing the candidate structures, (2) making a compact representation using PCA, and (3) selecting spanning compounds, and finally (4) making the final design of the library of combined building blocks.

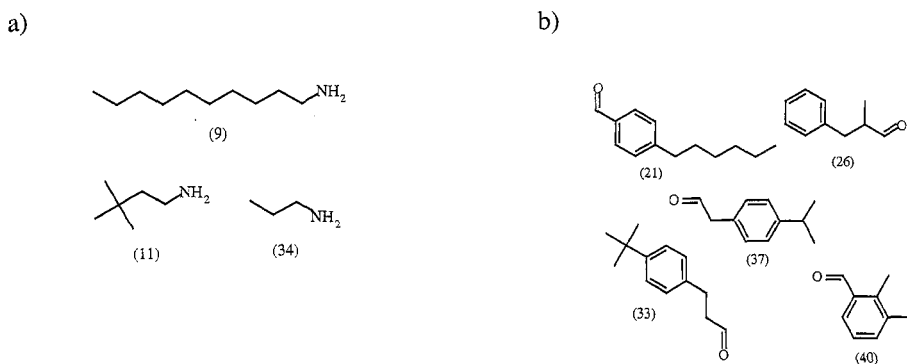
To allow a selection of compounds, a quantitative description of their structures must first be made. Lundstedt *et al.* investigated amines for synthetic objectives [9] and described  $n_1 = 35$  primary amines by means of  $K_1 = 11$  descriptors, including their  $\text{pK}_a$ , molecular weight and volume, and  $\log P$ . A PCA of the resulting  $35 \times 11$  matrix (centered and scaled to unit variance) gave one significant component. Hence, the selection of primary amines can be considered as a one dimensional problem, and three compounds would suffice to give a representative set; one with a low, one with a medium, and one with a high score value. The PC score values and the selected compounds are shown in Figures 6 a and 7 a.



**Figure 6.** a) (left) shows a bar chart of the score values resulting from the PCA of the  $35 \times 11$  amine descriptor matrix with the three selected compounds indicated on the line plot under the bar chart. Analogously, b) (right), shows the plot of the two PC scores of the 44 aromatic aldehydes together with the five selected compounds (rings).

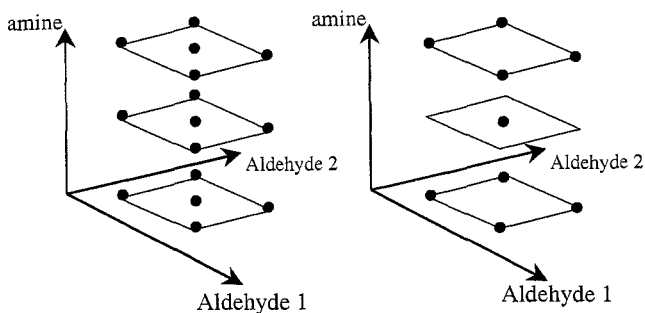


Similarly, the 44 aromatic aldehydes are characterized by  $K=54$  descriptors by means of simple quantum chemical and molecular mechanical calculations [36]. Here the PCA of the resulting  $44 \times 54$  matrix (centered and scaled to unit variance) gave two significant components. Hence, five compounds selected according to a factorial design plus a center point in the two PC scores would suffice to give a representative set. The PC score values and the selected compounds are shown in Figures 6 b and 7 b.



**Figure 7.** Building block libraries of the a) amines and b) aldehydes.

Finally, when sets of building blocks have been selected, these are combined to give the final library. Also this step can be made by means of statistical design, making the final library a representative subset of the full set of all combinations of the building blocks. This is done by considering each coordinate in the building block libraries (one in the amines and two in the aldehydes) as a quantitative variable in the final design. A linear model including interaction terms would have 7 terms (one constant, three linear "scores" and 3 cross-terms, interactions), and hence a final library with  $n_{\text{final}} = 9$  would constitute a minimal design. This is indicated in Figure 8.



**Figure 8.** The final design of the library is a combination of the building block coordinates (here PC scores) according to a sparse design. The full set of combinations of the two building blocks (left) gives an unnecessarily large library. A designed combination of each sets of building blocks gives a representative, spanning, library (right picture).

With this small example we have demonstrated that a surprisingly small subset of compounds (here  $n_{\text{final}} = 9$ ) will suffice as representative of the whole combinatorial library (here  $n_{\text{total}} = 1540$ ). In more complicated examples, the clustering of each building block library must be taken into account, but the resulting dramatic decrease in the numbers of final library compounds remains the same also in this situation [32,35].

After testing the resulting final library in a broad and deep set of biological tests, one can finally use the resulting data to construct a model relating the variation in structure ( $\mathbf{X}$ ) to the variation in biological activity ( $\mathbf{Y}$ ). This typically done using PLS as discussed in the next section. With the PLS model one can then predict interesting directions in the structural space for further exploration, thus having a rational basis for drug design.

### 5.3 Projections to Latent Structure by Partial Least Squares (PLS)

In sections 5 and 5.1, the idea of multivariate projections was briefly discussed. These projections (PCA and PLS) summarize a matrix  $\mathbf{X}$  (here describing structure) to a few independent scores,  $t_a$  ( $a=1,2,\dots,A$ ). PLS differs from PCA in that it makes use of a response matrix,  $\mathbf{Y}$ , to focus the PLS projection. Hence, the resulting score vectors ( $t_a$ ) differ from those of PCA, and are more correlated with the columns of  $\mathbf{Y}$ .

The advantages of PLS for relating a structure matrix  $\mathbf{X}$  to an activity matrix  $\mathbf{Y}$  are several compared with, for instance, traditional multiple regression. First, PLS can deal with very many structure descriptors even when  $N$  -- the number of compounds (rows in  $\mathbf{X}$  and  $\mathbf{Y}$ ) -- is small. Second, PLS can deal with noise, missing data, and inadequacies in the descriptor matrix ( $\mathbf{X}$ ). Third, PLS can simultaneously model several or all responses in the activity matrix,  $\mathbf{Y}$ , making the use and interpretation of the model simpler in comparison with the use of one model for each response.

The resulting PLS model is interpretable by means of its loadings and weights ( $w_a$ ) which show how the original structure descriptor variables are combined to form the scores,  $t_a$ . Additional diagnostics include residuals and their summaries, both for  $\mathbf{X}$  and for  $\mathbf{Y}$ .

PLS can be used also for classification. Then the  $\mathbf{Y}$ -matrix is set up to contain column of ones and zeros corresponding to the class membership of the compounds and  $\mathbf{X}$  contains a quantitative description of the structure. The scores resulting from the subsequent PLS analysis indicates the resolution of the classes, and the PLS-weights of the model indicates which variables that are important for the separation of the classes.

The use of PLS for modelling structure – activity relationships has been reviewed in several recent articles [37-39].

### 5.4 Some Bioinformatics Applications

The emerging field of bioinformatics [40,41] concerns relationships between the polymer sequences in genetic material (DNA or RNA) and 'proteins and biological "properties" of interest. These "properties" may be properties of the polymers themselves (folding, binding of substrates or inhibitors, *etc.*) or of the organisms carrying the polymers (*e.g.*, resistance to drugs, susceptibility to infection, genetically related defects, classification in genetic groups).

We here point out the utility of SMD and multivariate models also in these application types. Several interesting results of the use of these tools have already emerged. The first is the translation of amino acid sequence or nucleotide sequence to a quantitative representation. Hellberg *et. al.* described the 20 coded amino acids by 29 measured and calculated properties, and used PCA to derive three "principal property" (PP) scales ( $z_1$ ,  $z_2$ , and  $z_3$ ) for the amino acids [2]. They also showed that these scales

could be used to get a quantitative representation of the sequences of peptides and proteins, and that indeed this description was strongly related to biological properties of families of peptides and proteins [3,4]. Similar results have been shown by Fauchere *et. al.* [6]. Recently, this work was extended by Sandberg *et. al.* [5] to 87 amino acids (20 coded and 67 others) and totally 5 scales where the first three strongly resemble the original PP scales.

Hence, instead of describing peptide or nucleotide sequences by means of characters (Figure 9), we now have a pertinent quantitative description ( $X$ ) which then can be related to measured properties ( $Y$ ) for a family of sequences. Several examples are given in refs. [2,5, 42-45].

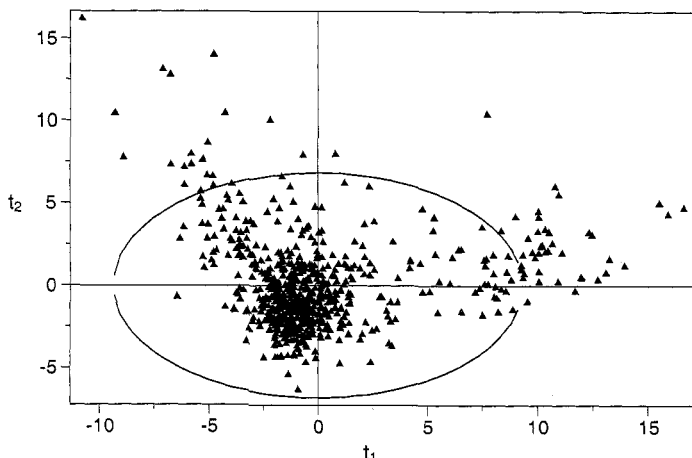
**Figure 9.** The traditional way to describe sequences as strings of characters. Here a set of signal peptides from ref. [45].

```

MKQSTIALALLPLLFPPVTKA
MKVMRTTVATVVAATLSMSAFSVFA
MKIKTGARILALSALTTMMFSASALA
MKRNAKTIIAGMIALAISHTAMA
MNTK GKALLAGLIALAFSNMALA
MNKKVLTLSAVMASMLFGAAAHA
MHKFTKALAAIGLAAVMSQSAMA
MFKTTLCALLITASCSTFA
MNMKKLATLVSVAVALSATV SANAMA
MKKLFASLALAAVVPVWA
MKFSATLLATLIAASVNA
MKLLQRGVALALLTFTLASETALA
MKSVLKVSALALTLFAVSSHA
MKMNKSLIVLCL SAGLLASAPGISLA
MKNRNRMI VNCVTASLMYYW SLPALA

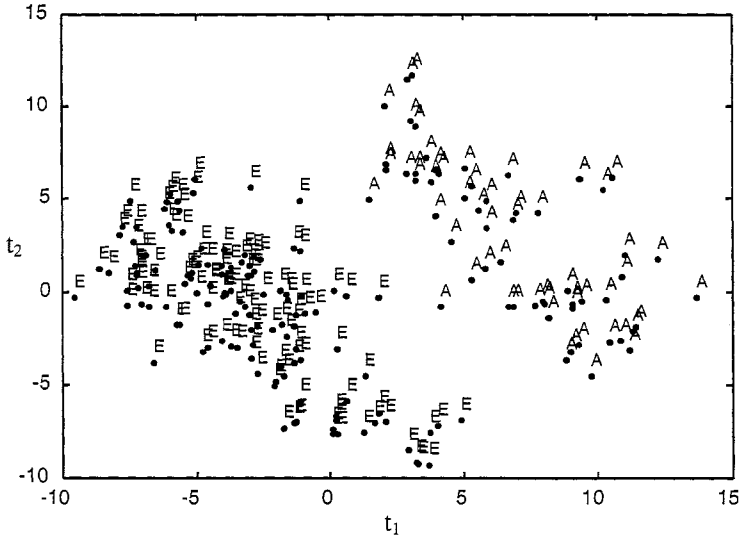
```

Second, the same group showed how to deal with sequences of varying length with tools borrowed from time series analysis, namely auto and cross-correlation spectra. These describe the variation of the PP's along the sequence of one polymer, and are translationally and alignment independent [44]. Sjöström, Wieslander, *et. al.* applied this to the classification of signal peptides of different lengths [45] and recently to the quantification and visualization of all proteins in an organism (Figure 10).



**Figure 10.** All proteins in *Mycoplasma pneumoniae*. The genome was first translated to amino acids, then each each position was translated to the three z-scale values. Auto and cross correlation spectra of the z-values along the sequences were calculated, and finally a PCA was made of the resulting matrix,  $X$ . The picture shows the first two PC scores of this analysis.

Finally, in a third "bioinformatics" example, we show the partial results of a PLS-discriminant analysis of two classes of bacteria -- E= eubacteria and A=archaeobacteria. N= 190 sequences of length 74 were translated to a numerical representation using the nucleoside scales recently developed by Sandberg *et. al.* [43]. Figure 11 shows the resulting discriminant scores and a clear separation between the two classes. The corresponding PLS weights indicate that the most important positions for the separation are 35-37 and 42-44, and that the principal property of importance in all these positions is the one of polarity.



**Figure 11.** A PLS-DA was made of the aligned tRNA sequences (length= 74) of E= eubacteria and A=archaeobacteria. Each RNA position was described by four values of the nucleotide principal property scales of Sandberg *et. al.* [40]. The figure shows the resulting X-scores ( $t_1$  and  $t_2$ ) of the different bacterial strains.

The tools of multivariate analysis – PCA and PLS – allow the development of a quantitative approach to bioinformatics. This starts with the translation of sequences to vectors of quantitative descriptors followed by modelling the relation between sequence and "biological properties" by means of PLS discriminant analysis for classification or ordinary PLS for the modelling of continuous properties. Whenever there is some kind of experimental control in the investigation, like for instance in site directed mutagenesis, one should use SMD for selecting representative molecules (peptides, proteins, nucleic acids, *etc.*) for the questions being asked. Thus, it would be impractical to modify one position at a time in these sequences. Only a planned modification of several positions in terms of a statistical design provides information about the joint influence of these positions on the properties of interest.

When there is little possibility for experimental intervention, sampling aspects are more dominating than those of design. Sampling is analogous to design, but instead one samples in a space of time, geography, age and sex of patients, *etc.*, in order to get representative and balanced data. Exactly the same principles as those used in design can be used to get a set of samples (objects, sequences, ..) that well span the abstract space of interest.

## 6. Conclusions

The complexity of chemical / biological systems relative to our limited brains, makes *modelling* the only feasible approach to their investigation and (partial) understanding. This is especially clear after the works of scientific giants such as Heisenberg, Schrödinger, Bohr, Dirac, and Gödel. Since all models are based on data (and theory), the quality and representativity of these data is essential for the reliability, usefulness, and interpretability of the models. The methodology to maximize quality and representativity of the X-data (here the structure descriptors) for a given modelling is called statistical experimental design. The only alternative to the use of design, is to have very large data sets, which is, at best, inefficient, and at worst confusing. Of course we also need good Y-data, *i.e.*, good and representative and therefore multivariate, measurements of the biological properties of the investigated systems. This is usually well understood. However combinatorial chemistry and HTS constitute an exception to this understanding.

When applied to the selection of molecules / polymers / this use of experimental design is called "Statistical molecular design", SMD. Without such design, modelling in the fields of QSAR and Combinatorial Chemistry is difficult to impossible. This is, in our view, a major explanation for the slow progress seen in these fields.

In bioinformatics there is usually little possibility for experimental intervention, and hence sampling aspects are more dominating than those of design. We just emphasize that sampling is analogous to design, but instead one samples in a space of time, geography, age and sex of patients, *etc.*, in order to get representative and balanced data. In this field, there is a great potential in making the models quantitative and multivariate, possibly along the lines outlined above.

The difficulties with the methods of statistical design and multivariate analysis are that they in the beginning seem counterintuitive and too mathematical. Since they are not yet taught much in university chemistry and biology, they have to be learnt outside the curriculum. This takes much motivation and insight, and hence the spread of these methods is still slow.

## Acknowledgements

Financial support from the Swedish Natural Science Research Council, SFR, the Commission of the European Union under the contract number ENV4-CT96-0221, Astra Hässle AB and Pharmacia & Upjohn AB to the Umeå Chemometrics Group are gratefully acknowledged.

The SAS (Swedish Acronym Society) is thanked for endless support and encouragement.

## References

1. C. Hansch, T. Fujita.  $\rho$ - $\sigma$ - $\pi$ -Analysis. A method for the correlation of Biological Activity and Chemical Structure, *J. Am. Chem. Soc.*, 1964, 86, 1616-1626
2. S. Hellberg, M. Sjöström, S. Wold, The Prediction of Bradykinin Potentiating Potency of Pentapeptides. An Example of a Peptide Quantitative Structure-Activity Relationship, *Acta Chem. Scand.*, 1986, B40, 135-140
3. S. Hellberg, M. Sjöström, B. Skagerberg, C. Wikström, S. Wold, On the design of multipositionally varied test series for quantitative structure-activity relationships, *Acta Pharm. Jugosl.*, 1987, 37, 53-65
4. J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, S. Wold, Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids, *Quant. Struct.-Act. Relat.*, 1989, 8, 204-209
5. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterisation of 87 Amino Acids, *J. Med. Chem.*, 1998, 41, 2481-2491

6. J.L. Fauchere, M. Charton, L.B. Kier, A. Verloop, V. Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology, *Int. J. Pept. Protein. Res.*, 1988, 32, 269-78
7. R. Carlsson, M. P. Prochazka, T. Lundstedt, Principal Properties for Synthetic Screening: Ketones and Aldehydes, *Acta Chem. Scand.*, 1988, B42, 145-156
8. T. Lundstedt, R. Carlsson, R. Shabana, Optimum Conditions for the Willgerodt-Kindler Reaction. 3. Amine Variation, *Acta Chem. Scand.*, 1987, B41, 157-163
9. R. Carlsson, M. P. Prochazka, T. Lundstedt, Principal Properties for Synthetic Screening: Amines, *Acta Chem. Scand.*, 1988, B42, 157-165
10. R. Carlsson, T. Lundstedt, Scope of Organic Synthetic Reactions. Multivariate Methods for Exploring the Reaction Space. An example by the Willgerodt-Kindler Reaction, *Acta Chem. Scand.*, 1987, B41, 164-173
11. R. Carlsson, Design and optimization in organic synthesis, Elsevier, Amsterdam, 1992
12. L. Eriksson, J. Jonsson, M. Sjöström, S. Wold. A strategy for Ranking Environmentally Occuring Chemicals, *Chemometrics and Intell. Lab. Syst.*, 1989, 7, 131-141
13. L. Eriksson, E. Johansson. Multivariate design and modelling in QSAR, *Chemometrics and Intell. Lab. Syst.*, 1996, 34, 1-19
14. L. Eriksson, E. Johansson and S. Wold, QSAR model Validation, SETAC Press, Pensacola, USA, In press, 1997
15. L. Eriksson, E. Johansson, M. Müller, S. Wold. Cluster-based Design in Environmental QSAR, *Quant. Struct.-Act. Relat.*, 1997, 16, 383-390
16. E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos, Measuring diversity: Experimental design of combinatorial libraries for drug discovery, *J. Med. Chem.*, 1995, 38, 110-114
17. R.D. Cramer, III, D.E. Patterson, J.D. Bunce. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins, *J. Am. Chem. Soc.*, 1988, 110, 5959-5967
18. P. J. Goodford. A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules, *J. Med. Chem.*, 1985, 28, 849-857
19. P. Goodford. Multivariate Characterisation of Molecules for QSAR Analysis, *J. Chemometrics*, 1996, 10, 107-117
20. A. Berglund, C. De Rosa, S. Wold. Alignment of Flexible Molecules at their Receptor Site Using 3D Descriptors and Hierarchical-PCA, *J. Comput. Aided Mol. Des.*, 1997, 11, 601-612
21. J.R. Broach, J. Thorner, High-throughput Screening for Drug Discovery, *Nature*, 1996, Suppl., 384, 14-16
22. P.F. de Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart, R. Phan-Than-Luu. D-optimal designs, *Chemometrics and Intell. Lab. Syst.*, 1995, 30, 199-210
23. E. Marengo, R. Todeschini. A new algorithm for optimal, distance-based experimental design, *Chemometrics and Intell. Lab. Syst.*, 1992, 16, 37-44
24. P.N. Craig, C.H. Hansch, J.W. Farland, Y.C. Martin, W.P. Purcell, R. Zahradnik, Minimal statistical data for structure function correlations, *J. Med. Chem.* 1971, 14, 447
25. C. Hansch, S.H. Unger, A.B. Forsythe. Strategy in drug design. Cluster analysis as an aid in the selection of substituents, *J. Med. Chem.*, 1973, 16, 1217-1222.
26. V. Austel. *Eur. J. Med. Chem.*, 1982, 17, 9-16
27. Å. Lindgren, M. Sjöström, S. Wold. QSAR Modelling of the Toxicity of Some Technical Non-Ionic Surfactants Towards Fairy Shrimps, *Quant. Struct.-Act.-Relat.* 1996, 15, 208-218
28. L-L. Uppgård, Å. Lindgren, M. Sjöström, S. Wold. Submitted *J. Surf. Deterg.*, 1998
29. S. Wold, M. Sjöström. 'Linear Free Energy Relationships as Tools for Investigating Chemical Similarity - Theory and Practice'. In *Correlation Analysis in Chemistry* (Ed. N.B. Chapman, J. Shorter) Plenum Publishing Corporation, 1978
30. J. E. Jackson, A Users Guide to Principal Components, Wiley, New York, 1991
31. S. Wold. Principal Component Analysis, *Chemometrics and Intell. Lab. Syst.* 1987, 2, 37-52
32. T. Lundstedt, P. M. Andersson, S. Clementi, G. Cruciani, N. Kettaneh. A. Linusson, B. Nordén, M. Pastor, M. Sjöström, S. Wold, 'Intelligent Combinatorial Libraries'. In *Computer-Assisted Lead Finding and Optimization* (Ed. H. van de Waterbeemd) Verlag Helvetica Chimica Acta, Basel, Switzerland, 1997, 191-208
33. A. Linusson, S. Wold, B. Nordén, In press. *Chemometrics and Intell. Lab. Syst.*, 1998
34. S. S. Young, D. M. Hawkins, Analysis of a 29 Full Factorial Chemical Library, *J. Med. Chem.*, 1995, 38, 2784-2788
35. P.M. Andersson, A. Linusson, S. Wold, M. Sjöström, T. Lundstedt, B. Nordén. 'Design of Small Libraries for Lead Exploration'. In *Molecular Diversity in Drug Design* (Ed. R. Lewis, P.M. Dean) In press 1998
36. Tsar 3.11, Oxford Molecular Group, [www.oxmol.co.uk](http://www.oxmol.co.uk)

37. S. Wold, E. Johansson, M. Cocchi. 'PLS - Partial Least-Squares Projections to Latent Structures'. In 3D QSAR in Drug Design; Theory, Methods and Applications. (Ed. H. Kubinyi) ESCOM Science Publishers, Leiden, Holland, 1993, 523-550
38. S. Wold. 'PLS for Multivariate Linear Modeling'. In QSAR: Chemometric Methods in Molecular Design, Methods and Principles in Medicinal Chemistry, Vol 2., (Ed. H. van de Waterbeemd) Verlag Chemie, Weinheim, Germany, 1995, 195-218
39. F. Lindgren, M. Sjöström, S. Wold. PLS-modelling of detergency performance for some technical nonionic surfactants, *Chemometrics and Intell. Lab Syst.*, 1996, 32, 111-124
40. E. Marshall, Bioinformatics: Hot Property: Biologists Who Compute, *Science*, 272 (1996) 1730-1732
41. J. B. Grace, Bioinformatics: Mathematical Challenges and Ecology, *Science*, 275 (1996) 1861c-1865c
42. J. Jonsson, M. Sandberg & S. Wold, The Evolutionary Transition from Uracil to Thymine Balances the Genetic Code, *J. of Chemometrics*, 1996, 10, 153-170
43. M. Sandberg, M. Sjöström, J. Jonsson. A Multivariate Characterization of tRNA Nucleosides, *J. of Chemometrics*, 1996, 10, 493-508
44. S. Wold, J. Jonsson, M. Sjöström, M. Sandberg, S. Rännar, DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by PCA and PLS Projections to Latent Structures, *Anal. Chim. Acta*, 1993, 227, 239-253
45. M. Sjöström, S. Rännar, Å. Wieslander. Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances, *Chemometrics and Intell. Lab. Syst.* 1995, 29, 295-305

# QSAR STUDY OF PAH CARCINOGENIC ACTIVITIES: TEST OF A GENERAL MODEL FOR MOLECULAR SIMILARITY ANALYSIS

William C. Herndon, Hung-Ta Chen,  
Yumei Zhang, and Gabrielle Rum

Department of Chemistry  
The University of Texas at El Paso  
El Paso, Texas 79968

## INTRODUCTION

Several QSAR methodologies have been developed which make use of hierarchical sets of molecular descriptors, coupled with multilinear regression analysis of physical or biological properties. Our procedures advance through enumerations of types of atoms and bonds (level 1), rings and functional groups (level 2), larger structural fragments and steric interactions (level 3), and end by testing the addition of level 4 descriptors based on the results of semiempirical or *ab initio* molecular orbital calculations. Experimental properties (e.g., logP, boiling points, etc.) are an additional possible source of descriptors, not tested in the present work. In general, the levels of hierarchical structural descriptors are augmented and tested sequentially to obtain information regarding the lowest levels of description that are necessary for statistically significant rectification of a particular dependent variable property. High quality, structure/property and structure/activity relationships are normally found that use significant terms from several descriptor levels.<sup>1-5</sup> In previous work, we have also shown how various types of molecular structure codes or molecular descriptors can be used to calculate measures of molecular similarity.<sup>6-9</sup>

In this paper a more general, simpler protocol to obtain molecular similarity measures is outlined which can be used for arbitrary sets of compounds and descriptors, either globally or at any restricted level of molecular description. We then illustrate how the numerical values of similarity to particular compounds, chosen by statistical multilinear regression analysis, can function as independent variables in QSAR model equations. The methodology is tested by correlating a complex biological endpoint, consisting of results of animal studies of carcinogenic activities of polycyclic aromatic hydrocarbons containing a large variety of types of aromatic rings and hydrocarbon alkyl substituents. We also attempt to assess predictive capabilities of the overall protocol by using a robust modification of a cross-validation method in which the twelve most active and six least active compounds, i.e., 20% of the cases, are excluded from the QSAR model equation development.



## PAH CARCINOGENIC ACTIVITIES

The carcinogenic polycyclic aromatic hydrocarbons include a relatively large class of compounds which contain fused six-membered benzene rings and five-membered rings as well as alkyl substituents. The abbreviations PAH and PAHs will be used to designate both the pure aromatic structures and their alkyl derivatives. A detailed review of the extant animal assay data for PAH carcinogenicities was undertaken.<sup>5</sup> These data were generally obtained from an examination of results abstracted in the series "Survey of Compounds Which Have Been Tested for Carcinogenic Activity." Public Health Service Publication No. 149, 15 volumes and two supplements, 1951-1992. All volumes from inception of publication were examined.

Active PAHs consisted of 210 active compounds of 312 that were tested. An index of carcinogenicity was assigned to every compound where the latent period was measured (90 compounds). The carcinogenicity index is defined analogous to the Iball index,<sup>10</sup> proportional to the percent of animals developing cancer and inversely proportional to latent period. The proportionality factor was taken to be 100 and latent periods were measured in days. Values were averaged over all reported experiments. Studies using promoters were weighted using a factor of 0.5. The derived index (HZACT) for these 90 compounds is the dependent variable in the QSAR analysis which is given below. The names of the compounds and their HZACT values are given in Table 1, sorted by activity.

## MOLECULAR DESCRIPTORS

The lowest level of molecular descriptors, derived from molecular structure drawings, was comprised of counts of types of carbon atom groups based on the hybridization state of the carbon atom. Thus saturated carbon atoms were divided into the usual quaternary, tertiary, secondary, and primary groups. Aromatic  $sp^2$  CH and substituted C atoms were distinguished from olefinic  $sp^2$  atoms at this level. Indicator variables for 15 varieties of aromatic five and six-membered rings constituted the next level of parameters. Saturated aliphatic rings, few in number, were only represented by their level 1 constituent groups.

Functional group indicator variables are not required for the PAHs. However, early in the course of this investigation, we discovered that indicator variables for classification of the aromatic ring systems corresponding to the unsubstituted prototype structures led to significant improvements in statistical correlations of the derived HZACT index. In fact, model equations developed solely with levels 1-3 atom and ring descriptors provided terrible correlations of the derived HZACT index. Thus the use of descriptors signifying the type of pi-system substructure, i.e. benz[a]anthracene, benz[e]pyrene, cholanthrene, etc., was mandatory for obtaining statistically significant ( $R^2 > 0.5$ ) rectifications of activities.

The next descriptor level consisted of parameters derived from AM1 calculations using the QSAR keyword of the SPARTAN computational chemistry software package from Wavefunction, Inc. The descriptors used in this work were the calculated values of heats of formation, E(HOMO), E(LUMO), electronegativities, polarizabilities, hardness, molecular volumes, surface areas, ovalities, logP, and dipole moments. The Mulliken population analyses at particular bay-region atoms and bonds (charges and bond orders) were also coded but will not be used for the study reported here.

The final level of descriptors was comprised of three preselected, less intuitive structural parameters, each of which turned out to be a significant factor in this QSAR study. The identification of these descriptors was based on the following. Many of the PAHs under consideration are highly nonplanar,<sup>11</sup> due either to the presence of methyl groups in a bay-region or as a result of the molecule containing a benz[c]phenanthrene fjord

Table 1. Carcinogenicity activity indices for 90 PAHs (HZACT; See text for definition.)

7,8,12-Trimethylbenz[a]anthracene	146.5	6,7,12-Trimethylbenz[a]anthracene	36.2
Dibenzo[a,l]pyrene	122.6	Benzo[a]pyrene	33.8
2,3-Dimethylbenzo[a]pyrene	117.7	Dibenz[a,h]anthracene	33.6
Benz[a]aceanthrylene	110.1	7-Methyl-8,9-ace-1,2-benzanthracene	33.3
6,7,8-Trimethylbenz[a]anthracene	104.2	4H-Cyclopenta[def]chrysene	33.0
1,4-Dimethylbenzo[a]pyrene	103.1	7-Methylbenz[a]anthracene	31.4
5,9-Dimethyl-1,2-benzanthracene	102.6	6,7-Dimethyl-1,2-benzanthracene	29.2
Dibenz[a,j]aceanthrylene	102.0	1,12-Trimethylenechrysene	26.2
Benzo[b]fluoranthene	99.6	7,14-Dimethyldibenz[a,j]anthracene	25.4
1,3-Dimethylbenzo[a]pyrene	99.0	4,10-Ace-1,2-benzanthracene	22.6
3,12-Dimethylbenzo[a]pyrene	97.1	1,3,6-Trimethylbenzo[a]pyrene	22.4
4,9-Dimethyl-1,2-benzanthracene	91.8	5,11-Dimethyl-chrysene	21.8
9-Methyl-1,2,5,6-dibenzanthracene	88.9	12-Methylbenz[a]anthracene	18.7
Dibenzo[a,h]pyrene	78.8	1',9-Methylene-1,2,5,6-dibenzanthra	17.7
22-Methylcholanthrene	77.4	7,9,12-Trimethylbenz[a]anthracene	16.6
9,12-Dimethyl-1,2-benzanthracene	76.8	7-Methylbenzo[a]pyrene	16.4
11-Methylbenzo[a]pyrene	73.4	5,6-Cyclopenteno-1,2-benzanthracene	16.3
2-Methylbenzo[a]pyrene	73.4	8-Methylbenz[a]anthracene	14.3
Dihydro-20-methylcholanthrene	73.3	Dibenzo[a,e]aceanthrylene	12.7
5,10-Dimethyl-1,2-benzanthracene	72.7	Benzo[a]anthracene	12.3
4-Methylbenzo[a]pyrene	70.0	Indeno[1,2,3-hi]chrysene	11.9
1,2-Dimethylbenzo[a]pyrene	69.5	Dibenz[a,j]anthracene	11.6
Dibenzo[a,i]pyrene	67.9	Benzo[e]pyrene	11.3
4,10-Dimethyl-1,2-benzanthracene	64.9	1,2-Cyclopenteno-5,10-aceanthrene	11.0
3-Methylbenzo[a]pyrene	62.9	1',2'-Dihydro-4'-methyl-3,4-benzpyren	8.4
12-Methylbenzo[a]pyrene	60.4	Dibenz[a,c]anthracene	7.1
16,20-Dimethylcholanthrene	59.5	10-Methylbenzo[a]pyrene	6.8
4,5,10-Trimethylbenz[a]anthracene	57.7	3'-Methyl-1,2,5,6-dibenzanthracene	6.5
1,2,3,4-Tetrahydro-7,12-DMB[a]A	56.6	Dibenzo[a,e]pyrene	6.4
20-Methylcholanthrene	56.1	4-Methylbenz[a]anthracene	5.0
4,5-Dimethylbenzo[a]pyrene	54.5	Dibenzo[cd,lm]perylene	4.9
3-Methylcholanthrylene	54.4	7-Methylbenzo[ppq]picene	4.2
5-Methylchrysene	54.0	10-Methyldibenz[a,c]anthracene	3.6
6,8,12-Trimethylbenz[a]anthracene	52.3	Dibenzo[def,mno]chrysene	3.4
23-Methylcholanthrene	52.1	5-Methylbenz[a]anthracene	3.3
7,11-Dimethylbenz[a]anthracene	52.0	9-Methylbenz[a]anthracene	3.1
4,5-Dimethylbenz[a]anthracene	50.4	Benzo[a]napho[1,2-k]chrysene	2.5
Cholanthrene	49.0	Coronene	2.1
6,8-Dimethylbenz[a]anthracene	46.7	Anthracene	1.6
6-Methylbenzo[a]pyrene	45.0	10-Methylbenz[a]anthracene	1.3
Meso-dihydrocholanthrene	41.7	1-Methylbenzo[a]anthracene	1.2
2-Methyl-3,4-benzphenanthrene	41.6	Pyrene	1.1
8,9-Ace-1,2-benzanthracene	40.8	Perylene	0.8
1,6-Dimethylbenzo[a]pyrene	40.5	1,7,12-Trimethylbenz[a]anthracene	0.7
3,6-Dimethylbenzo[a]pyrene	38.4	3'-Methylcyclopentenophenanthrene	0.2

substructure. Therefore, indicator variables denoting these structural features, BAYC1 and NONPLANA, respectively, were included in the data matrix. The global descriptor DELTA21C, which has been identified in previous PAH carcinogenicity studies,<sup>12,13</sup> was also included. This parameter is the absolute value of the difference, 21 minus the number of carbon atoms. The importance of this parameter in the previous investigations has been postulated to be related to an optimum molecular size for expressing carcinogenicity. These parameters were introduced at two points in model equation development, first as data used in the similarity analysis (see below), and later as separate auxiliary parameters, highly significant for correlation of the PAH activity data.

## MEASURES OF MOLECULAR SIMILARITY

The starting point for the similarity analysis is the usual type of N-by-M data matrix, where N is the number of rows (compounds) and M is the number of columns containing numerical values of all descriptors. The Pearson correlation matrix of this data table is an M-by-M square matrix which describes the linear correlations of the descriptors with each other, based on the set of N compounds. In many previous applications, the Pearson correlation matrix has been utilized to select subsets of descriptors for use as trial independent variables in QSAR multilinear regression studies.

The Pearson correlation matrix methodology can also be employed to define a (molecular) similarity matrix for the set of N compounds as follows. In the first step, the descriptor data matrix is standardized by subtracting means and dividing by the standard deviations for each one of the descriptor columns. This puts all the descriptors on a common standard scale by removing the undue influence of descriptors with large outlying numerical values. Then, for N compounds, an N-by-N similarity matrix is defined to be the Pearson correlation matrix for the transpose of the standardized matrix of the M molecular structure descriptors. Each column in the new similarity matrix represents pairwise numerical values of similarity (positive values) or dissimilarity (negative values) to a single compound. Multilinear regression analysis is then used to identify statistically significant similarities and dissimilarities to a (small) set of reference molecules which correlate the activities of the entire set of PAHs. Thus, the similarities and dissimilarities to the reference molecules may provide independent variables for a quantitative similarity/activity model equation.<sup>8,9</sup>

Generality and ease of interpretation are two advantages of this approach for defining measures of molecular similarity. The general nature of the procedure is obvious and does not require amplification. To understand interpretation of the similarity values, it is sufficient to know that each pairwise Pearson similarity term is simply the slope of the linear regression equation relating the standardized descriptors of the two compounds.

## RESULTS, CORRELATION, AND PREDICTION

The use of the original structural and AM1 descriptors to develop multilinear models which correlate the HZACT index leads to several moderately successful equations, with 10 or more significant parameters, which will not be discussed due to the mandated space limitation. The focus of this report, use of the similarity measures as independent variables, gives improved correlation of the activity data with fewer parameters. However, we would also like to possess a knowledge of the predictive capabilities of the regression models developed in the present application and in several related studies under investigation. We understand that the use of multilinear regression models derived from large data sets of descriptors for prediction is questionable.<sup>14,15</sup> In our opinion the use of the usual cross-validation procedures (leave-one-out, leave-many-out, or leave a random-sample-out) also don't really test predictive validity.

A result will be presented below which employs an unusual type of cross-validated analysis, designed to provide a more stringent test of predictive capability. The model equation using similarity parameters as independent variables is obtained leaving out 12 compounds with the highest activities (Table 1, HZACT > 90) and also leaving out the six compounds with HZACT lower than 1.5. The model for prediction is developed using the remaining compounds possessing the intermediate activities. We infer that this procedure is a more robust measure of predictive capabilities than other types of leave-out tests, and that the resulting calculated activities for the most active and least active compounds may constitute authentic predicted values of activity.

The result for the analysis of the 72 compound development set is given in Table 2. The first four parameters listed after the constant term are similarity measures to benz[a]pyrene, dibenzo[ah]pyrene, cholanthrene, and 6,8,12-trimethylbenz[a]anthracene. Stepwise regression allows inclusion of four additional terms in the final acceptable data rectification, the size and nonplanar terms mentioned previously, and the count of methyl groups attached to aromatic rings (CH3ARO). The results are illustrated in Figure 1 which is a plot of all 90 experimental HZACT values versus calculated values using the 72 compound equation from Table 2.

Table 2. Regression analysis using similarity measures and indicator variables

DEP VAR : HZACT      N : 72      MULTIPLE R : 0.849      SQUARED MULTIPLE R : 0.721  
 ADJUSTED SQUARED MULTIPLE R : 0.673      STANDARD ERROR OF ESTIMATE : 18.646

VARIABLE	COEFFICIENT	STD ERROR	T	P(2 TAIL)
CONSTANT	18.212	4.992	3.648	0.001
S_BZAPYR	22.967	6.365	3.609	0.001
S_DIBZP2	74.199	13.250	5.600	0.000
S_CHOLAN	37.829	14.824	2.556	0.013
S_TRMBA3	69.503	11.559	6.013	0.000
CH3ARO	12.635	2.936	4.304	0.000
NONPLANA	24.542	5.976	4.107	0.000
BAYC1	-59.806	12.003	-4.983	0.000
DELTA21C	-4.384	1.929	-2.273	0.027

ANALYSIS OF VARIANCE					
SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	56898.232	8	6598.349	17.047	0.000

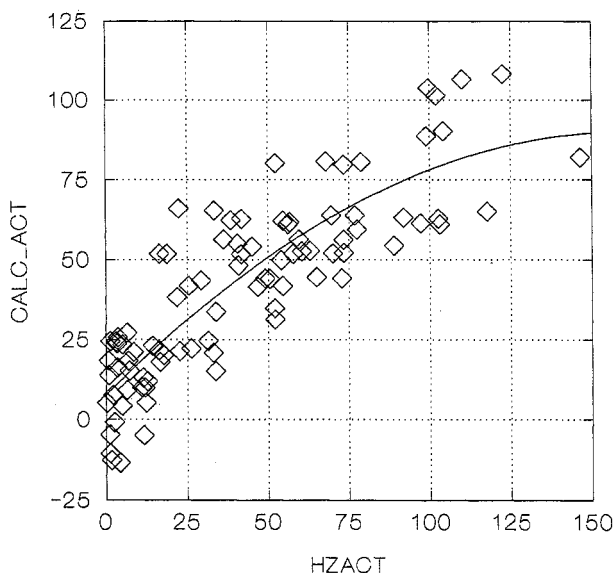


Figure 1. Experimental and calculated indices of carcinogenicity.

Five of the twelve most active compounds are calculated to be somewhat less active than the most active compounds of the development data set. The most active compound (Table 1, HZACT = 146.5) is predicted to have moderate/high activity (calculated HZACT = 82.0). One other high activity compound, #3 in Table 1, is also predicted poorly. The least squares quadratic line depicted in Figure 1 illustrates that the high activity compounds are generally predicted to have lower activities than observed. Six compounds with low activities, including two from the low activity validation set, have negative calculated activity indices, tantamount to prediction of inactivity. Mean deviations for the 72 compound correlation data set and the 18 compound predicted data set are 12.8 and 18.5 index units, respectively. One notes that predicted values are not as reliable as the correlated results.

The main goal of this work was to demonstrate that molecular similarity parameters, derived from a simple general similarity definition, could function as useful independent variables in QSAR studies. An ancillary, potentially useful finding is the modified cross-validation procedure employed in the analysis, which may be an effective tool for testing the predictive capabilities of QSAR model equations.

It should be mentioned that the results reported in this paper are to be regarded as preliminary. Both aspects of this investigation are undergoing further detailed study.

## ACKNOWLEDGMENTS

The financial support of the Welch Foundation of Houston, Texas, and the University of Texas Center for Environmental Resource Management (E. P. A. Superfund Research Program) is gratefully acknowledged.

## REFERENCES

1. M. Garbalena and W. C. Herndon, "Graph Theoretical Models for Enthalpic Properties of Alkanes." *J. Chem. Inf. Comp. Sci.*, **32**, 37-42 (1992).
2. W. C. Herndon and S. L. Knott, "Structure/Enthalpy Relationships for Hydrocarbons Containing Benzene Rings." *Polycycl. Arom. Compds.*, **11**, 229-236 (1996).
3. U. J. Urquidi, "Structure/Property and Structure/Activity Analyses of PCBs, PCDDs, and PCDFs." M. S. Thesis (Univ. of Texas at El Paso, Dec., 1994).
4. H.-T. Chen, "Structure/Activity Analyses of Antimalarial Compounds." M. S. Thesis (Univ. of Texas at El Paso, Dec., 1995).
5. Y. Zhang, "Studies of Aromatic Hydrocarbon Carcinogenicity." M. S. Thesis (Univ. of Texas at El Paso, Dec., 1996).
6. W. C. Herndon and S. H. Bertz, "Linear Notations and Molecular Graph Similarity." *J. Comp. Chem.*, **8**, 367-374 (1987).
7. A. J. Bruce, "Benzenoid Carcinogenicity and Abstract Definitions of Molecular Similarity." B. S. Honors Thesis (Univ. of Texas at El Paso, Aug., 1990).
8. G. Rum and W. C. Herndon, "Molecular Similarity Concepts 5. Analysis of Steroid-Protein Binding Constants." *J. Am. Chem. Soc.*, **113**, 9055-9060 (1991).
9. W. C. Herndon and G. Rum, "Three-Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids." in *QSAR and Molecular Modeling*, Prous Science Publishers, Madrid, 1996, pp. 380-384.
10. J. Iball, "The Relative Potency of Carcinogenic Compounds." *Am. J. Cancer*, **37**, 188-190 (1939).
11. W. C. Herndon, "On Enumeration and Classification of Condensed Polycyclic Benzenoid Aromatic Hydrocarbons." *J. Am. Chem. Soc.*, **112**, 4546-4547 (1990).
12. W. C. Herndon, "Quantum Theory of Aromatic Hydrocarbon Carcinogenesis". *Int. J. Quantum Chem.: Quantum Biology Symp.* No. 1, 123-134 (1974)
13. W. C. Herndon and L. V. Szentpaly, "Theoretical Model of Activation of Carcinogenic Polycyclic Benzenoid Aromatic Hydrocarbons. Possible New Classes of Carcinogenic Aromatic Hydrocarbons". *Journal of Molecular Structure (Theochem)*, **148**, 141-152 (1986).
14. P. P. Mager, "Biometrics in Medicinal Chemistry." in *QSAR in Design of Bioactive Compounds*, Prous Science Publishers, Madrid, 1984, pp. 433-442.
15. L. Eriksson, E. Johansson, and S. Wold "Quantitative Structure-Activity Relationship Model Validation." in *Quantitative Structure-Activity Relationships in Environmental Sciences-VII*, SETAC PRESS, 1997, Chpt. 26.

# COMPARATIVE MOLECULAR FIELD ANALYSIS OF AMINOPYRIDAZINE ACETYLCHOLINESTERASE INHIBITORS

Wolfgang Sippl,<sup>1</sup> Jean-Marie Contreras,<sup>2</sup> Yveline Rival<sup>2</sup> and Camille G. Wermuth<sup>2</sup>

<sup>1</sup> Institut für Pharmazeutische Chemie, Heinrich-Heine-Universität D-40225 Düsseldorf

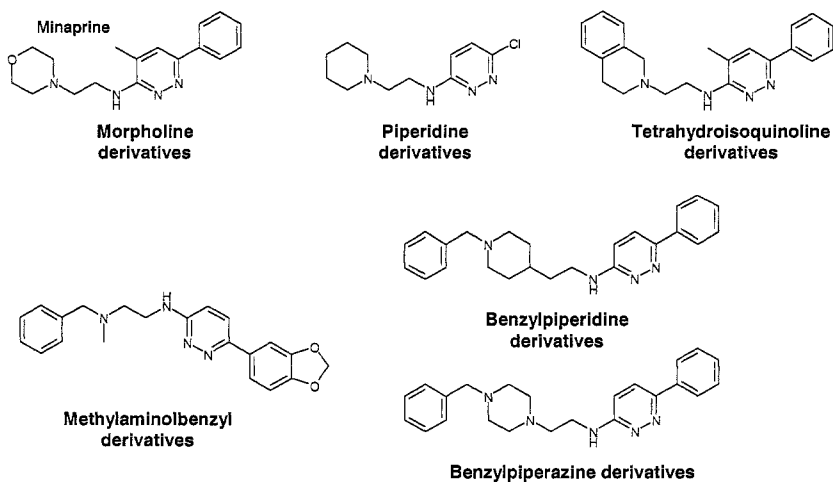
<sup>2</sup> Laboratoire de Pharmacochimie de la Communication Cellulaire Université Louis Pasteur Strasbourg, F-67401-Illkirch-Cedex

## INTRODUCTION

Modern methods for computer-assisted drug design fall into two major families - the indirect ligand-based methods, e.g. CoMFA or GOLPE and the direct receptor-based methods including molecular dynamics (MD) simulation, free energy perturbation (FEP) and the various docking procedures. Nowadays the ligand-based methods are widely used since they are computationally not demanding. The main problem of the ligand-based methods is the alignment of the investigated compounds. On the other hand the direct approach yields important information concerning the exact position of the ligands in the binding pocket. Since the MD and FEP methods are computationally intensive, they cannot be applied to large data sets. The faster docking programs on the other hand are at the moment not able to predict correctly the biological activity. One possibility to overcome these problems seems to be the combination of both approaches - merging the accuracy of the receptor-based strategies with the efficiency of modern 3D-QSAR techniques. This strategy has recently successfully applied by several groups<sup>1</sup>.

In the present study we report the application of such a combined approach to a series of aminopyridazine acetylcholinesterase (AChE) inhibitors<sup>2</sup>. AChE inhibitors are promising candidates for the treatment of Alzheimer's Disease, the fourth leading cause of death among the elderly in the industrial nations. Several AChE inhibitors are now undergoing clinical trials and recently, donepezil - a benzylpiperidine derivative - was introduced into therapy.

The starting point for the development of AChE inhibitors in our laboratory was the finding, that the antidepressant minaprine (figure 1) shows weak inhibition of AChE. Since minaprine has a unique structure among the known AChE inhibitors, it was taken as promising lead compound. The synthesized inhibitors can be classified into six different families, examples from each family are shown in figure 1.



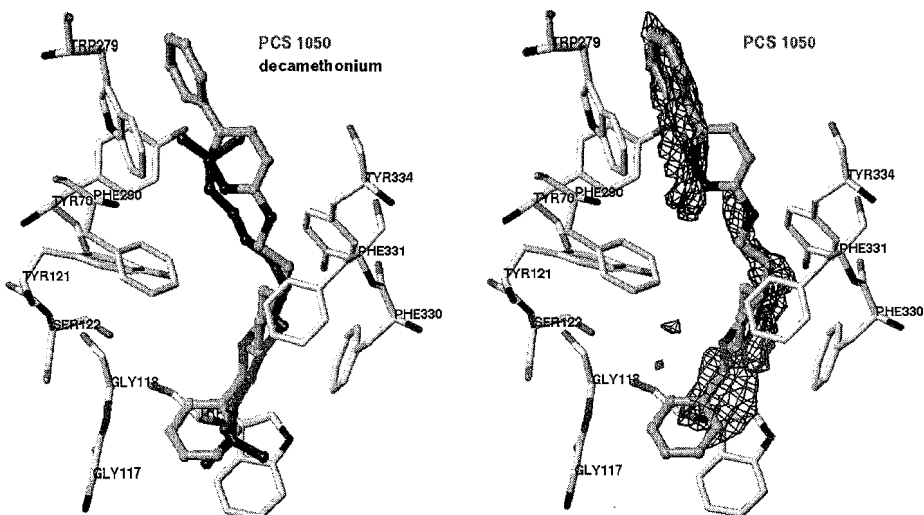
**Figure 1.** Examples of the investigated aminopyridazine AChE inhibitors.

## RESULTS AND DISCUSSION

The positioning of the molecules in a fixed lattice has been shown to be the most important input variable in comparative molecular field analysis. In order to obtain a realistic alignment of the investigated inhibitors we included the known crystal structures of AChE in our 3D-QSAR study. During the last few years four structures of AChE complexed with reversible inhibitors have been published (decamethonium, edrophonium, tacrine and huperzine). Unfortunately up to now no X-ray structure is available for AChE complexed with the potent benzylpiperidine inhibitors. Therefore we decided to use docking methods in order to determine the exact position of the inhibitors in the binding pocket.

The detailed inspection of the four AChE-inhibitor X-ray structures yielded crucial information concerning the orientation of the inhibitors in the binding pocket. AChE shows a nearly identical three-dimensional structure in all known X-ray structures. The active site is located 20Å 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 the phenyl ring 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 quite different. It seems improbable that a ligand-based method would be able to predict this alignment correctly.

In the next step we analyzed the binding pocket using the well-known program GRID. GRID generates a contour map of the interaction energy versus the three-dimensional position of the probe with respect to the crystal structure of the protein. This information can lead to the prediction of how various functional groups of the inhibitors will interact in a specific region within the active site. Several probes were used to analyze the active site of AChE. The results were compared with the positions of the co-crystallized inhibitors. We observed a nice agreement between the positions of the cationic head of the inhibitors and the contour maps obtained using the cationic trimethylammonium probe as well as between the location of the hydrophobic parts and the contour maps obtained using the hydrophobic DRY probe (figure 2). A detailed description of these results will be published elsewhere<sup>3</sup>. Encouraged by the good agreement between theoretically predicted and experimentally derived results we used the GRID contour maps as starting point for the docking of the aminopyridazine derivatives.



**Figure 2.** On the left side the predicted position of compound PCS1050 and the X-ray structure of decamethonium are shown. On the right side the favourable regions of interaction between the hydrophobic DRY probe and the active site are displayed for comparison (contour level -0.6 kcal/mol).

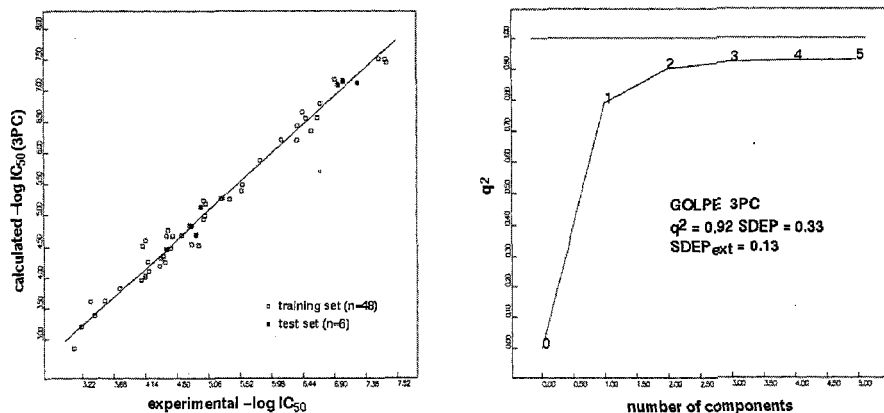
We started our docking analysis using compound PCS1050 - a potent and quite rigid inhibitor. First, a systematic conformational analysis was performed for this inhibitor. Second, program IXGROS, developed in our laboratory<sup>4</sup>, was applied in order to select the local minimum conformations. The resulting conformations were then docked individually into the binding pocket. The complexes were minimized keeping the protein atoms fixed and the complex with the most favourable energy was then taken as template for the docking of all the other inhibitors.

Figure 2 shows on the left side the predicted position of PCS1050 in comparison to the crystal structure of the complex with decamethonium. The hydrophobic parts of the inhibitors interact with an aromatic residue at the bottom of the gorge (Trp84), with three aromatic residues in the middle of the gorge (Phe330, Phe331 and Tyr334) and with two aromatic residues at the entrance of the gorge (Trp279 and Tyr70). No direct hydrogen bonds were observed for our inhibitors. It is possible that some water molecules bridge the distance between inhibitor and protein, as observed in the X-ray structure of AChE complexed with huperzine and tacrine. Electrostatic interaction appears mainly between the cationic head and Ser122 and Tyr121. The right side of figure 2 shows the agreement between the GRID contour maps derived from the hydrophobic DRY probe and the position of the inhibitors. Similar results were obtained for the other inhibitors under study.

The receptor-based alignment obtained by the docking procedure was further used as input for a comparative molecular field analysis. 48 aminopyridazine derivatives<sup>2</sup> were included in a GRID/GOLPE<sup>5</sup> analysis aimed to obtain information about the regions around the ligands which are important for the activity (see methods section for details). The model was validated using two randomly assigned groups of approximately the same size and then repeating the assignment 30 times. This cross-validation technique has been shown to yield better indices for the robustness and predictivity of a model than the normal leave one out method. To test the external predictivity we selected six newly synthesized inhibitors.

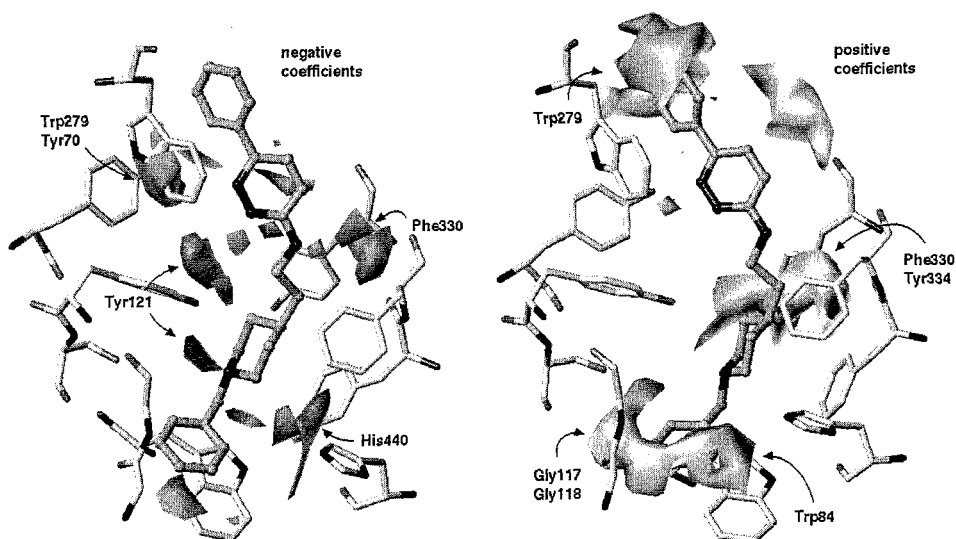
Figure 3 shows a plot of the experimental against calculated values and the values of the squared correlation coefficients ( $r^2$ ) and of the squared coefficients ( $q^2$ ) for different model dimensionalities. Three components were found to be significant ( $q^2 = 0.92$  and SDEP = 0.33). The derived model is highly predictive and robust also indicated by the prediction of the external test set (SDEP<sub>extern</sub> = 0.13).





**Figure 3.** GRID/GOLPE results for the manually derived alignment. Calculated vs experimental activity (left). Cross-validated squared correlation coefficients ( $q^2$ ) for different model dimensionalities (right).

Since the three-dimensional structure of our target is known, we were able to analyze the quality of the developed model by comparing the PLS coefficient maps of the inhibitors with the architecture of the active site. The regions which the model indicates as important for the activity should be close to the residues present in the binding pocket. Figure 4 shows on the left side the negative PLS coefficient maps and on the right side the positive PLS coefficient maps. Since we used the water probe the positive contour maps indicate the areas where polar interaction decrease activity and the negative contour maps show the regions where polar interaction increase activity. We observed a nice agreement between the maps and the positions of important amino acid residues in the active site. The three main positive fields are close to the important aromatic residues in the gorge. The negative maps are more widely distributed, but also for these maps a clear correlation was found between the location of the maps and the position of polar amino acid residues.



**Figure 4.** Comparison between the PLS coefficient maps and the location of important residues in the binding pocket (indicated by the arrows).

In the field of computer-aided drug design it is often recommended that a method can be applied to a large data set in a more or less unbiased automated way. Therefore, we started the development of a procedure able to automatically generate a 3D-QSAR model. The alignment of the compounds was performed using a combination of automated docking (AutoDock<sup>6</sup>) and geometry refinement (YETI force field<sup>7</sup>). Since most docking programs - including AutoDock - use simplified energy terms, the complex-ranking is not able to predict correctly the experimentally determined complex. Thus, a more sophisticated calculation method was chosen to refine the obtained protein-inhibitor complexes. We selected the YETI force field within PrGen since it has been shown to yield accurate results for protein-ligand complexes<sup>7</sup>. The complex possessing the most favourable interaction energy between protein and inhibitor was selected for the development of the inhibitor-alignment.

Before we applied the method to our aminopyridazine compounds the approach was validated using the X-ray structures of the four AChE-inhibitor complexes. Various AutoDock/YETI calculations have been performed using different docking and refinement conditions. An excellent agreement between the calculated complexes and the crystal structures was observed when we considered six structurally conserved water molecules during our docking studies. Not only are the rmsd between theoretically predicted and experimentally determined positions quite low (tacrine: 0.28Å; huperzine: 0.51Å; edrophonium: 0.71Å; decamethonium: 1.15Å), but also the positions found in the X-ray structure are in all cases those with the best interaction energy.

Encouraged by these results we applied the developed procedure to our data set of 48 aminopyridazine inhibitors. The automatically determined alignment is quite similar to the manually derived one, concerning the conformation of the inhibitors and the position of the cationic head. Differences occur in the relative alignment of the flexible inhibitors. A detailed analysis of the results is beyond the scope of this paper and an article devoted to this subject is in preparation<sup>3</sup>.

The automatically derived inhibitor-alignment was investigated using the already described GRID/GOLPE method. The resulting model shows a good correlation between experimental and predicted values. The  $q^2$  value - using the random group cross-validation is 0.86 and the SDEP is 0.45 using three components. Also the external predictivity is very good ( $SDEP_{\text{extern}} = 0.44$ ). Since the position of each inhibitor in the active site was calculated automatically the virtual testing of new compounds - not synthesized so far - seems to be a promising method for the design of new acetylcholinesterase inhibitors.

## COMPUTATIONAL METHODS

The crystal structure of minaprine retrieved from the Cambridge Structural Database was used as template to construct the inhibitors. All molecules were assumed to be mono-protonated under physiological condition and their molecular structures were generated accordingly using the SYBYL 6.3 software (Tripos Associates, St. Louis, USA).

To investigate the interaction potentials of the protein and inhibitor structures we performed a series of GRID (Molecular Discovery, Oxford, UK) calculations. The calculations were performed in order to search for binding sites complementary to the functional groups of the inhibitors.

The manual docking was performed using the SYBYL DOCK procedure taking into account the positions of the favourable GRID interaction fields in the binding pocket. No water molecules were considered during the manual docking. The resulting protein-inhibitor complexes were minimized keeping the protein atoms fixed.

The automated docking was performed applying the AutoDock program<sup>5</sup>. The obtained protein-inhibitor complexes were refined using the YETI force field within PrGen<sup>7</sup> (SIAT Biograph. Lab., Basel, Switzerland). The conformation of each inhibitor showing the most favourable interaction energy after the refinement was chosen for the inhibitor alignment.

The 3D-QSAR studies were carried out using the GOLPE4.0 program (Multivariate Infometric Analysis, Perugia, Italy). The 48 inhibitors of the training set were considered in the conformation found by the docking calculations. The biological activities (IC<sub>50</sub>) were determined using AChE from *Torpedo californica* and lie in the range between 850  $\mu$ M and 20 nM. They were transformed into -logIC<sub>50</sub> values. The energy calculations were performed with the GRID14 program, using the water probe. The size of the box was defined in such a way that it extends about 4Å from the structure of the inhibitors. A grid spacing of 1Å and an energy cut-off of +5 kcal/mol were used throughout the calculations. The advanced pretreatment method within GOLPE was applied to the X matrix in order to delete the non-informative variables. The X matrix was analyzed by PLS and variables were selected using the SRD/FFD method to improve the predictivity. Variables were grouped using 700 seeds, a cut-off distance of 1Å and a collapsing distance of 2Å.

## CONCLUSION

In this study the combination of ligand- and receptor-based methods has been successfully applied to a set of aminopyridazine derivatives with AChE inhibitor activities. We obtained highly predictive and robust models using a manually and an automated determined inhibitor-alignment. Besides the good predictivity, the models are also in close agreement with the known three-dimensional structure of the enzyme. The use of crystallographic data in the determination of the relative orientation of the studied inhibitors as an alignment tool is strongly supported by our results. The developed automated alignment-generation will be used in the future for the virtual testing of inhibitors not synthesized so far.

## Acknowledgments

We would like to thank Prof. H.-D. Höltje for providing computer facilities at the Heinrich-Heine-University Düsseldorf and Dr. G. Cruciani, University of Perugia for donating the GOLPE software.

## REFERENCES

1. C. Perez, M. Pastor, A.R. Ortiz and F. Gago, *Comparative binding energy analysis of HIV-1 protease inhibitors*, J. Med. Chem. 41, 836, (1998).
2. J.M. Contreras, Y. Rival, S. Chair and C.G. Wermuth, *Aminopyridazine bioisosteres of donepezil as acetylcholinesterase inhibitors*, J. Med. Chem., accepted (1998).
3. W. Sippl, J.M. Contreras, Y. Rival and C.G. Wermuth, *Comparative molecular field analysis of aminopyridazine acetylcholinesterase inhibitors*, in preparation.
4. H.-D. Höltje and G. Folkers. *Molecular Modelling*, VCH Publisher, Inc., New York (1996).
5. G. Cruciani and K.A. Watson, *Comparative molecular field analysis using GRID force field and GOLPE variable selection methods*, J. Med. Chem. 37, 2589 (1994).
6. D.S. Goodsell, G.M. Morris and A.J. Olson, *Automated docking of flexible ligands*, J. Mol. Recogn. 9, 1 (1996).
7. A. Vedani and D.W. Huhta, *A new force field for modeling metalloproteins*, J. Am. Chem. Soc. 112, 4759 (1990).

# THE INFLUENCE OF STRUCTURE REPRESENTATION ON QSAR MODELLING

Marjana Novič,<sup>1</sup> Matevž Pompe,<sup>2</sup> and Jure Zupan<sup>1</sup>

<sup>1</sup>National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

<sup>2</sup>Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia

## INTRODUCTION

In all kinds of QSAR studies it is very important how the chemical structure is represented. Usually a set of structural properties, calculated or extracted experimentally, is considered as a structure representation vector when compared and correlated to a biological property. Numerous attempts to suggest different structure representations reflect the vital importance of the structural coding problem in all kind of modelling procedures. Just a few examples are given for illustration in references<sup>1-7</sup>. One possible way of representing structures is by using a complete 3D structure information - atom type and coordinates. However, this representation suffers primarily from the lack of uniformity. Molecules containing different number of atoms  $N$  yield representations of matrices of various size ( $N \times 3$  or  $N \times 4$ ). Molecular descriptors originating from graph theory overcome the uniformity problem, they are also suitable because of their simplicity and often show good correlation with molecular properties<sup>8</sup> but the 3-D structural properties of compounds are lost. With the new "spectrum-like" structure code developed by Zupan et al.<sup>6,7</sup> the 3D representation is uniform, unique and reversible.

## METHODS AND DATA-SETS

### Molecular Descriptors

The methods for calculation of molecular descriptors will be briefly described. Descriptors used in the present study are all calculated either from the information about the connections between the atoms or from atomic 3D co-ordinates and information about atomic electronic properties. A set of  $m$  descriptors in a vector form  $X(x_1, \dots, x_m)$  is further on referred to as a *structure representation*.

**Topological descriptors** are derived from the topological characteristics of molecular graphs and describe the atomic connectivity in the molecule. All distances between arbitrary pairs of points in the graph are graph invariants independent of the numbering and links. One of the graph's invariants, characterizing many topological descriptors, is the *order* of each point in the graph equal to the number of links leaving the point, i.e., expressing how many neighbours are linked to the point. Topological descriptors, used here as components of structure representation vectors for the purpose of QSAR modelling, reflect specific structural features like size, shape, symmetry, branching, and cyclicity of the compounds they represent. Only a few most frequently used indices are listed here. The *Wiener index* is expressed in terms of the distance matrix and equates to the half-sum of all distance matrix entries. *Randic and Kier&Hall indices* (order 0-3) are calculated from co-ordination numbers of or from values of atomic connectivity. *Kier shape index* (order 1-3) depends on the number of skeletal atoms, the molecular branching and the ratio of the atomic radius and the radius of carbon atom in the  $sp^3$  hybridisation state. The *Kier flexibility index* is derived from *Kier shape index*. The *Balaban index* is defined by the number of edges in the molecular graph, by number of vertices, cyclometric number, and by distance degrees obtained by summation of *i*-th row and *i*-th column of the distance matrix. The *information content index and its derivatives* (order 0-2) are based on Shannon information theory. Modifications of *information contents index* are: structural information content, complementary information content and bond information content. All mentioned indices used in this study were calculated by CODESSA software<sup>9</sup> (for detailed description of indices see references in the CODESSA documentation).

**Geometric descriptors** are one of the possible structure representations that are also tested in the present study. These descriptors require 3D-atomic co-ordinates. Different values contributing to the set of geometric descriptors are calculated from atomic co-ordinates: *moments of inertia, shadow indices, molecular volume, molecular surface area, and gravitation indices*<sup>9</sup>.

**Electrostatic descriptors** in our investigation of QSAR models are added to the set of geometric descriptors. They reflect characteristics of the charge distribution of the molecule. The empirical partial charges are calculated by a method proposed by Zefirov<sup>9</sup>. Using partial charges, the following electrostatic descriptors are calculated: *minimum and maximum partial charges in the molecule, minimum and maximum partial charges of particular types of atoms, and polarity parameter*.

**3D descriptors for Spectrum-like representation** of molecular structure, defined by 3D-co-ordinates of its atoms, are obtained by a projection of all atomic centres of a molecule onto a sphere of arbitrary radius. An oriented structure is placed into an arbitrary large sphere. The projection beam from the central point of sphere causes a pattern of points on the sphere, where each point represents a particular atom. Then each point on the sphere is taken as the centre of a "bell-shaped" function with intensity related to the distance between the co-ordinate origin and a particular atom. As "bell-shaped" function of atom *i* we have taken Lorentzian curve with the form:

$$s_i(\varphi_j, \vartheta_l) = \frac{\rho_i}{(\varphi_j - \varphi_i)^2 + \sigma_i^2} + \frac{\rho_i}{(\vartheta_l - \vartheta_i)^2 + \sigma_i^2} ; \quad /1/$$

$$\varphi_j = \frac{2\pi}{k}, \dots, \frac{2\pi j}{k}, \dots, 2\pi ; \quad j = 1, k$$

$$\vartheta_l = \frac{\pi}{k/2}, \dots, \frac{\pi l}{k/2}, \dots, \pi ; \quad l = 1, k/2$$

where  $s_i(\varphi_j, \vartheta_i)$  is "spectrum intensity" related to atom  $i$ , while the parameters are:

- $\rho_i$  - distance between the center of the sphere and atom  $i$ ,
- $\varphi_i, \vartheta_i$  - polar and azimuthal angle of atom  $i$ ,
- $\sigma_i$  - atomic charge (extended by 1) on atom  $i$ ,
- $k$  - resolution of the representation (steps for indices  $j$  and  $l$ ).

The total intensity related to the entire molecule is then the sum of intensities belonging to individual atoms:

$$S(\varphi_j, \vartheta_l) = \sum_{i=1}^{n^{atom}} s_i(\varphi_j, \vartheta_l) \quad /2/$$

In practice the projections on three perpendicular equatorial trajectories rather than the projection on the entire sphere have been considered. In the case that the largest part of the skeletons of molecules in the study are planar only the projection on one trajectory (x-z plane) is taken into account. If Mulliken charges on atoms  $i$  are incorporated as  $\sigma_i + 1$  in equation /1/ the reversibility is not lost, however, recovering of atom positions from the code is more computer intensive.

## Modelling

**Multiple Linear Regression (MLR)** technique is successful in applications with linear relationship between the descriptors and the sought property. It is also effectively applicable for non-linear relations, if it is known which factors should be non-linear. The essence of MLR is to determine the coefficients at each factor to obtain the best overall relation of the real property and the property predicted by the linear equation (model). For the solution, one needs at least as many equations (objects with known properties) as there are factors, i.e. descriptors in each equation. In order to validate the obtained model with statistical parameters, more objects than factors must be available. In other words, the system has to be over-determined in order to be able to compare the errors due to the lack of fit (model errors) and experimental errors<sup>10</sup>.

**Counterpropagation Artificial Neural Network (CP ANN)**<sup>11</sup> modelling is based on a supervised learning method, although one part of the learning process involves elements of unsupervised learning. This means that for the learning procedure a set of input-target pairs  $\{X_s, T_s\}$  is required. In the case of the structure-property correlation problem the input  $X_s = (x_{s1}, x_{s2}, \dots, x_{sj}, \dots, x_{sm})$  is a structure representation of the  $s$ -th compound represented by  $m$  structural features or "variables". The corresponding target  $T_s = (t_{s1})$  is a one-component vector indicating the studied property of  $s$ -th compound. After the learning procedure, the ANN responds for each input structure representation  $X_s$  from the training set with the output  $Out_s$  identical to the target  $T_s$ .

## Data

Two data-sets are used in the study. The first one is a small set of 28 flavonoid derivatives<sup>12</sup>, inhibitors of the enzyme p56<sup>lck</sup>. The other data-set is a large collection containing 256 structurally diverse derivatives of 5-phenyl-3,4-diamino-6,6-dimethyldihydrotriazine inhibiting dihydrofolate reductase<sup>13,14</sup>.

## RESULTS

Chemical structures in both data-sets were initially represented by 3D coordinates of all atoms in the molecules determined for their minimal energy state, and with the connection tables describing all connectivities between the atoms in each molecule. In order to obtain uniform, equally dimensional structure representation vectors for the modelling purpose the initial representations were transformed in four different ways producing sets of:

- topological indices
- geometric and electronic indices
- *spectrum-like* code intensities
- *spectrum-like* code intensities modified by Mulliken charges

The two former representations are calculated by CODESSA software<sup>9</sup>, while the two latter ones are structural descriptors developed in the authors' laboratory.

*Topological code* of structure representation contains 38 descriptors, *geometric + electrostatic code* contains 87 descriptors, while both variations of *spectrum-like* representation of 28 flavonoid derivatives consist of 120 descriptors calculated with equation /1/ for the XY projection of molecular coordinates. The *spectrum-like* representation of the compounds from the second data-set consists of 180 descriptors, half of them are calculated for the XY and half for the XZ projection of molecular coordinates. With each of these four representations two modelling strategies were applied, i.e. multiple linear regression (MLR) and CP-ANN with Kohonen mapping strategy.

MLR is performed using the same software (CODESSA) as for calculation of topological, geometric and electronic structure descriptors. For each type of structure representation the procedure called *heuristic optimization* is applied to determine the descriptors giving the best correlation of modelled properties with the experimental ones. MLR modelling results for the set of 28 flavonoid derivatives are shown in Table 1.

Table 1. Prediction results of the best MLR models obtained for four different structure representations of the set of 28 flavonoid derivatives

Structure representation	MLR $r^2$	MLR $r^2$ (CV*)	MLR $s^2$	MLR F
Topological indices	0.77 <sup>a</sup> 0.90 <sup>b</sup>	0.55 <sup>a</sup> 0.68 <sup>b</sup>	0.107 <sup>a</sup> 0.061 <sup>b</sup>	14.5 <sup>a</sup> 15.0 <sup>b</sup>
Geometric + electrostatic indices	0.82 <sup>a</sup> 0.91 <sup>b</sup>	0.69 <sup>a</sup> 0.78 <sup>b</sup>	0.085 <sup>a</sup> 0.052 <sup>b</sup>	19.6 <sup>a</sup> 17.7 <sup>b</sup>
Spectrum-like structure representation	0.82 <sup>a</sup> 0.96 <sup>b</sup>	0.71 <sup>a</sup> 0.84 <sup>b</sup>	0.084 <sup>a</sup> 0.022 <sup>b</sup>	19.7 <sup>a</sup> 45.4 <sup>b</sup>
Spectrum-like structure representation + Mul. charges	0.94 <sup>a</sup> 0.98 <sup>b</sup>	0.88 <sup>a</sup> 0.95 <sup>b</sup>	0.027 <sup>a</sup> 0.011 <sup>b</sup>	70.0 <sup>a</sup> 90.6 <sup>b</sup>

\*Cross-validation using "leave-one-out" procedure

<sup>a</sup> five-factor MLR model

<sup>b</sup> ten-factor MLR model

Comparison of the results from Table 1 shows that the use of *spectrum-like* structure representation enables better correlation between chemical structure and biological property of the studied compounds than the use of topological and geometrical

descriptors. It is also seen that electronic descriptors improve modelling results. Additional useful information results from the choice of the reduced sets of descriptors of the structure-representation vectors. It is interesting to see which are the chosen five or ten descriptors in each model, especially in the case of *spectrum-like* representation vectors. By checking only the descriptors of the best model, i.e. ten-factor MLR model using *spectrum-like* structure representation modified by Mulliken charges, we can see that those parameters indeed describe the directions in the flavonoid molecule where 3' or 4' and 5, 6, 7, 8 substitutions<sup>12</sup> are located. In Figure 1 it is indicated in which directions the ten descriptors are chosen by the MLR procedure for reduction of representation parameters.

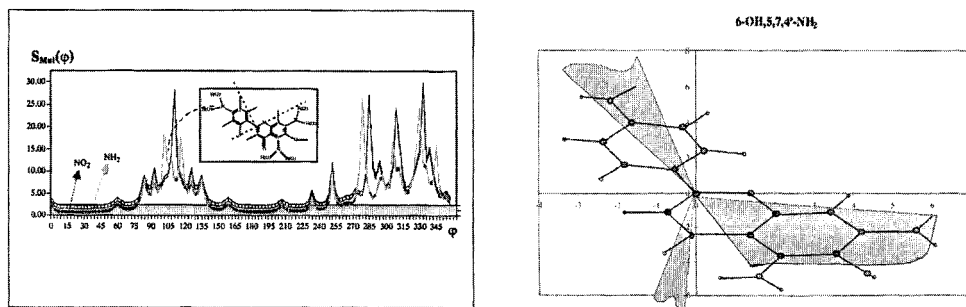


Figure 1. Two *spectrum-like* structure representations of flavonoid derivatives (6-OH,5,7,4'-NH<sub>2</sub> and 6-OH,5,7,4'-NO<sub>2</sub>) (left) and XY projection (right) of one of them. The shadowed areas correspond to the directions covered by 10 most representative descriptors chosen in optimization procedure for reduction of parameters in MLR modelling.

The next modelling approach applied in the present research is CP ANN. Only two of the four types of structure representations previously studied, i.e. *spectrum-like* structure representation and *spectrum-like* structure representation modified by Mulliken charges, were analysed. In order to compare the ANN results with those obtained by MLR models, the reduced sets of the same five and ten descriptors as determined in MLR study were used as structure representations. The parameters used for training the CP ANN were: learning rate  $a_{\max}=0.4$   $a_{\min}=0.05$ , 80 epochs, nontoroidal condition<sup>15</sup>. As it was expected, higher correlation coefficients were obtained for predictions of training samples, i.e., all 28 compounds from the data set. When the leave-one-out cross-validation procedure was performed, each compound was once excluded from the training set and the biological activity of this compound was then predicted on the basis of the model obtained with the rest ( $n-1$ ) of the compounds. For evaluation of the models the correlations obtained by cross-validation are more relevant and reflect the possibility of generalization of the proposed models, at least in the sense of variations of substituents in the group of compounds with the same skeleton.

The best model was obtained using ten-descriptors structure representation vector of *spectrum-like* structure representation modified by Mulliken charges. It has to be stressed that the selection of the descriptors for the reduced sets was not repeated in the ANN modelling approach. It was taken directly from the MLR optimization procedure. Correlation coefficient ( $r$ ) between the experimental and predicted biological activity with leave-one-out test is 0.92, while direct predictions from the model (retrieved values) are 100% correct, which means that the model recognises without an error the properties of all objects from the training set.



The modelling results in the case of dihydrofolate reductase (DHFR) inhibitors reveal quite a different situation. First, the data-set is very diverse and therefore it is more difficult to obtain one general model. The best correlation coefficients obtained with an optimised set of topological, geometrical, electrostatic and quantum-chemical indices was 0.84 for 30-factor MLR model and cross-validated correlation coefficient was 0.78. In the case of "spectrum-like" structure representation the correlation coefficient was 0.66 (0.56 in leave-one-out cross-validation). Even lower correlation between predicted and experimental activities was obtained with artificial neural network models. Correlation coefficients obtained by ten-fold cross-validation were 0.56 for "spectrum-like" representation and 0.65 for representation with structural indices. But the networks were trained with the optimised sets of 30 parameters determined in MLR procedure, which could be the source of the worse performance of ANN models. We expect better predictions from ANN models if the selection of parameters is made using ANN. The optimisation of structure representation parameter set using genetic algorithm is now in progress in our laboratory.

### Acknowledgment

The financial support of the Ministry of Science and Technology of Slovenia obtained by the Projects: J1 - 8900 and J1 - 0291 is gratefully acknowledged.

### REFERENCES

1. R. Todeschini, P. Gramatica: 3D Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM, *Quant. Struct.-Act. Relat.*, **16**, 113-119, (1997).
2. J.T. Clerc, A.L. Terkovich, Versatile topological structure descriptor for quantitative structure/property studies, *Anal. Chim. Acta*, **235**, 93-102, (1990).
3. J.H. Schuur, P. Selzer, J. Gasteiger, The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity, *J. Chem. Inf. Comput. Sci.*, **36**, 334-344, (1996).
4. S. Bauerschmidt, J. Gasteiger, Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species, *J. Chem. Inf. Comput. Sci.*, **37**, 705-714, (1997).
5. Y. Tominaga, I. Fujivara, Novel 3D Descriptors Using Excluded Volume: Application to 3D Quantitative Structure-Activity Relationships, *J. Chem. Inf. Comput. Sci.*, **37**, 1158-1161, (1997).
6. M. Novič, J. Zupan, A New General and Uniform Structure Representation, *Software-Entwicklung in der Chemie 10*, Johann Gasteiger (Ed.), Frankfurt am Main, pg. 47-58, (1996).
7. J. Zupan, M. Novič, General Type of a Uniform and Reversible Representation of Chemical Structures, *Anal. Chim. Acta*, **348**, 409-418, (1997).
8. M. Randić, M. Razinger, On characterization of 3D molecular structure, in: *From Chemical Topology to Three-Dimensional Geometry* (A. T. Balaban, Ed.), Plenum Press, New York, (1997).
9. A. R. Katritzky, V. S. Lobanov, M. Karelson, CODESSA 2.0, Comprehensive Descriptors for Structural and Statistical Analysis, Copyright (c) 1994-1996 University of Florida, U.S.A.
10. D.L. Massart, B.G. M. Vandengiste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a textbook*, Elsevier, Amsterdam, (1988).
11. R. Hecht-Nielsen, Counterpropagation Networks, *Appl. Optics*, **26**, 4979-4984, (1987).
12. M. Cushman, H. Zhu, L.R. Geahlen, J.A. Kraker, Synthesis and Biochemical Evaluation of a Series of Aminoflavones as Potential Inhibitors of Protein-Tyrosine Kinases p56, EGFr, p60. *J. Med. Chem.*, **37**, 3353-3362, (1994).
13. C. Silipo, C. Hansch, Correlation Analysis. Its Application to the Structure-Activity Relationship of Triazines Inhibiting Dihydrofolate Reductase, *J. Am. Chem. Soc.*, **97**, 6849, (1975).
14. F.R. Burden, B.S. Rosewarne, D.A. Winkler, Predicting Maximum Bioactivity by Effective Inversion of Neural Networks Using Genetic Algorithms, *Chemometrics and Intelligent Laboratory Systems*, **38**, 127-137, (1997).
15. J. Zupan, M. Novič, I. Ruisánchez: Kohonen and Counterpropagation Artificial Neural Networks in Analytical Chemistry, *Chem. Intell. Lab. System*, **38**, 1-23, (1997).

## THE CONSTRAINED PRINCIPAL PROPERTY (CPP) SPACE IN QSAR – DIRECTIONAL AND NON-DIRECTIONAL MODELLING APPROACHES

Lennart Eriksson,<sup>1</sup> Patrik Andersson,<sup>2</sup> Erik Johansson,<sup>1</sup> Mats Tysklind,<sup>2</sup>  
Maria Sandberg,<sup>1</sup> and Svante Wold<sup>3</sup>

<sup>1</sup>Umetri AB, POB 7960, 907 19 Umeå, Sweden, www.umetri.se

<sup>2</sup>Dept. Env. Chemistry, Umeå University, 901 87 Umeå, Sweden

<sup>3</sup>Institute of Chemistry, Umeå University, 901 87 Umeå, Sweden

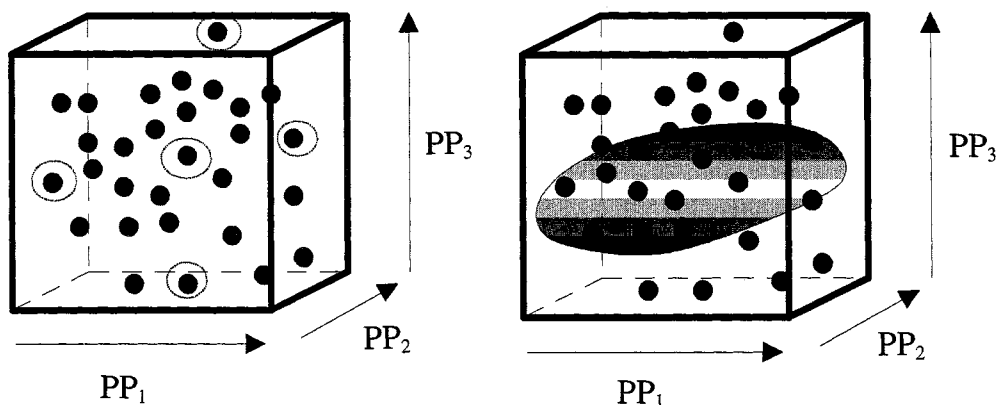
### INTRODUCTION

Multivariate design is useful for selecting informative training- and validation sets.<sup>1</sup> The essence of this approach is (i) to describe the compounds with many descriptors, (ii) to summarize these descriptors by means of principal component analysis<sup>2</sup> (PCA), and (iii) to create an informative multivariate design in the established PC-scores (“principal properties”, “PPs”). This approach has been used in many areas for selecting representative compounds, e.g., organic chemistry,<sup>3</sup> crystallization modelling,<sup>4</sup> environmental chemistry<sup>5</sup> and QSAR,<sup>6</sup> combinatorial chemistry,<sup>1</sup> and biopolymer sequence modelling.<sup>7</sup>

It is our aim to describe a limitation of the multivariate design approach in QSAR. This limitation arises when working with a biological response of a specific mechanism, which is elicited by a limited number of compounds distributed within a larger set of chemicals. In such a case, it is conceivable that the few biologically active compounds, with a specific combination of PPs, are grouped tightly together in the PP-space of the entire chemical class. This kind of constrained principal property (CPP) space is illustrated in Figure 1. Clearly, here only a limited portion of the PP-space is of relevance for QSAR, and it is not justifiable to select a training set covering the whole PP-space. Rather, it appears fruitful to select a training set located within the CPP-space. We shall discuss two procedures for doing this, which we call *directional* and *non-directional* modelling.

### ILLUSTRATION

Our illustration to the CPP-problem deals with poly-chlorinated biphenyls (PCBs). PCBs are widespread in the environment and a number of toxic and biochemical responses have been identified. Recently, the entire series of 209 PCBs was multivariately characterized by 52 chemical descriptors.<sup>8-10</sup> By means of PCA, this battery of descriptors was subsequently converted to a four-dimensional PP-space. The relevance of selecting representative PCBs based on this parametrization has been proven repeatedly.



**Figure 1.** Schematic illustration of a principal property (PP) space defined by three principal properties. (left) A multivariate design, symbolized by the encircled compounds, laid out in the entire PP-space. (right) A constrained region of a PP-space, which is poorly mapped by a multivariate design of the foregoing type. A design adapted to the constrained portion of the PP-space better applies.

In a recent article by Connor *et al.*, the CYP2B activity of 18 tri- to octachlorinated PCBs in female rat, was published.<sup>11</sup> Interestingly, these 18 biologically tested congeners exhibit multiple-ortho substitution and are located in a constrained part of the PCB PP-space (see below). This means that these 18 compounds share a specific combination of principal properties, a fact indicating the structural specificity of the biological response. We call this part of the PCB PP-space the “CYP2B-region”.

It is of interest to further explore the shape of the CYP2B-region and its distribution of compounds. We will do so by using multivariate analysis, and our goal is to understand (i) whether the 18 tested congeners are good representatives of the region, or (ii) whether they need to be supplemented with other PCBs to result in a better mapping of this region.

## MODELLING APPROACHES AND DATA ANALYTICAL METHODS

The first analysis approach, *non-directional modelling*, is based on using the *chemical* data of the 18 tested PCBs. PCA of this data set is used for defining local PPs. The remaining 191 PCBs are then fitted to this local model and classified as members or non-members. Those compounds which are classified as model (“class”) members have a combination of PPs resembling the 18 tested PCBs. Hence, they may be used to propose a suitable mapping set of the CYP2B-region. With the term *mapping set* we mean a series of compounds which can be used to explore the size and shape of the CYP2B-region.

The proposition of a mapping set corresponds to laying out a D-optimal design in the series of compounds fitting the local PCA model. This approach is *non-directional* in the sense that it allows the CYP2B-region to be explored in all directions for finding appropriate mapping set congeners. The reason for this non-directionality is that only chemical information of the PCBs are used in the modelling.

We consider the non-directional approach to be useful when the goal is to find more potent compounds. Ideally, one would like to identify potent chemicals being as diverse as possible, because this would allow the discovery of local sub-optima in biological potency. This approach is also of relevance if the goal is to guard for possible “new” or “unwanted” responses or side effects.

The second analysis approach, *directional modelling*, is also based on using the 18 tested compounds for training of a local model. However, in contrast to the foregoing approach, *chemical and biological* data are now used simultaneously. Thus, partial least squares<sup>12</sup> (PLS) regression is used for deriving a QSAR, accompanied by biological activity predictions for the 191 non-tested substances. Among the compounds which fit the QSAR, it is then possible to select appropriate compounds for a mapping set of the CYP2B-region. We realize that this approach is of a *directional* nature, because it allows the CPP-space to be investigated in a direction possibly encoding more potent CYP2B-inducers. With the directional approach, finding more potent compounds is the major goal. This strategy also works with several biological response variables.

In order to accomplish these two mapping approaches, we use the PCA<sup>3</sup> and PLS<sup>12</sup> methods, as implemented in SIMCA.<sup>13</sup> In order to propose representative compounds for mapping of the CYP2B-region, we use D-optimal design<sup>14</sup>, as implemented in MODDE.<sup>15</sup>

## RESULTS

Initially, a reference PCA of the entire 209\*52 (compounds\*chemical descriptors) data matrix gave a four-component model with  $R^2 = 0.78$  (explained variation) and  $Q^2 = 0.70$  (predicted variation according to cross-validation<sup>12,13</sup>). The first two components are dominant and account for 65% of the explained variation. A score plot of these is provided in Figure 2a. In this plot, the 18 tested congeners are highlighted with large triangles.

The framing of the 18 tested PCBs indicate the extent of the CYP2B-region. We can see that this region is embedded in the larger set of PCBs, and the question which arises is *where lie the pertinent borders of the CYP2B-region?* We wish to map this region according to the directional and non-directional modelling approaches, and produce an appropriate mapping set. The results of the two modelling approaches will be given below, and be graphically rendered in the PP-space of the 209 compounds.

The *non-directional approach* was commenced by computing a local PCA of the training set, that is, the 18 tested compounds. To make this mapping approach flexible three stoppage criteria were used, namely (i) retention of principal components (PCs) with an eigenvalue larger than 2, (ii) retention of PCs with an eigenvalue larger than 1, and (iii) cross-validation. As seen in Table 1, this leads to the use of 2, 4 and 7 principal components. Subsequently, all compounds, that is, the 18 in the training set and the 191 in the prediction set, were fitted to the PCA models of varying complexity.

**Table 1:** Summary of the non-directional mapping

Stop.Crit.	# Comp	Expl. Var.		Classification		D-optimal design			Geff	Cond No	Figure
		R2	Q2	Train.	Pred.	Model	#Cong	#New			
i, EIG = 2	2	0.68	0.44	18 (18)	79 (191)	Quadratic	8	5	80.1	5.5	2b
ii, EIG = 1	4	0.82	0.54	17 (18)	68 (191)	Linear	12	6	76.2	2.4	2c
ii, EIG = 1	4	0.82	0.54	17 (18)	68 (191)	Interaction	18	11	71	7.9	2d
iii, CV	7	0.93	0.62	17 (18)	45 (191)	Linear	13	2	75.3	2.9	2e

Stop. Crit. = stopping criterion used in the PCA modelling. # Comp = number of components in PCA model. R2 = explained variation.

Q2 = predicted variation. Train. = number of compounds of PCA training set fitting to the model

Pred. = number of compounds of PCA prediction set fitting to the model. Model = selected model for D-optimal design.

# Cong = number of PCB congeners selected by D-optimal design. # New = number of non-tested compounds among #Cong.

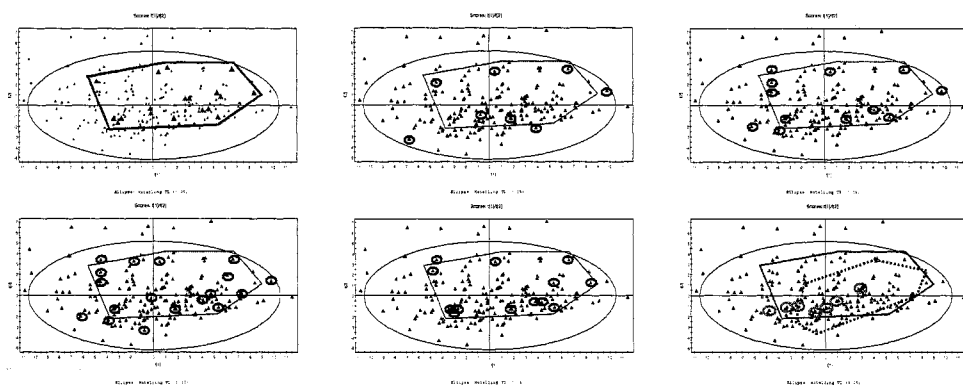
Geff = G-efficiency of D-optimal design. CondNo = condition number of D-optimal design. Figure = figure used in paper.

In the next step, D-optimal designs were laid out using as candidate sets all compounds fitting to the various PCA models. Four D-optimal designs were constructed, one supporting a quadratic model in two PCs, one a linear model in four PCs, one an

interaction model in four PCs, and one a linear model in seven PCs. These are summarized in Table 1 and the distribution of selected compounds plotted in Figures 2b-2e.

Subsequently, the *directional mapping* was started by calculating a PLS model based on the 18 tested compounds. This model contained four components and gave  $R^2 = 0.97$  and  $Q^2 = 0.59$ . In the next step, predictions of biological activity of the 18 training set and 191 prediction set compounds were conducted. We note that one compound in the training set, #163, is extreme in biological activity (BA). Its existence may, partly, shed some explanatory light on the moderate  $Q^2$ . The cross-validation procedure is unable to predict the behavior of #163, when omitted from model computation.

The obtained predictions can be used for a directional mapping of the CYP2B-region, which is summarized in Figure 2f. Again, the solid frame shows the distribution of the 18 tested PCBs. Seventeen of these compounds have a BA ranging from 4-102, and the BA for the extreme #163 is 195. Within the dotted area, prediction set compounds are found which (a) fit the model well and (ii) have a predicted BA above 105 and below 304. Predictions made inside the dotted area correspond to model interpolations. In addition, we have the seven encircled compounds, which did not fit the QSAR model. These are predicted to have a BA of 500+, and are thus substantially more potent than any of the actually tested compounds. The latter predictions correspond to model extrapolations.



**Figure 2.** Overview of results of non-directional and directional mapping. (a, upper left) Score plot of the reference PCA model. Large triangles denote the 18 tested congeners. (b, upper middle) Distribution of mapping set of D-optimal design supporting a quadratic model in two local PPs of the 18 tested compounds. (c, upper right) Same as (b), but with a linear model in four PPs. (d, lower left) Same as (b), but with an interaction model in four PPs. (e, lower middle) Same as (b), but with a linear model in seven PPs. (f, lower right) PLS modelling results. Solid frame demarcates distribution of the 18 tested compounds. Dotted frame indicates distribution of compounds fitting the PLS model, predicted to be more potent than the tested congeners. Seven encircled compounds, not fitting the model, are predicted to have BA >500.

## DISCUSSION

One interesting question in multivariate QSAR is how to formulate appropriate training- and validation sets. With a non-specific response, and with weak or no clustering of the compounds in a PP-space, it is often sufficient to lay out one single multivariate design. With a selective and specific endpoint, however, which usually correlate with a well-defined combination of PPs, the classical multivariate design approach ought to be modified. The reason for this is that such a response usually is elicited by a smaller set of compounds, which are grouped tightly within a larger PP-space. Hence, it is uninteresting to create a multivariate design in the entire PP-space, and a multivariate design adapted to the smaller, constrained part of the PP-space appears more appropriate.

We have here used a series of 18 PCBs to exemplify how biological performance may act as a constraining factor. If, in QSAR, these 18 PCBs were to be used for model training, one question of relevance would be to know their representativity of the CYP2B-region. There are different ways to probe the representativity of the 18 tested PCBs, and in this paper we have outlined two mapping approaches.

Initially, a reference PCA on the whole data set was calculated, and the distribution of PCBs in the first two dimensions are portrayed in Figure 2a. The solid frame indicates the size and extent of the CYP2B-region, and the solid triangles represent the tested PCBs. Evidently, the tested compounds display an unbalanced distribution. Hence, it may be anticipated that they are not optimally representative of the constrained region.

The non-directional mapping was based on PCA modelling of the 18 tested compounds. By means of three stopping rules, three alternative models of varying complexity were derived. One model had two components, one four, and one seven. In the next step, the remaining 191 congeners were used as a prediction set and were fitted to the three PCA models, the results of which are summarized in Table 1. We can see that in the case of seven components as few as 45 substances of the prediction set fit the model, whereas with two- and four-component models, 79 and 68 compounds fit, respectively.

The obtained classification results indicate that with seven components the model fits the CYP2B-inducers "tightly" compared to the other cases. Accordingly, only the prediction set compounds which show the highest degree of chemical similarity with the tested PCBs, are classified as class members. As a consequence, the D-optimal design laid out in this case, allows for the most conservative mapping set (cf. Figure 2e). In principle, the shape of the CYP2B-region is not explored outside the framed area. This is because only two of the 13 identified compounds are biologically untested.

Interestingly, it is possible to *decrease* the extent of chemical similarity and *increase* the extent of chemical diversity among the prediction set compounds which fit the various PCA models. This is accomplished by regulating the number of used principal components (PPs). Table 1 shows that when utilizing only two components, as many as 79 prediction set congeners are categorized as class members, and hence as potential CYP2B-inducers. The created D-optimal design encodes the most optimistic mapping set (cf. Figure 2b). Here, five out of eight chosen compounds are biologically untested, and we can see that these allow for an exploration of the CPP-space well outside the framed area.

Somewhere in between the two extremes portrayed in Figures 2b and 2e, we have the situations rendered in Figures 2c and 2d. The latter cases represent coverages of the CYP2B-region achieved by D-optimal designs in four PPs. Apparently, mapping sets are now proposed which allow for *some* extrapolation outside the framed area, but not as pronounced as in Figure 2b. Further, by tailoring the four factor D-optimal design towards a linear model (Figure 2c) or an interaction model (Figure 2d), it is possible to influence the investigation of the inner part of the CYP2B-region. A linear model in four factors seems more adequate than an interaction model, as the former gives a smaller mapping set.

Figures 2b-2e summarize the non-directional mapping. It is clear that this approach permits the mapping of the CYP2B-region to expand in all directions of the PP-space. In contrast to this, we have the directional mapping procedure founded on PLS regression. Here, use is made of the y-data, as a pointer for finding the combination of chemical properties predicted to represent the most potent compounds.

The PLS model was trained on the 18 tested compounds. In the ensuing step, the 191 prediction set congeners were fitted to the model and their CYP2B-induction potency predicted. Figure 2f represents a summary of the acquired results. The solid frame shows the portion of the PP-space in which the biologically tested compounds are found. With the exception of PCB#163, these have a biological activity (BA) range of 4-102. Congener #163 has a BA of 195. The dotted frame indicates another region in which PCBs predicted

to be generally more potent lie, and these have BA in the interval 105-304. Observe that these predictions correspond to model interpolation.

Furthermore, it is possible to consider predictions corresponding to model extrapolation. Such predictions are more uncertain than model interpolation forecasts, but may still be useful for identifying potent chemical structures. The seven PCBs encircled in Figure 2f are predicted to have a BA of 500+. These compounds occupy a small and narrow area, almost a curved line, in the PP-space, which strongly indicates that a specific combination of PCB PPs correlates with the investigated BA.

It is of interest to conduct a chemical interpretation of the acquired PLS model. An inspection of the PLS model coefficients (no plot provided) indicates that molecular polarization is one key element towards more potent compounds, because descriptors reflecting polarizability dominate the model. This interpretation is also supported by the distribution of PCBs in Figure 2f. The compounds lying within the dotted frame, that is, compounds predicted to be more active than the tested ones, are moderately polarized. Many of these congeners display di-ortho 2,6-substitution and have chlorine substituents on both rings. Furthermore, the seven encircled compounds, predicted to be very potent, are strongly polarized. Again, there is mainly di-ortho 2,6-substitution, but with the difference that chlorination is now predominantly found on one ring only.

In the light of this model interpretation, it is interesting to scrutinize what was made in the original publication (ref 11). Connors and coworkers concluded that di- and tri-ortho substituted PCBs exhibit the highest CYP2B-potency. But because they tested only one 2,6-disubstituted PCB, they might have missed the importance of this structural element for the modelled biological activity. Therefore, the future use of an appropriately tailored mapping set seems highly motivated.

## REFERENCES

1. T. Lundstedt, et al., Intelligent combinatorial libraries, in: *Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry*. H. van de Waterbeemd, B. Testa and G. Folkers, eds., Wiley-VCH, Weinheim (1997).
2. J.E. Jackson. *A User's Guide to Principal Components*, John Wiley & Sons, Inc., New York (1991).
3. Å. Nordahl and R. Carlson, Exploring organic synthetic procedures, *Top. Curr. Chem.* 166:1 (1993).
4. R. Granberg, *Solubility and Crystal Growth of Paracetamol in Various Solvents*, Ph.D. Thesis, Royal Institute of Technology, Stockholm, Sweden (1998).
5. E.U. Ramos, W.H.J. Vaes, H.J.M. Verhaar, and J.L.M. Hermens. Polar narcosis: designing a suitable training set for QSAR studies, *Environ. Sci. & Pollut. Res.* 4:83 (1997).
6. L. Eriksson, and J.L.M. Hermens, A multivariate approach to quantitative structure-activity and structure-property relationships, in: *The Handbook of Environmental Chemistry, Vol 2H, Chemometrics in Environmental Chemistry*, J. Einax, ed., Springer-Verlag, Berlin (1995).
7. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold, New chemical descriptors relevant for the design of biologically active peptides, *J. Med. Chem.* 41:2481 (1998).
8. M. Tysklind, P. Andersson, P. Haglund, B. van Bavel, and C. Rappe, Selection of polychlorinated biphenyls for use in quantitative structure-activity modelling, *SAR QSAR Env. Res.* 4:11 (1995).
9. P. Andersson, P. Haglund, and M. Tysklind, The internal barriers of rotation for the 209 polychlorinated biphenyls, *Environ. Sci. & Pollut. Res.* 4:75 (1997).
10. P. Andersson, P. Haglund, and M. Tysklind, Ultraviolet absorption spectra of all 209 polychlorinated biphenyls evaluated by principal component analysis, *Fresenius J. Anal. Chem.* 357:1088 (1997).
11. K. Connor, S. Safe, C.R. Jefcoate, and M. Larsen, Structure-dependent induction of CYP2B by polychlorinated biphenyl congeners in female Sprague-Dawley rats, *Biochem. Pharm.* 50:1913 (1995).
12. L. Eriksson, J.L.M. Hermens, E. Johansson, H.J.M. Verhaar, and S. Wold, Multivariate analysis of aquatic toxicity data with PLS, *Aquatic Sciences* 57:217 (1995).
13. SIMCA-P 7.0 and manual, Umetri AB, www.umetri.se.
14. P.F. De Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart, and R. Phan-Thau-Luu, D-optimal designs, *Chemom. Intell. Lab. Syst.* 30:199 (1995).
15. MODDE 4.0 and manual, Umetri AB, www.umetri.se.

# **Section III**

## **The Future of 3D-QSAR**



## HANDLING INFORMATION FROM 3D GRID MAPS FOR QSAR STUDIES

Gabriele Cruciani, Manuel Pastor and Sergio Clementi

Laboratory of Chemometrics  
University of Perugia  
06123 Perugia, Italy

### INTRODUCTION

3D-QSAR is an interesting and expanding discipline<sup>1-2</sup>. Nowadays software for 3D-QSAR methodologies and efficient algorithms to describe molecules and to predict biological activity are more accessible and easy to use<sup>3-5</sup>. Progress were done on the numerical description of the biological systems which now are more precise and detailed<sup>6</sup>.

In the past, most efforts were devoted to improve the numerical performance of the statistical models. However, one weak point of the methodology resides on the model interpretation. All the 3D-QSAR methodologies benefit from the use of two-dimensional and three-dimensional plots. However, with so many descriptors, the interpretation of two-dimensional and three-dimensional plots becomes messy and the structure-activity relationships are often very difficult to understand.

It is generally true that the interpretation phase is one of the most accurate validation of the model and that without a correct interpretation a model is completely useless.

The degree of complexity of a 3D-QSAR model depends on the data set under study. Even in the case of a simple data set, the interpretation of the results is often difficult and demanding. In order to interpret a 3D-QSAR model, the user should be able to understand the chemical model, the statistical model and the link between them. Sometimes this goal is very difficult to achieve. Reference 1 reports about 400 papers on 3D-QSAR field published over the last four years. Only few of them report a deep discussion on the interpretation phase, demonstrating the actual difficulties in this important aspect.

Interpretation is the only way we can improve our knowledge by our intelligence, that is much better than using only the *artificial intelligence* of the model.

### CHEMICAL MODEL: 3D GRID MAP

A 3D grid map may be viewed as a 3D matrix the elements of which are the attractive and repulsive forces, mapped by color coding, between an interacting partner and a target molecule. The majority of properties related with molecular interactions can be represented

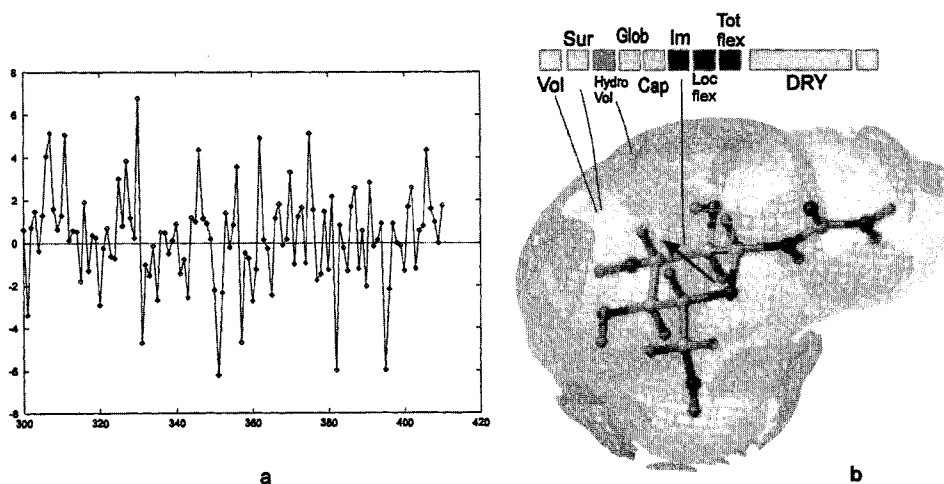
in a 3D grid map and thus these maps are useful to visualize large amount of molecular data and chemical information in a simple and comprehensive fashion.

The amount of information contained in a 3D grid map is related to the interacting molecular partners. Sometimes visual inspection is not sufficient since large amount of information is being coded and hidden in the sign and magnitude of the grid node forces, in the position of the grid nodes, in the grid nodes relationships and in other functional relationships. Specialized tools are required in order to help the user to interpret and to understand at best one particular problem.

3D-ACC<sup>7</sup> was proven to be an effective method for handling information from 3D grid maps of planar molecules. The method has the advantage to highlight the grid-nodes relationships, producing a new description which is practically independent from the location of the target molecules within the grid cage<sup>7</sup>. However, the new description produced by 3D-ACC is hard to understand. Although a 3D-QSAR model can be obtained, the usefulness of the model is limited by the difficult interpretation.

The VolSurf method<sup>8-11</sup> can be a good alternative to 3D-ACC. VolSurf is a computational procedure to explore the physicochemical property space from 3D grid maps. The basic concept of VolSurf is to compress the information present in 3D grid maps into few numerical descriptors very simple to understand and to interpret. Compression of the information can be better made if chemical knowledge is added to the process. VolSurf does so by selecting the most appropriate descriptors and parameterization according to the type of map under study. In the standard procedure, GRID<sup>12</sup> interaction fields with a water probe, a hydrophobic probe and a charged probe are used. However, other grid maps produced by different molecular mechanic or semiempirical methods can be used. VolSurf has the nice advantage of producing 2D descriptors using the 3D information embedded in the 3D maps. It is clear that not all the information can be transferred from 3D to 2D descriptors, but this is the only price to pay to obtain *lattice independent descriptors*.

In the following example (see Figure 1) part of a 3D-ACC transformation and a VolSurf transformation of a glucose analogue molecule are reported for comparison.



**Figure 1.** 3D-ACC (a) and VolSurf (b) numerical transformations of a 3D map for a glucose analogue molecule. The black vector on the glucose moiety ring represents one of the integrity moments calculated by VolSurf.

3D-ACC transformations are difficult to understand and not reversible: the spectra-like diagrams obtained cannot be transferred back into the original 3D grid map. Conversely, the VolSurf transformation is easier to understand, the descriptors have a clear chemical meaning and some of them can be projected back into the original 3D grid map from which they were obtained.

VolSurf is a sort of interface between the graphical representation of 3D grid maps and our need to produce from these maps useful numerical descriptors. The usefulness of this simple procedure was demonstrated in its practical applications in the fields of QSAR, 3D-QSAR and membrane penetration<sup>8-11</sup>.

## STATISTICAL MODELS

Principal Components Analysis (PCA) and Partial Least Squares (PLS) are chemometric tools for extracting and rationalizing the information from a multivariate description of a biological system. The complexity reduction and the data simplification are two of the most important features of such chemometric models<sup>13</sup>. PCA and PLS methods have the nice feature to condense the overall information into two smaller matrices, which in 3D-QSAR, show the molecule pattern (score plot) and the 3D descriptor pattern (loading plot)<sup>14</sup>.

However, while the interpretation of loading plots in classical QSAR analysis is simple and straightforward, in 3D-QSAR it becomes messy and apparently without a practical benefit. This is due to the huge amount of variables used in 3D-QSAR models and to the fact that the information in a 3D grid map resides in the actual position of each grid-node (variable) in the real 3D grid cage. The position-dependent information contained in each 3D variable is lost in the standard loading plots and so does the spatial correlation between the grid-nodes (variables). Since it is the pairwise comparison of loading and score plots that makes chemometric methods so powerful, when one of the two plots is useless no pairwise comparisons can be properly made.

In the following example, reporting the search for selectivity in Receptor-Based Drug Design<sup>15</sup>, a score plot and a loading plot were obtained. It should be noticed that all plots are linked to the actual 3D space of molecules by a PCA model. Any change in the position of a single object would produce changes in all the plots.

Since the relationships between such plots are quite complex, special tools are required in order to help the user to interpret and to fully understand this particular case study. Our proposal for this kind of tools will be reported later on in this chapter.

## MULTIBLOCK METHODS

Multiblock and hierarchical PCA and PLS methods have reported recently to be of interest for improving the interpretability of multivariate models in cases where the number of variables is large and there is some criteria which justifies the grouping into conceptually consistent blocks. This is the case in 3D-QSAR, where multiple blocks of 3D-descriptors are often produced. For example, the standard CoMFA procedure<sup>16</sup> describes the compounds with two blocks of variables: the steric field descriptor block and the electrostatic field descriptor block. Similar descriptions are produced in other models which use of lipophilic field<sup>17-18</sup> or GRID probes<sup>12</sup>.

In the field of ligand selectivity, in Structure-Based Drug Design, multiblock methods like the Consensus Principal Components Analysis CPCA can improve significantly the interpretability of multivariate models<sup>19</sup>.

Multiblock methods operate at two levels<sup>20</sup>: at the lower level each block of data is

treated separately, taking into account the variance of the field within each individual block. At the higher level the blocks are related to each other, retaining the independent information in each variable. The use of a two levels model allows to assess the importance of each grid-node within each descriptor block, while assessing simultaneously the relative importance of each block of descriptors.

Multiblock methods produce also plots that are linked one another. This is a further case in which appropriate tools are required in order to understand and take full advantage of these sophisticated methods in the field of 3D-QSAR.

## INTERACTIVE PLOTS

In all the different stages of 3D-QSAR modeling, the use of different kinds of plots can be very useful. Plots are used to visualize the molecules, to show the statistical relationships, to help the interpretation of the results. However, up to now, the plots were represented independently and not linked together at all.

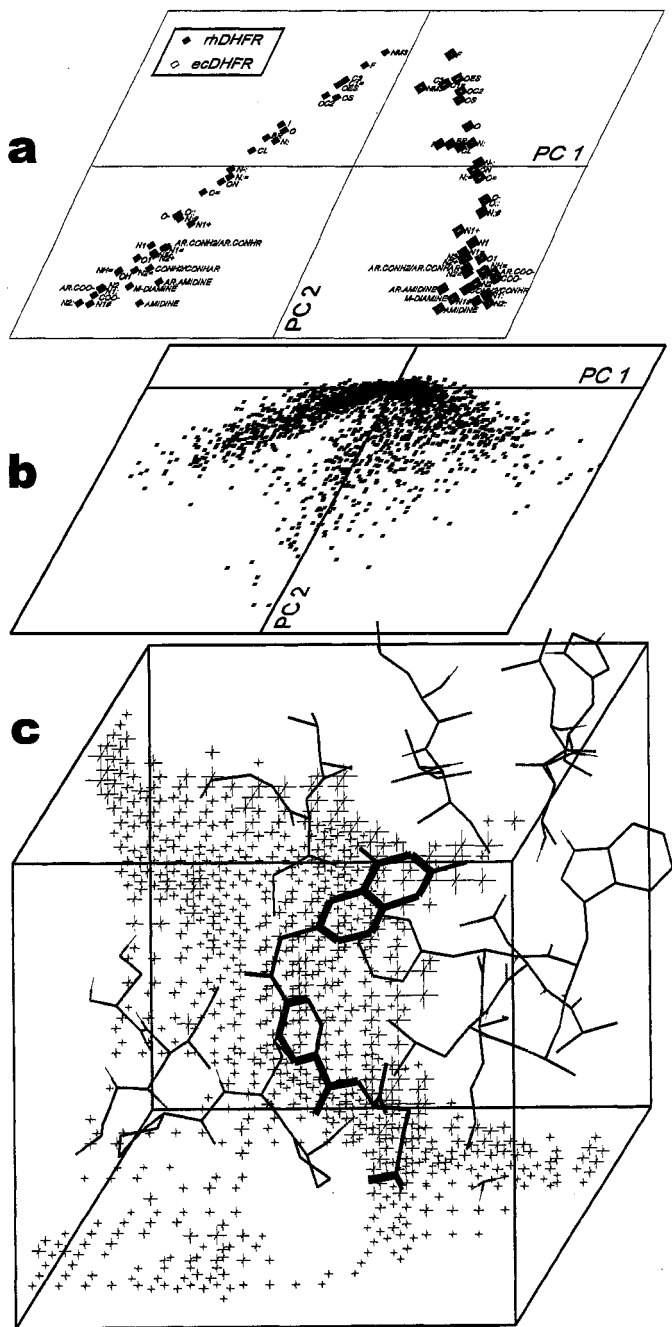
Interactive plots are couples of 2D and 3D representations of the model linked together interactively: the user can make selections of define virtual objects in one of the plots and see the effect of those actions reflected immediately in the linked plot. Interactive plots can be used to visualize how a change in a certain space (statistical space) is reflected in a different space (chemical space). For example, they can relate variables selected in a loading plot with the associated positions around the chemical structure or represent the field produced by a virtual compound placed somewhere in a PCA or PLS scores plot<sup>21-22</sup>.

Interactive plots were developed in our research group and their first implementation is already present in GOLPE version 4.0<sup>23</sup>. In these, the user can interact passively, by selecting positions in the plots in order to see these positions in a different space, or actively, by introducing "virtual objects", thus obtaining interactively the corresponding changes in the real 3D-space. With the help of those plots, important features of the chemicals can be easily highlighted, obtaining simple representations of the most important regions nearby the molecules.

The hidden link between many different plots can be directly evidenced with the help of interactive plots. For example, the relative contributions to activity or selectivity are immediately shown in the related 3D-space of chemicals. Moreover, interactive boundary translations<sup>21</sup> from 2D to 3D plot can lead to better interpretation of all the plots, namely 2D loading plots or multiblock loading plots.

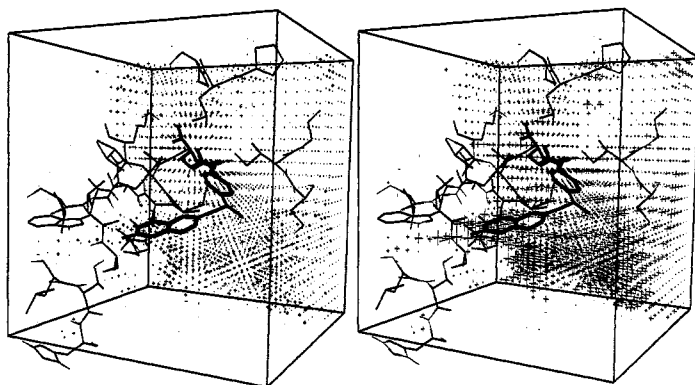
Figure 2 shows the score, loading and 3D grid map plots for two varieties of dihydrofolate reductase DHFR enzymes interacting with 41 GRID chemical probes<sup>12</sup>. The score plot represents the chemical probes while the loading plot represents 3D-grid point interaction energies between the probes and the two enzymes. Although this is the simplest example of selectivity in Structure-Based Drug Design (only two proteins are used) the interpretation of the plots is not straightforward. There is no direct way to answer practical questions like: what is the chemical meaning associated to the principal components axes in these plots? How these plots can be used to enhance selectivity or affinity in a ligand molecule? Since interactive plots are a sort of blackboard where the user can actively interact, the action of moving a virtual object over the plot is perfectly allowed. For this reason the user can move an object in the score plot from positions near to the fluorine probe (labeled as F) at the top of the Figure 2a (high value in PC2), to the bottom of the Figure 2a (low value in PC2). Since the interactive plots are linked one another, the 3D grid map will change interactively. Some regions will change as a result of this modification of the PC2 thus being highlighted in the 3D grid map.

At the same time the user can also draw in the loading plot a polygon enclosing interesting loading coefficients. Once more, immediately the 3D grid plots will update reporting together with the movement in the second PC axis the corresponding region linked to the loading plot.



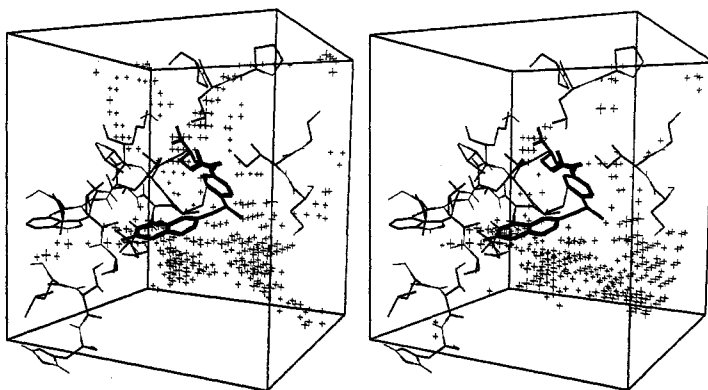
**Figure 2.** Score plot (a), loading plot (b) and actual 3D grid map (c) for the dataset of selectivity in Structure-Based Drug Design. All the plots are linked together, so any change in one of the plots is reflected by a change in the others.

The movement of the virtual object from positions near to the fluorine probe to the final position at the bottom of the plot produces modifications in the 3D grid maps of the two enzymes which are reported in Figure 3. From this figure it can be seen that the highlighted regions are the same in the two grid plots, but the intensity of the interaction fields is different, thus showing that the second PC express mainly the different magnitude of the interaction energies. While the fluorine probe interacts weakly with both enzymes, the N2: probe interacts strongly with them. Clearly it is possible to conclude that PC2 reports affinity regions in which the probe are ordered according to their ability to interact with common parts in both enzymes.



**Figure 3.** The movement of the object in the score plot of Figure 2 along the PC2 axis highlights the same regions in both proteins. The interaction fields are larger for probes in the lower part of the score plot.

Conversely, if a virtual object near to the hydroxyl (OH) probe is moved from its position in the human DHFR on the left to a new position on the right of the plot along the PC1 axis, the corresponding modification on the actual 3D space of proteins will show different regions. Figure 4 represents those positions in the lattice binding site where the probes establish strong selective interactions. It is important to point out that the position of these selective regions depends on the chosen probe, as clearly demonstrated by interactive plots when different objects representing different probes are selected and moved in the plot along the PC1 axis.

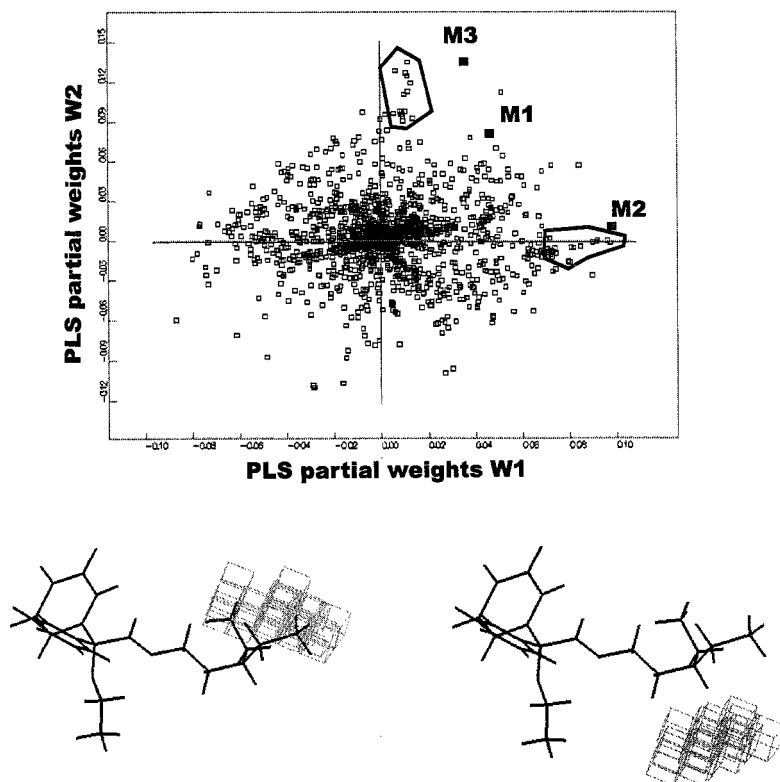


**Figure 4.** The movement of the object in the score plot of Figure 2 along the PC1 axis highlights different regions in the proteins. These are selective regions of interaction.

Gratteri et al.<sup>24</sup> reported a data set consisting on ninety M1, M2 and M3 muscarinic antagonist compounds. Among them, the most diverse compounds were superimposed and described with GRID OH probe, producing a data matrix consisting on nineteen chemicals described by more than 14.000 interaction energy variables (X descriptor matrix) plus the biological activity against the three M1, M2 and M3 receptor subtypes (Y response matrix).

The 3D-QSAR model was developed mainly to evidenciate the structural features required in a ligand in order to make a selective interaction with a specific receptor subtype. Figure 5 shows the interactive PLS partial weight plot and the corresponding 3D grid plot for such a data matrix.

The mutual position of the three responses M1, M2 and M3 in the partial weight plot evidenciates that it is not possible to increase one of the three biological activities without simultaneously increasing the others. Probably this is the maximum amount of information which one can obtain from a plain PLS partial weight plot. However, more information can be extracted from interactive plots. In fact, with the interactive boundary translation procedure, some of the variables reported in the W1-W2 plot of Figure 5 can be translated into the real space of chemicals. The user will decide the variables to be translated simply drawing a polygon enclosing them in the partial weight plot. In Figure 5, the two regions containing variables nearby M2 and M3 receptor subtypes will result in showing in the 3D space the corresponding two regions important for receptor selectivity.



**Figure 5.** Partial weight plot and real 3D grid plot for a muscarinic antagonist compound. Interactive boundary translation highlights the two regions for M2 and M3 selectivity.

## CONCLUSIONS

Since the work of Cramer<sup>16</sup> (CoMFA) the 3D-QSAR field has changed dramatically. New computational procedures were used to describe molecules like GRID<sup>12</sup>, CoMSIA<sup>4</sup>, CLIP<sup>17-18</sup>, CoMPA<sup>25</sup>, HINT<sup>26</sup>. New procedures were used to compute the statistical models like GOLPE<sup>27</sup> and SAMPLS<sup>28</sup> and new procedures were published for handling 3D regions<sup>5,29,30</sup>. Successful attempts to obtain information from a 3D receptor structure in 3D-QSAR were developed (COMBINE)<sup>31-32</sup> and new statistical tools for working with multiblock matrices were produced<sup>19-20</sup>.

These procedures have in common the fact that all of them use 3D grid maps of descriptors and plots to highlight the information contained in the data. In all the procedures interpretation is the crucial step.

The present work addresses the important problem of the interpretation phase from two different aspects. When difficulties arise from the superposition phase a proper compression of the information from the 3D grid map into a condensed vector of descriptors, easy to use and to interpret, can be really advantageous. When the problem is mainly related with the model interpretation then interactive plots can be helpful to increase the amount of useful information extracted from any 3D-QSAR model.

## ACKNOWLEDGMENTS

We thank our colleagues P.Gratteri, S.Scapecchi, M.N.Romanelli, F.Melani for the data regarding the muscarinic antagonist compounds. The Italian founding agencies of MURST and CNR are thanked for financial support to G.C. and S.C.; Rhone-Poulenc Rorer for financial support to M.P.

## REFERENCES

1. K.H. Kim, G. Greco, E. Novellino, in *3D-QSAR in drug design: Vol 3*, H. Kubinyi, G. Folkers, Y.C. Martin Ed.s, Kluwer Academic Publisher, Dordrecht, The Netherlands, 1998, pp.257-316.
2. A.J. Hopfinger, S. Wand, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, Construction of 3D-QSAR models using the 4D-QSAR analysis formalism, *J.Am.Chem.Soc.*, 119:10509, (1997).
3. W.J. Dunn III, A.J. Hopfinger, in *3D-QSAR in drug design: Vol 3*, H. Kubinyi, G. Folkers, Y.C. Martin Ed.s, Kluwer Academic Publisher, Dordrecht, The Netherlands, 1998, pp.167-182
4. G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, *J.Med.Chem.*, 37:4130, (1994).
5. M. Pastor, G. Cruciani, S. Clementi, Smart region definition (SRD): a new way to improve the predictive ability and interpretability of 3D QSAR models, *J.Med.Chem.*, 40:1455 (1997).
6. M. Pastor, G. Cruciani and K.A. Watson. A strategy for the incorporation of water molecules present in a ligand-binding site into a 3D-QSAR analysis, *J. Med. Chem.* 40:4089 (1997).
7. S. Clementi, G. Cruciani, D. Riganelli, R. Valigi, G. Costantino, M. Baroni, S. Wold. Autocorrelation as a tool for a congruent description of molecules in 3D-QSAR studies, *Pharm. Pharmacol. Lett.* 3:5 (1993).
8. G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, G. Musumarra, GRID characterization of heteroaromatics, in: *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, F. Sanz Ed., J.R. Prous Sci., 493-495, (1995)
9. S. Clementi, G. Cruciani, P. Fifi, D. Riganelli, R. Valigi, G. Musumarra, A new set of principal properties for heteroaromatics obtained by GRID, *Quant. Struct.-Act. Relat.*, 15:108 (1995).
10. W. Guba and G. Cruciani, Molecular Field-Derived Descriptors for the Multivariate Modeling of Pharmacokinetic Data, this conference.
11. R. Mannhold, G. Cruciani, H. Weber, H. Lemoine, A. Derix, C. Weichel, M. Clementi, 6-varied benzopyrans as potassium channel activators: synthesis, vasodilator properties and multivariate analysis, *work in preparation*.
12. P.J. Goodford, Computational procedure for determining energetically favourable binding sites for biologically important macromolecules, *J.Med.Chem.*, 28:849 (1985).



13. S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.*, 2:37 (1993)
14. G. Cruciani, S. Clementi, GOLPE, philosophy and applications in 3D-QSAR. In: *Advances computer-assisted techniques in drug discovery*, Vol. 3, Waterbeemd H.v. der, ed., Weinheim VCH, 61-88 (1995)
15. M. Pastor and G. Cruciani, A novel strategy for improving ligand selectivity in receptor-based drug design, *J. Med. Chem.*, 38:4637 (1995).
16. R.D. III Cramer, D.E. Patterson, J.D. Bunce, Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroid to Carrier Proteins, *J. Am. Chem. Soc.*, 110:5959 (1988).
17. P. Gaillard, P-A. Carrupt, B. Testa, A. Boudon, Molecular lipophilicity potential, a tool for 3D QSAR: methods and applications, *J. Comput.-Aided Mol. Des.*, 8:83 (1994)
18. CLIP 1.0, Institute of Medicinal Chemistry, University of Lausanne, 1996
19. J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, *Journal of Chemometrics*, 1998 in press.
20. M.C. De Rosa and A. Berglund, A new method for predicting the alignment of flexible molecules and orienting them in a receptor cleft of known structure, *J. Med. Chem.* 41:691 (1998).
21. G. Cruciani, S. Clementi, M. Baroni, M. Pastor, in *Rational Molecular Design in Drug research*, T. Liljefors, F. Jorgensen, P. Krosgaard-Larsen Ed.s, Alfred Benzon Symposium 42, 1998, pp. 87-97 (1998).
22. G. Cruciani, Chemometrics in 3D-QSAR and Structure-Based Drug Design, Second European Workshop in Drug Design, Siena, Italy (1998).
23. GOLPE 4.0, Multivariate Infometric Analysis, Viale dei castagni 16, Perugia, Italy (1998).
24. P. Gratteri, G. Cruciani, S. Scapecchi, M.N. Romanelli, F. Melani, 3D-QSAR, GRID descriptors and chemometric tools in the development of selective antagonist of the muscarinic receptor, 12<sup>th</sup> European Symposium on Quantitative Structure-Activity Relationships, Denmark, Copenhagen (1998).
25. P. Floersheim, J. Nozulak, H.P. Weber, Experience with Comparative Molecular Field Analysis, In: Trends in QSAR and Molecular Modelling '92, Wermut ed., ESCOM, Leiden, 227 (1993).
26. G.E. Kellog and D.J. Abraham, *J. Mol. Graph.*, 10:212 (1992).
27. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems, *Quant. Struct.-Act. Relat.*, 12:9 (1993).
28. B.L. Bush, R.B. Nachbar, *J. Comp.-Aided Molec. Des.*, 7:587 (1993).
29. S.J. Cho, A. Tropsha, Cross-validated  $r^2$ -guided region selection for comparative molecular field analysis: a simple method to achieve consistent results, *J. Med. Chem.*, 38:1060 (1995).
30. U. Norinder, Single and domain mode variable selection in 3D QSAR applications, *J. Chemometr.* 10:95 (1996).
31. A.R. Ortiz, M.T. Pisabarro, F. Gago and R. Wade, Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 38:2681 (1995).
32. C. Pérez, M. Pastor, A.R. Ortiz and F. Gago, Comparative binding energy (COMBINE) analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design, *J. Med. Chem.* 41:836 (1998).

## GAUSSIAN-BASED APPROACHES TO PROTEIN-STRUCTURE SIMILARITY

Jordi Mestres,<sup>1</sup> Douglas C. Rohrer,<sup>2</sup> and Gerald M. Maggiora<sup>2</sup>

<sup>1</sup> Department of Molecular Design & Informatics, N.V. Organon  
5340 BH Oss, The Netherlands

<sup>2</sup> Computer-Aided Drug Discovery, Pharmacia & Upjohn  
Kalamazoo, MI 49001

### INTRODUCTION

The number of protein structures available in the PDB<sup>1</sup> (7415 in July 22, 1998) is constantly growing and it is expected to increase even more rapidly in coming years.<sup>2</sup> This tremendous body of information is certainly having an impact in ligand design and modelling where the knowledge of the crystal structure of the target protein, or a closely related protein, makes a significant difference in the process of ligand optimization.<sup>3</sup> When a crystal structure of the target protein is unavailable, ligand optimization must rely on indirect approaches based on the similarity between the structures of the ligands themselves.<sup>4</sup> However, if a three-dimensional structure of the target protein is available docking methods can be applied,<sup>5</sup> which have the potential of providing important information on the interaction between the ligand and the residues of the given receptor site.

Unfortunately, at the initial stages of a drug design project, availability of the target protein crystal structure is unlikely. Nevertheless, the PDB often contains several crystal structures of proteins related to the protein of interest. Comparative analyses between the available protein structures<sup>6</sup> provide a means for revealing common structural features among the proteins of the family and, at the same time, identify those regions where the structures are more diverse. Previous studies have shown that the structural motif of a protein family is much better conserved than the amino acid sequence<sup>7</sup> and that identification of conserved structural features can be used in deriving, by homology, a model structure of more distant members of a family.<sup>8</sup> The construction of a structural model of the target protein in conjunction with a deep understanding of the similarities and dissimilarities with other members of the family is of key importance in ligand optimization. Such a model provides many clues for suggesting modifications to the ligand that can potentially enhance binding with the target protein and also improve selectivity and specificity with respect to other family members.

A similarity comparison of protein structures requires first finding the optimum three-dimensional alignment.<sup>9</sup> The difficulties in locating the optimum alignment between protein structures depending on their topology have been widely discussed and even the existence of a unique optimum alignment solution has been questioned.<sup>10</sup> Moreover, it is evident that any derived structural alignment will ultimately depend on the measure used to quantify the similarity. The need to adequately address these many inherent ambiguities has led to the continuous search for improved methods which produce sequence-independent means to identify and quantify protein-structure similarity.

The aim of this contribution is to present the use of a Gaussian-based approach for assessing protein-structure similarity. A description of the methodological aspects is presented next, followed by a discussion on the potential usefulness of performing not only rigid alignments but also flexible alignments between protein structures.

## METHODOLOGY

The present approach is based on the use of Gaussian functions to represent the structure of proteins, as implemented in the program MIMIC.<sup>11</sup> A Gaussian function,  $g_k$ , centered at position  $\mathbf{R}_k$  is given by

$$g_k(\mathbf{r}) = \alpha_k \cdot \exp(-\beta_k |\mathbf{r} - \mathbf{R}_k|^2), \quad (1)$$

where the coefficient,  $\alpha_k$ , and the exponent,  $\beta_k$ , determine the value of its maximum height at the origin and its decayment, respectively. In general, each atom,  $a_i$ , can be represented by a number of Gaussian functions

$$a_i(\mathbf{r}) = \sum_{k \in i}^n g_k(\mathbf{r}). \quad (2)$$

At this atomic level, the number of Gaussian functions,  $n$ , used to represent each atom will depend on the accuracy desired to reproduce the atomic electron density. Each amino acid,  $A_I$ , in a protein is represented as

$$A_I(\mathbf{r}) = \sum_{i \in I}^N a_i(\mathbf{r}) \quad (3)$$

and depends on the number of Gaussian centers,  $N$ , used to define the structure of each amino acid  $I$ . Normally, at the amino acid level, the centers are taken as the positions of the atoms constituting the amino acid. In general, however, these centers need not to correspond atom positions. Such "off-center" functions can still reproduce adequately the steric shape of an amino acid while reducing the number of Gaussian functions used. Finally, the structural characteristics of a protein,  $P$ , will be given by

$$P(\mathbf{r}) = \sum_{I \in P}^M A_I(\mathbf{r}). \quad (4)$$

At the protein level,  $M$  will normally correspond to the number of amino acids of the protein. However, in general,  $M$  could be optimized in order to obtain the minimum number of Gaussian functions that, placed strategically, still reproduce the structural characteristics of the protein. The final number of Gaussian functions employed will be a compromise between the level of structural detail desired and the amount of computing time required to evaluate the similarities. In this work,  $M$  will be the number of amino acids of the protein and a single-Gaussian amino acid approach ( $N=1$ ) with each function centered at the positions of the  $C_\alpha$  carbons ( $n=1$ ) will be used. An analysis of the dependency of the protein-structure similarity on the amino acid Gaussian representation can be found elsewhere.<sup>12</sup>

Note that the use of a Gaussian-based representation to evaluate protein-structure similarities is in fact a resolution-based approach. For a given  $\alpha_k$ , different  $\beta_k$  parameters in eq. (1) will lead to different values of structural similarity. On one side, for very small  $\beta_k$  values (low resolution) every protein structure would look almost alike, whereas on the other side, very large  $\beta_k$  values (high resolution) would result in no overlap at all between the structural representations and, thus, every protein structure would be essentially unique. In between these two limit cases there is a long range of possibilities and ultimately,  $\beta_k$  values should be user-customizable. In the present study, the coefficient  $\alpha_k$  and the exponent  $\beta_k$  in eq. (1) are optimized for each atom to reproduce its van der Waals steric volume as originally implemented in the program MIMIC.<sup>11</sup>

Once a Gaussian representation of the protein structure is defined, for a given protein superposition the structural similarity between two proteins  $A$  and  $B$ ,  $S_{AB}$ , is assessed by evaluating the overlap integral,  $Z_{AB}$ , between their respective Gaussian-based structure representations,  $P_A$  and  $P_B$  as defined by eq. (4),

$$Z_{AB} = \int P_A(\mathbf{r}) P_B(\mathbf{r}) d\mathbf{r} \quad (5)$$

which can be then normalized using a cosine-like index

$$S_{AB} = \frac{Z_{AB}}{(Z_{AA} \cdot Z_{BB})^{1/2}} \quad (6)$$

The values of  $S_{AB}$  in eq. (6) range from 0 to 1. A value of 1 is achieved only in the limit case of identity; any dissimilarity between the two proteins will be reflected in a value smaller than 1. Rigid alignments between pairs of protein structures are obtained by optimizing  $S_{AB}$  in all translational ( $\mathbf{t}$ ) and rotational ( $\theta$ ) degrees of freedom using standard gradient-seeking procedures.<sup>12</sup> In addition, flexible alignments can be performed by allowing  $P_B$  to adapt its conformation to that of  $P_A$  and, in this case all torsional angles in  $P_B$  ( $\tau_B$ ) will be included as degrees of freedom in the  $S_{AB}$  optimization. Therefore, optimization of  $S_{AB}$  through flexible alignment of  $P_B$  to  $P_A$  can be expressed as

$$S_{AB}(\mathbf{t}, \theta, \tau_B) = \frac{Z_{AB}(\mathbf{t}, \theta, \tau_B)}{(Z_{AA} \cdot Z_{BB}(\tau_B))^{1/2}} \quad (7)$$

## FLEXIBLE ALIGNMENT OF PROTEIN STRUCTURES

One of the advantages of a Gaussian-based representation of protein structures is that it provides a fuzzier representation of the spatial location of atoms, which conforms with the inherent uncertainty of atomic positions in protein crystallography. Such a fuzzy description provides a means for optimizing the alignment of protein structures without the need for specifying residue-by-residue correspondences. This is specially important when performing structural alignments between proteins with low sequence identity. In addition to the ability to optimize the alignment of rigid protein structures, one of the proteins can be allowed to relax its conformation thus providing a flexible alignment. The usefulness of performing unbiased sequence-independent flexible alignments between protein structures will be underlined in the remaining of the paper.

The first application of flexible alignments is for the analysis of protein domain movements.<sup>13</sup> Most large proteins are built from domains whose movements provide excellent examples of protein flexibility. In most cases, the presence of a bound substrate promotes a closed conformation whereas its absence favors an open conformation. Therefore, analyses of domain movements can provide structural information that will lead to a better understand of the induced fit in protein recognition. As an illustrative example, Figure 1 shows the result of flexibly aligning two structures of P450<sub>BM3</sub> in its open and closed conformations (PDB codes are 2BMH and 1FAG, respectively). The rigid alignment gave a structural similarity of 0.6496 (structures in black in Figure 1). From this alignment, constrained flexible alignments were performed to systematically increase similarities between the two conformations to 0.70, 0.75, 0.80, and 0.85 (structures in grey in Figure 1), which can be taken as snapshots of the transition from one conformation to the other from a structural similarity point of view. Following this '*similarity coordinate*' it is possible to analyze the movement that takes place at each stage and track the amino acids responsible for this movement.



**Figure 1.** Flexible alignment between the ligand-free and the ligand-bound conformations of P450<sub>BM3</sub>.

A second promising use of flexible alignments is for automatically deriving structure-based sequence alignments between members of a given protein family.<sup>14</sup> In cases where the structures of the different family members present conformational changes, even the availability of a global rigid alignment is not revealing enough to translate it into a structure-based sequence alignment and a great deal of visual inspection usually needs to be done. Identification of common structural patterns is normally required to translate structural alignments into sequence alignments. Flexible alignment of protein structures can potentially indeed identify common structural patterns between proteins, thus providing a simple means for deriving structure-based sequence alignments in an unbiased automatic manner. Once a structure-based sequence alignment has been obtained, sequence-structure consensus regions among the different members of a protein family can be identified. This type of information can be then used as a constraint criterion in homology modeling of other family members for which a crystal structure has not been resolved yet.<sup>14</sup>

In conclusion, Gaussian-based approaches to protein-structure similarity emerge as a promising non-biased sequence-independent method for, first, obtaining protein structure alignments and, second, for analyzing and maximally exploiting the similarity information contained in those alignments (both at the structure and sequence levels) towards the construction of protein homology models with a higher value of predictability.

## REFERENCES

1. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F.J. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures, *J. Mol. Biol.* 112:535 (1977).
2. S.E. Brenner, C. Chothia, and T.J.P. Hubbard, Population Statistics of Protein Structures: Lessons from Structural Classifications, *Curr. Opin. Struct. Biol.* 7:369 (1997).
3. a) N.C. Cohen (Ed.). *Molecular Modeling in Drug Design*, Academic Press, San Diego (1996); b) P.S. Charifson (Ed.). *Practical Application of Computer-Aided Drug Design*, Marcel Dekker, New York (1997).
4. a) M.A. Johnson and G.M. Maggiora (Eds.). *Concepts and Applications of Molecular Similarity*, Wiley, New York (1990); b) H. Kubinyi (Ed.). *3D QSAR in Drug Design: Theory, Methods, and Applications*, ESCOM, Leiden (1993); c) P.M. Dean (Ed.). *Molecular Similarity in Drug Design*, Blackie, London (1995).
5. a) J.M. Blaney and J.S. Dixon, A Good Ligand is Hard to Find: Automated Docking Methods, *Perspect. Drug Discov. Design* 1:301 (1993); b) T.P. Lybrand, Ligand-Protein Docking and Rational Drug Design, *Curr. Opin. Struct. Biol.* 5:224 (1995); c) G. Jones and P. Willett, Docking Small-Molecule Ligands into Active Sites, *Curr. Opin. Biotechnol.* 6:652 (1995).
6. a) J.P. Overington, Comparison of Three-Dimensional Structures of Homologous Proteins, *Curr. Opin. Struct. Biol.* 2:394 (1992); b) C. Orengo, Classification of Protein Folds, *Curr. Opin. Struct. Biol.* 4:429 (1994).
7. a) C. Chothia and A.M. Lesk, The Relation Between Divergence of Sequence and Structure in Proteins, *EMBO J.* 5:823 (1986); b) J. Hubbard and T.L. Blundell, Comparisons of Solvent Inaccessible Cores of Homologous Proteins: Definitions Useful for Protein Modelling, *Protein Eng.* 1:159 (1987); c) C. Sander and R. Schneider, Database of Homology-Derived Protein Structures and Structural Meaning of Sequence Alignment, *Proteins* 9:56 (1991); d) T. Flores, C.A. Orengo, D. Moss, and J.M. Thornton, Conformational Characteristics in Structurally Similar Protein Pairs, *Protein Sci.* 2:1811 (1993).
8. a) W.R. Taylor, Identification of Protein Sequence Homology by Consensus Template Alignment, *J. Mol. Biol.* 188:233 (1988); b) G.J. Barton and M.J.E. Sternberg, Flexible Protein Sequence Patterns: A Sensitive Method to Detect Weak Structural Similarities, *J. Mol. Biol.* 212:389 (1990); c) M. Hilbert, G. Bohm, and R. Jaenicke, Structural Relationships of Homologous Proteins as a Fundamental Principle in Homology Modelling, *Proteins* 17:138 (1993).
9. a) W.R. Taylor and C.A. Orengo, Protein Structure Alignment, *J. Mol. Biol.* 208:1 (1989); b) J. Rose and F. Eisenmenger, A Fast Unbiased Comparison of Protein Structures by Means of the Needleman-Wunsch Algorithm, *J. Mol. Evol.* 32:340 (1991); c) G. Vriend and C. Sander, Detection of Common Three-Dimensional Substructures in Proteins, *Proteins* 11:52 (1991); d) N.N. Alexandrov, K. Takahashi, and N. Go, Common Spatial Arrangements of Backbone Fragments in Homologous and Non-Homologous

- Proteins, *J. Mol. Biol.* 225:5 (1992); e) H.M. Grindley, P.J. Artymiuk, D.W. Rice, and P. Willett, Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm, *J. Mol. Biol.* 229:707 (1993); f) L. Holm and C. Sander, Protein Structure Comparison by Alignment of Distance Matrices, *J. Mol. Biol.* 233:123 (1993); g) K. Diederichs, Structural Superposition of Proteins with Unknown Alignment and Detection of Topological Similarity Using a Six-Dimensional Search Algorithm, *Proteins* 23:187 (1995); h) A. Falicov and F.E. Cohen, A Surface of Minimum Area Metric for the Structural Comparison of Proteins, *J. Mol. Biol.* 258:871 (1996).
10. a) F. Zu-Kang and M.J. Sippl, Optimum Superimposition of Protein Structures: Ambiguities and Implications, *Folding & Design* 1:123 (1996); b) A. Godzik, The Structural Alignment Between Two Proteins: Is There a Unique Answer?, *Protein Sci.* 5:1325 (1996).
  11. a) J. Mestres, D.C. Rohrer, and G.M. Maggiora, MIMIC: A Molecular-Field Matching Program. Exploiting Applicability of Molecular Similarity Approaches, *J. Comput. Chem.* 18:934 (1997); b) J. Mestres, D.C. Rohrer, and G.M. Maggiora, A Molecular Field-based Similarity Approach to Pharmacophoric Pattern Recognition, *J. Mol. Graphics Mod.* 15:114 (1997); c) J. Mestres, D.C. Rohrer, and G.M. Maggiora, A Molecular Field-based Similarity Study of Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors, *J. Comput.-Aided Mol. Design*, in press.
  12. J. Mestres, P.D.J. Grootenhuys, D.C. Rohrer, and G.M. Maggiora, A Gaussian-Based Approach to Protein Structure Similarity. Application to Matrix Metalloproteinases and Cytochromes P450, to be submitted.
  13. a) M. Gerstein, A.M. Lesk, and C. Chothia, Structural Mechanisms for Domain Movements in Proteins, *Biochemistry* 33:6739 (1994); b) W.L. Nichols, G. Rose, L.F.T. Eyck, and B.H. Zymm, Rigid Domains in Proteins: An Algorithmic Approach to their Identification, *Proteins* 23:38 (1995); c) W. Wriggers and K. Schulten, Protein Domain Movements: Detection of Rigid Domains and Visualization of Hinges in Comparisons of Atomic Coordinates, *Proteins* 29:1 (1997).
  14. J. Mestres, unpublished results.

## MOLECULAR FIELD-DERIVED DESCRIPTORS FOR THE MULTIVARIATE MODELING OF PHARMACOKINETIC DATA

Wolfgang Guba<sup>1</sup> and Gabriele Cruciani<sup>2</sup>

<sup>1</sup>Hoechst Marion Roussel  
Building G 838  
Chemical Research  
65926 Frankfurt am Main  
Germany

<sup>2</sup>Laboratory of Chemometrics  
Department of Chemistry  
University of Perugia  
Via Elce di Sotto, 8  
06123 Perugia, Italy

### INTRODUCTION

The optimization of pharmacokinetic properties is still one of the greatest challenges in lead optimization, and for the most part it is based on trial and error. As pharmacokinetics is closely linked with physicochemical properties, experimental design and quantitative structure-property modeling are key factors to systematically explore physicochemical property space and to establish stable, predictive models for lead optimization. However, experimental measurements of relevant parameters are often time-consuming, difficult and expensive. Furthermore, in vitro/in vivo approaches require the synthesis of compounds and cannot be used for the prioritization of synthesis targets.

Within this general context novel descriptors will be introduced which are derived from GRID<sup>1</sup> molecular interaction fields with the H<sub>2</sub>O and the hydrophobic DRY probe. These descriptors assign physicochemical attributes to a 3D structure and, therefore, are suitable to select physicochemically representative subsets from a pool of candidates. Three examples will be used to demonstrate how the information contents of these descriptors can be used to correlate the 3D structures of molecules with the intestinal absorption of drugs in humans, brain-blood partitioning and renal vs. hepatic clearance. A more detailed description will be given in a future publication.

To be useful for (physicochemical) lead optimization a descriptor has to fulfill 3 requirements:

- it must be applicable to various classes of compounds
- the result of a classification or quantitative structure-property relationship (QSPR) must be interpretable in structural terms
- the calculation of descriptors must not be a rate-limiting step in model building.



## THE VOLSURF<sup>2</sup> DESCRIPTORS

The interaction of molecules with biological membranes is mediated by surface properties such as shape, electrostatics, hydrogen-bonding and hydrophobicity. Therefore, the GRID<sup>1</sup> forcefield was chosen to characterize potential polar and hydrophobic interaction sites by the H<sub>2</sub>O and DRY probe, respectively, and to transform this information into a quantitative scale by calculating the volume of the interaction contours. As outlined in Table 1, 28 descriptors from the H<sub>2</sub>O and 8 descriptors from the DRY probe are generated (they will be referred to as VolSurf<sup>2</sup> descriptors). The first 4 parameters describe the size and the shape of the molecule, the descriptors 5-12 indicate polar interaction sites at 8 different energy levels and descriptors 21-28 calculate the concentration of polar interactions on the molecular surface. The "integy moment" (13-20) is defined in analogy to the dipole moment and describes the distance of the center of mass to the barycenter of polar interaction sites at a given energy level. If the integy moment is high, there is a clear separation between polar and hydrophobic interaction sites. If the integy moment is small, the polar moieties are either close to the center of mass or they are at opposite ends of the molecule and the resulting barycenter is close to the center of the molecule. Descriptors 29-36 indicate interactions with the hydrophobic probe at 8 different energy levels, which have been adapted to the energy range of the DRY probe. Summing up, the 3D structure is translated into physicochemically meaningful descriptors without the need for an alignment. Thus, size, shape, hydrogen-bonding and hydrophobicity can be quantitatively differentiated within a series of molecules. The resulting collinearity of descriptors is properly taken into account by multivariate statistics (principal component analysis<sup>3</sup> (PCA), partial least squares projections to latent structures<sup>4</sup> (PLS)).

**Table 1.** Definition of VolSurf<sup>2</sup> parameters. Descriptors 1-28 are generated with the H<sub>2</sub>O probe, descriptors 29 - 36 are calculated with the DRY probe

numbering	definition
1	total volume (0.25 kcal/mol)
2	total surface (0.25 kcal/mol)
3	total volume ( $V_{tot}$ ) / total surface ( $S_{tot}$ )
4	globularity ( $S_{tot}/S_e$ ; $S_e$ : surface area of equivalent sphere with $V_{tot}$ )
5 - 12	$V_-$ for interactions with the H <sub>2</sub> O probe at 8 energy levels (-0.2, -0.5, -1.0, -2.0, -3.0, -4.0, -5.0, -6.0 [kcal/mol])
13 - 20	integy moment: proportional to distance between barycenter of $S_{tot}$ and $V_-$ (at the above energy levels)
21 - 28	capacity: $V_-/S_{tot}$ (at the above energy levels)
29 - 36	$V_-$ for interactions with the DRY probe at 8 energy levels (-0.2, -0.4, -0.6, -0.8, -1.0, -1.2, -1.4, -1.6 [kcal/mol])

The physicochemical significance of the VolSurf descriptors was first determined by a comparison with the principal properties of amino acids (known as *z-scales*<sup>5</sup>), which were extracted by PCA from a multiproperty matrix containing 29 experimental measurements for each amino acid. The first principal component is interpreted as hydrophilicity, *z2* is associated with steric bulk, and *z3* describes electronic properties. For the calculation of VolSurf descriptors each amino acid in its neutral form was built in standard geometry and energy-minimized. The PLS analyses of each of the three *z-scales* as y-variable and the

VolSurf descriptors as X-matrix yielded one-dimensional models for the correlation of the VolSurf descriptors with  $z_1$  and  $z_2$ . As it can be seen in the PLS  $t$ - $u$  score plots and the corresponding loading plots (Fig. 1),  $z_1$  is positively correlated with capacity and polarity and negatively correlated with size, integrity moment and dispersion (first energy level of the DRY probe). Large hydrophobic amino acids (W, F, I, L) are well separated from polar amino acids (E, D, S) in the score plot. In the case of  $z_2$ , a correlation with size and hydrophobicity can be observed. In the corresponding score plot small amino acids (G, A) are separated from large amino acids (W, R). The correlation with  $z_3$  is quite low, but  $z_3$  explains a minor proportion of the original multiproperty matrix. In conclusion, the VolSurf parameters are very efficient descriptors of hydrophilicity, steric bulk and hydrophobicity as derived from the physicochemical characterisation of amino acids.

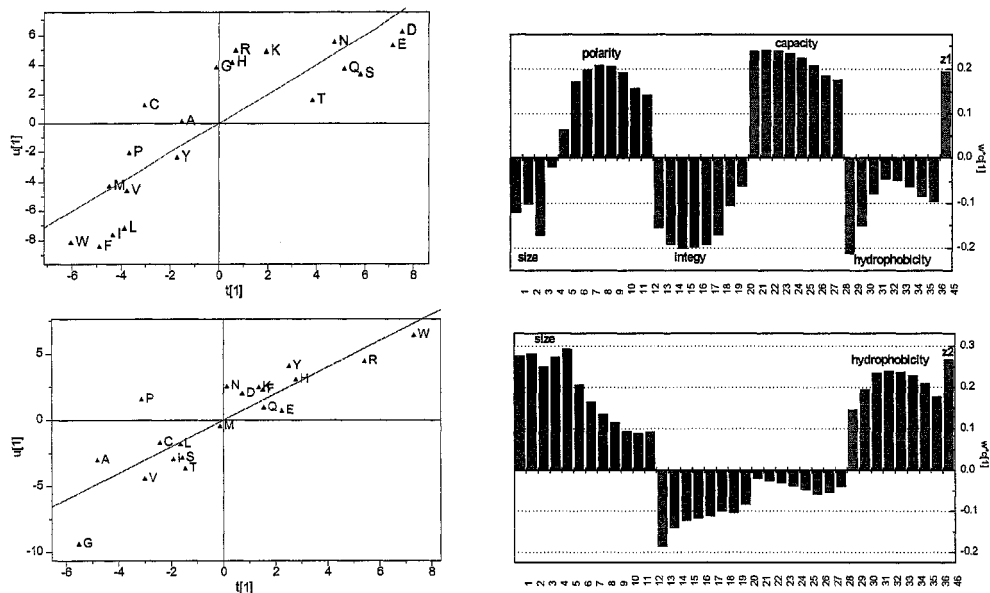


Fig. 1. PLS scores  $t_1$  and  $u_1$  and the corresponding partial weights plot for the correlation of VolSurf descriptors with  $z_1$  (top) and  $z_2$  (bottom).

## CORRELATIONS WITH PHARMACOKINETIC DATA

In the following, the application of the VolSurf descriptors for the multivariate modeling of pharmacokinetic data will be demonstrated using 3 literature examples. For all the statistical calculations SIMCA-S<sup>6</sup> was used.

In the first example the calculation was performed for a series of passively absorbed drugs with reliable data on human intestinal absorption (%HIA) covering a range of absorption values from 0.3 - 100 %. This data set has recently been analyzed by Luthman et al.<sup>7</sup> using dynamical averaging of polar surfaces. The aim of the present study was to compare dynamical averaging with the use of single conformers and to examine how the treatment of ionizable groups (charged vs. neutral) impacts the statistical quality of the correlations. Furthermore, the question should be addressed if a standard 2D-3D conversion combined with energy minimization is sufficient or if a conformational search for a minimum energy conformation is required.

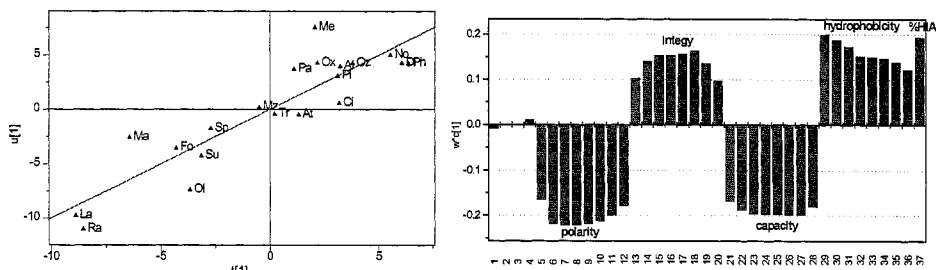
As it can be seen in Table 2, the explained variance  $r^2$  and crossvalidated  $q^2$  (determined by LOO-CV<sup>8</sup>) hardly differ between the various protocols, and the low

complexity of the models (maximum 2 components) allows for a straightforward interpretation. The multivariate modeling by the VolSurf descriptors is hardly influenced by conformational averaging and ionization of charged groups. Also the search for energetic minimum conformations only marginally improved the statistical quality. Taking into account computational efficiency, the protocol applying simple 2D-3D conversion and energy minimization of neutral molecules is fully sufficient and will be the basis for the following analysis.

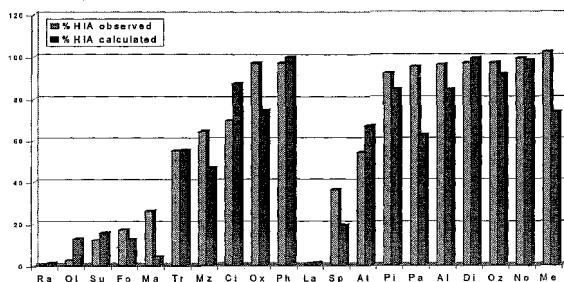
**Table 2.** Comparison of protocols for the correlation of VolSurf parameters with human intestinal absorption.

	neutral	charged
Boltzmann averaging	$q^2 = 0.76$ ( $A = 1$ ); $r^2 = 0.80$	
minimum energy conformation	$q^2 = 0.71$ ( $A = 1$ ); $r^2 = 0.8$	$q^2 = 0.80$ ( $A = 2$ ); $r^2 = 0.90$
2D-3D conversion followed by energy minimization	$q^2 = 0.73$ ( $A = 2$ ); $r^2 = 0.86$	$q^2 = 0.75$ ( $A = 2$ ); $r^2 = 0.89$

From the PLS  $t$ - $u$  score plot (the *logit* transformation was applied to %HIA) and the corresponding partial weights plot (Fig. 2) it can be deduced that hydrophobicity and high integrity moments are positively correlated with human intestinal absorption, whereas polarity and a high concentration of polar interaction sites on the molecular surface are detrimental to absorption. The high loadings of the integrity moments can be tentatively associated with the anisotropy of biological membranes which have to be crossed during absorption. It must be stressed that the interpretation of this model is valid only within the physicochemical property space of this data set. The predictive power of the model is demonstrated in Fig. 3. Applying optimal, distance-based experimental design<sup>9</sup> in principal components space the data set was divided in a training set of 10 compounds (RMS error = 13.4 %) and an external prediction set of the remaining 10 molecules (RMS error = 16.0 %). Summing up, an excellent and easily interpretable model could be obtained with low computational requirements.

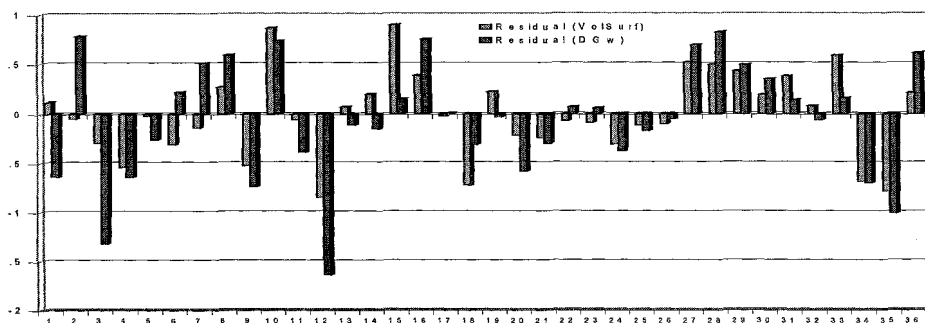


**Fig. 2.** PLS score plot ( $t_1$  vs.  $u_1$ ) and the corresponding partial weights plot for the correlation of VolSurf descriptors with human intestinal absorption (%HIA).



**Fig. 3.** Comparison of observed with calculated %HIA. The external prediction set ranges from lactulose (La) to metoprolol (Me).

In the case of brain-blood partitioning computationally efficient and predictive descriptors are of great importance, because the experimental determination is difficult, lengthy and expensive. For this study a data set was used which had originally been reported by Young<sup>10</sup> and has recently been analyzed via solvation free energy calculations by Lombardo<sup>11</sup> et al. The 30 Young<sup>10</sup> compounds were included in the training set, and the same 6 compounds as in the publication by Lombardo<sup>11</sup> et al. were used as an external prediction set. In contrast to the work by Abraham<sup>12</sup> and Lombardo<sup>11</sup> compounds **3** and **12** were not excluded as outliers. As it can be seen in the plot (Fig. 4) of the residuals of  $\log C_{\text{brain}}/\log C_{\text{blood}}$  (logBB) **12** remains an outlier whereas **3** is well modeled using VolSurf descriptors. The standard error of the model is around 0.5 log units and compares very favorably to the computationally far more expensive calculation of solvation free energies. This also applies to the predictivity of the model as determined by the external prediction set. In addition, the VolSurf descriptors are far superior in interpreting the model in structural terms. As it can be deduced from the PLS score and partial weights plot (data not shown), a good brain penetration can be obtained with a high volume to surface ratio, dispersive interactions and high integrity moments as exemplified by **8**. If a compound is strongly separated in polar and hydrophobic regions with a small integrity moment such as **19**, logBB is greatly diminished.



**Fig. 4.** Comparison of the logBB residuals (difference between measured and calculated values) of the free energy calculations (marked by "DGw") and the VolSurf correlations. The external prediction set is formed by compounds **31** - **36**.

In the last example, the VolSurf descriptors will be used to explain the preference towards renal or hepatic excretion in physicochemical terms. Because drug candidates often fail due to fast elimination via excretion into the bile, the rational modulation of the excretion behavior is of great interest. The correlation of the percentage of renal and hepatic excretion and of the ratio renal vs. hepatic excretion (as published by Fleck and Bräunlich<sup>13</sup>)

with the VolSurf descriptors yielded a two-dimensional model. From the PLS score and partial weights plot of the first component it becomes evident that low molecular weight and high polarity favor renal excretion and high molecular weight and hydrophobicity result in hepatic excretion. This information can then be used to design molecules with modified properties in order to optimize the elimination route.

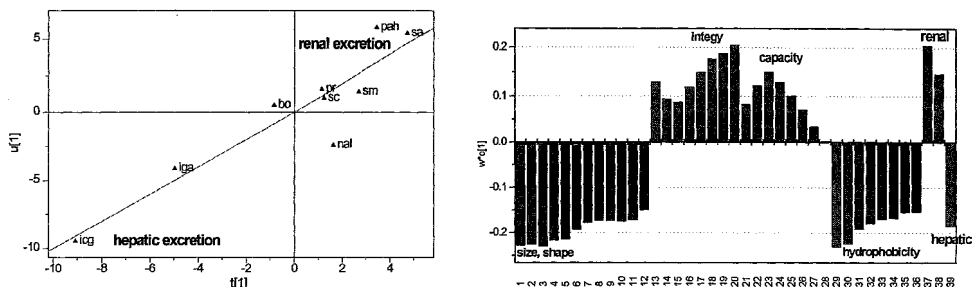


Fig. 5. PLS score plot ( $t_1$  vs.  $u_1$ ) and the corresponding partial weights plot for the correlation of VolSurf descriptors with renal and hepatic excretion of xenobiotics.

## CONCLUSION

Pharmacokinetic properties of a drug often depend on a variety of physicochemical parameters and, therefore, require a multivariate description. The novel VolSurf descriptors quantitatively characterize size, shape, polarity and hydrophobicity as determined with the GRID H<sub>2</sub>O and DRY probes and are independent of an alignment. Because the VolSurf parameters only encode physicochemical properties, they are not suitable if active transport or extensive biotransformation and metabolism are involved. However, if the pharmacokinetic phenomena to be modeled are linked with physicochemical properties, the VolSurf descriptors are ideally suited for lead optimization to explore physicochemical property space by experimental design and to interpret quantitative structure-property relationships in terms of molecular structure. A further application can be envisioned as virtual screen in library design.

## REFERENCES

- GRID v. 16, Molecular Discovery Ltd., West Way House, Elms Parade, Oxford, 1997.
- G. Cruciani, M. Pastor, VolSurf, version 0.0.2, MIA (Multivariate Infometric Analysis), 1998.
- I. T. Jolliffe, *Principal Component Analysis*, Springer Verlag, Ney York, 1986.
- S. Wold, E. Johansson, M. Cocchi, *PLS - Partial Least-Squares Projections to Latent Structures*. In: *3D QSAR in Drug Design*, ESCOM, Leiden, 1993, pp. 523-550.
- S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, *J. Med. Chem.* **1987**, *30*, 1126-1135.
- SIMCA-S, version 6.01, UMETRI AB, Umeå, 1997.
- K. Palm, P. Stenberg, K. Luthman, P. Artursson, *Pharm. Res.* **1997**, *14*, 568-571.
- S. Wold, *Technometrics* **1979**, *20*, 379-405.
- E. Marengo, R. Todeschini, *Chem. Intel. Lab. Syst.* **1992**, *16*, 37-44.
- R.C. Young, R.C. Mitchell, T.H. Brown, C.R. Ganellin, R. Griffiths, M. Jones, K.K. Rana, D. Saunders, I.R. Smith, N.E. Sore, T.J. Wilks, *J. Med. Chem.* **1988**, *31*, 656-671.
- F. Lombardo, J.F. Blake, W.J. Curatolo, *J. Med. Chem.* **1996**, *39*, 4750-4755.
- M.H. Abraham, H.S. Chadha, R.C. Mitchell, *J. Pharm. Sci.* **1994**, *83*, 1257-1268.
- C. Fleck, H. Bräunlich, *Arzneim.-Forsch./Drug Res.* **1990**, *40(II)*, 942-946.

## VALIDATING NOVEL QSAR DESCRIPTORS FOR USE IN DIVERSITY ANALYSIS

Robert D. Clark, Michael Brusati, Robert Jilek, Trevor Heritage  
and Richard D. Cramer

Tripos, Inc.  
1699 S. Hanley Road  
St. Louis, MO 63144

### INTRODUCTION

A descriptor is a mathematical function which maps chemical structures into the set of real numbers or into a real-valued vector. When functions take on single values which characterize a molecule as a whole, they can be classed as one dimensional descriptors, as can substituent parameters used to partition physical chemical effects among the substructures of which the molecules being studied are comprised. Partition coefficients (*e.g.*, logP and ClogP) and molar refractivity (MR) are examples of 1D descriptors. 2D descriptors such as atom pair and substructural fingerprints and connectivity indices explicitly incorporate contributions from molecular connectivity. 3D descriptors, in contrast, include contributions from effects (*e.g.*, through-space interactions) which are dependent on the conformation or the position of a molecule, or both. The most commonly employed 3D descriptors are molecular fields (CoMFA).<sup>1</sup>

Many descriptors have proven themselves useful over the years for delineating quantitative structure/activity relationships (QSARs). With the recent shift in pharmaceutical and agrochemical discovery and development towards combinatorial chemistry and high-throughput screening, it has become necessary to examine the potential usefulness of established QSAR descriptors in light of the subtly different demands of diversity analysis. Where QSAR analysis seeks to define local relationships among a sharply delimited, more or less congeneric set of compounds, diversity analysis seeks to assess how well-dispersed a set of compounds is across a broad region of structural space.

If it is to be useful in quantitating molecular diversity, a descriptor must exhibit a generalized neighborhood property with respect to biochemical properties – *i.e.*, most compounds which look similar when "viewed" through the lens of the descriptor in question should be biochemically similar as well.<sup>2</sup> In other words, proximity in the descriptor space must be a sufficient condition for similarity in biochemical space. Note that there is no requirement that compounds which are proximal in terms of biochemistry map to the same areas in the descriptor space – structural similarity is a *sufficient* condition for biochemical similarity but not a *necessary* condition.

Both UNITY fingerprints<sup>2</sup> and topomeric molecular fields<sup>3,4</sup> have been shown to exhibit good neighborhood behavior, whereas several other QSAR descriptors examined do not consistently do so.<sup>2</sup> Here we describe an extension of the method for determining the range over which a particular descriptor exhibits good neighborhood behavior (the neighborhood radius) and for assessing its statistical significance and report the preliminary results for the validation of molecular holograms,<sup>5</sup> Eigen VAlue (EVA) profiles,<sup>6</sup> and molecular fields obtained using inertial field orientation (IFO-CoMFA).<sup>4</sup>

## DATASETS

The datasets used here are expanded from three of the ten cited in the original neighborhood analysis paper.<sup>2</sup> In particular, the two sets of melatonin antagonists from Garratt *et al.*<sup>7</sup>, which were considered separately in the original, were consolidated into a single set of 19 compounds (2a-j, 6a-h, and 6j). The seven oxazolidinone NK<sub>1</sub> Substance P antagonists drawn from Lewis *et al.* were augmented with the other 14 compounds described therein.<sup>8</sup> Similarly, the six amidoisoxazole endothelin inhibitors<sup>9</sup> excluded from the original neighborhood validation study because they lacked the 3,4-dimethyl-5-amido core structure were included here.

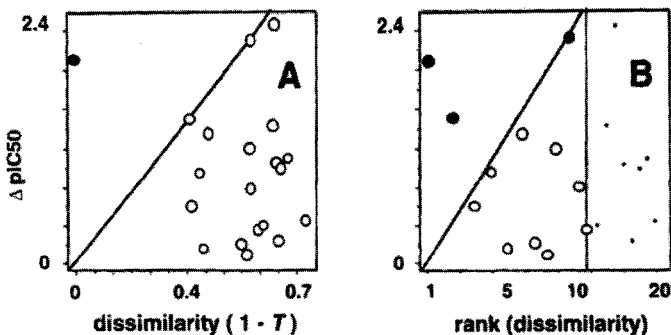
## NEIGHBORHOOD PLOT ANALYSIS

### Rank Transform

As neighborhood validation was originally formulated, pairwise differences in biochemical activity were plotted against the corresponding intercompound distances (or dissimilarities). That diagonal was then identified for which the density of points below and to the right of the diagonal (i.e., in the lower right trapezoid (LRTrap)) was highest. An enhancement and a  $\chi^2$  score were calculated from the actual density of points below the line and the density expected were the points to be evenly distributed across the entire area covered by the plot.<sup>2</sup> If a descriptor exhibits good neighborhood behavior, there will be relatively few instances where small distances in the metric space are associated with large differences in biochemical activity. If so, the upper left triangle (ULT) will be underpopulated and a substantial enhancement (the ratio of the density of points in the LRTrap to an even distribution) will be observed.

Some datasets give misleading statistics when processed in this way, however. Figure 1A is a plot for just such a dataset – in this case, using UNITY fingerprints for the side chains from seven NK<sub>1</sub> antagonists drawn from Lewis *et al.* dataset. A large gap exists between the origin and the second smallest dissimilarity in the dataset, so only one data point shows up in the ULT. As a result, values for enhancement and  $\chi^2$  (1.63 and 4.64, respectively) are obtained which are misleadingly high with respect to the corresponding significance thresholds (1.10 and 3.84). The method can be extended to handle such cases by applying a rank transform to the dissimilarity (or distance) axis (Figure 1B). This eliminates the gap along the ordinate axis, adding points to the ULT. To make the statistical estimates slightly more conservative, points on the line are assigned to the ULT (filled circles in Figure 1).

That limiting diagonal is chosen which has the highest  $\chi^2$  statistic. All points in the LRTrap were counted as "good" (open circles in Figure 1) in the original method, whereas only those directly under the diagonal (in the lower right triangle, or LRT) count in the modified validation method; the datapoints to the right of the vertical are outside the neighborhood radius and are ignored (Figure 1B). This modification makes it possible to analyze datasets which include many quite dissimilar compounds but does not materially affect the conclusions for well-behaved datasets.



**Figure 1.** Neighborhood plots for side-chain fingerprints. (A) Original neighborhood analysis with simple dissimilarity as ordinate. (B) Modified plot of the same data, using the rank of the distance as the ordinate.

Plotting the square root of the rank weights the smallest distances most heavily, as is appropriate. It also gives a good linear "edge" to plots for most datasets and generates radii quantitatively consistent with those obtained using the original method.

### Distribution Factor

In the original formulation of neighborhood analysis,<sup>2</sup> an even spread of data points was used as the reference distribution when calculating the  $\chi^2$  statistic with respect to a random distribution. Subsequent experience has shown that this is often not appropriate. In fact, small distances and dissimilarities necessarily predominate when *all* pairwise differences are considered. As a result, even a random distribution of points can produce a depopulated ULT for some datasets, in which unrealistically high  $\chi^2$  values can be obtained. There is generally an offsetting effect when all pairwise distances for the descriptor are used as well, but this is lost when the rank transform is applied. To counteract such possible distortions in the revised neighborhood analysis calculations, the fraction of the population expected to fall in the ULT is calculated directly for the actual distribution of differences in biochemical activity. This is accomplished by drawing chords through each actual difference in biochemical activity and summing the fractions falling in the upper left triangle (ULT) (Figure 2).

Typical values for the distribution factor range from 13 to 28% versus the 50% distribution factor used previously. These are used to calculate the number of points expected to fall in the LRT ( $\text{LRT}_{\text{exp}}$ ), which yields more realistic  $\chi^2$  values.

A side effect of this correction is that the enhancement statistic is no longer a meaningful statistic. An enrichment can be calculated instead, which is given by:

$$\text{enrichment} = 1 - \text{ULT} / \text{ULT}_{\text{exp}} \quad (1)$$

The upshot for the dataset shown in Figure 1 is that the sample is simply too small to yield an accurate neighborhood radius, and the statistics reflect this fact:  $\chi^2$  is 0.05 instead of 4.64, and enrichment is only 0.047 (vs the enhancement of 1.63 found using the original method<sup>2</sup>). Any enrichment over 0.300 is potentially useful.



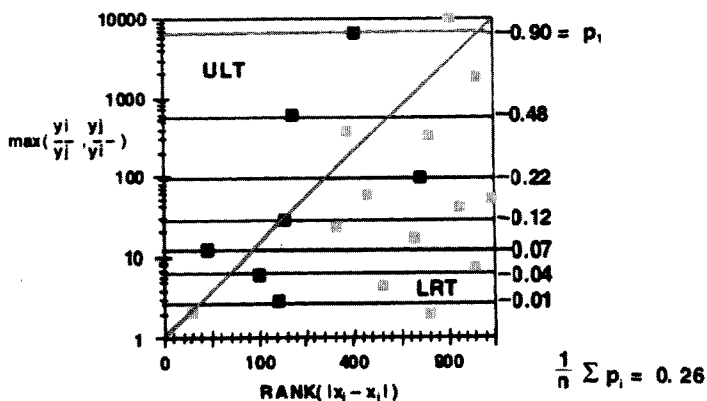


Figure 2. Calculation of the distribution factor. Only the dark symbols have been included in the illustrative summation shown.

Analysis of the full Lewis *et al.* dataset using the 2D fingerprints for each complete molecule, on the other hand, produces well-behaved, "classic" plots with both the original and the refined methods (Figure 3).

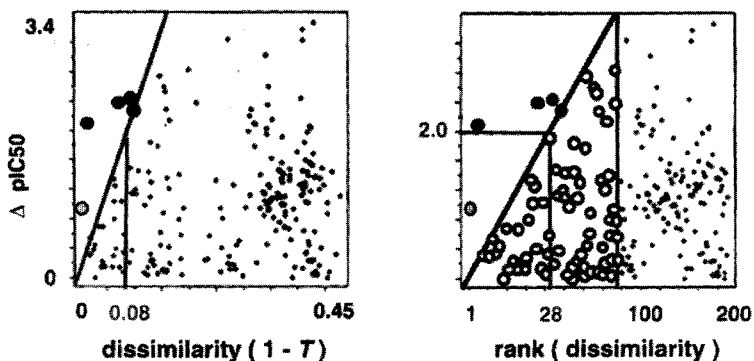


Figure 3. Neighborhood plots for whole-molecule UNITY fingerprints from Lewis *et al.*<sup>8</sup>

This is generally the case – descriptor and dataset combinations which exhibit good distributions using the original method also give good results using the modified approach. Many combinations which were clearly problematic by visual inspection as analyzed originally, however, are handled cleanly by the modified method.

The neighborhood radius (here, 0.08) is back-calculated from a rank for which the limiting difference in biochemical activity is 2 log units. Figure 3 shows that the limiting line drawn from the origin through the corresponding point when the data are graphed with simple dissimilarity as the ordinate is quite reasonable for such well-behaved datasets.

## DESCRIPTORS

A molecular hologram is an indexed, integer-valued vector in which each element is the count of all substructural fragments in a molecule which map to the corresponding index. The mapping index for any particular fragment is given by the modulus of its cyclic redundancy check (CRC) value with respect to the length of the hologram.<sup>2</sup> The length, range of fragment length considered, and criteria for distinguishing fragments – *e.g.*, whether or not connecting bond types are considered – can all be varied.

EVA profiles<sup>6</sup> are obtained by convoluting gaussian envelopes centered about the normal mode vibrational frequencies calculated for the molecule in question using AM1. For the analyses used here, profiles were obtained using a  $\sigma$  of 10 cm<sup>-1</sup> and a sampling interval ( $\delta$ ) of 5 cm<sup>-1</sup> across the frequency range from 0 to 4000 cm<sup>-1</sup>.

Topomeric alignments are done using common substructures,<sup>2</sup> whereas inertial field orientations make use of the principal axes and dipole moments of each molecule.<sup>4</sup> CONCORD® conformations were used as starting points for all compounds except for those from the Lewis *et al.* dataset, for which *S*-axial configurations were used for consistency.

ClogP and CMR were calculated using software from BioByte Inc., Pomona CA. Atom pair and UNITY® fingerprints were calculated using the Selector® module of SYBYL®, as were molecular holograms, EVA profiles and molecular fields. Random descriptors were generated uniformly across the interval from 0 to 100.

## RESULTS

All three datasets were evaluated with respect to a variety of descriptors, with the results shown in Tables 1 and 2. A "good" descriptor will give high values for both enrichment and  $\chi^2$ , and will have a radius large enough with respect to the natural scale of the descriptor to be useful. EVA distances in the Lewis *et al.* dataset, for example, range from 0.35 to 1.0, so a radius of 0.54 gives reasonable resolution. The radius of 0.000 found for CMR when applied to the Krystek dataset is almost significant, but not very useful.

Holograms generally perform very well regardless of how they are calculated, which is reasonable in that they are very high-resolution measures of *structural* similarity. Note that the maximum pairwise distances here range from 32 to 150, depending on length.

**Table 1.** Neighborhood validation of molecular holograms

dataset	hologram length	fragment lengths	connections	radius <sup>1</sup>	enrichment	$\chi^2$
Garratt	83	4 – 7	–	7.4	0.797	10.91
			+	7.5	0.880	15.28
		5 – 10	–	12.9	0.941	17.77
			+	12.5	0.946	19.63
Krystek	97	5 – 9	+	30.2	0.651	16.14
	257	5 – 9	+	30.3	0.516	9.66
Lewis	97	4 – 7	–	28.3	0.599	5.55
		5 – 10	–	66.2	0.465	1.38

<sup>1</sup> Distance in bin counts corresponding to 2 log units difference in biochemical activity.

Trends for fingerprints and topomeric fields mirror those already reported,<sup>2</sup> with the caveat that the compounds in the datasets analyzed here are relatively flexible and so are less than ideal subjects for these metrics. EVA and IFO-CoMFA have good, though not spectacular, statistical profiles; note, however, that both can be used in cases where topomeric CoMFA is inappropriate – *e.g.*, when no common core structure is present.

**Table 2.** Neighborhood validation of other descriptors

descriptor	measure	dataset	radius <sup>1</sup>	enrichment	$\chi^2$
random	Euclidean	Garratt	0.245	0.569	0.37
ClogP	Euclidean	Garratt	0.139	0.179	1.07
		Krystek	0.164	0.198	3.18
		Lewis	0.519	0.320	1.96
CMR	Euclidean	Garratt	0.078	0.112	0.01
		Krystek	0.000	0.411	3.54
		Lewis	0.369	0.086	0.10
UNITY fingerprints	Tanimoto	Garratt	0.022	0.658	8.79
		Krystek	0.024	0.522	6.71
		Lewis	0.079	0.537	3.02
Atom Pairs	Tanimoto	Garratt	0.036	0.597	8.22
		Krystek	0.208	0.106	0.88
		Lewis	0.168	0.630	6.74
EVA	Euclidean	Garratt	0.511	0.445	6.23
		Krystek	0.523	0.560	15.93
		Lewis	0.540	0.676	10.68
Topomeric CoMFA (sterics)	Euclidean	Garratt	79	0.362	5.03
		Lewis	80	0.524	3.43
IFO-CoMFA (sterics)	Euclidean	Garratt	97	0.488	15.82
		Krystek	115	0.488	5.94
		Lewis	153	0.554	3.38

<sup>1</sup> Radii are in units natural to each descriptor, e.g., kcal/mol for molecular fields.

In many respects, these 3D descriptors are complementary to holograms and to each other. In particular, they are more generalized and less structure-specific than holograms or fingerprints, which makes it possible to identify important similarities between structurally distinct compounds. Note, too, the good performance of EVA on the Lewis *et al.* dataset, which is so diverse in structure that molecular fields have a difficult time, as do holograms. Clearly, it can be useful to consider both target and chemistry class when choosing a metric to use for a particular diversity analysis.

## REFERENCES

- H. Kubinyi, ed., *3D QSAR in Drug Design*, ESCOM, Leiden (1993).
- D.E. Patterson, R.D. Cramer, A.M. Ferguson, R.D. Clark and L.E. Weinberger, Neighborhood behavior: a useful concept for validation of molecular diversity descriptors, *J. Med. Chem.* **39**, 3049-3059 (1996).
- R.D. Cramer, R.D. Clark, D.E. Patterson and A.M. Ferguson, Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers, *J. Med. Chem.* **39**, 3060-3069 (1996).
- R.D. Clark, A.M. Ferguson and R.D. Cramer, Bioisosterism and molecular diversity, in: *3D QSAR in Drug Design, Vol. 2*, H. Kubinyi, G. Folkers and Y.C. Martin, eds., Kluwer/ESCOM, Dordrecht (1998).
- W. Tong, D.R. Lewis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage and D.M. Sheehan, Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor, *J. Chem. Inf. Comput. Sci.* **38**, 669-677 (1998).
- A.M. Ferguson, T. Heritage, P. Jonathon, S.E. Pack, L. Phillips, J. Rogan and P.J. Snaith, EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis, *J. Comput.-Aided Mol. Des.* **11**, 143-152 (1997).
- P.J. Garratt, R. Jones and D.A. Tocher, Mapping the melatonin receptor. 3. Design and synthesis of melatonin agonists and antagonists derived from 2-phenyltryptamines, *J. Med. Chem.* **38**, 1132-1139 (1995).
- R.T. Lewis, A.M. Macleod, K.J. Merchant, F. Kelleher, I. Sanderson, R.H. Herbert, M.A. Cascieri, S. Sadowski, R.G. Ball and K. Hoogsteen, Tryptophan-derived NK1 antagonists: conformationally constrained heterocyclic bioisosteres of the ester linkage, *J. Med. Chem.* **38**, 923-933 (1995).
- S.R. Krystek, Jr., J.T. Hunt, P.D. Stein and T.R. Stouch, Three-dimensional quantitative structure-activity relationships of sulfonamide endothelin inhibitors, *J. Med. Chem.* **38**, 659-668 (1995).

**Section IV**  
**Prediction of Ligand-  
Protein Binding**

## STRUCTURAL AND ENERGETIC ASPECTS OF PROTEIN-LIGAND BINDING IN DRUG DESIGN

Gerhard Klebe, Markus Böhm, Frank Dullweber,  
Ulrich Grädler, Holger Gohlke, and Manfred Hendlich

Philipps-University Marburg  
Department of Pharmaceutical Chemistry  
Marbacher Weg 6, D 35032 Marburg, Germany

### Introduction

The interaction of a low-molecular weight ligand with a receptor protein is a process of mutual molecules recognition. This process, first defined by Jean-Marie Lehn in 1973 <sup>1</sup>, serves in biological systems a particular purpose, e.g. an enzymatic transformation, a substance transformation, an allosteric regulation or a specific signal transduction. Drugs are a particular class of low-molecular weight ligands that try to interfere with such processes by means of a specific high-affinity binding to the protein receptor under consideration. They establish their biological function, e.g. as an enzyme inhibitor, an allosteric effector, a receptor agonist or antagonist, a channel blocker or as a competitor in a transportation or transduction process. Prerequisite for specific recognition at the receptor can be associated with a high geometrical complementarity of ligand and binding site and with a strong negative free energy of binding in aqueous solution <sup>2</sup>.

### Knowledge-based Approaches to Protein-Ligand Recognition Principles

Over the last years we have witnessed a dramatic increase in the number of well-resolved protein-ligand complexes. They can be used as a knowledge base to learn about the fundamental principles of how proteins and ligands recognize each other. They provide multiple answers to questions such as: how do ligand-functional groups prefer to interact with particular active-site residues or which molecular building blocks are favorably accommodated in certain active-site cavities? Such queries can only be addressed to the known data if a computerized system is available that allows to retrieve such information. The recently developed ReliBase tool <sup>3,4</sup> makes protein and ligand information simultaneously accessible.

For example, one might be interested in short contacts between protein peptide groups and aromatic moieties in ligands. Amide groups are potent hydrogen-bond forming partners within the plane of the amide bond. A variety of structures can be found where the N-H bond dipole is oriented along the normal on the plane through the ligand's phenyl group. In contrast, perpendicular to the amide plane, the amide bond shows mainly

hydrophobic properties. Accordingly, a slit-type groove, e.g. the opening between two parallel  $\beta$ -sheets, can accommodate aromatic groups of ligands. In other examples, one of the flanking amide groups is replaced by a cluster of neighboring aromatic moieties showing a preferred edge-to-face arrangement among the benzene rings.

Besides the retrieval of recognition patterns between ligand moieties and protein building blocks the database can be used to compile contact preferences between ligand functional groups and protein residues<sup>5</sup>. Docking and de-novo design methods try to predict the putative binding of novel molecules to a given protein binding pocket. This process requires information about possible interaction patterns between functional groups of ligands and active-site amino-acid residues. Ligands usually possess several rotatable bonds, accordingly they can adopt multiple conformations of nearly equal energy. Conformational transitions change the shape of molecules<sup>6,7</sup>. As a consequence their recognition properties are altered. Accordingly, computational approaches to ligand docking and de novo design have to consider molecular flexibility. Information about both, conformational preferences and mutual functional group recognition patterns can be retrieved from crystal structures of protein ligand complexes. The results from these complexes are limited, either in the total number of examples available (presently about 7000) and in the accuracy of the structure determination (resolution mostly beyond resolving individual atomic positions). For this reason the database of small organic crystal structures has been evaluated<sup>5,8,9</sup>, however not without collecting in parallel evidence that results from small molecule data resemble those from protein ligand complexes<sup>5</sup>.

Recognition sites, favorable in space for ligand functional groups to interact with a protein, can be extracted from composite crystal-field environments<sup>10</sup>. These are obtained as composite picture from many crystal packings by superimposing the common functional group together with the positions of every individual contacting group present in all examples. Meanwhile a comprehensive collection of these composite crystal-field environments can be found in IsoStar<sup>9</sup>.

Within the spatial regions indicated in these distributions, sets of discrete interaction centers are generated. These centers are subsequently exploited in the de novo design tool Ludi<sup>11</sup> or in the docking program FlexX<sup>12</sup>. Ludi has its strength in the search of small molecule fragments as initial ideas for possible lead structures. Since FlexX can consider full conformational flexibility, also larger ligands can be docked successfully into the protein binding site to suggest possible leads. Conformational flexibility is considered in FlexX by evaluating conformational library information derived from crystal data. Torsion angles exhibited by common molecular fragments in crystals correspond to conditions adopted in a structured molecular environment. These are similar to those present at the binding site of a protein. After placing the base fragment, FlexX follows an incremental built-up procedure to grow a ligand into the active site of a protein<sup>12</sup>.

### **Computer-based Lead Finding for t-RNA Guanosine Transglycosylase Inhibitors**

Tools such as Ludi and FlexX can be used as alternative strategy to experimental high throughput screening for lead discovery. The latter approach requires a well-established and reliable HTS assay and access to a large database containing prospective lead compounds. The search for inhibitors of t-RNA guanosine transglycosylase (TGT) is an example where neither an appropriate assay nor a sufficiently large database is available to us. However, the crystal structure of this enzyme has been solved to 1.8 Å resolution<sup>13</sup>. TGT plays a key role in shigella dysentery. This is a frequent infection in the third world causing the death of more than 500.000 infants per year<sup>14</sup>. One way of therapy is the administration of antibiotics, however, resulting in a total loss of the entire intestinal flora and rapid resistance is acquired versus the established antibiotics. The infection is induced by shigella bacteria that are closely related to E.coli. They cause rapid inflammation of the

intestinal mucosa and receive their virulence via the transfer of pathogenity coding gens. It has been shown that strong reduction of virulence is achieved through the loss of activity of TGT<sup>15</sup>. The enzyme is involved in quenosine biosynthesis. Quenosine is a modified guanine base that is introduced into t-RNA. For the development of a selective antibiotic the fact is important that quenosine biosynthesis is not essential for E.coli and shigella, however the latter loose pathogenity upon down regulation.

Crystals of the apo-form of TGT could be soaked with preQ<sub>1</sub>, a weak substrate analog inhibitor. To elucidate the outlined therapeutic concept, more potent and selective inhibitors are required. Accordingly, based on the preQ<sub>1</sub> structure we embarked into a computer screening for putative small molecule inhibitors as first ideas for possible leads. Using the program Ludi a variety of ligands is suggested, all with a scoring well in the range of trypsin inhibitors of similar molecular weight proven to actually bind to this serine proteinase<sup>16</sup>. Some of the proposed compounds could be purchased and assayed. They suggest inhibition of TGT. Successful cocrystallization with the enzyme has established binding of 2,3-dihydroxy benzoic acid, one of the Ludi hits suggested to accommodate the guanosine recognition site. The obtained binding geometry of this ligand will be a starting point for a subsequent design cycle to develop larger and more potent inhibitors.

### Scoring of Putative Hits in Lead Finding

Crucial in all virtual computer screening experiments is the relative ranking of the suggested hits. In docking applications, as described above, the binding affinity has to be predicted correctly. This is a free energy quantity composed by an enthalpic and an entropic term<sup>2</sup>. Whereas the former contribution mainly results from interactions between the molecules (including water!) involved, the latter quantity changes with the degree of ordering of the system. In any case, it has to be remembered that only differences in the inventory matter between the common bound state and a situation where all interacting partners are individually solvated.

At best, such a required scoring function is developed from physics resulting in a master equation considering per se all contributing effects. Although being intellectually the most convincing approach, no satisfactory method has yet been reported that is precise enough and at the same time computationally affordable.

More successful and explicitly incorporated into the above-mentioned design tools Ludi and FlexX are scoring functions resulting from regression analyses of experimental data. In such functions a number of empirically derived terms is fitted to a data set of experimental observations<sup>2,17</sup>. Usually the obtained scoring schemes are fast to evaluate and, as long as they are developed on physical concepts, some fundamental understanding with respect to the binding process is provided. However, as common to all regression analyses, the derived scoring function can only be as precise and generally valid as the data used are relevant and complete to consider all contributing and discriminating effects.

At first, this fact calls for precise experimental data to characterize the ligand binding process (s. below). However, a closer inspection of binding modes generated by docking tools such as FlexX<sup>12</sup> or Dock<sup>18</sup>, performed on test cases with experimentally resolved binding modes suggest the following: often enough binding geometries are generated closely approximating experiment however they are ranked higher than other obviously artificial solutions. This refers to a weakness in the scoring function derived only at experimental structures. Accordingly, penalty terms to reject computer-generated artifacts are missing.

One possible way would be the development of selective filters to discard inappropriate binding modes. However, since again these filters learn at arbitrarily selected case studies, their general applicability is in question. As an alternative, we decided to develop a scoring function based on database knowledge. Following the ideas of an inverse

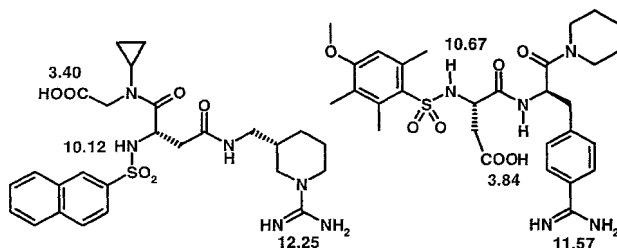
Boltzmann distribution, it is assumed that only those binding modes are favorable that agree to normal distributions of occurrence frequencies among particular interatomic contacts<sup>19</sup>.

For the analysis, contact distances of 1.0 Å up to 8 Å between distinct atom types in a ligand and a protein have been evaluated statistically using ReLiBase. Subsequently, the occurrence frequencies have been translated into statistical potentials. These distance dependent pair potentials have been calibrated to the total distance distribution considering all atom types. Significant deviations to shorter contacts from the mean all-atom distribution are observed for hydrogen-bonding groups whereas preferentially van der Waals contacting groups show reduced frequency and accordingly unfavorable potential at short distances. Besides, we have incorporated for each atom type a solvent accessible surface dependent potential considering ligand and protein to solvent interactions. This potential punishes the exposure of hydrophobic groups to the solvent or of polar functional groups to nonpolar counter parts. On the opposite, it favors mutual contacts between polar groups or tolerates unchanged solvation of polar ligand functional groups carried over from the solvated to the bound state.

The derived scoring function is fast to compute. For a set of test examples with crystallographically determined binding modes all FlexX-generated geometries with small rmsd (with respect to the native binding mode) fall into a narrow window scored as favorable. With increasing geometric deviation also reduced affinity is suggested. This observation gives confidence that also docked geometries where no X-ray reference is known will be ranked as favorable. Hopefully they are reliable enough to describe the actual binding mode.

## Experimental Characterization of the Ligand Binding Process

Nevertheless, as mentioned above, our present understanding of binding modes and the thermodynamics driving ligand binding is still rather scarce. Experimental approaches to learn more about the energetics are based on the temperature dependent evaluation of binding affinity. Assuming a temperature-independent binding enthalpy and entropy over a range of perhaps 40°C van't Hoff plots allow to separate enthalpic and entropic contributions. In such experiments all effects will cancel out that are comparable at the various temperatures. However, the assumed temperature independence will hardly be given<sup>20</sup>. An alternative is isothermal titration calorimetry (ITC)<sup>21</sup>. The heat produced upon binding is directly measured and the shape of the titration curve gives access to the dissociation constant  $K_D$ <sup>22</sup>. Using trypsin and thrombin as model systems, we titrated the binding of different ligands. Important enough, the dissociation constant obtained by ITC corresponds within the experimental errors to  $K_i$  values resulting from photometric assays using chromogenic substrates.



**Figure 1.** Three different pKa values obtained for napsagatran (left) and CRC220 (right).



More difficult to interpret is the heat produced during the isothermal binding process. It contains the binding enthalpy, however, other phenomena involved in the binding process are overlaid. For example, we investigated the binding of napsagatran, a potent thrombin inhibitor from Roche<sup>23</sup>, to trypsin and thrombin. Studying this inhibitor from different buffer solutions, a distinct amount of heat is produced. This effect can be explained by an imposed protonation step. Subsequent potentiometric titrations reveal three titratable groups with different  $pK_a$  values in aqueous solution (Fig. 1). To better characterize the involved protonation step, the ethyl ester derivative of napsagatran has been studied. Titration data show that no comparable protonation step is involved. Accordingly, it has to be concluded that the carboxylate group of napsagatran takes up a proton upon thrombin binding. A related thrombin inhibitor CRC 220, developed by Behring<sup>24</sup>, has been studied. Compared to napsagatran, this inhibitor contains similar functional groups that could become protonated upon binding (Fig. 1). Especially the carboxylate group in the central aspartate moiety is slightly more basic compared to that in napsagatran in aqueous solution. Isothermal titration experiments with CRC 220 show that no protonation step parallels the binding step to thrombin. The deviating behavior of CRC 220 and napsagatran can only be explained once their binding modes are compared in detail. As the crystal structure shows, the carboxylate of napsagatran is placed close to Ser 195 toward the oxyanion hole in thrombin<sup>23</sup>. In contrast, the aspartate in CRC 220 is oriented away from the binding site toward the surrounding solvent environment and likely it is hydrogen-bonded via its anti-lone pair to the NH of Gly 219<sup>24</sup>. Accordingly, on a first glance, its local environment remains rather similar to bulk solvent conditions. In agreement, no protonation of its carboxylate is observed. The local dielectric conditions around the carboxylate in napsagatran are strongly modified upon binding. The partial negatively charged environment shifts the  $pK_a$  substantially, in consequence protonation is observed.

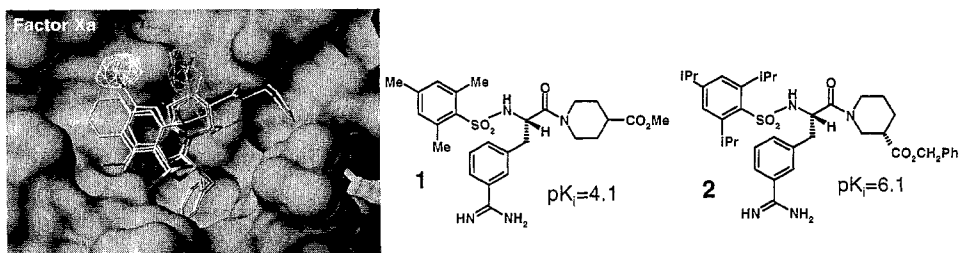
The obtained results are not surprising. Nature extensively exploits this concept of local  $pK_a$  tuning of amino-acid residues to enable particular enzymatic mechanisms. However, the results leave the modeler in a quite uncomfortable situation. The prediction of protonation states is by no means satisfactorily solved. They are already difficult enough to handle under aqueous solution conditions. The described example points to substantial locally induced environment effects. On the long run, they have to be considered in computational methods since, e.g. in a docking experiment, the change from a hydrogen-donor functional group to an acceptor group could completely reverse the binding mode and perturb the relative affinity scoring.

### **Correlation of Ligand Properties with Binding Affinity and Selectivity**

Often enough in relevant drug design projects the 3D structure of the target protein is not available, however, various ligands with deviating binding affinity are known. This discrimination in affinity is related to the capabilities of how these ligands can interact with an – unfortunately unknown – receptor. Accordingly, in order to compare such ligands – at least relative to each other – methods are required that can quantify and rank the putative interaction properties these ligands can experience at a binding site. At best, such methods provide tools to map the correlation results back onto molecular structure in order to elucidate where to alter a particular skeleton to improve binding affinity. This aspect is of special importance if 3D QSAR is used to assist the design of novel affinity-improved ligands<sup>25</sup>.

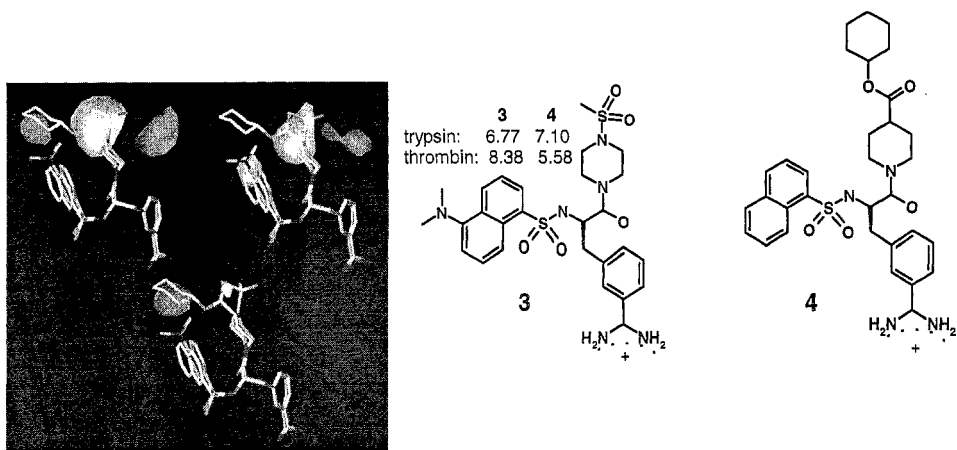
Comparative molecular field analyses are one approach to endure such comparisons. Prerequisite is a reasonable superposition model of the considered molecules that – at best – approximates the actually observed binding modes in the protein. For our study, we wanted to uncouple the conclusions resulting from the correlation model with effects

arising from uncertainties in the superposition model. Accordingly, we selected a data set of inhibitors binding with different affinities to the three related serine proteinases thrombin, trypsin and factor Xa<sup>26</sup>. Since the crystal structures of the three proteins are known, a relative alignment of the ligands can be defined with high reliability.



**Figure 2.** Contribution map of steric properties for factor Xa data. Steric occupancy of the white contoured region increases affinity whereas the gray contoured area should be sterically avoided. The weak binding inhibitor **1** places its COOMe group in the latter unfavorable region whereas **2** occupies the favorable area with its iPr group.

Two different comparative field methods have been applied. In both approaches, molecular property fields are evaluated between a probe atom and each molecule of a data set at the intersections of a regularly spaced grid. The widely used CoMFA method<sup>27</sup> calculates steric and electrostatic properties according to Lennard-Jones and Coulomb potentials. The alternative CoMSIA approach<sup>28</sup> determines molecular similarity considering various physicochemical properties in space. Both methods reveal significant correlation models with high  $q^2$  values and convincing predictive power. CoMSIA could be demonstrated to perform slightly better and to be of higher robustness. However, more important, the resulting contribution maps from the latter approach are much clearer and can be intuitively interpreted to map and pin down those features responsible for affinity and selectivity differences among the superimposed ligands. In Figure 2, the steric properties derived from the factor Xa affinity data are displayed. Areas indicated by white contours correspond to regions where steric occupancy with bulky groups will increase affinity. Areas encompassed by black isopleths should be sterically avoided, otherwise reduced affinity can be expected. Different contour diagrams are revealed for the two other enzymes. The black contour on the right (next to the catalytic center) is sterically unfavored in factor Xa. A favorable region is indicated in the distal pocket. Two molecules, displayed together with the latter map, occupy these regions differently. The less active **1** orients its methyl ester group into the disfavored region whereas the more active **2** fills the white contoured area by its p-isopropyl substituent (Fig. 2).



**Figure 3.** Steric contribution maps for thrombin (upper left), trypsin (upper right) and the selectivity discriminating map (center below). Steric occupation of the gray contoured area in the latter map indicates decreasing affinity towards thrombin. Inhibitor **4** with higher affinity towards trypsin places its terminal cyclohexyl moiety into this area.

To better elucidate the selectivity-discriminating criteria operating in the data set under consideration, we performed an additional analysis with the thrombin and trypsin data. We used the affinity differences between thrombin and trypsin for all 72 inhibitors as dependent property in CoMSIA. The obtained correlation model is of convincing statistical significance and shows some predictive power. Subsequently, we consulted the contribution maps derived from these affinity differences. The steric “selectivity map” (Fig. 3) shows one area to be sterically avoided in order to discriminate selectivity toward enhanced thrombin binding. Fulfilling this criterion, binding affinity toward thrombin will increase relatively to trypsin. Two inhibitors are shown together with this map. The inhibitor **3** possesses higher affinity toward thrombin and leaves the indicated area unoccupied. The inhibitor **4** with higher affinity toward trypsin places its terminal cyclohexyl moiety into this affinity-discriminating area. Additional features can be extracted from the other property maps. Comparing the local shape differences of the thrombin versus trypsin binding site, it is interesting to note that both contours highlighted in the steric and electrostatic selectivity-indicating maps fall next to the 60 loop. This loop occurs as a special characteristic in thrombin, accordingly it is reasonable that areas where affinity between both enzymes is discriminated fall close to this 60 loop. Obviously, contour diagrams derived from a CoMSIA analysis based on binding affinity differences highlight plausible spatial characteristics associated with structural differences responsible for selectivity discrimination.

## REFERENCES

1. J.M. Lehn, Supramolecular Chemistry – Scope and perspectives. molecules, supramolecules, and molecular devices (Nobel Lecture), *Angew. Chem. Int. Ed. Engl.* 27:89 (1988).
2. H.J. Böhm, and G. Klebe, What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs, *Angew. Chem. Int. Ed. Engl.* 35:2588 (1996).
3. K. Hemm, M. Hendlich, and K. Aberer, Constituting a receptor ligand information base from quality-enriched data, in: Proceedings from the Third International Conference on Intelligent Systems for Molecular Biology, ISBN 0-929280-83-0, 170 (1995).
4. <http://www.2.ebi.ac.uk:8081/home.html>
5. G. Klebe, The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands, *J. Mol. Biol.* 237:212 (1994).
6. G. Klebe: Toward a more efficient handling of conformational flexibility in computer-assisted modelling of drug molecules, *Persp. Drug Discov. and Design* 3:85 (1995).

7. G. Klebe, T. and Mietzner, A fast and efficient method to generate biologically relevant conformations, *J. Comput.-Aided Mol. Design* 8:583 (1994).
8. F.A. Allen, O. Kennard, and R. Taylor, Systematic analysis of structural data as a research technique in organic chemistry, *Acc. Chem. Res.* 16:146 (1983).
9. I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor, and M. Verdonk, IsoStar: a library of information about nonbonded interactions, *J. Comput.-Aided Mol. Design* 11:525 (1997).
10. R. Taylor, A. Mullaley, and G.W. Mullier, Use of crystallographic data in searching for isosteric replacements: composite crystal-filed environments of nitro and carbonyl groups, *Pestic. Sci.*, 29:197 (1990).
11. H.J. Böhm, The computer program LUDI: a new method for the de novo design of enzyme inhibitors, *J. Comput.-Aided Mol. Design* 6:61 (1992).
12. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261:470 (1996).
13. C. Romier, K. Reuter, D. Suck and R. Ficner, Crystal structure of tRNA-guanine transglycosylase from *Zymononas mobilis*: RNA modification by base exchange, *EMBO J.* 15:2850 (1996).
14. J.E. Rohde, Selective primary health care: strategies for control of disease in the developing world. XV. Acute diarrhea, *Rev. Infect. Dis.* 6:840 (1984).
15. J.M. Durand, N. Okada, T. Tobe, M. Watari, I. Fukuda, T. Suzuki, N. Nakata, D. Komatsu, M. Yoshikawa and C. Sasakawa, *vacC*, a virulence-associated chromosomal locus of *Shigella flexneri*, is homologous to *Tgt*, a gene encoding tRNA-guanine transglycosylase (Tgt) of *Escherichia coli* K12, *J. Bacteriol.* 176:4627 (1994).
16. H.J. Böhm, LUDI: rule-based automatic design of new substituent for enzyme inhibitors leads, *J. Comput.-Aided Mol. Design* 6:593 (1992).
17. H.J. Böhm, The development of a simple empirical Scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure, *J. Comput.-Aided Mol. Design* 8:243 (1994).
18. I.D. Kuntz, J.M. Blaney, S.J. Oatley, R.L. Langridge, and E.T. Ferrin, A geometric approach to macromolecular-ligand interactions. *J. Mol. Biol.* 161:269 (1982).
19. I. Bahar, and R.L. Jernigan, Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, *J. Mol. Biol.* 266:195 (1977).
20. H. Naghibi, A. Tamura, and J.M. Sturtevant, Significant discrepancies between van't Hoff and calorimetric enthalpies, *Proc. Natl. Acad. Sci. USA* 92:5597 (1995).
21. T. Wisemann, S. Williston, J.F. Brandts, and L.N. Lin, Rapid measurement of binding constants and heat of binding using a new titration calorimeter, *Anal. Biochem.* 179:131 (1989).
22. D.R. Bundle, and B.W. Sikurskjold, Determination of accurate thermodynamics of binding by titration calorimetry, *Methods Enzym.* 247:288 (1994).
23. K. Hilpert, J. Ackermann, D.W. Banner, A. Gust, K. Gubernator, P. Hadváry, L. Labler, K. Müller, G. Schmid, T.B. Tschopp, and H. van de Waterbeemd, Design and synthesis of potent and highly selective thrombin inhibitors, *J. Med. Chem.* 37:3889 (1994).
24. M. Reers, R. Koschinsky, G. Dickneite, D. Hoffmann, J. Czech, and W. Stüber, Synthesis and characterisation of novel thrombin inhibitors based on 4-aminidophenylalanine, *J. Enzyme Inhib.* 9:61 (1995).
25. G. Klebe, Comparative molecular similarity indices analysis: CoMSIA, *Persp. Drug Discov. Design* 12:87 (1998).
26. M. Böhm, J. Stürzebecher, and G. Klebe: 3D-QSAR analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin and factor Xa, *submitted* (1998).
27. R.D. Cramer III, D.E. Patterson, and J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110:5959 (1988).
28. G. Klebe, U. Abraham and T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 37:4130 (1994).

# USE OF MD-DERIVED SHAPE DESCRIPTORS AS A NOVEL WAY TO PREDICT THE IN VIVO ACTIVITY OF FLEXIBLE MOLECULES

## The Case of New Immunosuppressive Peptides

**Abdelaziz YASRI\***, Michel Kaczorek and Roger LAHANA

<sup>1</sup>Synt:em, Parc Scientifique Georges Besse, 30000 Nîmes, France

and **Gérard Grassy**

Centre de Biochimie Structurale, UMR CNRS 9955, INSERM U414, Université Montpellier  
I, 15 avenue Charles Flahault, F-34060 Montpellier, France

and **Roland Buelow**

Sangstat Medical Corporation, Menlo Park, California

In a first report, we used the « In Silico Screening » rational design for the identification of a new immunosuppressive peptides. The molecule predicted to be best, coded as RDP1258, displayed an immunosuppressive activity approximately 1000 times higher than the lead compound: 30% of mice heart allografts survived for more than 100 days, with a dose 80 times lower than that of the lead compound.

Therapy with the rationally designed peptides described here also resulted in upregulation of HO-1 activity in vivo which was shown to inhibit several immune effector functions. However, a cyclized RDP1258 peptide while being able to inhibit HO-1 in vitro, had no effect on HO-1 expression in vivo. These data suggest that flexibility of the peptides is indeed required for immunomodulatory activity in vivo.

In this study we have examined the correlation between the in vitro and in vivo data for the immunosuppressor peptide RDB1258. Our strategy was based on the use of a virtual combinatorial library combined to molecular dynamic simulations. The diversity of the built library was assessed by using the conformational autocorrelation method associated with cluster analysis method. A set of 9 different peptidic sequences were subjected to a molecular dynamics simulation study. The comparisons of the conformational spaces via the conformational autocorrelation method combined to the principal component analysis of the derived peptides to RDP1258 suggested that some of them are predicted to be in vivo active peptides, whereas some other peptides are predicted inactive.

---

\* To whom correspondence should be addressed

## Introduction

In a first study, we successfully applied the In Silico Screening method to the rational design of the immunosuppressive peptide RDP1258<sup>1</sup>. It was based on a peptide derived from the  $\alpha$ 1 helix of HLA-B2702<sup>2,3,4</sup>. This peptide was shown to prolong heart graft survival in mice.

Recently, characterization of L- and D-isomers of 2702.75-84 derived peptides resulted in the identification of hemeoxygenase-1 (HO-1)<sup>5</sup>, also known as hsp32, as a receptor for these immunosuppressive peptides. The peptides inhibited HO-1 activity in vitro. In vivo administration of the peptide resulted in upregulation of hemeoxygenase activity, a phenomenon common to all HO-1 inhibitors. Upregulation of hemeoxygenase was shown to inhibit several immune effector functions including cell mediated cytotoxicity and cell proliferation, and to prolong mouse heart allograft survival<sup>6</sup>. Upregulation of HO-1 was also shown to inhibit an inflammatory response, while inhibition of HO-1 increased such a response<sup>7,8</sup>. Therapy with the rationally designed peptides described here also resulted in upregulation of HO-1 activity in vivo (Iyer and Buelow, unpublished results). However, a cyclized RDP1258 peptide while being able to inhibit HO-1 in vitro, had no effect on HO-1 expression in vivo. These data suggest that flexibility of the peptides is indeed required for immunomodulatory activity in vivo.

The rational design of the peptides described in the first study was based on activity in a mouse heart allograft transplantation model. In fact, upregulation of HO-1 following administration of 2702.75-84-derived peptides was only demonstrated upon completion of the described rational approach. The observation that the designed peptides are more potent inhibitors of HO-1 in vitro and more potent inducers of HO-1 expression in vivo, support the hypothesis that the peptides immunomodulatory activity is due to an interaction with HO-1. Upregulation of HO activity may therefore provide novel strategies to modulate immune responses in vivo.

The aim of this work was to design new peptides based on the structure of RDP1258 peptide to study the interaction between Allotrap and HO-1 and to set up a predictive system for the in vivo activity of Allotrap. Some peptides derived from RDP1258 were designed by mutating systematically the sequence of RDP1258 from L to D forms. This was achieved by building a virtual combinatorial library and evaluating its diversity. Molecular dynamics simulations were applied to the selected peptides in order to compare their explored conformational spaces.

## **Materials and Methods**

- *Molecular Modeling of the Combinatorial Set*

The combinatorial explosion was performed using Combex (Syntem, Nîmes, France). All molecules were generated using the SMILES convention, and then converted into a 3D structure using Corina (Oxford Molecular Group, Oxford, UK). This was performed by mutating the nine positions of RDP1258 systematically to D forms. This has resulted in 512 different structures.

- *Vectorial Description of the Combinatorial Set*

Conformational description of the generated structures was performed by using the conformational autocorrelation method implemented in TSAR V3 software (Oxford Molecular Group, Oxford, UK). Each 3-D structure is associated to an ACV (Autocorrelation Vector).

- *Clustering of the Combinatorial Set*

We applied cluster analysis and principal component analysis methods implemented in the TSAR software to classify the generated structures (i.e., their associated ACVs). The barycenter structures of the obtained clusters were extracted and compared with the structure of RDP1258. The distances between the average structures of each cluster and RDP1258 structure were evaluated by using the nearest neighbor method implemented in TSAR software.

- ***Molecular Dynamics Simulation Protocols***

The MD simulations, performed using AMBER 4.11, used 1050 ps in duration for each peptide solvated with a box water with periodic conditions. The dielectric constant was set to the unit. The temperature of the system was first gradually increased from 10 to 300 K, during a time period of 20 ps and a constant temperature, during simulation, was maintained at  $300 \pm 10$  K by coupling to an external bath with a relaxation time of 0.1 ps. The chosen time step was 1 fs. The computational time was approximately 0.5 hour per ps. A 10 angströms residue-based cutoff was used for all non-bonded interactions. The non-bonded pair list was updated every 10 fs and the coordinates were collected every 1 ps during the trajectories resulting in a set of 1050 conformations for each trajectory. In all trajectories, no constraints were applied to the atoms. No cross terms were used in the energy expression.

- ***Trajectory analysis***

Each conformation is associated with a 3D-ACV<sup>9</sup>. A set of 3D-ACVs is calculated for each MD run, and then processed using multivariate statistics. In order to be able to compare the multiple 3D-ACVs representing the trajectories of the set of molecules to analyze, a principal components analysis is applied to each of these multiple 3D-ACVs in order to reduce the dimensionality of the data set to a smaller number (in our case, a mere 2D space) and also to project on to a common space all the trajectories of all the molecules<sup>10</sup>. In this reduced space, each molecule is represented by a set of dots (i.e., their conformations throughout the MD simulation), which is called its conformational space. Molecules can then be compared one to each other in terms of conformational spaces. These comparisons were validated by calculating the conformational radius of gyration during the trajectories.



## Results and Discussions

A set of 512 different structures were generated from RDP1258 by mutating the positions 1, 2, 3, 4, 5, 6, 7, 8, and 10 systematically from L to D forms. The molecular diversity of the generated structures were assessed by the 3-D ACV description combined to multivariate statistical analysis.

### • *Structural Diversity*

Cluster analysis was performed on the whole combinatorial set, at 25 % of maximal amalgamation distance in the conformation sample, we could easily distinguish 19 clusters. If the barycentres of each cluster are calculated, then the main conformational diversity obtained by the combinatorial building from RDP1258 reduce to a smaller number of structures or data. Table I shows the sequences of the calculated barycentres.

**Table I.** Amnio acid sequences of cluster barycentres obtained from cluster analysis. NLE: Norleucin; capital letters: L-amino acid; small letters: D-amino acid

Barycentre	Sequence									
RDP1258	R	NLE	NLE	NLE	R	NLE	NLE	NLE	G	Y
BC_sym-1	R	NLE	NLE	NLE	R	NLE	NLE	nle	G	y
BC_sym-2	r	NLE	NLE	NLE	R	NLE	NLE	NLE	G	Y
BC_sym-3	r	NLE	NLE	NLE	R	NLE	NLE	nle	G	y
BC_sym-4	R	NLE	NLE	NLE	R	NLE	NLE	NLE	G	y
BC_sym-5	R	nle	NLE	nle	R	nle	NLE	nle	G	Y
BC_sym-6	r	nle	NLE	nle	R	nle	NLE	nle	G	Y
BC_sym-7	r	nle	NLE	nle	R	nle	NLE	nle	G	y
BC_sym-8	r	NLE	NLE	nle	R	nle	NLE	nle	G	Y
BC_sym-9	r	nle	NLE	nle	R	nle	NLE	NLE	G	y
BC_sym-10	R	nle	nle	nle	R	nle	NLE	nle	G	y
BC_sym-11	R	nle	nle	nle	R	nle	nle	nle	G	Y
BC_sym-12	R	NLE	nle	NLE	r	nle	NLE	NLE	G	Y
BC_sym-13	r	NLE	NLE	NLE	R	NLE	NLE	nle	G	y
BC_sym-14	r	nle	nle	nle	r	nle	nle	nle	G	y
BC_sym-15	r	NLE	NLE	NLE	r	NLE	NLE	NLE	G	Y
BC_sym-16	R	nle	nle	nle	R	nle	nle	nle	G	Y
BC_sym-17	r	nle	nle	nle	R	nle	nle	nle	G	Y
BC_sym-18	R	nle	nle	nle	R	nle	nle	nle	G	y
BC_sym-19	r	NLE	NLE	NLE	R	NLE	NLE	NLE	G	y

The structural similarity between RDP1258 and the obtained barycenters was evaluated by their distances in the hyperspace of the whole 3-D ACV components. This was done by nearest neighbor method as implemented in TSAR software. We chose to keep as similar

structures all the barycenters with a distance to RDP1258 structure lower than 4 units. This resulted in 8 peptides whose sequences are summarized in table II.

**Table II.** Amino acid sequences of the most nearest peptides to RDP1258

Peptide	Sequence									
RDP1258	R	NLE	NLE	NLE	R	NLE	NLE	NLE	G	Y
BC_sym-1	R	NLE	NLE	NLE	R	NLE	NLE	nle	G	y
BC_sym-3	r	NLE	NLE	NLE	R	NLE	NLE	nle	G	y
BC_sym-7	r	nle	NLE	nle	R	nle	NLE	nle	G	y
BC_sym-11	R	nle	nle	nle	R	nle	nle	nle	G	Y
BC_sym-15	r	NLE	NLE	NLE	r	NLE	NLE	NLE	G	Y
BC_sym-16	R	nle	nle	nle	R	nle	nle	nle	G	Y
BC_sym-18	R	nle	nle	nle	R	nle	nle	nle	G	y
BC_sym-19	r	NLE	NLE	NLE	R	NLE	NLE	NLE	G	y

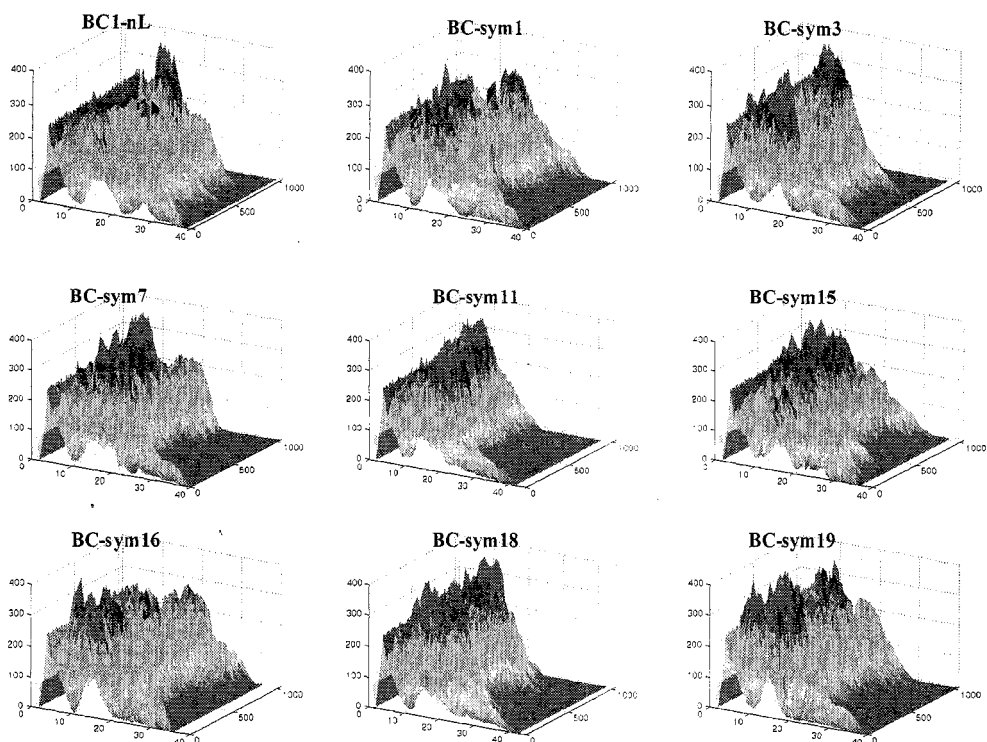
### • *Molecular Dynamics Simulation*

The selected 8 peptides were subjected to molecular dynamics simulation in order to compare their dynamic behavior to the in vivo active peptide, RDP1258. These comparisons were performed via the conformational autocorrelation method combined with principal component analysis as well as by the molecular radius of gyration calculated during the trajectories.

### *Global Dynamic Behavior*

A simple examination of the 3-D ACV profile of the different trajectories (figure 1), readily reveals differences or similarities between the trajectories.

Within the same trajectory, the profile of the 3-D ACV may undergo considerable changes reflecting the conformational diversity explored by the peptide. Some trajectories are represented with 3-D ACV profiles undergoing reversible change (trajectories RDP1258, BC-sym3, BC-sym7, BC-sym15, and BC-sym16). The peptides simulated in these trajectories fluctuate between different conformations and may therefore be more flexible molecules. On the other hand, some trajectories show profile of their 3-D ACV representing an irreversible change from the half-time of the simulations (trajectories BC-sym11 and BC-sym18).

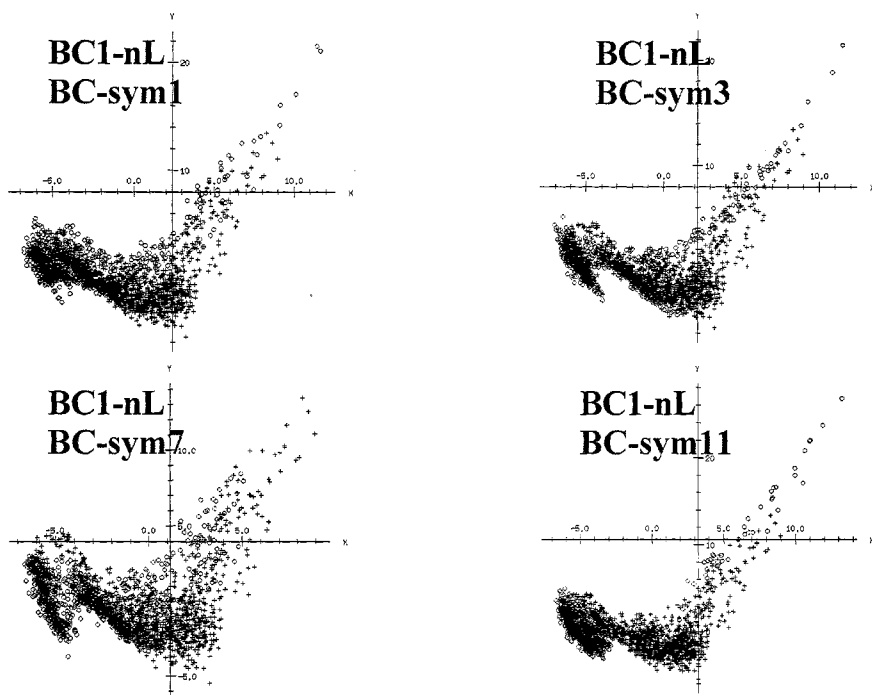


**Figure 1.** Three-dimensional plots of 3-D ACVs associated with the conformations generated during RDP1258 and its peptide derivatives. X axis: Interatomic Distance ( $\text{\AA}$ ), Y axis: Simulation Time (ps), and Z axis: Atom Pairs.

### *Conformational Space Comparison*

Principal component analysis was performed on the 3-D ACVs associated with the conformations visited in the trajectory of RDP1258 peptide. The principal components (PCs) are arranged in the order of their contributions to the total variance, i.e. the first (PC1) contributes by 61.2 % to the total variance, the second (PC2) 18.4 %, and the third (PC3) 8.9 %. Figure 2-1-A shows 3-D ACVs associated with RDP1258 trajectory projected into the plane defined by the first two PCs. Because PC1 and PC2 together contribute 79.6 % to the total variance, Figure 2-1-A must give a fairly accurate representation of the nature of the conformational space explored by RDP1258 peptide. This trajectory starts on the right-hand

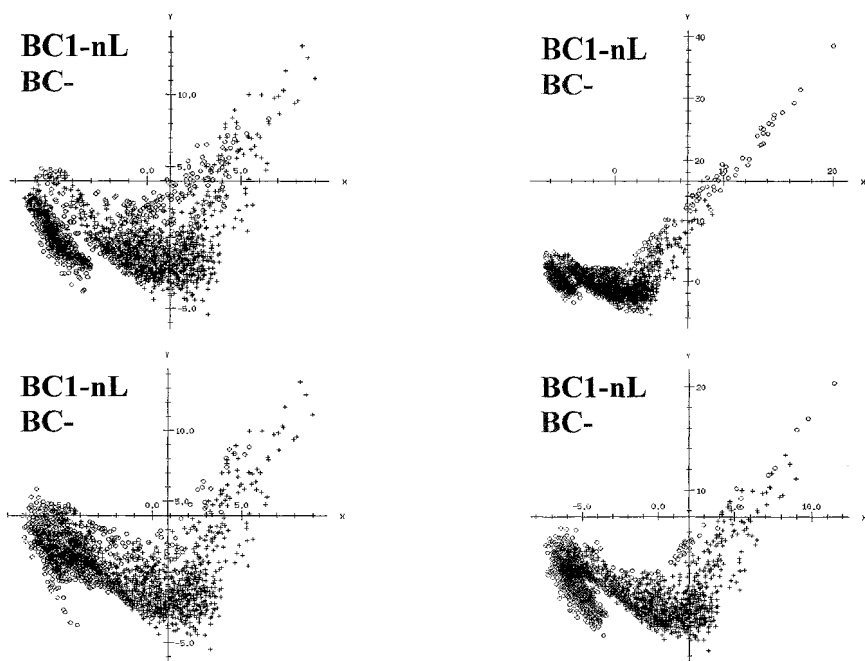
side of the plane and ends on the left-hand side. A clear clustering effect is visible in the middle of the plane. Before the end of the trajectory, the molecule leaves the cluster and starts to form another one by taking an intermediate path. This transition is illustrated by the decrease of the molecular radius of gyration.



**Figure 2-1.** Conformational Space analyses of RDP1258 (black cross) and its derivatives (blue cross).

The projection of the cloud of points associated with trajectories of the RDP1258 derivatives on the principal plane defined by the two first principal components is reported in figures 2-1 and 2-2. Globally, all the peptides followed the same trajectory as RDP1258. Some of them show conformational space which resemble RDP1258's one (trajectories of BC-sym3 and BC-sym7) suggesting similar dynamic behavior. As RDP1258 peptide, BC-sym3 and BC-sym7 peptides show reversible transitions between stretched and compacted conformations according to their radius of gyration. The similarity of trajectories to BC-nL is partially

existing for BC-sym15, BC-sym18 and BC-sym19 (figure 2-2). On the other hand, trajectories of BC-sym11 and BC-sym16 follow antagonist pathway by going rapidly to more compacted conformations as shown by the radius of gyration (figure 2-1, figure 2-2).



**Figure 2-2.** Conformational Space analyses of RDP1258 (black cross) and its derivatives (blue cross). Snapshots correspond to the radius of gyration.

These comparisons clearly show that BC-sym3 and BC-sym7 peptides explore a larger conformational spaces by presenting a high flexibility allowing them to make transitions between different conformations and reproducing consequently the dynamic behavior of RDP1258. On the other hand, BC-sym11 and BC-sym16 peptides present reduced flexibility and an antagonist dynamic behavior to RDP1258.

Through the comparison of conformational space of RDP1258 and its derivatives peptides, BC-sym1, BC-sym3, and BC-sym7 are predicted to be in vivo active peptides, whereas BC-sym11 and BC-sym16 peptides are predicted inactive. BC-sym15, BC-sym18, and BC-sym19 peptides could show intermediate in vivo activity.

## Conclusions

In this study we have examined the effect of stereoisomeric point mutations on the dynamic behavior of the immunosuppressor peptide RDP1258. Our strategy was based on the use of the virtual combinatorial library combined to molecular dynamic simulations.

The diversity of the built library was assessed by using the conformational autocorrelation method associated with cluster analysis method. A set of 9 different peptidic sequences (RDP1258, BC\_sym-1, BC\_sym-3, BC\_sym-7, BC\_sym-11, BC\_sym-15, BC\_sym-16, BC\_sym-18, and BC\_sym-19 ) were subjected to a molecular dynamics simulation study. The comparisons of the conformational spaces via the conformational autocorrelation method combined to the principal component analysis of the derived peptides to RDP1258 suggested that BC-sym1, BC-sym3, and BC-sym7 are predicted to be in vivo active peptides, whereas BC-sym11 and BC-sym16 peptides are predicted inactive. BC-sym15, BC-sym18, and BC-sym19 peptides could exhibit intermediate in vivo activity.

## References

1. G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h and R. Buelow, "In Silico Screening" applied to the rational design of immunosuppressive compounds. *Nature Biotech.*, 748-752 (1998).
2. L. Gao, J. Woo, and R. Buelow, Both L- and D-isomers of HLA class I heavy chain derived peptides prolong heart allograft survival in mice. *Heart and Lung Transplantation*, 15: 78-87 (1996).
3. R. Buelow, P. Veyron, C. Clayberger, P. Pouletty, and J.L. Touraine, Prolongation of skin allograft survival in mice following administration of Allotrap. *Transplantation*, 59:455-460 (1995).
4. J. Woo, S. Iyer, M.C. Cornejo, L. Gao, C. Cuturi, Soullou and R. Buelow, Immunosuppression by HLA Class I heavy chain (amino acid 75 to 84). derived peptides is independent of binding to Hsc70. *Transplantation*, 64:1460-1467 (1997).

5. S. Iyer, J. Woo, M.C. Cornejo, L. Gao, W. McCoubrey, M. Maines, and R. Buelow, Characterization and biological significance of immunosuppressive peptide D2702.75-84 (E>V) binding protein: Isolation of heme oxygenase. *J Biol Chem*, 273: 2692-2697 (1998).
6. J. Woo, S. Iyer, M.C. Cornejo, N. Mori, L. Gao, and R. Buelow, Stress induced immunosuppression: Inhibition of Cellular immune effector functions following overexpression of heat shock protein 32. *Transplantation Immunology*, (1998) in press.
7. D. Willis, A.R. Moore, R. Frederick, and D.A. Willoughby, Heme oxygenase: A novel target for the modulation of the inflammatory response. *Nature Med.*, 2: 87-90 (1996).
8. M. Laniado-Schwartzmann, N.G. Abraham, M. Conners, M.W. Dunn, R.D. Levere, and A. Kappas, Heme oxygenase induction with attenuation of experimentally induced corneal inflammation. *Biochem. Pharmacol.*, 53(8):1069-1075 (1997).
9. P. Broto, G. Moreau, and Van C. Dycke, Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem. - Chim. Theor.*, 19:61-70 (1984).
10. A. Yasri, L. Chiche, J. Haiech, and G. Grassy, Rational choice of molecular dynamics simulation parameters through the use of conformational autocorrelation 3-D Method. Application to Calmoduline flexibility study. *Protein Engineering*, 9(11): 959-976 (1996).

# A VIEW ON AFFINITY AND SELECTIVITY OF NONPEPTIDIC MATRIX METALLOPROTEINASE INHIBITORS FROM THE PERSPECTIVE OF LIGANDS AND TARGET

Hans Matter and Wilfried Schwab

**Hoechst Marion Roussel**

Chemical Research, G 838

D-65926 Frankfurt am Main, Germany

## INTRODUCTION

The destruction of articular cartilage is a major pathological event in Osteoarthritis (<sup>1</sup>), ultimately leading to the loss of joint function. Proteoglycan aggregates (*aggrecan*) are the preferred cartilage components for proteolytic attack under pathological conditions. Different cleavage sites for MMP-3 and MMP-8 have been identified at the interglobular aggrecan region. These enzymes belong to the family of matrix metalloproteinases (MMPs) - zinc endopeptidases involved in tissue remodeling and turnover of cartilage and bone. In the pathological case the degenerative potential of MMPs against components of the extracellular matrix is not longer controlled by specific tissue inhibitors. Thus MMPs are attractive targets for the treatment of arthritis and tumor progression.

While structure-based design is focussed on protein-ligand interactions, it does not always lead to predictive models. In contrast, 3D-QSAR models with acceptable statistical parameters do not necessarily reflect the topological features of the binding site. In this study we successfully combined both approaches to understand biological activity and selectivity of 90 nonpeptidic MMP inhibitors (<sup>2</sup>). The availability of MMP-3 and MMP-8 x-ray structures (<sup>3,4</sup>) led to the design of rigid 1,2,3,4-tetrahydroisoquinoline derivatives with appropriate functional groups complementary to the S1' pocket and hydroxamates or carboxylates as Zn<sup>2+</sup> binding groups. Subsequently various 3D-QSAR models identified binding regions, where sterical, electronical or hydrophobic effects play a dominant role in protein-ligand interaction. In addition to this ligands' view, a technique based on PCA of multivariate GRID descriptors (<sup>5</sup>) uncovered major differences of both protein binding sites (<sup>6</sup>). Those results led to a consistent picture allowing further prediction of novel, selective inhibitors.

## 3D-QSAR FOR MMP-3 and MMP-8 AFFINITY

For a reliable alignment, a reference compound was manually docked into MMP-8 and minimized, while treating sidechains within 5 Å as flexible. Automatic docking with a genetic algorithm (FlexiDock <sup>7</sup>) produced similar results. All other compounds were superimposed onto this template and minimized using a rigid protein. This alignment produced consistent

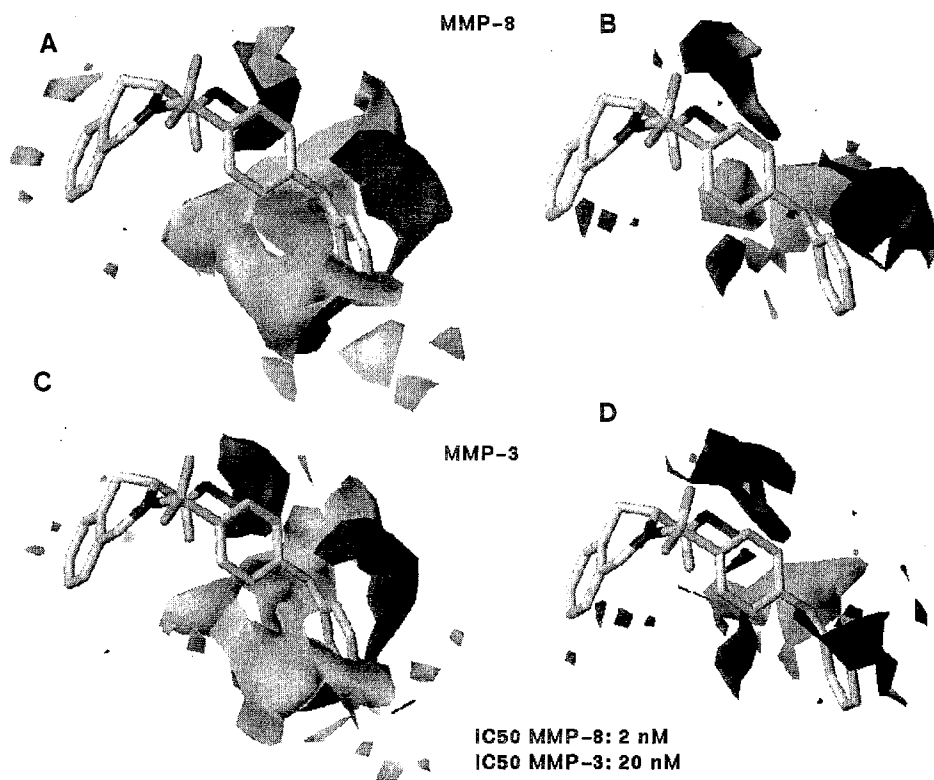


models (CoMFA<sup>8</sup>, CoMSIA<sup>9</sup>, GRID/Golpe), which were confirmed by statistical methods and interpretation in terms of binding site topologies. Finally the binding mode was validated by a 1.7 Å X-ray structure of a reference compound in complex with MMP-8 (<sup>2</sup>).

**Table 1.** Summary of 3D-QSAR models for MMP-3 and MMP-8 affinity<sup>a)</sup>

	q <sup>2</sup>	SD Comp	r <sup>2</sup>	Validation
<b>MMP-8:</b>				
CoMFA (2A)	0.569	0.685	5	0.905 2 CV , Randomize, Grid Var.
CoMFA (1A)	0.516	0.726	5	0.911
CoMSIA (2A)	0.478	0.763	7	0.924 2 CV , Randomize
CoMSIA (1A)	0.447	0.786	7	0.924
<b>MMP-3:</b>				
CoMFA (2A)	0.563	0.629	6	0.944 2 CV , Randomize, Grid Var.
CoMFA (1A)	0.432	0.717	5	0.917 2 CV , Randomize
CoMSIA (2A)	0.413	0.738	8	0.957 2 CV , Randomize
CoMSIA (1A)	0.382	0.757	8	0.954
GOLPE_FFD	0.795	0.413	5	0.967 LTO, 5RG
GOLPE_SRD	0.789	0.419	5	0.964 LTO, 5RG

<sup>a)</sup> q<sup>2</sup>: crossval. r<sup>2</sup> using leave-one-out; SD: standard dev. of error in leave-one-out; Comp.: optimal number of components; r<sup>2</sup>: non-crossval. regression coeff.; Validation 2 CV: crossval. using 2 random groups 100 times; Randomize: randomization of y-block; Grid Var.: shifting the alignment within fixed grid box; LTO: leave-two-out; 5RG: crossval. using 5 random groups 20 times.



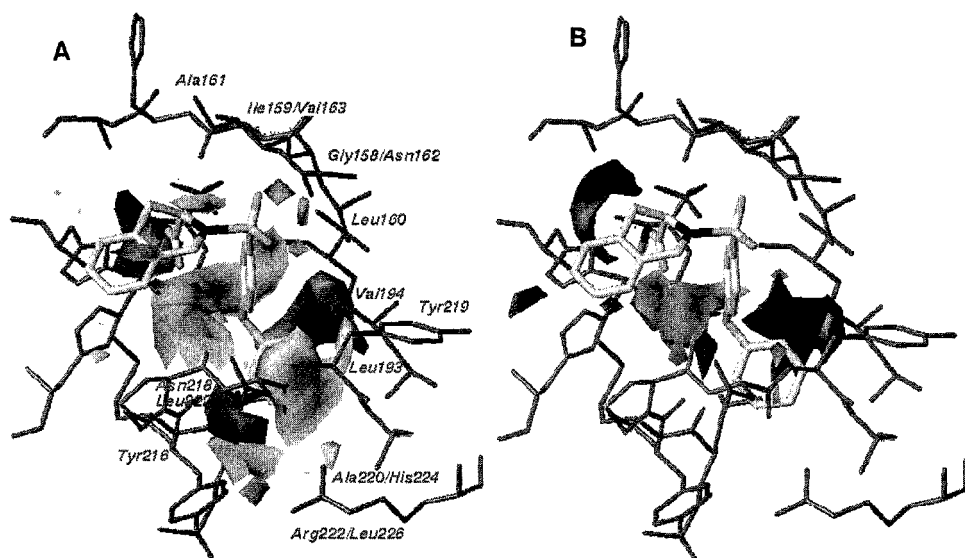
**Figure 1.** CoMFA steric and electrostatic *std\*coeff* fields (< 85 %, > 15 % contribution) for MMP-8 (A,B) and MMP-3 (C,D) with a potent MMP-3/8 inhibitor.

A CoMFA model for MMP-8 affinity with an  $r^2(\text{cv})$  of 0.569 (5 comp.) and an  $r^2$  of 0.905 (Table 1, CoMFA (2A)) was obtained. Changing the grid spacing from 2 to 1 Å produced similar results. Moreover, we investigated the effect of the alignment relative to the grid by moving all compounds in x,y and z direction. The resulting  $r^2(\text{cv})$  values show a minor dependence on the orientation (0.42 to 0.58). Randomizing biological activities revealed the significance of the original model: The mean  $r^2(\text{cv})$  for 50 trials is -0.13 (SD 0.11). Multiple PLS analyses with two randomly chosen cross-validation groups and original activities lead to only a slightly lower mean  $r^2(\text{cv})$  value of 0.438 (SD: 0.08), suggesting a stable, predictive model. Similar results are observed with CoMSIA (table 1).

CoMFA and CoMSIA models for MMP-3 affinity produced models of similar significance (only 85 compounds, table 1), the same validation techniques were used to support the finding of stable and predictive models, like a CoMFA model with an  $r^2(\text{cv})$  of 0.563 (6 comp.) and an  $r^2$  of 0.944.

### COMPARISON WITH RECEPTOR TOPOLOGY

The steric and electrostatic  $\text{std}^*\text{coeff}$  fields for MMP-8 (fig. 1A, B) and MMP-3 (fig. 1C, D) are similar, they are displayed with a potent inhibitor containing a hydroxamate and a biphenylether as main binding elements ( $\text{IC}_{50}$  MMP-8: 2 nM; MMP-3: 20 nM). For steric fields (fig. 1A, C) dark contours are related to favourable steric bulk, while light grey contours indicate regions, where bulk will lower the activity. In the electrostatic maps, dark contours represent regions, where positive charge is favourable.



**Figure 2.** Comparison of CoMFA steric (A) and electrostatic (B)  $\text{std}^*\text{coeff}$  fields to experimental MMP-8 binding site topologies. Differing residues in MMP-3 are indicated.

These models correspond to experimental binding topologies, as obvious from fig. 2 with CoMFA derived steric (A) and electrostatic (B) contour maps mapped onto MMP-8. The residue numbering refers to MMP-8, mutations in MMP-3 are indicated. As results are similar to MMP-3, the question arises, how to explain experimental selectivities.

Enhancing steric bulk in S1' increases affinity for both targets. A dark contour at the distal phenyl ring highlights a hydrophobic cleft formed by Tyr219, Leu193 and Val194, which is filled with water in MMP-8 x-ray structures, when S1' is not completely occupied. Unfavourable steric interactions at the S1' entrance are indicated by light grey contours, while the steric requirements at the S1' bottom (Arg222) are obvious. The preferred geometry for zinc-binding can be deduced, showing that the optimal oxygens distance for zinc coordination is better realized in hydroxamates. Finally the decrease of activity, when inverting the C3 chirality, is indicated by a light grey region at the upper side of the tetrahydroisoquinoline.

Additional 3D-QSAR models were generated using interaction energies from ligands to a phenolic OH probe (GRID) with a GOLPE<sup>(10)</sup> variable selection. The effect of individual variables on model predictivity based on a FFD design matrix points exactly to relevant variables (GOLPE\_FFD, table 1). Moreover a method for grouping descriptors into regions of neighboring 3D variables with similar statistical and chemical information was applied to enhance chemical relevance of results<sup>(11)</sup>. This "smart region definition" (SRD) procedure works by extracting a subset of highly informative X variables and partitioning the space around the molecules among them (GOLPE\_SRD, table 1). Identified regions, containing single pieces of information, are used at a later stage for the FFD based variable selection. The final PLS model after SRD contained 1049 from initially 25740 variables, a  $r^2(\text{cv})$  of 0.789 (5 comp.) and an  $r^2$  of 0.964 was obtained. This model was validated by leave-two-out ( $r^2(\text{cv})$ : 0.787) and 5 crossvalidation groups ( $r^2(\text{cv})$  mean: 0.743). A model without SRD led to comparable results (table 1), both models provide consistent insights into favourable interactions, which complement information obtained by interpreting CoMFA fields.

## UNDERSTANDING LIGAND SELECTIVITY

Although previous models can explain MMP-8 and MMP-3 affinity, no relevant information was obtained for selectivity. However, 3D protein structures provide extremely valuable input, which we used to extract important ligand-protein interactions and selectivity regions. 3D structures for MMP-8 (*Ijap*) and MMP-3 (*Isln*) were superimposed using an iterative procedure (rmsd of 0.41 Å for C $\alpha$  of 117 of 157 residues), their binding sites were characterized by interaction energies to functional groups using GRID. This matrix was then analysed using PCA on only favourable interactions.

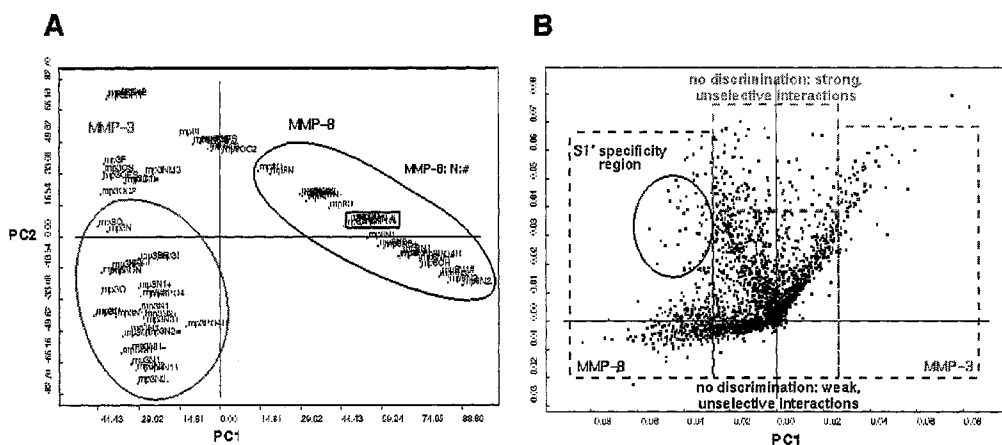
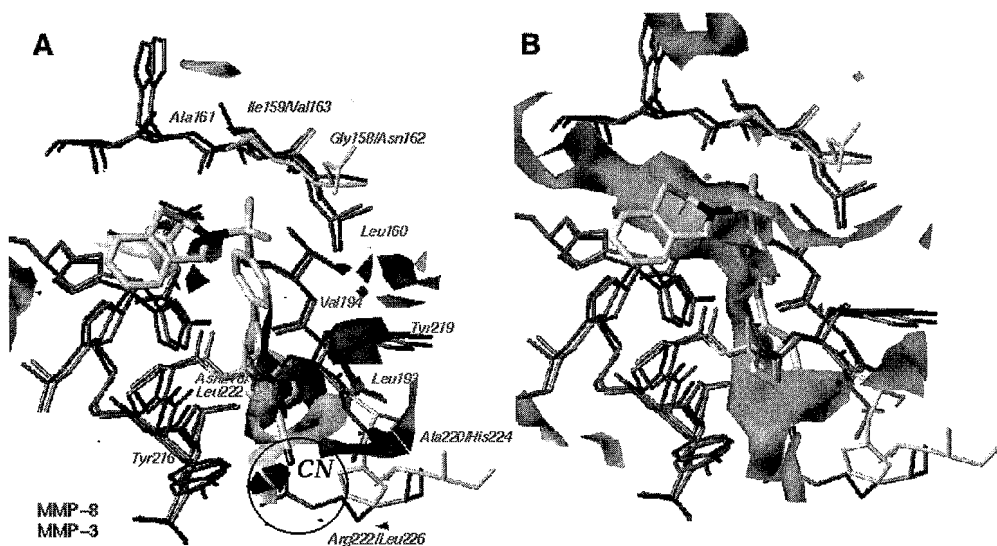


Figure 3. 2D score (left) and loadings (right) plot of PC1 versus PC2 for the final PCA model using grid interaction energies between various probes and both targets.

A significant two component model results for 38 probes, the first PC explains 33.5 % of the variance, the second 27.8 %. The score plot (fig. 3, left) represents objects in the X matrix - interactions of a GRID probe with a target. The clustering of objects into two groups indicates that PC1 discriminates between both proteins (MMP-3 negative PC1 scores, *mp3*; MMP-8 positive PC1 scores, *mp8*), while PC2 is related to non-selective ligand-protein interactions, ranking the probes by their ability to interact with common binding site regions (negative PC2 scores refer to stronger interactions).

Variables with high absolute PC1 loadings (fig. 3, right) indicate binding regions with different interaction behaviour. The greater the horizontal spread of variables in PC1 is (fig. 3, right), the more relevant this variable is for discrimination. Several regions in the binding site can be identified: On the left the MMP-8 selectivity region with high PC1 and PC2 scores points to strong, selective interactions (fig. 3, right). These variables circled in fig. 3 are located in the S1' pocket (fig. 4A). Similar variables in fig. 3 (right) point to less dominant MMP-3 selectivity regions. In contrast, low values for PC1 and PC2 indicate regions with weak, unselective interactions, while high PC2, but low PC1 values refer to strong, unselective interactions (fig. 4B). For designing selective compounds, it is preferable to use chemical groups with higher absolute PC1 scores.

In fig. 4A, 3D loadings contour maps are indicating selectivity regions, where appropriate substituents would increase the desired property. Selectivity regions for MMP-8 are indicated by dark contours, while substitutions in light grey regions would improve selectivity towards MMP-3. For interpretation a selective MMP-8 inhibitor is shown in fig. 4 with a biphenylether S1' moiety and a p-cyano substituent at the distal ring (IC<sub>50</sub> MMP-8: 10 nM; MMP-3: 1000 nM).



**Figure 4.** 3D contour map of PC1 (A) and PC2 (B) loadings for the PCA model. PC1 highlights selectivity regions (dark: MMP-8, grey: MMP-3), PC2 explains affinity regions.

The CN group directly points to a dark MMP-8 selectivity region at the bottom of the S1' pocket close to Arg222. This preference is reflected by the position of a N:# (sp nitrogen with lone pair) probe in fig. 3 (right). Thus a discrimination between MMP-8 and MMP-3 can be achieved by adequate placement of functional groups in this and related regions according

to the ranking of functionalities in fig. 3 (left). In MMP-3, Arg222 is replaced by Leu226 and the S1' pocket is not occluded, which corresponds to substitutions in some inhibitors. Other highlighted S1' regions also suggest the chemical significance of this model to explain selective protein-ligand interactions.

When comparing sequences for both MMPs, not only the bottom of the S1' pocket differs, but also Ile159 and Gly158 are replaced by Val and Asn, and Asn218 and Ala220 of the upper rim of S1' by Leu and His. As the sidechain of the Asn218-Leu222 replacement points to the outside of S1', this is not a major selectivity regions, which corresponds to the 3D contour maps. In contrast, the imidazole ring of His224 forms a part of the S1' pocket, changing the binding requirements for ligands, which is reflected by additional dark contours.

In fig. 4B the regions for unselective strong ligand recognition are shown as contour maps (positive loadings from fig. 3, right), revealing that the S1' pocket plus the region left to Ala161 (Pro-Leu-Gly-NHOH binding region, unprimed P1-P2-P3) are of high importance for MMP-affinity. Those maps are in agreement with all results obtained from CoMFA and CoMSIA studies (see above), leading to a consistent picture explaining both affinity and selectivity of MMP-3 and MMP-8 inhibitors.

## CONCLUSIONS

Protein-ligand interactions are difficult to describe, there is no single approach resulting in a complete picture of all forces. Some useful insights to this problem for MMP inhibitors are presented here: Our study combines structure-based design with 3D-QSAR to understand MMP-3 and MMP-8 affinity. As hydroxamates lead to more potent inhibitors than other zinc-binding groups, the latter require additional features for compensation. Our final 3D-QSAR results are not only able to reveal the optimal zinc-binding geometry, but also the optimal S1' complementarity for MMP-3 or -8 inhibitors. Moreover a recently described method provided further understanding of different experimentally observed ligand selectivities. Detailed SAR information for these inhibitors is obtained, which is in agreement with all experimental data for binding site topologies, and thus provide clear guidelines and activity predictions for designing and optimizing MMP-3/8 inhibitors in related series.

## REFERENCES

- <sup>1</sup> Murphy, G.; Hembry, R.M. *J. Rheumatol.*, **1992**, *19*, 61-64.
- <sup>2</sup> Matter, H.; Schwab, W.; Barbier, D.; Billen, G.; Haase, B.; Neises, B., Schudok, M.; Thorwart, W.; Schreuder, H.; Brachvogel, V.; Lönze, P. *J. Med. Chem.* **1999**, submitted.
- <sup>3</sup> Stams, T.; Spurlino, J.C.; Smith, D.L.; Wahl, R.C.; Ho, T.F.; Qoronfleh, M.W.; Banks, T.M.; Rubin, B. *Nature Struct. Biol.* **1994**, *1*, 119-123.
- <sup>4</sup> Grams, F.; Reinemer, P.; Powers, J.; Kleine, T.; Pieper, M.; Tschesche, H.; Huber, R.; Bode, W. *Eur. J. Biochem.* **1995**, *228*, 830-841.
- <sup>5</sup> Goodford, P.J. *J. Med. Chem.* **1985**, *28*, 849-857.
- <sup>6</sup> Pastor, M.; Cruciani, G. *J. Med. Chem.* **1995**, *38*, 4637-4647.
- <sup>7</sup> Jones, G.; Willett, P.; Glen, R.C. *J. Mol. Biol.* **1995**, *254*, 43-53.
- <sup>8</sup> Cramer III, R.D.; Patterson, D.E.; Bunce, J.E. *J. Am. Chem. Soc.*, **1988**, *110*, 5959-5967.
- <sup>9</sup> Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130-4146.
- <sup>10</sup> Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant.-Struct.-Act. Relat.* **1993**, *12*, 9-20.
- <sup>11</sup> Pastor, M.; Cruciani, G.; Clementi, S. *J. Med. Chem.* **1997**, *40*, 1455-1464.

## ON THE USE OF SCRF METHODS IN DRUG DESIGN STUDIES

Modesto Orozco<sup>1</sup>, Carles Colominas<sup>1</sup>, Xavier Barril<sup>1</sup>, and  
F. Javier Luque<sup>2</sup>

<sup>1</sup> Departament de Bioquímica i Biologia Molecular. Facultat de Química. Universitat de Barcelona. Martí i Franquès 1. Barcelona 08028. Spain

<sup>2</sup> Departament de Físicoquímica. Facultat de Farmàcia. Universitat de Barcelona. Avgda Diagonal sn. Barcelona 08028. Spain

### INTRODUCTION

Self consistent reaction field (SCRF) methods have been largely used to examine solvent effects in chemical interactions. These methods<sup>1</sup> are designed to determine solvation free energy, which is the reversible work necessary to transfer a molecule from gas phase to solution (considering the same reference states, typically 1M). In SCRF methods such a work is computed (see Eq. 1) as the addition of three elemental contributions: i) the work necessary to build up the solute cavity in the solvent (cavitation term), ii) the work needed to generate the uncharged solute in the pre-formed cavity (van der Waals term), and iii) the work spent in generating the solute charge distribution in solution (the electrostatic term).

$$\Delta G^{\text{solv}} = \Delta G^{\text{cav}} + \Delta G^{\text{vW}} + \Delta G^{\text{ele}} \quad (1)$$

The steric contributions (cavitation and van der Waals) can be easily represented by means of empirical relationships with the surface, volume, or other size-related property of molecules.<sup>1</sup> The electrostatic contribution can be determined following different algorithms, all of them being based on the theory of polarizable fluids.<sup>1</sup> One of the most rigorous methods for the computation of  $\Delta G^{\text{ele}}$  was developed by Miertus, Scrocco and Tomasi (MST<sup>1,2</sup>), and has been successfully used for the study of different systems in solution.<sup>1,2</sup> Current versions of the MST method combined with suitable algorithms for the determination of the steric term leads to estimates of the free energy of solvation with errors below 1 kcal/mol for different solvents including water (see Figure 1).<sup>3-5</sup>

According to the MST algorithm,  $\Delta G^{\text{ele}}$  is determined within the Quantum Mechanical (QM) framework as shown in Eq. 2, where the solvent-adapted wavefunction of the solute is determined by solving a non-linear pseudo-Schrödinger equation (see Eq. 3). The perturbational operator  $\hat{V}_R$  is determined by solving Laplace equation with suitable boundary conditions (see Eq. 4).

$$\Delta G^{\text{ele}} = \langle \Psi^{\text{sol}} | \hat{H}^0 + \frac{1}{2} \hat{V}_R | \Psi^{\text{sol}} \rangle - \langle \Psi^0 | \hat{H}^0 | \Psi^0 \rangle \quad (2)$$

where the indexes sol and 0 refer to solution and gas phase, and  $V_R$  stands for the perturbational operator representing the solvent reaction field generated by the solute charge distribution.

$$(\hat{H}^0 + \hat{V}_R) \Psi = E \Psi \quad (3)$$

$$\sigma_m = -\frac{\epsilon-1}{4\pi\epsilon} \left( \frac{\delta(V_\sigma + V_\rho)}{\delta\mathbf{n}} \right)_m \quad (4)$$

where  $\epsilon$  is the dielectric constant of the solvent,  $\mathbf{n}$  is the vector normal to the surface element  $m$ , and the indexes  $\sigma$  and  $\rho$  refer to solvent and solute.

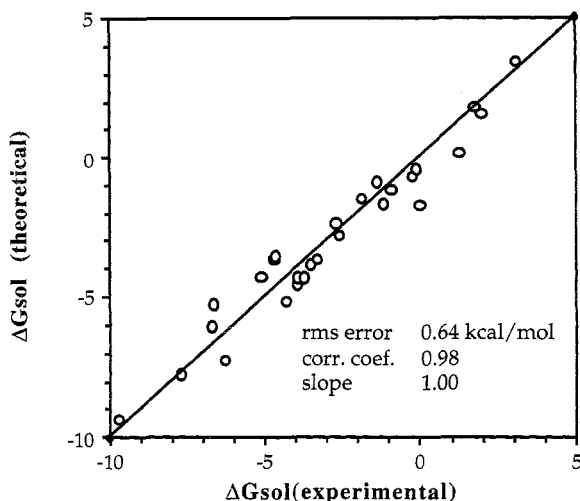


Figure 1. Correlation between MST 6-31G(d) and experimental free energies of hydration.

In this paper we will present new ideas on the use of the MST method in molecular modeling studies. We will present firstly the use of a modified version of MST in docking studies in condensed phases. Secondly, we will show a new method, based on the MST formalism for the fragmental description of the solvation properties of molecules.

## DOCKING IN CONDENSED PHASES

The determination of the best possible arrangement of two molecules (for instance a drug and a receptor) is of major importance in molecular modeling studies. This justifies the tremendous effort focused on the development of docking algorithms, which are expected to determine the best fitting between two complementary molecules in the absence of experimental data. Docking programs generally incorporate Monte Carlo (MC) or Molecular Dynamics (MD) algorithms, which are used to sample extensively the inter-molecular configurational space.

Docking programs are very valuable in molecular modeling studies, and provide often very reasonable results. However, they have two obvious shortcomings arising

from: i) the total or partial neglect of intramolecular contributions to binding, and ii) the total or partial neglect of solvent effects. The neglect of the effect of polar solvents like water in binding can lead to erroneous results. For instance, Figure 2 represents the H-bond dimerization of two formic acid molecules. High level QM calculations<sup>6</sup> indicate that the interaction is favorable in the gas phase ( $\Delta G = -2.5$  kcal/mol, at the G2 level). It is clear that any docking program will find that the structure at the right of Figure 2 is the most stable configuration of the formic acid dimer. However, when the same reaction is studied in aqueous solution, a free energy of dimerization of +5.3 kcal/mol (from G2, SCRF and MC-FEP calculations) is found.<sup>6</sup> This demonstrates that H-bond dimerization of formic acid does not occur in water.

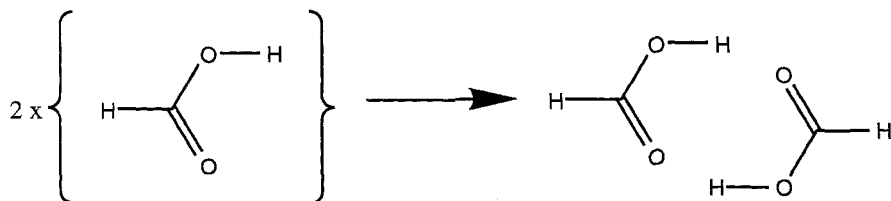


Figure 2. H-bond dimerization of formic acid.

In order to obtain suitable samplings of the configurational space accessible to dimers in solution we have developed a Monte Carlo-continuum model based on the MST algorithm and in classical force-fields (used to compute the inter-molecular energies). The method couples a Metropolis-Monte Carlo technique with a fast *quasi*-classical version of the MST algorithm, where in order to increase the computational efficiency, the electrostatic contribution to solvation is computed using Eq. 5 for each configuration.<sup>7</sup>

$$\Delta G_{\text{ele}} = \frac{1}{2} \sum_i \sum_m \frac{Q_i^0 Q_m^{\text{sol}}}{R_{im}} \quad (5)$$

where indexes *i* and *m* refers to atoms and to small surface elements of the solute cavity, and indexes 0 and sol refers to the values of the charges in gas phase and solution.

The MC-MST method allows us to obtain a complete sampling of the configurational space of dimers in gas phase or any solvent with higher efficiency than discrete methods which expend a lot of CPU time sampling solvent movements. The MC-MST method can be used with single or multiple copies strategies. The first strategy needs shorter equilibration periods, but the later guarantees a better and less biased sampling. The "multiple copies" approach is based in parallel MC runs using 20 copies of one monomer which are placed randomly around a central monomer. Each copy is allowed to interact with the central monomer and solvent, but not with other copies.

As an example we analyzed the configurational space of 4-oxo-pyrimidine dimer in gas phase and aqueous solution ( $T=298\text{K}$ , 1M in both cases). Simulations were run using a multiple copy approach (20 copies) for a total of 200000 configurations, both in gas phase and aqueous solution. ESP and ESPF charges were used to represent the electrostatic potential of the solutes, and empirical Lennard Jones parameters are used to represent their van der Waals properties.

In the gas phase the most populated configurations are those corresponding to a double H-bond, while in solution such configurations are low populated. This is clear from inspection of Figure 3 which corresponds to the last snapshots of the multiple copies runs in gas phase and solution. It is also clear in the density maps shown in Figure 4, which represents the contours corresponding to regions of the space of large probability (15 times that expected for a 1M solution) to find a 4-oxo-pyrimidine



molecule (for a common reference system defined by the other monomer). It is clear that such changes are related to the disturbing effect of water which makes more difficult the formation of solute-solute H-bonds.

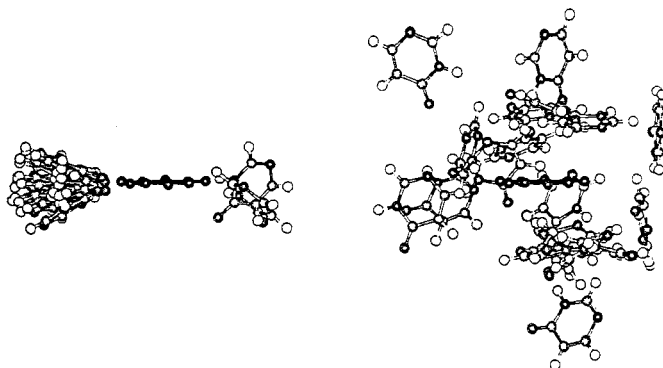


Figure 3. Representation of the last snapshot of the MC-MST simulation in the gas phase (left) and aqueous solution (right).

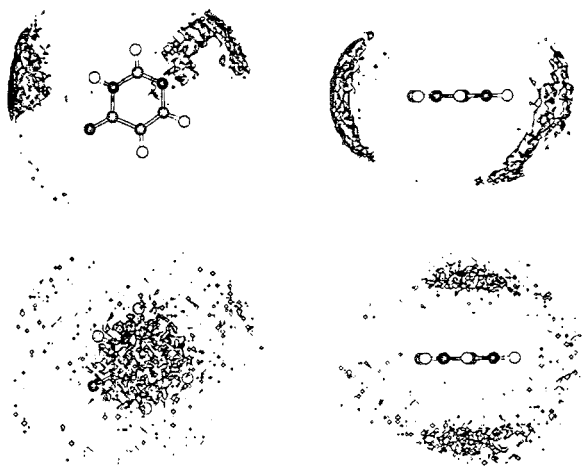


Figure 4. Representation of the regions of large probability (15 times over the background) to find a 4-oxo-pyrimidine molecule. A common reference system defined by the central molecule has been used. Results in top of Figure are in gas phase, and those in the bottom refer to aqueous solution.

## FRACTIONAL REPRESENTATION OF SOLVATION

The determination of the the solvation, and transfer free energies of molecules is of major importance in drug design. This has led to the development of different approaches for the determination of solvation/transfer properties of molecules.<sup>8</sup> However, few methods allow the partition of solvation/transfer free energies into molecular fragments. Such information is very important for the determination of the hydrophobic pattern of molecules which is known to play a key role for a proper drug-receptor binding.

We have recently developed a rigorous QM approach based on the MST algorithm which allows the partition of the total free energy of solvation into surface elements, which can be then grouped into molecular subunits. The method is based on the use of a first order perturbational treatment of the basic MST equations.<sup>7</sup> Accordingly, the electrostatic contribution to the free energy of solvation can be computed as shown in Eq. 6.

$$\Delta G_{ele} = \langle \Psi^0 | \frac{1}{2} V^{sol}(\rho^{sol}) | \Psi^0 \rangle \quad (6)$$

Eq. 6 allows for the rigorous partition of the electrostatic free energy of solvation in surface elements (M) as shown in Eq. 7. Calculation of fractional contributions to the total free energy of solvation is then simple (see Eq. 8) since the steric contribution is directly related to molecular surface areas. Furthermore, fractional contributions to transfer free energies can be also computed using Eq. 9.

$$\Delta G_{ele} = \sum_{j=1}^M \Delta G_{ele}^j = \frac{1}{2} \sum_{j=1}^M q_j^{sol} \langle \Psi^0 | \frac{1}{|r_j - r|} | \Psi^0 \rangle \quad (7)$$

$$\Delta G_{sol}^i = \sum_{i=1}^N \Delta G_{cav}^i + \sum_{i=1}^N \Delta G_{vw}^i + \sum_{i=1}^N \Delta G_{ele}^i \quad (8)$$

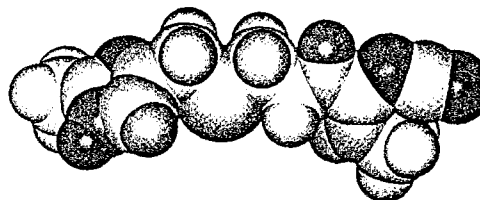
$$\Delta G_{transfer}^{\omega \rightarrow \tau} = \sum_{i=1}^N \Delta \Delta G_{cav}^i + \sum_{i=1}^N \Delta \Delta G_{vw}^i + \sum_{i=1}^N \Delta \Delta G_{ele}^i \quad (9)$$

where

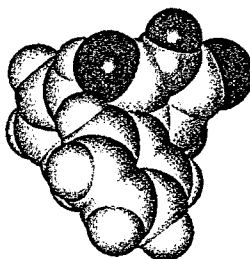
$$\Delta \Delta G = \Delta G(\tau) - \Delta G(\omega)$$

The method can be used in combination with any QM approach, and a quasi-classical version of Eq. 7 has been also developed which allow a very fast calculation of the hydrophobic/hydrophilic pattern of molecules.

The use of the *fractional-MST* method allows us to obtain hydrophobic/hydrophilic profiles like those shown in Figure 5. This type of information is very useful to determine the most polar/apolar regions of molecules, as well as to detect changes in hydrophobicity/hydrophilicity in a given region of the space due to changes in other regions of the space.



Cimetidine



Phenytoin

Figure 5. Fractional contributions to the free energy of hydration of cimetidine and phenytoin. The darker the color, the larger the contribution to  $\Delta G_{\text{hyd}}$ .

## REFERENCES

1. J.Tomasi and M.Persico. Molecular Interactions in Solution: An overview of Methods based on continuous distributions of solvent. *Chem. Rev.* 94: 2027 (1994).
2. S.Miertus, E.Scrocco and J.Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* 55: 117 (1981).
3. M.Orozco, M.Bachs and F.J.Luque. Development of optimized MST/SCRF methods for semiempirical calculations. *J.Comput.Chem.* 16:563 (1995).
4. F.J.Luque, Y.Zhang, C.Alemán, M.Bachs, J.Gao and M.Orozco. Solvent effects in chloroform solution: parametrization of the MST/SCRF continuum model. *J.Phys.Chem.* 100: 4269 (1996).
5. F.J.Luque, M.Bachs, C.Alemán and M.Orozco. Extension of MST/SCRF method to organic solvents: ab initio and semiempirical parametrization for neutral solutes in  $\text{CCl}_4$ . *J.Comput.Chem.* 17: 806 (1996).
6. C.Colominas, J.Teixidó, J.Cemeli, F.J.Luque and M.Orozco. Dimerization of carboxylic acids: reliability of theoretical calculations and the effect of solvent. *J.Phys.Chem.B.* 12: 2269 (1998).
7. F.J.Luque, J.M.Bofill and M.Orozco. New strategies to incorporate the solvent polarization in SCRF and FEP simulations. *J.Chem.Phys.* 107: 1291 (1997).
8. Y.C.Martin. *Quantitative Drug Design. A Critical Introduction*. Marcel Decker. New York 1982.

# 3D-QSAR STUDY OF 1,4-DIHYDROPYRIDINES REVEALS DISTINCT MOLECULAR REQUIREMENTS OF THEIR BINDING SITE IN THE RESTING AND THE INACTIVATED STATE OF VOLTAGE-GATED CALCIUM CHANNELS

Klaus-Jürgen Schleifer, Edith Tot and Hans-Dieter Höltje

Heinrich-Heine-University Düsseldorf, Institute for Pharmaceutical Chemistry, Universitätsstrasse 1, D-40225 Düsseldorf, Germany

## INTRODUCTION

Voltage-gated calcium channels (VGCC) are transmembrane proteins that mediate the calcium influx in response to membrane depolarization and thereby initiate cellular activities such as secretion, contraction, and gene expression. According to pharmacological and electrophysiological results they may be divided into the distinct L-, N-, P/Q-, R-, and T-type subfamilies. While all VGCC are composed of the pore-forming  $\alpha_1$  subunits, the disulfide-linked  $\alpha_2\delta$  subunits and the intracellular  $\beta$  subunits, only the skeletal muscle L-type channel has an additional transmembrane  $\gamma$  subunit. A second special feature of L-type channels is their unique reaction to the calcium entry blockers such as 1,4-dihydropyridines (DHP), phenylalkylamines and benzothiazepines that are therapeutically used against hypertension, angina pectoris and supraventricular arrhythmias, and the exceptional DHP channel activators (Bay k 8644, RS30026, CGP 28392 or Bay y 5959). However it is not the unique L-type  $\gamma$  subunit which is the physiological target of these compounds, but specific regions of the  $\alpha_1$  subunit.

Regardless of antagonistic or agonistic effect, the receptor affinity of the modulators is dependent of the actual channel mode. While at polarized membranes (-70 mV to -90 mV) the channels are in the closed resting state, depolarization (starting at -30 mV for L-type VGCC) leads to an oscillation between the opened and the inactivated state. All DHP derivatives show lower affinity to their binding site in the resting state in relation to the opened or inactivated mode, but for DHP antagonists this behaviour is more pronounced in relation to the channel opening DHP activators.

In order to find some reasonable explanations for this different binding behaviour of structural closely related DHP antagonists and agonists, the aim of the present study was to construct selective pseudoreceptor models of the resting as well as the inactivated state of L-type VGCC.

## METHODS

### DHP Generation

All investigated DHP derivatives were generated within the BUILDER module of the SYBYL software package (Tripos Associates, Inc.) and energy minimized applying the conjugate gradients algorithm. To consistently yield geometry optimized ligand and receptor molecules, all ligands were re-optimized within the PrGen software (Biographics Laboratory) applying the implemented YETI force field. A following semiempiric AM1 single point calculation was performed to yield accurate ESP atomic charges for all ligands.

### Pseudoreceptor Modelling

The pseudoreceptor modelling software PrGen was used to generate atomistic binding site models for a series of pharmacologically active DHP derivatives. Within this routine, a coupling constant of 1.0 and a maximal allowed rms of 0.1 kcal/mol for the predicted versus experimental dissociation constants of all correlation-coupled receptor and ligand minimizations was chosen. The target rms deviation was limited to a maximum of 0.130 kcal/mol. Both the training set and the test set structures were relaxed inside the receptor cavity without constraints applying 10 trails of a Monte-Carlo procedure. Solvation energies of all ligands were calculated according to Still et al. (1990) and entropy corrections were considered following Searle and Williams (1992). Biological binding data of the pure DHP enantiomers showing either antagonistic or agonistic activities were taken from Zheng et al. (1992).

Taking into account the Gibbs-Helmholtz equation, conversion of experimental dissociation constants  $K_d$  to free energies of binding were calculated as follows:  $\Delta G^0 = R \cdot T \cdot \ln(K_d) \cong 1.419 \text{ (kcal/mol)} \cdot \lg(K_d)$  at 37° Celsius.

## RESULTS AND DISCUSSION

### Pharmacophore Generation

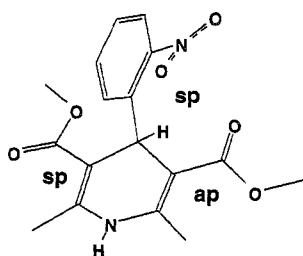


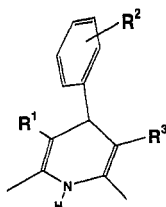
Figure 1. Nifedipine X-ray structure

In order to construct a common pharmacophore of all investigated DHPs, we considered 34 X-ray structures from Cambridge Structural Database. Taking the X-ray structure of nifedipine as an example, the carbonyl oxygens of the almost coplanar arranged ester side chains may be oriented in a synperiplanar (Z)-conformation (sp) or an antiperiplanar (E)-conformation (ap) relative to the double bonds of the boat-like DHP ring (Figure 1). Also for the relative spatial orientation of the 2'-nitro group and the hydrogen in position C4, the terms sp and ap are used if both are pointing to the same or opposite side, respectively.

While the sp conformations for the left-hand side (C3) and the 4-phenyl substituent (C4) as the bioactive orientation are well established (Goldmann and Stoltefuß, 1991), the right-hand side is usually described in literature as non-essential. On the other hand there are known inactive lactone fused DHP with a frozen ap oriented C5 carbonyl oxygen (Kwon et al., 1989), whereas an unrestricted carboxylate at the same position shows full activity. This

clearly demonstrates the essential sp orientation of the carbonyl oxygen also for the right-hand side of DHP. Therefore all molecules were superimposed over the common 1,4-dihydropyridine ring in a sp/sp/sp arrangement.

**Table 1.** Investigated DHP derivatives with their corresponding experimentally determined ( $\Delta G_{\text{exp}}$ ) and via pseudoreceptor modelling predicted ( $\Delta G_{\text{calc}}$ ) free energies of binding in the resting (r.s.) and the inactivated state (i.s.) in kcal/mol. Lower six compounds (\*) represent the test set derivatives



Derivative	R <sup>1</sup>	R <sup>2</sup>	R <sup>3</sup>	$\Delta G_{\text{exp}}$ r.s.	$\Delta G_{\text{calc}}$ r.s.	$\Delta G_{\text{exp}}$ i.s.	$\Delta G_{\text{calc}}$ i.s.
nifedipine	COOCH <sub>3</sub>	2'-NO <sub>2</sub>	COOCH <sub>3</sub>	-10.502	-10.381	-13.184	-13.058
3CN	COOCH <sub>3</sub>	3'-CN	COOCH <sub>3</sub>	-9.708	-9.784	-12.108	-12.294
4Cl	COOCH <sub>3</sub>	4'-Cl	COOCH <sub>3</sub>	-8.209	-8.176	-8.964	-9.021
III	NO <sub>2</sub>	2'-OCF <sub>2</sub> H	COOCH <sub>3</sub>	-9.571	-9.660	-10.474	-10.499
IV	COOCH <sub>3</sub>	2'-OCF <sub>2</sub> H	NO <sub>2</sub>	-9.264	-9.161	-10.564	-10.430
IX	NO <sub>2</sub>	2'-CF <sub>3</sub>	H	-6.967	-6.949	-7.634	-7.583
X	H	2'-CF <sub>3</sub>	NO <sub>2</sub>	-7.364	-7.375	-7.741	-7.867
XIII	NO <sub>2</sub>	2'-CF <sub>3</sub>	NO <sub>2</sub>	-8.256	-8.453	-9.110	-9.216
XIV	NO <sub>2</sub>	2'-OCF <sub>2</sub> H	NO <sub>2</sub>	-7.817	-7.718	-8.660	-8.471
H*	COOCH <sub>3</sub>	H	COOCH <sub>3</sub>	-8.576	-12.083	-10.387	-14.251
3OMe*	COOCH <sub>3</sub>	3'-OCH <sub>3</sub>	COOCH <sub>3</sub>	-7.819	-12.767	-8.461	-14.877
I*	NO <sub>2</sub>	2'-CF <sub>3</sub>	COOCH <sub>3</sub>	-9.704	-9.566	-10.641	-10.432
II*	COOCH <sub>3</sub>	2'-CF <sub>3</sub>	NO <sub>2</sub>	-8.803	-9.294	-10.296	-11.284
XI*	NO <sub>2</sub>	2'-OCF <sub>2</sub> H	H	-7.860	-6.965	-8.277	-7.795
XII*	H	2'-OCF <sub>2</sub> H	NO <sub>2</sub>	-7.422	-7.158	-7.783	-6.509

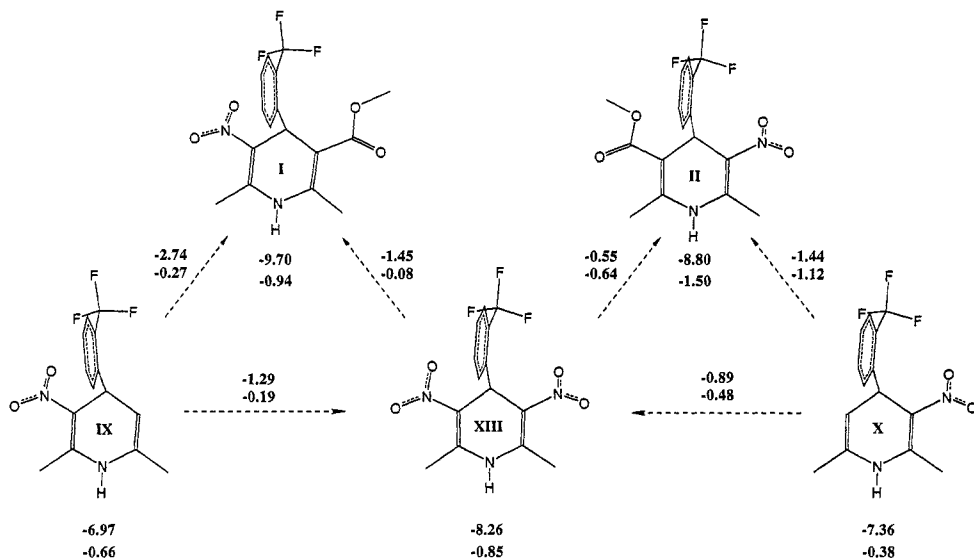
### Pseudoreceptor Model of the Resting State

To generate reasonable pseudoreceptor models we considered the experimentally detected amino acid residues crucial for high affinity binding at L-type VGCC. On the other hand, since the goal of this study was not to imitate the *real* binding cavity but to find minimum requirements for an accurate binding not only explicitly determined amino acids but also residues showing same characteristics were allowed.

Taking the 4-aryl moiety as a mirror axis, all investigated DHP possess an almost symmetric construction showing either a nitro or a carboxylate substituent at C3 or C5. Therefore, residues of a hypothetical binding site might almost equally be positioned at either side. To overcome this dilemma a careful comparison of the effects caused by same substituents at opposite sides was carried out (Figure 2).

Closer examination of the binding affinities of compounds IX and X in the resting state reveals the nitro group at the right-hand side to be more important for binding (X:  $\Delta G$  -7.36 kcal/mol) than positioned at the opposite side (IX:  $\Delta G$  -6.97 kcal/mol). The same tendency is observed by insertion of a second nitro group yielding compound XIII.

Following path X→XIII the binding energy increases by -0.89 kcal/mol while the way IX→XIII which generates the nitro group at the right-hand side yields an energy gain of -1.29 kcal/mol. Even more striking are the changes from IX→I and X→II. While the additional methyl ester at C5 (I) increases the binding affinity by -2.74 kcal/mol, the same substitution at the left-hand side (II) yields only -1.44 kcal/mol. To look at derivative XIII, exchange of a nitro group against the methyl ester at the right side (XIII→I;  $\Delta G$  -1.45 kcal/mol) and the left side (XIII→II;  $\Delta G$  -0.55 kcal/mol), respectively, also indicates the importance of the C5 substituents for the resting state.



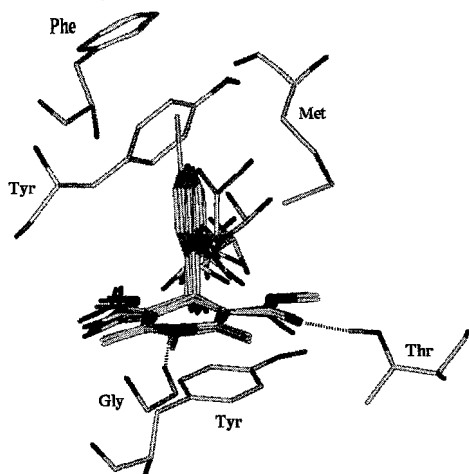
**Figure 2.** Comparative study of binding affinities. Upper values: free energies of binding in the resting state (below the structures); energy gain caused by different substitution in the resting state (arrows). Lower values: gain of binding energy after channel activation (inactivated/opened state) [all values in kcal/mol].

In the light of these observations, we placed a crucial threonine as hydrogen donor at the sp oriented right-hand side of the pharmacophore. The NH function of the DHP ring was saturated by a carbonyl oxygen of the glycine backbone. A methionine was located axially beside and a phenylalanine on top of the substituted 4-phenyl ring. Two additional tyrosines were placed below the 1,4-dihydropyridine ring and parallel to the 2'- and 3'-moieties, respectively (Figure 3).

A receptor equilibration was carried out by minimizing all residues of the crude pseudoreceptor keeping the ligands of the training set fixed. In the following step a correlation-coupled receptor minimization followed by free ligand relaxation was used to obtain a satisfactory correlation of  $R=0.99$  ( $rms=0.097$  kcal/mol) between experimental and predicted binding energies. To overcome local minima of the ligands a Monte-Carlo search was performed to find the best adjustment within the binding cavity.

The quality of this pseudoreceptor model was validated by replacing the training set with the test set ligands followed by an unrestricted Monte-Carlo relaxation. Thereafter, free energies of binding were predicted for these ligands using the linear regression obtained with the training set yielding a rms of 2.51 kcal/mol (Table 1). As can be seen, the unsatisfactory result for the complete test set showing a deviation of more than one  $K_d$  unit is mainly caused by the unsubstituted derivative H (-3.51 kcal/mol) and the 3'-OMe DHP

(-4.95 kcal/mol). Exclusion of those outliers yields a rms of 0.532 kcal/mol, representing an uncertainty factor (UF) of 2.37 ( $=10^{0.532/1.419}$ ).



**Figure 3.** Pseudoreceptor model of the resting state. For clarity only NH and OH hydrogens are displayed (dashed lines indicate hydrogen bonds).

Since H is the only unsubstituted 4-phenyl derivative and test set molecules usually may only be predicted correctly if there are related derivatives in the training set, receptor equilibration was repeated including H into the training set. But surprisingly no sufficient correlation was found ( $R=0.884$ ), indicating once again the exceptional role of compound H. Closer inspection of the individual ligand/receptor complexes revealed no detectable interactions to explain such high receptor affinities. This makes it difficult to understand why the only unsubstituted derivative H generates more attractive interactions in relation to the substituted derivatives, all the more if one considers that both tyrosines of the pseudoreceptor model generate strong attractive interactions to the 4-phenyl substituents.

To draw the conclusion from these findings, it is unlikely that PrGen really calculates to high interaction energies of the above mentioned outliers, but quite the contrary, that the programme is not able to accurately determine the binding energies of all other derivatives. In this case, at least one force must be relevant for ligand binding that is not recognized by the force field. Since all molecules of the first approach possess an electron withdrawing substituent inducing an electron impoverished 4-phenyl moiety, a natural suspicion of that "unrecognized force" might be a charge transfer interaction. To proof this hypothesis, three separate complexes, composed of compound H/pseudoreceptor (H/PR), 3'-CN/pseudoreceptor (CN/PR) and nifedipine/pseudoreceptor (nif/PR), respectively, were extracted and used as input for quantum chemical AM1 calculations. Due to convergence problems in course of the computation, the model had to be reduced by the phenylalanine and one tyrosine residue. Computation of the HOMOs and LUMOs indicates striking differences between the complexes. While in all cases the HOMO is localized at the methionine that is placed beside the 4-phenyl ring, the LUMO of nifedipine, LUMO+1 of CN/PR and only LUMO+5 of H/PR -as the energetically most favourable unoccupied molecular orbitals- are localized in front of the HOMO at the 4-phenyl ring. Careful calculation of the orbital energies reveals significant distinctions yielding energy differences for corresponding HOMOs and LUMOs of 7.73 eV, 8.15 eV and 8.92 eV for nif/PR, CN/PR and H/PR, respectively. Since small energy differences between HOMOs and LUMOs are essential for electron donor acceptor interactions, the results are in agreement with a charge transfer hypothesis.

In order to proof the selectivity of the pseudoreceptor model representing the resting state, the whole receptor generation was repeated using the same ligand molecules but experimental data of the channel in the opened/inactivated state (Table 1). In spite of a correlation of  $R=0.99$  (rms=0.115 kcal/mol) for the training set the prediction for the test set molecules with a rms of 5.928 kcal/mol demonstrated the inability of an accurate correlation. Again exclusion of derivatives H ( $\Delta G$  -9.39 kcal/mol) and 3'-OMe ( $\Delta G$  -8.83 kcal/mol) yields a smaller deviation of 2.033 kcal/mol (Table 1). Nevertheless, compared to the results applying experimental data of the channel in resting state (rms 0.532 kcal/mol



vs. 2.033 kcal/mol) the uncertainty factor raises from 2.37 to 27.08, indicating a sufficient distinction between these channel modes.

### Pseudoreceptor of the Opened/Inactivated State

In order to gain hints about the varied binding site characteristics induced by channel activation, a careful interpretation of figure 2 gives helpful information. Substitution of a nitro against a carboxylate group on the right-hand side (XIII→I) yields an energy gain of -0.08 kcal/mol in relation to the resting state. The same exchange at the left-hand side yields an additional energy of -0.64 kcal/mol. Insertion of a methyl carboxylate group at the right-hand (IX→I,  $\Delta\Delta G = -0.27$  kcal/mol) and the left-hand side (X→II,  $\Delta\Delta G = -1.12$  kcal/mol), respectively, reflects still more profoundly the essential meaning of the left-hand side for ligand binding in the inactivated state.

Considering these observations, it seemed to be reasonable to place a hydrogen donor in form of a second threonine at that side for a simulation of this channel mode (Figure 4).

And in fact, this simple variation yields a correlation of  $R=0.99$  (rms=0,123 kcal/mol) for the pseudoreceptor model of the inactivated state with a rms of 0.848 kcal/mol (Uf: 3.96) for the prediction of the residual four test set derivatives. Naturally, also for this model the suspected charge transfer interactions were observed leading to deviations of -3.86 kcal/mol and -6.42 kcal/mol for H and the 3'-OMe DHP, respectively.

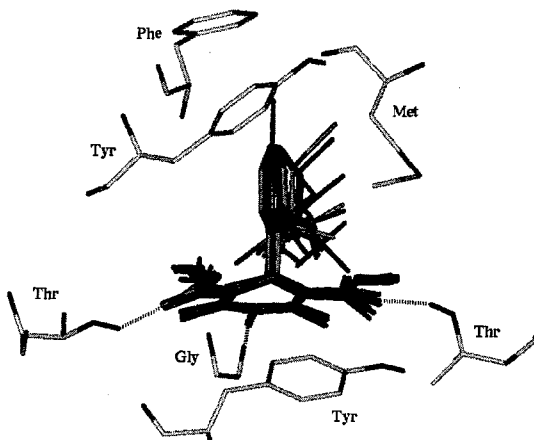


Figure 4. Pseudoreceptor model of the channel in the inactivated state. For clarity only NH and OH hydrogens are displayed (dashed lines indicate hydrogen bonds).

Even though a transfer of these theoretically derived findings to a realistic binding site is quite speculative, the observed motions of the channel during transition from the resting to the opened state could explain the generation of an additional contact region for DHP causing increased binding affinities.

### REFERENCES

- Goldmann, S., and Stoltefuß, J., 1991, 1,4-Dihydropyridine: Einfluß von Chiralität und Konformation auf Calcium-antagonistische und -agonistische Wirkung, *Angew. Chem.* 103:1587.
- Kwon, Y.W., Franckowiak, G., Langs, D.A., Hawthorn, M., Joslyn, A., and Triggle, D.J., 1989, Pharmacologic and radioligand binding analysis of actions of 1,4-dihydropyridine activators related to Bay K 8644 in smooth muscle, cardiac muscle and neuronal preparations, *Naunyn-Schmiedeberg's Arch. Pharmacol.* 339:19.
- Searle, M.S., and Williams, D.H., 1992, The cost of conformational order: entropy changes in molecular associations, *J. Am. Chem. Soc.* 114:10690.
- Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T., 1990, Semianalytical treatment of solvation for molecular mechanics and dynamics, *J. Am. Chem. Soc.* 112:6127.
- Zheng, W., Stoltefuß, J., Goldmann, S., and Triggle, D.J., 1992, Pharmacologic and radioligand binding studies of 1,4-dihydropyridines in rat cardiac and vascular preparations: stereoselectivity and voltage dependence of antagonist and activator interactions, *Mol. Pharmacol.* 41:535.

## PHARMACOPHORE DEVELOPMENT FOR THE INTERACTION OF CYTOCHROME P450 1A2 WITH ITS SUBSTRATES AND INHIBITORS

Elena López-de-Briñas,<sup>1</sup> Juan J. Lozano,<sup>1</sup> Nuria B. Centeno,<sup>1</sup> Jordi Segura,<sup>2</sup> Marisa González,<sup>2</sup> Rafael de la Torre<sup>2</sup> and Ferran Sanz<sup>1,\*</sup>

<sup>1</sup>Research Group on Medical Informatics,

<sup>2</sup>Research Unit of Pharmacology,

Institut Municipal d'Investigació Mèdica (UAB), c/ Dr. Aiguader 80, E-08003 Barcelona (Spain)

### INTRODUCTION

The cytochromes P450 are a superfamily of isoenzymes that catalyse the metabolism of a large number of compounds of both endogenous and exogenous origins.<sup>1</sup> Cytochrome P450 1A2 (CYP1A2) is a member of the CYP1 family that is responsible for the metabolism of several planar highly conjugated compounds.

Among the substrates of this cytochrome, there are several important substituted xanthenes like caffeine,<sup>2</sup> as well as heterocyclic aromatic amines (HCA) present in cooked food meat and fish. The metabolism of the HCA has biological importance because they exert a genotoxic activity after their N-oxidation by cytochrome P450 1A2.<sup>3</sup> Other specific substrates are 7-ethoxyresorufin and phenacetin. On the other hand, several quinolones, which could be interesting in therapeutics because they are potent antibacterials, present the side-effect of being competitive inhibitors of the metabolism of other P450 1A2 substrates like caffeine.<sup>4</sup>

The P450 1A2 substrates exhibit a wide structural variability as it can be observed in Figure 1. The main objective of the present study was to find veiled

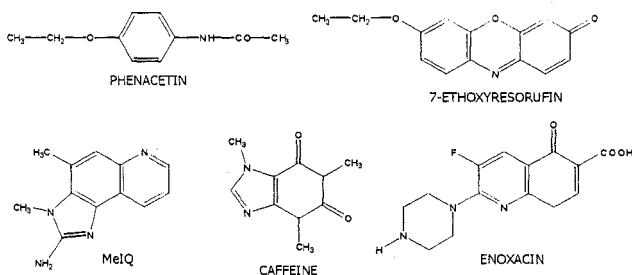


Figure 1. Some substrates of cytochrome P450 1A2.

similarities between the mentioned compounds that could explain their common biological activity as substrates of the cytochrome P450 1A2. The study was carried out on the basis of the analysis of the molecular electrostatic potential (MEP) distributions of the compounds.

\* To whom correspondence has to be addressed

## MOLECULAR ELECTROSTATIC POTENTIAL ANALYSIS

The MEP distributions of the considered compounds were computed at the quantum mechanical level using the wavefunctions resulting from full geometrical optimisations using the GAUSSIAN software with the 3-21G basis set. The MEP distributions were computed and analysed with the MEPMIN module<sup>5</sup> of MEPSIM package.<sup>6</sup> MEPMIN detects the MEP minima of a molecule and finds the geometrical relationships between them. In the case of compounds with several low energy conformations that generated different MEP distributions (phenacetin and 7-ethoxyresorufin), they were analysed separately. The MEP maps of the considered compounds in their main molecular plane are shown in Figures 2-6.

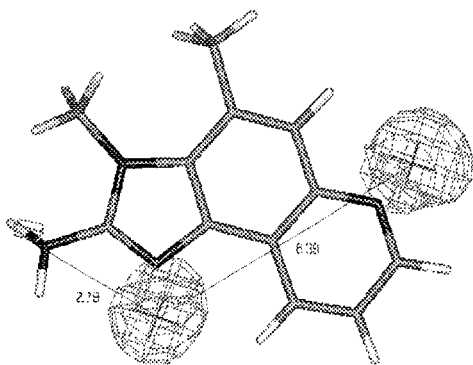


Figure 2. MEP map of MeIQ (the most active HCA).

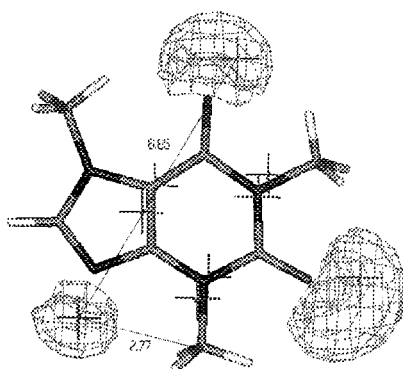


Figure 3. MEP map of caffeine.

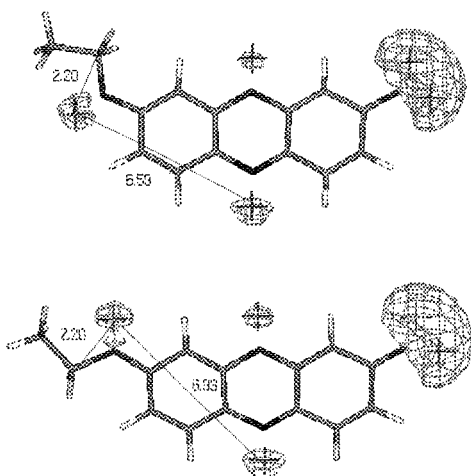


Figure 4. MEP maps of two low energy conformations of 7-ethoxyresorufin.

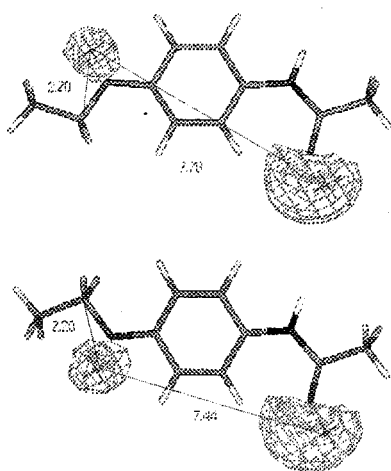


Figure 5. MEP maps of two low energy conformations of phenacetin.

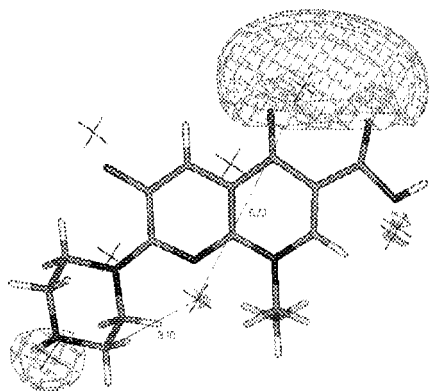


Figure 6. MEP map of enoxacin.

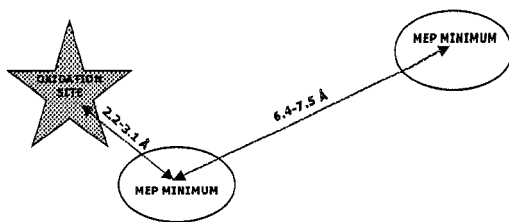


Figure 7. Scheme of the proposed pharmacophore for the substrates of cytochrome P450 1A2.

The observation of MEP maps like those shown in Figures 2-6 allowed us to define the pharmacophore presented in Figure 7. It indicates that the CYP1A2 substrates have two deep zones of negative MEP located at opposite sides of the molecular structure and separated by a distance that ranges from 6.4 to 7.5 Å. Furthermore, one of these zones is located at a distance of 2.2-3.1 Å of the group that is oxidated by the cytochrome P450 1A2. The fitting of the above mentioned substrates on the basis of the pharmacophore is shown in figure 8.

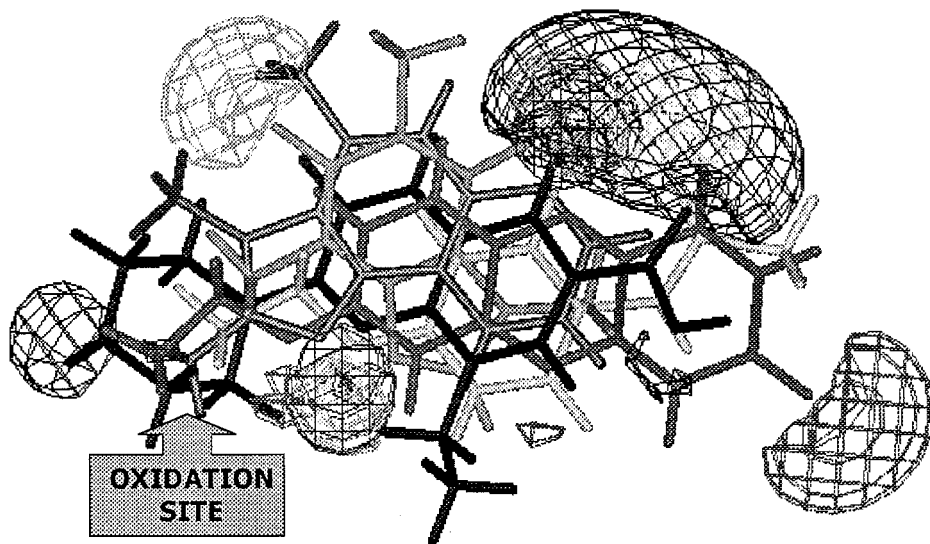
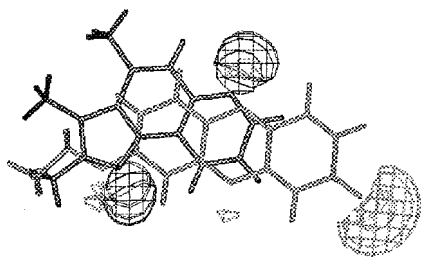
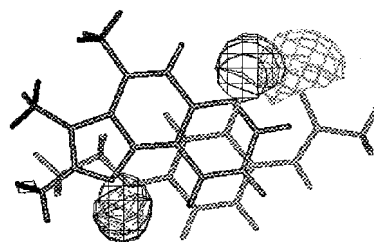


Figure 8. Fitting of CYP1A2 substrates on the basis of the proposed pharmacophore

Another procedure of analysing the similarity of MEP distributions is by means of the use of the MEPCOMP program,<sup>7</sup> which is also integrated in the MEPSIM package.<sup>6</sup> MEPCOMP performs an automatic search of the alignment of two compounds looking for a maximum of a similarity coefficient between the corresponding MEP distributions. It has to be pointed out that MEPCOMP takes into account the whole MEP distributions and not only the position of the minima as it happened in the MEPMIN approach. In the present study, we used the MEPCOMP program to test if the above mentioned relative positions of the compounds (see Figure 8) agreed with optimal alignments after MEPCOMP processes. Figures 9 and 10 show the alignments proposed by MEPCOMP in the comparisons of MeIQ with 7-ethoxyresorufin and phenacetin. In these two examples, MEPCOMP supplied relative positions that agreed with the manually proposed on the basis of the pharmacophore.

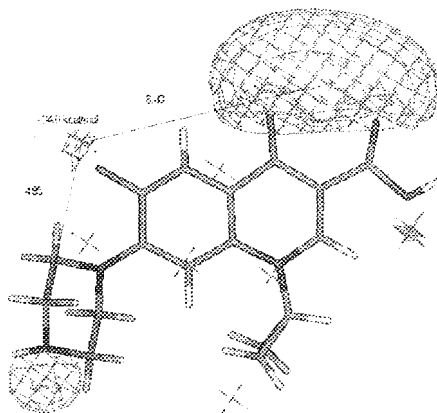


**Figure 9.** Alignment proposed by MEPCOMP in the comparison of MeIQ vs 7-ethoxyresorufin

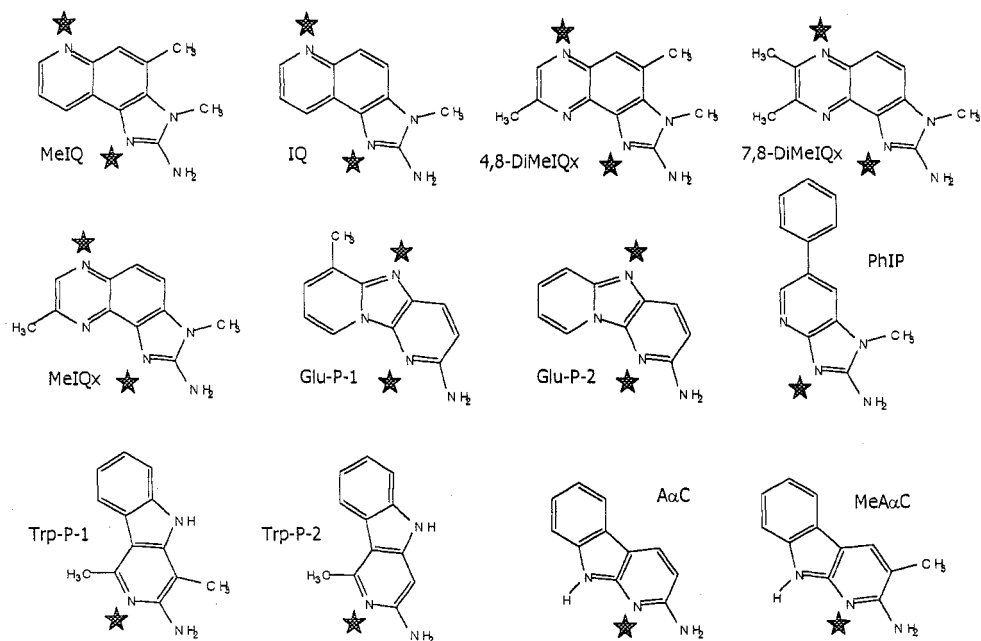


**Figure 10.** Alignment proposed by MEPCOMP in the comparison of MeIQ vs phenacetin.

An additional challenge for the proposed pharmacophore was to observe if it could contribute to explain differences in activity within congeneric series of compounds. A first positive result on this issue arose from the comparison of the MEP maps of enoxacin (Figure 6) and ciprofloxacin (Figure 11), two quinolonic antibacterials which are more and less active at the cytochrome P450 1A2 respectively. In both cases it is possible to define the proposed pharmacophore, but in the case of ciprofloxacin it shows flawed features like the need of rotating the piperazine ring from its lowest energy conformation in order to reach the proposed distance between one of the MEP minima and the group to be oxidated. Another weak feature of ciprofloxacin is the fact that both minima are at the same side of the molecular structure, and the last defect is the smaller magnitude of the MEP minimum that is close to the oxidation site, in comparison to the rest of the compounds. This magnitude is 34.0 kcal/mol in the case of ciprofloxacin, whereas it is 48.5 kcal/mol in the case of enoxacin and even greater in the rest of the studied substrates.



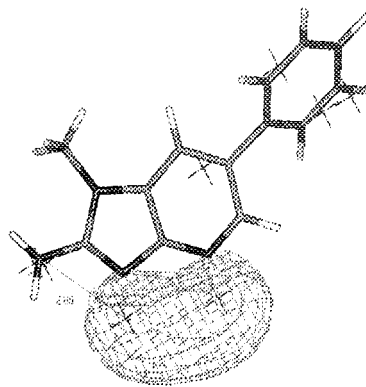
**Figure 11.** MEP map of ciprofloxacin.



**Figure 12.** Mutagenic heterocyclic amines.<sup>8</sup> Stars indicate the approximate locations of minimum MEP zones.

A second positive result that we obtained on the same issue, relates with the food heterocyclic amines that are activated to mutagens by cytochrome P450 1A2. If we observe the MEP distributions of the series of such amines experimentally studied by Wakabayashi et al.,<sup>8</sup> we can see that all of them have a deep zone of negative MEP at the relevant distance (almost three Ångstroms) of the amino group that is N-oxidated by the cytochrome (Figure 12). Furthermore, the three less active compounds (PhIP, AαC, MeAαC) lack of the second proposed zone of minimum MEP, while the five most active molecules (MeIQ, IQ, 4,8-DiMeIQx, 7,8-DiMeIQx, MeIQx) possess both zones. As an example of the MEP distribution of a weakly active amine, Figure 13 shows the MEP map of PhIP. It has to be pointed out that in this case, the MEP distribution not only lacks of one of the negative MEP zones but the one close to the oxidation site is more extended that in the rest of studied compounds (see Figures 2-6).

We have been successfully using the present pharmacophoric model in other kinds of theoretical studies. For instance, we have used the minima positions as possible solvation positions in docking simulations.<sup>9</sup> It has to be pointed out that we have found interesting coincidences between the pharmacophore and the results of other approaches that we have been using to study the same problem. For instance, we have carried out docking simulations of the series of heterocyclic amines using the AUTODOCK 2.4<sup>10</sup> software and a 3D model of cytochrome P450 1A2 previously obtained.<sup>9</sup> The automatic docking processes generated two clusters of interaction positions



**Figure 13.** MEP map of PhIP.

that included the amines having two or only one minimum MEP zones respectively.<sup>11</sup> Using the alignment of the 12 amines resulting of the AUTODOCK computations, we have performed COMBINE<sup>12</sup> and GRID/GOLPE<sup>13</sup> analyses that yielded excellent predictive indexes ( $q^2$  approximately equal to 0.8 in two PC models).<sup>11</sup>

## CONCLUSIONS

We have proposed a MEP-based pharmacophore that could facilitate the qualitative prediction of the capability of compounds to interact with cytochrome P450 1A2. This possible application has a notable interest in the drug development process. On the other hand, the agreement that we have found between the proposed model (MEP-based pharmacophore), and the results obtained using other approaches (docking simulations, 3D-QSAR studies) gives us an increased confidence in all of them. The control of the agreement between the results obtained using several independent methods should be a normal working strategy to increase the reliability of the theoretical models.

## Acknowledgements

This research was supported in part by CICYT (SAF 93-0722-C02-02) and CESCA grants.

## REFERENCES

1. S.D. Black and M.J. Coon. *Adv. Enzymol.* 60:35 (1987).
2. J. Segura, D.J. Roberts and E. Tarrús. *J. Pharm. Pharmacol.* 41:129 (1988).
3. T. Shimada, M. Iwasaki, M.V. Martin and F.P. Guengerich. *Cancer Res.* 49:3218 (1989).
4. V. Fuhr, G. Strobl, F. Manaut, E.-M. Anders, F. Sörgel, E. López-de-Briñas, D.T.W. Chu, A.G. Pernet, G. Mahr, F. Sanz and H. Staib. *Mol. Pharmacol.* 43:191 (1993).
5. F. Sanz, F. Manaut, J. José, J. Segura, M. Carbó, and R. de la Torre. *J. Mol. Struct. (THEOCHEM)* 170:171 (1988).
6. F. Sanz, F. Manaut, J. Rodríguez, E. Lozoya and E. López-de-Briñas. *J. Comput.-Aided Mol. Design* 7:337 (1993).
7. F. Manaut, F. Sanz, J. José and M. Milesi. *J. Comput.-Aided Mol. Design* 5:371 (1991).
8. K. Wakabayashi, M. Nagao, H. Esumi and T. Sugimura. *Cancer Res.* 52(suppl):2092s (1992).
9. J.J. Lozano, E. López de Briñas, N.B. Centeno, R. Guigó and F. Sanz. *J. Comput.-Aided Mol. Des.* 11:39 (1997).
10. G.M. Morris, D. Goodsell, R. Huey and A.J. Olson. *J. Comput.-Aided Mol. Des.* 10:293 (1996).
11. See the chapter of J.J. Lozano et al. in the same book.
12. C. Pérez, M. Pastor, A.R. Ortiz and F. Gago. *J. Med. Chem.* 41:836 (1998).
13. M. Baroni, G. Constantino, G. Cruciani, D. Riganelli, S. Valigri and S. Clementi. *QSAR* 12:9 (1993).

**Section V**  
**Computational Aspects of**  
**Molecular Diversity and**  
**Combinatorial Libraries**



# ANALYSIS OF A LARGE, HIGH-THROUGHPUT SCREENING DATA USING RECURSIVE PARTITIONING

S. Stanley Young<sup>1</sup> and Jerome Sacks<sup>2</sup>

<sup>1</sup>Glaxo Wellcome Inc.  
5 Moore Drive  
RTP, North Carolina 27709

<sup>2</sup>National Institute of Statistical Science  
P.O. Box 14006  
RTP, North Carolina 27709-4006

## ABSTRACT

As biological drug targets multiply through the human genome project and as the number of chemical compounds available for screening becomes very large, the expense of screening every compound against every target becomes prohibitive. We need to improve the efficiency of the drug screening process so that active compounds can be found for more biological targets and turned over to medicinal chemists for atom-by-atom optimization. We create a method for analysis of the very large, complex data sets coming from high throughput screening, and then integrate the analysis with the selection of compounds for screening so that the structure-activity rules derived from an initial compound set can be used to suggest additional compounds for screening. Cycles of screening and analysis become *sequential screening* rather than the mass screening of all available compounds. We extend the analysis method to deal with multivariate responses. Previously, a screening campaign might screen hundreds of thousands of compounds; sequential screening can cut the number of compounds screened by up to eighty percent. Sequential screening also gives SAR rules that can be used to mathematically screen compound collections or virtual chemical libraries.

## INTRODUCTION

The basic techniques of drug discovery are rapidly changing. Many more targets are being identified through the human genome project and other genetic initiatives. Vast numbers of compounds are available for testing. A large pharmaceutical company has hundreds of thousands of compounds in inventory for this initial testing. Over one million compounds are available from commercial sources. Combinatorial synthesis is making additional millions of compounds available. Testing has also changed. Robots can rapidly deal with many, very small samples. Testing is giving rise to data sets with hundreds of thousands of biological results.

Can we take advantage of all this data? The need is to relate chemical structural features to biological results. But the representation of chemical structures in a form suitable for analysis is complex. A molecule is a set of connected atoms. The connections can be complex and the atoms can take on different characteristics depending on which

and how other atoms are connected to them. It is possible to develop large numbers of structural descriptors, thousands to millions! Compounds typically contain rotatable bonds and are typically very flexible. They elicit their effect by binding to proteins that are also flexible.

But, even more problematical, the various compounds can bind in different ways and even in different places to product their effects. We have a mixture problem. Analysis should take into account that compounds can act through different mechanisms; most analysis methods fail in these circumstances. To be successful we need to identify the different classes and the features important for each class.

In this paper we describe how recursive partitioning, RP, can be extended to deal with very large data sets where each sample is described with a large number of bivariate, 0/1, descriptors. We describe how RP can be used in sequential screening to improve the efficiency of the screening process. See Figure 1.

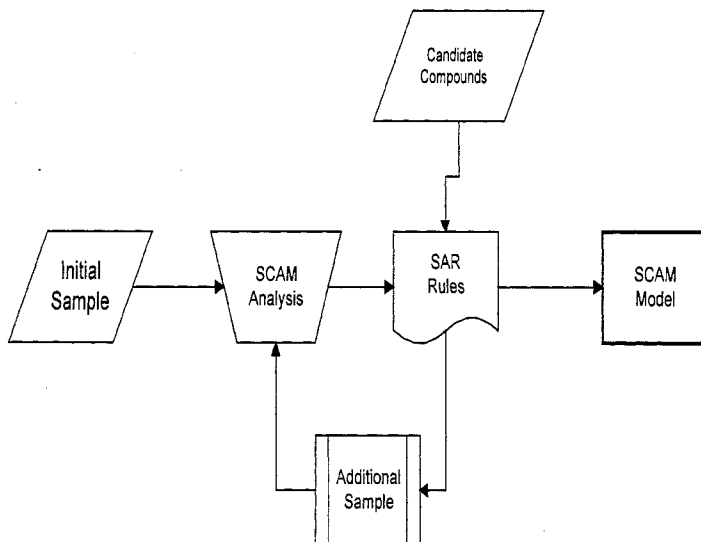


Figure 1. Sequential screening is accomplished by screening an initial set of compounds and computing a SCAM analysis. This analysis gives SAR rules. These rules are used to select an additional sample of compounds from a collection. The additional sample and initial sample are used to compute new SAR rules using SCAM. Several cycles give a final SCAM model.

In sequential screening an initial set of compounds is analyzed by our specialized RP program, SCAM, to give SAR rules. These SAR rules can be used to select additional compounds for screening. The screening and analysis cycle searches out active compounds and gives a model for active compounds.

## THE DATA

We describe each chemical structure by determining the presence or absence of structural features, atom pairs [Carhart et al., 1986], topological torsions [Nilakantan et al., 1987], and atom triples in the structure, hydrogens omitted. Atom triples are an extension of atom pairs See Figure 2.

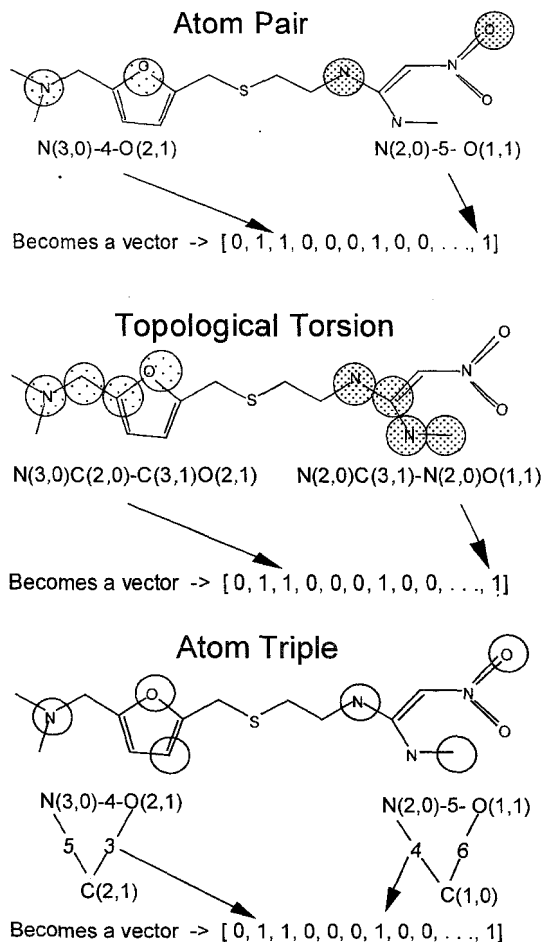


Figure 2. Examples of atom pairs, topological torsions and atom triples. Atoms are typed by their name, the number of non-hydrogen connections and the number of shared  $\pi$  electrons and the shortest path between the atoms. We describe each compound with vector, a bit string. A 1 indicates that a feature is present and a 0 indicates that it is not.

An atom triple consists of three atoms characterized by their atomic number, the number of non-hydrogen connections and the number of  $\pi$  electrons. Three topological distances are used, the shortest path between each pair of atoms in the triple. Redundant atom triples are eliminated and unique triples for the data set are enumerated. Each chemical structure is characterized by a bit string noting the presence and absence of each possible feature, atom pair, topological torsion, and/or atom triple, Figure 3.



We chose to modify the FIRM method of Hawkins and Kass[1982] from multi-way splits to two-way splits using a two-group Student's t-test. The t-test with the smallest p-value is used for the split. Each daughter group is split until either the t-test is not significant, adjusted for multiple testing, or the sample size becomes too small. Splitting with a t-test is much faster than the RP algorithms of CART and C4.5.

The matrix of descriptors can be very large, 100k by 2M is not unusual (Several million t-tests are examined at each split!) We recognize that the descriptor matrix is sparse, mostly 0s so we store only where the 1s are located, saving storage. Once a group is split we can recursively compute summary statistics needed for the t-test from the parent node and the node with the smaller number of observations thus gaining speed. Our algorithms operate in real time on a UNIX workstation enabling interactive analysis. Also, attention is paid to multiple testing. We adjust the analysis to reflect the number of variables examined at each split to control the possibility of a false split. Our RP method determines molecular structural features associated with biological activity for each terminal node, de-convoluting the original mixture and finding the important molecular features for each.

## EXAMPLE

We present an analysis of a set of 1650 compounds of widely varying structure tested for monoamine oxidase, MAO, inhibition [Brown and Martin,1996]. It is known that there are at least two types of compounds acting through completely different mechanisms. The activity of each compound was scored 0, 1, 2, 3 with 0 indicating no activity and 3 the most active. The analysis of this data set using atom triples as descriptors is given in Figure 5.

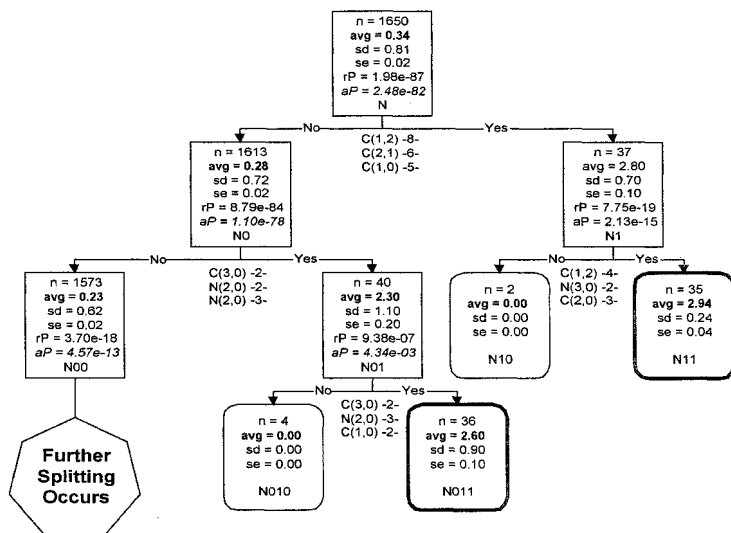


Figure 5. A recursive partitioning tree can be read in the following way. Each node gives summary statistics for the objects in the node, e.g., number of observations, average, standard deviation, standard error of the mean and two p-values for the splitting—the raw or unadjusted p-value and the p-value adjusted to reflect the number of variables examined. Below each node is the feature used to split the node. The nodes are named reflecting the binary splitting. Splitting stops when the adjusted p-value is not statistically significant.

Among all the atom triples, the triple, C(1,2)-8-C(2,1)-6-C(1,0)-5 splits off 37 compounds with an average activity of 2.80, from the remaining compounds. Following node N1, two inactive compounds were split off giving 35 compounds with an average activity of 2.94. Node N011 with 36 compounds also contains highly active compounds. By tracing the rules that give rise to the splits leading to a terminal node, chemical structural features are identified. Figure 6 gives a typical compound from nodes N110 and N011. These two compounds indeed act through different mechanisms, one binding to a co-factor of MAO and the other binding to MAO.

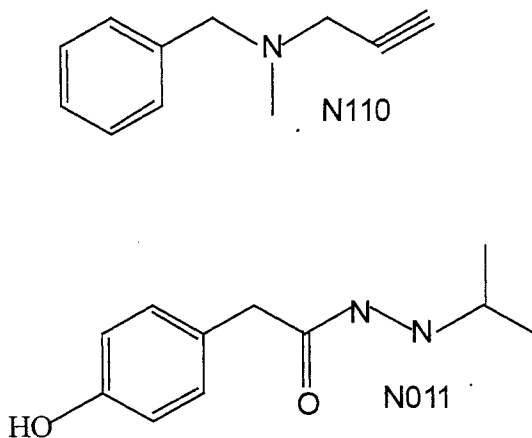


Figure 6. Compounds selected from two active nodes, N110 and N011.

## SEQUENTIAL SCREENING

This method is being used as part of a sequential screening strategy. The number of compounds available for testing is exceedingly large. The idea is to select a small number of compounds for initial screening, analyze the results, and then predict which of the available, untested compounds should be screened next. To study the effectiveness of sequential screening we conducted extensive simulations using large data sets of tested compounds.

But there are many questions about how to optimize sequential screening. How large should the initial screening set be? How should the compounds be selected? How should the compounds be described? (There are many types of descriptors other than atom triples.) What analysis method should be used? How many cycles of selection should be completed? We examined many factors using a factorial design [Box, Hunter, and Hunter, 1978]. For example, we looked at initial sample sizes of 5,000 or 10,000, selected at random or selected to be chemically diverse. We looked at follow-up samples of size 2,500 or 5,000. As a measure of effectiveness we formed a ratio of the number of good compounds found, the number in the top 100 or top 350 compounds from the ~71,000 compounds in the data set, relative the number of good compounds expected to be found by chance.

We learned the general characteristics that give good results for sequential screening. The initial set should be five to ten thousand compounds. If the assay has good precision, then even smaller initial sets work well. Two cycles of selection work well. Etc.

The most surprising result was that how the initial set of compounds was selected seemed unimportant. In particular, a random selection of compounds was as effective as a carefully selected diverse set for starting the sequential screening process. In virtually all areas of experimentation, carefully selecting the learning/training set gives the most information relative to effort expended. It seemed reasonable to us that selecting compounds as different from one another as possible would sample more of the available chemical space and we would be less likely to miss an important region. Since our belief

was not confirmed, we decided to re-test the proposition in a second set of about fifty thousand compounds. Some of the results are given in Table 1.

Table 1. The initial set of compounds for sequential screening was selected based on chemical diversity and at random. Given is the ratio of active compounds found by sequential screening relative to the number expected by chance. A ratio of one would indicate that sequential screening is no better than chance. We formed this ratio for the top 100 and top 350 compounds. And we repeated the experiment four times for each set of conditions. The four replicate values are ordered by size to facilitate comparisons

	Diversity		Random	
	Top100	Top 350	Top100	Top 350
	3.08	4.21	3.36	3.66
	3.14	4.31	3.36	4.33
	3.64	4.59	3.37	4.41
	<u>3.66</u>	<u>4.64</u>	<u>3.40</u>	<u>4.42</u>
Average	3.38	4.44	3.37	4.21

We give the ratio of active compounds found by sequential screening relative to the number expected by chance. A ratio of one would indicate that sequential screening is no better than chance. We formed this ratio for the top 100 and top 350 compounds. And we repeated the experiment four times for each set of conditions. It is clear that sequential screening greatly improves the ability to find good compounds, compounds in the top 100 or 350 in the collection. It is also clear that the performance of sequential screening is not improved by the selection of a diverse set of initial compounds. Why is diversity selection not better than random selection? Examine Figure 7.

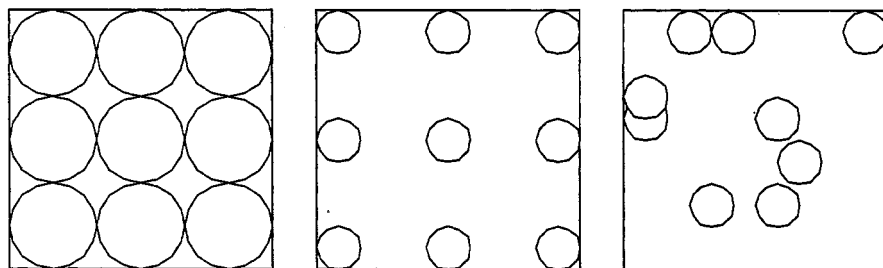


Figure 7. How important is the initial design? Should the compounds be well-spaced? And how well do they cover the space?

Suppose that each compound covers only a small space, second figure versus the first. Then, random compounds are unlikely to overlap so the random compounds cover the same amount of space as the carefully selected compounds, third figure versus the second [Young et al., 1996]. Since it can be difficult and time consuming to carefully select a diverse set of compounds, time and effort can be saved by taking a simple random sample.

## MULTIVARIATE SCAM

Often, several response variables are measured on each compound. It would be useful to find structural features associated with the profile of responses of a compound. One interest is finding features associated with selective compounds. The Student t-test can be replaced with the Hotelling  $T^2$  to solve this problem.

## CONCLUSION

The big payoff of this work is that our modified version of recursive partitioning can be used to greatly increase the efficiency of drug discovery; by screening five to ten percent of a collection we can find thirty to fifty percent of the most active compounds. Important to the chemist is the fact that the method also gives the reasons for activity and that the method finds multiple classes of active compounds. The RP tree shows the chemist which features are important and which are not, so that atom-by-atom synthetic modification is efficient. The method also gives rules to search large structure data bases to suggest additional compounds for screening, either from actual collections or from virtual libraries.

The term "data mining" is used to describe the process of examining large, amorphous data sets with the idea that useful information can be extracted. The data sets usually come from operations and were not collected for decision making. In business situations, billing records are an example. In our situation, compounds are screened and the usual practice is to simply note the few most active compounds and ignore the rest of the data. Data mining is difficult. We think our success comes from applying knowledge from three subjects, statistics, computer science, and the subject matter, chemistry.

**Patents are pending on portions of this work.**

### Data set and code availability

Abbott Laboratory will make an electronic copy of this data set available; contact Daniel W. Norbeck, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, IL 60064-3500.

A recursive partitioning code, FIRM, is available for nominal charge. Contact Dr. D. M. Hawkins, [doug@stat.umn.edu](mailto:doug@stat.umn.edu).

## REFERENCES

- Box, G.E., Hunter, W.G., and Hunter, S., 1978, *Statistics for Experimenters*. J. Wiley & Sons, New York.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., 1984, *Classification and Regression Trees*, Wadsworth
- Brown, R.D. and Martin, Y.C., 1996, Use of Structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36:572-584.
- Carhart, R.E., Smith, D.H., Venkataraghavan, R., 1985, Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* 25:64-73.
- Hawkins, D.M., 1995, *FIRM Formal Inference-based Recursive Modeling*. Release 2. University of Minnesota: St. Paul, MN.
- Hawkins, D.M. and Kass, G.V., 1982, Automatic Interaction Detection. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press, pp. 269-302.
- Hawkins, D.M., Young, S.S., and Rusinko, A., 1997, Analysis of a large structure-activity data set using recursive partitioning. *Quant. Struct.-Act. Relat.* 16:296-302.
- Kass, G.V. 1980, An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29:119-127.
- Nilakantan, R., Bauman, N., Dixon, J. S., Venkataraghavan, R., 1987, Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* 27:82-85.
- Rusinko, A. III, Farnen, M.W., Lambert, C.G., Brown, P.L., and Young, S.S. Analysis of a large structure-activity data set using recursive partitioning. (submitted for publication).
- Quinlan, J.R., 1993, *C4.5: Programs for Machine Learning*. Morgan Kaufmann
- Young, S.S., Farmer, M.W., and Rusinko, A. III, 1996, Random versus rational. Which is better for general compound screening. <http://www.netsci.org/science/screening/feature09.html>.



## 3D STRUCTURE DESCRIPTORS FOR BIOLOGICAL ACTIVITY

Johann Gasteiger\*, Sandra Handschuh, Markus C. Hemmer,  
Thomas Kleinöder, Christof H. Schwab, and Andreas Teckentrup  
Computer-Chemie-Centrum, Universität Erlangen-Nürnberg  
D-91052 Erlangen, Germany  
<http://www2.ccc.uni-erlangen.de>

Jens Sadowski  
BASF AG Drug Design, ZHB/W - A30  
D-67056 Ludwigshafen, Germany

Markus Wagener  
N.V. Organon, Computational Medicinal Chemistry  
PO Box 20  
NL-5340 BH Oss, Netherlands

### ABSTRACT

Novel ways of coding the structure of chemical compounds are presented and their use for correlating biological activity is explored. These structure codes take account of the three-dimensional arrangement of the atoms in a molecule, or consider molecular surface properties. These molecular representations have been studied with large datasets; various applications to biological activity studies and the definition of chemical diversity will be presented.

### INTRODUCTION

Methods for the prediction of biological activity have to rely on prior information, and have to employ inductive learning methods to derive models for the relationships between biological activity and chemical structure from previous observations. The development of combinatorial chemistry and high-throughput screening serves nothing else but to more rapidly provide data on biological activity for a wider range of compounds. These data have

then to be analyzed by modeling techniques. The design of chemical libraries has to focus, sooner or later, on obtaining compounds exhibiting biological activity.

Thus, in both approaches, rational drug design and combinatorial chemistry, the study of the relationships between chemical structure and biological activity is of central importance. These relationships are sought to be unraveled by learning techniques such as statistical and pattern recognition methods or neural networks. In this endeavor, the representation of chemical structure plays a major role.

## HIERARCHY OF PRESENTATIONS

Chemists have developed a variety of methods for representing and communicating structure information. The most widely used, international language is the structural formula; it is still the method of choice when representing chemical reactions. For a more in-depth analysis, three-dimensional molecular models are built, either by mechanical molecular model kits, or, increasingly by computer modeling. A variety of representations is available, from framework, through ball and stick, to space-filling models. An even more refined analysis of molecules, particularly when studying biological activity, has to consider molecular surfaces, surface properties, and molecular potentials and fields.

All these various representations of chemical structures have to be translated into a form amenable to computer manipulation. A further requirement set by the use of learning methods is that molecules have to be represented by the same number of descriptors, irrespective of their size, the number of atoms in a molecule. Only then can datasets of different molecules be automatically processed by statistical methods or neural networks.

In the following, we will present various techniques for encoding these different forms that the chemists use for structure representation. We will briefly mention the encoding of the constitution (topology) of a molecule but mainly concentrate on the representation of 3D structures and of molecular surfaces. These different encoding methods have been developed for the different requirements made by the intended applications. Furthermore, the kind of coding method to be chosen will also be strongly dictated by the size of the datasets that have to be studied.

## NEURAL NETWORKS

Our group has a long history of applying chemometrics, such as statistical or pattern recognition methods, to understand chemical information. In recent years, however, we have largely concentrated on using neural networks for this purpose because of the great potential of neural networks for projection, clustering, and modeling.<sup>1,2</sup>

A detailed discussion of neural networks clearly is beyond the scope of this presentation; a textbook on applications in chemistry is available.<sup>1</sup> Suffice to say that neural networks can do both unsupervised and supervised learning. For unsupervised learning we quite extensively use the self-organizing maps introduced by Kohonen. Kohonen networks are powerful similarity perception and clustering techniques. An overview of the use of Kohonen networks in drug design has recently appeared.<sup>3</sup> Basically, two types of uses of Kohonen networks in drug design have been developed, a 1:1 and an n:1 application.<sup>3</sup> In the 1:1 approach, one molecule is mapped into one network; a typical example is the analysis of molecular surfaces by a Kohonen map. In the n:1 approach, a set of n molecules is mapped into one network so as to study their similarity or diversity, or their different biological

activity (see following sections). Supervised learning is usually performed by feedforward networks with backpropagation learning or by counterpropagation networks.<sup>1,2</sup>

## PHYSICOCHEMICAL PROPERTIES

The binding of a ligand to its receptor may depend on a variety of physicochemical properties such as electrostatic potential, hydrophobicity, and hydrogen bonding potential. The coding of molecular structures should therefore incorporate these physicochemical effects. Methods for the calculation of a range of properties such as partial atomic charges,<sup>4,5</sup> measures of the inductive effect,<sup>6</sup> resonance,<sup>5</sup> or polarizability effect<sup>7</sup> have been developed. These procedures are empirical in nature and are therefore quite rapid and can be applied to large datasets. These methods have been collected in the program package PETRA (Parameter Estimation for the Treatment of Reactivity Applications).<sup>8</sup>

## CODING THE CONSTITUTION

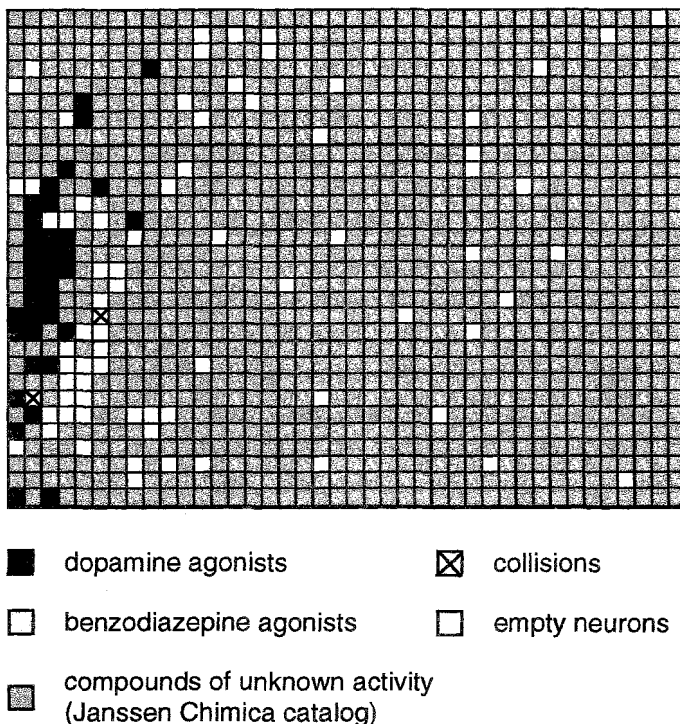
The structural formula can be considered as a mathematical graph; graph theory has therefore played a major role in the computer handling of structure information. However, the representation of a molecule as a graph, as a list of atoms and bonds does not fulfill the requirement for a fixed number of descriptors irrespective of the size of a molecule. In many applications, molecules are represented by lists of fragments, in the form of bit strings, the presence or absence of a certain functional group indicated by a 1 or 0. Such representations of a structural formula are often called 2D descriptors, however, they do not carry any direct 2D information; they are only a reflection of the constitution of a molecule, and therefore should be called topological descriptors, at most.

We have sought for methods that allow one to encode various physicochemical properties of the atoms in a molecule, such as partial charges, polarizability, etc. Our approach rests on autocorrelation functions (eq 1) introduced for structure handling by G. Moreau quite some time ago.<sup>9</sup>

$$A(d) = \sum_{j=i+1}^N \sum_{i=1}^{N-1} \delta_{ij}(d) p(i) p(j) \quad (1)$$

A value for the autocorrelation function A, at a certain topological distance (number of bonds), d, is calculated by summation over all products of a certain property, p, of atoms i and j having the required distance, d.

A range of properties such as partial atomic charges,<sup>4,5</sup> measures of the inductive effect<sup>6</sup>, resonance<sup>5</sup>, or polarizability effect<sup>7</sup> were calculated by rapid empirical methods contained in the program PETRA (Parameter Estimation for the Treatment of Reactivity Applications).<sup>8</sup> With seven such properties, p, and seven topological distances, d = 2...8, each molecule was represented by a 49-dimensional vector. It could be shown that such a representation can distinguish between dopamine agonists and benzodiazepine agonists.<sup>10</sup> The separation of these two types of molecules was even maintained after projection of this 49-dimensional space into two dimensions by a Kohonen neural network. Of even more importance is the fact that dopamine and benzodiazepine agonists could still be distinguished when contained in a datafile of more than 8,000 compounds of a chemical supplier catalog.



**Figure 1.** Kohonen map of 40x30 neurons trained with the topological autocorrelation vector of 112 dopamine agonists, 60 benzodiazepine agonists, and 8,323 structures from a chemical supplier catalog.

Figure 1 shows that dopamine and benzodiazepine agonists could nearly completely be separated (there are only two neurons (collisions) obtaining both types of compounds). Furthermore, these compounds populate only a limited area of the entire range of organic compounds covered by the chemical supplier catalog (Janssen Chimica). The two types of compounds were found in limited and separated regions of a Kohonen map.<sup>10</sup> Thus, this study showed where benzodiazepine and dopamine agonists have to be sought and in which region of chemical space no such activity is to be anticipated.

### 3D STRUCTURE

The study of the relationships between biological activity and the 3D structure of a molecule on a broad scale has been made possible by the advent of universal and efficient automatic 3D generators.<sup>11</sup>

The program system CORINA<sup>12,13</sup> developed in our group provides excellent 3D models as was shown by a comparison with X-ray structures.<sup>13</sup> CORINA is applicable to large molecules and large datasets. Figure 2 shows a 3D model of a molecule containing 999 atoms that was automatically built by CORINA from a connection table.<sup>13</sup>

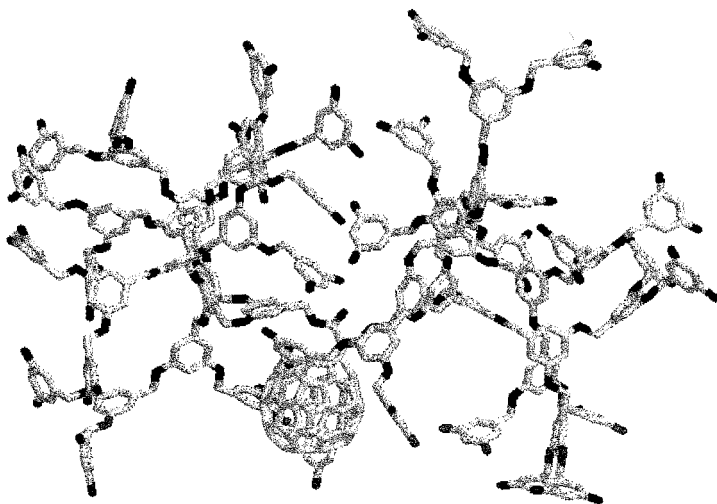


Figure 2. 3D model built by CORINA.

To illustrate the broad scope of CORINA, the database of the National Cancer Institute that was recently made public was automatically converted by CORINA. Of the 237,771 connection tables 99.8% could be converted into a 3D structure and all this took only 18,092s (0.08s / molecule) on an SGI R10000. Even the entire Beilstein file with nearly 7 million structures has been converted into 3D, again with a conversion rate of over 99%.

With a 3D structure accessible for practically any organic molecule, the problem is then, how to encode the 3D structure under the restriction of a fixed number of variables, independent of the number of atoms in a molecule.

Building on equations used for obtaining the 3D structure of a molecule from electron diffraction experiments the encoding procedure embodied in eq 2 was developed.<sup>15</sup>

$$I(s) = \sum_{j=i+1}^N \sum_{i=1}^{N-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \quad (2)$$

In this equation,  $I(s)$  is the intensity of the scattered electron beam at observation angles  $s$ ,  $A_i$  and  $A_j$  are atomic properties such as atomic number, or partial charges, and  $r_{ij}$  is the distance between the atoms  $i$  and  $j$ ;  $N$  is the number of atoms in the molecule.

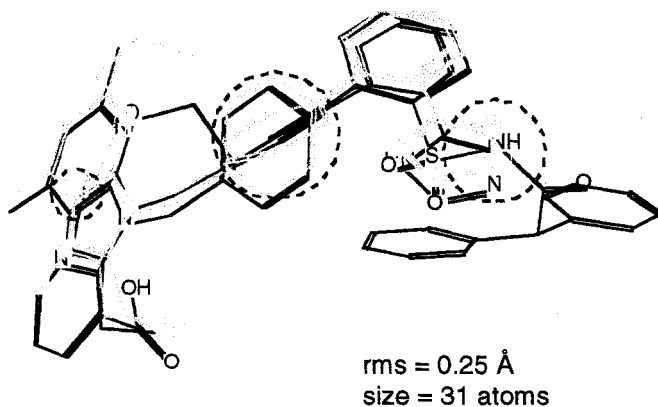
In electron diffraction, the intensity is measured and the 3D structure as given by all distances  $r_{ij}$  is derived from the intensities on the basis of eq 2. In our approach, we have turned the equation around, inputting the 3D structure of a molecule in the form of the distances  $r_{ij}$  and calculating  $I(s)$ . Furthermore, these values of  $I(s)$  are calculated only at discrete, equidistant values of  $s$ , providing a fixed, predefined number of values of  $I(s)$  which are then used as an encoding of the 3D structure of a molecule. This molecular representation was called 3D-MoRSE Code (3D Molecule Representation of Structures based on Electron diffraction).<sup>15</sup>

This 3D-MoRSE code was mainly used for the simulation of infrared spectra. However, it could also be demonstrated that this code shows great promise for correlating structure with biological activity. Dopamine D1 agonists could be separated from dopamine D2 agonists on the basis of the 3D-MoRSE code by a Kohonen network.<sup>15</sup> Furthermore, 31 steroids binding to the corticosteroid binding globulin (CBG) receptor could be clustered according to this activity in a Kohonen network.

Recent work has shown that a structure code based on radial distribution functions which is quite similar to the 3D-MoRSE code can indeed be transformed back into 3D space.<sup>16</sup> This potential of a structure code for regaining the 3D structure opens exciting possibilities.

## CONFORMATIONAL FLEXIBILITY

Indeed, biological activity is intimately tied to the 3D structure of molecules and therefore the goal in structure-activity studies should be to account for the 3D structure of molecules. However, many studies of other groups have shown that quite often 3D descriptors do not offer additional benefits to topological descriptors. The reason for this is that most molecules are flexible, attaining a variety of conformations and one cannot be sure whether the conformation investigated actually is the one picked up at the receptor. The problem is aggravated in those cases where the 3D structure of the receptor is not known - and these are still the majority of the cases. It is our belief that one can derive knowledge on the structural requirements of a ligand for binding to a receptor through systematic investigation of a series of ligands binding to one and the same receptor. In one of such approaches we have developed a method that searches for the maximum common 3D substructure (3D-MCSS) of a series of molecules through superposition of these molecules.<sup>16</sup> This superposition is achieved by the combination of a genetic algorithm with a steepest descent optimizer, the directed tweak method.<sup>18</sup> In this process, allowance is made for conformational flexibility in order to maximize the superposition. Figure 3 shows the superimposition of four ligands that inhibit the angiotensin II receptor.



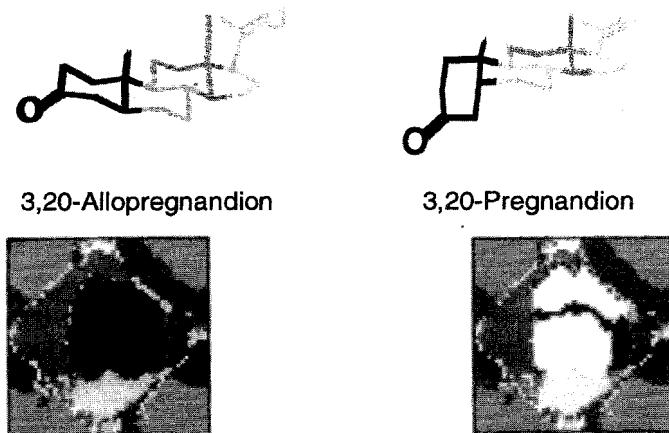
**Figure 3.** The superposition of four angiotensin II antagonists. Pharmacophore points are indicated by circles of broken lines.

## MOLECULAR SURFACES

Molecules interact with each other at molecular surfaces. This is particularly true for the interaction of a ligand binding to its receptor. The investigation of molecular surfaces, the coding of surface properties, is therefore of primary importance.

A 2D description of surface properties of a molecule was obtained by projection of molecular surfaces into a two-dimensional map by a Kohonen network.<sup>18</sup> It was shown that such maps of the electrostatic potential on a molecular surface can be used to distinguish between muscarinic and nicotinic agonists.<sup>20</sup>

The Kohonen network stores the three-dimensional coordinates of points on the molecular surface. It has been shown that such a network can be used for quantifying shape similarities in a series of compounds.<sup>21, 22</sup> The Kohonen network of one molecule can be used as a template for shape comparison. Another molecule can be sent through this network to show the differences of the shapes of these two molecules. Figure 4 illustrates this for the comparison between the molecular surfaces of 3,20-allopregnandion and 3,20-pregnandion.



**Figure 4.** Kohonen map of the molecular surface of 3,20-allopregnandion taken as template, and that of 3,20-pregnandion sent through this network. The areas where the shapes differ show up as white areas.

Autocorrelation can be used to encode surface properties. In an autocorrelation function,  $A(d)$ , a summation is made over all products of a certain property  $p$  at a point  $x$  and a second point at a distance of  $d$  from  $x$  (eq 2).<sup>4</sup> For the distance parameter,  $d$ , all distances within a certain range, e.g., between 3 and 4 Å are collected in one autocorrelation value.  $L$  is the number of distances.

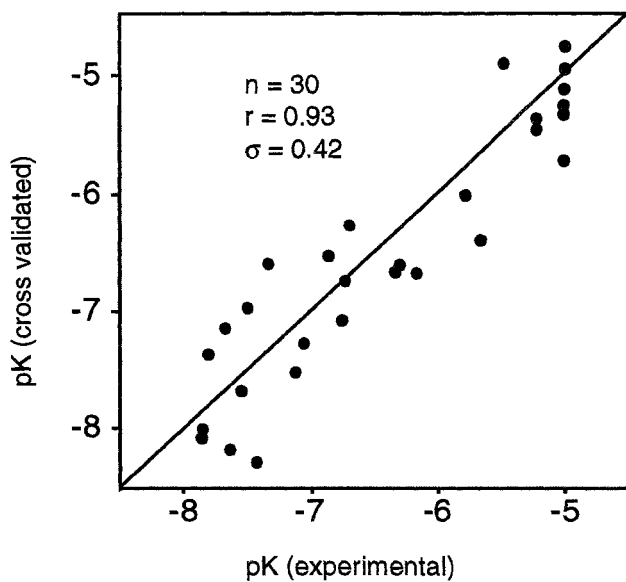
$$A(d) = \frac{1}{L} \sum_x p(x) \cdot p(x+d) \quad (3)$$

We have shown that autocorrelation of the electrostatic potential on the van der Waals surfaces into 12 descriptors provides excellent descriptors for the modeling of affinity of steroids for binding to the corticosteroid binding globulin (CBG) receptor.<sup>23</sup> In fact, the prediction of CBG binding affinity is better than the one achieved by the widely used CoMFA method. Figure 5 shows these results.

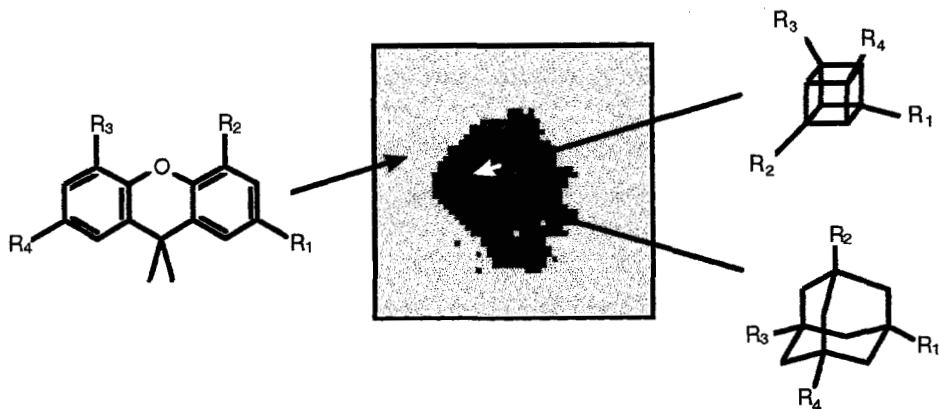
In a similar manner, autocorrelation of the hydrophobicity potential of a series of 78 polyhalogenated aromatic compounds can quantitatively model the binding of these molecules to the cytosolic Ah receptor.<sup>23</sup>

The same encoding method, autocorrelation of the molecular electrostatic potential (MEP) into 12 descriptors was used for the definition of diversity and similarity of combinatorial libraries.<sup>24</sup> As an application, the experiments of Rebek et al.<sup>25</sup> searching for a trypsin inhibitor, were investigated. Reaction of dimethylxanthene, carrying four acid chloride substituents, with 19 different amino acids provided a library of a maximum of 65,341 compounds. The same experiment with the cubane skeleton carrying four acid chloride residues provided up to 11,191 compounds. The autocorrelation of the MEP of these compounds showed that these two sets of structures have to be considered as diverse. On the other hand, a library of 11,191 compounds obtained from these 19 amino acids and adamantane carrying four acid chloride groups is, on the basis of the autocorrelation vector of MEP, highly similar to the library of cubane compounds.<sup>24</sup> Figure 6 shows a Kohonen map of 50x50 neurons trained with 65,341 dimethylxanthene, 11,191, cubane, and 11,191 adamantane derivatives. The dimethylxanthene compounds are quite well separated from the other molecules, whereas the cubane and adamantane libraries quite extensively overlap.





**Figure 5.** Predictive power of CBG activity by a feedforward neural network trained with steroids represented by 12-dimensional autocorrelation descriptors.



**Figure 6.** Kohonen map of three different libraries showing both similarity and dissimilarity of compounds.

This encoding method has also great benefits for planning a strategy for deconvolution experiments.<sup>23</sup> These investigations showed that with the methods developed here such large datasets can be handled with a moderate amount of computation times.

## SUMMARY

Approaches to the encoding of molecular structures have been developed that allow the investigation of datasets of diverse molecules by learning methods. These structure representations form a hierarchy of increasing sophistication. The level used will largely be dictated by the size of the dataset to be investigated. Representations of the constitution will be applied to datasets comprising millions of structures, whereas representations of molecular surface properties can still be chosen for datasets comprising 100,000 and more structures.

Even with large datasets these methods are rapid enough to be performed on small workstations with computation times of a few hours. Consideration of conformational flexibility is presently limited to smaller sets of structures.

## ACKNOWLEDGEMENTS

The methods were developed by able coworkers whose names appear in the references. Our research was recently funded in particular by the Federal Minister of Research and Development (BMBF). Cooperation with Merck KGaA (G. Barnickel, S. Anzali, M. Krug), BASF AG (G. Klebe, T. Mietzner) and University of Tübingen (A. Zell, H. Siemens) is gratefully acknowledged.

## REFERENCES

1. J. Zupan, J. Gasteiger, *Neural Networks for Chemists: An Introduction*, VCH-Verlag, Weinheim, (1993).
2. J. Gasteiger, J. Zupan, *Neural Networks in Chemistry*, *Angew. Chem.* 105:510 (1993), *Angew. Chem. Intern. Ed. Engl.* 32:503 (1993).
3. S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup, M. Wagener, The Use of Self-Organizing Neural Networks in Drug Design, in: *3D QSAR in Drug Design - Volume 2*, p. 273-299, H. Kubinyi, G. Folkers, Y. C. Martin, Eds., Kluwer/ESCOM, Dordrecht, NL, (1998).
4. J. Gasteiger, M. Marsili, Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges, *Tetrahedron* 36: 3219 (1980).
5. J. Gasteiger, H. Saller, Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept, *Angew. Chem.* 97:699 (1985), *Angew. Chem. Intern. Ed. Engl.* 24:687 (1985).
6. M. G. Hutchings, J. Gasteiger, Residual Electronegativity - An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines, *Tetrahedron Lett.* 24:2541 (1983).
7. J. Gasteiger, M. G. Hutchings, Quantification of Effective Polarisability. Applications to Studies of X-Ray, Photoelectron Spectroscopy and Alkylamine Protonation, *J. Chem. Soc. Perkin 2*, 559 (1984).
8. <http://www2.ccc.uni-erlangen.de/PETRA>.
9. G. Moreau, P. Broto, Autocorrelation of molecular structures: Application to SAR studies, *Nouv. J. Chim.* 4:757 (1980).
10. H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists, *J. Chem. Inf. Comput. Sci.* 36: 1205 (1996).
11. J. Sadowski, J. Gasteiger, From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders, *Chem. Reviews* 93:2567 (1993).
12. J. Gasteiger, C. Rudolph, J. Sadowski, Automatic Generation of 3D-Atomic Coordinates for Organic Molecules, *Tetrahedron Comput. Method.* 3:537 (1992).
13. J. Sadowski, J. Gasteiger, G. Klebe, Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures, *J. Chem. Inf. Comput. Sci.* 34:1000 (1994).
14. CORINA can be accessed on the internet:  
<http://www2.ccc.uni-erlangen.de/services/3d.html> and <http://www.mol-net.de>
15. J. H. Schuur, P. Selzer, J. Gasteiger, The Coding of the Three-dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure - Spectra Correlations and Studies of Biological Activity, *J. Chem. Inf. Comput. Sci.* 36:334 (1996).
16. M. C. Hemmer, V. Steinhauer, J. Gasteiger, The Prediction of the 3D Structure of Organic Molecules from Their Infrared Spectra, *Vibr. Spectr.*, in press.
17. S. Handschuh, M. Wagener, J. Gasteiger, Superposition of Three-Dimensional Chemical Structures Allowing for Conformational Flexibility by a Hybrid Method, *J. Chem. Inf. Comput. Sci.* 38:220 (1998).
18. T. Hurst, Flexible 3D Searching: The Directed Tweak Technique, *J. Chem. Inf. Comput. Sci.* 34:190 (1994).
19. J. Gasteiger, X. Li, C. Rudolph, J. Sadowski, J. Zupan, Representation of Molecular Electrostatic Potentials by Topological Feature Maps, *J. Am. Chem. Soc.* 116:4608 (1994).

20. J. Gasteiger, X. Li, Mapping the Electrostatic Potential of Muscarinic and Nicotinic Agonists with Artificial Neural Networks, *Angew. Chem.* 106:671 (1994), *Angew. Chem. Int. Ed. Engl.* 33:643 (1994).
21. S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger, J. Polanski, The Comparison of Geometric and Electronic Properties of Molecular Surfaces by Neural Networks: Application to the Analysis of Corticosteroid Binding Globulin Activity of Steroids, *J. Comput.-Aided Mol. Design* 10:521 (1996).
22. J. Polanski, J. Gasteiger, M. Wagener, J. Sadowski, The Comparison of Molecular Surfaces by Neural Networks and its Application to Quantitative Structure Activity Studies, *Quant. Struct. -Act. Relat.* 17:27 (1998).
23. M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks, *J. Am. Chem. Soc.* 117:7769 (1995).
24. J. Sadowski, M. Wagener, J. Gasteiger, Assessing Similarity and Diversity of Combinatorial Libraries by Spatial Autocorrelation Functions and Neural Networks, *Angew. Chem.* 107:2892 (1995), *Angew. Chem. Int. Ed. Engl.* 34:2674 (1995).
25. T. Carell, E. A. Wintner, A. Bashir-Hashemi, J. Rebek, A Novel Procedure for Synthesis of Libraries containing Small Organic Molecules *Angew. Chem.* 106:2159 (1994), *Angew. Chem. Int. Ed. Engl.* 33: 2059 (1994).
- T. Carell, E. A. Wintner, A. Bashir-Hashemi, J. Rebek, A Solution-phase Screening Procedure for the Isolation of Active Compounds from a Library of Molecules, *ibid*, *Angew. Chem* 106:2162 (1994), *J. Angew. Chem. Int. Ed. Engl.* 33:2062 (1994).

## FRAGMENT-BASED SCREENING OF LIGAND DATABASES

Christian Lemmen<sup>1</sup> and Thomas Lengauer<sup>1</sup>

<sup>1</sup>GMD – German National Research Center for Information Technology  
Institute for Algorithms and Scientific Computing (SCAI)  
Schloß Birlinghoven, 53754 Sankt Augustin, Germany

### Introduction

Lead structure identification in databases of substantial size is a difficult task in computer-aided drug design. Several approaches to this problem have been suggested in the literature [1], including fingerprints, substructure matching, rigid-body superposition, grid- and graph-based field overlay, and flexible fitting. Although 2D molecular descriptors outperformed their 3D analogues in a comparative study [2], current research efforts focus on 3D approaches. Despite recent advances [3], especially the appropriate consideration of molecular flexibility remains a challenging problem. In general, two kinds of approaches have been used to tackle this problem. Either a multi-conformer database is compiled in a preprocessing step, at the expense of increased memory requirements, or conformers are generated on the fly at the expense of increased computational costs. However, both of these approaches are only compromises, since in any case only a limited number of conformers can be considered. We propose a novel strategy for database screening on the basis of molecular fragments. From the computational point of view, a fragment-based approach is especially effective for two reasons. First, fragments often comprise only a small number of atoms and the runtime usually depends linearly on this quantity. Thus, a fragment-based search should be quite fast. Second, fragments frequently show only limited flexibility, thus allowing to ignore the flexibility in a first order approximation.

Klopman [4] demonstrated that fragment-based discrimination of active and inactive molecules is frequently possible. However, while Klopman employs substructure search and focuses on the combination of absent/present fragments in order to discriminate between active and inactive molecules, we focus on the similarity search on the basis of fragments. Since the physico-chemical characteristics rather than the atomic structure of a fragment determines its activity, or contribution to an activity, it appears appropriate to rate activity on the basis of fragment similarity.

The result of our fragment-based database screening is twofold: it comprises a similarity score and a structural alignment witnessing this score that is worthwhile to be analyzed separately. Our similarity approach called RIGFIT is based on a Fourier space alignment technique described previously [5]. In a preliminary screening experiment we already demonstrated the usefulness of RIGFIT in two respects. First, the approach allows for a user-adjusted tradeoff between accuracy and efficiency. This allows for rapid processing at low resolution which is appropriate for screening experiments. Second, since RIGFIT optimizes rotation and translation in two successive independent steps, the latter of which is extremely fast, it allows for dense sampling of starting points for translational optimization. This, in turn, is a prerequisite for successful fragment handling. Utilizing a computational shortcut known from *Fast Fourier Transform*, we recently further increased the RIGFIT performance by about a factor

of three. RIGFIT superposes a fragment with another molecule in about 1 second/structure\* on a common day workstation. An optional second screening step may be performed by our flexible superpositioning tool FLEXS [6]. FLEXS superposes pairs of molecules and takes full flexibility of one of the structures into account. The average computing time of FLEXS is in the range of 1 minute/structure†. RIGFIT is implemented as part of FLEXS which is available on the WWW‡.

## Methods

Both our search engines are described in detail elsewhere [5, 6]. The idea is to use these tools in a two-step screening approach. Here, we will focus on the rapid fragment-based screening that may be performed using RIGFIT. The application scenario we consider can be described as the following task. Given a ligand exhibiting a desired property and a 3D database to be screened, detect analogs to the query structure in the database.

### Fragment based screening

RIGFIT optimizes the common volume of two molecules expressed by various Gaussian functions associated to different physico-chemical properties. The basic algorithmic idea in RIGFIT originates from X-ray crystallography, and uses the concept of the *Patterson function*. One way to approach the well known *phase problem* in crystal structure determination is to consider Patterson densities instead of real space electron densities. Since the Patterson function contains only information about interatomic distances, this description is independent of the translation of the molecule. By transforming the Gaussians to Fourier space and neglecting the phases artificially, we mimic the *molecular replacement* approach of X-ray structure determination and reveal a translation-independent description of the molecules. To compare two molecules, we evaluate the similarity measure proposed by Hodgkin [7]. Since the measure derived is invariant under translation, rotation can be optimized separately [8]. Thus, the six-dimensional search (as performed, e.g., in the SEAL system [9, 10]) is divided into two successive three-dimensional searches which inherently speeds up the optimization process. After determining the local optima of the rotation function, we optimize the translation in a second independent step. This optimization is carried out in Fourier space, utilizing the *convolution theorem*, and is extremely efficient.

For the purpose of efficiency, we employ another concept originating from X-ray crystallography. We treat our molecules as being located in a virtually infinite lattice of replications. In this way, the Fourier transform of a real space density function becomes discrete, i.e. it is non-zero only for integral points in Fourier space (*Laue vectors*). Thus, the computation of an integral simplifies to the summation of function values for the Laue vectors. Furthermore, if not the entire set of Laue vectors but rather a spherical region around the origin is considered, the high-frequency contributions are removed from the Fourier series. As a consequence, the computational costs for the summation decrease and the scoring function becomes smoother. Thus, it is possible to trade off accuracy in the density description against computing speed. During objective function evaluation in Fourier space, it is necessary to calculate exponential terms  $e^{hx}$  on a regular grid of Laue vectors  $h$ . Thus, effectively, one needs to evaluate  $e^{(h_0+n\Delta h)x}$  with a constant offset  $\Delta h$  (grid spacing). We transform

$$E_n := e^{(h_0+n\Delta h)x} = e^{h_0x} \cdot \prod_n e^{\Delta hx}$$

and precompute  $E_0 := e^{h_0x}$  and  $\Delta E := e^{\Delta hx}$ . This enables us to substitute computationally expensive, successive exponential calculations by a single multiplication

$$E_n = E_{n-1} \cdot \Delta E.$$

Some computational overhead is necessary in order to provide a sequence of exponential calculations that best benefits from the novel incremental evaluation strategy. Note that

---

\*i.e., about 86,400 structures/day

†i.e., about 1440 structures/day

‡<http://cartan.gmd.de/FlexS>

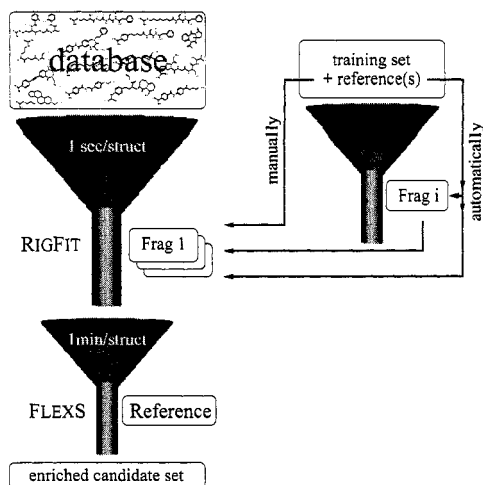


Figure 1. The two step screening strategy comprises RIGFIT as a first efficient filter, using fragment-based superpositioning. The optional second filter FLEXS is based on flexible alignment. The selection of a screening fragment (right hand side) may either be performed manually, automatically using a set of rules, or automatically based on the performance on a training set.

in three dimensions at least three increments  $\Delta h_x$ ,  $\Delta h_y$ , and  $\Delta h_z$  have to be considered. However, for the usual range of numbers of Laue vectors that we consider (from 50 up to 400), the computational savings amount to about a factor of three.

### Optional flexible fitting

Our approach FLEXS follows a combinatorial approach to solving the ligand superposition problem [11]. It allows to fit a flexible *test ligand* onto a rigid *reference ligand* applying the following protocol: First, the flexible ligand is decomposed into small and relatively rigid portions (fragments). Second, an anchor fragment of the test ligand is selected. Third, using a discrete surface approximation, possible positions of the anchor on top of the reference molecule are determined. Finally, in an iterative incremental construction procedure, the anchor placements are extended by adding the remaining fragments of the test ligand step by step considering a discrete set of possible conformations for each fragment. The number of partial placements, generated in this way, grows exponentially with the number of added fragments. A greedy strategy is applied in each iteration to select a suitable subset of placements which is carried into the next iteration.

### Screening strategy

A complete screening is carried out by the procedure illustrated in Figure 1. RIGFIT is used to superimpose a given fragment (or a set of fragments) onto every structure in the database. This results in a similarity score and a structural alignment. Sorting the structures by score (or any combination of scores for the different fragments) and applying a minimum threshold allows to filter out some fraction of the database. Supposedly, the molecules with similar fragment characteristics are top-ranking and subjected to a second screening step using FLEXS as the filter. Depending on the size of the database to be screened (cf. below), we extracted the 5 - 10% top ranking hits for the second filter.

### Determining a screening fragment

We experimented with three different approaches to determining a screening fragment. First, the user may specify one or more fragments of the query structure. Second, the fragment selection may be performed automatically by the program. Finally, if a set of active molecules (training set) is available, a test-DB is utilized to train the automatic fragment selection procedure, mentioned above, in order to suggest those fragments that discriminate best between the training set and the remainder of the test-DB. Subsequently, this selection of fragments

is processed as above. An illustration of the fragment selection procedure is given in the right half of Figure 1.

## Results

We have used two databases for our tests. The first one is based on a data set assembled by Briem et al. [12]. It contains 972 ligands from the MDDR DB<sup>§</sup> which are grouped into five activity classes<sup>¶</sup> and a sixth class of randomly selected drug molecules. We augmented this data set with a set (RGD) of 12 fibrinogen receptor antagonists taken from the literature. The second database is the complete NCI database with 121,491 molecules<sup>||</sup>.

We performed six filter experiments on each of these databases, one for each of the activity classes RGD, ACE, 5HT3, HMG, PAF, and TX2. In each case we selected one reference molecule in the class and screened the database for molecules that are similar to the reference molecule. We tried this procedure with several reference molecules in each activity class. Here we report results on reference molecules that worked especially well.

Table 1 summarizes the results of our tests. Column 1 of the table names the class and depicts the reference structure. The fragments of the reference molecule that are used to screen the database with RIGFIT have a white background. The remainder of the molecule is before a grey background.

Columns 2 and 3 show enrichment factors achieved by using our screening procedure on the two databases. We calculate enrichment factors with the formula

$$E_A(p) = \frac{N_A(p)/p}{N_A/N}$$

Here,  $N_A(p)$  is the number of molecules of a certain activity class  $A$  among the top-ranking  $p\%$  compounds of the database (hit-rate).  $N_A$  is the number of molecules of activity class  $A$  in the entire database and  $N$  is the size of the database. Thus,  $E_A(p)$  counts the factor of how many more active molecules are found among the first  $p\%$  of the database after screening, than according to a uniform distribution of the active molecules across the database. The charts in Table 1 display the enrichment factors achieved with RIGFIT (solid line), the enrichment factors achieved with the DAYLIGHT fingerprints\*\* (short dashes), the maximum enrichment factor possible (dotted line), and the fraction of all active molecules contained in the respective portion of the database (long dashes). The  $y$ -axis on the left hand side shows the enrichment factor. The percentage level for the fourth curve is provided by the  $y$ -axis on the right hand side. Additionally, the decrease/increase in performance using FLEXS as a second filter is indicated by small arrows originating from the RIGFIT curve. The respective percentage fraction of the database considered is given along the  $x$ -axis.

In addition to providing the charts, we calculated the median  $M_A$  of the ranks of the active molecules

$$M_A = R_A\left(\frac{N_A + 1}{2}\right) \text{ if } N_A \text{ odd, and } M_A = \frac{1}{2}\left(R_A\left(\frac{N_A}{2}\right) + R_A\left(\frac{N_A}{2} + 1\right)\right) \text{ if } N_A \text{ even.}$$

Herein,  $R_A(c)$  is the rank of the  $c$ -th active compound that we find in the list of all compounds as we go through it by decreasing rank. Thus, the median displays the rank by which 50% of the active molecules are covered. The changes in the rank of the median from using only the first to using both filters is provided in the upper right hand corner of each chart. In column 4 we show an interesting *new* molecule, i.e. a molecule that we found by visual inspection of the high ranking structures that did not belong to the activity class from which the reference molecule was taken.

It can be seen that in three cases the curve resulting from RIGFIT screens (solid lines) is above the curve corresponding to DAYLIGHT fingerprint based screening (short dashes), in two cases it is below (PAF, TXA2), and in one case both methods perform similarly (HMG).

<sup>§</sup>MDL Information Systems Inc., San Leandro, CA, USA. *MACCS Drug Data Report (MDDR)*

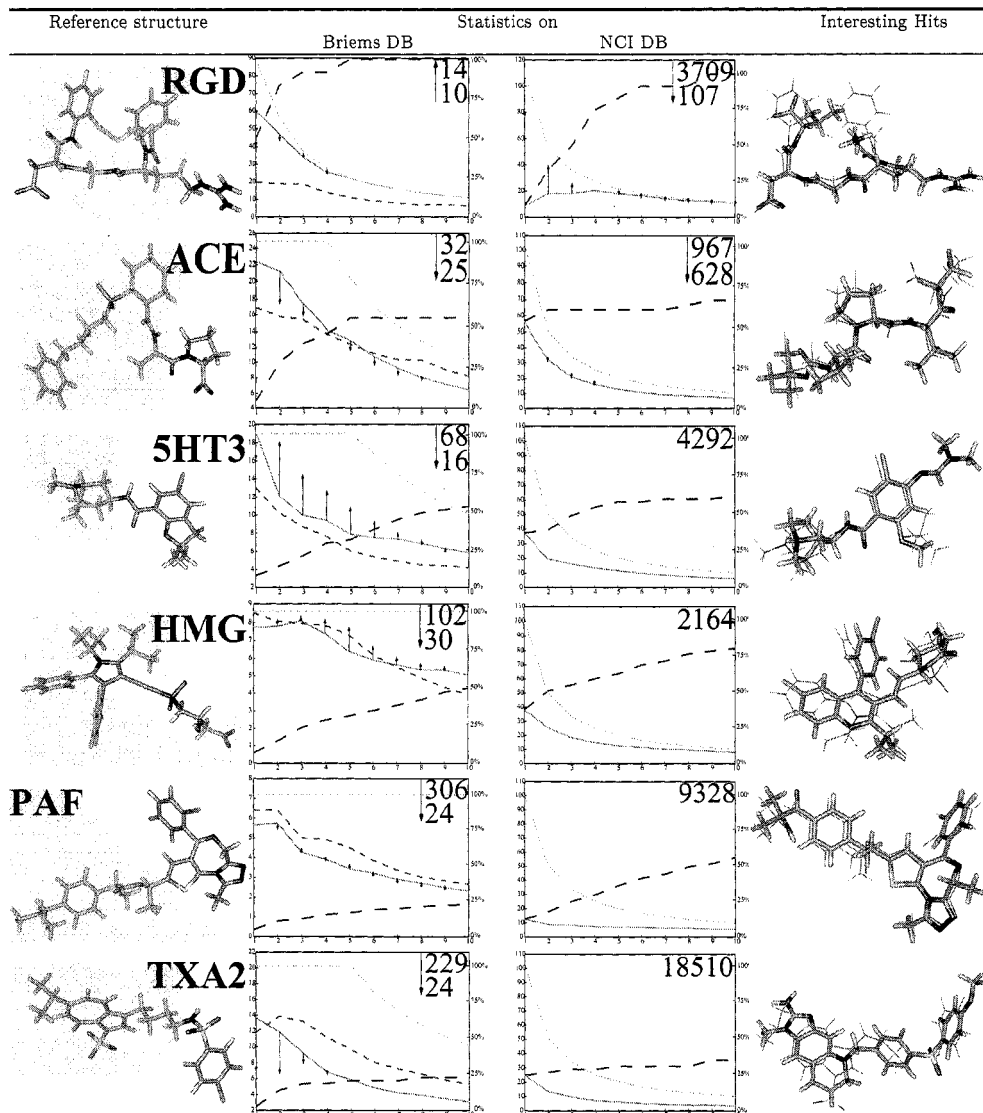
<sup>¶</sup>40 angiotensin-converting enzyme (ACE) inhibitors, 136 PAF receptor antagonists, 49 Thromboxane A2 (TXA2) receptor antagonists, 114 HMG CoA reductase inhibitors, 52 5HT3 receptor antagonists

<sup>||</sup>Before conversion to SYBYL mol2 and validation of structures, the NCI DB release from 07/01/98 (<http://epnws1.ncifcrf.gov:2345/dis3d/3ddatabase/pubstruc.html>) contained 126,710 structures

\*\*DAYLIGHT Inc., Mission Viejo, California, USA. *DAYLIGHT Software Manual*, 1994



Table 1. Screening Results



The first column shows the reference molecule (one for each activity class) and the selected screening fragments (non-shaded region). Columns two and three display the overall result statistics on Briems dataset and our compiled version of the NCI database, respectively (see the text for a discussion of the different curves). The fourth column shows an interesting hit found during the screening experiments (sticks model) that does not belong to the original activity class, superposed by FLEXS onto the reference ligand (lines model).

In the latter case, however, the performances are already close to their limit. Generally, the performance is quite high. The best enrichments are found with the RGD dataset. In this case, the selection of fragments has been performed manually and comprises two functional groups, carboxylate and guanidinium. However, also in the 5HT3 example the maximum performance is reached within the first percentile of the database. The weakest performance (with an enrichment of 13.7) was found in the TXA2 example. The best of all fragments (a benzol-sulfonamide group) in this case is still relatively unspecific and obviously not capable to display the key characteristic of this dataset. The top 100 hits of the RIGFIT screen have been subjected to flexible superpositioning onto the reference structure by FLEXS. The enrichments of the ACE and TXA2 examples dropped, while those on 5HT3 and HMG increased, and those on RGD and PAF remained almost unaltered. However, it is difficult to compare the

enrichments on the same scale here. The median of the ranks (provided in the upper right corner) of five out of six examples drops (indicated by the arrow). Thus, we reveal further enrichment even on this very small subset of 100 examples. We found that the structure containing the best screening fragment need not necessarily be the best possible choice of a reference structure for flexible superpositioning.

Generally, results carry over convincingly to the large dataset. With 56.4, 38.1, 37.3, 25.0, 12.6, and 9.1, respectively, remarkable enrichments are achieved. Of course, the maximum possible performance appears to be inaccessible here. However, the curves showing the accumulated active hits found (long dashes), as well as the median indicate a comparatively high rate of active molecules within a fraction of the database that is small enough to allow for applying FLEXS. We performed the second filter step on two examples (ACE and RGD) using the 5,000 top ranking candidates. Again, the arrows indicate the increase/decrease in performance. It can be seen that the enrichments of the RGD example significantly increase. In both cases also the median decreases significantly, thus indicating for further substantial enrichment by the second filter.

Among the hits not contained in the original activity classes (column 4 in the table) the most interesting one is part of the random set contained in Briems database. It ideally fits on top of the RGD reference structure. We found that this compound is labeled with 'platelet aggregation properties' in the MDDR and thus can be assumed to bind to the fibrinogen receptor. However, several other interesting hits have been found for the other examples showing the potential of our method to detecting analogs in a database.

## Conclusions and outlook

Fragment-based database screening has proven to be fast and effective. So far, we did not exploit the positioning of a fragment provided by the RIGFIT approach. Also, we made only limited use of combining fragment scores. It is to be expected that both these options further increase the performance of our approach. Also, one could employ a standard basis set of fragments and use the RIGFIT scores as a fingerprint, thus coding the fragment characteristics of a molecule.

## References

- [1] P. Willett. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J. Mol. Recognition*, 8:290-303, 1995.
- [2] R.D. Brown and Y.C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36:572-584, 1996.
- [3] D.A. Thorner, D.J. Wild, P. Willett, and P.M. Wright. Similarity searching in files of three-dimensional chemical structures: Flexible field-based searching of molecular electrostatic potentials. *J. Chem. Inf. Comput. Sci.*, 36:900-908, 1996.
- [4] G. Klopman. MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.*, 11:176-184, 1992.
- [5] C. Lemmen, C. Hiller, and T. Lengauer. RIGFIT: A new approach to superimpose ligand molecules. *J. Comput.-Aided Mol. Des.*, 12, 1998. In press.
- [6] C. Lemmen and T. Lengauer. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Des.*, 11:357-368, 1997.
- [7] E.E. Hodgkin and G. Richards. Quantum biology symposium 14. *Int. J. Quant. Chem., Quantum Biol. Sympos.*, 14:105-110, 1987.
- [8] M.G. Rossmann and D.M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.*, 15:24-31, 1962.
- [9] S.K. Kearsley and G.M. Smith. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.*, 3:615-633, 1990.
- [10] G. Klebe, T. Mietzner, and F. Weber. Different approaches toward an automatic structural alignment of drug molecules: Application to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput.-Aided Mol. Des.*, 8:751-778, 1994.
- [11] C. Lemmen, T. Lengauer, and G. Klebe. FLEXS: A method for fast flexible ligand superposition. *J. Med. Chem.*, 1998. Submitted.
- [12] H. Briem and I.D. Kuntz. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.*, 39:3401-3408, 1996.

## THE COMPUTER SIMULATION OF HIGH THROUGHPUT SCREENING OF BIOACTIVE MOLECULES

Frank R. Burden<sup>1</sup> and David A. Winkler<sup>2</sup>

<sup>1</sup>Department of Chemistry, Monash University, Clayton 3168, Australia

<sup>2</sup>CSIRO Division of Molecular Science, Private Bag 10, Clayton South MDC, Clayton 3169, Australia

### INTRODUCTION

As the activity of synthesised bioactive compounds increases, it becomes more difficult to discover new chemical entities with substantial advantages. The average number of compounds synthesized in order to obtain a commercial candidate has risen from 10,000 to around 40-50,000. The recently-developed combinatorial methods greatly increase the numbers of compounds synthesized and tested but generate very large amounts of data. Clearly it has become very important to find new methods for extracting useful molecular design information from these large quantities of structure-activity data. The data sets which derive from combinatorial chemistry and high throughput screening are often so massive that QSAR is the method of choice. The method, using multivariate statistics, was developed by Hansch and Fujita<sup>1</sup>, and it has been successfully applied to many drug and agrochemical design problems

QSAR has advantages of speed and simplicity and it can, in some cases, account for some transport and metabolic processes which occur once the compound is administered. Hence, the method is often applicable to the analysis of *in vivo* data. However classical QSAR has limitations in that it cannot handle stereoisomers, cannot correlate compounds where the base structure varies widely, cannot implicitly handle non-linear dependencies and interaction terms between the parameters, and QSAR analyses can be difficult to interpret in terms of mechanism at the molecular level. New QSAR methods have been developed recently which overcome some of these shortcomings. This paper discusses several of these novel molecular representations, the use of Bayesian regularised neural networks in SAR, and the application of these to bioactive compound design and the simulation of combinatorial discovery by the screening of large existing databases. Some of these are huge with up to  $10^{12}$  compounds though this may be only a small fraction of a combinatorial universe containing more than  $10^{100}$  compounds.

### SIMPLE MOLECULAR REPRESENTATIONS

Many types of molecular representation have been proposed, from Hansch parameters to chemical graph-based methods<sup>2,3,4</sup>. Recently several new representations have been devised: atomistic counts<sup>5</sup>; molecular eigenvalues<sup>6</sup>; E-state fields<sup>7</sup>; topological

autocorrelation vectors<sup>8</sup>; various molecular fragment-based hash codes<sup>9,10</sup>; and molecular holograms<sup>11</sup>. These representations may have advantages in speed of computation, in more accurately representing molecular properties most relevant to receptor activity, or in being more generally applicable to diverse chemical classes acting at a common receptor, than the traditional representations.

### **Atomistic Representations**

In this deceptively simple approach<sup>5</sup>, molecules are represented simply by counting the numbers of atoms of specific elemental type, with specific numbers of connections (a measure of the hybridization). Although simple this representation is adequate to encode not only physicochemical parameters, such as lipophilicity and molar refractivity, but also biological activity (DHFR inhibition<sup>5</sup>). The fact that steric and lipophilic factors are often important in drug receptor interactions provides a partial explanation as to how such a simple representation may work.

### **Molecular Eigenvalues**

A previous version of this eigenvalue index<sup>12</sup> has been developed further by Pearlmann to become the BCUT (Burden, CAS, University of Texas) index<sup>13</sup>. In the present context the eigenvalue indices can be thought of as quantifying the most electronegative and electropositive atoms in the the molecules. This comes about because the diagonal elements of the modified adjacency matrix have been ascribed atom specific values while the off-diagonal elements have values proportional to the bond orders. The diagonalization process in effect ascribes the bond electrons back to the atoms (the trace of the matrix remaining invariant).

### **Molecular Multipole Moments**

Both of the above representations are simple to implement for very large numbers of compounds with diverse structures. However, a recent paper by Platt and Silverman<sup>14,15</sup> introduced a third general representation which is intuitively appealing. They generated the zero-, first- and second-order molecular multipole moments with respect to atomic mass, and atomic charge. We are currently working on a variant of Silverman & Platt's method which makes use of the principal pseudo-moments of inertia and associated axes (all atoms have unit mass) and also generating analogous 'lipophilic' molecular multipole moments (hydropoles), by utilizing the hydrophobic atom constant approach of Abraham and Leo<sup>16,17</sup>.

### **Molecular Hologram Generation**

A very recent development is the molecular hologram, which is derived from a common strategy to increase the efficiency of database searching by translation of chemical structure representations into binary bit strings, known as fingerprints. Several approaches to fingerprinting have been implemented within commercial software<sup>18</sup>.

The PLS technique is then used to generate a statistical model that relates the descriptor variables (occupancy numbers of the bins in the hologram) to an observable property, for example the biological activity expressed as  $-\log IC_{50}$ . The predictive power of the model is determined by using statistical cross validation using a number of cross validation groups. For the final model, the QSAR analysis is redone with the number of components set to the optimal number of components

Molecular holograms, eigenvalue descriptors, molecular multipole moments, chemical graph theory, and several other developments have significantly improved the mathematical description of molecules for use in SAR studies and rational design.

## **BAYESIAN REGULARISED ARTIFICIAL NEURAL NETWORKS (BRANNs)**

Artificial neural networks are computer-based mathematical models developed to have analogous functions to idealised simple biological nervous systems. They consist of layers of processing elements (neurodes), which are considered to be analogous to the nerve cells (neurons) and these are interconnected to form a network which simulates a parallel computer<sup>30</sup>.

We have recently investigated the use of Bayesian regularization in artificial neural nets<sup>11</sup>. Using Bayesian regularisation<sup>19</sup> removes the need to supply a validation set since it minimizes a linear combination of squared errors and weights. It also modifies the linear combination so that at the end of training the resulting network has good generalization qualities. It has also been suggested that there is no need for a testing set since the application of the Bayesian statistics provides a network that has maximum generalisation. Our study used a network architecture with 3 hidden nodes which proved to be more than sufficient in all cases with the Bayesian regularisation method estimating the number of effective parameters. The concerns about overfitting and overtraining are also removed by this method so that the production of a definitive and reproducible model is attained. A minor problem of instability which is caused by the provision of randomised weights at the start of training which can be overcome by training a number of nets with different starting weights and selecting the best standard error of prediction. It has been found that those networks that converge to finite weights produce near identical answers.

## **Simulation of Combinatorial Discovery**

The development of combinatorial chemistry, and the resultant large increases in the numbers of chemical entities screened for drug activity, has resulted in a paradigm shift in the way new drug leads are discovered. It is now possible to generate millions of chemical analogues in a relatively short time with greatly reduced effort. Rapid screening techniques have necessarily emerged to keep pace with the generation of new combinatorial libraries. It is now routine to carry out 40-50,000 screening events per week with a small number of staff<sup>20</sup>.

As powerful as these new methods of combinatorial and mass screening are, they are still only capable of accessing a very small region of the 'universe' of possible chemistries. Estimates of the 'universe' of chemical compounds that it is possible to synthesize by combinatorial methods range from  $10^{60}$  -  $10^{400}$ , numbers so vast that only a minute fraction could conceivably be generated and tested by combinatorial methods. This recognition is driving the quest for methods of simulation of combinatorial synthesis and

high throughput screening *in silico*. Methods to allow exploration of much larger region of combinatorial space would be of considerable interest in allowing a focusing of the combinatorial chemistry effort into chemical species with inherent novelty and receptor efficacy.

Recently, Ho and Marshall<sup>21</sup> described a technique for generating very large databases, representing a 'virtual' combinatorial library, using a procedure they called DBMaker using permutations of SMILES strings. Tripos Associates have used an alternative approach which exploits simple chemistries and commercially-available building blocks to generate a 3D database. This ChemSpace<sup>®</sup> database contains approximately 1 *trillion* chemical structures for use in similarity and pharmacophore searches, approximately 50,000 times more than all the compounds in CAS.

We have utilized the concept of a QSAR model as a 'virtual receptor' to allow rapid screening of these 'virtual' combinatorial libraries. We are working towards the development of computationally cheap, simple molecular representations for use in these studies involving large data sets which, coupled with developments in neural networks, will facilitate the generation of receptor surrogates with useful properties. Our implementation of virtual receptors involves a trained neural network and simple molecular representations which would be capable of rapidly evaluating large numbers of compounds for possible activity against the receptor type. Such virtual receptors would be useful screening paradigms for finding leads in large virtual databases. We have investigated this possibility using ANNs and found the approach feasible. More recently a number of different approaches to the library design and virtual screening have been reported<sup>24-29</sup>. Tropsha and his group have developed a method (Focus-2D) for searching virtual libraries for structures similar to biologically active compounds using simulated annealing and topological descriptors<sup>24</sup>. Shi et al have used genetic function approximations to carry out QSAR studies in the NCI database which describe antitumour activity patterns<sup>25</sup>. Screening of virtual libraries using 3D steric and electronic grids has been reported by Lui et al<sup>26</sup>. Horvath<sup>27</sup> has automated the conformational analysis and active site docking of a 2500 library of potential trypanothione reductase inhibitors. Vedani, Dobler and Zbinden<sup>28</sup> have developed a quasi atomistic receptor model for use in screening of libraries which defines a pseudo receptor surface with properties which adapt to the requirements of the training set. Polanski has used a self organizing map to derive a receptor-like neural network which could be used to screen virtual libraries. A flexible pharmacophore model of another receptor type has recently been described using a genetic algorithm approach<sup>29</sup>.

The problem of defining a virtual receptor, once a suitable set of measurements of biological activity of compounds at the receptor is compiled, involves: generation of a suitable molecular representation for the compounds whose activity have been determined; mapping of molecular representation to biological activity.

### **Benzodiazepine Virtual Receptor**

We applied some of the new molecular representations, and Bayesian regularised neural networks to the concept of virtual receptors. We used a combination of three simple representations (RKA): the atomistic representation<sup>5</sup>, which superficially appears to disregard much information such as topology and stereochemistry together with the Randic<sup>2,3</sup> and Kier & Hall<sup>4</sup> indices.

A data set was compiled from the literature<sup>30-38</sup> which consisted of 300 compounds of diverse structure: benzodiazepines, arylpyrazolo-quinolines,  $\beta$ -carboline, imidazo-pyridazines, cyclopyrrolones. These were broken up into two sets: 30 compounds would serve as the test set, whilst the other 300 compounds would form the training set. The test set was chosen using a k-means clustering method to ensure a good representation of the total set.

The standard error of predictions (SEPs) for the test set was 0134(scaled) with R=0734. Numerous ANN architectures were tested; the network with the lowest cross-validated SEP was deemed to be the optimal architecture. Rather surprisingly, the model does not suffer when positional information is removed from the representation (ie the position of substitution is ignored). Indeed, the best model using the atomistic approach was positionally-independent. The model obtained using the atomistic representation provides an SEP comparable to the model using the functional group representation.

The resulting virtual receptor was used to screen a virtual combinatorial library simulated by 110,000 compounds from the NCI chemical database. Several compounds were predicted to have a higher affinity for the benzodiazepine receptor than any in the training or test sets, and their activity is currently under investigation.

## REFERENCES

1. Hansch, C. and Fujita, T.,  $\rho$ - $\sigma$ - $\pi$  analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* (1964) 86 1616
2. Randic, M., On Characterisation of Molecular Branching. *J. Amer. Chem. Soc.* (1975) 97,6609
3. Randic, M. and Trinajstić, N., In search of graph invariants of chemical interest. *J. Molec. Struct.* (1993) 300,551-571
4. Kier, L.B. and Hall, L.H. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modelling, *Molecular Connectivity in Structure-Activity Analysis*, J. Wiley and Sons, New York, 1986.
5. Burden, F.R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct.-Act. Relat.* (1996) 15, 7-11
6. Burden, F.R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.-Act. Relat.* (1997) 16,309-314
7. Kier, L.B., Hall, L.H. in *Reviews in Computational Chemistry*, K.B. Lipkowitz and D.B. Boyd (Eds.) VCH Publishers, NY (1995) Volume 2, p374
8. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J., Gasteiger, J. Locating Biologically Active Compounds in Medium-sized Heterogeneous Datasets by Topological Autocorrelation Vectors. *J. Chem. Inf. Comput. Sci.* (1996) 36, 1205-1213
9. Winkler, D.A., Burden F.R., Watkins, A. Atomistic Topological Indices Applied to Benzodiazepines using Various Regression Methods. *Quant. Struct.-Activ. Relat.* (1998) in press.
10. Brown, R.D., Martin, Y.C. J., The Information Content of @d and #d Structural descriptors Relevant to Ligand-Receptor Binding. *Chem. Inf. Comput. Sci.* (1997) 37, 1-9
11. Winkler, D.A. and Burden, F.R. Holographic QSAR of Benzodiazepines. *Quant. Struct.-Activ. Relat.* (1998b) 17, 224-231
12. Burden, F.R. Molecular Identification Number for Substructure Searches *J. Chem. Inf. Comput. Sci.* (1989) 29, 225-27.
13. Pearlman, R. S.; Stewart, E. L.; Smith, K. M.; Balducci, R. Novel Software Tools for Combinatorial Chemistry and Chemical Diversity. Paper given at the 1997 Charleston Conference *Advancing New Lead Discovery*, Isle of Palms, SC (March 1997).
14. Platt, D.E.; Silverman, B.D. Registration, Orientation, and Similarity of Molecular Electrostatic Potentials through Multipole Matching. *J. Computat. Chem.* (1996), 17, 358-66.
15. Silverman, B.D.; Platt, D.R. Comparative molecular moment analysis (CoMMA): 3D - QSAR without molecular superposition. *J. Med. Chem.* (1996) 39, 2129-40

16. Abraham, D.J.; Leo, A.J. Extension of the Fragment Method to calculate Amino Acid Zwitterion and Sidechain Partition Coefficients. *Proteins: Struct. Funct. Genetics* (1987), 2, 130-152.
17. Kellogg, G.E., Semus, S.F., Abraham, D.J. - A New Method of Empirical Hydrophobic Field Calculation for CoMFA J. Comput.-Aided Molecular Design. *J. Comput.-Aided Mol. Des.* (1991) 5, 545-552.
18. Tripos Associates, 1699 South Hanley Road, Suite 303, St. Louis, MO 63144. HQSAR Software v 1.0. Tripos Associates: (<http://www.tripos.com/products/hqsar.html>)
19. MacKay, D.J.C. A Practical Bayesian Framework for Backprop Networks, *Neural Computation*, (1992) 4, 415-447
20. Rouvray, D.H. Making the Right Connection. *Chem. Brit.* (1993b) June, 495-498.
21. Ho, C.M.W. and Marshall, G.R. J. Comput.-Aided Mol. Des. (1995) 9 65-86
22. Rusinko, A., Sheridan, R.P., Nilakanta, R., Haraki, K.S., Bauman, N., Venkataraghavan, R. J. Chem. Inf. Comput. Sci. (1989) 29 251-255.
23. Maddalena, D. and Johnston, G.A.R., Prediction of Receptor Properties and Binding Affinities of Ligands to Benzodiazepine/GABAA Receptors Using Neural Networks. *J. Med. Chem.* 38, 715-724 (1995)
24. Zheng, W., Cho, S.J., Tropsha, A. Rational combinatorial library design. 1. Focus 2D: A new approach to the design of targeted combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, 1998, 38, 251-258. 25. Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H., Weinstein, J.N. *J. Chem. Inf. Comput. Sci.* (1998) 38, 189-99.
26. Lui, D., Jiang, H., Chen, K., Ji, R. A new approach to design virtual combinatorial library with genetic algorithm based on 3D grid property. *J. Chem. Inf. Comput. Sci.* (1998) 38, 233-42.
27. Horvath, D. A virtual screening approach applied to the search for trypanosome reductase inhibitors. *J. Med. Chem.* 1997, 40, 2412-2423.
28. Vedani, A., Dobler, M., Zbinden, P. Quasi-atomistic receptor surface models: A bridge between 3-D QSAR and receptor modelling, *J. Am. Chem. Soc.* 1998, 120, 4471-4477.
29. Polanski, J.J. The receptor-like neural network for modelling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.*, 1997, 37, 553-561
30. Zhang, W., Koehler, K.F., Harris, B., Skolnick, P., Cook, J.M., Synthesis of benzo-fused benzodiazepines employed a probes of the agonist pharmacophore of benzodiazepine receptors. *J. Med. Chem.* (1994) 37, 745-757.
31. Harrison, P.W., Barlin, G.B., Davies, L.P., Ireland, S.J., Matyus, P., Wong, M.G. Syntheses, pharmacological evaluation and molecular modelling of substituted 6-alkoxyimidazo[1,2-b]pyridazines as new ligands for the benzodiazepine receptor. *Eur. J. Med. Chem.*, (1996), 31, 651-662
32. Davies, L.P., Barlin G.B., Ireland, S.J., Ngu, M.M.L. Substituted Imidazo[1,2-b]ptiazines. New Compounds with Activity at entral and peripheral Benzodiazepine receptors., *Biochem. Pharmacol.* (1992) 44, 1555-1561.
33. Barlin, G.B., Davies, L.P., Davis, R.A., Harrison, P.W., Imidazo[1,2-b]pyridazines. XVII\* Synthesis and central nervous system activity of some 6-(alkylthio and chloro)-3-(methoxy, unsubstituted and benzamidomethyl)-2-aryl-imidazo[1,2-b]pyridazines containing methoxy, methylenedioxy and methyl substituents. *Aust. J. Chem.* (1994) 47, 2001-2012.
34. Fryer, R.I., Zhang, P., Rios, R., Gu, Z-Q, Basile, A.S., Skolnick, P., Structure-activity relationship studies Computer-aided molecular modelling, synthesis and biological evaluation of 8-(benzyloxy)-2-phenylpyrazolo[4,3-c]quinoline as a novel benzodiazepine receptor agonist ligand.
35. Wang, C-G, Langer, T., Kamath, P.G., Gu, Z-Q, Skolnick, P, Fryer, R.I., Computer-aided molecular modelling, synthesis and biological evaluation of 8-(benzyloxy)-2-phenylpyrazolo[4,3-c]quinoline as a novel benzodiazepine receptor agonist ligand. *J. Med. Chem.* (1995) 38, 950-957.
36. Hollinshead, S.P., Trudell, M.L., Skolnick, P., Cook, J.M. Structural requirements for agonist actions at the benzodiazepine receptor: studies with analogues of 6-(benzyloxy)-4-(methoxymethyl)-b-carboline-3-carboxylic *J. Med. Chem.* (1990) 33, 1062-1069.
37. Allen, M.S., Hagen, T.J., Trudell, M.L., Coddig, P.W., Skolnick, P., Cook, J.M., Synthesis of novel 3-substituted b-carbolines as benzodiazepine receptor ligands: Probing the benzodiazepine pharmacophore *J. Med. Chem.* (1988) 31, 1854-1861.
38. Yokoyama, N., Ritter, B., Neubert, A.D., . 2-Arylpyrazolo[4,3-c]quinolin-3-ones: Novel agonist, partial agonist and antagonist benzodiazepines, *J. Med. Chem.* (1982) 25, 337-339.



**Section VI**  
**Affinity and Efficacy**  
**Models of G-Protein**  
**Coupled Receptors**

## 5-HT<sub>1A</sub> RECEPTORS MAPPING BY CONFORMATIONAL ANALYSIS (2D NOESY/MM) AND "THREWAYMODELLING" (HASL, CoMFA, PARM)

Maria Santagati<sup>(a)</sup>, Arthur Doweyko<sup>(b)</sup>, Andrea Santagati<sup>(a)</sup>, Maria Modica<sup>(a)</sup>, Salvatore Guccione<sup>(a)</sup>, Hongming Chen<sup>(c)</sup>, Gloria Uccello Barretta<sup>(d)</sup>, Federica Balzano<sup>(d)</sup>

<sup>(a)</sup> *Dipartimento di Scienze Farmaceutiche, Università di Catania, viale Andrea Doria 6, Ed. 12, I-95125 Catania, Italy*

<sup>(b)</sup> *Macromolecular Modeling-CADD, Bristol-Myers Squibb, Pharmaceutical Research Institute, PO Box 4000, Princeton, NJ 08543*

<sup>(c)</sup> *Laboratory of Computer Chemistry, Institute of Chemical Metallurgy, Chinese Academy of Science, Beijing 100081, P. R. China*

<sup>(d)</sup> *Centro CNR di Studio per le Macromolecole Stereordinate ed Otticamente Attive, Università di Pisa, via Risorgimento 35, I-56126 Pisa, Italy*

---

The precise function of the 5-HT receptors remains undefined, and progress toward this has been hampered by the lack of selective ligands.

Direct interactions with the 5-HT<sub>1A</sub> receptor *via* selective ligands may have beneficial effects in a large number of diseases including a number of neuropsychiatric disorders (anxiety and depression).

The findings regarding the conformational analyses for a small set of [[[Arylpiperazinyl]alkyl]thio]thieno[2,3-d]pyrimidinone derivatives as high-affinity, selective 5-HT<sub>1A</sub> receptor ligands are reported. These include NMR analyses and molecular modeling approaches which were complemented by the use of three 3D-QSAR methodologies: PARM (PseudoAtomic Receptor Model), HASL (Hypothetical Active Site Lattice), and CoMFA (Comparative Molecular Field Analysis). The use of PARM (see chapter: APPLICATION OF PARM TO CONSTRUCTING AND COMPARING 5-HT<sub>1A</sub> AND  $\alpha_1$  RECEPTOR MODELS) represents an introduction of a novel paradigm, which is compared and contrasted using the same training/test set of 15/8 thienopyrimidinones.

All three methodologies were found to provide predictive models.

---

A wide array of biologically active substances, including neurotransmitters, hormones and neuropeptides, produce their biological effects by interacting with receptors that couple with G proteins. The G proteins, or guanine-nucleotide-binding regulatory proteins are generally localized to the inner surface of the plasma membrane. Trimeric ( $\alpha$ ,  $\beta$ ,  $\gamma$ ) G proteins relay signals from transmembrane receptors to intracellular enzymes and ion channels, thereby mediating vision, smell, taste and the actions of many hormones and neurotransmitters.

Although there are, as yet, no three dimensional crystal structure data available for GPCRs, the generally accepted view is that GPCRs contain seven  $\alpha$ -helical transmembrane domains linked by hydrophilic loops with an extracellular N-terminus and cytoplasmatic C-terminus<sup>1,2</sup>.

Serotonin modulates many processes in the mammalian peripheral and central nervous system through its interactions with at least 14 receptor subtypes, all but one (5-HT<sub>3</sub> subtype) of which are G protein (heterotrimeric GTP-binding protein)-coupled.

The 5-HT<sub>3</sub> subtype is a ligand-gated ion channel that shares functional and structural similarities with nicotinic acetylcholine receptors<sup>3,4</sup>.

The serotonin receptor subtype 5-HT<sub>1A</sub> has been cloned (genomic clone, G21, transiently expressed in monkey kidney cells) and is constituted by 421 amino acids arranged in seven helices<sup>5</sup>.

To our knowledge only recently have reasonable 5-HT<sub>1A</sub> receptor models been reported<sup>6</sup> and no HASL<sup>7,8</sup> and PARM<sup>9</sup> applications have been applied to 5-HT<sub>1A</sub> receptors.

The aim of this work, based on a combination of comparative conformational analysis by molecular mechanics and NMR spectroscopy (2D NOESY) is to define those features most critical to the design of selective, high affinity 5-HT<sub>1A</sub> ligands, taking into consideration the anchoring role of the scaffold heteroaromatic portion<sup>10</sup> connected to the "canonical" pharmacophoric arylpiperazine moiety. In addition, it was of interest to compare these features in both 5-HT<sub>1A</sub> and *alpha* receptors.

## Experimental

### NMR conformational analysis

Compounds **19**, **20**, **21** (69, 70, 71)<sup>10</sup> have been characterized by 2D NOESY and HETCOR analyses. Their stereochemistry has been determined by analyzing the intermolecular dipolar-dipolar interactions by means of 2D NOESY spectroscopy.

In the case of compound **20**, the methylene protons, belonging to the piperazine moiety, originate n.O.e.s on the aromatic protons of the *o*-methoxyphenyl substituent and on the chain methylene groups. Moreover, methoxy protons produce n.O.e. only on the aromatic proton adjacent to them and no dipolar interaction is originated by NH<sub>2</sub> or methyl protons.

Therefore, **20** assumes a conformation in which the chain linked to the sulfur atom is in a zig-zag planar arrangement, bringing the piperazine and aromatic rings far away from the thienopyrimidinone moiety. The two substituents linked to the two nitrogen of the piperazine ring mainly assume a pseudoequatorial arrangement which prevents any spatial proximity between the aromatic protons of the methoxyphenyl substituent at one nitrogen of the piperazine ring and the piperazine methylene directly linked to the other nitrogen. The aromatic moiety is mainly perpendicular to the plane of the piperazine ring and the methylene protons directly linked to the sulfur atom are far away from the amino group (Fig 1).

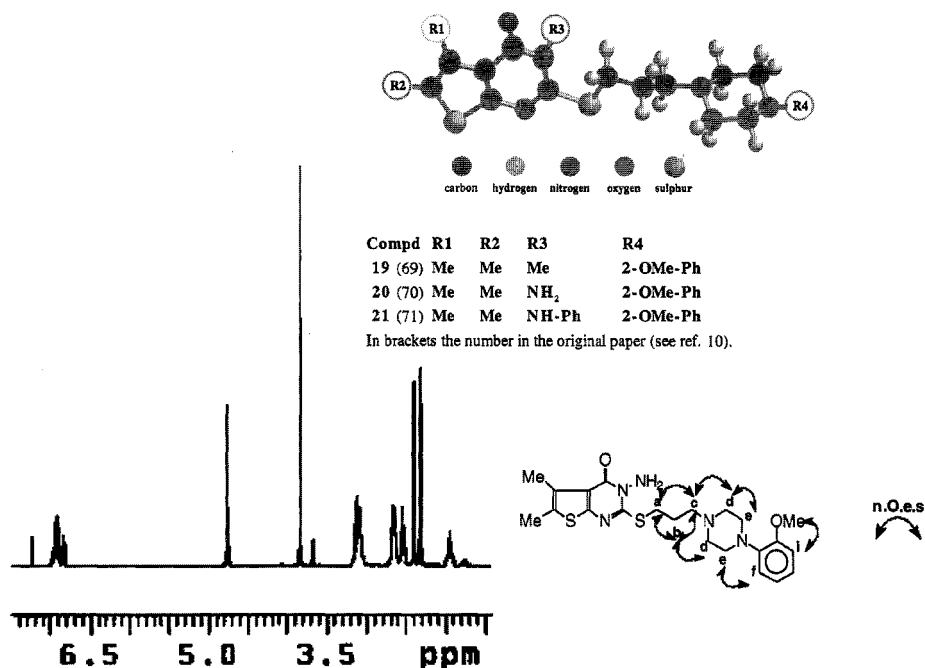


Fig 1 <sup>1</sup>H NMR spectrum of compound 20 in CDCl<sub>3</sub>

The <sup>13</sup>C longitudinal relaxation times (T<sub>1</sub>), determined for the two kinds of methylene carbons, are according to a rigid structure tumbling isotropically. Analogous NOESY analyses and <sup>13</sup>C relaxation time determinations have been also carried out on other related compounds (19, 21). Their stereochemical and dynamic features are similar to those already discussed for 20 (Fig 1-3).

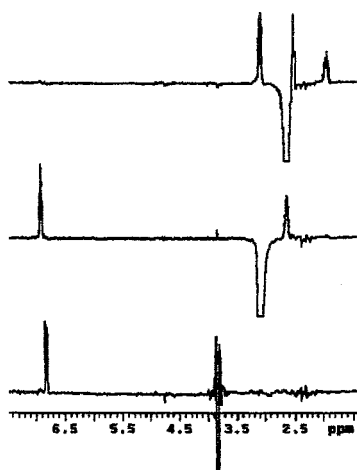


Fig 2 2D NOESY (300 MHz, CDCl<sub>3</sub>, 25°C, t=0.6 s) of compound 20

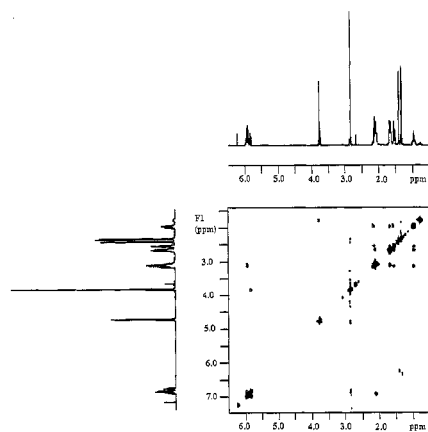
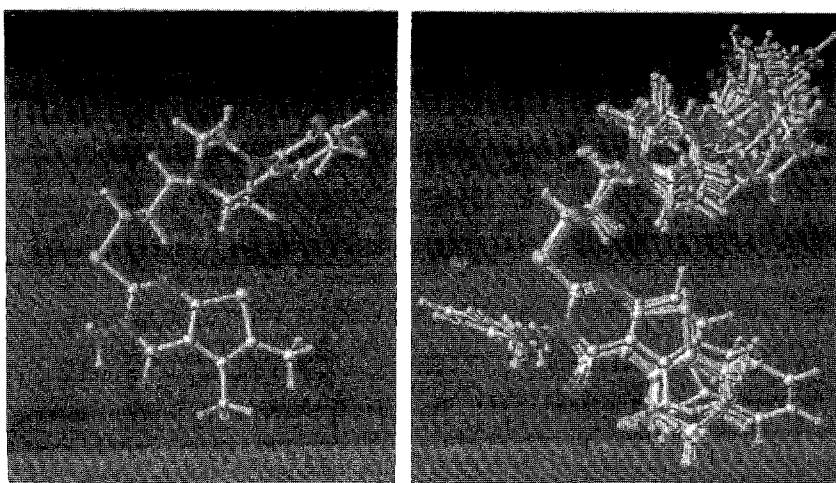


Fig 3 2D NOESY (300 MHz, CDCl<sub>3</sub>, 25 °C, t=0.6 s) of compound 20 a)methoxyl protons, b)He, c)Hd

### IC50 Measurements

The *in vitro* affinity for 5-HT<sub>1A</sub> and  $\alpha_1$ -AR was evaluated by radioligand binding assay on hippocampus and cortex of male CRL: CD(SD)BR-COBS rats weighing about 150 g, respectively as previously described<sup>10</sup>.

**Molecular Overlays.** The 23 molecules<sup>10</sup> used in the present investigation were built using SYBYL and SPARTAN 5.03 molecular modeling software<sup>11,12</sup>. The molecules were first geometry optimized using molecular mechanics (Tripos force field 5.0). Then each molecule underwent a systematic conformational search with each rotatable bond undergoing 10 degree steps until a global minimum energy conformation was found. The lowest energy conformation and atomic partial charges of each molecule were determined using PM3 within the SYBYL MOPAC module. All molecules were superimposed upon molecule 20 acting as a common template (Fig 4 a,b). Fifteen molecules were selected to act as the training set, while the remaining eight were used as the test set.



**Fig 4** (a)The template molecule (20) ; (b) the overlays of all 23 molecule

### 3D-QSAR

**CoMFA.** Conventional CoMFA<sup>13</sup> (Comparative Molecular Field Analysis) was carried out using the QSAR option within SYBYL<sup>11</sup> (Version 6.3) as configured on a SGI R10000 workstation (operating under IRIX64, release 6.4). Default settings were used except where otherwise noted. The steric and electrostatic energies were calculated using sp<sup>3</sup> carbon probes with a +1 charge. Grid spacing was set to 2.0 Å within the defined region and extending beyond van der Waals envelopes of all molecules by at least 4.0 Å. The CoMFA QSAR equations were determined using the PLS option with optimal number of components determined by a leave-one-out cross-validation procedure wherein the number of components yielding the lowest standard error of prediction were chosen.

**HASL.** Hypothetical Active Site Lattice (HASL, version 3.30)<sup>7,8</sup> computation was implemented on an SGIO2 R5000 (IRIX6.3) and SGI R10000 workstation (IRIX6.4).

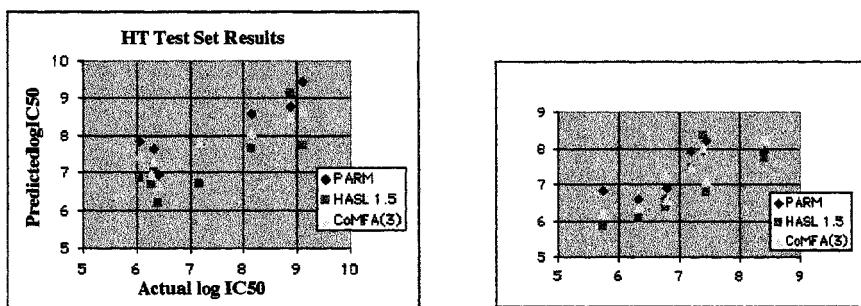
Default atom types were used ( $H=-1,0,+1$ ) and the grid spacing was 1.5 Å. Models were iteratively solved to an average error in prediction of 0.001 log units activity. Several new programs were written to provide a graphical display of resulting lattice-based binding models within SYBYL.

**PARM.** PARM is a program developed in-house which can build an atomic pseudo-receptor model by using a genetic algorithm<sup>9</sup> (see chapter:APPLICATION OF PARM TO CONSTRUCTING AND COMPARING 5-HT<sub>1A</sub> AND  $\alpha_1$  RECEPTOR MODELS).

### Results and Discussions

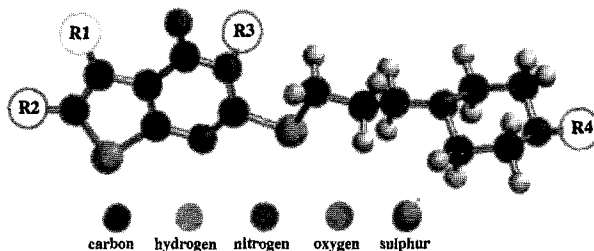
Preliminary investigations using CoMFA, HASL and PARM models derived from a 15-member training set to predict a 8-membered test set indicated that these models were essentially equivalent in their predictive strengths (*Tables I and II*).

In order to fully investigate the structural properties most closely associated with activities, it was of interest to develop models utilising all 23 analogues. CoMFA models were derived using the full 23 compound data set: 5-HT<sub>1A</sub> (5-components,  $r^2 = 0.991$ ,  $r^2_{cv} = 0.261$ ) and  $\alpha_1$ -AR (4-components,  $r^2 = 0.970$ ,  $r^2_{cv} = 0.652$ ). In addition, a difference model was also developed wherein the difference,  $pIC_{50}(5-HT_{1A}) - pIC_{50}(\alpha_1-AR)$ , was correlated to structure: 5 components,  $r^2 = 0.978$ . The cross-validated  $r^2$  in this analysis was very poor. Although the cross-validated correlation coefficient was low for the HT and difference models, this may not be a significant issue since the number of analogues is small, and it may well be that each molecule is a significant contributor to the model. Using the full 23 compound data set, three models were developed using the HASL paradigm for 5-HT<sub>1A</sub>,  $\alpha_1$ -AR and the difference data set at 1.5 Å, in each case yielding models containing 427 lattice points iteratively solved to average errors in prediction < 0.001 pIC<sub>50</sub> units. In the PARM model building paradigm, the results of the test set were used to guide model selection. The fifteenth and fourth models were found to have the best predictions for the 5-HT<sub>1A</sub> and  $\alpha_1$ -AR data sets, respectively. These two models were chosen to be analyzed (**Fig 5**). The computational results of these models for the two data sets are listed in chapter: APPLICATION OF PARM TO CONSTRUCTING AND COMPARING 5-HT<sub>1A</sub> AND  $\alpha_1$  RECEPTOR MODELS.



**Fig 5** Graphs of the three models derived from 15 molecules predicting 8

The CoMFA computational results of these models for the two data sets are listed in *Tables I and II*.

Table I Test set calculated statistics for HT<sub>1A</sub> receptors

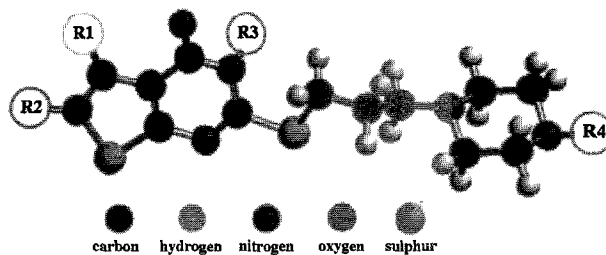
Compd	R1	R2	R3	R4	Actual	PARM	Error	HASL 1.5	Error	CoMFA (5) <sup>b</sup>	Error	CoMFA (3) <sup>b</sup>	Error
1 (43)*	Me	Me	2-CIPh	H	6.337	7.62	1.283	7.11	0.773	7.43	1.093	7.27	0.933
6 (50)*		-(CH <sub>2</sub> ) <sub>4</sub> -	2-CIPh	H	6.074	7.823	1.749	6.87	0.796	7.36	1.286	7.43	1.356
9 (56)*		-(CH <sub>2</sub> ) <sub>4</sub> -	1-naphthyl	H	6.431	6.939	0.508	6.20	0.231	6.56	0.129	6.66	0.229
10 (57)*		-(CH <sub>2</sub> ) <sub>4</sub> -	2-pyrimidinyl	H	6.297	6.888	0.591	6.71	0.413	7.17	0.873	6.91	0.613
15 (66)*		-(CH <sub>2</sub> ) <sub>4</sub> -	Me	2-OMe-Ph	8.155	8.596	0.441	7.64	0.515	8.01	0.145	8.08	0.075
16 (67)*		-(CH <sub>2</sub> ) <sub>4</sub> -	NH <sub>2</sub>	2-OMe-Ph	8.886	8.76	0.126	9.13	0.244	8.42	0.466	8.50	0.386
22 (72)*	Me	Me	Me	2-pyrimidinyl	7.187	7.743	0.556	6.70	0.487	8.02	0.833	7.81	0.623
24 (74) <sup>a</sup> *	Me	Me	NH <sub>2</sub>	2-OMe-Ph	9.097	9.468	0.371	7.73	1.367	9.07	0.027	8.67	0.427
<b>SD*</b>							<b>0.860</b>		<b>0.700</b>		<b>0.750</b>		<b>0.690</b>

\*In brackets the number in the paper (see ref. 10).

<sup>a</sup>The piperazine ring has been replaced by a piperidine nucleus.

<sup>b</sup>Number of components in PLS analysis.

Table II Test set calculated statistics for  $\alpha$ 1-AR



Compd	R1	R2	R3	R4	Actual	PARM	Error	HASL 1.5	Error	CoMFA (5)	Error	CoMFA (3)	Error
1 (43)*	Me	Me	2-ClPh	H	6.793	6.886	0.093	6.520	0.273	6.450	0.343	6.53	0.263
6 (50)*		-(CH <sub>2</sub> ) <sub>4</sub> -	2-ClPh	H	6.775	6.581	0.194	6.350	0.425	7.060	0.285	7.22	0.445
9 (56)*		-(CH <sub>2</sub> ) <sub>4</sub> -	1-naphtyl	H	6.352	6.594	0.242	6.060	0.292	6.150	0.202	6.37	0.018
10 (57)*		-(CH <sub>2</sub> ) <sub>4</sub> -	2-pyrimidinyl	H	5.741	6.830	1.089	5.820	0.079	6.120	0.379	6.16	0.419
15 (66)*		-(CH <sub>2</sub> ) <sub>4</sub> -	Me	2-OMe-Ph	7.194	7.919	0.725	7.470	0.276	7.530	0.336	7.47	0.276
16 (67)*		-(CH <sub>2</sub> ) <sub>4</sub> -	NH <sub>2</sub>	2-OMe-Ph	7.409	7.924	0.515	8.350	0.941	7.980	0.571	8.00	0.591
23 (72)*	Me	Me	Me	2-pyrimidinyl	7.444	8.217	0.773	6.770	0.674	7.050	0.394	7.11	0.334
24 (74) <sup>a*</sup>	Me	Me	NH <sub>2</sub>	2-OMe-Ph	8.398	7.893	0.505	7.710	0.688	8.280	0.118	8.26	0.138
<b>SD*</b>							<b>0.610</b>		<b>0.530</b>		<b>0.350</b>		<b>0.350</b>

\*In brackets the number in the paper (see ref. 10).

<sup>a</sup>The piperazine ring has been replaced by a piperidine nucleus.

<sup>b</sup>Number of components in PLS analysis.



In the 5-HT<sub>1A</sub> receptor model, for the training set, the conventional correlation coefficient is 0.962, the cross-validated correlation coefficient is 0.906 and its prediction deviation to the test set is 0.860. On the other hand, for the  $\alpha_1$ -AR receptor model, the conventional correlation coefficient is 0.977 and the cross-validated one is 0.945. Its test set prediction deviation is less than that of the 5-HT<sub>1A</sub> receptor model and it is only 0.61. So the computation results show that for the training set, both of the receptor models have a good correlation between the interaction energy and the bioactivity. They also have some predicting ability (based on test set performance), although the molecules in the prediction set were not involved in the process of building the model.

The receptor models derived from PARM and HASL methods were converted to files compatible with SYBYL software in order to visualize their characteristics and make direct comparisons to CoMFA fields. The PARM models for 5-HT<sub>1A</sub> and  $\alpha_1$ -AR are shown in Figs 6 and 7, respectively. In those figures, the unconnected atoms distributed in the space represent a receptor model which simulates the receptor pocket. The ligands are docked in the middle of the receptor model in order to compute the interaction energy between ligands and receptor model. The colours of the pseudo receptor atoms are the same as those of atoms normally defined in SYBYL. The largest atom types having the smoke-grey colour in the receptor model represent the void space there.

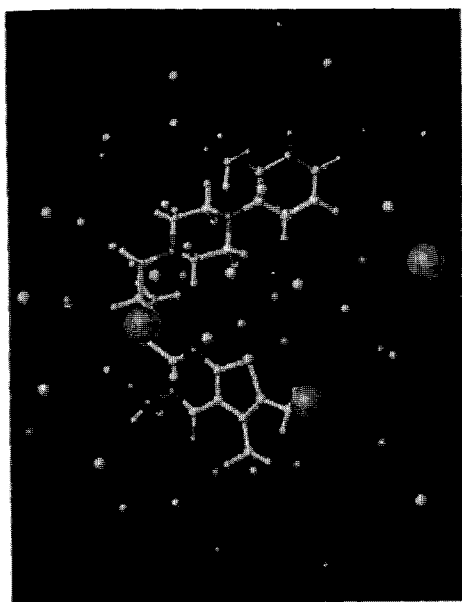


Fig 6 The PARM 5-HT<sub>1A</sub> receptor model

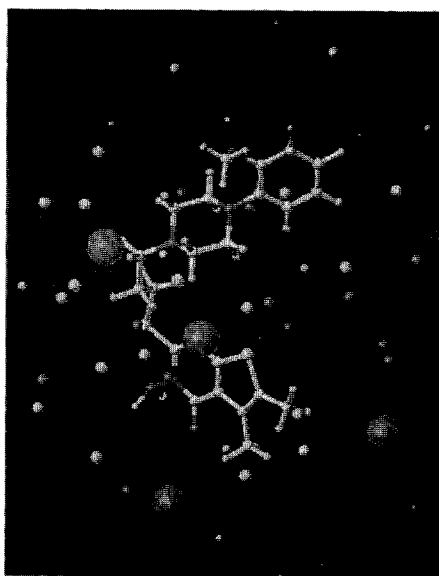


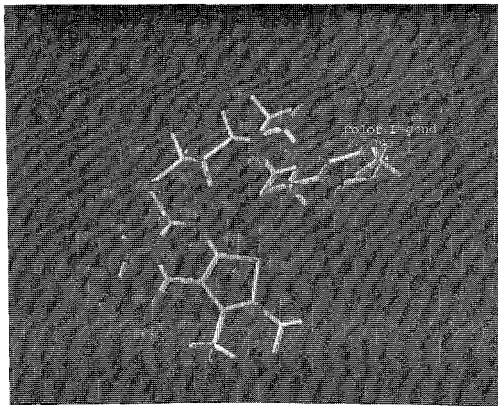
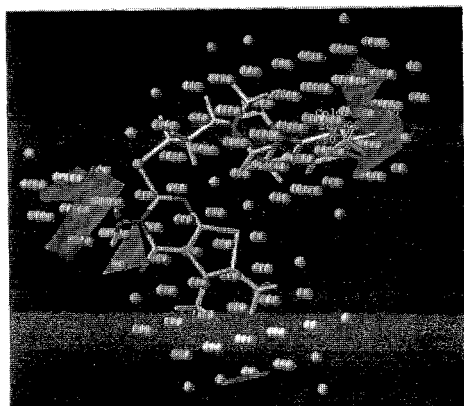
Fig 7 The PARM  $\alpha_1$ -AR receptor model

The colours of the pseudo receptor atoms are the same as those of atoms normally defined in SYBYL. The largest atom types having the smoke-grey colour in the receptor model represent the void space there. The molecule which is docked in the middle of the receptor pocket is the template molecule (20). At the region near the R3 position, there exist several negatively charged O and N atoms. These pseudo-receptor atoms may act as hydrogen bond acceptors in their interaction with the R3 position of the ligand.

**Fig 6** illustrates the 5-HT<sub>1A</sub> PARM receptor model. We can see that at the region near the R3 position, there exist several negatively charged O and N atoms. These pseudo-receptor atoms may act as hydrogen bond acceptors in their interaction with the R3 position of the ligand. The molecule which is docked in the middle of the receptor pocket is the template molecule (**20**). The model further incorporates a large void space between the receptor atoms and the R3 substituent of molecule **20**, suggesting that binding affinity can be improved by placing a bulky group which has positive charge in the R3 position. This is the same conclusion which we can draw from an analysis of the CoMFA and HASL 5-HT<sub>1A</sub> models illustrated in **Fig 8**.

It is generally accepted that 5-HT<sub>1A</sub> receptor agonists and antagonists bind their protonated amino sites to the highly conserved aspartate (Asp 129) on transmembrane helix 3 (TM3)<sup>5</sup>.

In the R4 position, all the molecules have a substituted aromatic ring. In the PARM model there is one positively charged hydrogen atom (proton) near the methoxy group of the phenyl ring to act as a hydrogen bond acceptor. This observation is consistent with the conclusion from the CoMFA contour map that a negatively charged group on position 2 of the aromatic ring will enhance affinity. Near the 3,4,5 positions of the aromatic ring, there exist some negatively charged N, O pseudo-receptor atoms which suggests, that in these positions, the ligands should have some positively charged groups to improve affinity. This observation is consistent with the CoMFA 5-HT<sub>1A</sub> analysis shown in **Fig 9**, but is not present in the HASL model. There are some differences between the PARM model and CoMFA, for example, some neutral pseudo-receptor atoms, the void space surrounding the R1 and R2 region and several negative atoms near R1. The relevancy of these differences is difficult to ascertain, as all the molecules have some neutral substitutions in these positions which may not have a direct effect on affinity.

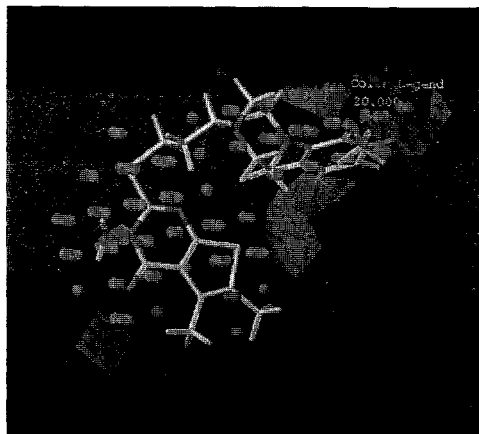
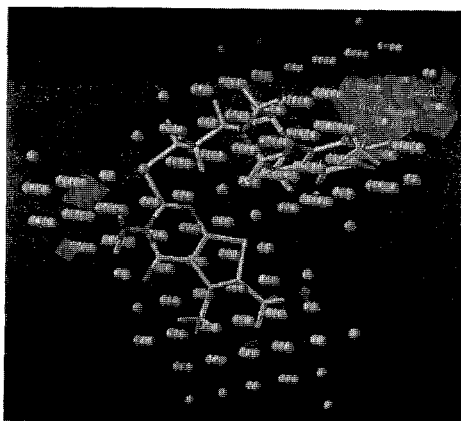


**Fig 8** CoMFA and HASL 5-HT<sub>1A</sub> Models: CoMFA Steric Fields; HASL Neutral Atom Types. Substituted phenyl region is identified with strong positive steric effects by both methods, while the strongest HASL negative effects are found at extensions of the R1 and R2 regions. The steric extension into the R3 region is identified by CoMFA as positive, while HASL relegates it to a mild positive effect.

**Fig 9** CoMFA and HASL 5-HT<sub>1A</sub> Models: CoMFA Electrostatics ; HASL Electron-Rich Atom Types and Electron-Poor Atom Types. The HASL model indicates that both electron-poor and electron-rich atoms are found to positively enhance affinity at the R3 region, while the CoMFA electron-poor field is also found significant at that position.

CoMFA Electrostatics and HASL electron-rich atom types are both found significant in the phenyl substituted region.

The features of the  $\alpha_1$ -AR PARM model are shown in Fig 7. Some differences between the 5-HT<sub>1A</sub> and  $\alpha_1$ -AR receptor models are apparent. As was observed in the 5-HT<sub>1A</sub> model, a negatively charged atom near the R3 position is present, however, there are also two neutral C atoms here. Thus it would appear that the effect of an electrostatic interaction at this region in the PARM  $\alpha_1$ -AR model is not as important as indicated in the PARM 5-HT<sub>1A</sub> receptor model. In the CoMFA analysis of the  $\alpha_1$ -AR data set, the electrostatic region is not present at the R3 position. In addition, there is a large void space between the ligand and receptor model which could include a sterically bulky group. These conclusions are further supported by the HASL  $\alpha_1$ -AR model which illustrates a positive effect by bulky R3-substituents (Fig 10) and supports some mixed introduction of electron-rich/poor atom types at R3 (Fig 11). At the R4 position of the PARM  $\alpha_1$ -AR model there is a positively charged hydrogen atom near position 2 of the phenyl ring, which would increase affinity with a negatively charged group there. Proximate to the methoxy group and the region near positions 3,4,5 of the aromatic ring, there are several negatively charged pseudo-receptor atoms. Thus, the ligand should have some positively charged group near that position in order to enhance affinity. In addition, two negatively charged atoms near the R1 and R2 position exist which would complement a positive group there to increase affinity. This conclusion is further supported by the CoMFA and HASL models.



**Fig 10** CoMFA and HASL  $\alpha_1$ -AR Models: CoMFA Steric Fields and HASL Neutral Atom Types. HASL yielded a similar dependence of affinity to steric bulk as identified in the 5-HT<sub>1A</sub> model, while positive contributions from CoMFA steric fields are now limited to the substituted phenyl region.

**Fig 11** CoMFA and HASL  $\alpha_1$ -AR Models: CoMFA electrostatics and HASL Electron-Rich and Electron-Poor Atom Types (colors are defined as in Figure 9). HASL identified the R3 region in a similar way as in the 5-HT<sub>1A</sub> model, however with less emphasis on the electron-rich effect. Parallel effects are once again observed for the phenyl-substituted region as previously observed in the 5-HT<sub>1A</sub> model.

In conclusion, this investigation of modeled 5-HT<sub>1A</sub> and  $\alpha_1$ -AR receptor features based on analyses of 23 molecules illustrated the effective use of three 3D-QSAR methodologies namely, PARM, CoMFA and HASL. These methods were found to provide reasonable predictive strength, and in addition, pointed to ligand features which were found to be significant to binding affinity. The comparison of a field-based method (CoMFA), an occupancy method (HASL) and the pseudo-receptor method (PARM) provided a number of consistent observations across the three paradigms. The use of more than one 3D-QSAR analysis was found to be an exceptionally useful approach to the identification of ligand features most likely to be significant in enhancing affinity, i.e. to uncover additional sites of interactions not apparent in single models (*forthcoming paper*).

No information on the pharmacological activity of the subject 5-HT<sub>1A</sub> selective ligands [[(Arylpiperazinyl)alkyl]thio]thieno[2,3-d]pyrimidinone derivatives by M. Santagati *et al*<sup>10</sup> is still available to categorize the ligands as agonists or antagonists. Different alignments for agonists/antagonists, would lead to better discriminating models, provided that the statistical quality of our models is already satisfactory.

Further pharmacological studies are ongoing to solve the problem.

**Acknowledgements.** Financial support (40%) from Italian MURST and technical support from Dr. M. Mabilia (S.IN Soluzioni Informatiche s.as - via Salvemini, 9, I-36100 Vicenza - Italy) are gratefully acknowledged.

S. Guccione thanks Prof. Thierry Langer for training him to the molecular modeling and Prof. Eric Walters for the helpful discussion and directions.

## References

1. H. E. Hamm, The many faces of G protein signaling, *J. Biol. Chem.*, **273**: 669 (1988).
2. S. R. Sprang, G Protein mechanism: insights from structural analysis, *Annu. Rev. Biochem.*, **66**: 639 (1997).
3. Data from Trends in Pharmacological Sciences: RECEPTOR & ION CHANNEL NOMENCLATURE SUPPLEMENT, (1998).
4. L. J. England, J. Imperial, R. Jacobsen, A. G. Craig, J. Gulyas, M. Akhtar, J. Rivier, D. Julius, B. M. Olivera, Inactivation of a serotonin-gated ion channel by a polypeptide toxin from marine snails, *SCIENCE*, **281**: 575 (1998) and enclosed references.
5. J. Hoflack, S. Trumpp-Kallmeyer, M. F. Hilbert, Molecular modelling of G-protein coupled receptors, in *3D QSAR in Drug Design. Theory methods and applications*, H. Kubinyi, ed., ESCOM Science Publishers, Leiden (1993).
6. P. Gaillard, P-A Carrupt, B. Testa and P. Schambel, Binding of Arylpiperazines, (Aryloxy)propanolamines, and Tetrahydropyridylindoles to the 5-HT<sub>1A</sub> Receptor: contribution of the molecular lipophilicity potential to three-dimensional quantitative structure-activity relationship models, *J. Med. Chem.*, **39**: 126 (1996) and enclosed references.
7. HASL (Hypothetical Active Site Lattice), Hypothesis Software, PO Box 237, Long Valley, NJ 07853. (email: [hyposoft@cris.com](mailto:hyposoft@cris.com); [www.cris.com/~Hyposoft](http://www.cris.com/~Hyposoft)).
8. A. M. Doweyko, The Hypothetical Active Site Lattice. An approach to modelling active sites from data on inhibitor molecules, *J. Med. Chem.*, **31**: 1396 (1988).
9. H. M. Chen, J. J. Zhou, G. R. Xie, PARM: A genetic evolved algorithm to predict bioactivity, *J. Chem. Inf. Comput. Sci.*, **38**: 243 (1998).
10. M. Modica, M. Santagati, F. Russo, L. Parotti, L. De Gioia, C. Selvaggini, M. Salmona and T. Mennini, [[(Arylpiperazinyl)alkyl]thio]thieno [2,3-d]pyrimidinone derivatives as high-affinity, selective 5-HT<sub>1A</sub> receptor ligands, *J. Med. Chem.*, **40**: 574 (1997).

11. SYBYL Molecular Modelling Software, Tripos Inc., 1699 S. Hanley Raod, Suite 303, St. Louis, MO 63144.
12. SPARTAN 5.03 Molecular Modelling Software, Wavefunction Inc., 18401 Von Karman Avenue, Suite 370, Irvine, CA 92612.
13. R. D. Cramer, III, D. E. Patterson, J. D. Bunce, Comparative molecular field analysis (CoMFA).1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, **110**: 5959 (1988).

## DESIGN AND ACTIVITY ESTIMATION OF A NEW CLASS OF ANALGESICS

Slavomir Filipek and Danuta Pawlak

Department of Chemistry  
University of Warsaw  
1 Pasteur St, 02-093 Warsaw  
Poland

### INTRODUCTION

Our ability to use rational design for the generation of useful peptides is depends on our ability to determine the specific relationships of molecular structure to biological activity. Opiates and opioid peptides display a large spectrum of biological activities, including analgesia, respiratory depression, euphoria, hypothermia, tolerance, physical dependence etc. There are at least three different receptor classes ( $\mu$ ,  $\delta$ ,  $\kappa$ ) differing from one another in their structural requirements towards opioid ligands. In order to associate a particular receptor class with a distinct biological function, it is of great importance to develop opioid receptor ligands with high activity and selectivity for a particular receptor type. Unfortunately, none of the endogenous opioid peptide is very selective for a particular receptor class. The lack of selectivity observed with most naturally occurring opioid peptides and with many of their linear analogs is most likely due to their structural flexibility which permits conformational adaptation to more than one opioid receptor type. Flexible molecules of peptides assume many conformations in solution. One of the available conformation or closely related family of conformations is responsible for the biological activity of the peptide. To determine which conformations are important it is necessary to confine conformational space accessible to flexible peptides. The most drastic restriction of the overall conformation freedom can be achieved through peptide cyclization<sup>1,2</sup>. Cyclization through covalent linkage of two side-chains has been performed by disulfide bond formation or by amide bond formation. In particular, cyclization of enkephalin *via* side chains of appropriately substituted amino acid residues have been successful<sup>3-9</sup>.

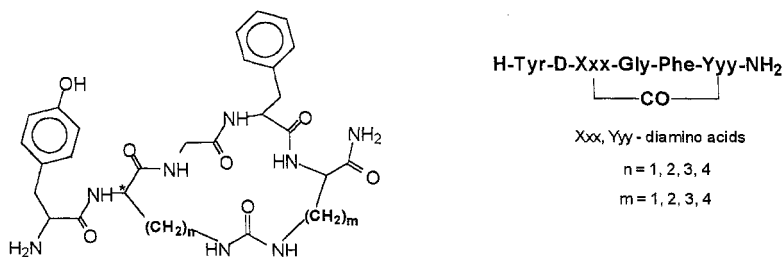


Figure 1. Structural formulas of trial set of molecules.

In the present paper we describe an introductory prognosis of activity and selectivity of a series of cyclic enkephalin analogues in which ring formation was achieved *via* ureido group incorporating the side-chain amino groups of diamino acids (Fig. 1).

## METHODS

The molecular structures were obtained using the Insight II package release 95.0<sup>10</sup> installed on Silicon Graphics computers. This package was also used for generating input data for Molecular Dynamics method in the Discover program<sup>10</sup>. Simulations were conducted in vacuum in 300 K temperature for a period of 1 ns with 1 fs step. The standard Insight II force field cff91 was chosen. Partial charges were calculated by the Gasteiger and Marsili method. Conformations were picked every 1 ps, that seems to be sufficiently long time period to assure the lack of correlation between subsequent conformations. All compounds appeared in the unprotonated form. None of conformations were minimized prior to cluster analysis.

Cluster analysis were done with APEX\_CLUST algorithm<sup>10</sup> from Insight II package. Clustering was conducted for all reference and trial compounds. Due to a great conformational flexibility of the molecular structures the inter-cluster distance parameter was set to a rather large value of  $d_0 = 3 \text{ \AA}$ . Usually eight central conformers from best clusters were sufficient for all reference and test compounds. Conformations from weakly populated clusters consist mainly of higher energy structures and, therefore, should be rejected. APEX-3D module<sup>10</sup> from InsightII package was used to construct models of  $\mu$ - and  $\delta$ -receptor ligand activities and determine activities of novel compounds. Despite of small number of conformers, a number of generated pharmacophores overwhelmed possibilities of the program to analyze them.

Two methods were used to establish a relationship between structure of reference compounds and their biological activities (QSAR) and then estimate activity of test molecules: Molecular Field Analysis (MFA) and Receptor Surface Analysis (RSA) modules both in Cerius<sup>2</sup> package<sup>11</sup>. Both methods require properly aligned set of molecules. As a reference set of molecules 24 cyclic peptide analogs were chosen (Table 1). All of them were minimized in the Universal Force Field<sup>12</sup> in conjunction with the Charge Equilibration in Cerius<sup>2</sup> package. All the molecules were fitted to the model of  $\mu$ -selective receptor pharmacophore<sup>13, 14</sup> based on  $\mu$ -selective opiates such as PET (PEO).

**Table 1.** A reference set of molecules:  $\mu$  and  $\delta$  activities taken from<sup>3, 6, 7, 15, 16</sup>

H-Tyr-c[N <sup>6</sup> -D-A <sub>2</sub> pr-Gly-Phe-Leu-]	H-Tyr-D-Pen-Gly-Phe-L-Cys-NH <sub>2</sub>	H-Tyr-c[N <sup>6</sup> -D-Lys-Phe-Ala-]
H-Tyr-c[N <sup>7</sup> -D-A <sub>2</sub> bu-Gly-Phe-Leu-]	H-Tyr-D-Pen-Gly-Phe-D-Cys-NH <sub>2</sub>	H-Tyr-c[N <sup>6</sup> -D-Orn-Phe-Ala-]
H-Tyr-c[N <sup>6</sup> -D-Orn-Gly-Phe-Leu-]	H-Tyr-D-Pen-Gly-Phe-D-Cys-OH	H-Tyr-c[N <sup>7</sup> -D-A <sub>2</sub> bu-Phe-Ala-Leu-]
H-Tyr-c[N <sup>6</sup> -D-Lys-Gly-Phe-Leu-]	H-Tyr-D-Pen-Gly-Phe-L-Cys-OH	H-Tyr-c[N <sup>6</sup> -D-Orn-Phe-Gly-]
H-Tyr-D-Lys-Gly-Phe-Glu-NH <sub>2</sub>	H-Tyr-D-Pen-Gly-Phe-D-Pen-OH	H-Tyr-D-Cys-Phe-Asp-Cys-Val-Gly-NH <sub>2</sub>
H-Tyr-D-Orn-Phe-Asp-NH <sub>2</sub>	H-Tyr-D-Pen-Gly-Phe-L-Pen-OH	H-Tyr-D-Cys-Phe-Asp-Pen-Val-Gly-NH <sub>2</sub>
H-Tyr-D-Asp-Phe-Orn-NH <sub>2</sub>	H-Tyr-D-Cys-Gly-Phe-D-Pen-OH	H-Tyr-D-Pen-Phe-Asp-Pen-Val-Gly-NH <sub>2</sub>
H-Tyr-D-Lys-Phe-Glu-NH <sub>2</sub>	H-Tyr-D-Cys-Gly-Phe-L-Pen-OH	H-Tyr-D-Pen-Phe-Asp-Pen-Nle-Gly-NH <sub>2</sub>
H-Tyr-D-Glu-Phe-Lys-NH <sub>2</sub>		H-Tyr-D-Pen-Phe-Glu-Pen-Val-Gly-NH <sub>2</sub>

For an alignment of reference as well as trial molecules a flexible fitting was used. We initially used default method - subgraph search. This method initially uses rigid fitting to determine the best set of atom matches and then flexible fitting was executed using this set. At the end of the process manual fitting was employed. The conformer generated for each moving model was minimized after alignment has been complete.

Molecular Field Analysis (MFA) quantifies the interaction energy between a probe molecule and a set of aligned target molecules in a QSAR. To generate an energy field (probe map), a probe molecules is placed at a random location then moved about a target molecule within a defined 3D grid. For each molecule two fields were generated: one with a proton probe (H<sup>+</sup>) and one with an

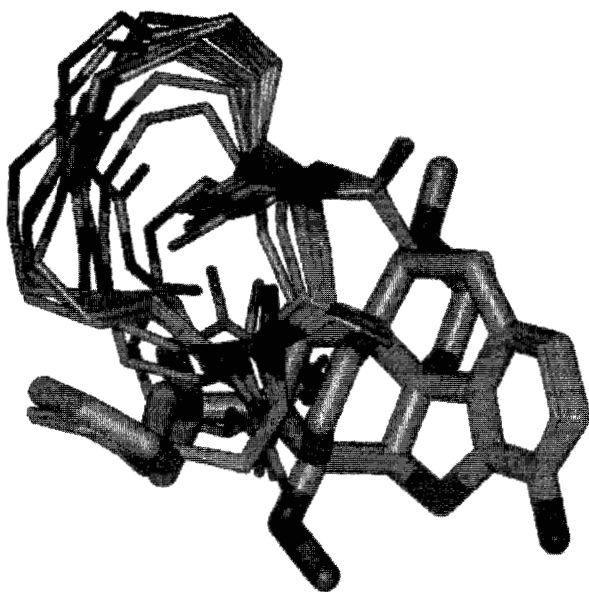


Figure 2. An alignment of trial set of molecules.

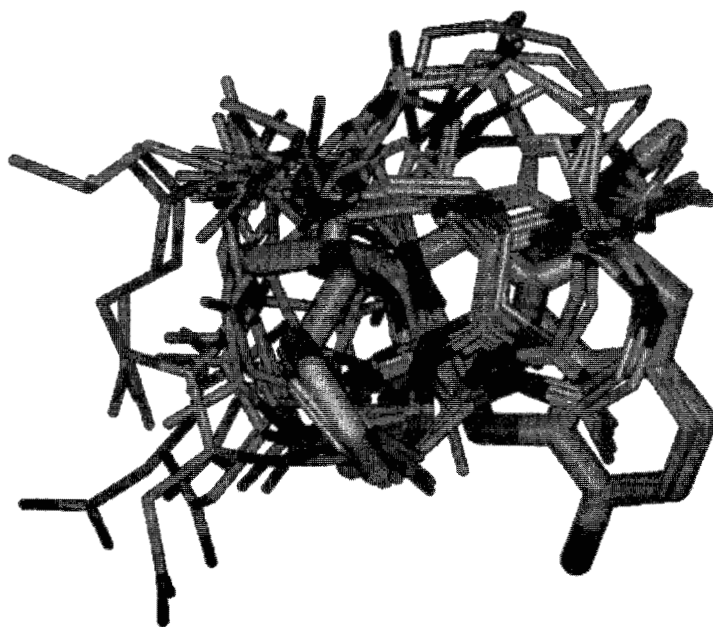


Figure 3. An alignment of reference set of molecules.



uncharged methyl probe (CH<sub>3</sub>). Each calculation uses a cubic grid with 2Å-spacing. Only Molecular Field points with highest variance were used as molecular descriptors and Genetic Function Approximation (GFA) was used as a regression method. A major advantage of this approach is that a collection of diverse, small models is generated that all have roughly the same high predictability.

The second module - Receptor Surface Analysis - creates hypothetical models that characterize the active site of the receptor based on the construction of surfaces to represent spatial and electrostatic properties of a receptor active site. Molecules are minimized within the receptor surface model and interaction energies are calculated, which allows the evaluation of new candidate compounds.

μ- and δ-receptor models were created from seven most active ligands of μ- and δ-receptor respectively. The interaction energy calculated by RSA module between a molecule and a receptor surface model were used to develop QSAR regression. RSA calculates molecule - receptor model interaction energies at a receptor surface, and these energies serve as input for the calculation of QSAR relationship. The default selection in RSA module (the same as in MFA) was used: 90% points with lowest variance were excluded from calculations. A standard GFA method was used to perform statistical calculations.

## RESULTS AND DISCUSSION

The aim of our investigations was to construct activity model of cyclic peptide analogs, and, on the basis of the model, try to predict of biological activity of array of analogs. As a reference set of molecules 24 cyclic enkephalins and deltorphins were chosen. Cyclic molecules have much less degrees of freedom than their linear analogs but still possess great deal of flexibility especially when linear long side groups are connected to macrocycle. Besides of flexibility there are many of possible pharmacophore atoms or groups in these molecules, so programs looking for potential pharmacophore sites generate too many of artificial pharmacophores. We tried with a program APEX-3D in InsightII<sup>10</sup> package but over 50,000 of generated pharmacophores overwhelmed possibilities of the program to analyze them.

We therefore decided to switch to the pharmacophore model developed by Brandt et al.<sup>13,14</sup> for μ-selective ligands based on μ-selective opiates such as PET (PEO). The model was simplified to adapt to our reference set of molecules. Some of pharmacophore sites were excluded and four sites of well known great importance for activity were remained – two benzene rings of phenylalanine and tyrosine, nitrogen atom of an amino group of tyrosine and hydroxyl group in tyrosine. Both reference and trial set of molecules were generated and minimized with the same procedures as indicated in the Method section. All of them were fitted to the μ-selective ligands receptor pharmacophore model based on mentioned above four pharmacophore sites with proper distance to each other. After fitting procedure all molecules were minimized once again to achieve local minimum conformation closest to the fit found before.

The conformations are shown in Fig. 2 and Fig. 3 for trial and reference set of molecules respectively. They are superimposed on PEO (fat stick model) to show their compact structures similar to opiate ligands. Besides, negative electrostatic potential of cyclic peptides is compatible to one generated by opiates (not shown).

MFA method provided a QSAR regression with correlation coefficient  $R = 0.94$  and least square error of prediction  $LSE = 0.27$ . Majority of the trial molecules were predicted to be μ-selective ligands as were known from preliminary experiments<sup>17</sup>.

Although the model was developed for μ-selective ligands we tried to examine δ-selectivity based on the same model to check if it is working. The only difference this time we used δ activity values. The regression we obtained were characterized with  $R = 0.94$ ,  $LSE = 0.19$ . Results indicated that trial molecules are not δ-selective in general and the model based on μ-selectivity and simplified is working quite good. Some modifications are required to rearrange pharmacophore sites to more properly describe δ-selective receptor model in the future.

The second method used was Receptor Surface Analysis (RSA). Both reference and trial sets of molecules were fitted to the μ-selective ligands receptor pharmacophore model as for MFA method. But this time no minimization was conducted after fitting procedure. A receptor model surface was generated for seven most μ active molecules from reference set. At each surface point the potentials were mapped based on complementarity between a respective molecule potential and

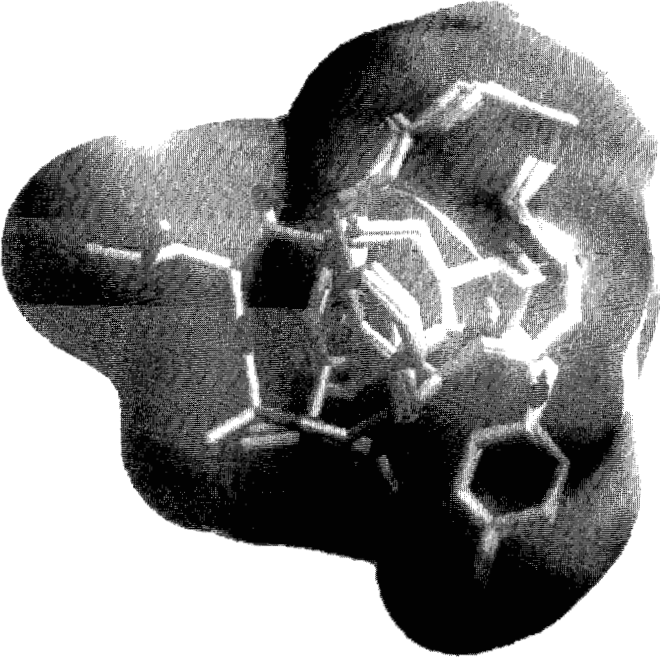


Figure 4. Receptor model for  $\mu$ -receptor ligands.

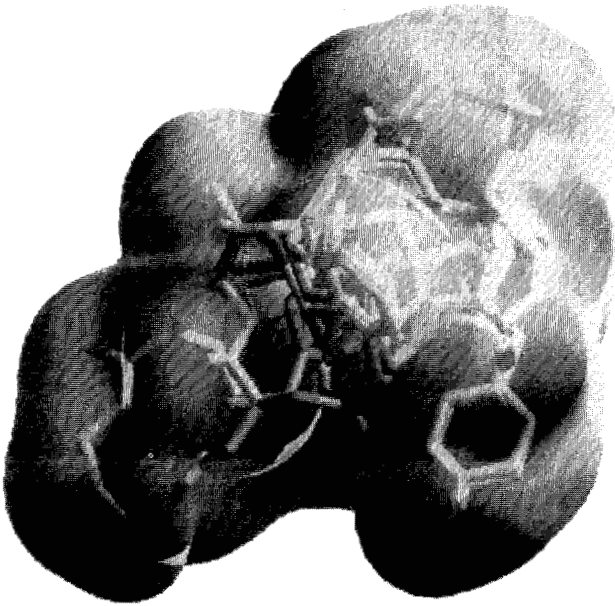


Figure 5. Receptor model for  $\delta$ -receptor ligands.

a surface point. The resulting receptor with seven ligands used to construction of the receptor is shown in Fig. 4. All molecules from reference and trial sets were minimized inside the receptor surface and total interaction energies at each surface points of the receptor were analyzed by statistical method. Resulting QSAR regression provided  $R = 0.86$  and least square error of prediction  $LSE = 0.57$ . Rather low correlation coefficient suggests that further modifications of construction of the receptor are desirable.

The same procedure was conducted for trial construction of  $\delta$ -selective receptor model from seven most active  $\delta$ -receptor ligands from reference set of molecules. The receptor surface together with seven ligands used for construction is shown in Fig. 5. The QSAR regression resulted in  $R = 0.85$ ,  $LSE = 0.45$ . RSA method is more sensitive to inaccuracy in alignment than MFA and hence smaller correlation coefficients.

Models of  $\mu$  and  $\delta$ -receptors differ mainly in shape of a tail (left parts of Fig. 4 and 5). In case of  $\mu$ -receptor the tail is constructed from hydrophobic residues of most potent  $\mu$  receptor ligands whereas in case of  $\delta$ -receptor the hydrophilic extensive residues elongate the tail. It would mean that either there are hydrophilic residues in that part of the receptor or that place is accessible by a solvent. In the latter case open model of the  $\delta$ -receptors should be employed. To explain uncertainties in both models a greater set of reference molecules will be selected.

## ACKNOWLEDGEMENTS

This work was supported by the University of Warsaw (BW-1383/5/97).

The computational task was partly done in the ICM computer center, University of Warsaw.

## REFERENCES

1. P.W. Schiller, *The Peptides, Analysis, Synthesis, Biology*, Vol. 6, S. Udenfriend and J. Meienhofer, eds., Academic Press, Orlando, FL (1984).
2. V.J.Hruby, and G.G. Bonner, *Methods in Molecular Biology*, Vol. 35, M.W. Pennington and B.M. Dunn, eds., Humana Press Inc., Totowa, NJ (1994).
3. J. DiMaio, and P.W. Schiller, A cyclic enkephalin analog with high *in vitro* opiate activity, *Proc. Natl. Acad. Sci. USA*, 77:7162 (1980).
4. J. DiMaio, T.M.-D. Nguyen, C. Lemieux, and P.W. Schiller, Synthesis and pharmacological characterization in vitro of cyclic enkephalin analogues: effect of conformational constraints on opiate receptor selectivity, *J. Med. Chem.* 25:1432 (1982).
5. P.W. Schiller, B. Eggimann, J. DiMaio, C. Lemieux, and T.M.-D. Nguyen, Cyclic Enkephalin Analogs containing a cystine bridge, *Biochem. Biophys. Res. Commun.* 101:337 (1981).
6. P.W. Schiller, T.M.-D. Nguyen, L. Masiak and C. Lemieux, A novel cyclic opioid peptide analog showing high preference for  $\mu$ -receptors, *FEBS Lett.* 191:231 (1985).
7. H.I. Mosberg, R. Hurst, V.J. Hruby, J.J. Galligan, T.F. Burks, K. Gee, and H.I. Yamamura, [D-Pen<sup>2</sup>, L-Cys<sup>5</sup>]enkephalinamide and [D-Pen<sup>2</sup>, D-Cys<sup>5</sup>] enkephalinamide, conformationally constrained cyclic enkephalinamide analogs with delta receptor specificity, *Biochem. Biophys. Res. Commun.* 106:506 (1982).
8. H.I. Mosberg, R. Hurst, V.J. Hruby, K. Gee, H.I. Yamamura, J.J. Galligan, and T.F. Burks, Bis-penicillamine enkephalins possess highly improved specificity toward  $\delta$  opioid receptors, *Proc. Natl. Acad. Sci. USA* 80:5871 (1983).
9. P.W. Schiller, T.M.-D. Nguyen, L. Maziak, and C. Lemieux, A novel cyclic opioid peptide analog showing high preference for  $\mu$ -receptors, *Biochem. Biophys. Res. Commun.* 127:558 (1985).
10. Insight II, Release 95.0, Biosym/MSI, San Diego, 1995.
11. Cerius<sup>2</sup> release 3.5 (1997), MSI Molecular Simulations Inc., San Diego.
12. A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard, and W.M. Skiff, UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations, *J. Amer. Chem. Soc.*, 114:10024 (1992).
13. W. Brandt, A. Barth, and H.D. Höltje, A new consistent model explaining structure (conformation)-activity relationships of opiates with  $\mu$ -selectivity, *Drug Design and Discovery*, 10:257 (1993).
14. W. Brandt, C. Mrestani-Klaus, H. Schinke, K. Neubert, A. Barth, R. Schmidt, P.W. Schiller, and H.D. Höltje, The  $\mu$  opioid receptor binding pharmacophore conformation of ornithine containing cyclic  $\beta$ -casomorphin analogues and related peptides, *Quantitative Structure-Activity Relationship*, 14:417 (1995).
15. S. Ro, Q. Zhu, C.W. Lee, M. Goodman, K. Darlak, A.F. Spatola, N.N. Chung, P.W. Schiller, A.B. Maimberg, T.L. Yaksh, and T.F. Burks, Highly potent side chain - main chain cyclized dermorphin-deltorphin analogues: an integrated approach including synthesis, bioassays, NMR spectroscopy and molecular modelling, *J. Peptide Sci.* 3:157 (1995).
16. A. Misicka, A.W. Lipkowski, R. Horvath, P. Davis, H.I. Yamamura, F. Porreca, and V.J. Hruby, Design of cyclic deltorphins and dermenkephalins with a disulfide bridge leads to analogues with high selectivity for delta-opioid receptors, *J. Med. Chem.* 37:141 (1994).
17. D. Pawlak, N.N. Chung, P.W. Schiller, and J. Izdebski, Synthesis of a novel side-chain to side-chain cyclized enkephalin analogue containing a carbonyl bridge, *J. Peptide Sci.* 3:277 (1997).

## UNIFIED PHARMACOPHORIC MODEL FOR CANNABINOIDS AND AMINOALKYLINDOLES

Joong-Youn Shim<sup>1</sup>, Elizabeth R. Collantes<sup>1,3</sup>, William J. Welsh<sup>1</sup>,  
and Allyn C. Howlett<sup>2</sup>

<sup>1</sup>Department of Chemistry and Center for Molecular Electronics University of Missouri-St. Louis, St. Louis, MO 63121

<sup>2</sup>Department of Pharmacological and Physiological Science, Saint Louis University School of Medicine, St. Louis, MO 63104

<sup>3</sup>Current address: Monsanto Life Science Company, St. Louis, MO 63167

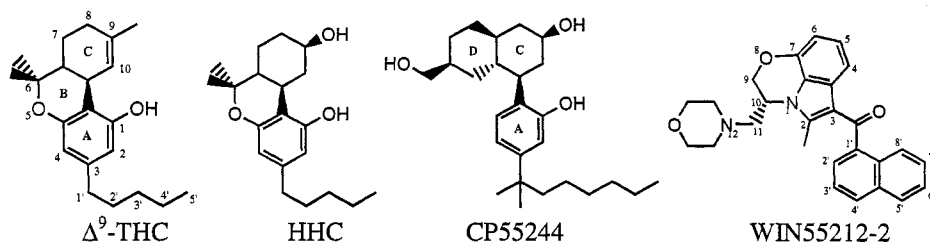
### INTRODUCTION

Despite their obvious structural dissimilarities, the cannabinoids and aminoalkylindoles (AAIs) are known to exhibit similar *in vitro* and *in vivo* cannabimimetic activities<sup>1-6</sup>. This body of evidence suggests that the cannabinoids and AAIs interact with the same cannabinoid receptor and share at least some regions in common when bound to the receptor to elicit the cannabinoid activity. In support of this hypothesis, binding studies have shown that they compete for the same binding regions of the CB<sub>1</sub> cannabinoid receptor<sup>7</sup>.

Structure-activity relationship (SAR) studies of the cannabinoids<sup>8-11</sup> and AAIs<sup>12-14</sup> have identified pharmacophoric elements common to both classes of compounds: (1) a lipophilic and/or sterically bulky group (i.e., the C3 side chain in the cannabinoids and the C3 aryl group in the AAIs) that appears to be a structural prerequisite for cannabinoid activity<sup>10</sup>; (2) a polar oxygen atom (i.e., the C1 hydroxyl group of the phenolic A-ring in the cannabinoids and the C3 carbonyl oxygen in the AAIs) that may form a hydrogen bond with the receptor<sup>15</sup>; and (3) the cyclic ring system (i.e., the cyclohexyl C-ring in the cannabinoids and the indole ring in the AAIs).

In order to understand the similarity in cannabimimetic activity of the cannabinoids and AAIs in terms of their common pharmacophoric features, two superimposition models have been proposed<sup>16-18</sup>. The Huffman model<sup>17</sup>, developed by superimposing a structurally modified analog of WIN55212-2 with  $\Delta^9$ -THC (Figure 1), assumes a similar functionality between the C3 side chain of the cannabinoids and the N1 aminoalkyl side chain of the AAIs. In this model, the benzene ring in the indole moiety of WIN55212-2 is not overlaid with any part of  $\Delta^9$ -THC, thus

implying that this benzene ring is unimportant. This decision may be unwarranted in light of the sharp decrease in both in vivo and in vitro activities of a series of pyrrole analogues<sup>19</sup> versus the corresponding AAI analogues. The Huffman model was derived without precise structural information, and suffers from a lack of consideration of the polar nature of the heterocyclic N1 aminoalkyl side chain of the AAIs<sup>3</sup>.



**Figure 1.** The cannabinoids and AAIs used for developing superimposition models.

The Makriyannis model<sup>18</sup> was derived by superimposition of HHC with WIN55212-2 whose structure was ascertained from interpretation of 2D-NMR spectra and MD simulations. In this model, the C3 aryl group of the AAI is superimposed on the C3 side chain of the cannabinoid, and the N atom of the N1 side chain of the AAI was positioned to nearly coincide with the hydroxyl group of the cyclohexyl C-ring of the cannabinoid. Nevertheless, the rationale for which specific atoms were fitted in the superimposition remains unclear.

Although largely incompatible, the Huffman and Makriyannis models both align the C1 hydroxyl group of the cannabinoids with the C3 aryl oxygen of the AAIs. This point of agreement suggests a common hydrogen-bonding interaction with the corresponding region of the receptor's binding site<sup>15</sup>. Eissenstat et al.<sup>12</sup> recently proposed a model in which the C1 hydroxyl group of the cannabinoids overlays the N1 side chain of the AAIs, based on the observation that the C9 hydroxyl group of the cannabinoids functioned differently from the morpholino N in the pravadoline series<sup>14</sup>. As yet, however, no unified superimposition model for both cannabinoids and AAIs has been generally accepted.

In the present work, novel superimposition models were developed based on 3-D pharmacophore mapping of two highly potent CB<sub>1</sub> cannabinoid receptor agonists CP55244 and WIN55212-2. Initial conformations corresponded to those ascertained by Tong et al.<sup>20</sup> and Shim et al.<sup>21</sup>, respectively. CP55244 possesses an additional pharmacophoric element which is not found in the classical ABC tricyclic cannabinoids like  $\Delta^9$ -THC and HHC<sup>11</sup>. The D-ring methanol extension (comparable to the hydroxypropyl in CP55940) forms a potential hydrogen bonding site which may confer the extremely high potency exhibited by CP55244. The superimposition models so derived confirm earlier speculation about certain key pharmacophoric elements common to both the cannabinoids and AAIs.

## COMPUTATIONAL METHODS

The highly potent CP55244 ( $K_i = 0.11$  nM) and WIN55212-2 ( $K_i = 1.1$  nM) were selected to represent the cannabinoids and AAIs, respectively<sup>11,21</sup>. The conformations of the cannabinoids and AAIs were taken from our previously derived CoMFA models<sup>20,21</sup>. For WIN55212-2,

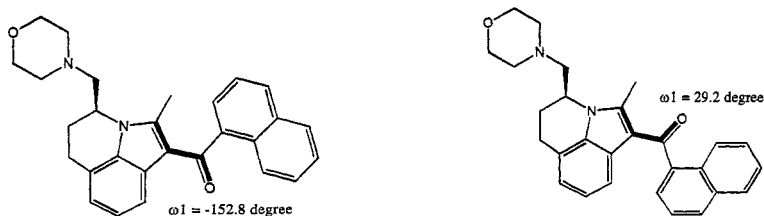
additional conformations were explored by conducting a systematic search of the torsion angles  $\omega_1(\text{C}2=\text{C}3-\text{C}=\text{O})$  and  $\omega_2(\text{O}=\text{C}-\text{C}1'-\text{C}2')$ .

The DISCO module [DISTANCE COMPARISON (DISCO) technique<sup>23</sup>], accessed through the molecular modeling program Sybyl (version 6.2)<sup>24</sup>, was employed to extract the common pharmacophoric elements from the cannabinoids and AAIs. DISCO generates superimposition models by matching common features after identifying certain predefined pharmacophoric features, i.e., hydrophobic center, donor site, acceptor site, donor atom, acceptor atom, for each compound. Based on these superimposition models, the corresponding pharmacophoric elements were identified. Superimposition models of CP55244 and WIN55212-2 were compared and evaluated using the following criteria: (1) root-mean-square (RMS) fit of corresponding pharmacophoric elements, (2) proper orientation and overlap of the C3 dimethylheptyl side chain of CP55244 with the C3 aroyl moiety of WIN55212-2 (which was deemed critical for tight binding), (3) the number of pharmacophoric elements, and (4) the degree of overlap of molecular volumes. The superimposition models were also chosen to ensure proper orientation and overlap of the C3 dimethylheptyl side chain. WIN55212-2 was used as the reference compound for fitting as it contains a greater number of pharmacophoric features than CP55244. The superimposition models selected by DISCO were further refined by fitting WIN55212-2 to CP55244 using the “field fit” option in Sybyl in which the partial atomic charges for the electrostatic interactions were calculated using the Gasteiger-Marsili formalism<sup>25</sup>. To compare with the present superimposition models, the Huffman and Makriyannis models were reconstructed by superimposing CP55244 and WIN55212-2 in the conformations considered in the present report using the same alignment atoms as described in the respective original papers.

## RESULTS AND DISCUSSION

### Superimposition of CP55244 and WIN55212-2

DISCO<sup>23</sup> was employed to help identify the corresponding pharmacophoric elements in the cannabinoids (represented by CP55244) and the AAIs (represented by WIN55212-2). DISCO found two separate low-energy AAI conformers, designated Z and C, that differ with respect to the torsion angle  $\omega_1(\text{C}2=\text{C}3-\text{C}=\text{O})$ . The value of  $\omega_1$  is  $-152.8^\circ$  in the Z form and  $29.2^\circ$  in the C form (Figure 2). With WIN55212-2 in the Z form, DISCO identified five pharmacophoric features: (i) two around the C1 phenolic oxygen of CP55244 and the C3 carbonyl oxygen of WIN55212-2 (oxygen as the acceptor atom and a donor site), (ii) one hydrophobic center for the C-ring of CP55244 and the benzene ring of the indole of WIN55212-2, and (iii) two around the D-ring hydroxyl group of CP55244 and the morpholino oxygen of WIN55212-2 (oxygen as an acceptor atom and a donor site). With WIN55212-2 in the alternative C form, DISCO identified three pharmacophoric features: (i) two around the C1 phenolic oxygen of CP55244 and the C3 carbonyl oxygen of WIN55212-2 (oxygen as an acceptor atom and two donor sites), and (ii) one around the C9 hydroxyl oxygen of CP55244 and the morpholino nitrogen of WIN55212-2 as a donor atom. Both models displayed a high degree of overlap between the C3 side chain of CP55244 and the C3 aroyl moiety of WIN55212-2, consistent with the notion that a hydrophobic moiety is important for cannabimimetic activity<sup>10,12-14,26</sup>. In addition, both models would predict that addition of hydrophobic substituents to the second ring of the naphthyl group (i.e., 6' or 7' position) in WIN55212-2 enhances binding potency.



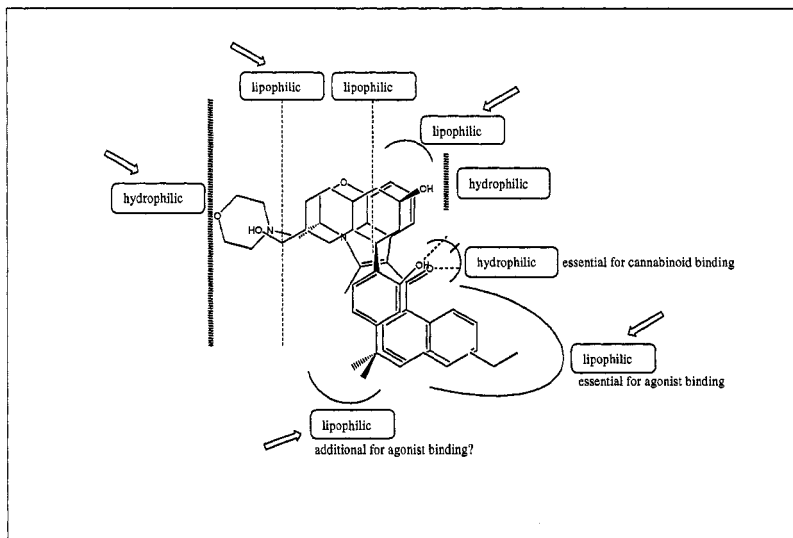
**Figure 2.** Illustration of the AAI WIN55212-2 in both the Z form (left) and C form (right).

Analysis of the present two superimposition models (hereafter called the Z and C models) provides some interesting comparisons. The molecular volume overlap is only slightly larger for the Z model (156 Å<sup>3</sup>) than for the C model (142 Å<sup>3</sup>). In the Z model, the morpholino oxygen of WIN55212-2 is aligned with the D-ring hydroxyl group of CP55244. In the C model, the morpholino nitrogen of WIN55212-2 is aligned with the C-ring hydroxyl group of CP55244. The C model bears a likeness to the Makriyannis model<sup>18</sup>; however, the corresponding Z model shows no similarity to the Huffman model<sup>17</sup>. The C model better emphasizes the importance of the C-ring C9 hydroxyl group in the cannabinoids and its similarity to the morpholino nitrogen in the AAIs. However, the C9 hydroxyl group of the cannabinoids may be not essential for potent binding<sup>10,22</sup>. Furthermore, the cannabinoid C9 hydroxyl and the AAI morpholino nitrogen interact with different receptor binding sites<sup>12</sup>. Noting that the sensitivity of K<sub>i</sub> to the length of the lipophilic alkyl N1 substituent for a series of AAIs was similar to that observed for the cannabinoids with respect to variation in length of the lipophilic C3 alkyl side chain, Huffman et al.<sup>17</sup> overlaid these two lipophilic groups in their superimposition model. This choice would imply that the hydrophilic side chain of the AAIs (e.g., the O atom in the morpholino ring of WIN55212-2) is not critical. However, the Z model could explain why AAIs with an N1 side chain of four to seven carbons exhibit high potency. Earlier workers have proposed that a specific hydrophobic region of the receptor borders the B and D rings of cannabinoids<sup>10,11,27</sup>. Consistent with this notion, the Z model superimposes the N1 side chain of the AAIs on the hydrophobic substituents attached to the B and D rings of the cannabinoids. By virtue of its ability to resolve this apparent inconsistency, the Z model may be superior to alternative superimposition models in terms of accommodating the structurally dissimilar cannabinoids and AAIs inside the same critical binding site of the CB<sub>1</sub> cannabinoid receptor.

### Proposed Cannabinoid Receptor Map

Based on our superimposition models and the known SAR for the cannabinoid and AAIs, we have constructed a pharmacophoric map for the cannabinoid CB<sub>1</sub> receptor appropriate to both the cannabinoids and AAIs (illustrated in Figure 3 for the Z model). Similar to the one proposed by Howlett et al.<sup>27</sup> from the SAR for bi- and tricyclic nonclassical cannabinoids, this receptor map depicts the pharmacophoric elements required for cannabimimetic activity including those common to both the cannabinoid and AAI compounds. The map also shows those pharmacophoric elements that are specific for each compound, such as a lipophilic receptor site near the benzene ring of the AAIs and a hydrophilic receptor site next to the C9 hydroxyl of the cannabinoids. Inspection of our proposed CB<sub>1</sub> cannabinoid receptor map reveals that WIN55212-2 could be accommodated inside the binding site in either the C and Z models. Both conformations of WIN55212-2 seem capable of satisfying those interactions with the receptor deemed necessary for tight binding. In

fact, the region of the receptor binding site in contact with the C and D rings of the cannabinoids and with the indole and N1 side chain of the AAIs appears to possess the proper distribution of hydrophilic and lipophilic sites with respect to the C6-C7 axis of the cannabinoids and the C3-C(carbonyl) of the AAIs (i.e., complementary hydrophilic-lipophilic sites on top left and top right of Figure 3). So either conformation of WIN55212-2 could function accordingly.



**Figure 3.** Putative pharmacophoric model of the CB<sub>1</sub> cannabinoid receptor showing possible interactions with both cannabinoid and AAI agonists (Z conformation).

## ACKNOWLEDGMENTS

This work was supported in part by National Institute on Drug Abuse (NIDA) grants R01-DA06312 and K05-DA00185 to ACH.

## REFERENCES

1. S. J. Ward, E. Baizman, M. Bell, S. Childers, T. D'Ambra, M. Eissenstat, K. Estep, D. Haycock, A. Howlett, D. Luttinger, M. Miller, and M. Pacheco, In *Problems of Drug Dependence*; Harris, L. S., Ed.; NIDA Research Monograph, National Institute on Drug Abuse: Rockville, MD, 425 (1990).
2. C. C. Felder, J. S. Veluz, H. L. Williams, E. M. Briley, and L. Matsuda, *Mol. Pharmacol.* 42: 838 (1993).
3. A. C. Howlett, B. Berglund, and L. S. Melvin, *Curr. Pharmaceut. Des.* 1:343 (1995).
4. R. Mechoulam, A. Breuer, R. U. C. Jarbe, A. J. Hiltunen, and R. Glaser, *J. Med. Chem.* 33:1037 (1990).
5. B. R. Martin, D. R. Compton, B. F. Thomas, W. R. Prescott, P. J. Little, R. K. Razdan, M. R. Johnson, L. S. Melvin, R. Mechoulam, and S. J. Ward, *Pharmacol. Biochem. Behav.* 40:471(1991).
6. J. E. Kuster, J. I. Stevenson, S. J. Ward, T. E. D'Ambra, and D. A. Haycock, *J. Pharmacol. Exp. Ther.* 264:1352 (1993).



7. K. Yamada, K. C. Rice, J. L. Flippen-Anderson, M. A. Eissenstat, S. J. Ward, M. R. Johnson, and A. C. Howlett, *J. Med. Chem.* 39:1967 (1996).
8. R. K. Razdan, *Pharmacol. Rev.* 38:75 (1986).
9. L. S. Melvin and M. R. Johnson, In *Structure-Activity Relationships of the Cannabinoids*; NIDA Research Monograph 79; National Institute on Drug Abuse: Rockville, MD; 31 (1987).
10. L. S. Melvin, G. M. Milne, M. R. Johnson, B. Subramaniam, G. H. Wilken, and A. C. Howlett, *Mol. Pharmacol.* 44:1008 (1993).
11. L. S. Melvin, G. M. Milne, M. R. Johnson, G. H. Wilken, and A. C. Howlett, *Drug Design and Discovery* 13:155 (1995).
12. M. A. Eissenstat, M. R. Bell, T. E. D'Ambra, E. J. Alexander, S. J. Daum, J. H. Ackerman, M. D. Gruett, V. Kumar, K. G. Estep, E. M. Olefirowicz, J. R. Wetzel, M. D. Alexander, J. D. Weaver, III, D. A. Haycock, D. A. Luttinger, F. M. Casiano, S. M. Chippari, J. E. Kuster, J. I. Stevenson, and S. J. Ward, *J. Med. Chem.* 38:3094 (1995).
13. T. E. D'Ambra, K. G. Estep, M. R. Bell, M. A. Eissenstat, K. A. Josef, S. J. Ward, D. A. Haycock, E. R. Baizman, F. M. Casiano, N. C. Beglin, S. M. Chippari, J. D. Grego, R. K. Kullnig, and G. T. Daley, *J. Med. Chem.*, 35:124 (1992).
14. M. R. Bell, T. E. D'Ambra, V. Kumar, M. A. Eissenstat, J. L. Herrmann, Jr., J. R. Wetzel, D. Rosi, R. E. Philion, S. J. Daum, D. J. Hlasta, R. K. Kullnig, J. H. Ackerman, D. R. Haubrich, D. A. Luttinger, E. R. Baizman, M. S. Miller, and S. J. Ward, *J. Med. Chem.* 34:1099 (1991).
15. P. H. Reggio, K. V. Greer, and S. M. Cox, *J. Med. Chem.* 32:1630 (1989).
16. P. H. Reggio, A. M. Panu, and S. Miles, *J. Med. Chem.* 36:1761 (1993).
17. J. W. Huffman, D. Dai, B. R. Martin, and D. R. Compton, *Bioorganic & Medicinal Chem. Lett.* 4:563 (1994).
18. X.-Q. Xie, M. Eissenstat, and A. Makriyannis, *Life Sci.* 56:1963 (1995).
19. J. A. H. Lainton, J. W. Huffman, B. R. Martin, and D. R. Compton, *Tetrahedron Lett.* 36:1401 (1995).
20. W. Tong, E. R. Collantes, A. C. Howlett, and W. J. Welsh, *J. Med. Chem.* 41:4207 (1998).
21. J. Y. Shim, E. R. Collantes, W. J. Welsh, B. Subramaniam, A. C. Howlett, M. A. Eissenstat, and S. J. Ward, *J. Med. Chem.* 41:4521 (1998); J.-Y. Shim, E. R. Collantes, W. J. Welsh, and A. C. Howlett, In *ACS Symposium Series on Rational Drug Design*, American Chemical Society (in press).
22. J. W. Huffman, S. Yu, V. Showalter, M. E. Abood, J. L. Wiley, D. R. Compton, B. R. Martin, R. D. Bramblett, and P. H. Reggio, *J. Med. Chem.* 39:3875 (1996).
23. DISCO: Y. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, and P. A. Pavlik, *J. Comp.-Aid. Mol. Des.* 7:83 (1993).
24. The molecular modeling program Sybyl is a product of Tripos, Inc., St. Louis, MO 63144.
25. J. Gasteiger and M. Marsili, *Tetrahedron* 36:3219 (1980).
26. R. S. Rapaka and A. Makriyannis, In *Structure-Activity Relationships of the Cannabinoids*; NIDA Research Monograph 79; National Institute on Drug Abuse: Rockville, MD (1987).
27. A. C. Howlett, M. R. Johnson, L. S. Melvin, and G. M. Milne, *Mol. Pharmacol.* 33:297 (1988).

## CHEMOMETRIC DETECTION OF BINDING SITES OF 7TM RECEPTORS

Monica Clementi<sup>1</sup>, Sara Clementi<sup>1</sup>, Sergio Clementi<sup>1</sup>, Gabriele Cruciani<sup>1</sup>,  
Manuel Pastor<sup>1</sup> and Jonas E. Nilsson<sup>2</sup>

<sup>1</sup> Department of Chemistry  
University of Perugia  
Perugia, Italy

<sup>2</sup> Pharmacia & Upjohn  
Uppsala, Sweden

### INTRODUCTION

Over the last few years much attention was paid to 3D-QSAR studies<sup>1,2</sup> owing to the incorporation of information derived from molecular modelling techniques such as GRID<sup>3</sup>, CoMFA<sup>4</sup> and many others into the chemometrics procedures aimed at deriving a statistical model. Our research group has contributed to the field by suggesting the GOLPE<sup>5</sup> approach, which is based on a severe validation criterion and on a region selection procedure<sup>6</sup> in order to produce easily interpretable results and reliable predictions of the activity. Such studies are generally carried out on small molecules acting as ligands or enzyme inhibitors. On the contrary, relatively less attention was devoted to a quantitative description of the receptors.

Since Wold and coworkers published the principal properties (PP.s) of aminoacids (AA.s), peptide description in traditional QSAR studies has been done describing the structural variation within a series of related peptides by arranging the PP.s according to the AA sequence<sup>7</sup>. The PP.s of AA.s represented a great improvement in peptide QSAR since they permit both a quantitative description of peptides and the use of experimental design criteria using few orthogonal variables (z scales in ref. 7), for selecting a few informative molecules in each series. The three PP.s used can be chemically interpreted as scales of the hydrophobicity (z1), size (z2) and electronic properties (z3). Accordingly, each peptide is described by a vector containing the triplets of PP.s for the AA sitting at positions 1, 2, etc., and therefore, peptides of different length require a different number of descriptors.

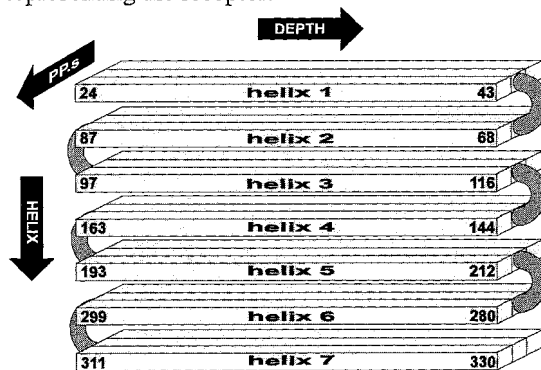
However, each descriptor is considered as a single column in a QSAR table and continuity constraints arising by the AA sequence are not explicitly formulated. In order to overcome this traditional QSAR limitation, the same Authors have later suggested to describe the AA sequence in peptides by their auto and cross covariance (ACC) based description, which depends only upon the sequence of AA.s and not on the length of the peptide<sup>8</sup>. Our research group has already used the former technique for studying peptides<sup>9</sup> and the ACC transforms for describing both peptidic fragments<sup>10</sup> and 3D structures<sup>11</sup>. This

paper will attempt to describe 7TM receptors by such approaches and to check whether this description can be related to their biological behaviours. Indeed, it should be pointed out that the same description, either adding triplets of PP.s for each AA of the sequence<sup>7, 9</sup> or the ACC tems<sup>8, 10</sup>, can be used to characterize short peptides acting as ligands or long peptidic fragments of whole proteins. Of course in the former case the objective is to find out the AA sequence with optimal response, whereas in the latter the aim is detecting structural features that are common and/or that are important for modulating the response.

Furthermore our group has developed a new generation of PP.s for AA.s<sup>12</sup>, obtained by a multivariate characterization on 3D and 2D data derived by GRID with the strategy just reported for heteroaromatics<sup>13</sup>. The first PP can be taken as a measure of the side-chain polarity, the second describes hydrophobicity effects and the third one indicates hydrogen-bonding capabilities: positive values indicate polar, hydrophobic/large and hydrogen-bond donor properties respectively.

## RECEPTOR DESCRIPTION

**Depth Description.** In principle each seven helix trans membrane (7TM) receptor can be represented by a 3-mode array of data where the elements of the 3-way matrix are the PP values of AA.s and the modes represent: (1) the sequence of AA.s in each helix limited to 20 AA.s<sup>14</sup>, (2) the 7 helices, (3) the 3 PP.s. A set of receptors then constitutes a 4-mode data set. The order of AA.s for each helix follows the depth of each AA along the TM channel starting from the outer part of the cell: accordingly, while for the odd helices the order is that reported in the normal sequence, for the even helices the order is opposite to the usual sequence numbering. Figure 1 illustrates the organization of the data matrix with respect to AA.s and helices. The 3-way matrix can be thereafter deconvoluted into a vector of 420 elements, thus representing the receptor.



**Figure 1.** Characterization of a 7TM receptor in terms of depth from the cell surface.

**ACC Description.** The idea of autocorrelation or autocovariance transforms of the data, together with Fourier transforms, have been developed to account for dependencies between consecutive observations. Wold et al.<sup>8</sup> described peptides by the ACC functions along the sequence from NH to CO. On using two descriptors (PP scales) for each aminoacid one obtains four nearest neighbour ACC.s: the autocorrelation between PP1 for AA(i) and PP1 for AA(i+1) and between PP2 for AA(i) and PP2 for AA(i+1), as well as the cross-correlations between PP1 for AA(i) and PP2 for AA(i+1), which is different from PP2 for AA(i) and PP1 for AA(i+1). With three descriptors the number of ACC terms becomes 9, while with a single descriptor there is a single ACC term.

These parameters account only for the nearest neighbour interaction, which is also called lag 1. The same transformation applied to the next nearest neighbour provides four chains belonging to the two aminoacids separated by one further aminoacid. The number of lags can be increased up to the length of the shortest peptide minus one. To simplify the description and calculations Wold and coworkers used the auto-and cross-covariance functions instead of auto-and cross-correlations, since they are the same after scaling. Their work has already shown advantages and drawbacks of modelling by ACC transforms. On one side such a description is independent of length and alignment so that peptides of different length can be described in a congruent way. Moreover, ACC transforms permit to model consistent dependencies between neighbouring sequence positions, and therefore to find out the need for the simultaneous presence of certain structural features at some fixed distances. However, interpretation and understanding of ACC models may be quite difficult. In fact we suggested<sup>12</sup> different modifications of the original application of ACC on biopolymers aimed at a better understanding of QSAR models. The interpretation of the original ACC models was difficult since the AC (Auto Covariance) and CC (Cross Covariance) terms were obtained as the sums of all possible AA.s interactions and it was not easy to interpret the results in terms of which interaction is the most important in a QSAR analysis.

During our work with ACC transform for developing 3D-ACC.s<sup>11</sup> we came across a number of possible drawbacks with the original ACC approach, which might lead to unclear interpretation. The first point regards the different leverage that different ranges for each descriptor to the individual elements of the ACC vector. Normalizing each descriptor between -1 and +1, without any subsequent scaling of the QSAR table, gives a straightforward way of understanding the magnitude of each individual interaction.

**Table 1.** Receptor affinities (pK<sub>i</sub>)<sup>16</sup>

	<b>Receptor</b>	<b>chlorpromazine</b>	<b>haloperidol</b>	<b>clozapine</b>
1	D1	7.49	8	7.07
2	D2L	8.52	9.3	6.98
3	D3	8.4	8.7	6.52
4	D4	7.47	8.7	7.5
5	D5		7.57	6.66
6	5HT1A	5.44	5.77	6.06
7	5HT2A	8.15	7.13	8.1
8	5HT2C	7.92	5.24	7.92
9	5HT6	8.4	<5.3	8.4
10	M1	7.6	5.77	8.7
11	M2	6.82	5.6	7.68
12	M3	7.17	<5.52	7.89
13	M4	7.4	5.57	7.92
14	M5	7.38	5.74	8.43
15	H1		5.44	8.22
16	α1B	8.3	7.34	8.15
17	α2B		6.44	8.1
18	β1		<5	<5

The second, most important drawback with the original ACC is related to the use of the algebraic sums of all members for each element. The meaning of signs is chemically recognizable for each scale: a positive PP1 means polar, whereas a negative PP1 means non polar, a positive PP2 means large and hydrophobic and a negative PP2 means small and hydrophilic, a positive PP3 means H-donor and a negative PP3 H-acceptor. In the algebraic sum the positive-positive interactions are added to the negative-negative interactions, while the positive-negative interactions are subtracted to the sum. This approach might be

appropriate for handling time-dependent continuity constraints, but it mixes up chemically different interactions that should be kept well separated for the safety of chemical interpretation. Therefore, in order to have pure effects (the interaction hydrophobic-hydrophobic is  $PP2(+)*PP2(+)$ ) we decided to keep disjoint the different kinds of interactions bearing the same chemical meaning: positive-positive, negative-negative, positive-negative and negative-positive for the auto covariance and the cross-covariance terms: using three descriptors for each lag we have 36 ACC terms: 12 for the AC and 24 for the CC terms. We suggested to call DACC (Disjoint Auto and Cross Covariance transform) this modified transformation.

A third point regards the magnitude of each element of the original ACC vector: it should be noticed that each element is given by the average value of the individual products which constitute each term of sum. Because of this, the same value can be obtained either from several values of intermediate magnitude or from one large value and a number of small values. Describing a peptide in terms of the interactions between aminoacids in the sequence may be a useful tool for labelling the really strong interactions that the peptide can offer to bind to a receptor. To this end, the presence of one strong interaction is by far more important than the sum of several weaker interactions. On the other end, the presence of two such interactions means that there are two possibilities to bind to the receptor with the same strength.

In view of these considerations it seemed appropriate to keep for each element of the ACC vector only the maximum term of the sum instead of their average. Consequently we suggested to use the MACC1 transformation (Maximum Auto and Cross Covariance in one direction). However, if a peptide can bind to a protein in two ways some of the ACC terms should be considered together, since they describe the same type of interaction. In particular,  $ACi+-$  and  $ACi-+$ ,  $CCij--$  and  $CCji--$ ,  $CCij++$  and  $CCji++$ ,  $CCij+-$  and  $CCji+-$ , and  $CCij+-$  and  $CCji+-$  are equivalent. Consequently, in this shrunked description with three descriptors we have 21 terms, i.e. 9 AC and 12 CC terms for each lag and we suggested to call this modification MACC2 (two directions).

## RESULTS AND DISCUSSION

Following our previous interest in the field<sup>15</sup>, we worked out a data set collected from literature regarding the binding activities of three molecules used in the treatment of schizophrenia<sup>16</sup> (Table 1) towards all subtypes of dopamine and muscarinic receptors, plus a number of serotonin, histamine and adrenergic receptor subtypes<sup>14</sup>. We determined a few PCA and PLS models, both within and between receptor types, in terms of the 7TM description of AA depth (420 variables = 7 helices x 20 AA.s x 3 PP.s) and of the MACC2 transform (588 variables = 7 helices x 21 MACC2 terms x 4 lags). Objectives of the study are comparing the two descriptions for distinguishing receptor subtypes and providing chemometric tools to find out common features and features responsible for modulating the binding affinities.

**PCA Results.** Figure 2 illustrates the principal components results of the two descriptions. It is clear that in both cases there are clusters of subtypes belonging to the same receptor. However, the two descriptions are different as shown by the relative position of the five subtypes for both D and M receptors.

**PLS Analysis.** All six PLS models (3 molecules x 2 descriptors) explain over 70-80% of the y variance by a single latent variable, although the clustering of objects prevents from performing a proper validation according to the GOLPE criteria.

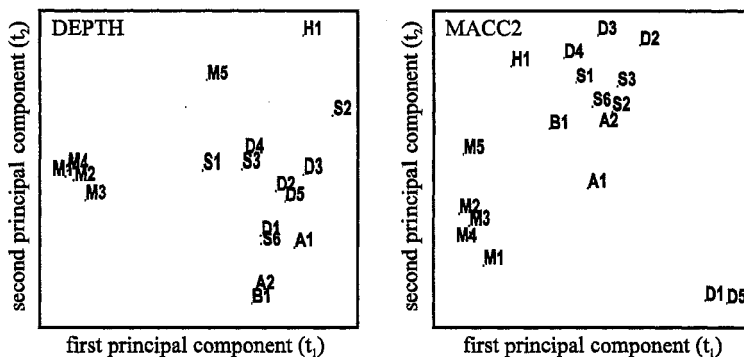


Figure 2. PC score plots of receptors: a) depth model, b) MACC2 model.

The interpretation of the latent variable allows to detect the structural features which are common for all receptor subtypes as well as those responsible for modulating the binding affinities.

**Depth Models.** By the depth models it is possible to spot the AA.s sitting in the same position (because they show no variance): none throughout the series and four within the D receptor subtypes (Leu in positions 12 and 17 of helix 2, Ala in position 14 of helix 2 and Pro in position 6 of helix 4). However from the PLS weights it is possible to recognize the most important AA.s responsible for ranking the affinities: AA5 of helix 4 (which is Val or Leu for D receptors and Ala for M receptors) for chlorpromazine and haloperidol, AA18 of helix 2 (which is Asp or Ser for D receptors and Leu for M receptors) for haloperidol and clozapine, etc. Results are much clearer within a single series of subtypes: for the D subtypes it is always possible to detect a few AA.s distinguishing between the higher and lower affinities of each ligand. However, the results obtained by this type of description are strictly dependent upon the problem formulation and we know that this approach does not take into account the homology studies leading to the alignment of helices of different receptors. As in the latter case the chemometric results might be used to check the importance of the detected AA.s by site-directed mutagenesis.

**MACC2 Models.** More suitable results are obtained by the MACC2 description which takes into account only the AA sequence, independently of alignment. Since the relative affinities of the different subtypes are roughly reversed for haloperidol and clozapine the relevant features are the same but with opposite signs. Particularly important, among others, appear to be several interactions at lag 2 and lag 4 in helix 3 and at lag 1 and 4 in helix 4; significantly the same interactions exhibit high loading also in the chlorpromazine model. The detailed analysis limited to the D receptors focus on the differences between the highest and lowest affinities: for clozapine D4 differs from the others because of seven interactions in helices 1, 2 and 3, that are always lower than in the other subtypes; for chlorpromazine the pair D2/D3 differs from the pair D1/D4 because of eight interactions in helices 2, 3 and 4.

## CONCLUSIONS

The paper has shown that both the depth and MACC2 description can be used to characterize receptor subtypes and to detect structural features responsible for the biological

activity. However none of them is completely satisfactory.

In fact the depth description depends too heavily upon the way helices are aligned: our suggestion, i.e. starting the helix from the cell surface, is not in agreement with the widely recognized need for alignment according to homology criteria. Indeed our problem formulation may be inconsistent because it was by these criteria that it was postulated which is the AA starting the helix. Moreover, to the end of describing a whole 7TM receptor, it seems that the relative position of the seven helices of the same receptor should be much more important than aligning the same helices of different receptors.

On the other hand, the stability of the MACC2 models and the reasonable interpretations that can be extracted from them, render these results very stimulating. However, there is no reason why the AA interactions described by MACC2 should be restricted within each individual helix.

Because of these reasons we will attempt in the near future to describe by the MACC2 transform the interactions among AA.s belonging to different helices. This procedure should turn out to be independent of alignment, taking into account only the relative position of the AA.s within the 7TM channel. We hope to report these results at the next QSAR Symposium and good, homogeneous, data sets are welcome in order to check the soundness of the new method.

## REFERENCES

1. H. Kubinyi, *3D-QSAR in Drug Design*, Vol I, ESCOM, Leiden (1993).
2. H. Kubinyi, J. Folkers and Y.C. Martin, *3D-QSAR in Drug Design*, Vols II and III, KLUVER/ESCOM, Dodrecht (1998).
3. P.J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, *J.Med.Chem.*, 28: 849 (1985).
4. R.D. Cramer, D.E. Patterson and J.D. Bunce, CoMFA. Effect of shape on binding of steroids to carried proteins, *J.Am.Chem.Soc.*, 110: 5959 (1988).
5. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, GOLPE: an advanced chemometric tool for handling 3D-QSAR problems, *Quant.Struct.Act.Relat.*, 12: 9 (1993).
6. M. Pastor, G. Cruciani and S. Clementi, Smart region definition (SRD): a new way to improve the predictive ability and interpretability of 3D-QSAR models, *J.Med.Chem.*, 40: 1455 (1997).
7. S. Hellberg, M. Sjöström, B. Skagerberg and S. Wold, Peptide Quantitative Structure-Activity Relationships: a multivariate approach, *J.Med.Chem.*, 30: 1127 (1987).
8. S. Wold, J. Jonnson, M. Sjöström, M. Sandberg and S. Rännar, DNA and peptide sequences and chemical processes multivariately modelled by PCA and PLS, *Analytica Chimica Acta*, 277: 239 (1993).
9. P. Rovero, D. Riganelli, D. Fruci, S. Viganò, S. Pegoraro, R. Revoltella, G. Greco, R. Butler, S. Clementi and N. Tanigaki, The importance of secondary anchor residue motifs of HLA class I proteins: a chemometric approach, *Molec.Immunol.*, 31: 549 (1994).
10. A.M. Davis, D.R. Flower, N. Gesmantsel, M. Clementi, G. Cruciani, M. Pastor and S. Clementi, Protein secondary structure prediction, Abstracts of 11th European Symposium on QSAR, Lusanne 1996, paper P1A.
11. S. Clementi, G. Cruciani, D. Riganelli, R. Valigi, G. Costantino, M. Baroni and S. Wold, Autocorrelation as a tool for a congruent description of molecules in 3D-QSAR studies, *Pharm.Pharmacol.Lett.*, 3: 5 (1993).
12. R. Valigi, M. Clementi, S. Clementi, G. Cruciani, M. Pastor, unpublished.
13. S. Clementi, G. Cruciani, P. Fifi, D. Riganelli, R. Valigi and G. Musumarra, A new set of Principal Properties for heteroaromatics obtained by GRID, *Quant. Struct.-Act. Relat.*, 15: 108 (1996).
14. S. Watson, S. Arkininstall, *The G-Protein Linked Receptor Facts Book*, Academic Press, London (1994).
15. J. Nilsson, H. Wikström, A. Smilde, S. Glase, T. Pugsley, G. Cruciani, M. Pastor and S. Clementi, GRID/GOLPE 3D quantitative structure-activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D<sub>3</sub> receptor subtype, *J.Med.Chem.*, 40: 833 (1997).
16. J. Nilsson PhD thesis, Multiway calibration in 3D QSAR. Applications to dopamine receptor ligands, University of Groningen, The Netherlands (1998).

**Section VII**  
**New Methods in Drug**  
**Discovery**



# SPECMAT: SPECTRA AS MOLECULAR DESCRIPTORS FOR THE PREDICTION OF BIOLOGICAL ACTIVITY

R. Bursi\* and V.J. van Geerestein  
Molecular Design & Informatics  
NV Organon  
P.O. Box 20, 5340 BH Oss, The Netherlands  
e-mail: r.buma@organon.oss.akzonobel.nl

## INTRODUCTION

A proper choice of molecular descriptors or molecular fingerprints is decisive for successful application of computational methods in QSAR and Lead Discovery.<sup>1-3</sup>

In the present study, we investigated the usage of molecular spectra as descriptors for the prediction of the biological activities of molecules. Several considerations suggested this choice: i) spectra are unique fingerprints of the chemical composition and structure of molecules; ii) spectra are observables, i.e., measurable properties of molecules; as such they are reproducible (under the same experimental conditions) and invariant to translations and rotations (in gas phase and solution); in simulations iii) no alignment of the molecular fragments are needed and iv) in comparison to many other descriptors, no calculations of charges are needed.

A program, SpecMat,<sup>4</sup> was developed which could read in spectra and transform them into matrices ready to be analysed by multivariate regression analysis techniques (PLS<sup>5,6</sup>) in SYBYL.<sup>7</sup> A QSAR validation study on a congeneric set of progestagens was performed here by means of experimentally determined Mass and <sup>1</sup>H NMR and simulated Infrared and <sup>13</sup>C NMR spectra. The results were compared with CoMFA<sup>8</sup> results.

## METHODOLOGY

A set of 45 progestagens was selected for this study.<sup>4</sup> Care was taken that sufficiently diverse structures were present in the selection. For two compounds (ORG 1002, 959) the

---

\*To whom correspondence should be addressed.

corresponding  $\alpha,\beta$  isomers (ORG L1310, OE59) were also considered. A total of 47 progestagens was thus analysed.

### SpecMat Data Conversion

Spectral data were obtained in several formats, i.e. GAUSSIAN<sup>9</sup> output files or in J-CAMP<sup>10</sup> format. In all cases conversion was performed to present the data in the format required by SYBYL. The technical details of SpecMat, simulated Infrared and <sup>13</sup>C NMR spectra, CoMFA (rigid alignment) and experimental Mass and <sup>1</sup>H NMR spectra will be given somewhere else.<sup>4</sup>

## RESULTS AND DISCUSSION

A set of 45 progestagens was investigated by means of SpecMat and CoMFA. To avoid possible correlation problems, the two  $\alpha,\beta$  isomers mentioned in the database selection section were used for predictions only. The models obtained by means of SpecMat and CoMFA on the whole set of 45 progestagens are given in Table 1.

**Table 1.** Comparison of SpecMat and CoMFA for a set of 45 progestagens

	q <sup>2</sup>	s	# comp	r <sup>2</sup>	s	F
CoMFA	0.395	0.588	4	0.865	0.278	64.21
IR SIM	0.533	0.516	4	0.987	0.086	762.71
MASS EXP	0.484	0.536	3	0.924	0.204	166.96
<sup>1</sup> H NMR EXP	0.548	0.515	5	0.969	0.135	243.60
<sup>13</sup> C NMR SIM	0.395	0.613	5	0.987	0.088	587.45

In this study, simulated (SIM) Infrared and <sup>13</sup>C NMR and experimental (EXP) Mass and <sup>1</sup>H NMR spectra were considered as well as CoMFA (standard steric and electrostatic fields). The best statistics (q<sup>2</sup> and s values) are provided by the SIM IR and the EXP <sup>1</sup>H NMR and Mass spectra. CoMFA yield statistics comparable to the SIM <sup>13</sup>C NMR spectra. In general, the different q<sup>2</sup> and s values show that every descriptor provides a different description of the structure-activity relationship of this data set.

**Table 2.** Comparison of SpecMat and CoMFA for a set of 38 progestagens (training set)

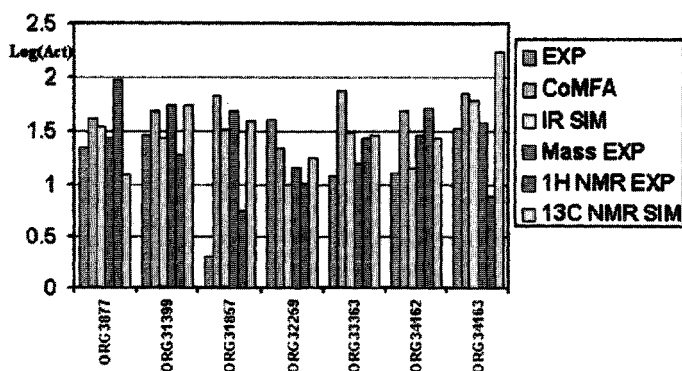
	q <sup>2</sup>	s	# comp	r <sup>2</sup>	s	F
CoMFA	0.519	0.561	4	0.914	0.238	87.6
IR SIM	0.638	0.481	3	0.986	0.096	771.9
MASS EXP	0.624	0.489	3	0.939	0.198	173.0
<sup>1</sup> H NMR EXP	0.524	0.567	5	0.973	0.134	233.2
<sup>13</sup> C NMR SIM	0.397	0.619	3	0.957	0.165	255.1

We have also assessed the quality of the model by its predictive power. For this purpose the data set was split in two subsets, a training and a test set. The criterion followed in this

separation was extracting from the data set (training set) the most diverse compounds, i.e., compounds with unique substituents and/or features and placing them in the test set. This led to a training set of 38 compounds and a test set of 7 compounds. These seven compounds are: ORG 3877, 31399, 31857, 32259, 33363, 34162 and 34163.

Statistics were repeated on the new training set and the resulting models were used to predict the activities of the test set. The results are given in Tables 2. Predictions of the test set are displayed in Figure 1.

For all descriptors the statistics obtained on the training set improved with the exception of SIM <sup>13</sup>C NMR which yield a 3 components model instead of 5, but with a smaller  $r^2$  value.



**Figure 1.** Descriptors predictions of the test set. Log(Act) is the decimal logarithm of *in vitro* binding affinities (EXP).

Concerning the test set, on the average Mass EXP, IR SIM and CoMFA produce the best predictions for all molecules except one: ORG 31857 which is badly predicted by all descriptors except EXP <sup>1</sup>H NMR.

**Table 3.** Descriptors Combinations: whole set (45 molecules)

Descr <sub>1</sub>	Descr <sub>2</sub>	q <sup>2</sup>	s	# c	r <sup>2</sup>	s
CoMFA	Mass	0.529	0.518	4	0.969	0.133
CoMFA	IR	0.511	0.518	4	0.975	0.119
CoMFA	<sup>1</sup> H NMR	0.640	0.453	4	0.974	0.122
Mass	IR	0.509	0.517	2	0.936	0.187
Mass	<sup>1</sup> H NMR	0.514	0.520	3	0.968	0.134
IR	<sup>1</sup> H NMR	0.602	0.471	3	0.970	0.129
<sup>1</sup> H NMR	<sup>13</sup> C NMR	0.598	0.479	4	0.993	0.063

Orthogonal descriptors provide better statistics when they are used in combination than when they are used individually, as long as they are capable of describing the activities under

consideration.<sup>11</sup> SpecMat and CoMFA, and spectra with spectra were therefore combined to verify how useful the combination of descriptors can be for this data set. The results for the whole set (45 compounds) are given in Table 3. The best models were obtained with the following combinations: CoMFA and <sup>1</sup>H NMR, IR and <sup>1</sup>H NMR and <sup>1</sup>H NMR and <sup>13</sup>C NMR.

For these descriptors, statistics are significantly better than the statistics of the individual models.

### Isomers Predictions

A further objective of this study was to investigate whether these descriptors could reproduce and distinguish between the activities of  $\alpha$ ,  $\beta$  isomers. Two pairs of isomers were available at the time of this study: ORG 1002 and L1310 differ by an  $\alpha$  and  $\beta$  methyl substitution on position 10; ORG 959 and OE59 by an  $\alpha$  and  $\beta$  methyl substitution on position 6. ORG 1002 and 959 are active and are part of the training set, while the corresponding isomers are inactive and constitute the test set.

As given in Table 4, ORG 1002 and 959, i.e., the isomers present in the model, were generally well predicted by all descriptors. In particular, <sup>1</sup>H NMR, IR and Mass yield the best predictions. When descriptors were combined, the best combinations were given by: CoMFA and <sup>1</sup>H NMR = IR and <sup>1</sup>H NMR > Mass and <sup>1</sup>H NMR.

**Table 4.** Isomers predictions by SpecMat, CoMFA and descriptors combinations. ORG 1002 and 959 are part of the training set (model) and ORG L1310 and OE59 are the corresponding  $\alpha$ ,  $\beta$  isomers. Log(Act) is the decimal logarithm of *in vitro* binding affinities (EXP)

	ORG 1002	ORG L1310	ORG 959	ORG OE59
Exp	0.65	-1.16	0.78	-0.41
CoMFA	0.40	0.95	0.63	0.98
IR	0.61	1.54	0.71	1.06
Mass	0.53	1.04	0.75	0.93
<sup>1</sup> H NMR	0.57	1.14	0.78	0.58
<sup>1</sup> H NMR+CoMFA	0.71	1.32	0.88	0.23
<sup>1</sup> H NMR+IR	0.60	1.32	0.92	0.57
<sup>1</sup> H NMR+Mass	0.49	0.93	1.10	0.85

As the SpecMat and CoMFA predictions of activities of isomeric compounds were quantitatively not very accurate, we looked for predictions of the active-inactive experimental trends only.

The experimental trend for ORG L1310 was not reproduced by any descriptor. This might well be due to the methyl  $\alpha$  substitution of position 10 which involves a non negligible structural change in the steroidal skeleton and which is not known by the training set. For ORG OE59 the experimental trend was well reproduced by <sup>1</sup>H NMR and the combination of <sup>1</sup>H NMR with IR, and especially with CoMFA. In this case the combination of <sup>1</sup>H NMR and CoMFA is clearly superior to any other descriptor or combination of them.

Although general conclusions cannot be reached because of the limited number of isomers considered, some trends can be identified: when combined, <sup>1</sup>H NMR and CoMFA seem to be the best descriptors to reproduce the experimental trends of isomers. The differentiation of  $\alpha$ - $\beta$  isomers in IR spectra is generally very difficult. In our simulated IR spectra, both

isomers have also identical intensity profiles, since no real intensities were calculated. The experimental Mass spectra of  $\alpha$ - $\beta$  isomers are different in intensities, but not in peaks positions. These differences, however, are clearly not sufficiently large to be picked up by PLS.

## CONCLUSIONS

The objective of this study was an initial investigation of the use of spectra as molecular descriptors for activities predictions. Experimental Mass,  $^1\text{H}$  NMR spectra and simulated  $^{13}\text{C}$  NMR and Infrared spectra were used to predict the potency of a congeneric set of 47 progestagens, among which were 2 pairs of  $\alpha$ ,  $\beta$  isomers. The analyses show good statistical correlation's of the training sets and good predictions of the test sets. Bearing in mind the limited size of the data set considered, these descriptors were found to perform at least as well as CoMFA. Descriptor's combinations and especially the combination of  $^1\text{H}$  NMR and CoMFA seem to be capable of predicting some differences between  $\alpha$ ,  $\beta$  pairs of isomers.

Obviously, larger and more diverse data sets need to be investigated for further assessment of the approach. A non-congeneric data set of estrogens is currently under investigation. SpecMat predictions are fast and at least as reliable as CoMFA predictions. Moreover, SpecMat does not require molecular superposition. Therefore, SpecMat is very suitable for activity prediction and ranking of large series of compounds. It is less obvious at this point how a SpecMat model can be translated directly into chemical suggestions for structure optimization. For now, we see a real potential for SpecMat in Lead Discovery and to a lesser extent in Lead Optimisation. Exploiting this latter direction to its fullest potential will require further and extensive investigation.

## ACKNOWLEDGEMENTS

The authors are grateful to: Dr. P. Verwer (CAOS/CAMM Centre, University of Nijmegen) for his contribution to the development of SpecMat; Dr. E. Kellenbach, T. Dao, T. v. Wijk, Dr. P. v. Hoof, and Dr. C. Funke (Analytical Chemistry for Development, Organon) for measuring  $^1\text{H}$  NMR and Mass spectra, for providing  $^{13}\text{C}$  NMR simulated spectra and for helpful discussions; Dr. J.-R. Mellema and Dr. C. Thijssen-van Zuylen (Analytical Chemistry for Research, Organon) for helpful discussions.

## REFERENCES

1. C. Hansch and A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC (1995) Vol 1.
2. H. Kubinyi. *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden (1993).
3. H. Kubinyi, G. Folkers, and Y.C. Martin, Eds., *3D QSAR in Drug Design: Recent Advances*, ESCOM (Kluwer), Leiden (1997).
4. Submitted publication.
5. P. Geladi and B.R. Kowalski, Partial least squares regression: a tutorial, *Anal. Chim. Acta*, 185:1 (1986).
6. S. Wold, E. Johansson, and M. Cocchi, Partial least squares projections to latent structures, in *3D QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi, Ed., ESCOM, Leiden (1993).
7. SYBYL, Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.
8. R.D. Cramer III, D.E. Patterson, and J.D. Bunce, Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, 110:5959 (1988).
9. GAUSSIAN 94, M.J. Frisch, G.W. Trucks, H.B. Schlegel, P.M.W. Gill, B. G. Johnson, M.A. Robb, J.R. Cheeseman, T.A. Keith, G.A. Petersson, J.A. Montgomery, K. Raghavachari, M.A. Al-Laham, V.G. Zakrzewski, J.V. Ortiz, J.B. Foresman, J. Cioslowski, B.B. Stefanov, A. Nanayakkara, M. Challacombe, C.Y. Peng, P.Y. Ayala, W. Chen, M.W. Wong, J.L. Andres, E.S. Repogle, R. Gomperts, R.L. Martin, D. J. Fox, J.S. Binkley, D.J. Defrees, J. Baker, J.P. Stewart, M. Head-Gordon, C. Gonzalez, and J.A. Pople, Gaussian, Inc., Pittsburgh PA (1995).
10. R.S. McDonald and P.A. Wilks Jr., *Appl. Spec.*, 42:151 (1988).
11. K. Esbensen, S. Schonkopf and T. Midtgaard, *Multivariate Analysis in Practice*, Wennbergs Trykkeri AS, Trondheim (1996).

## HYDROGEN BOND CONTRIBUTIONS TO PROPERTIES AND ACTIVITIES OF CHEMICALS AND DRUGS.

Oleg A. Raevsky,<sup>1</sup> Klaus J. Schaper,<sup>2</sup> Han van de Waterbeemd,<sup>3</sup> and James W. McFarland<sup>4</sup>

<sup>1</sup>Institute of Physiologically Active Compounds of Russian Academy of Sciences, 142432, Chernogolovka, Moscow region, Russia

<sup>2</sup>Borstel Research Institute, D-23845 Borstel, Germany

<sup>3</sup>Pfizer Central Research, Sandwich, CT 13 9NJ, England

<sup>4</sup>Reckon.dat consulting, Old Lyme, CT 06371

Hydrogen bonding plays an important role in many chemical and biological processes, but this interaction is complex and has been difficult to quantify in correlation analysis. One of the better ways to describe hydrogen bonding strength is to use the thermodynamic parameters of H-bond formation: free energy ( $\Delta G$ ), enthalpy ( $\Delta H$ ), entropy ( $\Delta S$ ) and H-bond binding constant ( $K$ ). These are connected to each other by the following relationships:

$$\Delta G = -RT \ln K = \Delta H - T\Delta S \quad (1)$$

It is possible to estimate the values of these parameters in the framework of a multiplicative approach based on equations (2) - (4)<sup>1,2</sup>:

$$\Delta G = k'C_a C_d + k''_0 \quad (2)$$

$$\Delta H = kE_a E_d \quad (3)$$

$$\log K = k''\alpha\beta + k''_0 \quad (4)$$

where  $C_a$  and  $C_d$  are free energy H-bond acceptor and donor factors,  $E_a$  and  $E_d$  are enthalpy H-bond acceptor and donor factors, and  $\alpha$  and  $\beta$  are H-bond donor and acceptor binding

constants. Some limitations of this multiplicative approach inspired us to construct an overall H-bond acceptor scale on the basis of equation (5)<sup>3</sup>:

$$\Sigma C_a^0 = 0.266\alpha - \log P \quad (5)$$

The utility of eqs. (2) - (5) depends on the existence of vast, readily accessible experimental thermodynamic data, and a program to estimate thermodynamic parameters for new chemical compounds. HYBOT (Hydrogen Bond Thermodynamics), described in detail by Raevsky<sup>2</sup>, is such a database and program.

On the basis of the previously noted factors, it is possible to construct QSAR descriptors for H-bonding. Table 1 summarizes information about such descriptors and identifies the computer programs that calculate and use those descriptors. These programs were created in the Department of Computer-Aided Molecular Design, Institute of Physiologically Active Compounds, Russian Academy of Sciences.

Table 1. Hydrogen bond descriptors and the programs used to generate them.

Symbol	Type	Descriptor	Program
C <sub>a</sub> max	2D	Free energy factor for the strongest H-bond acceptor atom in the molecule	HYBOT-PLUS
C <sub>d</sub> max	2D	Free energy factor for the strongest H-bond donor atom in the molecule	HYBOT-PLUS
ΣC <sub>a</sub>	2D	Sum of C <sub>a</sub> values for all H-bond acceptors in the molecule	HYBOT-PLUS, SLIPPER
ΣC <sub>d</sub>	2D	Sum of C <sub>d</sub> values for all H-bond donors in the molecule	HYBOT-PLUS
ΣC <sub>ad</sub>	2D	Sum of absolute values for C <sub>a</sub> and C <sub>d</sub> of all H-bond acceptors and donors	HYBOT-PLUS
FRG	2D	Fragments with classified H-bond factor values	MOLDIVS
HB++	3D	Interaction intensities of H-bond acceptors at (i) A	MOLTRA
HB--	3D	Interaction intensities of H-bond donors at (i) A	MOLTRA
HB++	3D	Interactions intensities of H-bond acceptor with donors at (i) A	MOLTRA
SIS++	3D	Similarity Indices of Spectra of H-bond acceptors interactions	CONFAN
SIS--	3D	Similarity Indices for the Spectra of H-bond donor interactions	CONFAN
SIS +-	3D	Similarity Indices for the Spectra of H-bond donor and acceptor interactions	CONFAN

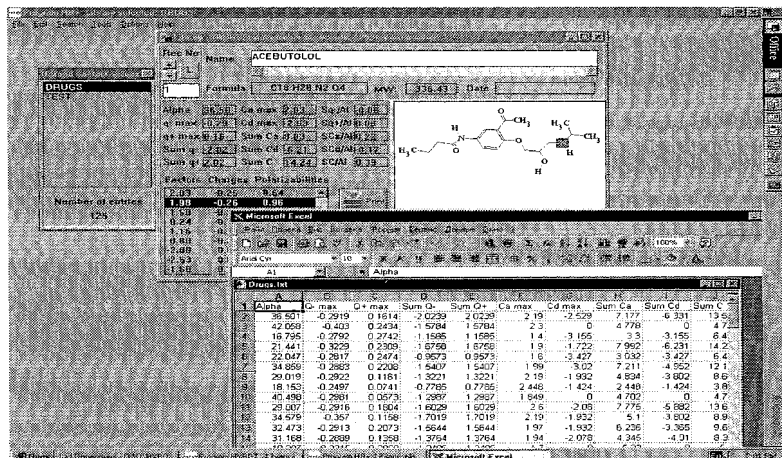


Fig 1. Results for acetubutolol calculated by HYBOT-PLUS.



H-bonds are not the only interatomic interactions; therefore, the program HYBOT-PLUS was created to calculate descriptors for steric and electrostatic forces. In all, the program calculates 15 molecular descriptors: molecular polarizability ( $\alpha$ ), maximal negative charge ( $q_{\max}^-$ ), maximal positive charge ( $q_{\max}^+$ ), sum of negative and positive charges ( $\sum q^-$  and  $\sum q^+$ ),  $C_a$ max,  $C_d$ max,  $\sum C_a$ ,  $\sum C_d$ ,  $\sum C_{ad}$ ,  $\sum q^-/\alpha$ ,  $\sum q^+/\alpha$ ,  $\sum C_a/\alpha$ ,  $\sum C_d/\alpha$  and  $\sum C_{ad}/\alpha$ . In addition it computes the polarizability, partial atomic charge and H-bond factor values for each atom in a molecule. The program uses the Structural Editor or MOL and SDF files for structural input, and Excel spreadsheets to report the results (Fig. 1).

These descriptors together with acid/base parameters are valuable for predicting chemical properties of compounds (free energy of hydration, solubility in water and other solvents, lipophilicity and permeability).

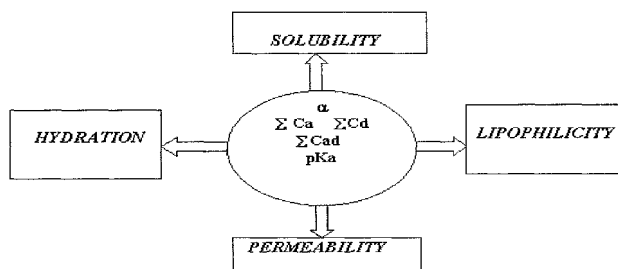


Fig 2. The scheme of relationships physico-chemical parameters with chemicals properties.

Examples of successful correlations are presented below.

#### Free energy of hydration

$$\Delta G_{\text{hyd}} = -1.05 (\pm 0.86) - 6.80 (\pm 0.55) C_a + 3.72 (\pm 0.50) C_d \quad (n=223, R=0.922, s= 3.29, F = 625)$$

(6)

#### Water solubility of liquid neutral compounds

$$\log S_w = -0.258 (\pm 0.017) \alpha + 1.08 (\pm 0.10) C_a - 0.20 (\pm 0.09) C_d \quad (n=142, R=0.953, s = 0.38, F = 452)$$

(7)

#### Lipophilicity

$$\log P_{\text{oct-water}} = 0.266 (\pm 0.030) \alpha - 1.00 (\pm 0.10) \sum C_a \quad (n=2850, R=0.970, s=0.23)$$

(8)

## Permeability

*Human red cell basal permeability (BP) of alcohols, water, urea and thiourea*<sup>4</sup>:

$$\log BP = -0.70 (\pm 0.64) + 1.08 (\pm 0.16) C_d \quad (n=10, r=0.983, s=0.43, r_{cv} = 0.976)$$

(9)

*Permeation of non-electrolytes through cells of the alga Chara ceratophylla*<sup>4</sup>:

$$\log Per = 0.83 (\pm 0.57) + 0.59 (\pm 0.12) C_d \quad (n=27, r=0.903, s=0.49, r_{cv} = 0.885)$$

(10)

*Human skin permeability coefficients (log  $k_p$ ) of phenols*<sup>4</sup>:

$$\log k_p = -8.72 (\pm 2.79) + 0.67 (\pm 0.15) C_d + 2.47 (\pm 1.28) \log MW \quad (n=17, R=0.949, s=0.20, r_{cv}=0.915)$$

(11)

*Caco-2 cell permeability of drugs*<sup>5</sup>:

$$\log k_p = -4.10 (\pm 0.57) + 0.005 (\pm 0.002) \log MW - 0.20 (\pm 0.03) C_{ad} \quad (n=17, R = 0.883)$$

(12)

*Lecithin Saposomes permeability of phenols*:

$$\log k_p = 0.78 (\pm 0.79) + 0.171 (\pm 0.034) \alpha - 0.69 (\pm 0.13) C_a - 0.15 (\pm 0.13) C_d \quad (n=26, R=0.947, s=0.22, F=63.3)$$

(13)

Because these new H-bond descriptors were successful in predicting lipophilicity and solubility, we were able to create the computer program SLIPPER (Solubility, LIPophilicity, PERmeability)<sup>6</sup>. The current version of the program calculates complete compound profiles of pKa-lipophilicity and pKa-water solubility on the basis of polarizability, H-bond factors and pKa.

In addition, these H-bond descriptors can be useful in QSAR correlations of various biological activities. For example, there is the case of tadpole narcosis (the biological data are taken from<sup>7</sup>):

$$\log (1/C) = 0.98 (\pm 0.22) + 0.221 (\pm 0.018) \alpha - 0.73 (\pm 0.08) \sum C_a \quad (n=100, R=0.932, s=0.38, F=323)$$

(14)

Further QSAR models for toxicity to *Daphnia magna* are presented in equations (15-20). The biological data are taken from<sup>8</sup>.

Common narcotic models:

$$\log 1/EC_{50} = 1.07 (\pm 0.64) + 0.23 (\pm 0.05) \alpha \quad (n = 35, r = 0.86, s = 0.34, F = 97)$$

(15)

$$\log 1/EC_{50} = 1.46 (\pm 0.61) + 0.22 (\pm 0.042) \alpha - 0.31 (\pm 0.19) C_a \quad (n=35, R=0.90, s=0.30, F=69)$$

(16)

Non-Polar narcotic models:

$$\log 1/EC_{50} = 0.91 (\pm 0.65) + 0.25 (\pm 0.05) \alpha \quad (n = 23, r = 0.92, s = 0.30, F = 117)$$

(17)

$$\log 1/EC_{50} = 1.36 (\pm 0.84) + 0.23 (\pm 0.05) \alpha - 0.57 (\pm 0.72) C_a \quad (n=23, R=0.93, s=0.29, F = 65)$$

(18)

Polar narcotic models:

$$\log 1/EC_{50} = 5.33 (\pm 0.58) - 1.01(\pm 0.038) C_a \quad (n = 12, R = 0.88, s = 0.25, F = 36)$$

(19)

$$\log 1/EC_{50} = 4.17 (\pm 1.51) + 0.07 (\pm 0.09) \alpha - 0.87 (\pm 0.38) C_a \quad (n=12, R=0.92, s= 0.23, F= 24)$$

(20)

These new physico-chemical descriptors also can be used as estimates for similarity among chemicals and diversity in databases by yet another new computer program: MOLDIVS (MOLEcular DIVersity and Similarity)<sup>9</sup>.

To construct predictive QSAR models it is necessary to take the three dimensional properties of compounds into account. In 1987 Raevsky<sup>10</sup> proposed that 3D structures could be described by the spectra of interatomic interactions. In this approach each pair of atoms in a molecule gives a line in the spectrum for any type of interaction. A line's position corresponds to the distance between the two atoms while its intensity corresponds to the product of physico-chemical parameters associated with those atoms. Atomic vibrations transform lines into bands; thus spectra of interatomic interactions are superpositions of all such bands. The computer programs MOLTRA (MOLEcular Transform Analysis) and CONFAN (CONFormation Analysis) calculate the following interaction spectra: van-der-Waals, positive charges between each other, negative charges between each other, positive charges with negative ones, H-bond acceptors between each other, H-bond donors between each other and H-bond acceptors with H-bond donors. Examples of these spectra are presented on the left side of Fig.3. In principle, each point of such spectra can be used as a 3D descriptor. For example, in a QSAR study on the inhibition of phosphorylation of polyGAT by  $\alpha$ -substituted benzylidenemalononitrile-5-S-aryltyrphostins, it was found that the interactions of two H-bond donors at the distance 7.4 Å played an important role.

$$\log 1/IC_{50} = 0.64 (\pm 1.88) + 1.74 (\pm 0.10) LUMO + 0.39 (\pm 0.10) HB-7.4 \quad (n=12, R=0.912, s=0.37, F=22.3)$$

(21)

Other valuable 3D H-bond descriptors can be estimated by quantitatively comparing the same type of spectra for all compounds in the training set. Any spectral region and all possible distances may be considered. For example, take the case of the inhibition of dihydrofolate reductase by the 15 most active 4,6-diamino-1,2-dehydro-2,2-dimethyl-1-(*p*-phenyl)-s-triazines in a particular series. Comparing each compound using the Similarity Indexes of Spectra of H-bond acceptor interactions (SIS++) and other properties establishes the following relationships:

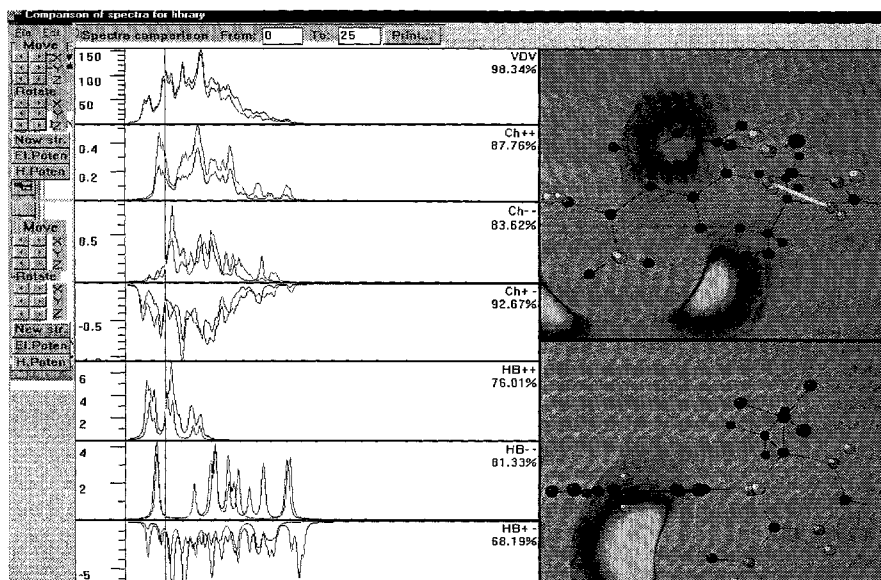


Fig.3. Spectra of interatomic interactions, Similarity Indexes of Spectra of interatomic interactions and H-bond potentials for any pair of compounds.

$$\log 1/K_i = 13.7 (\pm 2.3) - 4.33 (\pm 1.43) \log \text{SIS}++ \quad (n = 15, r = 0.849, s = 0.50, F = 42.6) \quad (22)$$

$$\log 1/K_i = 14.6 (\pm 4.1) + 0.076 (\pm 0.040) \alpha - 6.06 (\pm 2.05) \log \text{SIS}++ \quad (n=15, R = 0.940, s = 0.37, F = 45.5) \quad (23)$$

The right side of Fig. 3 shows molecular H-bond potentials for two compounds; Goodford's approach<sup>11</sup> was used. However, here we used enthalpy H-bond factor values (eq. 3) of concrete atoms of two molecules which are interacting between each other.

In summary, we have developed a method for the quantitative description of H-bonding that is founded on large databases of thermodynamic parameters and H-bond factors (the program HYBOT). Supplementing this are other programs. HYBOT-PLUS calculates H-bond descriptors, polarizabilities and partial atomic charges. SLIPPER estimates important properties as water solubility and the lipophilicity-pKa profile. Based on structural fragments, MOLDIVS affords a measure for similarities and diversities among a set of compounds. MOLTRA calculates 3D H-bond descriptors. CONFAN estimates similarities among H-bond donors and acceptors. These programs afford new and quantitative descriptors for H-bonding. Combined with two other important terms for interatomic interactions (steric and electrostatic forces) they can be used broadly in Drug Design and QSAR.

#### ACKNOWLEDGMENT

A great deal of the work briefly presented here was carried out at Department of Computer-Aided Molecular Design of Institute of Physiologically Active Compounds of Russian Academy of Sciences. It is a pleasure to acknowledge the contributions made by Drs. V.Grigor'ev, S.Trelalin, V.Gerasimenko, A.Razdolsky, E.Trepalina.

## REFERENCES

1. O.A. Raevsky, Quantification of non-covalent interactions on the basis of the thermodynamic hydrogen bond parameters, *J.Phys.Org.Chem.*, 10: 404 (1997).
2. O.A. Raevsky, Hydrogen bond strength estimation by means of the HYBOT program package, in *Computer-Assisted Lead Finding and Optimization*, H. van de Waterbeemd, B. Testa, G. Folkers, eds., Basel: Wiley-VHC, Basel ( 1997).
3. O.A. Raevsky, V.Ju. Grigor'ev, Lipophilicity estimation on the basis of polarizability and H-bond ability, *Chim.-Farm.z. (Rus.)*, 32 (1998).
4. O.A. Raevsky, K.-J. Schaper, Quantitative estimation of hydrogen bond contribution to permeability and absorption processes of some chemicals and drugs, *Eur.J.Med.Chem.*, 33: ppp (1998).
5. H. van de Waterbeemd, G .Gamenisch, G. Folkers, O.A. Raevsky, Estimation of Caco-2 cell permeability using calculated molecular descriptors, *Quant.Struct.-Act.Relat.*, 15: 480 (1996);
6. O.A. Raevsky, S.V. Trepalin, L.P. Trepalina, this volume. Slipper - a new program for water solubility, lipophilicity and permeability prediction, *This volume* .
7. C.Hansch, A.Leo, D.Hockman, Exploring QSAR, Am. Chem. Soc., Washington (1995).
8. Y.H.Zhao, M.T.D.Cronin, J.C.Dearden, Quant. Struct.-Act. Relat., Quantitative structure-activity relationships of chemicals, acting by non-polar narcosis-theoretical considerations, 17:131 (1998).
9. V.A. Gerasimenko, S.V. Trepalin, O.A. Raevsky, MOLDIVS - a new program for molecular similarity and diversity calculations, *This volume*.
10. O.A Raevsky, QSAR description of molecular structure, in *QSAR in Drug Design and Toxicology*, D. Hadzy, B. Jerman-Blazic, eds., Elsevier, Amsterdam ( 1987).
11. R.C. Wade, K.J. Clark, P.J. Goodford, Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe group with the ability to form two hydrogen bonds, *J.Med.Chem.*, 36:148 (1993).

**Section VIII**  
**Modeling of Membrane**  
**Penetration**

## PREDICTING PEPTIDE ABSORPTION

Lene H. Krarup<sup>1</sup>, Anders Berglund<sup>3</sup>, Maria Sandberg<sup>4</sup>,  
Inge Thøger Christensen<sup>2</sup>, Lars Hovgaard<sup>1</sup> and Sven Frokjaer<sup>1</sup>

<sup>1</sup>Department of Pharmaceutics

<sup>2</sup>Department of Medicinal Chemistry

Royal Danish School of Pharmacy, DK-2100 Copenhagen, Denmark

<sup>3</sup>Chemometrics Research Group, Umeå University, Sweden

<sup>4</sup>Umetri AB, Umeå, Sweden

## INTRODUCTION

An important prerequisite for a drug to be active is that it is able to reach its site of action. The preferred and most widely used route of drug administration is the oral route, and by far the most common mechanism of absorption from the gastrointestinal tract is passive diffusion through the intestinal epithelial cells. This process depends heavily on the solute's ability to diffuse through the lipophilic phospholipids of the cellular membrane. If a new drug candidate - even with optimized potency and selectivity for a target molecule - lacks this ability, it has little chance of reaching the market place. As a consequence, the optimization of absorption properties of drug candidates has become integrated in the early stages of drug discovery during recent years. The aim is to be able to predict the absorptive properties as early as possible; preferentially by calculated molecular properties as that may obviate the synthesis of poorly absorbed molecules.

The transport may be modelled as a partitioning between an aqueous and a lipidic phase. The traditionally used  $\log P_{\text{octanol}}$  value is mainly useful within homologous series, but has failed in a number of cases, e.g.  $\beta$ -blocking agents<sup>1,2</sup> and peptides<sup>3</sup>. Recently, more generalized computational methods using e.g. molecular surface properties or ab initio methods have been successful in predicting the absorption of small molecule drugs<sup>1,2,4,5,6</sup>. However, in the case of peptides or peptide-like molecules the understanding of the molecular factors governing absorption is still somewhat lacking. From studies on homologous peptide series it is known, that reducing the number of potential hydrogen bonds e.g. by methylating the backbone amide increases absorption<sup>3</sup>. The so-called desolvation energy hypothesis explains this in terms of the hydrogen bonds between the solute and the surrounding water which must be broken before the solute can pass through

the lipophilic membrane. There is also evidence to suggest that solution conformation may play a role<sup>7,8</sup>

Membrane partitioning may be described as a contribution from a cavity term, e.g. size of the solute and a polarity term including hydrogen bonding<sup>9</sup>. The overall charge is also of importance but may be accounted for by calculating the unionized fraction from the pKa value as it is generally accepted that only the unionized fraction is capable of passing the membrane. Tentatively, the cavity term is calculated as the size of the total water accessible surface area (TWASA), the polarity term as the polar water accessible area (PWASA), and pKa for the charge. In previous studies on low molecular weight molecules, the polar surface area has been shown to be inversely correlated to absorption<sup>1,2,4</sup> and blood-barrier permeation<sup>9</sup>. As also suggested by these studies, the molecular surface area varies with conformation so it may be necessary to take that into account also in the case of peptides.

So far, the understanding of peptide absorption has mainly been qualitative. Therefore, the overall aim of this study is to contribute to a more quantitative understanding of peptide absorption by a systematic study of the relationship between sequence, conformation and desolvation energy. More specifically, the aim was to test if calculated molecular properties of a peptide can be used to predict its interaction with phospholipids.

## **MATERIALS AND METHODS**

### **Design of model peptides**

It was decided to vary the amino acid sequence of 20 tetrapeptides according to a statistical design plan. All four positions were varied simultaneously in order to avoid the pitfalls of changing one factor at a time. Furthermore, the effect of methylating the backbone amide nitrogen of residue 2 and 4 was tested. Considering only the 20 naturally occurring amino acids and the combinations of the two N-methylations, 640 000 sequences exist. From this large experimental space, 20 representative peptides were selected by means of D-optimal design<sup>12</sup>, which allowed to take into account a number of structural constraints.

Each amino acid in the peptides was described by three orthogonal z-scale descriptors, z1-z3<sup>10</sup>. These descriptors are an updated version of the z-scales previously published by Hellberg et al<sup>11</sup>, and now includes five principal properties (z1-z5) for 87 natural and unnatural amino acids. The interpretation of the three first properties is the same as before, i.e. z1 describes hydrophilicity, z2 size/polarizability and z3 is interpreted as electronic effects.

A D-optimal algorithm is an exchange algorithm, which picks out experimental runs for which the determinant of the  $XX'$  matrix is maximized<sup>12</sup>. The D-optimal algorithm was repeated several times and the training set with the best combination of high G-efficiency and well-distributed parameters for polar surface area and total size was selected (Fig. 1). The G-efficiency is the efficiency of the design compared to a factorial design which by definition is 100%. The obtained G-efficiency of 46.3% is slightly below 50% which by rule of thumb is acceptable for D-optimal designs. Nevertheless, this is the best design obtainable for this strongly reduced and constrained design problem.



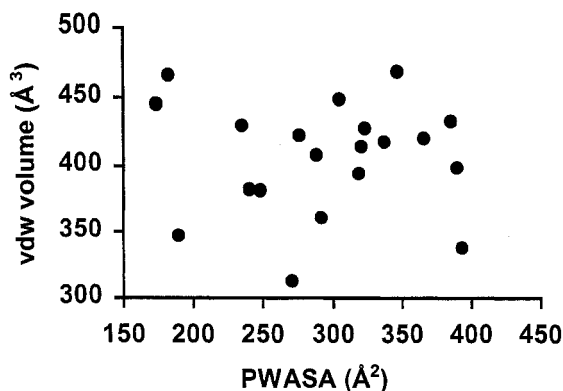


FIG. 1. Distribution of PWASA (polar water accessible surface area) and van der Waal's volume for the training set. G-efficiency for this D-optimal design was 46.3 %.

## Synthesis

The 20 peptides were synthesized by solid phase synthesis using Fmoc chemistry. Their identity was confirmed by LC-MS, PD-MS and amino acid analysis by standard procedures within Novo Nordisk.

## Peptide/phospholipid Interactions

The ability of the peptides to interact with phospholipids was studied in two chromatographic systems with phospholipids as the stationary phase. System no. 1 was a commercially available Immobilized Artificial Membrane chromatography column (IAM.PC.DD.)<sup>13</sup>, consisting of silica particles with covalently linked phosphatidylcholine. System no. 2 was the technique of Immobilized Liposome (IL) chromatography<sup>14</sup> in which liposomes are imbedded into the matrix of a chromatographic gel. The columns were mounted on a standard HPLC system and the retention time (Rt) of the peptides was recorded. The capacity factors, k'IAM and Ks were calculated as follows:

$$k'IAM = (Rt_{\text{peptide}} - Rt_{\text{citric acid}}) / Rt_{\text{citric acid}}$$

$$Ks = (Rt_{\text{peptide}} - Rt_{\text{dichromate}}) / (\text{molar amount of phospholipids})$$

## Theoretical Characterization

The polar as well as the total water accessible surface area for one extended, low-energy conformation of each peptide was calculated by means of Savol<sup>15</sup>.

## Statistical Analysis

The Savol parameters and the z-scales were used as x-variables and were supplemented with indicator variables (0/1) for the N-methylations and for presence of positive/negative charge at pH 7.4. The Y-matrix was the logarithmically transformed k'IAM and Ks values. Partial Least Squares Projection to Latent Structures (PLS) as

implemented in Simca-P® was used to correlate the x-matrix to the y-matrix. The number of significant components was determined by means of cross-validation expressed as  $Q^2$ .

## RESULTS

In table 1, the results of the statistical analysis using the various parameter sets are listed. In all cases - except model 2 employing just the z-scales and the indicator variables - quite good, low-dimensional models are obtained with  $R^2$ -values around 0.8-0.95 and  $Q^2$ -values around 0.75.

Table 1. Statistical quality of the different models

Model No	Savol	z-scales	Ind	N	$R^2$	$Q^2$
1	x		x	1	0.823	0.752
2		x	x	1	0.667	0.392
3		(Expanded)	x	2	0.911	0.723
4	x	x	x	2	0.946	0.762

Savol: PWASA, TWASA, PWASA/TWASA

z-scales: z1 (hydrophilicity), z2 (size/polarizability), z3 (electronic effects) for each amino acid

Ind: Indicator variables (0/1) for N-methylation and +/- charge

N: number of significant PLS components

$R^2$ : Explanatory value

$Q^2$ : Predictive value according to cross-validation (leave 1/7 out)

In Figure 2, the PLS coefficients for model 1 are shown. The negative effect of PWASA and fraction PWASA is equally large and it is possible to use either of them in the model with only minor decreases in the predictive power  $Q^2$ . Even though it is not possible to distinguish which is the most important factor, it can be concluded that the same factors (PWASA and fraction PWASA) previously shown to be governing

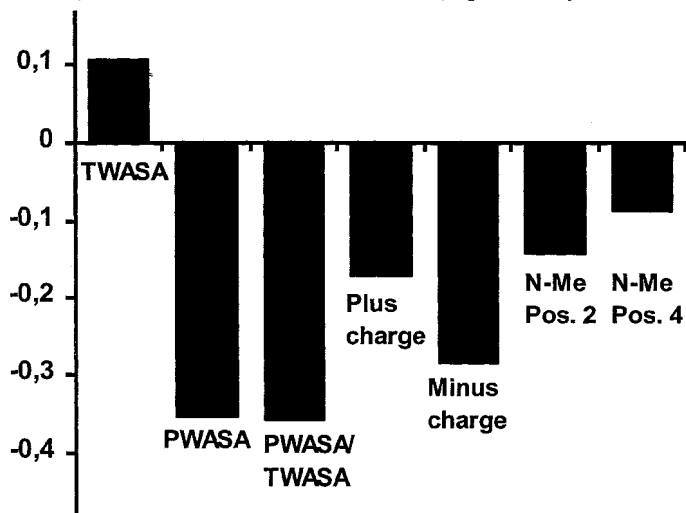


FIG.2 PLS coefficients for log  $K_s$  in model 1.  $R^2 = 0.823$ ,  $Q^2 = 0.752$ . Legend: TWASA = Total Water Accessible Surface Area, PWASA = Polar Water Accessible Surface Area, N-Me : N-methylation in position 2 or 4 as indicated. A similar coefficients plot was obtained for log  $k'IAM$ .

absorption in Caco-2 cells<sup>1,2</sup> and humans<sup>4</sup> also have major influence on the peptide-phospholipid interactions.

As would be expected, a full charge has a negative impact on phospholipid interactions. Furthermore, a negative charge has a stronger effect than a positive one. A

similar observation was made by Pauletti et al<sup>16</sup> studying IAM retention times of peptides carrying positive, negative or no charges.

The effect of N-methylation in position 2 and 4 is modest, but, surprisingly, negative. However, in this one-dimensional model, these effects are mixed up with effects of the molecular size parameters and, therefore, a detailed interpretation should be cautious. It may be speculated that a positive effect of N-methylation is only seen when the neighbouring amino acids are lipophilic and/or the overall hydrophilicity of the peptide is low. In order to further elucidate the matter, a new, extended design with room to determine interaction effects should be made.

When calculating the molecular size parameters for this model, no attempts were made to correct for the conformational variability of the parameters. The quality of the models is so good that it is doubtful if they can be improved significantly by doing so. The reason for this may be that the peptides of this study are overall too hydrophilic: no matter how they twist around they expose hydrophilic parts.

Using the z-scales plus the indicator variables gives a very weak model (model 2 in Table 1). Thus, it was necessary to improve the model by including some interaction terms and squared terms (significance based on VIP - Variable Importance for the variation in the x- and the y-matrix) to obtain the much better model 3. The necessity for the term expansion points out that the information contained in the linear combination of the z-scales describing individual amino acids is not sufficient for describing the whole-peptide properties responsible for phospholipid interactions. The validity of the suggested interdependencies between sequence and N-methylation has to be tested by a new external validation set.

If combining the Savol parameters and the z-scales (model 4) the z-scale model needs only four square terms and it is thus, intuitively, easier to interpret. From this model it is possible to estimate the effect of an amino acid exchange if in the same time calculating the Savol parameter. From the PLS coefficients for model 4 (not shown) it can be inferred that in all four positions a positive  $z_1$  value, i.e. a hydrophilic amino acid, decreases the ability to interact with phospholipids. The absolute size,  $z_2$ , of the individual amino acids is of minor importance as is  $z_3$ , the electronic effects. However, the latter seems to be of largest importance in position 4, where also a quadratic term was found to be significant.

It should be emphasized, though, that the models described here are not completely optimized. All main terms are included for illustrative purposes and we are thus running the risk of modelling "noise". Probably the predictive power of the models could be improved by excluding some of the insignificant terms.

Future absorption studies in Caco-2 cells will define the relationship between the measured lipophilicity measures and absorption which presumably is sigmoidal. Although not giving all the answers to the questions involved in peptide absorption it is a step on the way towards a more quantitative understanding of this process.

## ACKNOWLEDGEMENTS

Farideh Beigi and Per Lundahl, Uppsala University, Sweden are thanked for assistance with the Immobilized Liposome chromatography. This study was supported by a grant from Novo Nordisk A/S, Fertin A/S, Nycomed Denmark and the Danish Research Academy.

## REFERENCES

1. L.H. Krarup, I.T. Christensen, L. Hovgaard, and S. Frokjaer. Predicting drug absorption from molecular surface properties based on molecular dynamics simulations. *Pharm Res* 15:972-978 (1998).
2. K. Palm, K. Luthman, A. Ungell, G. Strandlund and P. Artursson. Correlation of drug absorption with molecular surface properties. *J Pharm Sci* 85:32-39 (1996).
3. R.A. Conradi, A.R. Hilgers, N.F.H. Ho and P.S. Burton. The influence of peptide structure on transport across Caco-2 cells. *Pharm Res* 9:435-438 (1992).
4. K. Palm, P. Stenberg, K. Luthman and P. Artursson. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm Res* 14:568-571 (1997).
5. H. van de Waterbeemd, G. Camenish, G. Folkers and O.A. Raevsky. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant Struct- Act Relat* 15:480-490 (1996).
6. U. Norinder, T. Österberg, and P. Artursson. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parameterization and PLS statistics. *Pharm Res* 14:1786-1791 (1997).
7. G.T. Knipp, D.G. Vander Velde, T.J. Siahaan and R.T. Borchardt. The effect of  $\beta$ -turn structure on the passive diffusion of peptides across Caco-2 cell monolayers. *Pharm Res* 14:1332-1340 (1997).
8. G.M. Pauletti, F.W. Okumu, S. Gangwar, T.J. Siahaan, V.J. Stella, and R.T. Borchardt. Esterase-sensitive cyclic prodrugs of peptides: evaluation of an acylalkoxy promoiety in a model hexapeptide. *Pharm Res* 13:1615-1623 (1996).
9. H. van de Waterbeemd and M. Kansy. Hydrogen-bonding capacity and brain penetration. *Chimia* 46:299-203 (1992).
10. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* 41:2481-2491 (1998).
11. S. Hellberg, M. Sjöström, B. Skagerberg, and S. Wold. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem* 30:1126-1135 (1987).
12. M. Baroni, S. Clementi, G. Cruciano, N. Kettaneh-Wold, and S. Wold. D-optimal designs i QSAR. *Quant Struct- Act Relat* 12:225-331 (1993).
13. C. Pidgeon, S. Ong, H. Liu, X. Qiu, M. Pidgeon, A.H. Dantzig, J. munroe, W.J. Hornback, J.S. Kasher, L. Glunz, and T. Szczerba. IAM chromatography: an in vitro screen for predicting drug membrane permeability. *J Med Chem* 38:590-594 (1995)
14. F. Beigi, I. Gottschalk, C.H. Hägglund, L. Haneskog, E. Brekkan, Y. Zhang, T. Österberg and P. Lundahl. Immobilized liposome and biomembrane partitioning chromatography of drugs for prediction of drug transport. *Int J Pharm* 164:129-137 (1998).
15. R.S. Pearlman<sup>(\*)</sup> and J. M. Skell, SAVOL3: Numerical and analytical algorithms for molecular surface area and volume, software distributed by the author, College of Pharmacy, University of Texas, Austin TX 78712, USA.
16. G.M. Pauletti, F.W. Okumu, R.T. Borchardt. Effect of size and charge on the passive diffusion of peptides across Caco-2 cell monolayers via the paracellular pathway. *Pharm Res* 14:164-168 (1997).

## **PHYSICOCHEMICAL HIGH THROUGHPUT SCREENING (pC-HTS): DETERMINATION OF MEMBRANE PERMEABILITY, PARTITIONING AND SOLUBILITY**

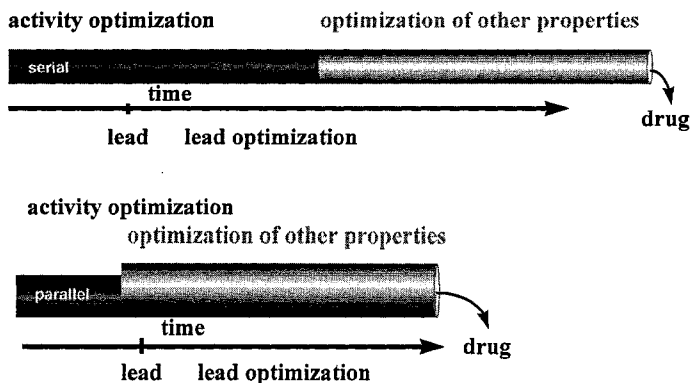
Manfred Kansy, Krystyna Kratzat, Isabelle Parrilla,  
Frank Senner, and Björn Wagner

F. Hoffmann-La Roche Ltd., Pharma Research  
Molecular Structure Research,  
4070 Basel, Switzerland

### **INTRODUCTION**

Combinatorial and parallel chemistry and genomics in combination with high-throughput screening (HTS) are capable in increasing the number of lead compounds identified in lead discovery programs. Successful application of high-throughput-technologies in biological screening demonstrates that lead identification itself is often not the time limiting step in drug development. Bottlenecks occur due to missing fast secondary assays as well as the lack of high speed and quality prediction tools. These tools might focus on many aspects of bioavailability such as absorption, protein binding, metabolic stability and toxicity. Although today screening for biological activity is fast, the entire process of lead optimisation is performed in the traditional serial way rather than in parallel (see Figure 1). Future drug discovery and development should preferably proceed with the application of parallel strategies.

It is therefore necessary to refine the existing methodologies for experimental measurements of relevant properties as well as to identify new parameters that are closely correlated with relevant aspects of *in vivo* bioavailability. In addition to experimental methods, computational approaches have to be refined to give adequate estimation of relevant compound properties so that extensive compound sets can be assessed reliably prior to synthesis.



**Figure 1.** Multidimensional optimisation strategies in lead discovery and development. Traditional versus more modern parallel approach.

## MOLECULAR PROPERTIES AND BIOAVAILABILITY

Several molecular properties are known to have major impacts on the bioavailability of a drug<sup>1</sup> ( see Table 1). Unfortunately, determinations of these properties by standard methods are time consuming and not suited for parallel HT-processing. Therefore computer programs have been developed which allow the fast calculation of relevant parameters. Computational tools can be applied prior to synthesis, so that the number of synthesised compounds can be reduced. Lipophilicity, size, solubility and ionisation constants of a molecule can be calculated by those programs. However there are limitations in the application of computational methods, due to inaccurate or incomplete parameterisation. This prevents correct calculations of important parameters such as distribution coefficients (log D) and pH dependent solubilities. Physicochemical High Throughput Screening (pC-HTS) can support the fast, standardised determination of molecular properties related to bioavailability for hundreds of compounds. pC-HTS can thus be considered to be an important factor in lead optimisation.

**Table 1.** Molecular properties with impact on bioavailability

• lipophilicity
- log P/ log D
• pKa
• solubility
- dissolution rate
• size
• hydrogen bonding
• existence of specific structural moieties relevant for
- metabolic instability
- active transport

## PHYSICOCHEMICAL HIGH THROUGHPUT SCREENING (pC-HTS) IN THE DESCRIPTION OF SPECIFIC ASPECTS BIOAVAILABILITY

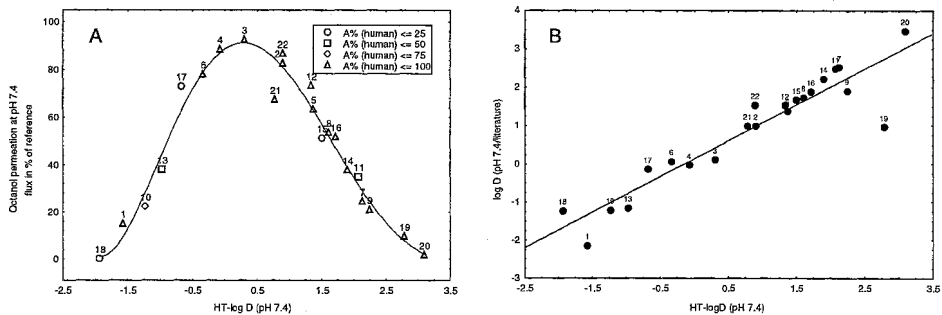
Several molecular properties have been identified which influence the absorption process of an orally administered compound. They include dissolution rate, solubility, ionisation ( $pK_a$ ), lipophilicity, hydrogen bonding and membrane permeability.

The octanol/water partition coefficient ( $\log P/D$ ) is often used as an estimate of drug permeation of barriers such as membranes. Parabolic/bilinear or hyperbolic/sigmoid relationships between permeation or absorption rates and lipophilicities have been described. Kubinyi<sup>2</sup> could show that the bilinear relationship is a useful model in the description of passive transport behaviour of congeneric compound series. Optimal permeation rates of octanol barriers are found for lipophilicities around zero. Maximum permeation of membranes is generally observed in a more lipophilic range ( $\log D$  0.5 - 3.0). Nowadays these simple rules are applied in the pharmaceutical industry in the selection process of potential leads (<http://www.glaxowellcome.co.uk/science/drugmet/chap7.html>). The capability of a compound to pass an organic layer such as octanol, should *per se* help identify compounds which have a potential to be passively absorbed. For a structurally diverse compound set we have analysed the rates of permeation through an octanol layer at pH 7.4 (Table 2/ Figure 2).

**Table 2.**  $\log D_{pH\ 7.4}$  (octanol) values determined by our HT-lipophilicity assay, performed on microtiter plates, and permeation rates in % at pH 7.4 derived by the octanol permeation assay (OPA)

No	name	HT-log D pH 7.4	OPA perm. rates	No	name	HT-log D pH 7.4	OPA perm. rates
1	salicylic acid	-1.58	15.05	12	hydrocortison	1.33	73.25
2	chloramphenicol	0.9	82.66	13	sulpiride	-0.98	38.02
3	warfarin	0.3	92.55	14	diltiazem	1.9	37.86
4	theophylline	-0.08	88.56	15	guanabenz	1.5	51.26
5	coumarin	1.37	63.27	16	corticosterone	1.71	51.64
6	metoprolol	-0.34	78.03	17	sulphasalazine	-0.68	72.95
7	imipramine	2.13	24.42	18	ceftriaxone	-1.94	0.17
8	dexamethason	1.6	53.48	19	nitrendipine	2.78	9.62
9	verapamil	2.24	20.94	20	felodipine	3.09	1.77
10	furosemid	-1.24	22.23	21	alprenolol	0.78	67.38
11	proscillaridin	2.07	35.02	22	propranolol	0.89	86.72

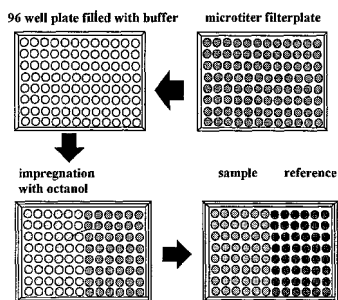
We obtained a good correlation between flux values and  $\log D$  from in house HT- $\log D$  partition coefficient measurements. Apart from the value for nitrendipine(19),<sup>6</sup> all  $\log D$  values were in accordance with known literature values (Figure 2). The prediction of human absorption data and Caco-2 permeation rates by diffusion measurements through impregnated filters has been described by several groups.<sup>7,8,9</sup> The transfer of these procedures into an HT-assay allowing the measurement of hundreds of compounds a day is simple, and is schematically described in Figure 3.



**Figure 2.** Permeation of an octanol layer at pH 7.4 as a simple HT-assay for absorption prediction for a diverse set of compounds. The permeation of a compound through the octanol layer is described by the percentage permeation (% flux). The flux values were calculated considering the UV absorption of the acceptor compartment after 15 hours and that of a reference well with same concentration containing no membrane barrier.

A: log D values determined by an in house HT-assay for lipophilicity (HT-log D) performed on a microtiter plate are depicted against results of permeation measurements through octanol. Human absorption rates (A%) as described in the literature<sup>3,5</sup> are included.

B: Comparison of log D values (literature<sup>3,4</sup>) against values derived by a HT-assay for lipophilicity determination.

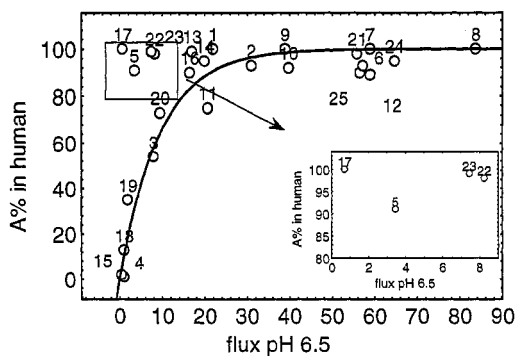


**Figure 3.** Schematic depiction of the Octanol Permeation Assay (OPA).

Due to the high solubility of water in octanol, the latter does not behave as a real barrier, so highly charged compounds can be mis-classified in their passive absorption by simple octanol permeation experiments.

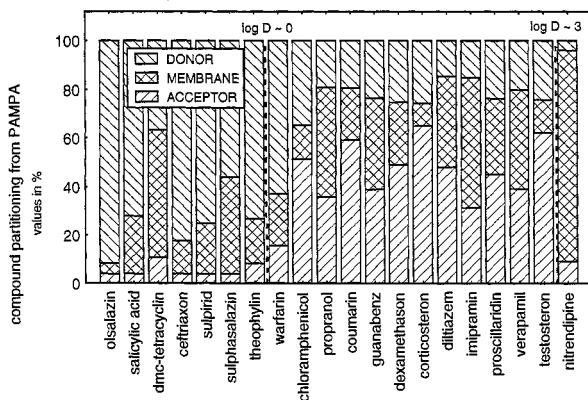
Artificial membranes allow a better determination of the permeation characteristics of such compounds. Thompson<sup>10,11</sup> and co-workers could show that stable bilayers, so called micro-BLM (Black Lipid Membranes), can be formed on a specific filter material. These membranes can be used in drug transport studies, as recently shown in our laboratory<sup>12</sup> (see Figure 4). First promising results demonstrate that such studies allow reasonable estimations of passive absorption capabilities. Passive transcellular transport is the focus of the parallel artificial membrane permeation assay (PAMPA)<sup>12</sup>. Combination of our HT-screen with cell culture models for active and paracellular transport might be advantageous in the prediction of paracellular and active transport.





**Figure 4.** PAMPA flux at pH 6.5 versus human absorption. Reproduced with the permission from the publisher<sup>12</sup>.

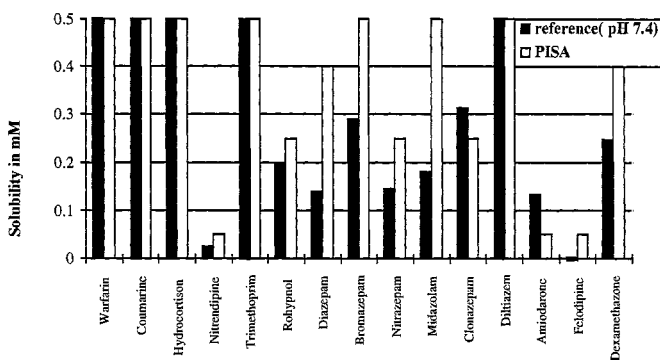
An estimate of the amount of compound which partitions into the membrane can be derived by considering the mass balance in the donor and acceptor compartments of the artificial membrane permeation assay. In Figure 5 the distribution of structurally and physicochemical diverse compounds in the different compartments of the PAMPA system is displayed. Compounds with  $\log D_{\text{octanol}}$  between 0 and 3 show increased permeation rates. More lipophilic compounds show a strong affinity to the artificial membrane (see nitrendipine).



**Figure 5.** Compartment distribution derived by the Parallel Artificial Membrane Permeation Assay (PAMPA) for a set of diverse compounds (pH 7.4, after 15 h).

Solubility and the dissolution rate also have major impacts on the *in vivo* absorption process. At first sight, the experimental determination of aqueous solubility appears to be a

straightforward task. In practice however, there are many pitfalls that should not be overlooked. The crystal form of the solute, its particle size, the ionic strength and pH of the solvent are only a few parameters that may influence the result of a solubility measurement. These can make thermodynamic solubility measurements a tedious task. In a simplification of the correct solubility measurement, turbidity determinations<sup>13</sup> can be applied as a first rough estimate of solubility. Although this method is not appropriate for the determination of high aqueous solubilities, compounds with low solubility can be easily identified. Small solubility differences can be detected, allowing the early identification and selection of compounds with a superior solubility profile (see Figure 6).



**Figure 6.** Comparison of the results determined by the Parallel Incremental Solubility Assay (PISA-turbidity measurement) versus the corresponding solubility determined by a conventional method. The bar graph for warfarin, coumarin, hydrocortison, imipramin, and diltiazem reflect the maximum solubilities achievable under the selected test conditions for the turbidity measurement (final DMSO concentration 1 %, initial concentration of the stock solution 50 mM). The corresponding solubilities determined by the standard procedure are higher, as depicted for these compounds.

The combined application of the above HT-screening technologies in the determination of molecular properties has a major impact on modern drug discovery. Significant reductions in time and costs in the development and optimisation of potential lead compounds may be realised. Beside the benefits of the application of physicochemical HT-screening for current discovery programs, a further advantage lies in the fast generation of large standardised datasets. In the long-term, data bases constructed from such datasets are a precondition for the development and improvement of high quality prediction tools. These databases will, in turn, provide the basis for enhanced virtual screening.

## ACKNOWLEDGMENTS

We would like to acknowledge Drs. Daniel Bur and David Banner for the critical review of the manuscript.

## REFERENCES

1. M. Kansy, Molecular properties, in: Structure-Property Correlations in Drug Research, H. van de Waterbeemd, ed., R.G. Landes Company, Austin (1996).
2. H. Kubinyi, Lipophilicity and biological activity. *Arzneim.-Forsch./Drug Res.*, 29(II): 1067-1080 (1979).
3. P. Artursson, J. Karlsson, Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (CACO-2) cells, *Bioch. Biophys. Res. Com.* 175:880-885 (1991).
4. MedChem97 database. Daylight Chemical Information Systems, Inc. 27401 Los Altos, Mission Viejo, CA 92691 USA
5. J.G. Hardman, L.E. Limbird, P.B. Molinoff, R.W. Ruddon, A. Goodman Gilman, *The Pharmacological Basis of Therapeutics*, MacGraw-Hill, New York (1995).
6. D.C. Pang, N. Sperelakis, Uptake of Calcium Antagonistic drugs into muscles as related to their lipid solubilities, *Biochem. Pharmacol.*, 33:821 (1984).
7. T. Fujita, J. Iwasa, C. Hansch, A new substituent constant,  $\pi$ , derived from partition coefficients, *J. Am. Chem. Soc.*, 86:5175-5180 (1964).
8. G. Camenish, G. Folkers, H. van de Waterbeemd, Comparison of passive drug transport through Caco-2 cells and artificial membranes, *Int. J. Pharm.* 147:61-70(1997).
9. R.N. Smith, C. Hansch, M.M. Ames, Selection of a reference partitioning system for drug design work, *J. Pharm. Sci.* 64:599-606 (1975).
10. M. Thompson, U.J. Krull, P.J. Worsfold, The structure and electrochemical properties of a polymer-supported lipid biosensor. *Anal. Chim. Acta*, 117:133-145 (1980).
11. M. Thompson; R.B. Lennox, R.A. McClelland, Structure and Electrochemical Properties of Microfiltration Filter-Lipid Membrane Systems. *Anal. Chem.* 54:76-81 (1982).
12. M. Kansy, F. Senner, K. Gubernator, Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes, *J. Med. Chem.* 41: 1007-1010 (1998).
13. P. Li, R. Vishnubajjala, S.E. Tabibi, S.H. Yalkowsky, Evaluation of in vitro precipitation methods, *J. Pharm. Sci.* 87: 196-198 (1998).

## UNDERSTANDING AND ESTIMATING MEMBRANE/WATER PARTITION COEFFICIENTS: APPROACHES TO DERIVE QUANTITATIVE STRUCTURE PROPERTY RELATIONSHIPS

Wouter H. J. Vaes,<sup>1</sup> Eñaut Urrestarazu Ramos,<sup>1</sup> Henk J. M. Verhaar,<sup>1</sup> Christopher J. Cramer,<sup>2</sup> and Joop L. M. Hermens<sup>1</sup>

<sup>1</sup> Research Institute of Toxicology (RITOX)  
Utrecht University  
P.O. Box 80176  
3508 TD, Utrecht  
The Netherlands

<sup>2</sup> Department of Chemistry and Supercomputer Institute  
University of Minnesota  
207 Pleasant Street SE, Minneapolis  
Minnesota 55455-0431

### INTRODUCTION

Since Meyer (1899) and Overton (1901) discovered a relationship between general anesthetic potency and oil/water partition coefficients, partition coefficients have been used as parameters in pharmacological and toxicological studies. The partition coefficient between *n*-octanol and water ( $\log K_{ow}$ ) has been used most often, occasionally referring to it as a surrogate for membrane/water partitioning.

Nevertheless, recently we showed that the assumption that  $\log K_{ow}$  is a good surrogate for membrane/water partitioning should be used cautiously (Vaes *et al.*, 1997). Therefore, the aim of this contribution is twofold: -to gain insight in the interactions that are involved in membrane/water partitioning, -and to obtain information about the differences between *n*-octanol/water and membrane/water partitioning. For this purpose we developed structure-property relationships for a set of twenty-eight chemicals.

### METHODS

This paper describes the methodologies that were developed to measure membrane/water partition coefficients. Subsequently, models are described that give insight

in the partition behavior, but also enable the prediction (or estimation) of membrane/water partition coefficients.

### Measurement of Membrane/Water Partition Coefficients

Membrane/water partition coefficients were determined by a negligible depletion extraction method using solid phase microextraction (SPME). SPME was introduced by Arthur and Pawliszyn (1990) and uses polymer coated fibers as extraction device. In short, the SPME apparatus is placed in an aqueous sample, and the organic molecules partition from the aqueous phase to the polymeric coating. Subsequently, the polymer coated fiber is transferred from the solution to the injector of a gas chromatograph. In the injector, the chemicals are thermally desorbed from the fiber, after which they can be analyzed. SPME was used as a negligible depletion extraction, which means that the extracted amount from the sample is negligibly small to not disturb equilibria between the aqueous phase and, in this case, phospholipid vesicles (Vaes *et al.*, 1997). Negligible depletion SPME was used to measure the freely available concentration of the compound of interest in a solution with and without phospholipid vesicles, from which the partition coefficients was determined.

### Calculation of Parameters to Develop Quantitative Structure-Property Relationships

Three approaches were chosen to develop QSPRs for membrane/water partitioning. First, a model was derived based on only calculated descriptors. Second, a model was developed to correct the octanol/water partition coefficient to obtain the membrane/water partition coefficients, based on quantum chemical descriptors. Last, structural fragment values were derived with the same purpose.

**Model I.** Jin and Hopfinger (1996) described that membranes could be considered as consisting of three different phases. The interior of the membrane resembles alkane-solvents, close to the interface there is a region with high rigidity, while at the interface hydrogen bonding properties dominate. For these three regions, specific molecular descriptors were derived. The hexadecane solvation energy ( $\Delta G_{S,C16}$ ) for the interior, the molecular volume (MV) for the rigid part, and hydrogen bonding parameters, as described by Cramer *et al.* (1993) ( $Q$ ,  $Q^+$ ,  $\epsilon_{HOMO}$ ,  $\epsilon_{LUMO}$ ), are compared to the aqueous solvation energy ( $\Delta G_{S,aq}$ ) by a PLS regression analyses, where membrane/water partition coefficients are used as the dependent variable.

**Model II.** Hydrogen bonding parameters, described in the earlier section, were chosen to correct  $\log K_{ow}$ 's to obtain  $\log K_{mw}$ 's.  $\log K_{ow}$ ,  $Q$ ,  $Q^+$ ,  $\epsilon_{HOMO}$ ,  $\epsilon_{LUMO}$  were used as independent, and  $\log K_{mw}$  as dependent variables in this PLS regression.

**Model III.** Structural fragments were chosen from the molecular structures, taking all polar fragments. Chosen structural fragments were alcohol (alOH), phenol (arOH), aliphatic amine (alNH<sub>3</sub><sup>+</sup>), aromatic amine (arNH<sub>2</sub>), aromatic nitro-group (arNO<sub>2</sub>), aliphatic ester (alC(=O)O) and aromatic ester (arC(=O)O). Structural fragment values were entered in the data matrix as dummy variables counting the occurrence of each structural fragment in each molecule. The fragment values were derived using PLS regression on the dummy variables according to the following Free-Wilson analysis:

$$\log K_{mw} - \log K_{ow} = \sum_j b_j X_j$$

where  $b_j$  is the value of structural fragment  $j$ , where  $j$  runs over all structural fragments that occur  $X$  times in the molecule.

## RESULTS

For the models, and a more complete discussion, see Vaes *et al.* (1998).

**Model I.** ( $R^2 = 0.87$ ,  $Q^2 = 0.81$ ) The autoscaled pseudo regression coefficient shows that  $\Delta G_{S,C16}$  is the dominant factor in describing the membrane solvation energy and therefore the largest portion of partitioning is governed by the hydrophobic regions (mostly the inner zone and probably partly the intermediate zone). A negative coefficient for  $\epsilon_{LUMO}$  shows that good electron accepting capabilities of the solute (low  $\epsilon_{LUMO}$ ) increases the membrane-water partition coefficient. Additionally, the positive coefficient for  $Q^+$  indicates that good proton donating capabilities (high  $Q^+$ ) also interact favorably with the membrane, thereby increasing  $\log K_{mw}$ . The negatively charged groups in the phospholipid, i.e. the carbonyl-groups and the phosphorous group, thus interact favorable with acidic protons of the solute. On the other hand, a highly negative  $Q^-$  results in repulsions due to the highly negative charged groups in phospholipids, and decrease the membrane-water partition coefficient. Concluding, high membrane-water partition coefficients, result from hydrophobic chemicals (low  $\Delta G_{S,C16}$  and high  $\Delta G_{S,aq}$ ) with good hydrogen bond donating capabilities (low  $\epsilon_{LUMO}$ , high  $Q^+$ ) and low hydrogen bond accepting capabilities (low absolute value of  $Q^-$ ).

**Model II.** ( $R^2 = 0.97$ ,  $Q^2 = 0.94$ ) The model shows very clearly that the differences between *n*-octanol and phospholipid membranes are governed by hydrogen bonding interactions. A low  $\epsilon_{LUMO}$  and a high  $Q^+$  show a more favorable interaction with phospholipids than with *n*-octanol, while  $Q^-$  interacts weakly and unfavorably with phospholipids. This implies that phospholipids are better electron donors than *n*-octanol, while the opposite is valid for the electron accepting capabilities. *n*-Octanol does carry an acidic proton which might interact favorably with negative groups on the solute. Since phospholipids do not have any acidic hydrogen that can be shared by electron donation of a solute, a positive sign of the coefficient of  $Q^-$  makes sense.

**Model III.** ( $R^2 = 0.91$ ) Results from this model are in accordance with model II. Some examples: -phenols are good hydrogen bond donors and thus have a positive contribution to  $\log K_{mw}$ , -esters have low hydrogen bond donor, but good accepting capabilities, therefore they have a negative contribution to  $\log K_{mw}$ .

## CONCLUSIONS

The partitioning behavior of organic chemicals to phospholipids can be modeled using physico-chemical and quantum-chemical descriptors that account for hydrophobicity as well as hydrogen bonding capabilities. Differences between the *n*-octanol-water and membrane-water partition coefficients can be almost completely explained by differences in hydrogen bonding capabilities of the solvents. The influence of one being a bulk phase and the other being a highly organized bilayer seems to be of minor importance. In addition, this study provides structural fragment values for adjusting  $\log K_{ow}$  to obtain  $\log K_{mw}$  for phenol, aniline, nitrobenzene, alcohol, amine and ester groups.

## ACKNOWLEDGEMENTS

The financial support of the Dutch Ministry of Housing, Spatial Planning and Environment, project number 94230302, and the Basque Government, Department of Education, Universities and Investigation, is gratefully acknowledged. This work was partially carried out within the framework of the EC project Fate and Activity Modeling of Environmental Pollutants Using Structure-Activity Relationships (FAME) under contract ENV4-CT96-0221.

## REFERENCES

- Arthur, C. L., and Pawliszyn, J., 1990, Solid phase microextraction with thermal desorption using fused silica optical fibers. *Anal. Chem.* 62:2145-2148.
- Cramer, C. J., Famini, G. R., and Lowrey, A. H., 1993, Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationship. *Acc. Chem. Res.* 26:599-605.
- Jin, B., and Hopfinger, A. J., 1996, Characterization of lipid membrane dynamics by simulation: 3. Probing molecular transport across the phospholipid bilayer. *Pharmaceut. Res.* 13:1786-1794.
- Meyer, H., 1899, Zur Theorie der Alkoholnarkose, welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung. *Arch. Exp. Pathol. Pharmacol.* 42:109-118.
- Overton, E., 1901, Studien über die Narkose Zugleich ein Beitrag zur Allgemeine Pharmakologie, Jens, Verlag von Gustav Fischer, Germany.
- Vaes, W. H. J., Urrestarazu Ramos, E., Hamwijk, C., Van Holsteijn, I., Blaauboer, B. J., Seinen, W., Verhaar, H. J. M., and Hermens, J. L. M., 1997, Solid phase microextraction as a tool to determine membrane/water partition coefficients and bioavailable concentrations in *in vitro* systems. *Chem. Res. Toxicol.* 10:1067-1072.
- Vaes, W. H. J., Urrestarazu Ramos, E., Verhaar, H. J. M., Cramer, C. J., and Hermens, J. L. M., 1998, Understanding and estimating membrane/water partition coefficients: Approaches to derive quantitative structure property relationships (QSPR). *Chem. Res. Toxicol.* 11:847-854.

## PREDICTION OF HUMAN INTESTINAL ABSORPTION OF DRUG COMPOUNDS FROM MOLECULAR STRUCTURE

M. D. Wessel,<sup>1</sup> P. C. Jurs,<sup>2</sup> J. W. Tolan,<sup>3</sup> and S. M. Muskal<sup>3</sup>

<sup>1</sup> Pfizer Central Research, Eastern Point Road, Groton, CT 06340

<sup>2</sup> Chemistry Department, Penn State University, University Park, PA 16802

<sup>3</sup> Affymax Research Institute, 3410 Central Expy., Santa Clara, CA 95051

### INTRODUCTION

Prediction of human intestinal absorption (HIA) is a major goal in the development of oral drugs. The application of combinatorial chemistry methods to drug discovery has dramatically increased the demand for rapid and efficient models for estimating HIA and other biopharmaceutical properties. While experimental methods for measurement of intestinal absorption have been developed and are used widely, computational approaches provide an attractive alternative.

Quantitative structure-property relationship (QSPR) methods have been used to model many chemical and biological properties of organic compounds. Computational models have also been reported for such biopharmaceutical properties as %HIA, blood-brain barrier, skin and ocular permeation, pharmacokinetics, and metabolism.<sup>1,2,3</sup> However, these studies all involved sets of structural analogs, and models based on limited chemical space generally lack predictive value outside their structural classes. Broadly applicable QSPR models of biopharmaceutical properties must be built using compounds which cover both a wide range of the property being modeled as well as of chemical structure space.

The QSPR methodology used in this study consists of three main parts: representation of molecular structure, feature selection, and mapping. The QSPR relationship is developed from a set of compounds with known %HIA values. The compounds are encoded with calculated structural descriptors, which are mathematical representations of chemical structure. Once the structures have been encoded, the subset of descriptors that best encodes the property of interest is sought with feature selection methods employing the genetic algorithm (GA)<sup>4</sup> coupled with computational neural networks.<sup>5</sup> Feature selection is

---

Excerpted with permission from *J. Chem. Inf. Comput. Sci.* 1998, 38, 726-735. Copyright 1998 American Chemical Society



a necessary step because of the large numbers of descriptors available (more than one hundred per compound). Once a subset of descriptors is found, the descriptors are then mapped to the property of interest, using either a linear regression equation or a non-linear computational neural network. These mapping methods effectively provide a mechanism for linking the chemical structures to their corresponding %HIA values.

## EXPERIMENTAL

The computations for this work, with two exceptions, were performed at Penn State University on a DEC 3000 AXP Model 500 workstation. Those calculations involving HyperChem,<sup>6</sup> were performed on a Pentium PC. The 3-dimensional model-building, utilizing CORINA (Version 1.6)<sup>7</sup> as well as the molecular fragment extractions and presence/absence determinations, were performed on a SGI Challenge-L at Affymax Research Institute. The ADAPT software system was used for all computations except those discussed above and those involving computational neural networks. The neural network software was developed independently at Penn State University.

### Data Set

The set of 86 drugs and drug-like compounds and their experimental %HIA values were taken from literature sources. These compounds are listed in Table 1 of reference 8 with their experimental %HIA values and literature references. Much of the literature uses the term "percent absorbed" imprecisely to mean either percent intestinal absorption (%HIA) or absolute oral bioavailability, which can be lower than %HIA due to first-pass hepatic metabolism. Therefore, each reference was reviewed to ensure that intestinal absorption values were used in this modeling effort, and furthermore that these values were not dose-dependent and involved only healthy clinical populations. The subset of 64 compounds with %HIA less than 100% comprise all the literature examples we were able to find which met these criteria. The remaining 22 compounds, with 100% HIA, were randomly selected from the large number of publications on well-absorbed oral drugs. The proportion of 100% HIA compounds was kept low to lessen overloading the training set with high %HIA values. The data set contains a large amount of structural diversity. Of the 86 compounds, 22 absorb at 100%, 47 have absorption values at 90% or higher, and 71 compounds (or about 83% of the total data set) absorb at 50% or higher. Only 15 absorb below 50%. While the entire range spanned is 0-100%, this data set is heavily biased towards large values of absorption given the tendency towards successfully-developed orally active drug compounds.

The 86 compounds were divided into a training set of 76 compounds and an external prediction set of 10 compounds. The external prediction set spans the range of 5% to 100% HIA. The compounds in the external prediction set were never used during the model development process but were reserved to validate potential models.

The structures of the 86 compounds were extracted from the ARI database with ISIS<sup>9</sup> and transferred to the DEC Alpha workstation where they were entered into ADAPT. CORINA was used to generate accurate 3-dimensional geometries.

### Descriptor Generation and Analysis

A total of 162 descriptors was generated for each of the 86 compounds using ADAPT descriptor development routines. The descriptors fall into three general catego-

ries: topological, electronic, and geometric. Topological descriptors are derived from information about the two-dimensional structure of the molecule. Graph theory was applied to the 2-D structures to generate a multitude of topological indices. Other topological descriptors, such as atom counts, bond counts, and molecular weight, were also derived from the 2-D structural representations. Electronic descriptors were calculated with MOPAC using the AM1 Hamiltonian. Electronic descriptors include partial atomic charges and the dipole moment. Geometric descriptors, including moments of inertia, surface area, and volume, are derived from three-dimensional geometries of the molecules. An additional class of descriptors can be derived by combining electronic and geometric information to form hybrid descriptors. By combining the molecular surface area with partial atomic charges, charged-partial surface area (CPSA) descriptors can be calculated. The same can be done with certain atom types (H, N, O, F, S) to calculate hydrogen bonding specific descriptors. Of the 162 ADAPT descriptors calculated, 84 were topological, six were electronic, 29 were geometric, and 43 were hybrid descriptors.

In addition, a large number of substructure fragment descriptors were also generated. These fragment descriptors were binary strings that indicated the presence or absence of 566 important substructure features or fragments. A pool of more than 3200 functional group fragments was excised from over 7000 drug and drug-like molecules found in MDL's Comprehensive Medicinal Chemistry database (CMC 97.1) using the first-order functional group extraction algorithm developed by Sello.<sup>10</sup> These 3200 basis-set functional groups were made more general by changing all single bonds to single or aromatic bonds and all double bonds to double or aromatic bonds, respectively. A total of 566 fragments from the basis-set pool was found in at least one, but not all, of the 86 compounds in the working set. The fragment descriptors were augmented to the pool of ADAPT descriptors. Thus, each compound was represented by 728 descriptors in all.

Objective feature selection was used to discard descriptors which contained redundant or minimal information, reducing the pool to 127 members. Of these, there were 61 fragment descriptors, 25 topological descriptors, 21 CPSA/H-Bonding descriptors, and 20 geometric descriptors.

## Computational Neural Networks

The neural network type most used for quantitative structure-property relationships is the three-layer, fully-connected, feed-forward neural network. This network consists of a multi-layer system of neurons, with each neuron in a given layer fully connected to all neurons in the two adjacent levels. The objective of a neural network is to map a set of input data to a particular set of output data. In this case, molecular structure descriptors, linearly scaled to the range (0,1), serve as input data, and the %HIA values serve as output data. The connections between neurons are known as weights. A neural network is trained to map a set of input data to a corresponding set of output data by iterative adjustment of the weights. In this study, the networks were trained using a quasi-Newton optimization algorithm. Detailed discussions of the type of neural network and the training algorithm used in this study have been published previously.<sup>5</sup>

A feature selection routine which combines the genetic algorithm with a neural network fitness evaluator<sup>11</sup> was used for this study. The GA/neural network routine selects subsets from the reduced descriptor pool that support good non-linear models by using neural networks to evaluate each potential subset. The genetic algorithm uses the rms error to find a good subset of descriptors. This forces the algorithm to find descriptor subsets that minimize the number of large outliers, at the possible expense of overall model quality. The genetic algorithm used in this study also incorporates the PRESS sta-

tistic to improve the chances of finding a general and predictive model. In any optimization procedure similar to the one described here, the starting conditions can greatly influence the final results. This is largely due to the multivariate nature of the problem. Therefore, it should not be surprising that the "best" subset of descriptors found by the GA will differ from run to run. It is also fully expected that as more and more compounds are added to the training set, the GA will find different, but perhaps qualitatively overlapping, subsets of descriptors.

## RESULTS AND DISCUSSION

The 127-member reduced pool of descriptors was fed to the GA/neural network feature selection routine for the purpose of developing a non-linear model. The original regression training set was split randomly into a neural network training set of 67 compounds and a cross-validation (CV) set of nine compounds. The original 10-member external prediction set was used to validate any neural network models. The CV set was used to monitor overtraining of the network, and the training set was used to actually train the network. The CV set and training set rms errors are used by the GA to determine a cost function that relates directly to the overall quality of a particular subset.

To decrease the possibility of chance effects influencing neural network training, the ratio of observations to total adjustable parameters should be at or above 2.0.<sup>12</sup> A neural network consisting of six input neurons (descriptors), four hidden neurons, and one output neuron (target, %HIA), a 6-4-1 architecture, was used since it produced the maximum number of adjustable parameters recommended for a data set of this size. For this 6-4-1 architecture, the ratio of training set observations to adjustable parameters was 67/33, or 2.03.

Using this 6-4-1 network architecture, the GA routine searched the reduced descriptor pool for subsets that supported good models. Several models with good cost functions were found by the GA routine. The best subset of descriptors was then studied and further optimized for network performance.

After the genetic algorithm runs were completed, several sets of weights and biases were then found in separate CNN trainings. The set that produced the best training set and cross-validation set errors was then validated with the external prediction set. The six descriptors that comprised the best subset found by the GA are as follows: NSB, the number of single bonds; SHDW-6, the normalized 2-D projection of the molecule on the YZ plane; CHDH-1, the charge on donatable hydrogen atoms; SAAA-2, surface area of hydrogen bond acceptor atoms divided by the number of hydrogen bond acceptor atoms; SCAA-2, surface area multiplied by the charge of hydrogen bond acceptors divided by the number of hydrogen bond acceptors; GRAV-3, the cube root of the gravitational index. Of the six descriptors, one is a topological descriptor, three are hydrogen bonding descriptors, and two are geometric descriptors. Pairwise correlations were calculated for the six descriptors, and the mean value is 0.21 and the highest correlation coefficient between any two of these six descriptors is 0.63. The six descriptors span the following ranges: NSB (3 to 35), SHDW-6 (0.36 to 0.76), CHDH-1 (0.00 to 1.30), SAAA-2 (3.91 to 38.23), SCAA-2 (-0.28 to -18.38), GRAV-3 (8.76 to 15.75). Of the six descriptors in the final model, none were binary fragment descriptors.

The descriptors in this model do not encode a causal relationship between structure and %HIA. However, it is useful to examine qualitatively the possible meaning of each descriptor. The NSB descriptor is encoding single bonds, and this may be an indication of the amount of structural flexibility. The SHDW-6 and GRAV-3 descriptors are

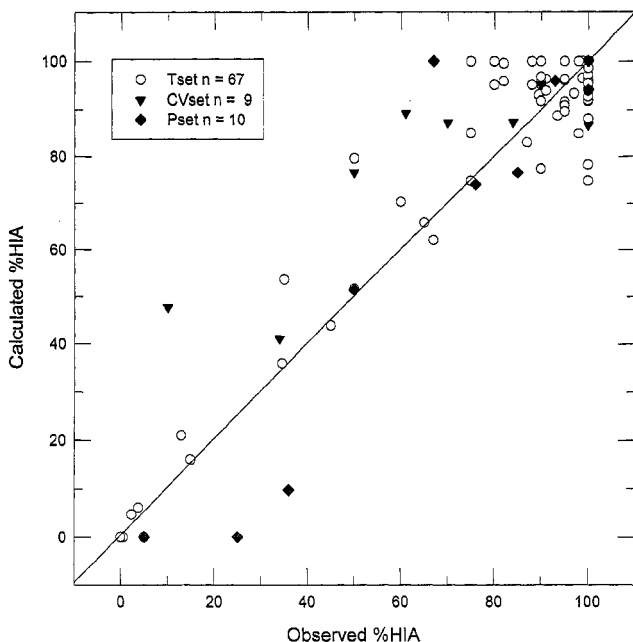


Figure 1. Plot of calculated percent human intestinal absorption (cHIA) versus observed %HIA for the training set, cross-validation set, and prediction set compounds. [Adapted from Figures 1 and 2 of *J. Chem. Inf. Comput. Sci.* 38:726 (1998) with permission of the American Chemical Society.]

encoding molecular size, shape and bulk properties. These size descriptors may be important with respect to the ability of the drug to penetrate cell membranes. The three remaining descriptors are all hydrogen bonding descriptors. These can be thought of as indicators of the degree of the lipophobic and lipophilic character of a drug compound in biological environments.

The training set rms error for this six-descriptor neural network model was 9.4 %HIA units. The mean absolute error (mae) was 6.7 %HIA units. These values were calculated after all output values from the network greater than 100% or less than 0% were fixed at 100% or 0%, respectively. The CV set rms error was 19.7 %HIA units (mae 15.4 %HIA units). Figure 1 shows a plot of cHIA vs. observed %HIA for the training and cross-validation sets. There is a good fit to the 1:1 correlation line. Validation of the model was performed using the 10-compound external prediction set. The rms error for the external prediction set was 16.0 %HIA units (mae 11.0 %HIA units), a good validation of the model. Figure 1 also shows a plot of cHIA vs. observed %HIA. It is likely that the overprediction of absorption values above 50% is mainly due to the original bias in the training set.

Chance correlations can influence the development of QSPRs. To ensure that chance effects did not influence the current study, a randomized test was performed. The dependent variables of each of the compounds in the training set and cross-validation sets were scrambled randomly and the GA was run again. The prediction set rms error from the randomized model was 41.7 %HIA units, as opposed to 16.0 %HIA units from the

real model. The cost function for the scrambled data is 50% higher than that for the real data, which indicates that the model built from the real data was not based on chance.

Intestinal absorption of drug compounds depends on complex biological processes (including passive membrane penetration, active transport mechanisms and metabolism in the gastrointestinal tract) and on compound physicochemical properties (including solubility, dissolution rate and dissociation constants). Therefore, we do not expect that a QSPR model derived from 76 diverse compounds will be a highly precise and rugged predictive tool. A much larger training set, presently unavailable in the published literature, would be required to build a model based not only on structural diversity but also on diverse biological and physicochemical properties. This model is intended to serve as a tool for both individual and compound library design to significantly improve the likelihood of overall increased %HIA of compounds selected for synthesis. As shown in Figure 1, this model does not produce an exact rank ordering, but it clearly differentiates the well-absorbed compounds from the poorly-absorbed ones.

## CONCLUSIONS

A six-descriptor non-linear computational neural network model has been developed for the estimation of %HIA values for a data set of 86 drug and drug-like compounds. The six descriptors in the final model are varied measures of structure. The training set rms error was 9.4 %HIA units, and the CV set rms error was 19.7 %HIA units. Based on the rms errors of the training and CV sets, it is clear that a link between structure and %HIA does exist. However, the strength of that link is best measured by the quality of the external prediction set. With an rms error of 16.0 %HIA units, and a good visual plot, the external prediction set assures the quality of the model. Given the structural diversity and bias of the data set, this is a good first attempt at modeling human intestinal absorption using QSPR methods.

A basic QSPR for estimation of %HIA values of drug and drug-like compounds is presented in this paper. The model can be used as a potential virtual screen, or property estimator. With a larger data supply less biased towards the high end values of %HIA, a more successful model could likely be developed. This study illustrates the potential of using QSPR methods to aid in the drug development process.

## REFERENCES

1. M.H. Abraham, H.S. Chadha, R.C. Mitchell, Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain, *J. Pharm. Sci.* 83:1257 (1994).
2. S.C. Basak, B.D. Gute, E.R. Drewes, Predicting blood-brain transport of drugs: a computational approach. *Pharm. Res.*, 13:775 (1996).
3. R.O. Potts and R.H. Guy, A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity, *Pharm. Res.* 12:1628 (1995).
4. C.B. Lucasius and G. Kateman, Understanding and using genetic algorithms part 1. concepts, properties and context, *Chemom. Intell. Lab. Sys.* 19:1 (1993).
5. L. Xu, J.W. Ball, S.L. Dixon, P.C. Jurs, Quantitative structure-activity relationships for toxicity of phenols using regression analysis and computational neural networks, *Environ. Toxicol. Chem.* 13:841 (1994).
6. Hypercube, Inc. 1115 NW 4th Street, Gainesville, FL 32601-4256.
7. J. Sadowski and J. Gasteiger, From atoms to bonds to three-dimensional atomic coordinates: automatic model builders, *Chem. Rev.* 93:2567 (1993).

8. M. D. Wessel, P. C. Jurs, J. W. Tolan, S. M. Muskal, Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure, *Jour. Chem. Inf. Comput. Sci.* 38:726 (1998).
9. MDL Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577
10. G.A. Sello, New definition of functional groups and a general procedure for their identification in organic structures, *J. Am. Chem. Soc.* 114:3306 (1992).
11. *Computer-Assisted Development of Quantitative Structure-Property Relationships and Design of Feature Selection Routines*; Wessel, M. D.; Ph. D. Dissertation, The Pennsylvania State University, 1997; Chapter 3.
12. D.J. Livingstone and T. Manallack, Statistics using neural networks: chance effects, *J. Med. Chem.* 36:1295 (1993).

**Section IX**  
**Poster Presentations**

**Poster Session I**  
**New Developments and**  
**Applications of**  
**Multivariate QSAR**



## FREE-WILSON-TYPE QSAR ANALYSES USING LINEAR AND NONLINEAR REGRESSION TECHNIQUES

Klaus-Jürgen Schaper

Research Center Borstel

D-23845 Borstel, Germany; e-mail: kschaper@fz-borstel.de

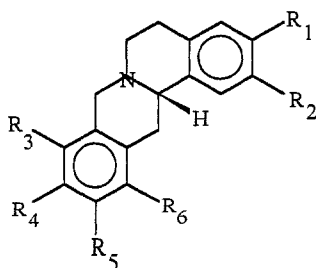
Tang *et al.*<sup>1,2</sup> recently published brain dopamine receptor affinity data of tetrahydroprotoberberine (THPB) derivatives (Table). These compounds contain only a small number of substituents in several positions of the parent molecule. For this type of data typically a Free-Wilson analysis is used for the derivation of QSARs. Meaningful results can be obtained by this type of analysis only if the activity contributions of single substituents are independent from each other or, in other words, if the contributions are constant and additive.

Inspection of the THPB data set showed that the substituent activity contributions seemed to be non-additive. This feeling was supported by a Free-Wilson (F-W) analysis (reference molecule: no. 6). Starting with 9 indicator variables the stepwise regression analysis resulted in an equation with only *one* remaining variable ( $I_{OH \rightarrow OMe}^{R2}$ ) that seemed to be statistically significant ( $r=0.583$ ; Leave-one-out (LOO) cross-validation:  $r=0.358$ ).

To recognize whether this result was indeed due to non-additivity or simply due to large experimental errors in the activity data a 'nonlinear Free-Wilson analysis' was performed using an Artificial Neural Network (ANN) and the same 9 indicator variables as independent variable input.

By applying different techniques for the reduction of the size of the network<sup>3,4</sup> it was found that only 2 of the 9 binary descriptors did not contribute significantly to the description of the THPB affinity data. Using a back-propagation neural network with 2 hidden layer neurons, 7 indicator variables and 13 network weights an excellent fit was obtained for the observed 15 affinity values ( $r=0.990$ ). As 13 weight values had been determined from only 15 compounds one has to consider that the network may be simply memorizing the observed data. To recognize whether the network indeed has predictive power a LOO cross-validation procedure was run using the same seven descriptors and ANN as before. A satisfactory correlation ( $r=0.810$ ) was found between observed and predicted activities. Thus the network was able to recognize that the *simultaneous* presence or absence of two or more substituents leads to high or low activity.

A 3-D / 4-D plot (not shown) of calculated activities vs. indicator variables is able to show the variability / non-additivity of substituent effects.



Brain dopamine D<sub>2</sub> receptor affinities (pK<sub>i</sub>) of tetrahydroprotoberberine derivatives

	R <sup>1</sup>	R <sup>2</sup>	R <sup>3</sup>	R <sup>4</sup>	R <sup>5</sup>	R <sup>6</sup>	Obsd. <sup>1,2</sup>	calc. F-W	calc. <sup>a</sup> ANN	pred. <sup>a</sup> ANN
1	OMe	OMe	OMe	OH	H	H	6.19	5.81	6.14	6.14
2	OH	OMe	OMe	OMe	H	H	6.26	5.81	6.12	6.10
3	OH	OMe	OMe	OH	H	H	6.22	5.81	6.12	6.11
4	OMe	OH	OMe	OMe	H	H	6.00	6.68	6.26	6.30
5	OMe	OH	OH	OMe	H	H	6.75	6.68	6.72	6.59
6	OMe	OH	OMe	OH	H	H	7.07	6.68	6.99	6.77
7	OMe	OMe	OMe	OMe	H	H	6.06	5.81	6.12	6.14
8	OMe	OMe	OH	OMe	OMe	H	5.13 min.	5.81	5.01	5.03
9	<b>O-CH<sub>2</sub>-O</b>		OMe	OMe	H	H	6.13	6.68	6.26	6.26
10	OH	OH	OMe	OMe	H	H	6.22	6.68	6.13	6.11
11	OH	OH	OH	OH	H	H	6.13	6.68	6.16	6.38
12	OMe	OH	H	OH	OMe	Cl	6.87	6.68	6.89	7.65
13	OMe	OMe	OH	OMe	H	Cl	5.66	5.81	5.71	5.80
14	OMe	OMe	OH	OMe	H	Br	5.16	5.81	5.17	5.62 <sup>b</sup>
15	OMe	OH	OH	OMe	H	Cl	8.24 max.	6.68	8.17	6.96

<sup>a</sup> calculated/predicted using 13 ANN weights;

<sup>b</sup> predicted without I<sup>R6</sup><sub>H→Br</sub> (because of singularity)

## REFERENCES

1. Y. Tang, K.-X. Chen, H.-L. Jiang, G.-Z. Jin, R.-Y. Ji, Molecular modeling of interactions between tetrahydroprotoberberines and dopamine receptors, *Acta Pharmacol.Sin.* 17:8 (1996).
2. Y. Tang, K. Chen, H. Jiang, G. Jin, R. Ji, 3D-QSAR study on tetrahydroprotoberberines using comparative molecular field analysis approach, 4<sup>th</sup> China-Japan Joint Symposium on Drug Design and Development, Xi'an, China, Oct.4-7, Abstracts, pp.115-118 (1995).
3. High correlation coefficients between weights of different hidden layer neurons indicate similar information transferred to different neurons, thus at least one of them may be deleted (Schaper, unpubl.).
4. T. Aoyama, H. Ichikawa, Reconstruction of weight matrices in neural networks - a method of correlating outputs with inputs, *Chem.Pharm.Bull.* 39:1222 (1991).

## QSAR STUDIES OF PICRODENDRINS AND RELATED TERPENOIDS - STRUCTURAL DIFFERENCES BETWEEN ANTAGONIST BINDING SITES ON GABA RECEPTORS OF INSECTS AND MAMMALS

Miki Akamatsu<sup>1</sup>, Yoshihisa Ozoe<sup>2</sup>, Taizo Higata<sup>2</sup>, Izumi Ikeda<sup>2</sup>, Kazuo Mochida<sup>2</sup>, Kazuo Koike<sup>3</sup>, Taichi Ohmoto<sup>3</sup>, Tamotsu Nikaido<sup>3</sup>, and Tamio Ueno<sup>1</sup>

<sup>1</sup>Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan

<sup>2</sup>Department of Life Science and Biotechnology, Shimane University, Matsue, Shimane 690-8504, Japan

<sup>3</sup>School of Pharmaceutical Sciences, Toho University, Funabashi, Chiba 274-8510, Japan

### INTRODUCTION

$\gamma$ -Aminobutyric acid (GABA), an inhibitory neurotransmitter, binds to the GABA<sub>A</sub> (ionotropic) receptor, to regulate the central nervous system of vertebrates. Insects have similar ionotropic receptors with different pharmacological properties, and, as a result, their GABA receptors represent promising targets for insecticides. Recently, 3D-QSAR analyses for insecticidal activity (against houseflies) and competitive activity against the specific [<sup>35</sup>S]*tert*-butylbicyclophosphorothionate (TBPS) binding (to rat brain membranes) of some picrotoxinin-type GABA antagonists, including  $\gamma$ -BHC, endosulfan, bicyclopophates, dioxatricyclododecenes (DTD) and related compounds, were carried out<sup>1</sup> using comparative molecular field analysis (CoMFA). The CoMFA results showed that similarities and dissimilarities in sterically and electrostatically favourable and forbidden regions on the molecule were reflected in the insecticidal and rat-receptor binding activities.

Picrodendrins<sup>2,3</sup> are a series of terpenoids which have been recently isolated from the Euphorbiaceae plant, *Picrodendron baccatum* (L.) Krug & Urban. Some of these terpenoids have been reported<sup>2</sup> to competitively inhibit the specific binding of the noncompetitive GABA antagonist [<sup>35</sup>S]TBPS to rat-brain membranes in a manner which is similar to the compounds previously used in CoMFA. The structure of picrodendrin Q which is the most potent of these is shown in Figure 1.

In this study, we examined the inhibition of the specific binding of [<sup>35</sup>S]TBPS and [<sup>3</sup>H]4'-ethynyl-4-*n*-propylbicycloorthoobenzoate (EBOB), the noncompetitive antagonist of GABA receptors, to rat-brain membranes as well as the binding inhibition of [<sup>3</sup>H]EBOB to housefly-head membranes using QSAR methods. Based on the results obtained, we infer structural differences in the noncompetitive antagonist binding sites of mammalian and insect GABA receptors.

## RESULTS AND DISCUSSION

The QSAR analysis of inhibition of the binding of [<sup>3</sup>H]EBOB to housefly-head membranes for picrodendrins and related compounds,  $\log [1/IC_{50}(M)]$ , yielded Eq. 1.

$$\log [1/IC_{50}(M)] = -2.870 q(C16) - 1.419 (I_{4OH} + I_{8OH}) + 6.933 \quad (1)$$

$$n = 12 \quad s = 0.427 \quad r = 0.877$$

where  $q(C16)$  is the atomic charge on the carbon atom 16, and  $I_{4OH}$  and  $I_{8OH}$  are indicator variables for the presence of the OH group at the 4- and 8-position, respectively. Eq. 1 shows that the electronegativity of the carbon atom 16 and the presence or absence of the 4- and 8-hydroxyl groups are important determinants of the potency of nor-diterpenes in housefly receptors. In the case of rat receptors, the number of available active nor-diterpenes was not sufficient to perform a quantitative analysis. However, the negative charge on the carbon 17-carbonyl oxygen atom appeared to be important. These findings indicate that significant differences exist between the structures of the complementary binding sites in mammalian and insect GABA receptors.

The superposition of picrodendrins onto GABA antagonists used for the previous CoMFA was carried out using the Superimpose and Field-fit procedures of SYBYL Ver. 6.4 to obtain 10 alignments. On the basis of those alignments, the inhibitory activity of the [<sup>35</sup>S]TBPS binding to rat-brain membranes was analyzed for picrodendrin A, B, M, O and Q along with a variety of GABA antagonists. A reliable CoMFA equation was obtained when one of the alignments was used. Picrodendrin Q and one of DTD compounds in the alignment are shown in Figure 2. Picrodendrin B, in which the structure of the  $\gamma$ -butyrolactone moiety is different from the others, was excluded from the equation because the measured activity was much higher than predicted. The equation showed contour maps which were similar to those drawn according to our previous CoMFA equation.<sup>1</sup>

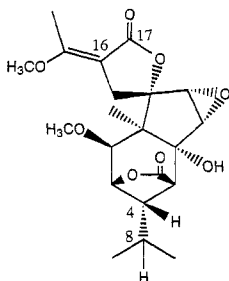


Figure 1. Structure of picrodendrin Q

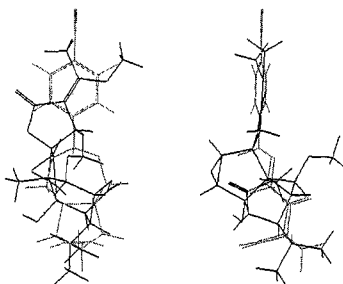


Figure 2. Superposition of picrodendrin Q (black) and one of DTD compounds (gray)

## REFERENCES

1. M. Akamatsu, Y. Ozoe, T. Ueno, T. Fujita, K. Mochida, T. Nakamura, and F. Matsumura, Sites of action of noncompetitive GABA antagonists in houseflies and rats: three-dimensional QSAR analysis, *Pestic., Sci.* **49**:319 (1997).
2. Y. Ozoe, H. Hasegawa, K. Mochida, K. Koike, Y. Suzuki, M. Nagahisa, and T. Ohmoto, Picrodendrins, a new group of picrotoxane terpenoids: structure-activity profile of action at the GABA<sub>A</sub> receptor-coupled picrotoxinin binding site in rat brain, *Biosci. Biotech. Biochem.* **58**:1506 (1994).
3. A.M. Hosie, Y. Ozoe, K. Koike, T. Ohmoto, T. Nikaido, and D.B. Sattelle, Actions of picrodendrin antagonists on dieldrin-sensitive and -resistant *Drosophila* GABA receptors, *Br. J. Pharmacol.* **119**:1569 (1996).

## Molecular lipophilicity descriptors: a multivariate analysis

Raimund Mannhold and Gabriele Cruciani

Institut für Lasermedizin, AG Molekulare Wirkstoff-Forschung, Heinrich-Heine-Universität, Universitätsstr.1, D-40225 Düsseldorf, Deutschland and Dipartimento di Chimica, Laboratorio di Chemiometria, Università di Perugia, Via Elce di Sotto, 8, I-06123 Perugia, Italia

### Introduction

The importance of lipophilicity in QSAR and drug design demands for the availability of quick, precise and reproducible experimental approaches to quantify this physico-chemical property. The Hansch group introduced the determination of log P in the octanol-water system as the standard. The need to derive lipophilicity data for steadily increasing numbers of compounds initiated the search for both experimental and computational alternatives to octanol-water partitioning. Calculation approaches are either atom-based or use fragments; in recent time attention is paid to the impact of 3D-aspects on lipophilicity. Application of calculation approaches demands a validity check with experimental data. In this study 2 experimental ( $\log P_{\text{Oct}}$ ,  $R_{\text{Mw}}$ ) and 17 calculation approaches (fragmental, atom-based, based on molecular properties) are investigated by regression and principal component analysis (PCA) for 159 molecules including simple structures and more complex drugs.

### Materials and Methods

**Experimental data:**  $\log P_{\text{Oct}}$  values are from (1-4); chromatographic lipophilicity data were derived by RP-TLC, according to (5). For the simple compounds they were published. (6).

**Calculated data:** lipophilicity was calculated with 17 programs; *fragmental methods:*  $\Sigma f$ -SYBYL, CLOGP, PROLOGP\_cdr, SANALOGP\_EO, SANALOGP\_ER, *atom-based methods:* MOLCAD, Tsar 2.2, PROLOGP\_atomics, CHEMICALC-2, SMILOGP; *methods based on molecular properties:* HINT, BLOGP, ASCLOGP. PROLOGP\_comb combines a fragmental (Rekker) and an atom-based approach (Ghose/Crippen). **Statistical analysis:** PCA was performed with GOLPE (7), version 3.1, on a Silicon Graphics workstation. The MREG option of SIMCA 3B (8) was used for regression analysis.

### Results

PCA of the entire database exhibits a clustering of chemical groups, preciseness of clustering corresponds to chemical similarity (Fig. 1). Thus, **diversity searching** in databases might effectively be performed by PCA on the basis of calculated log P.

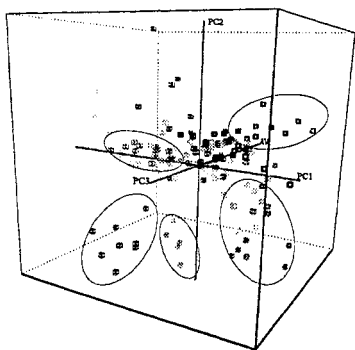


Fig. 1. Score plot of a PCA with the reduced data set ( $n=153$ ). Chemical groups are colour-coded and one can detect a clustering for imidazoles (dark-blue), halo-benzenes (magenta), benzamides (grey) and phenothiazines containing a piperazine side chain (dark-violet) or lacking this moiety (light-violet).

For broadly defined groups such as aromatic acids (black) or neutral aromates (yellow) clustering is less pronounced.

The comparative validity check of experimental and calculative procedures by regression analysis and PCA was performed with a chemically balanced, reduced data set ( $n = 55$ ) representing 11 chemical groups with 5 members, each.

Regression of experimental data ( $\log P_{\text{oct}} \leftrightarrow R_{\text{Mw}}$ ) proves that RP-TLC values can be used as valid and equivalent substitutes for  $\log P$ :

$$R_{\text{Mw}} = 0.994 (\pm 0.03) \log P_{\text{oct}} - 0.009 (\pm 0.09); n = 55; r = 0.995; s = 0.169; F = 5710$$

Regression of calculative versus experimental lipophilicity data exhibits a superiority of fragmental over atom-based methods and approaches based on molecular properties, as indicated by correlation coefficients, slopes and intercepts.

Present data indicate it worthwhile to unravel separate quality rankings for various chemical classes of interest; the KOWWIN program, shown to be excellent for drug molecules, exhibits a reduced predictivity for simple organics.

In addition, PCA (Fig. 2) revealed that fragmental methods (Rekker-type, KOWWIN, KLOGP) sense the compound ranking in  $\log P$  to almost the same extent as experimental methods. For atom-based procedures and CLOGP, both the comparability of absolute values and the sensing of the compound ranking in the database are slightly less. This trend is more pronounced for the methods based on molecular properties, with the exception of BLOGP.

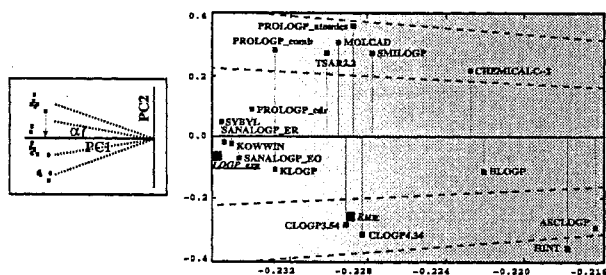


Fig. 2. PCA with the reduced, balanced set ( $n = 55$ ): loading plot of the first versus the second component. The information content is twofold. The distance between the projection of the data points (arrow in the left scheme) for experimental and calculative procedures onto the first component indicates the similarity in absolute values for experiment and calculation. The deviation of the loading direction of a given variable from the direction of the first PC (angle  $\alpha$  in the left scheme) reflects the similarity between a calculation procedure and the experimental approach in reflecting the compound ranking in  $\log P$  data within a database.

## References

- 1) Hansch, C., Leo, A. and Hoekman, D., Exploring QSAR. Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, DC, 1995
- 2) Mannhold, R., Dross, K. and Rekker, R.F., Quant. Struct.-Act. Relat. 9 (1990) 21.
- 3) Mannhold, R., Rekker, R.F., Sonntag, C., ter Laak, A.M., Dross, K. and Polymeropoulos, E.E., J. Pharm. Sci. 84 (1995) 1410
- 4) Taylor, P.J. and Cruickshank, J.M., J. Pharm. Pharmacol. 37 (1985) 143
- 5) Dross, K., Sonntag, Ch. and Mannhold, R., J. Chromatogr. A 638 (1993) 287
- 6) Dross, K., Sonntag, Ch. and Mannhold, R., J. Chromatogr. A 673 (1994) 113
- 7) Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S. Quant. Struct.-Act. Relat. 12 (1993) 9.
- 8) Wold, S., University of Umeå, 1983

## WORLD WIDE WEB-BASED CALCULATION OF SUBSTITUENT PARAMETERS FOR QSAR STUDIES

Peter Ertl

Novartis Crop Protection AG  
Lead Discovery  
CH-4002 Basel, Switzerland  
peter.ertl@cp.novartis.com

In the classical Hansch-Fujita correlation analysis<sup>1</sup> properties of molecules under study are quantified with help of hydrophobic, electronic and steric substituent parameters. In actual calculations these parameters are usually manually extracted from various data tables<sup>2</sup>. This approach, however, has numerous disadvantages, most notably low quality of data for rare functional groups and unavailability of parameters for many important substituents. Therefore a web-based system enabling interactive calculation of important substituent parameters for any organic functional group was developed. Properties calculated include hydrophobic, electronic and steric ones. Hydrophobic properties are represented by octanol-water partition coefficient ( $\pi$  constant) and molar refractivity. Both these parameters are calculated according to the methodology of Ghose and Crippen<sup>3</sup> based on the sum of atomic hydrophobicity contributions. Comparison of calculated and experimental<sup>2</sup>  $\pi$  constants for a set of 256 substituents yielded the following correlation (see also Fig. 1).

$$\pi_{\text{exp}} = 0.9916 * \pi_{\text{calc}}$$

$$n = 256, r^2_{\text{CV}} = 0.794, r^2 = 0.798, s = 0.540, \text{av.abs.error} = 0.410, F = 1005.6$$

The electron-donating and withdrawing power of substituents is characterized by theoretical parameters compatible with the Hammett  $\sigma$  constants. These are calculated according to the methodology developed in-house<sup>4</sup> from simple quantum chemical data. Comparison of calculated  $\sigma$  constants with experiment<sup>2</sup> for 368 organic substituent provided the correlation shown below (see also Fig. 1).  $q\gamma$  and  $q\delta$  are charges at the two terminal atoms of butadienylyl probe attached to the substituent (for details of the methodology see<sup>4</sup>).

$$\sigma_{\text{meta}} = 6.4274 + 14.9465 * q\gamma + 20.4036 * q\delta$$

$$n = 368, r^2_{\text{CV}} = 0.714, r^2 = 0.722, s = 0.142, \text{av.abs.error} = 0.101, F = 474.3$$

$$\sigma_{\text{para}} = 5.7509 + 9.8838 * q\gamma + 20.8919 * q\delta$$

$$n = 368, r^2_{\text{CV}} = 0.746, r^2 = 0.752, s = 0.186, \text{av.abs.error} = 0.135, F = 553.7$$

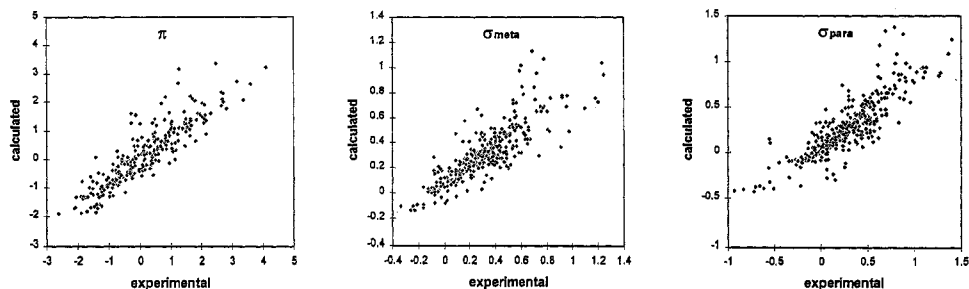


Fig. 1. Calculated vs. experimental substituent's  $\pi$ ,  $\sigma_{meta}$  and  $\sigma_{para}$  constants.

Steric properties of substituents are represented simply by their topological size (number of nonhydrogen atoms) and maximal topological length. In our experience these parameters are sufficient to characterize steric requirements of substituents. The addition of more sophisticated parameters (e.g., STERIMOL), however, would be straightforward.

Users interact with the system through the simple web interface<sup>5</sup>. In the entry part of the program the substituent for which data should be calculated is created with the help of our molecular editor written in Java. The editor creates a SMILES code for the substituent, which is passed to the CORINA<sup>6</sup> 3D geometry builder. Then AM1 calculation<sup>7</sup> is run to calculate charges. Other in-house programs calculate the  $\sigma$  constants from them, and estimate other substituent properties. Despite this relatively complex processing, the response is very fast and data are delivered within 2 - 3 seconds.

The processing engine behind the program may be called also directly (without the graphic interface) just by referencing to the address of the cgi script with substituent's SMILES as a parameter. In this way it is possible to calculate data for a large number of substituents in a "batch" mode. By using this technique, data for more than 80 000 functional groups used in substituent similarity searches or in the design of targeted combinatorial libraries with desired properties has been generated.

The module described here is a part of the Novartis web-based molecular modelling system<sup>8</sup> which delivers powerful and easy to use modelling capabilities directly to the desk of synthetic organic chemist. Numerous successful application of calculated substituent parameters in pesticide design at Novartis indicate that these data are becoming a really powerful alternative to the classical substituent parameters originated from experimental measurements.

## REFERENCES

1. C. Hansch and A. Leo, *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*, ACS Washington, DC (1995).
2. C. Hansch, A. Leo, D. Hoekman, *Exploring QSAR, Hydrophobic, Electronic, and Steric Constants*, ACS Washington, DC (1995).
3. V.M. Viswanadhan, A.K. Ghose, G.R. Revankar, R.K. Robins, *J.Chem.Inf.Comput.Sci.* 29, 163-172, (1989).
4. P. Ertl, *Quant.Struct.-Act.Relat.* 16, 377-382, (1997).
5. P. Ertl, *J.Molec.Graph.Model.* in press
6. J. Sadowski and J. Gasteiger, *Chem.Rev.* 93, 2567-2581, (1993).
7. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.P. Stewart, *J.Am.Chem.Soc.* 107, 3902-3909, (1985).
8. Ertl and O. Jacob, *THEOCHEM*, 419, 113-120, (1997), see also <http://www.elsevier.com/homepage/saa/eccc3/paper6>



## COMBINE AND FREE-WILSON QSAR ANALYSIS OF NUCLEAR RECEPTOR-DNA BINDING

Sanja Tomić<sup>1,2</sup>, Lennart Nilsson<sup>3</sup> and Rebecca C. Wade<sup>1</sup>

<sup>1</sup>EMBL, Meyerhofstr. 1, D-69012 Heidelberg, Germany

<sup>2</sup>Ruđer Bošković Institute, Bijenička 54, P.O.B. 1016, HR-10001 Zagreb, Croatia

<sup>3</sup>Center for Structural Biochemistry, Karolinska Institute, S-141 57 Huddinge, Sweden

### INTRODUCTION

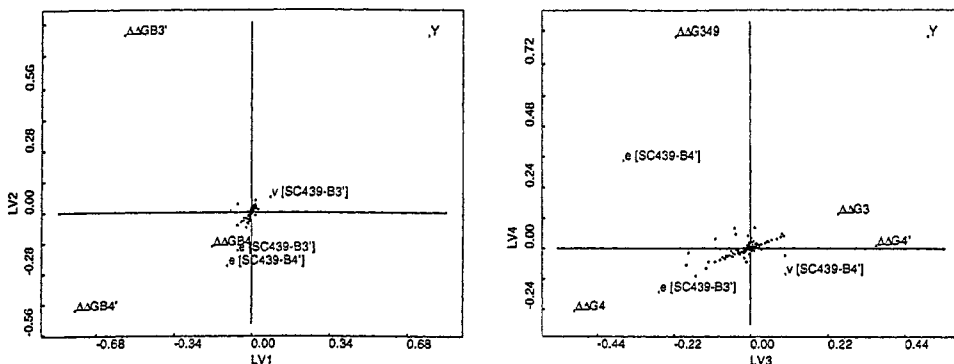
Specific binding of transcription factors to DNA is crucial for gene regulation. We have studied the specificity of binding of transcription factors from the nuclear receptor family to DNA using two QSAR methods: a) a Free-Wilson like method and b) Comparative Binding Energy (COMBINE)<sup>1</sup> analysis.

We used the experimental data of Zilliacus *et al.* (Zilliacus *et al.*, 1995b)<sup>2</sup> who studied how substitution of an amino acid at a single position of the DNA-binding domain (DBD) modulates DNA binding specificity. They measured the interaction of 20 mutant glucocorticoid receptor DBDs, which differ in the amino acid at position 439, with 16 different response elements.

### RESULTS

A Free-Wilson-like QSAR analysis was performed earlier<sup>3</sup> on a smaller set of complexes of DNA and glucocorticoid receptor DBD mutants with three variable positions. We applied the same analysis to the present data set of 320 complexes. We then compared the results with those from COMBINE 3D-QSAR analysis which was used to obtain physical insight into the features important for binding. For this purpose inter- and intra-molecular interaction energies per residue, changes of surface area, free energies of solvation of amino acid side-chains and the mutated bases, and side chain rotational entropy upon binding were analyzed.

The most important features for binding specificity are: the change of the solvation free energy of the mutated bases and of the mutated amino acid residue and the electrostatic and the van der Waals interaction energies of the side chain of the mutated residue with the mutated bases, see Fig. 1.



**Figure 1.** The partial weights plot for the X variables and the activity in the first two latent variables (left), and in the third and the fourth latent variables (right) obtained for the entire data set (320 objects) using the intermolecular interaction energies (van der Waals and electrostatic) and the free energies of solvation of amino acid side-chains and the mutated bases. Legend: SC = side chain, B = base, v [...] and e [...] = the van der Waals and the electrostatic interaction energy, respectively between the groups specified in the brackets,  $\Delta\Delta G$  = the change of relative free energy of solvation of amino SC and the mutated bases.

From the results obtained after the variable selection by the fractional factorial design, it seems that specificity of binding of transcription factor DBDs is not regulated solely by the residue at position 439, but also with the residues at some other positions, *i.e.* Lys-422 and Arg-427. The electrostatic interaction energies between these residues and mutated nucleotides appear as important descriptors of latent variables after the variable selection. However, it is clear that there are additional features important for the specificity of binding not included in this model, *e.g.* differences in interfacial hydration.

## REFERENCES

1. A.R. Ortiz, M.T. Pisabarro, F. Gago, and R.C. Wade, Prediction of drug binding affinities by comparative binding energy analysis, *J Med Chem* **38**(14), 2681-2691 (1995).
2. J. Zilliacus, A.P. Wright, J. Carlstedt-Duke, L. Nilsson, and J.A. Gustafsson, Modulation of DNA-binding specificity within the nuclear receptor family by substitutions at a single amino acid position, *Proteins: Struct., Func. and Genetics* **21**(1), 57-67 (1995b).
3. J. Zilliacus, A.P. Wright, U. Norinder, J.A. Gustafsson, and J. Carlstedt-Duke, Determinants for DNA-binding site recognition by the glucocorticoid receptor, *J. Biol. Chem.* **267**, 24941-24947 (1992).

## QSAR MODEL VALIDATION

Erik Johansson<sup>1</sup>, Lennart Eriksson<sup>1</sup>, Maria Sandberg<sup>1</sup> and Svante Wold<sup>2</sup>

<sup>1</sup> Umetri AB, POB 7960, 90719 Umeå, Sweden, [www.umetri.se](http://www.umetri.se)

<sup>2</sup> Institute of Chemistry, Umeå University, Sweden

In any modelling including QSAR, it is very important to validate the resulting model. For this purpose one should consider the QSAR model's predictive ability. Several tools are available for QSAR validation. The most demanding manner is by (I) external validation which consists of making predictions for an independent set of compounds not available during model training. External validation, however is often difficult in QSAR because it takes time and money to make new compounds and alternatives to external validation are hence of great interest. The alternatives discussed here are (II) cross-validation, (III) splitting the data into a training and a test set where both are present at modelling, and (IV) response permutation tests.

In the validation of a QSAR model it is essential to understand the nature of the data. Below, we will list four limiting cases of data structures. This classification is based on two principles, that is, whether the validation set compounds are (i) inside or outside the domain of the model (the training set) and (ii) inside or outside the biological activity range of the model

### **Limiting case A: Inside Model, Inside Y range**

This limiting case is characterised by a known and closed set of compounds, e.g. polychlorinated biphenyls. It is also likely that a representative training set has been selected and we know/believe that our training set contains the least potent and the most potent compounds. This is a fairly uncommon situation in QSAR but is common in e.g. NIR calibrations.

### **Limiting case B: Inside Model, Outside Y range**

Again, a known and closed set of compounds, e.g., polychlorinated biphenyls with a representative training set. Within this set of compounds, we want to find the most potent ones. This is a common situation that appears within toxicity and environmental studies<sup>1</sup>.

### **Limiting case C: Outside Model, Inside Y range**

The third limiting case might be a drug design study, where the goal is to make compounds for a patent. This patent should cover as many potent compounds as possible around a chosen candidate drug (CD), but these compounds need not be more potent than the CD. Here the structural domain is in principle endless and we want to make mild extrapolations<sup>2</sup>.

### **Limiting case D: Outside Model, Outside Y range**

The last limiting case also occurs within pharmaceutical industry. This is the most demanding case and we want to make new and unique compounds that have a higher potency than any of the existing compounds.

For limiting cases A and B, (II) cross validation and (III) splitting the data into a training and a test set, will be proper methods for QSAR validation. However for limiting cases C and D, when we extrapolate in the descriptor space, (II) cross validation and (III) splitting the data into a training and a test set, can be very misleading unless the training set is properly designed. The validation here requires a firm definition of the range of the model and assessment of the appropriateness of the validation set. The best way to define the scope and limitations of this approach is through the use of multivariate design. In addition for limiting case D we need to verify that we can predict outside the Y range. Therefore it is recommended that the two most active compounds are left outside the training set and placed in the test set.

A major problem in all the four limiting cases is that (II) cross validation and (III) splitting the data into a training and a test set, are sensitive to clustering and might give incorrect results if the data are grouped or clustered. As chemical compounds are discrete entities we know that our data always are more or less clustered. This problem is most pronounced the smaller number of compounds that are left out for prediction and worst for leave one out cross-validation and leave one-out predictions. Response permutation tests (IV)<sup>3</sup> is an additional validation tool that always should be employed as a complement as it has been found to give adequate warning for a number of clustered data sets.

### **References**

1. M. Tysklind, P. Andersson, P. Haglund, B. van Bavel and C. Rappe, Selection of polychlorinated biphenyls for use in quantitative structure-activity modelling, SAR QSAR Env. Res. 4:11 (1995).
2. L. Eriksson, P. Andersson, E. Johansson, M. Tysklind, M. Sandberg and S. Wold, The constrained principal property (CPP) space in QSAR- Directional and Non-directional modelling approaches, This volume.
3. L. Eriksson, E. Johansson, and S. Wold, QSAR model validation", in QSAR in Environmental Chemistry – VII Proceedings of the 7<sup>th</sup> International Workshop on QSAR in Environmental Sciences, June 24-28, 1996, Elsinore Denmark. F.Chen & G. Schuurmann (Eds) SETAC Press Pensacola, Florida, (1997), pp. 381-397

## QSPR PREDICTION OF HENRY'S LAW CONSTANT: IMPROVED CORRELATION WITH NEW PARAMETERS

John C. Dearden, Shazia A. Ahmed, Mark T.D. Cronin and Janeth A. Sharra

School of Pharmacy and Chemistry  
Liverpool John Moores University  
Byrom Street  
Liverpool L3 3AF  
England

### INTRODUCTION

Henry's law constant (H) is the air-water partition coefficient, and as such is important in modelling the environmental distribution of chemicals. Several quantitative structure-property relationship (QSPR) studies have been made of Henry's law constant; we recently (Dearden et al., 1997) developed such a QSPR for 294 diverse compounds from a consideration of the fundamental processes occurring during air-water partitioning, and using only calculated parameters:

$$\log H = -0.294 \text{HB}_N - 0.957 \text{HB}_I - 1.86 \Delta\text{MR} + 0.998 \log P - 1.11 \text{MR} + 0.356 \text{BI}_{\text{dw}}/100 + 0.229 {}^4\chi_p^v + 0.579 \quad (1)$$

$$n = 294 \quad r^2(\text{adj}) = 0.874 \quad s = 0.769 \quad F = 292.5$$

where:  $\text{HB}_N$  = total number of hydrogen (H) bonds that a molecule can form with water;  $\text{HB}_I$  = sum of indicator variables for H-bond donation and acceptance;  $\Delta\text{MR}$  = excess molar refractivity;  $\log P$  = calculated logarithm of octanol-water partition coefficient;  $\text{MR}$  = calculated molar refractivity;  $\text{BI}_{\text{dw}}$  = Bonchev index; and  ${}^4\chi_p^v$  = 4th order valence-corrected path molecular connectivity.

Equation 1, whilst giving reasonable predictions, has a rather high standard error, and our current work has been directed towards reducing this. We have used new quantitative hydrogen bonding parameters and also a parameter that reflects conformational entropy change, which is likely to be important for any process involving a phase change.

## METHOD

We used the same training set of 294 compounds and test set of 48 compounds as before (Dearden et al., 1997). Molecular connectivities were calculated using MOLCONN-X (Hall Associates, Quincy, MA) and the energy of the lowest unoccupied molecular orbital ( $E_{LUMO}$ ) was calculated using MOPAC 6. Four H-bonding parameters (free energy, enthalpy, and Kamlet  $\alpha$  and  $\beta$  values) were obtained from the HYBOT software (Raevsky, 1997). The number of rotatable bonds per molecule ( $B_R$ ) was used as a measure of conformational entropy; these values were kindly calculated by Dr. R.S. Pearlman of the University of Texas. An in-house genetic algorithm program was used to select the best combination of parameters.

## RESULTS AND DISCUSSION

The best seven-parameter equation that we could obtain was:

$$\log H = 2.31 {}^4\chi_c - 1.14 MR - 1.00 HB_I + 0.304 E_{LUMO} - 1.76 \alpha + 0.137 B_R + 1.09 \log P - 0.44 \quad (2)$$

$$n = 294 \quad r^2(\text{adj}) = 0.907 \quad s = 0.669 \quad F = 400.4$$

where:  ${}^4\chi_c$  = 4th order cluster molecular connectivity.

This is a considerable improvement over equation 1, especially with regard to the standard error. Using our test set of 48 different compounds, equation 2 gave a good correlation between observed and predicted log H values; however, one compound, *cis,trans*-cyclohexadec-8-en-1-one, was an outlier, due probably to the fact that Pearlman's program treated this compound as having no conformational flexibility. Removal of this compound gave:

$$\log H_{\text{obsd}} = 0.984 \log H_{\text{pred}} - 0.120 \quad (3)$$

$$n = 47 \quad r^2(\text{adj}) = 0.928 \quad s = 0.702 \quad F = 576.8$$

This is a better result than that obtained using Syracuse Research Corporation's HENRYWIN software (Meylan and Howard, 1992).

## REFERENCES

- Dearden, J.C., Cronin, M.T.D., Sharra, J.A., Higgins, C., Boxall, A.B.A., and Watts, C.D., 1997, The prediction of Henry's law constant: a QSPR from fundamental considerations, in: *Quantitative Structure-Activity Relationships in Environmental Sciences - VII*, F. Chen and G. Schuurmann, eds., SETAC Press, Pensacola, FL, pp. 135-142.
- Meylan, W.M. and Howard, P.H., 1992, Henry's law constant program manual, Lewis Publishers, Boca Raton, FL.
- Raevsky, O.A., 1997, Hydrogen bond strength estimation by means of the HYBOT program package, in: *Computer-Assisted Lead Finding and Optimization*, H. van de Waterbeemd, B. Testa and G. Folkers, eds., Wiley-VCH, Weinheim, pp. 369-378.

## QSAR OF A SERIES OF CARNITINE ACETYL TRANSFERASE (CAT) SUBSTRATES.

G. Gallo\*, M. Mabilia<sup>o</sup>, M. Santaniello\*, M.O. Tinti\*, P. Chiodi\*

\*Direzione Ricerche Sigma -Tau, Via Pontina km 30,400, I-00040  
Pomezia (RM) Italy

<sup>o</sup>S.IN - Soluzioni Informatiche S.a.s., Via Salvemini 9, I-36100 Vicenza,  
Italy

### INTRODUCTION

Carnitine acyl transferases are a family of enzymes that differ with respect to subcellular localization and substrate specificity. Carnitine acetyl transferase (CAT) is mainly found in the mitochondrial matrix where it is postulated to play a key role in stabilizing the CoA-SH / CoA-SAc ratio.<sup>1</sup> CAT catalyses the reversible reaction:



that has an equilibrium constant equal to 0.6. The kinetic enzymatic mechanism for CAT follows a random-order equilibrium reaction where Michaelis constant ( $K_m$ ) approximates true dissociation constant ( $K_s$ ) and binding of one substrate has little or no effect on binding of the second.<sup>2</sup> The aim of this work is to study a set of acyl-CoA derivatives that includes linear, branched and cycloalkyl, and unsaturated substituents.<sup>2,3,4</sup> A QSAR approach is used to investigate the influence of such substituents on kinetic parameters,  $K_m$  and  $V'max$ .

### MATERIALS AND METHODS

Kinetic parameters,  $K_m$  and  $V'max$  for a series of acyl-CoA derivatives - with linear (from C2 to C10), branched (isovaleryl and isobutyryl), cycloalkyl (C3 and C4) and unsaturated substituents (acryloyl, sorboyl and pent-4-enoyl) are taken from literature<sup>2,3,4</sup> with the exception of branched analogs that were tested in house as described by<sup>2</sup>.  $V'max$  are expressed as % with respect to the natural substrate, acetylCoA.

The descriptors employed in this study originally included: Verloop, connectivity indexes, moments of inertia, 3D-shape properties - all generated using TSAR<sup>TM</sup> software<sup>5</sup>- and electronic properties. In particular, energy values for the lowest unoccupied  $\Pi^*$  molecular orbital ( $E_{\Pi^*}$ ) and sum of squares of  $\Pi^*$  coefficient on the carbonyl C ( $\sigma^2_{\Pi^*}$ ), are calculated with the semi-empirical quantum mechanics method AM1 (Mopac software<sup>6</sup>) on R-(C=O)S-CH<sub>3</sub> fragments. Classification analyses and multiple linear regression (MLR) methods, available in TSAR<sup>TM</sup> software<sup>5</sup>, were employed to carry out this QSAR study.

## RESULTS AND DISCUSSION

QSAR equation could be obtained only for a subset of substrates with equal or very similar values, so branched and linear derivatives, with more than 8 C atoms in the alkyl chain, were excluded ( $K_m > 150$  and  $< 15$  respectively). When the acylCoA chain length varies from C<sub>2</sub> to C<sub>7</sub>,  $K_m$  values are invariant ( $38 \pm 6 \mu\text{M}$ ) while  $V'_{max}$  decreases 10-fold over the same length range with a logarithmic trend.

The following regression equation shows the linear relationships between kinetic parameter and chain length, expressed as Verloop L, for substrates with close CAT affinity:

$$\text{Log}(V'_{max}) = -0.18 L + 2.57$$

$$N = 7, s = 0.10, R^2 = 0.96, Q^2 = 0.92$$

CAT seems to contain a hydrophobic region that interacts with the side chain of long linear acyl-CoA compounds. Furthermore, branching in  $\alpha$  or  $\beta$  position to carbonyl might be responsible for steric hindrance.

Cycloalkyl derivatives are well tolerated but their activity, expressed as  $V'_{max}$ , is reduced. The parameter B3, one of the Verloop substituent size descriptors orthogonal to L, is necessary to describe cyclic derivatives:

$$\text{Log}(V'_{max}) = -0.31 B4 - 0.92 B3 + 4.47$$

$$N = 9, s = 0.22, R^2 = 0.88, Q^2 = 0.65$$

Similar equations can be obtained replacing B4 with L or B5: these descriptors are highly correlated for the set of substituents here considered.

Non-conjugated unsaturated compound, pent-4-enoylCoA, do not require additional parameters to L to explain its  $V'_{max}$ , while  $\alpha,\beta$ -unsaturated carbonyl derivatives require either a binary variable or an electronic descriptor in addition to L.

$$\text{Log}(V'_{max}) = -0.20 L + 3.18 c^2_{\pi^*} + 0.77$$

$$N = 10, s = 0.12, R^2 = 0.97, Q^2 = 0.88$$

The velocity difference does not result from an inductive effect, since an  $\alpha,\beta$ -unsaturated carbonyl is more electrophilic than a saturated one. So, on the basis of mesomeric effect, it has been speculated that an  $\alpha,\beta$ -unsaturated carbonyl might undergo reversible conjugate addition, which would compete with acyl transfer thus lowering  $V'_{max}$ . By the same token, the  $c^2_{\pi^*}$  value for  $\alpha,\beta$ -unsaturated carbonyl group is lower than the value for saturated carbonyl moiety, thus indicating a lower reactivity of the former one.

The following equation can finally be obtained if linear, cyclic and unsaturated analogs are considered together:

$$\text{Log}(V'_{max}) = -0.19 L - 0.87 B3 + 2.56 c^2_{\pi^*} + 2.74$$

$$N = 12, s = 0.20, R^2 = 0.91, Q^2 = 0.73$$

1. L. L. Bibier. *Current Concepts in Carnitine Research*, Carter A. L., CRC Press (1992)
2. J.F.A. Chase *Biochem. J.* 99:32 (1966)
3. W.J. Colucci, and R.D. Gandour *Bioorg. Chem.* 16:307 (1988)
4. P.C. Holland, A.E. Senior, and H.S.A. Sherratt *Biochem. J.* 136:173 (1973)
5. TSAR™ 3.1 Oxford Molecular Group Inc., Beaverton.
6. Mopac v.6.0 QCPE n°455 by J.J.P. Stewart.



## “CLASSICAL” AND QUANTUM MECHANICAL DESCRIPTORS FOR PHENOLIC INHIBITION OF BACTERIAL GROWTH

S. Shapiro<sup>1</sup> and D. Turner<sup>2</sup>

<sup>1</sup>Institut für orale Mikrobiologie und allgemeine Immunologie  
Zentrum für Zahn-, Mund- und Kieferheilkunde der Universität Zürich  
Plattenstrasse 11, Postfach, CH-8028 Zürich 7, Switzerland

<sup>2</sup>Krebs Institute for Biomolecular Research and Department of  
Information Studies, The University of Sheffield, Western Bank,  
Sheffield S10 2TN, United Kingdom

In connection with studies on effects of plant products on bacterial physiology,<sup>1-3</sup> minimal inhibitory concentrations (MICs) for ~150 monohydroxy phenols and related compounds were obtained for planktonic monocultures of three physiologically and ecologically diverse oral bacteria: *Porphyromonas gingivalis*, *Streptococcus sobrinus*, and *Selenomonas artemidis*. MIC values indicate that “activity space” is thoroughly and regularly covered for *P. gingivalis* and *Str. sobrinus*, the least and most active compounds having neighbours with very similar MIC values, though the coverage is less good for *S. artemidis*. High bacteriostatic activity is associated with (i) the basic phenol skeleton, (ii) presence of a single non-polar bulky substituent in the *ortho* position, and (iii) high hydrophobicity. A phenolic recognition site has been proposed consisting of a broad hydrophobic channel with a pocket *ortho* to the phenolic hydroxy moiety, and a vicinal hydrogen acceptor pointing away from the hydrophobic pocket.<sup>4</sup>

Phenolic “global” energy minima were obtained using GMMX 1.5,<sup>5</sup> and structure-activity data fitted to four QSAR paradigms to try to obtain equations with good log <sup>1</sup>/<sub>MIC</sub> predictivities and to elucidate the nature of the interactions between phenols and their target sites. Kier-Hall molecular connectivity indices produced statistically robust QSARs for *P. gingivalis* and *Str. sobrinus* where the first-order valence path index <sup>1</sup>χ<sup>v</sup> was the dominant variable, though results for *S. artemidis* were less satisfactory.<sup>4</sup> Structure-MIC data were also examined using Hansch analysis (*vide infra*), Famini’s theoretical linear solvation energy relationships (TLSEs),<sup>6</sup> and Todeschini’s weighted holistic invariant molecular (WHIM) descriptors.<sup>7</sup> Again, satisfactory equations were obtained for *P. gingivalis* and *Str. sobrinus*, but not for *S. artemidis*.<sup>8</sup> For the Hansch and TLSE QSARs the dominant descriptors are log P and V<sub>mc</sub> (+ π<sub>1</sub>), respectively. <sup>1</sup>χ<sup>v</sup>, V<sub>mc</sub>, and π<sub>1</sub> strongly correlate with log P, underscoring the importance of hydrophobicity for the antibacterial action of phenolics. The TLSE equations also suggest the importance of ligand transport to the relevant

cellular site *vis-à-vis* bioactivity.<sup>8</sup> WHIM descriptors were the *least* satisfactory in terms of statistical quality and interpretability for our data sets (results not shown).<sup>8</sup>

EVA analysis was also applied to the data sets. EVA is a descriptor derived from the *EigenVA* values of a classical normal coordinate analysis.<sup>9-10</sup> GMMX-optimised structures were re-optimised using MOPAC 6.0 (AM1; EF optimiser, GNORM = 0.01, MMOK where necessary; separate runs for FORCE calculations). EVA models were obtained for a Gaussian kernel width ( $\sigma$ ) of 10 cm<sup>-1</sup> using unscaled frequencies. Data reduction was accomplished using partial least squares (PLS), and the best, most parsimonious models chosen using jack-knife cross-validation.

Robust QSARs for *P. gingivalis* and *Str. sobrinus* were obtained with 5-6 latent variables (LVs), though (not unexpectedly) results were less satisfactory for *S. artemidis* (*vide infra*). The following information is noteworthy: (i) the functional group stretching region (1600-2200 cm<sup>-1</sup>) is underrepresented in the LVs, whereas the hydrogen stretching region (2700-3800 cm<sup>-1</sup>) is heavily weighted; (ii) there is a heavy weighting of variables centered around 1400 cm<sup>-1</sup>, an area populated by various C-H bending and scissoring vibrations as well as O-H bending vibrations.<sup>11-13</sup> There is very much intra- and interstructural overlap of EVA Gaussians around 1400 cm<sup>-1</sup>. EVA descriptor values tend to be maximal in regions with high kernel overlap; this results in substantial univariate variance around 1400 cm<sup>-1</sup>, which may explain the dominance of this region in the analysis.

Efforts are underway to use path analysis for "reverse-engineering" EVA QSARs to guide the design of molecular entities with enhanced pharmaceutical properties.

## HANSCH ANALYSIS EQUATIONS

***P. gingivalis*:**  $\log^1/\text{MIC} = 0.736 \log \mathbf{P} + 0.064 \text{L/B}_1 + 0.015 \text{polar S} - 2.957$   
 ( $\pm 0.027$ ) ( $\pm 0.025$ ) ( $\pm 0.004$ ) ( $\pm 0.182$ )  
 VIF: 1.2 1.1 1.3  
 n = 124  $r^2_{(\text{adj})} = 0.867$   $\sigma = 0.373$  F = 267.29  $R^2_{50\%} = 0.859$

***S. artemidis*:**  $\log^1/\text{MIC} = 0.611 \log \mathbf{P} - 0.106 \text{B}_3 + 0.007 \text{unsat'd S} - 2.006$   
 ( $\pm 0.029$ ) ( $\pm 0.038$ ) ( $\pm 0.001$ ) ( $\pm 0.146$ )  
 VIF: 1.5 1.5 1.0  
 n = 110\*  $r^2_{(\text{adj})} = 0.843$   $\sigma = 0.274$  F = 195.84  $R^2_{50\%} = 0.837$

***Str. sobrinus*:**  $\log^1/\text{MIC} = 0.744 \log \mathbf{P} + 0.057 \text{L/B}_1 + 0.017 \text{polar S} - 3.263$   
 ( $\pm 0.026$ ) ( $\pm 0.022$ ) ( $\pm 0.004$ ) ( $\pm 0.173$ )  
 VIF: 1.2 1.0 1.2  
 n = 111  $r^2_{(\text{adj})} = 0.889$   $\sigma = 0.331$  F = 293.91  $R^2_{50\%} = 0.897$

## EVA RESULTS

***P. gingivalis*:**  $\log^1/\text{MIC} = 1.709 \text{LV}_1 + 1.992 \text{LV}_2 + 1.293 \text{LV}_3 + 3.139 \text{LV}_4 + 1.473 \text{LV}_5 + 0.075$   
 ( $\pm 0.111$ ) ( $\pm 0.267$ ) ( $\pm 0.301$ ) ( $\pm 0.663$ ) ( $\pm 0.491$ ) ( $\pm 0.053$ )  
 % variances explained: LV<sub>1</sub> = 63.2; LV<sub>2</sub> = 14.9; LV<sub>3</sub> = 4.9; LV<sub>4</sub> = 6.1; LV<sub>5</sub> = 2.4  
 n = 124  $r^2_{(\text{adj})} = 0.916$   $\sigma = 0.296$  F = 268.38  $R^2_{50\%}$  (3 LVs only) = 0.799

***S. artemidis*:**  $\log^1/\text{MIC} = 1.408 \text{LV}_1 + 1.285 \text{LV}_2 + 1.011 \text{LV}_3 + 2.527 \text{LV}_4 + 1.021 \text{LV}_5 - 0.139$   
 ( $\pm 0.131$ ) ( $\pm 0.225$ ) ( $\pm 0.274$ ) ( $\pm 0.557$ ) ( $\pm 0.432$ ) ( $\pm 0.045$ )  
 % variances explained: LV<sub>1</sub> = 53.5; LV<sub>2</sub> = 15.1; LV<sub>3</sub> = 6.2; LV<sub>4</sub> = 9.7; LV<sub>5</sub> = 2.6  
 n = 110\*  $r^2_{(\text{adj})} = 0.871$   $\sigma = 0.238$  F = 148.10  $R^2_{50\%}$  (3 LVs only) = 0.658

**Str. sobrinus:**  $\log^1/\text{MIC} = 1.753 \text{ LV}_1 + 1.786 \text{ LV}_2 + 1.425 \text{ LV}_3 + 3.250 \text{ LV}_4 + 0.928 \text{ LV}_5 + 1.717 \text{ LV}_6$   
 $(\pm 0.084) \quad (\pm 0.202) \quad (\pm 0.233) \quad (\pm 0.511) \quad (\pm 0.343) \quad (\pm 0.602)$   
 $- 0.188$   
 $(\pm 0.039)$

% variances explained:  $\text{LV}_1 = 68.5$ ;  $\text{LV}_2 = 12.3$ ;  $\text{LV}_3 = 5.9$ ;  $\text{LV}_4 = 6.5$ ;  $\text{LV}_5 = 1.1$ ;  
 $\text{LV}_6 = 1.3$

$n = 111$   $r^2_{(\text{adj})} = 0.956$   $\sigma = 0.209$   $F = 396.97$   $R^2_{50\%} (3 \text{ LVs only}) = 0.841$

$\log P = n$ -octanol/water partition coefficient;  $L$ ,  $B_1$ ,  $B_3 = \text{STERIMOL}$  parameters as per ref. 8;  $S = \text{molecular surface area}$ ;  $R^2_{50\%} = \text{cross-validation statistic calculated as per ref. 8}$ .

\*nb: 5 worst outliers omitted

## REFERENCES

1. S. Shapiro, A. Meier, and B. Guggenheim, The antimicrobial activity of essential oils and essential oil components towards oral bacteria, *Oral Microbiol. Immunol.* 9:202 (1994).
2. S. Shapiro and B. Guggenheim, The action of thymol on oral bacteria, *Oral Microbiol. Immunol.* 10:241 (1995).
3. S. Shapiro, The inhibitory action of fatty acids on oral bacteria, *Oral Microbiol. Immunol.* 11:350 (1996).
4. S. Shapiro and B. Guggenheim, Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity, *Quant. Struct.-Act. Relat.* 17 (1998, in press)
5. F.L. Tobiason and R.W. Hemingway, Predicting heterocyclic ring coupling constants through a conformational search of tetra-*O*-methyl-(+)-catechin, *Tetrah. Lett.* 35: 2137 (1994).
6. R.F. Famini and L. Y. Wilson, Using theoretical descriptors in linear solvation energy relationships, *Theor. Comput. Chem.* 1:213 (1994).
7. R. Todeschini and P. Gramatica, 3D-Modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies, *Quant. Struct.-Act. Relat.* 16:120 (1997).
8. S. Shapiro and B. Guggenheim, Inhibition of oral bacteria by phenolic compounds. Part 2. Correlations with molecular descriptors, *Quant. Struct.-Act. Relat.* 17 (1998, in press)
9. A.M. Ferguson, T. Heritage, P. Jonathon, S.E. Pack, L. Phillips, J. Rogan, and P.J. Snaith, EVA: a theoretically-based molecular descriptor for use in QSAR/QSPR analysis, *J. Comput.-Aided Mol. Design* 11:143 (1997).
10. D.B. Turner, P. Willett, A.M. Ferguson, and T. Heritage, Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application, *J. Comput.-Aided Mol. Design* 11:409 (1997).
11. J.R. Jakobsen and J.W. Brasch, Far infrared studies of the hydrogen bond of phenols. *Spectrochim. Acta* 21:1753 (1965).
12. J.H.S. Green, D.J. Harrison, and W. Kynaston, Vibrational spectra of benzene derivatives — XIV. Mono substituted phenols, *Spectrochim. Acta* 27A:2199 (1971).
13. T.N. Pliev, Complex method of identification of molecular structures of substituted phenols from their IR, UV, and NMR spectra. *J. Appl. Spectrosc.* 47:1259 (1987).

## HYDROGEN BOND ACCEPTOR AND DONOR FACTORS, C<sub>a</sub> AND C<sub>d</sub>: NEW QSAR DESCRIPTORS

James W. McFarland<sup>1</sup>, Oleg A. Raevsky<sup>2</sup> and Wendell W. Wilkerson<sup>3</sup>

<sup>1</sup>reckon.dat consulting, Old Lyme, Connecticut 06371

<sup>2</sup>The Institute of Physiologically Active Compounds of the Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russia,

<sup>3</sup>Dextron Corporation, Bear, Delaware 19710

### INTRODUCTION

Hydrogen bonding has been widely recognized as an important contributor to the forces binding a drug to its receptor, and also as one of the physical properties associated with lipophilicity and cell permeability. Until lately, H-bonding ability mainly has been described in QSAR problems by the use of indicator variables, e.g. the presence or absence of a H-bond donor (1 or 0). Over the past two decades, Raevsky and coworkers<sup>1</sup> have prepared a large database (>12,000 entries) of thermodynamic measurements on H-bonding systems. From these data, the Russian team developed a method to estimate the H-bond acceptor and donor strengths of various chemical moieties. Both the thermodynamic database and the method are available as software called HYBOT (HYdrogen BOND Thermodynamics).

We will show here that H-bond acceptor and donor factors, C<sub>a</sub> and C<sub>d</sub> as calculated by HYBOT, are superior to indicator variables in QSAR analyses. One example concerns the ability of diverse compounds to penetrate skin<sup>2</sup>. Another deals with data related to certain cyclic ureas in their ability to inhibit HIV protease<sup>3</sup>.

### RESULTS AND DISCUSSION

In an unpublished result, Lien<sup>4</sup> found that skin penetration of 23 diverse compounds correlated with measured log P (MLOGP) and molecular weight (MW), ( $R^2 = 0.96$ ). He also considered the number of H-bond sites in each molecule (HB) and phenolic character (I, 1 or 0), but these played no role in the final outcome. To this same dataset we added the descriptors: sum of H-bond acceptor and donor factors [ $\Sigma C_a$  and  $\Sigma C_d$ ], and calculated molar refractivity and calculated log P (CMR and CLOGP, MedChem Software, Pomona College, Claremont, CA). When we used forward stepwise regression, we found skin penetration

correlated with MLOGP and  $\Sigma C_a$  ( $R^2 = 0.96$ ); with backward stepwise regression it correlated with CLOGP,  $CLOGP^2$ , CMR and  $\Sigma C_d$  ( $R^2 = 0.97$ ). These results suggested that there was high collinearity among all the descriptors considered. Therefore, we analyzed the data by PLS. The PLS model (two components) gave a good correlation ( $R^2 = 0.96$ ), but more importantly the loadings plot showed clearly defined groupings of correlated descriptors: MW with CMR; MLOGP with CLOGP;  $\Sigma C_a$  and  $\Sigma C_d$  with HB; and a group of one, **I**. Hence, among these compounds, skin penetration depends on molecular size, lipophilicity, H-bonding capacity and phenolic character, a result in keeping with other permeability studies.<sup>5</sup>

Wilkerson et al.<sup>3</sup> recently reported QSAR results on some symmetrical cyclic urea HIV protease inhibitors. The inhibitors were modified by varying R in two identical CONHR groups. When R was 2-pyridyl or an analog thereof, there was a dramatic increase in potency. It was hypothesized that a critical H-bond formed from the enzyme to the N atom in the heteroaromatic ring. Following the example of Wilkerson et al.<sup>3</sup>, we used an indicator variable (**I**) for this H-bond effect in 30 compounds and found the relationship:

$$\begin{aligned}
 -\log Ki &= 0.09(\pm 0.03)CLOGP^2 - 0.78(\pm 0.32)CLOGP - 0.95(\pm 0.39)mv \\
 &\quad - 0.006(\pm 0.002)MW + 1.47(\pm 0.26)\mathbf{I} + 11.77 \\
 n = 30 \quad R^2 &= 0.68 \quad s = 0.41 \quad F_{5,24} = 10.11 \quad p < 0.00003
 \end{aligned}$$

where Ki is the inhibition constant and mv is 1/100th of the molecular volume.

Because the CONHR group also has the potential to be a H-bond donating group and because of our interest in H-bond factors, we next used  $C_aN$  to estimate the H-bond acceptor strength of the "2-pyridyl" N atom and  $C_dNH$  for H-bond donating capacity of the amide group. These H-bond factors were determined for each of the 30 compounds by HYBOT. Regression analysis gave the following best outcome:

$$\begin{aligned}
 -\log Ki &= 0.10(\pm 0.02)CLOGP^2 - 0.95(\pm 0.26)CLOGP - 0.78(\pm 0.31)mv \\
 &\quad - 0.008(\pm 0.002)MW - 0.48(\pm 0.12)C_dNH + 0.51(\pm 0.09)C_aN + 11.55 \\
 n = 30 \quad R^2 &= 0.80 \quad s = 0.33 \quad F_{6,23} = 15.52 \quad p < 0.000001
 \end{aligned}$$

In this result the H-bond factors superseded **I**. Hence, a substantial improvement in the correlation between log Ki and its physical properties was obtained by using *quantitative* estimates of H-bond capacity at potential key H-bonding sites.

## REFERENCES

1. O.A. Raevsky, V. Grigor'ev, E. Mednikova, QSAR H-bonding descriptions, in: *Trends in QSAR and Molecular Modelling 92*, C.G. Wermuth, ed., ESCOM, Leiden, (1993).
2. R.L. Scheuplein, R.L. Bronaugh, in: *Biochemistry and Physiology of the Skin*, L.A. Goldsmith, ed., Oxford University Press, Oxford (1983), p. 1255 ff.
3. W.W. Wilkerson, E. Akamike, W.W. Cheatham, A.Y. Hollis, R.D. Collins, I. DeLucca, P.Y.S. Lam, Y. Ru, HIV Protease Inhibitory Bis-benzamide Cyclic Ureas: A Quantitative Structure-Activity Relationship Analysis, *J. Med. Chem.*, 39:4299 (1996).
4. E.J. Lien, C-QSAR Program, BioByte Corp. 201 West 4th St., Suite 204, Claremont, CA 91711, USA, Dataset name: BIO\_1700.
5. H. van de Waterbeemd, G. Camenisch, G. Folkers, O.A. Raevsky, Estimation of Caco-2 Cell Permeability using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.*, 15:480 (1996), and references therein.

## Development and Validation of a Novel Variable Selection Technique with Application to QSAR Studies

Chris L. Waller<sup>a</sup> and Mary P. Bradley<sup>b</sup>

<sup>a</sup>OSI Pharmaceuticals, Inc.  
106 Charles Lindbergh Blvd.  
Uniondale, NY 11553

<sup>b</sup>Rhone-Poulenc Agro  
2 TW Alexander Drive  
Research Triangle Park, NC 27711

Variable selection is typically a time-consuming and ambiguous procedure in performing quantitative structure-activity relationship (QSAR) studies on over-determined (regressor-heavy) data sets. A variety of techniques including stepwise and partial least squares/principle components analysis (PLS/PCA) regression have been applied to this common problem. Other strategies, such as neural networks, cluster significance analysis, nearest neighbor, or genetic (function) or evolutionary algorithms have also been evaluated. A simple random selection strategy that implements iterative generation of models, but directly avoids cross-over and mutation, has been developed and is implemented herein to rapidly identify from a pool of allowable variables, those which are most closely associated with a given response variable. The FRED (fast random elucidation of determinants) algorithm begins with a population of offspring (models) composed of a fixed, or variable, number of randomly selected variables. Iterative elimination of descriptors leads naturally to subsequent generations of more fit offspring (models). In contrast to common genetic and evolutionary algorithms, only those descriptors determined to contribute to the genetic make-up of less fit offspring (models) are eliminated from the descriptor pool. After every generation, a new random increment line search of the remaining descriptors initiates the development of the next generation of randomly constructed models. An optional algorithm which eliminates highly correlated descriptors in a stepwise manner prior to the development of the first generation of offspring greatly enhances the efficiency of the FRED algorithm. A FRED analysis on a set of antifilarials published by Selwood (n=31 compounds, k=53 descriptors) demonstrates the ability of the algorithm to rapidly identify determinants of biological outcome from a large collection of highly intercorrelated

variables (see Figure 1.). A comparison of the results of a FRED analysis of the Selwood data set with those obtained using alternative algorithms reveals that this technique is capable of identifying the same “optimal” solutions in an efficient manner.

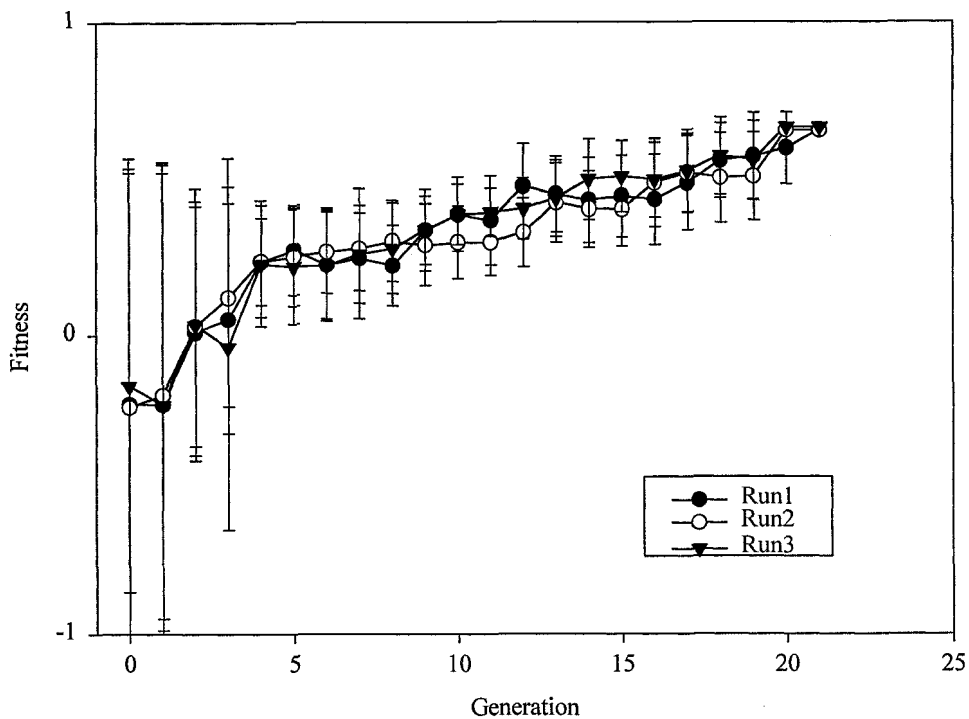


Figure 1. Evolution of FRED-derived QSAR Models

## QSAR STUDIES OF ENVIRONMENTAL ESTROGENS

M.G.B. Drew<sup>1</sup>, N.R.Price<sup>2</sup>, H.J. Wood<sup>1</sup>

<sup>1</sup>Chemistry Department, University of Reading  
Whiteknights, Reading, RG6 6AD, UK

<sup>2</sup>Central Science Laboratory, MAFF  
Sand Hutton, York, YO4 1LZ, UK

### INTRODUCTION

There is now considerable concern about the accumulation of estrogenic substances within the environment. In addition to proprietary steroidal hormones, a variety of structurally diverse chemicals are thought to mimic estrogens. These have been shown to have adverse reproductive effects on invertebrates, and hence perturb ecological dynamics. Since the array of potential estrogens in the environment is diverse and it may not prove possible to assess all compounds in the laboratory, some way of assessing likely estrogenic potential is needed. Consequently, it is desirable to formulate simple predictive models in order to direct experimental studies.

### METHODS

Experimental data for 20 polychlorinated biphenyls (PCBs) and 48 estradiol derivatives were taken from the literature<sup>1,2,3</sup>. The activity values for the group of estradiols were measured as RBA units, the ratio of test compound to labelled estradiol bound to the receptor. Activity values for the PCBs were also converted to RBA units. Starting structures were derived from crystallographic data<sup>4</sup>, and were assigned atomic charges using the QEq method<sup>5</sup>. Lowest energy conformations were established from molecular mechanics minimisation using the Dreiding field<sup>6</sup>, and the molecules were then superimposed using alignment of relevant 6-membered rings.

Regression models were derived separately for each group of compounds, based upon both molecular descriptors and molecular field analysis (MFA) data. Seventy-two descriptors were calculated for each molecule using the TSAR software package<sup>7</sup>, including electrostatic quantities, topological and connective indices, and parameters derived from atomic and molecular log P values. In addition, Cerius<sup>2</sup> <sup>5</sup> was used



to carry out molecular field analyses using a minimised hydroxyl probe within a 2Å spaced grid of suitable dimensions.

After removing significantly correlated variables from both types of data, the molecular descriptors were used to construct separate stepwise regression models for each group of compounds. In addition, this technique was used to predict activity values for groups of molecules containing both PCBs and estradiols. QSAR equations were also derived from the MFA data, using the genetic algorithm implementation in Cerius<sup>2</sup> and a partial least squares (PLS) regression for the separate groups of PCBs and estradiols respectively. These two statistical tools are combined in the GPLS module of Cerius<sup>2</sup> - the genetic algorithm selects a specified number of significant MFA descriptors to be incorporated into the PLS analysis. This algorithm was used to model the binding behaviour of the combined training group of PCBs and estradiols. Crossvalidation of the predictive equations for the training set was performed, and the models were also tested using external test sets (8 PCBs, 17 estradiols).

## RESULTS

Stepwise regression models using the descriptor data for the separate groups of molecules gave reasonable  $r^2$  values ( $\geq 0.721$ ). The activity values of the combined training set of compounds were also successfully modelled using this regression technique, with  $r^2 = 0.853$  (crossvalidated  $r^2 = 0.779$ ), and good predictions were also made for the combined test set ( $r^2 = 0.954$ ). It was observed that these QSAR equations gave more accurate predictions for the activity of the PCBs, reflecting the smaller group and relative structural simplicity of these compounds.

Generally, models using MFA data were found to give  $r^2$  values slightly greater than those generated via stepwise regression. The GPLS algorithm was also used to construct a predictive QSAR equation using the combined sets of molecules, giving  $r^2$  values of 0.839 and 0.962 for the training and test sets respectively. Although these values are not significantly better than those obtained by stepwise regression, the operating parameters of the genetic algorithm aspect of this technique may be adjusted to include more of the MFA data in the PLS analysis, possibly improving the model.

## REFERENCES

- [1] K Connor et al. Hydroxylated polychlorinated biphenyls (PCBs) as estrogens and anti - estrogens: structure - activity relationships. *Toxicol. App. Pharmacol.* 145: 111 (1997).
- [2] SP Bradbury, OG Mekenyan and GT Ankley. Quantative structure - activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity: an assessment of conformer flexibility. *Env. Toxicol. Chem.* 15: 1945 (1996).
- [3] TG Gantchev, H Ali and JE van Lier. Quantative structure - activity relationships / comparative molecular field analysis (QSAR/CoMFA) for receptor binding properties of halogenated estradiol derivatives. *J. Med. Chem.* 37: 4164 (1994).
- [4] Cambridge Database Service, Daresbury Laboratory, Warrington, UK.
- [5] Cerius<sup>2</sup>, May 1997, Molecular Simulations Inc., San Diego.
- [6] SL Mayo, BD Olafson and WD Goddard III. Dreiding: a generic forcefield for molecular simulations *J. Phys. Chem.* 94 (1990).
- [7] Tsar v2.4, Oxford Molecular Ltd. Oxford Science Park, Oxford, UK.

# QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP OF ANTIMUTAGENIC BENZALACETONES AND RELATED COMPOUNDS

Chisako Yamagami<sup>1</sup>, Noriko Motohashi<sup>1</sup> and Miki Akamatsu<sup>2</sup>

<sup>1</sup>Kobe Pharmaceutical University, Kobe 658-8558, Japan

<sup>2</sup>Graduate School of Agriculture, Kyoto University  
Kyoto, 606-8502, Japan

## INTRODUCTION

In the course of our study of the antimutagenic activity produced by  $\alpha$ ,  $\beta$ -unsaturated carbonyl compounds such as curcumin and cinnamaldehyde, we found that benzalacetones(I) have an antimutagenic effect on UV-induced mutagenesis<sup>1</sup>. This finding prompted us to study quantitatively their structure-activity relationship. In this study, we performed the 3D-QSAR analysis of the antimutagenic potency of (I) to examine the sterically and electrostatically favorable regions for activity.

## MATERIALS AND METHODS

**Compounds:** Compounds tested were (I) with X-substituents such as, H, halogens, alkyls, OR, CN, NO<sub>2</sub>, OH, NMe<sub>2</sub>, and NMe<sub>3</sub><sup>+</sup>. Among compounds tested, 22 compounds were active and subjected to the QSAR analysis. The 2-OMe, 2-Me, 3-Me, 4-Me, 4-OAc, and 4-NMe<sub>3</sub><sup>+</sup> derivatives were not active, and the 4-NMe<sub>2</sub> derivative was found to be mutagenic. Preparation of the compounds is described elsewhere.<sup>2</sup>

**log P values:** The log P values of all compounds tested were measured by the shake-flask method.<sup>2</sup>

**Assay for antimutagenic activity:** Bio-antimutagenic activity was assayed by observing mutagenesis induced by UV-irradiation at 254nm (1 J/m<sup>2</sup>) in *E. coli* WP2s (*uvrA*<sup>-</sup> *trpE*). The reverse mutations (M) and viable cells (V) irradiated were measured after the incubation at 37°C for 3 days using semi-enriched minimal agar plates with various amounts of test compound. The spontaneous revertants (S) and viable cells of untreated were measured simultaneously. The induced mutation frequency (IMF, Trp<sup>+</sup> revertants/10<sup>7</sup> cells) were calculated using the equation, IMF = 10<sup>7</sup> (M - S)/10<sup>5</sup>V. The IC<sub>50</sub> value, which represents the dose reducing the mutation frequency to 50% of the control values, was calculated from the results of assays at various sample concentrations.

**3D-QSAR:** Structures of the 22 compounds were fully optimized by using the AM1 method in the MOPAC 93 program package incorporated in an ANCHOR II modeling system (Fujitsu). For the optimized coordinates, the esp charges were calculated. Using the esp charges obtained, the pIC<sub>50</sub> values were analyzed three-dimensionally using the comparative molecular field analysis (CoMFA) module of the SYBYL software package (Ver. 6.4). The five atoms, 1-5, in (I) were superposed.

## RESULTS AND DISCUSSION

The analysis for all compounds yielded Eq. 1,

$$pIC_{50} = 3.298 + [\text{CoMFA field terms}] \quad (1)$$

$$CN = 5, n = 22, s = 0.118, r^2 = 0.950, s_{CV} = 0.317, r_{CV}^2 = 0.639$$
$$RC_{steric} = 0.341, RC_{electro.} = 0.659$$

where  $CN$  and  $RC$  are the number of components and relative contributions, respectively. The parameters with the subscript  $cv$  represent those from the leave-one-out cross validation. Interestingly, addition of  $\log P$  term to Eq. 1 did not improve the correlation, presenting a striking contrast to the antimutagenic effect on  $\gamma$ -induced mutagenesis in *Salmonella typhimurium* TA2638 which was found to correlate well with  $\log P$  (equation not shown). Eq. 1 indicates that the electronic factor is the most important. Figure 1 shows that electron-withdrawing substituents at the 4-position is favorable for activity. With the compounds used, clear steric requirements could not be obtained.

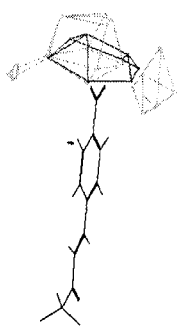
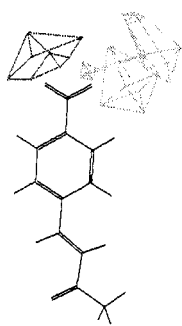
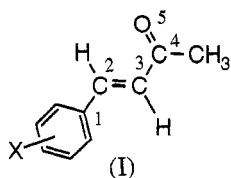


Figure 1

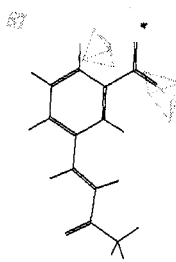


Figure 2

**Figure 1.** Orthogonal views of the electrostatic field map for (I) according to Eq. 1 with the 4-NO<sub>2</sub> derivative inserted. The contours surround regions where a negative (light gray lines) or positive (dark gray lines) electrostatic potential increases the activity.

**Figure 2.** Orthogonal views of the steric field map for (I) according to Eq. 1, with the 3-NO<sub>2</sub> derivative inserted. The contours surround regions where a higher steric bulk increases (light gray lines) or decreases (dark gray lines) the activity.

## REFERENCES

1. N. Motohashi, Y. Ashihara, C. Yamagami, and Y. Saito, Antimutagenic effects of dehydrozingerone and its analogs on UV-induced mutagenesis in *Escherichia coli*, *Mutat. Res.*, 377: 17(1997).
2. C. Yamagami, N. Kishida, S. Ka, M. Horiuchi, and N. Motohashi, Hydrophobicity parameters of antimutagenic benzalacetones and related compounds, *Chem. Pharm. Bull.*, 46: 274(1998).

# MULTIVARIATE REGRESSION EXCELS NEURAL NETWORKS, GENETIC ALGORITHM AND PARTIAL LEAST-SQUARES IN QSAR MODELING

Bono Lučić and Nenad Trinajstić

The Rugjer Bošković Institute  
P.O. Box 1016  
HR-10001 Zagreb, Croatia

## INTRODUCTION

Depending on the mathematical approach used in the QSAR analysis, the final models may be quite different in their complexity, accuracy, stability and predictability. This comparative study is undertaken in order to see which of the most frequently used methods is most effective in searching for the 'best' models. A couple of problems we have to solve in QSAR modeling: the first is related to the selection of the most relevant descriptors, and the second is the generation of the most reliable models. Contrary to other methods, our procedure, which is based on multiregression (MR) analysis, solves both of these problems simultaneously. In the present study the Selwood data set, which has become a standard for testing QSAR, is used.<sup>1</sup> This data set was already used for the determination of QSAR models by applying the neural networks (NNs),<sup>2</sup> genetic algorithm (GA),<sup>3</sup> and partial-least squares (PLS).<sup>4</sup> Three types of MR models are generated: (1) the best possible MR models; (2) the MR model with ordered orthogonalized descriptors; (3) the nonlinear MR models (take into account linear descriptors, their squares and cross-product terms).

## RESULTS

The Selwood data set contains the series of 31 antifilarial antimycin analogues.<sup>1</sup> For each compound 53 physicochemical descriptors were calculated. The quality of models is indicated by the  $R$ ,  $R_{cv}$  (leave-one-out cross-validated correlation coefficient),  $S$ ,  $S_{cv}$  (cross-validated standard error),  $F$  and  $Q^2$ .

The best possible linear MR models (with the highest  $R$  and  $R_{cv}$ ) with 4, 5, 6, and 7 descriptors were selected from the data set of  $N$  descriptors (Table 1, A).

The best ordered orthogonalized MR model with 3 significant descriptors ( $\Omega_{50}$ ,  $\Omega_{52}$ ,  $\Omega_4$ ) was obtained by the orthogonalization of 7 descriptors (from the model in Table 1, A) in the following order:  $d_{38}$ ,  $d_{13}$ ,  $\underline{d}_{50}$ ,  $d_{11}$ ,  $d_{48}$ ,  $\underline{d}_{52}$ ,  $\underline{d}_4$  (Table 1, B). This order was selected as the

best one, after the orthogonalization in all possible orderings was carried out.

Nonlinearities were introduced through cross-product terms of the initial descriptors. These 1431 descriptors were considered with the initial 53 descriptors, and the best possible nonlinear MR models were selected. The statistical parameters of the best model with 4 descriptors are given in Table 1, C.

**Table 1.** Statistical parameters of multiregression, NN<sup>2</sup>, GA<sup>3</sup> and PLS<sup>4</sup> models.

$I^a$	$R$	$S$	$R_{cv}$	$S_{cv}$	$F$	$Q^2$	descriptors
<b>A. The best possible <u>linear</u> MR models</b>							
4	<b>0.863</b>	0.410	<b>0.817</b>	0.470	19.0	<b>0.665</b>	$d_{12}, d_{38}, d_{50}, d_{52}$
5	<b>0.904</b>	0.347	<b>0.842</b>	0.446	22.5	<b>0.699</b>	$d_4, d_{11}, d_{39}, d_{50}, d_{52}$
6	<b>0.924</b>	0.311	<b>0.871</b>	0.403	23.2	<b>0.754</b>	$d_4, d_{11}, d_{38}, d_{48}, d_{50}, d_{52}$
7	<b>0.928</b>	0.304	<b>0.878</b>	0.394	20.2	<b>0.766</b>	$d_4, d_{11}, d_{13}, d_{38}, d_{48}, d_{50}, d_{52}$
<b>B. The best <u>ordered orthogonalized</u> MR model with 3 significant descriptors</b>							
<i>orthogonalization ordering: <math>d_{38}, d_{13}, d_{50}, d_{11}, d_{48}, d_{52}, d_4</math></i>							
3			<b>0.898</b>	0.359		<b>0.807</b>	$\Omega_{50}, \Omega_{52}, \Omega_4$
<b>C. The best possible <u>nonlinear</u> MR model (with cross-products)</b>							
4	<b>0.942</b>	0.273	<b>0.907</b>	0.346	51.2	<b>0.818</b>	$d_4 \cdot d_{46}, d_5 \cdot d_{50}, d_{11} \cdot d_{51}, d_{38} \cdot d_{47}$
<b>D. NN<sup>2</sup>, GA<sup>3</sup> and PLS<sup>4</sup> models, respectively</b>							
3	<b>0.919</b>		<b>0.866</b>				$d_{27}, d_{38}, d_{50}$
6	<b>0.920</b>		<b>0.849</b>		22.0		$d_4, d_5, d_6, d_{11}, d_{39}, d_{50}$
10	<b>0.910</b>	0.376			24.0	<b>0.694</b>	$d_4, d_5, d_{11}, d_{17}, d_{36}, d_{38}, d_{39}, d_{40}, d_{50}, d_{52}$

<sup>a</sup> Number of descriptors in the model

## CONCLUSIONS

It is evident that the best MR models generated in this report are better and simpler (contain a smaller number of optimized parameters) than those obtained by NNs, GA and PLS. Especially important is the comparison between the nonlinear MR model (Table 1, C) and NN model (Table 1, D), which shows that by application of MR one can obtain better nonlinear QSAR model than with NNs. Additionally, the comparison with GA and PLS techniques shows that the MR based selection of descriptors produces better results.

## REFERENCES

1. D.L. Selwood, D.J. Livingstone, J.C.W. Comley, A.B. O'Dowd, A.T. Hudson, P. Jackson, K.S. Jandu, V.S. Rose, and J.N. Stables, Structure-activity relationships of antifoliar antimycin analogues: a multivariate pattern recognition study, *J. Med. Chem.* 33:136 (1990).
2. S.-S. So and M. Karplus, Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural network, *J. Med. Chem.* 39:1521 (1996).
3. D. Rogers and A.J. Hopfinger, Application of genetic function approximation to quantitative structure-activity relationships and structure-property relationships, *J. Chem. Inf. Comput. Sci.* 34:854 (1994).
4. H. Kubinyi, Evolutionary variable selection in regression and PLS analyses, *J. Chemometrics* 10:119 (1996).
5. B. Lučić, S. Nikolić, N. Trinajstić, and D.; Juretić, The structure-property models can be improved using the orthogonalized descriptors, *J. Chem. Inf. Comput. Sci.* 35:532 (1995).
6. B. Lučić and N. Trinajstić, New developments in QSPR/QSAR modeling based on topological indices, *SAR and QSAR in Environ. Res.* 7:45 (1997).

## STRUCTURE - ACTIVITY RELATIONSHIPS OF NITROFURAN DERIVATIVES WITH ANTIBACTERIAL ACTIVITY

José Ricardo Pires,<sup>1</sup> Astréa Giesbrecht,<sup>2</sup> Suely L. Gomes,<sup>3</sup> and Antonia T. do-Amaral<sup>1</sup>

<sup>1</sup> Departamento de Química Fundamental, Instituto de Química

<sup>2</sup> Departamento de Farmacologia, Instituto de Ciências Biomédicas

<sup>3</sup> Departamento de Bioquímica, Instituto de Química

Universidade de São Paulo

Caixa Postal 26077, 05599-970 São Paulo, Brazil

### INTRODUCTION

The antimicrobial activity of nitroaromatic compounds requires and is related to an enzymatic reduction of the nitro group *in vivo*, yielding toxic species<sup>1</sup>. In the present work, QSAR analysis of nine 5-X-substituted 2-(5-nitro-2-furfurilidene)-3-oxo-2,3-dihydrobenzofuranes, (*set I*), were performed in order to gain an insight into their physico-chemical features which describe the antibacterial activities, evaluated for a Gram-positive and a Gram-negative bacteria: *Staphylococcus aureus* (ATCC-25923) and *Caulobacter crescentus* (NA 1000), respectively. In addition, five 5'-X-substituted 1-(2-hydroxy-phenyl)-3-(5-nitro-2-furyl)-2-propen-1-ones, (*set II*), and their corresponding acetylated analogs, (*set III*), were included in the analysis to verify the role of the benzofuran ring on the activity.

### MATERIALS AND METHODS

All the compounds (*set I*: X = -H; -CH<sub>3</sub>; -C<sub>2</sub>H<sub>5</sub>; -*n*-(CH<sub>2</sub>)<sub>2</sub>CH<sub>3</sub>; -Cl; -Br; -OCH<sub>3</sub>; -CN and -NO<sub>2</sub>; *sets II* and *III*: X = -H; -CH<sub>3</sub>; -C<sub>2</sub>H<sub>5</sub>; -Cl and -NO<sub>2</sub>) were prepared by methods found in literature and identified by their <sup>1</sup>H-NMR and <sup>13</sup>C-NMR spectra. The electronic, lipophilic, molar refractivity-related and steric descriptors used in the analysis were, respectively: E, the reduction potential measured by cyclic voltametry; E<sub>LUMO</sub>, the energy of the lowest unoccupied molecular orbital, calculated by AM1 using MOPAC 6.0 (QCPE); σ Hammett or Σ and ℞ Swain-Lupton electronic substituent constants found in literature<sup>2</sup>; log P<sub>o/w</sub>, the logarithm of partition coefficient, obtained either by RP-HPLC<sup>3</sup> measurements or by CLOGP, v 1.0.0., Biobyte, Corp. (kindly provided by A.Leo, Pomona College); π and MR substituent constants found in literature<sup>2</sup> and V, the molecular volume calculated by Sybyl 6.4. The indicator variable, I<sub>(ab)</sub>, has been introduced to indicate the presence (I=0) or absence (I=1) of the benzofuran ring in the structure. The biological parameter was chosen to be IC<sub>50</sub>. Besides, for *Caulobacter crescentus* the reduction of the nitrofuran derivatives by NADPH was

studied on aerobic conditions and catalysed by non-purified extracts. The data obtained were analysed using traditional QSAR, using the BILIN-program (kindly provided by H. Kubinyi, BASF-AG, Ludwigshafen). All the molecular modelling approaches were performed with Sybyl, 6.4. (Tripos Ass.) on an IRIS 2 - R10000.

## RESULTS AND DISCUSSION

For set I, the QSAR models indicate a linear and significative negative contribution of the electronic term (expressed either by  $\sigma_p$ , E or by  $\mathfrak{I}$  and  $\mathfrak{R}$ ) for the antibacterial activity evaluated against *S. aureus*, equation (1). Similar models have been derived for *C. crescentus*.

$$\log(1/IC_{50}) S. aureus = -1.304 (\pm 0.42) \sigma_p + 3.427 (\pm 0.17) \quad (1)$$

$n = 9, r = 0.941, s = 0.189, F = 54.062, r_{CV}^2 = 0.829, s_{PRESS} = 0.231$

When the analysis is extended to more flexible analogs, the QSAR models reveal that the antibacterial activity, against *S. aureus*, is mainly described by two factors: the electronic one (expressed by either the  $\mathfrak{I}$  and  $\mathfrak{R}$  constant or by E) and the steric and/or conformational one (expressed by the indicator variable  $I_{ab}$ ), equation (2). For *C. crescentus*, similar models have been derived. It was also observed that the increase of antibacterial activity is followed by a decrease of the nitrofurantoin derivative reduction maximum rate.

$$\log(1/IC_{50}) S. aureus = -0.8 (\pm 0.4) \mathfrak{I} - 1.6 (\pm 0.7) \mathfrak{R} - 1.0 (\pm 0.2) I_{ab} + 3.3 (\pm 0.2) \quad (2)$$

$n = 19; r = 0.970; s = 0.197; F = 78.429; r_{CV}^2 = 0.905; s_{PRESS} = 0.246$

The lower activities (~ 10 times) observed in sets II and III, when compared with set I, for *S. aureus* and *C. crescentus*, could be explained by conformational requirements not fulfilled for the former ones, considering that compounds in three sets with common substituents have only slightly varying reduction potential as well as the other studied physico-chemical descriptors, when they were analysed in pairs. The obtained MSA and CoMFA models reveal important structural features influencing the antibacterial activity and allow us to draw a physical interpretation of the indicator variable  $I_{ab}$  derived by the traditional QSAR models.

## ACKNOWLEDGEMENTS

FAPESP (Brazil); CNPq (Brazil) and DFG/DAAD (Germany) provided financial support.

## REFERENCES

1. D. I. Edwards, DNA binding and nicking agents, in: *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*, P.G. Sammes, ed, Pergamon Press, Oxford (1990).
2. C. Hansch, A. Leo and D. Hoekman. *Exploring QSAR: Hydrophobic, Electronic and Steric Constants*, ACS Professional reference Book, ACS, Washington (1995).
3. A. T-do Amaral, A.C. Oliveira, R. Neidlein, M. Gallacci, L. Caprara and Y. Miyazaki, Physicochemical parameters involved in the lethal toxicity of N,N-[(dimethylamino) ethyl]4-substituted benzoate hydrochlorides: a QSAR study, *Eur. J. Med. Chem.* 32:433 (1997).

## QSAR APPROACH FOR THE SELECTION OF CONGENERIC COMPOUNDS WITH SIMILAR TOXICOLOGICAL MODES OF ACTION

Paola Gramatica<sup>1</sup>, Federica Consolaro<sup>1</sup>, Marco Vighi<sup>2a</sup>, Roberto Todeschini<sup>2b</sup>, Antonio Finizio<sup>2a</sup> and Michael Faust<sup>3</sup>

<sup>1</sup>Dep. Structural and Functional Biology.  
QSAR Research Unit (MI Chemometrics Group),  
Milano University, via Ravasi 2 - 21100 Varese (Italy)

<sup>2</sup>Dep. Environmental Sciences, Milano University  
Via Emanuelli, 15 - 20126 Milano (Italy)

<sup>a</sup>Ecotoxicology Group; <sup>b</sup>MI Chemometrics Research Group

<sup>3</sup>Institute of Cell Biology, Biochemistry and Biotechnology  
University of Bremen, Leobener Strasse - 28359 Bremen (Germany)

Preliminary Principal Component Analysis (PCA), MultiDimensional Scaling (MDS) and cluster analyses with different linkages and distances were performed on 31 triazines and 27 urea compounds (23 phenylureas *sensu stricto* and 4 similar compounds) using several molecular descriptors (structural, topological, 3D-WHIM<sup>1</sup>).

PCA shows that the second component (PC2) separates quite well triazines from phenylureas highlighting the capability of this approach to distinguish among these two different classes of chemicals. These results were confirmed using the two other approaches.

QSAR models were then developed on toxicity data on algae, available for 15 phenylureas and 18 triazines, using the whole set of 168 descriptors and the Genetic Algorithms approach to select the most relevant variables. The predictive capability of all models was tested by means of the *leave-one-out* and *leave-more-out* procedures with good results ( $Q^2=92\%$  and  $87\%$ ).

PCA, MDS and cluster analysis were finally performed using the independent variables of the best toxicity-models, allowing the highlighting of differences and similarities among substances, based on parameters significant in describing the toxic effect. By this approach, groups of similar compounds, with the same toxicological mode of action, were selected, in order to plan experimental assays to confirm the concept of additivity in the toxicity of mixtures.

### REFERENCES

- 1) R.Todeschini and P.Gramatica, "3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and Chemical Meaning of WHIM Descriptors", *Quant. Struct.-Act. Relat.* **16**, 113-119 (1997).



# STRATEGIES FOR SELECTION OF TEST COMPOUNDS IN STRUCTURE-AFFINITY MODELLING OF ACTIVE CARBON ADSORPTION PERFORMANCE: A MULTIVARIATE APPROACH

L-G. Hammarström<sup>1</sup>, I. Fångmark<sup>1</sup>, P.G. Jönsson<sup>1</sup>, P.R. Norman<sup>2</sup>,  
A.L. Ness<sup>2</sup>, S.L. McFarlane<sup>2</sup> and N.M. Osmond<sup>2</sup>

<sup>1</sup>FOA, Division of NBC Defence, SE-90182 Umeå, Sweden,  
<sup>2</sup>DERA, CBD Porton Down, Salisbury, United Kingdom.

## INTRODUCTION

Active carbon is the most universal adsorbent for organic vapours. A study with the following aims was initiated to aid in the design of optimum performance carbon filters:

- To identify important physical and chemical properties influencing the adsorption, and thus gain a better understanding of the adsorption process.
- To investigate the possibilities of developing predictive tools for filter performance under dry and humid conditions based on the physico-chemical properties of the adsorbate.

Most earlier models that account for carbon capacity and break-through times were derived from adsorption isotherms and kinetic equations. Carbon capacity, and affinity parameters ( $k$  or  $\beta$ ), have also been modelled by structure and property related descriptors.<sup>1-4</sup> Many previously developed models have involved compounds with a limited structural variation, i.e. homologous series.<sup>2,5</sup>

The first part of this work focuses on strategies for selection of training-sets. A totally empirical multivariate data analysis model, unbiased by physical laws, was developed.

## METHOD

Experimental filter performance data for 31 chlorinated hydrocarbons was selected for the study.<sup>6-8</sup> The selection of a training-set was made based on different variations in physical properties of the adsorbate. The following 22 compound properties, were used as descriptors:

Molecular weight (Mw)	Heat capacity (Cp)	Log solubility in water (logS)
Density (D)	Diffusion coefficient (Diff)	Log o/w partition coeff. (logP)
Boiling point (Bp)	Surface tension (St)	-Log Henry's law const. (pHL)
Melting point (Mp)	Viscosity constant (Vc)	Vdw interaction (graphite) (Eint)
Critical temperature (Tc)	Ionisation potential (Ip)	Molar volume (Mvol)
Critical pressure (Pc)	Refractive index (nD)	Molar refractivity (Mref)
Log vapour pressure (lgPv)	Dipole moment (Dipm)	
Heat of vaporisation (dHv)	Dielectric constant (Diel)	

The strategy was to select a minimum number of test compounds for modelling (a training-set); using the remaining compounds for model validation. Selection was based on principal component analysis<sup>9</sup> of the descriptor data set. Three significant components (principal properties, PPs) were obtained accounting for 87.8 % of the variation in data, according to the cross-validation criterion.

To investigate the influence of the selection of training-sets on the predictive power of the model, the following strategies were used: Fractional factorial design ( $2^{3-1}$ ) with two centre points (Model 1), the use of a homologous series, chloromethane to chlorohexane (Model 2), and a selection of six compounds with limited variation in the second PP (Model 3). Partial least squares (PLS)<sup>9</sup> models were calculated to establish a correlation between break-through data and the PPs for each of the training-sets. A prediction was done for the whole data set based on the three models (Table 1).

## RESULTS AND DISCUSSION

The ability to predict break-through data is good and similar for Model 1 and 3 due to the dominant influence of the first principal component on break-through characteristics. The volume of liquid adsorbed at 100 % break-through is very well predicted for medium volatile compounds when using factorial design. A selection based on a homologous series shows very bad performance.

Inspection of the PLS loading plot shows, as expected, that break-through time and adsorbed volume are highly influenced by the first principal component. Descriptors related to volatility (Bp, lgPv, Tc, dHv, Cp., etc.), molecular size (Mw, Mref and Mvol), other descriptors for intermolecular interactions (Eint, Vc, St, and Mp) as well as hydrophilicity (lgS and logP) are major contributors. Density, dipole moment, refractive index, dielectric constant, and ionisation potential dominate the second component.

**Table 1.** Summary of experimental and predicted break-through performance <sup>a</sup>

		t1(min)	t10 (min)	t50 (min)	v1 (ml/g)	v10 (ml/g)	v100 (ml/g)
Exp. <sup>b</sup>	Max	0.05	0.7	3.5	<0.001	0.001	0.009
	Min	110.3	132.3	160.6	0.507	0.564	0.698
Model 1	R <sup>c</sup>	<b>0.945</b>	<b>0.951</b>	<b>0.950</b>	<b>0.954</b>	<b>0.969</b>	<b>0.981</b>
	Dev <sup>d</sup>	10.21	12.51	15.30	0.053	0.043	0.048
Model 2	R	<b>0.233</b>	<b>0.217</b>	<b>0.214</b>	<b>0.362</b>	<b>0.374</b>	<b>0.564</b>
	Dev	63.35	71.64	63.03	0.321	0.347	0.231
Model 3	R	<b>0.923</b>	<b>0.935</b>	<b>0.945</b>	<b>0.908</b>	<b>0.931</b>	<b>0.956</b>
	Dev	10.90	11.53	13.34	0.055	0.056	0.055

<sup>a</sup> Break-through times (t) at 1, 10, and 50 % and adsorbed volume (v) at 1, 10, and 100 % break-through.

<sup>b</sup> Break-through times have been corrected for differences in carbon weight.

<sup>c</sup> Correlation coefficient

<sup>d</sup> Average absolute deviation from experimental value.

## CONCLUSIONS

Predictive structure-affinity models can be established for break-through profiles and filter capacity for halogenated hydrocarbons based on physical properties of the adsorbate, provided that the training-set is selected according to a factorial design in principal properties. Important descriptors are: boiling point, critical temperature, vapour pressure, heat of vaporisation, molecular weight, molar refraction, diffusion coefficient, logP, vdW-interaction energy, hydrophilicity, density, refractive index, dipole moment, and dielectric constant.

## REFERENCES AND NOTES

1. Nirmalakhandan, N.N. and Speece, R.E. *Environ. Sci. Technol.* **27**, 1512 (1993).
2. Prakash, J., Nirmalakhandan, N. and Speece, R.E. *Environ. Sci. Technol.* **28**, 1403 (1994).
3. Urano, K., Shigeaki, O. and Yamamoto, E. *Environ. Sci. Technol.* **16**, 10 (1982).
4. Wood, G.O. *Carbon* **30**, 593 (1992).
5. Harrison, B.H. and Narayan, S.B. Conference proceedings, CRDEC-SP-034, November 1990.
6. Nelson, G.A. and Harder, C.A. *Am. Ind. Hyg. Assoc. J.* **35**, 391 (1974).
7. Yoon, Y.H. and Nelson, J.H. *Am. Ind. Hyg. Assoc. J.* **53**, 303 (1992).
8. Compounds 20-33, 36-41, and 43-53 were used in the study.
9. Calculations were performed using Simca-S 6.0 from Umetri AB, Umeå, Sweden.

## DESIGN AND QSAR OF DIHYDROPYRAZOLO[4,3-c]QUINOLINONES AS PDE4 INHIBITORS

M. López, V. Segarra, M. I. Crespo, J. Gràcia, T. Doménech,  
J. Beleta, H. Ryder and J. M. Palacios

Almirall Prodesfarma, Research Centre, Cardener 68-74,  
08024, Barcelona (Spain)

### INTRODUCTION

The interest in therapeutic utility of PDE inhibitors is mostly focused on new agents which selectively inhibit the PDE4 family<sup>1</sup>. This subtype of phosphodiesterase is found in both respiratory smooth muscle and circulatory inflammatory cells. Its inhibition causes relaxation of the former as well as inhibition of the inflammatory response of the latter. Selective PDE 4 inhibitors, lacking adverse effects such as emesis, have potential utility in asthma therapy<sup>2</sup>.

Following our initial strategy based on the pharmacophore of compounds structurally related to niraquazone<sup>3</sup>, new series of 2,5-dihydro[4,3-c]pyrazoloquinolin-3-ones (DHPQ) have been designed and synthesized. The synthesis, SAR, and the antiasthmatic potential of these new PDE 4 inhibitors has been recently described<sup>4</sup>.

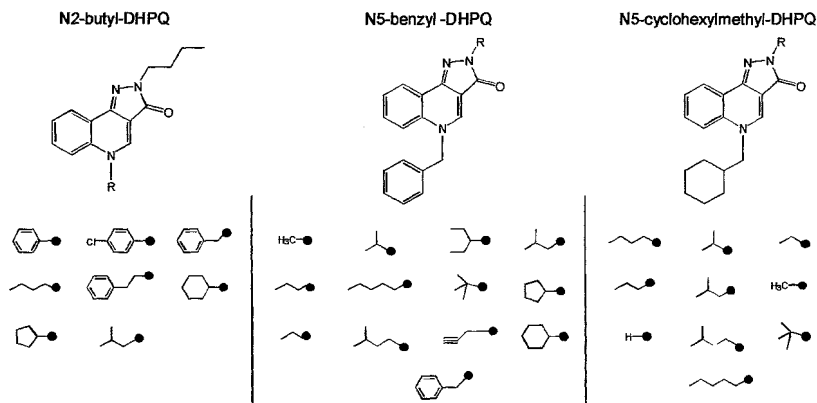
### METHODOLOGY

The Molecular graphics studies were carried out on an Alpha station 3000 using Chem-X software (Chemical Design Ltd, Oxford). Compounds included in this study were constructed of small fragments from Chem-X library. All structures were initially optimized using steepest descents and conjugate gradient methods. Charge distributions were calculated after semiempirical optimization using the MOPAC/AM1 method<sup>5</sup> of version 6.00.

QSAR studies were carried out using ChemStat module of Chem-X. The 2D descriptors, such as logP, Van der Waals volume, molar refractivity and Verloop steric parameters<sup>6</sup>, were automatically assigned by the Substituent Database module. Quantum mechanic descriptors such as atom charges, frontier orbitals coefficients (HOMO and LUMO), and superdelocalisabilities were calculated from the AM1 method. The LogP parameter was calculated by ChemLogP module that implements the method of Suzuki and Kudo<sup>7</sup>.

## RESULTS AND DISCUSSION

Several substitutions have been explored around the DHPQ moiety (Figure 1) obtaining three related chemical families: fixing an n-butyl at N<sup>2</sup> (N<sup>2</sup>-butyl-DHPQ) or fixing at the N<sup>5</sup> position a benzyl group (N<sup>5</sup>-benzyl-DHPQ) or a cyclohexylmethyl group (N<sup>5</sup>-cyclohexylmethyl-DHPQ).



**Figure 1.** Substitutions studied around the DHPQ moiety

In order to investigate which kind of substitution may enhance PDE4 inhibition, a QSAR study around the DHPQ moiety, has been carried out selecting several classical QSAR and quantum mechanical descriptors (see methodology section). Steric and lipophilic properties have shown good correlation with activity. No correlation was obtained with electronic and quantum parameters.

The N<sup>2</sup>-butyl DHPQ series (eq. 1) shows that activity depends on the length (L) and size (B1) of the substituents.

$$\text{Log}(1/IC_{50}) = -0.4037 L + 0.7607 B1 + 6.0334 \quad n=8 \quad R^2=0.96 \quad (1)$$

The activities of the N<sup>5</sup>-benzyl (eq. 2) and N<sup>5</sup>-cyclohexylmethyl series (eq. 3) can be expressed in terms of steric (B3) and lipophilic (Log P) parameters.

$$\text{Log}(1/IC_{50}) = 0.2976 B3 - 0.5636 (\text{LogP})^2 + 3.116 \text{LogP} + 0.8425 \quad n=13 \quad R^2=0.89 \quad (2)$$

$$\text{Log}(1/IC_{50}) = 0.3926 B3 - 0.4346 (\text{LogP})^2 + 3.360 \text{LogP} - 1.1408 \quad n=10 \quad R^2=0.81 \quad (3)$$

## REFERENCES

1. J.M.Palacios, J.Beleta and V.Segarra, *Il Farmaco*, 50: 819 (1995)
2. M.A.Giembycz, *Trends in Pharm.Sci.*, 1996, 17, 331 (1996)
3. M.López, V.Segarra, M.I.Crespo, J.Beleta and J.M.Palacios, *XIVth International Symposium on Medicinal Chemistry*. Maastricht, P 5.36 (1996).
4. Ll.Pagès, M.I.Crespo, J.Gràcia, C.Puig, V.Segarra, J.Bou, A.G.Fernández, T.Doménech, J.Beleta, H.Ryder and J.M.Palacios, *26th National Medicinal Chemistry Symposium*. Richmond Virginia, Abstract C-12 (1998).
5. J.Stewart, *QCPE Bulletin*, 9(1):10 (1989).
6. A.Verloop and W. Hoogenstraaten, *Med. Chem. (Academic)*, 11:165 (1976).
7. T. Suzuki and Y. Kudo, *J. Comput.-Aided Mol. Des.*, 4:155 (1990)

## QSAR BASED ON BIOLOGICAL MICROCALORIMETRY:

### On the study of the interaction between hydrazides and *Escherichia coli* and *Saccharomyces cerevisiae*

M. L. C. Montanari,<sup>1</sup> A. E. Beezer,<sup>2</sup> and C. A. Montanari<sup>1\*</sup>

<sup>1</sup>Núcleo de Estudos em Química Medicinal-NEQUIM. Departamento de Química - Universidade Federal de Minas Gerais - Campus da Pampulha - 31270-901 - Belo Horizonte - MG - Brazil

<sup>2</sup>Chemical Laboratory, University of Kent at Canterbury, CT2 7NH, UK

## INTRODUCTION

QSAR studies may rely upon the correctness of *quantitative* measurement of drug potencies, that generally starts with *in vitro* screening.<sup>1-3</sup> The screening of drugs using *biological microcalorimetry*<sup>4-6</sup> to derive quantitative biological potency values is a powerful tool for such studies. Yet, the lipophilicity measurement through partition coefficient, using the diffusion process of Taylor-Aris,  $\log P_{TA}$ , in the same cells used for the biological screening via biological microcalorimetry<sup>7</sup> has been carried out. Thus, in this paper we show an established QSAR between hydrazide potencies against *Escherichia coli* and *Saccharomyces cerevisiae* and  $\log P_{TA}$ .

## RESULTS AND DISCUSSION

The equations (1) and (2) state that there is a linear relationship between  $\log 1/D_{50}$  and  $\log P_{TA}$  for *Saccharomyces cerevisiae* and *Escherichia coli*. A negative slope for SARs involving *S. cerevisiae* has not been found before, but it is common for *Escherichia coli*. It appears, possibly, that a hydrophilic interaction, instead of a hydrophobic one, could play a role in the partitioning process.<sup>8,9</sup>

**Linear dependence of  $\log 1/D_{50}$ , for *Saccharomyces cerevisiae*, versus  $\log P_{TA}$ .**

$$\log 1/D_{(50) S.c} = -1.223 (\pm 0.67) \log P_{TA(S.c)} + 2.673 (\pm 0.49) \\ (n = 8; \quad r = 0.878; \quad s = 0.141; \quad F = 20.147; \quad r^2_{cv} = 0.532) \text{ (Equation 1)}$$

**Linear dependence of  $\log 1/D_{50}$ , for *Escherichia coli*, versus  $\log P_{TA}$ .**

$$\text{Log} 1/D_{(50) E.c} = -1.939 (\pm 1.03) \text{Log} P_{(TA)(E.c)} + 1.468 (\pm 0.96) \\ (n = 8; \quad r = 0.883; \quad s = 0.185; \quad F = 21.132; \quad r^2_{cv} = 0.535) \text{ (Equation 2)}$$

It is worthwhile recalling that the same set of compounds was used to derive the QSAR models presented in this paper. This prompted us to undertake a further development that is related to the relationship between the cell systems themselves. This is simply done by correlating the potencies for both cell systems, and Equation 3 shows the result.

#### Extrathermodynamic correlation between *E. coli* and *S. cerevisiae*

$$\log 1/D_{50}(Sc) = 0.489(\pm 0.16)\log 1/D_{50}(Ec) + 2.065(\pm 0.49)$$

(n = 8, r = 0.951, s = 0.093, F = 56.44, r<sup>2</sup><sub>cv</sub> = 0.788) (Equation 3)

Equation 3 states the existence of an extrathermodynamic relationship in the antimicrobial activity between the same series of compounds but different cellular systems, as taken from biological microcalorimetry.

### CONCLUSIONS

For the first time we have shown that log P<sub>TA</sub> and biological microcalorimetry can be used to derive QSARs. This seems to be a good alternative to the octanol/water system largely because the cell suspension is more "real" - a better representation of a natural system, and microcalorimetry is a promising tool for such QSAR studies. Overall, biological microcalorimetry is efficient, fast, and reproducible to better than 3%. It can be used instead of other techniques like agar diffusion or tube assays (serial dilution). *In vitro* screening can be performed in complex and defined medium using frozen cells. Calorimetric output can reveal biocide and biostatic compounds directly, and this is very important in order to control drug doses.

### ACKNOWLEDGMENTS

We would like to thank the following Brazilian Granting Agencies for supporting this research work: CAPES, CNPq, FAPEMIG, FINEP.

### REFERENCES

1. C. A. Montanari, *Química Nova*, 18:56 (1995).
2. C. A. Montanari, A. E. Beezer, J. P. B. Sandall, M. L. C. Montanari, J. Miller and A.M. Giesbrecht, *Rev. Microbiol.*, 23:274 (1992).
3. C. A. Montanari, M. L. C. Montanari, A. E. Beezer and A. M. Giesbrecht, *Química Nova*, 16:133 (1993).
4. M. L. C. Montanari, A. E. Beezer, J. P. B. Sandall and C. A. Montanari, *Int. J. Pharm.*, 85:199 (1992).
5. A. E. Beezer, J. C. Mitchell, R. M. Colegate, D. J. Scally, L. J. Twyman and R. J. Wilson, *Thermochimica Acta*, 250:277 (1995).
- 6 P. L. O. Volpe and C. A. Montanari, *Química Nova*, 20:125 (1997).
- 7 M. L. C. Montanari, C. A. Montanari, D. Piló-Veloso, A. E. Beezer, J. C. Mitchell and P. L. O. Volpe, *Quant. Struct-Act. Relat.*, 17:102 (1998).
- 8 T. Fujita, *Comprehensive Medicinal Chemistry, The Rational Design, Mechanistic Study & Therapeutic Applications of Chemical Compounds*, C. A. Ramsden (Ed.), Vol. 4, Pergamon Press, New York, p. 497 (1990).
- 9 J. K. Seydel, K.-J. Schaper, E. Wempe, and H. P. Cordes, *J. Med. Chem.* 19:483 (1976).

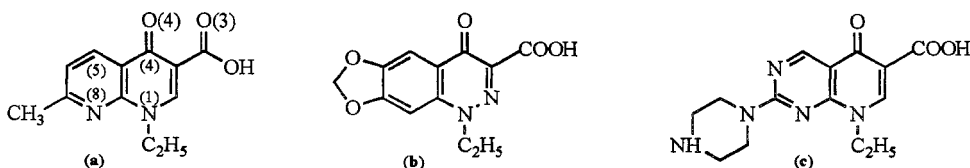
**CINNOLINE ANALOGS OF QUINOLONES:  
STRUCTURAL CONSEQUENCES OF THE N ATOM  
INTRODUCTION IN THE POSITION 2**

Marek L. Główka, Dariusz Martynowski, Andrzej Olczak, Alina Staszewska

Institute of General and Ecological Chemistry, Technical University of Łódź,  
ul. Żwirki 36, 90-924 Łódź, e-mail: marekgl@ck-sg.p.lodz.pl

**INTRODUCTION**

According to a mechanism of inhibition of DNA-gyrase complex by quinolones proposed by Shen [1], the drug binds guanine base of a single stranded bacterial DNA. One of the consequences of this model is that binding energy depends on negative partial charges on O(4) and O(3) atoms (Fig.1), i.e. the greater negative charges the stronger will be the hydrogen bonds with guanine. We assume that introduction of N atom(s) into aromatic rings system of quinolones should affect partial charges on the O(4) and O(3). As positions 3,4 and 7 in quinolones are usually substituted, the remaining positions for N substitution are 2,5,6 and 8 (Fig.1). Some of them exert significant effects on structural and electronic properties of the quinolone analogs.



**Figure 1.** Representatives of prominent analogs of quinolones used as antibacterial agents: nalidixic acid (a), cinoxacin (b) and pipemidic acid (c). Also shown is numbering system.

**Methods.** Calculations of the partial charges were performed with MOPAC (AM1) on an SGI computer for two carboxyl conformations; one with *intramolecular* hydrogen bond O(4)...H-O and the other one with parallel orientations of keto groups. Starting parameters were taken from crystal structures of nalidixic acid, pipemidic acid, cinoxacin (from literature) and other cinnoline analogs, synthesized by Dr Stańczak (Medical Academy, Łódź) and determined in our laboratory.

## RESULTS AND DISCUSSION

**Table 1.** Changes ( $\Delta$ ) in partial charges at O(4) and O(3) atoms (x100) resulting from introduction of additional N heteroatom(s) in quinolone ring system. Parent quinolone is 1-methyl-4-oxo-1,4-dihydro-3-quinolinecarboxylic acid

Conformation of carboxylic group	Parent quinolone	N2	N5	N6	N8	N2 N5	N2 N6	N2 N8	N2 N5 N6	N2 N5 N8	N2 N6 N8
with intramolecular hydrogen bond	O(4) -0.35	3	4	0	0	8	4	3	9	8	4
with parallel keto groups	O(3) -0.35	6	0	0	0	7	7	6	7	7	7
with intramolecular hydrogen bond	O(4) -0.29	1	5	1	0	5	1	1	2	6	6
with parallel keto groups	O(3) -0.33	0	1	1	0	1	1	1	1	1	1

The introduction of additional N heteroatom(s) into positions 2 and 5 of the quinolone ring system **decreases** binding energy of quinolones with DNA. The introduction of N atom(s) into positions 6 and/or 8 does not change partial charges on O(4) and O(3), which participate in hydrogen bonds with DNA. The changes are significant in case of the O(4) atom and may accumulate up to 25% of the partial charge.

**This explains generally lower antibacterial activity of cinnoline analogs of quinolones.**

**Table 2.** Changes ( $\Delta$ ) in partial charges at O(4) and O(3) atoms (x100) resulting from typical substitutions in positions 6 and 7 of quinolone antibacterials

Additional N atom in positions	Conformation of carboxylic group	Parent quinolone	6-F	7-N(CH <sub>3</sub> ) <sub>2</sub>	6-F, 7-N(CH <sub>3</sub> ) <sub>2</sub>
N2	with intramolecular hydrogen bond	O(4) -0.35	0	-2	-1
	with parallel keto groups	O(3) -0.35	0	-1	0
	with intramolecular hydrogen bond	O(4) -0.29	1	-1	0
	with parallel keto groups	O(3) -0.33	0	0	0
	with intramolecular hydrogen bond	O(4) -0.32	1	-1	-1
	with parallel keto groups	O(3) -0.29	0	0	-1
N2	with parallel keto groups	O(4) -0.28	2	-1	-1
	with parallel keto groups	O(3) -0.33	0	0	0

Typical fluoroquinolones, i.e. 6-F, 7-amine derivatives, are characterized by a slight increase (or invariability) of partial charges at O(4) and O(3) atoms as compared with those in **unsubstituted** quinolones. Particularly important is amine-type substituent at the 7 position, which always generates additional negative charge at O(4).

**This agrees with the observation that 7-amine group** (also morpholine, piperidine, piperazine, pyrrolidine) **is advantageous for antibacterial activity.**

### Acknowledgments

The authors thank the Polish State Committee for Scientific Research for financial support under the project 4.P05F.008.09.

### REFERENCES

1. L.S. Shen, L.A. Mitscher, P.N. Sharma, T.J. O'Donnell, D.W.T. Chu, C.S. Cooper, T. Rosen, and A.G. Pernet, Mechanism of inhibition of DNA gyrase by quinolone antibacterials: a cooperative drug-DNA binding model, *Biochemistry* 28:3886 (1989).



# JOINT CONTINUUM REGRESSION FOR ANALYSIS OF MULTIPLE RESPONSES

Martyn G. Ford<sup>a</sup>, David W. Salt<sup>ab</sup> and Jon Malpass<sup>a</sup>

<sup>a</sup> Centre for Molecular Design

<sup>b</sup> School of Computer Science and Mathematics  
University of Portsmouth  
Portsmouth, PO1 2EG, UK

## INTRODUCTION

The rationale behind developing a multiple response algorithm for continuum regression (CR) is to provide the user with a method of investigating how any number of responses change simultaneously given one set of physico-chemical properties. The background behind multiple response algorithms is well documented with such algorithms available for ordinary least squares regression (OLS) often referred to as multivariate linear regression (MVLR), partial least squares regression (PLS), sometimes known as PLS2 and principal components regression (PCR). Furthermore, an algorithm has been proposed by Brooks and Stone [1994], named joint continuum regression (JCR), which maintains a number of the properties of their formulation of the single response continuum regression [Stone and Brooks, 1990].

This report details the development of a multiple response continuum regression algorithm that maintains the pertinent features of the Portsmouth formulation of continuum regression [Malpass *et al*, 1995]. The report addresses the algebra behind the method, highlights the equivalence with other methods and illustrates the utility of the multiple response algorithm.

## The Portsmouth Formulation of Joint Continuum Regression

The strategy adopted for developing a Portsmouth formulation of JCR followed the approach adopted when CR-P was developed, *viz.* to maintain the essential structure of the GCF and the equivalence with MLR, PLS and PCR for  $\alpha = 0, 0.5$  and  $1$  respectively.

This can be achieved by taking the generic GCF

$$T = (\mathbf{c}'\mathbf{E}\mathbf{c})^{f(\alpha)}(\mathbf{c}'\mathbf{S}\mathbf{c})^{g(\alpha)} \quad (1)$$

and then to reformulate the power terms,  $f(\alpha)$  and  $g(\alpha)$ . Initially, the two terms used in CR-P were adopted, *i.e.*

$$f(\alpha) = 2 + 2\alpha - 4\alpha^2, \quad (2)$$

$$g(\alpha) = -1 + 2\alpha. \quad (3)$$

However, the GCF-SB comprises a 'covariance' ( $\mathbf{c}'\mathbf{E}\mathbf{c}$ ) term which is formulated such that the new component is squared, *i.e.* the value of  $f(\alpha)$  for JCR needs to be 1 to achieve equivalence with its value in CR-SB, where it is fixed at the value of 2. If JCR-P is to be formulated such that the structure of the generic GCF is retained then the value of  $f(\alpha)$  of equation 2.

In the CR-P GCF the power of the covariance term varies continuously with  $\alpha$ , so that it takes the value 2 for MLR and PLS and 0 for PCR. This means that we cannot directly apply the same formulation of the power terms of CR-P to JCR-P. However, if a common factor of 2 is taken out of equation 2 it is possible to achieve the necessary functions, *i.e.*  $f(\alpha)$  becomes  $f(\alpha) = 2(1 + \alpha - 2\alpha^2)$ . By considering the generic GCF with this power, it can be seen that the common factor of 2 is already accounted for by the covariance factor,  $\mathbf{c}'\mathbf{E}\mathbf{c}$ . Hence, the Portsmouth formulation of Joint Continuum Regression can be achieved by using the two powers

$$f(\alpha) = 1 + \alpha - 2\alpha^2 \quad (4)$$

$$g(\alpha) = -1 + 2\alpha \quad (5)$$

so yielding the alternative GCF

$$T = (\mathbf{c}'\mathbf{E}\mathbf{c})^{(1+\alpha-2\alpha^2)} (\mathbf{c}'\mathbf{S}\mathbf{c})^{(-1+2\alpha)} \quad (6)$$

The formulation was implemented as a SAS macro and validated using simulated data sets generated to give all accessible combinations of low ( $r=0.1$ ), medium ( $r=0.5$ ) and high ( $r=0.8$ ) correlation between the responses ( $y_s$ ), the predictors ( $x_s$ ) and their associations ( $y_x$ s). The results suggest that the Portsmouth formulation of Joint Continuum Regression yields reliable prediction, particularly whenever the associations between the  $y$  and  $x$  blocks are medium to high. Multicollinearities within the responses ( $y_s$ ) and within the predictors ( $x_s$ ) are overcome by the construction of optimised components with maximum values for the criterion function.

## REFERENCES

- Brooks, R. and Stone, M., 1994, Joint Continuum Regression for Multiple Predictands, *J. American Statistical Association*, **89**, 1374-1377
- Malpass J., Salt, D.W., Ford, M.G., Wynn, E.W. Livingstone, D.J., 1995, Continuum regression: A new algorithm for the prediction of biological activity. In *Methods & Principles in Medicinal Chemistry*, 3, Advanced computer assisted techniques in drug discovery. (Ed. H. van de Waterbeemd), 163-189, VCH Publishers, Weinheim.
- Stone, M. and Brooks, R., 1990, Continuum Regression: Cross Validated Sequentially Constructed Prediction Embracing OLS, PLS and PCR, *J. Roy. Statist. Soc. (B)*, **52**, 237-269

## **PUTATIVE PHARMACOPHORES FOR FLEXIBLE PYRETHROID INSECTICIDES**

Martyn G. Ford<sup>1</sup>, Neil E. Hoare<sup>1</sup>, Brian D. Hudson<sup>2</sup> Thomas G. Nevell<sup>1</sup> & John A. Wyatt<sup>3</sup>

<sup>1</sup>Centre For Molecular Design, University of Portsmouth, UK PO1 2QF

<sup>2</sup>GlaxoWellcome, Stevenage, SG1 2NY UK

<sup>3</sup>Tripos Associates, Milton Keynes, UK

### **INTRODUCTION**

In any study of structure/activity relationships based on computational chemistry, it is necessary to postulate an appropriate structure on which the study is to be based. Previous investigations have been based on low energy structures such as the experimentally determined crystal structure of a molecule, or a minimal energy structure calculated using molecular mechanics. Selection of these pharmacophores is somewhat arbitrary. In practice, a larger set of candidate structures should be considered in order to obtain the most appropriate structure. This is particularly true for flexible molecules such as pyrethroids for which a large number of conformations are possible.

The present work aims to study the molecular motions of pyrethroid insecticides using molecular dynamics simulations. The simulations have been partially validated by comparing the preferred orientations of the torsional bonds with the results obtained using different force field (Hudson et. al.). The structures sampled are used to identify sub-sets of conformations for consideration as possible representations of the active conformation. The use of comparative molecular field analysis (CoMFA), which relates the steric and electrostatic properties of molecules to biological activity is investigated as a basis for choosing the most appropriate pharmacophore.

### **CoMFA ANALYSIS OF PYRETHROID INSECTICIDES**

Although many factors are involved in drug-receptor interactions, the steric and electrostatic properties of the ligand are particularly important. The relationship between biological activity and steric and electrostatic effects can be investigated using CoMFA analysis. In the following study, the conformation of pyrethroids has been varied in an attempt to identify possible pharmacophoric structures. In order to achieve this objective, CoMFA was performed on 18 putative pharmacophores: 16 based upon the proposed preferred orientations of deltamethrin and a further two chosen to represent the major

clusters identified by the pooled cluster analysis for 40 pyrethroid insecticides. Each analysis has been performed on one of these putative structures using the 36 compounds for which killing activity was available or the 14 compounds for which knockdown activity was available. For the compounds for which knockdown activity was available, an additional 8 putative pharmacophores derived by including the additional orientation exhibited by T4 (as in QSAR1) were also investigated. This orientation is not accessed by the most active killing compound, deltamethrin, or by other potent Type II pyrethroids which possess an  $\alpha$ -cyano substituent, but which possess poor knockdown activity.

The results of the CoMFA analysis for 36 insecticidal compounds and the 14 knockdown compounds are presented in Tables 3-25 and 3-26. The orientations of the torsion angles T1-T5, the steric and electrostatic contributions to the model, the optimum number of crossvalidated (LOO) R2 are also given.

## PYRETHROID MODE OF ACTION

The dynamic behaviour of a pyrethroid may act (1) to disrupt ordered domains within the bilayer, and/or (2) to induce a more ordered arrangement of phospholipid molecules in the more disordered domains. In this respect, the nature of the pharmacophores proposed for knockdown and killing are of some interest. These structures can interconvert by rotations about T2 and T4: furthermore, there is little evidence of correlated effects between these torsional movements. One consequence of such an interconversion is that the dipole associated with the carbonyl attached to the ester linkage is rotated through approximately 120 degrees.

A study undertaken by Zeneca and reported earlier at an SCI symposium on membranes held in London has shown that in artificial bilayers devoid of sodium channel protein, pyrethroids were able to induce a reversal of dipole potential across the membrane. This perturbation was considered to involve displacement of the  $\beta$ -carbonyl of the glyceryl backbone of the phospholipid components of the membrane. The displacement may be caused by carbonyl-carbonyl dipole repulsion between the pyrethroid and the membrane. This would have two consequences, deformation (disordering) of the membrane and causing the pyrethroid to move through the vacuole created. As the pyrethroid moves, the process may be repeated as further phospholipid carbonyls are encountered and may therefore be a mechanism of transport to a receptor site. This has important consequences for the organisation of the bilayer, since rotations in this region of the phospholipids are known to increase or decrease their packing density within the bilayer and modify the ability of the phospholipid head groups to bind sodium or calcium (Houslay and Stanley). It is interesting to speculate, therefore, that pyrethroids act by inducing changes to the order of the bilayer as a result of interconversion between pyrethroid conformations with different dipole orientations at the carbonyl of the ester. Such interconversions would be expected to result in associated changes in the dipole orientation of the  $\beta$ -carbonyl of phospholipids in order to minimise local dipole interactions. This would have profound effects on many membrane properties including freedom of motion of transverse ion channels such as sodium and calcium channels.

## REFERENCES

- Hudson B.D., George A.R., Ford, M.G., Livingstone, D.J., 1992, The Structure Activity Relationships of Pyrethroid Insecticides. 2. The use of molecular dynamics for conformation searching and average parameter calculation. *J. Comp.-Aided Mol. Design*, 6, 191-201.
- Houslay M.D. and Stanley, K.K., 1982, *Dynamics of Biological Membranes; influence On Synthesis, Structure and Function*, Wiley, New York.

# PREDICTING MAXIMUM BIOACTIVITY OF DIHYDROFOLATE REDUCTASE INHIBITORS

Matevž Pompe<sup>1</sup>, Marjana Novič<sup>2</sup>, Jure Zupan<sup>1,2</sup>, and Marjan Veber<sup>1</sup>

<sup>1</sup>Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia

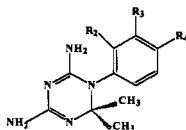
<sup>2</sup>National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

## INTRODUCTION

In the present study the correlation between chemical structures and inhibiting properties of 256 5-phenyl-3,4-diamino-6,6-dimethyldihydrotriazine derivatives, inhibitors of dihydrofolate reductase (DHFR), is investigated. The data-set has been studied by several researches in many different laboratories<sup>1-3</sup>. In the first studies<sup>1</sup>, the linear regression models were tested, and later on artificial neural network models were successfully applied<sup>2,3</sup>. In all mentioned studies, the compounds were represented on the same way, with physicochemical parameters: Hammett's  $\sigma$ , Hansch's  $\pi$  hydrophobicity parameter, molar refractivity MR, and additional indicator variables for presence or absence of specific structural features. In our study the chemical structures were represented by general codes, regardless of presumably important substituents' sites. Molecular descriptors were: topological, geometrical, electrostatic, and quantum-mechanical indices calculated with CODESSA<sup>4</sup> software package, and the "spectrum-like" structure representation<sup>5</sup>.

## RESULTS AND DISCUSSION

The data set comprises 132 compound sub-set tested on DHFR from Walker 256 carcinoma cells and the 113 compound sub-set tested on DHFR from L1210 leukemia tumors. The main skeleton:



is common to all compounds, while the substituents  $R_2$ ,  $R_3$  and  $R_4$  are varied. Eleven compounds contain non-hydrogen  $R_2$  substituent and were also included to the analysis. To

calculate molecular descriptors (topological, geometrical, electrostatic, and quantum-mechanical indices and "spectrum-like" structure representations) the optimised 3D structural co-ordinates and net atomic charges (for minimal energy state) were calculated by MOPAC software package. The descriptors employed in the study contain either the information about the connections between the atoms, symmetry, shape, branching, and cyclicity, or 3D co-ordinates and information about atomic electronic properties.

The whole molecule was translated into a set of different descriptors. Multiple linear regression model (MLR) and counterpropagation artificial neural networks (CP ANN) were used as modelling techniques. The selection of optimal number of structural descriptors was based on the best prediction capabilities of MLR model. The evaluation of prediction capabilities of the developed models was done by ten-fold or leave-one-out cross-validation procedures. At the end our results were compared with the study in which the "spectrum-like" structural code was used for the structural representation. The results of prediction capabilities are gathered in Table 1.

Table 1. Results of the prediction capabilities of different models

	*MLR Str. code 30 descriptors	*MLR indices 30 descriptors	**CP ANN Str. code 30 descriptors	**CP ANN Indices 30 descriptors
<b>r</b>	0.57	0.78	0.56	0.65
<b>b</b>			0.46	0.45
<b>RMS</b>	0.58	0.311	0.74	0.72

\* leave-one-out cross validation procedure

\*\* ten-fold cross validation procedure

From MLR and ANN models it can be concluded, that the large and diverse data-set treated homogeneously can only give satisfactory results if the model is able to organise the data into local sub-models, which would theoretically be able to predict properties of compounds being active on the basis of different reaction mechanism. MLR model does not meet such requirements at all. On the other hand CP ANN has been shown as a powerful grouping tool and their prediction capabilities can be improved by using different optimisation criteria for the selection of best subset of structural descriptors (e.g. genetic algorithm). This part of the research is still not finished.

### Acknowledgement

The financial support of the Ministry of Science and Technology of Slovenia obtained by the Projects: J1 - 8900 and J1 - 0291 is gratefully acknowledged.

### REFERENCES

1. C. Silipo, C. Hansch, Correlation Analysis. Its Application to the Structure-Activity Relationship of Triazines Inhibiting Dihydrofolate Reductase, *J. Am. Chem. Soc.*, **97**, 6849, (1975).
2. T.A. Andrea, H. Kalayeh, Applications of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors, *J. Med. Chem.*, **34**, 2824-2836, (1991).
3. F.R. Burden, B.S. Rosewarne, D.A. Winkler, Predicting Maximum Bioactivity by Effective Inversion of Neural Networks Using Genetic Algorithms, *Chemometrics and Intelligent Laboratory Systems*, **38**, 127-137, (1997).
4. A. R. Katritzky, V. S. Lobanov, M. Karelson, CODESSA 2.0, Comprehensive Descriptors for Structural and Statistical Analysis, Copyright (c) 1994-1996 University of Florida, U.S.A.
5. J. Zupan, M. Novič, General Type of a Uniform and Reversible Representation of Chemical Structures, *Anal. Chim. Acta*, **348**, 409-418, (1997).

## EVALUATION OF CARCINOGENICITY OF THE ELEMENTS BY USING NONLINEAR MAPPING

Alexander A. Ivanov

Zorge St., 36-117, Moscow, 125252, RF

E-mail: aai@aha.ru

### OBJECTIVES

The study was intend to select physicochemical parameters associated with carcinogenic properties of the elements, and to evaluate carcinogenicity for nine elements of the fourth period.

### DATA AND METHODS

The training set contains two groups of the elements. First group contains forty one elements without carcinogenic properties:<sup>1,2</sup> Na, Mg, Al, K, Ca, V, Ga, Br, Rb, Sr, Zr, Nb, Mo, In, Sn, Te, I, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Tb, Dy, Ho, Er, Yb, Lu, Hf, Ta, Re, Os, Ir, Pt, Au, Hg, Tl, Bi. Second group contains 4 elements: Be, Cr, Ni, As. These four elements or some their compounds are human carcinogens.<sup>1,2,3,4</sup> Cr, Ni, As are elements of the fourth period. Carcinogenicity was evaluated for nine elements of the fourth period (Sc, Ti, Mn, Fe, Co, Cu, Zn, Ge, Se).

Nineteen physicochemical parameters were used for selection of the predictors: Atomic radius, covalent radius, electronegativity (Pauling), electronegativity (Allred), electronegativity (Pearson), effective nuclear charge (Slater), effective nuclear charge (Clementi), effective nuclear charge (Froese-Fisher), thermal entropy, heat capacity, density of solid, thermal conductivity, molar volume, coefficient of linear thermal expansion, mass absorption coefficient( $\text{CuK}_\alpha$ ), mass absorption coefficient( $\text{MoK}_\alpha$ ), cross section for the thermal neutrons, electron affinity, ionization enthalpy.

Graph of radar type and nonlinear mapping<sup>5</sup> were used for selection of predictors. Nonlinear mapping was used for the evaluation of carcinogenicity.

Database program chosen for the study is Q&A for DOS (v.4). The data analysis programs are NS for Quattro Pro and SPSS for WIN (v.6.1).

## RESULTS

Atomic radius, covalent radius, thermal entropy, mass absorption coefficient ( $CuK_{\alpha}$ ) are predictors selected for the mapping. These four parameters were used for the mapping (see Figure 1). Every point on the graph corresponds to an element. The closer are the points the more similar are the elements according to four parameters. Ge is in the carcinogenic are. Sc, Ti, Zn are in the are of elements without carcinogenic properties.

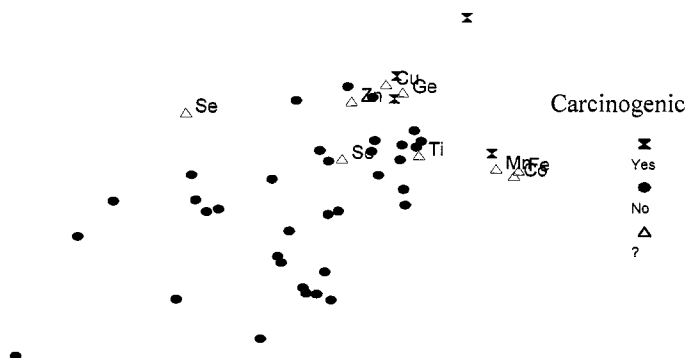


Figure 1. Evaluation of carcinogenicity for nine elements of the fourth period by using nonlinear mapping.

## CONCLUSIONS

1. Atomic radius, covalent radius, thermal entropy, mass absorption coefficient ( $CuK_{\alpha}$ ) are associated with carcinogenic properties of the elements.
2. Ge was considered human carcinogen.
3. Sc, Ti, Zn were considered noncarcinogenic in man.

## REFERENCES

1. J.Emsley. The Elements, Clarendon Press, Oxford (1991).
2. WebElements. <http://www.shef.ac.uk/chemistry/web-elements/>.
3. S.H.Swierenga, J.P.Gilman, J.R.McLean, Cancer risk from inorganics, *Cancer Metastasis Rev.*6:113-154(1987).
4. A.Leonard, G.B.Gerber, P.Jacquet, R.R.Lauwerys, Mutagenicity, carcinogenicity, and teratogenicity of industrially used metals, in: *Mutagenicity, Carcinogenicity, and Teratogenicity of Industrial Pollutants*, M.Kirsh-Volders, ed., Plenum Press, New York (1984).
5. J.W.Sammon, A nonlinear mapping for data structure analysis, *IEEE Trans.Comput.*18:401-409(1969).



**Poster Session II**  
**The Future of 3D-QSAR**

## PARTITION COEFFICIENTS OF BINARY MIXTURES OF CHEMICALS: POSSIBILITY FOR THE QSAR ANALYSIS

Miloň Tichý<sup>1</sup>, Marián Rucki<sup>1</sup>, Václav Bořek Dohalský<sup>1</sup>, Ladislav Feltl<sup>2</sup>

<sup>1</sup>National Institute of Public Health, Šrobárova 48, 10042 Praha 10, Czech Republic

<sup>2</sup>Faculty of Natural Sciences, Charles University, Albertov 2030, 12800 Praha 2, Czech Republic

Biological activity of both individual medical drugs and individual environmental contaminants may change if chemicals in mixtures.

Acute toxicity of several binary mixtures was determined in the whole spectrum of their composition, that is from one pure component to the second pure component: benzene - ethanol (inhibition), benzene - aniline (potentiation), aniline - phenol (additivity) and aniline - nitrobenzene (additivity)

Molar fraction (ratio) was used as a composition descriptor and R-plot for graphical representation of the dependence biological activity - mixture composition.<sup>1, 2, 3</sup> The inhibition of movement of *Tubifex tubifex* worms was measured as the acute toxicity expressed in ED50 (mol/L).<sup>1, 4</sup> The approach was inspired by the Raoult law and its positive and negative deviations in behaviour of mixtures of ideal gases,<sup>5</sup> Loewe and Muschnek isobols<sup>6</sup> and Finney test of additivity.<sup>7</sup>

The results are summarised in the Figs. 1 - 4 for four binary mixtures of volatile organic compounds as indicated. The figures are composed from three sections: A - showing interrelation between composition (expressed as molar fraction R) of gaseous (g) and liquid (l) phases, B - a plot of  $\log R_l/R_g$ , representing a partition of the two compounds between the phases, against the composition of the gaseous phase ( $\log R_g$ ), C - a dependence of the acute toxicity of the mixture on their composition again represented by the molar fraction R concerning one of the compounds (Bz - benzene, An - aniline).

A shape of the plots in the sections A and B indicate a nature of a dependence of the acute toxicity on a mixture composition shown in the section C: as far as the zero axis is crossed by the higher branch of the plot, an inhibition of the two compounds in their mixture occurs (Fig. 1) (it indicates also that the mixture possesses an azeotropic point), if not, a potentiation takes place (Fig. 2). The plots A and B regarding to the distribution of compounds between the phases are not clearly expressed if an additivity exists in acute toxicity of the mixture.

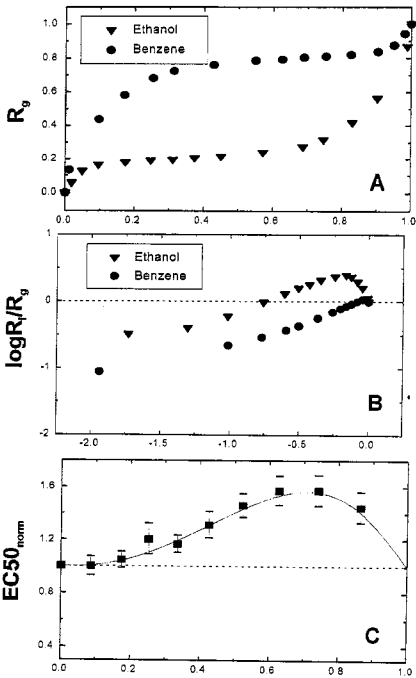


Fig. 1

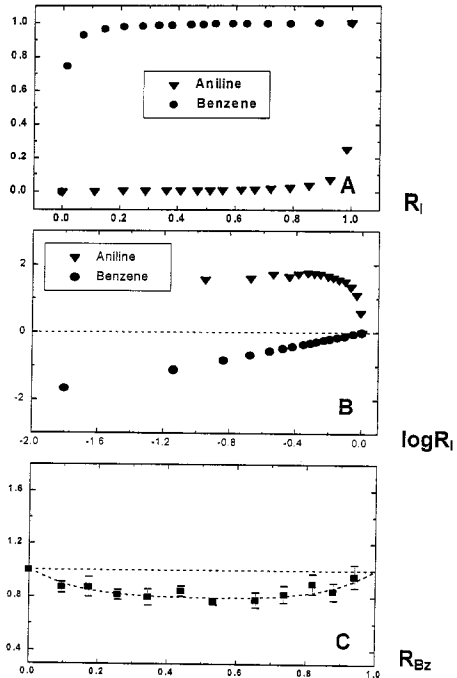


Fig. 2

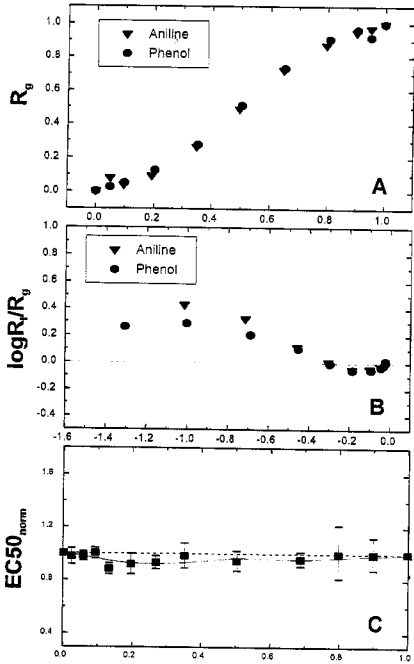


Fig. 3

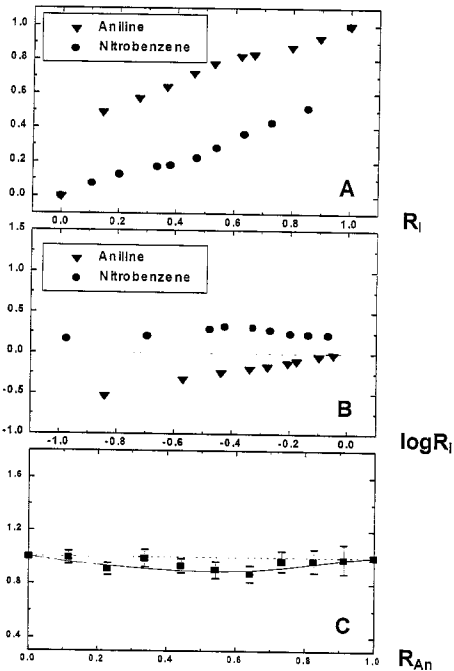


Fig. 4

The reproducibility of these phenomena was proved by a choice of the mixture tetrachloromethane with ethanol from physicochemical tables<sup>8</sup>. This mixture possesses the azeotropic point and the same plots A and B as in the Fig. 1 benzene with ethanol. The measured acute toxicity plotted as in the section Cs correspond to the inhibition of acute toxicity of the compounds being in the mixture.

Considering these results we suppose that besides metabolic reasons (interaction receptor - substrate, influence of the biotransformation, transport and distribution) for the deviation in activities of chemicals being in mixtures, a physicochemical interaction can be involved, too.

All the plots are possible to be expressed as mathematical functions. There is a hope for using this methodology for predicting acute toxicity of binary mixtures knowing their composition and physicochemical properties like boiling point, Henry constant, etc. The next step will be a study of the aqueous solutions of the mixtures, thus, ternary mixtures with low concentration of the chemical components. The step presented supports the idea that QSAR methodology can be useful for predicting biological properties of chemicals being in mixtures.

## REFERENCES

1. Tichý M., Rucki M., and Dohalský V.: *Composition Descriptor. Possible Use in the QSAR Analysis?* Presented at the 11th European Symposium on Quantitative Structure-Activity Relationships: Computer-Assisted Lead Finding and Optimization, Lausanne (1996), Switzerland.
2. M.Tichý, M. Cikrt, Z. Roth, and M. Rucki, QSAR analysis in mixture toxicity assessment, *SAR QSAR Environ. Res.* 9:in press (1998).
3. M. Tichý, M. Rucki, V. Bořek-Dohalský, L. Feltl, and Z. Roth, Possibilities of QSAR analysis in toxicity assessment of binary mixtures. In: *QSARs in Environmental Toxicology VIII*, in press (1998).
4. M. Tichý, and M. Rucki, Alternative methods of acute toxicity determination: inhibition of movement of *Tubifex tubifex* worms (in Czech, abstract in English), *Pracov. Léč.* 48: 225 (1996).
5. F.M. Rault (1886), in the textbook: *Physical Chemistry*, W.J.Moore, Chapter 7 and 8, the 4<sup>th</sup> edition, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1972.
6. S. Loewe, and H. Muischnek, Über Kombinationswirkungen. *Arch. Exp. Pathol. Pharmacol.* 114: 313 (1926).
7. D.J. Finney, The analysis of toxic tests of mixtures of poisons. *Ann. Appl. Biol.* 29: 82 (1942).
8. E. Lax, and K. Synowietz, *Taschenbuch für Chemiker und Physiker*. 3 Auflage, Springer-Verlag, Berlin, Heidelberg, New York (1964).

## ACKNOWLEDGEMENT

This work was supported by the grants of Grant Agency of Czech Republic No. 203/97/1027, of Grant Agency of Ministry of Education of Czech Republic No. OK280 and of COPERNICUS project C.E.U. No. CP94-1029.

## A CoMFA study on antileishmaniasis bisamidines

Carlos Alberto Montanari

Núcleo de Estudos em Química Medicinal-NEQUIM  
Departamento de Química, Universidade Federal de Minas Gerais,  
Campus da Pampulha, 31270-901, Belo Horizonte-MG, Brasil

### INTRODUCTION

The grooves of the DNA double helix are the principal interaction sites for many molecules.<sup>1,2</sup> The recognition is a global process that involves the overall structure and dynamics of the complex as well as hydrogen bonds, ion pairs, van der Waals and hydrophobic interactions.

We have previously shown that the antileishmaniasis activity against *Leishmania mexicana amazonensis* can be modelled for a set of pentamidine analogues interacting with B-DNA through their *isohelical pharmacophoric conformation*.<sup>3, 4</sup> The most potent compounds must have a bioisosteric change from -O- to -NH-, as depicted by the electrotopological index, S(i). However, no rationalisation of the previous studies have dealt with the importance of the amidine groups themselves, within pentamidine analogues. In order to circumvent this and try to understand their role as a major group used in the B-DNA molecular recognising process we have carried out a 3D QSAR study by using Comparative Molecular Field Analysis, CoMFA.<sup>5</sup>

Thus, this study reveals that for a set of 37 pentamidine analogues with antileishmaniasis activity, the receptor perturbational treatment is consistent with the hypothesis that a suitable sampling of the ligand steric and electrostatic interactions shall give an insight in the possible receptor interactions.

### METHOD

The CoMFA analysis were as follows: (i) generating the needed conformation for each molecule (including Coulombic terms); (ii) superposition; (iii) calculating the interaction energies; (iv) performing a PLS analysis and (v) graphical representation of the results. The Gasteiger-Marsili charges for 37 compounds were calculated using the physiological protonation state at the amidine group. The molecules were aligned by molecular weighted extent based on size and charge weighting factors using the automated similarity package,<sup>6</sup> (ASP). In order to carry out this alignment the "extended" X-Ray structure of pentamidine (Figure 1) and its "isohelical" conformation to B-DNA, (Figure 2), also from X-Ray structure, were used. The database were updated to explicitly describe similar conformations for all members into two separated data set.

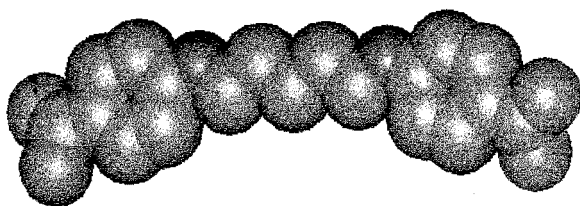


Figure 1: Pentamidine extended conformation.

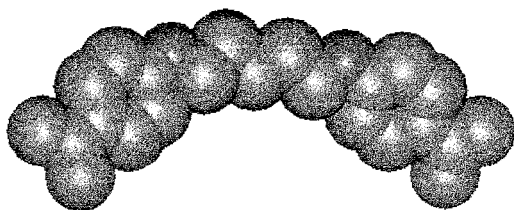


Figure 2. Isohelical pentamidine conformation to B-DNA

## RESULTS AND DISCUSSION

The similarities for the isohelical spatial orientation for all the pentamidine analogues prevail over the extended ones.<sup>3,4</sup> Nevertheless, it is quite plausible to assume that one must be cautious in the analysis of similarity matrices data. When the conformation at the receptor site is known, ASP similarity calculations seem to be a powerful way of describing it. However, the CoMFA analysis did not encompass these features when comparing both the isohelical and extended conformations. In spite of this, the final CoMFA model ( $r^2 = 0.984$ ,  $r^2_{cv} = 0.514$ ) shows the molecular features needed to describe the antileishmaniasis potencies as follows: (i) the amidine group can accommodate more bulk substituents; (ii) there is a need for more positive charge in the bisamidine linker, and (iii) the major contribution to potency is the steric field (70%).

A comparison between these results with the classical QSAR study<sup>3, 4</sup> demonstrates the model's predictive power by confirming some of the previous characteristics, but most certain do reveal the above ones which were not earlier disclosed.

## ACKNOWLEDGMENTS

The author would like to thank the Brazilian Grating Agencies CAPES, CNPq, FAPEMIG and FINEP, and Oxford Molecular for supporting this research work.

## REFERENCES

1. N. B. Boutonnet, X. Hui and K. Zakrzewska, *Biopolymers* 33:479 (1993).
2. M. E. A. Churchill and A. A. Travers, *TIBS* 16:92 (1991).
3. C. A. Montanari, M. S. Tute, A. E. Beezer and J. C. Mitchell, *J. Comp.-Aided Mol. Des.* 10:67 (1996).
4. C. A. Montanari and M. S. Tute, *Quant. Struct.-Act.Relat.*, 16:480 (1997)
5. Sybyl Version 6.3, Tripos, Inc.
6. ASP, Version 3.0, Oxford Molecular, Ltd.

## ANTILEISHMANIAL CHALCONES: STATISTICAL DESIGN AND 3D-QSAR ANALYSIS

Simon F. Nielsen<sup>1,2</sup>, S. B. Christensen<sup>1</sup>, A. Kharazmi<sup>3</sup> and Tommy Liljefors<sup>1</sup>

<sup>1</sup>Department of Medicinal Chemistry, Royal Danish School of Pharmacy, Universitetsparken 2, DK-2100 Copenhagen, Denmark, <sup>2</sup>State Serum Institute, Copenhagen, Denmark, <sup>3</sup>Department of Infectious Diseases, University Hospital, Copenhagen, Denmark

Leishmaniasis is an often lethal disease caused by various species of the protozoan parasite *Leishmania*. We have shown that chalcones *e.g.* Licochalcone A cure *leishmania* infections in mice. Unfortunately the chalcones are slightly toxic as shown by the inhibition of lymphocyte proliferation. This work describes the statistical design of substituted chalcones and calculation of 3D-QSAR models for the antiparasitic and toxic effects of the chalcones.

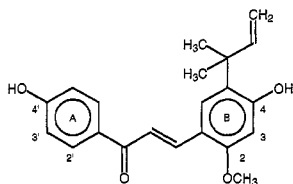


Figure 1. Licochalcone A.

Table 1. Properties of the 3D-QSAR models.

	Var.	$R^2$	$Q^2$
Initial model	39950	0.68	0.48
After pretreat.	4077	0.68	0.49
After variable selection	1365	0.73	0.63

In order to get a comprehensive data-material for 3D-QSAR analysis a number of substituted chalcones were designed by statistical methods. Sixty-two substituents, which can be introduced at aromatic positions in the chalcone skeleton were described by the four parameters  $MR$ ,  $\sigma_p$ ,  $\sigma_m$  and  $\pi$ . Principal Components Analysis was performed and the two first principal components which explain 89 % of the variance in the 4 original parameters were used for the statistical design. Using factorial design 24 chalcones were designed.

The biological data of these supplemented with data for 60 chalcones prepared for opening studies were used for the 3D-QSAR analyses (The antilymphocytic model are not described in detail here). Nine chalcones were chosen to form an external validation set.

The interaction energies between the energy-minimized compounds and three different probes (water, methyl and ammonium ion) were calculated by using the GRID program employing a grid spacing of 1 Å, which gave 57,200 variables for each compound.

The 3D-QSAR models were calculated by GOLPE. Using the Smart Region Definition procedure for variable preselection the number of variables were reduced to approximately 10% without reducing the quality of the models (Table 1). Subsequent Variable Selection removes variables which do not contribute, in a positive way, to the predictivity of the models, giving models of high quality. This was confirmed by the predictions of the external validation set (Figure 2).

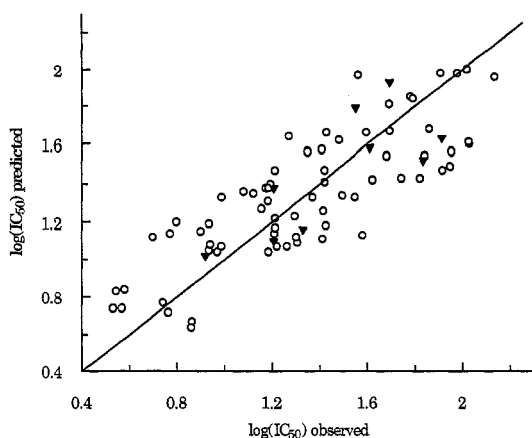


Figure 2. Observed and predicted antileishmanial activity; triangles represent validation set.

The coefficient plots for the three probes used in the GRID calculations are almost identical indicating that the difference in activity between the chalcones is mainly due to steric interactions with the target. The interpretation of the model is thereby simplified since the coefficient plot for the methyl probe contains almost all relevant information.

*Antileishmanial model.* The coefficient plots (Figure 3) show that substituents on the A-ring is mainly responsible for the difference in the antileishmanial activity of the chalcones. The negative coefficient regions around the 2'- and 3'-positions (ring A) indicate that substituents in these regions giving unfavorable interaction (positive interaction energies) with the methyl probe (e.g. bulky groups) will increase the activity of the compound. The positive coefficients illustrate regions around the molecule in which introduction of substituents are predicted to reduce the activity of the compounds. Thus, a bulky substituent in the 4'-position is predicted to reduce the antileishmanial activity of the compound.



Figure 3. Negative (left) and positive (right) coefficients for the antileishmanial activity.

*Antilymphocytic model.* In contrast to the coefficient plots for antileishmanial activity the plots for antilymphocytic activity shows that the antilymphocytic activity of the chalcones is influenced by substituents on the A as well as the B ring (data not shown).



# CHEMICAL FUNCTION BASED ALIGNMENT GENERATION FOR 3D QSAR OF HIGHLY FLEXIBLE PLATELET AGGREGATION INHIBITORS

Rémy D. Hoffmann<sup>1</sup>, Thierry Langer<sup>2</sup>, Peter Lukavsky<sup>2</sup>, Michael Winger<sup>2</sup>

<sup>1</sup> Molecular Simulations SARL, Parc Club Université, 20 rue Jean Rostand, F-91893 Orsay, France

<sup>2</sup> Institute of Pharmaceutical Chemistry, University of Innsbruck, Innrain 52a, A-6020 Innsbruck, Austria

## INTRODUCTION

It is well established that the quality of 3D QSAR experiments relies on one or multiple consistent 3D alignments for an ensemble of molecules as starting point for the calculations, especially when these molecules present a high degree of flexibility. Among the different pharmacophore identification techniques, the feature-based alignment methodology constitutes a useful approach [1]. To illustrate this methodology, we used a training set of 24 platelet aggregation inhibitors [2] (thromboxane A<sub>2</sub> receptor antagonists, TXRA / thromboxane synthetase inhibitors, TXSI) with affinities covering a range over four orders of magnitude for the receptor and two orders of magnitude for the enzyme.

## METHODS AND RESULTS

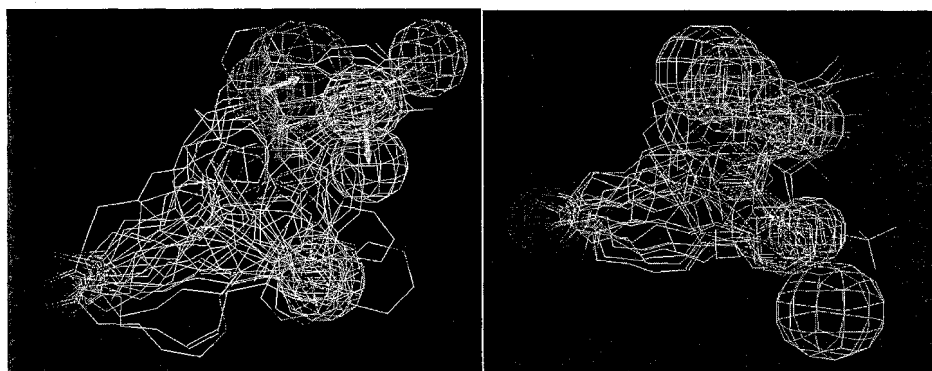
All molecular structures were edited within the CATALYST software [3] and minimized to their closest local energy minimum. Poled conformations [4] were generated using an energy cutoff of 15 kcal/mol. The molecules were then aligned according to the pharmacophore models generated using catHypo for TXRA. The H-bond acceptor, donor, hydrophobe, negative ionizable, and aromatic ring functions [5] were considered and only hypotheses containing five features were retained. All the other parameters were set to their default values. The TXSI training set suffers from i) a narrow activity range and ii) an unbalanced distribution of activity data. Therefore in this case the HipHop [6] method was used to generate the alignments, considering hypotheses containing a minimum of five features (negative ionizable, aromatic hydrophobes, hydrophobes and H-bond acceptor function). In this case, only the nine most active compounds were considered for model generation. All 24 molecules were then aligned on the generated pharmacophore models and then used as input for the 3D QSAR study.

The alignments for TXRA and TXSI are shown in Figures 1. A CoMFA[7] was performed using the standard atom probes (C.sp<sup>3</sup>, charge +1). In order to determine how well the model predicts data, each predictive value was cross-validated using initially five components resulting in a determination of the optimum number of components. The results of CATALYST and CoMFA QSAR activity prediction as well as the statistical

evaluation is shown in Table 1. As can be seen from the results listed, use of three and 2 components, respectively, is sufficient to obtain a satisfactory prediction. PLS analysis of the descriptors generated from the initial region without cross-validation afforded the final model with a conventional  $r^2$  of 0.97 for TXRA and  $r^2$  of 0.87 for TXSI together with a standard error of estimate of 0.2. The relative contribution of steric and electrostatic potential to the CoMFA regression equation was found to be 45.6 and 54.4 % steric and electrostatic, respectively. The standard deviation coefficient contour maps (Figures 2 and 3) derived from the final model, display the 3D CoMFA contributions of steric and electrostatic potentials. These contour maps indicate where the changes in fields are correlated with changes in binding affinity.

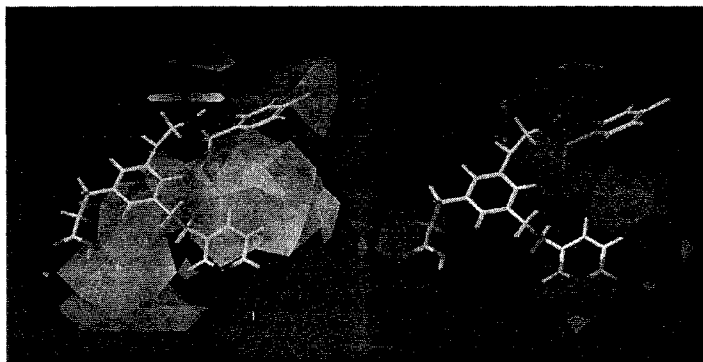
**Table 1.** Actual and predicted affinity data ( $pIC_{50}$ )

Compound	TXRA	TXRA	TXRA	TXSI	TXSI	Compound	TXRA	TXRA	TXRA	TXSI	TXSI
	$pIC_{50}$	Catalyst	CoMFA	$pIC_{50}$	CoMFA		$pIC_{50}$	Catalyst	CoMFA	$pIC_{50}$	CoMFA
<b>Ridogrel</b>	5,77	5,51	5,75	8,22	7,79	<b>15-12</b>	6,66	7,25	6,71	7,18	7,24
<b>162293</b>	6,57	6,96	6,76	7,32	7,31	<b>15-14</b>	7,44	7,52	7,43	7,00	7,11
<b>PL176</b>	5,22	5,72	5,06	7,80	7,75	<b>13-14b</b>	4,83	4,96	4,87	6,77	6,90
<b>PL91</b>	4,96	5,44	5,04	8,10	7,99	<b>13-14e</b>	4,27	4,96	4,08	7,40	7,18
<b>PL138</b>	5,82	5,72	5,76	7,41	7,54	<b>13-14g</b>	4,91	5,20	5,15	7,40	7,32
<b>PL137</b>	6,44	6,31	6,37	5,21	7,35	<b>13-23c</b>	7,40	6,17	5,07	7,10	7,02
<b>14-35</b>	7,92	7,19	7,89	8,40	8,13	<b>2-16</b>	5,41	4,96	5,38	7,11	7,36
<b>14-42</b>	7,18	7,42	7,19	8,40	8,35	<b>2-23</b>	5,37	4,96	5,35	7,30	7,37
<b>14-75</b>	6,64	6,39	6,60	8,40	8,74	<b>2-35</b>	7,52	7,55	7,35	6,47	7,34
<b>6-1R</b>	7,96	7,05	7,73	8,30	8,15	<b>16-7</b>	6,64	7,39	7,17	7,59	7,93
<b>6-2S</b>	8,00	8,31	8,20	8,46	7,55	<b>16-11</b>	8,52	8,79	8,24	8,05	7,92
<b>15-9</b>	7,03	7,16	6,81	7,10	7,14	<b>16-12</b>	5,95	6,57	6,15	8,15	8,23



**Figure 1.** Alignment generated for TXRA (left) and TXSI (right) from CATALYST

Whereas HipHop is only usable for a qualitative alignment generation based on common chemical features thus giving a suitable input for a further 3D QSAR analysis, catHypo itself represents a quantitative pharmacophore construction tool permitting to estimate the activity of molecules from their mapping on the hypothesis obtained (see Table 1). However, also in this case, the use of another 3D QSAR method (e.g. like CoMFA) gives additional and moreover complementary information on a given problem.



**Figure 2**

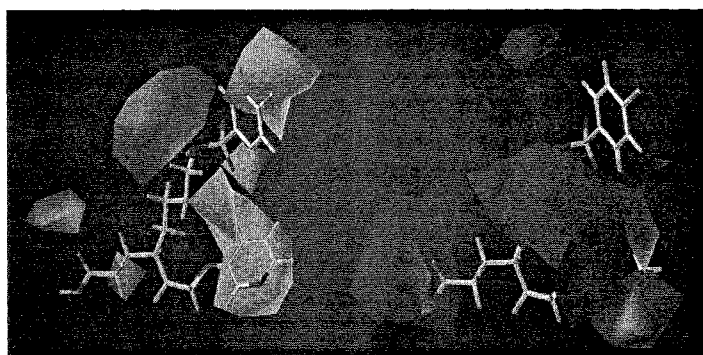
Sdev \* coefficient contour plots for the TXRA model derived by CoMFA (left: steric, right: electrostatic interaction contour maps)

$$r_{cv}^2 = 0.69$$

$$r^2 = 0.97$$

$$s = 0.2 \text{ (2 comp.)}$$

$$n = 24$$



**Figure 3**

Sdev \* coefficient contour plots for the TXSI model derived by CoMFA (left: steric, right: electrostatic interaction contour maps)

$$r_{cv}^2 = 0.43$$

$$r^2 = 0.87$$

$$s = 0.2 \text{ (2 comp.)}$$

$$n = 24$$

## CONCLUSIONS

The results of this work clearly indicate that CATALYST generated alignments are suitable inputs for the generation of 3D - QSAR models. Both CATALYST's hypothesis generation algorithm and the CoMFA method yield predictive interaction models. As information provided by both methods is complementary the combined use of CATALYST and CoMFA 3D - QSAR appears to be a promising approach in drug design.

## REFERENCES

- 1 T. Langer, R. D. Hoffmann, On the use of chemical function based alignments as input for 3D QSAR, *J. Chem. Inf. Comput. Sci.* 38: 325 (1998)
- 2 M. Winger, 3D QSAR Untersuchungen an neuen Thrombozytenaggregationshemmern, Diploma Thesis, University of Innsbruck, 1995 and references cited therein
- 3 Molecular Simulations Inc., San Diego, CA, USA (1996)
- 4 A. Smellie, S. L. Teig, P. Towbin, Poling: Promoting Conformational Coverage *J. Comp. Chem.* 16: 171 (1995)
- 5 J. Greene, S. D. Kahn, H. Savoj, P. Sprague, S. Teig, Chemical Function Queries for 3D Database Search, *J. Chem. Inf. Comput. Sci.* 34: 1297 (1994)
- 6 D. Barnum, J. Greene, A. Smellie, P. Sprague, Identification of common functional configurations among molecules *J. Chem. Inf. Comput. Sci.* 36: 563 (1996)
- 7 Tripos Ass., St. Louis, MO, USA (1996)

## 3D QSAR ON MUTAGENIC HETEROCYCLIC AMINES THAT ARE SUBSTRATES OF CYTOCHROME P450 1A2

Juan J. Lozano,<sup>1</sup> Manuel Pastor,<sup>2,3</sup> Federico Gago,<sup>2</sup> Gabriele Cruciani,<sup>3</sup>  
Nuria B. Centeno<sup>1</sup> and Ferran Sanz<sup>1,\*</sup>

<sup>1</sup>Research Group on Medical Informatics, Institut Municipal d'Investigació Mèdica (UAB), c/Doctor Aiguader 80, E-08003 Barcelona

<sup>2</sup>Dep. of Pharmacology, Univ. of Alcalá, E-28871 Alcalá de Henares

<sup>3</sup>Lab. di Chemiometria, Dip. di Chimica, Univ. di Perugia, I-06123 Perugia

### INTRODUCTION

Heterocyclic aromatic amines (HCA) present in cooked food, exert a genotoxic activity after metabolism (N-oxidation) by cytochrome P450 1A2<sup>1</sup>. Two different 3D-QSAR approaches (COMBINE<sup>2</sup> and GRID/GOLPE<sup>3</sup>) have been applied to a series of 12 HCAs showing different degrees of mutagenic activity (Figure 1).

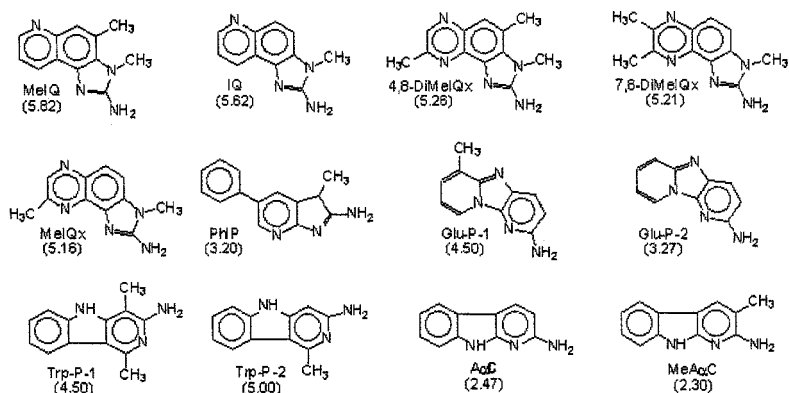


Figure 1. Chemical structures of the considered HCAs. Values within parenthesis are estimated mutagenicities<sup>4</sup>

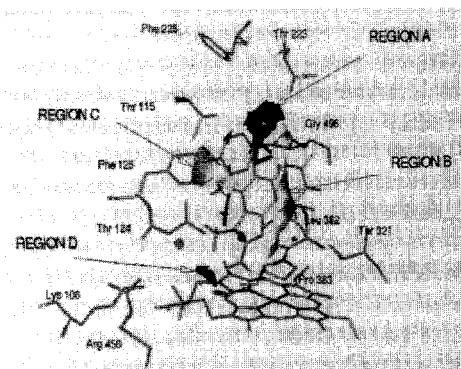
### COMBINE AND GRID/GOLPE ANALYSES

Solvated HCA-P450 1A2 complexes were obtained with AUTODOCK 2.4<sup>5</sup> using a

\* To whom correspondence has to be addressed

model of cytochrome P450 1A2 previously published.<sup>6</sup> Energies of the complexes were obtained after geometrical optimisation using AMBER 4.1.<sup>7</sup> In order to carry out the COMBINE analysis<sup>2</sup>, energies were partitioned on a *per residue* basis. Energy with absolute values lower than 0.05 kcal/mol were zeroed. Variables with SD < 0.05 among the compounds were not taken into account. Block unscaled weights scaling was applied.<sup>3</sup> After Fractional Factorial Design (FFD) variables selection, a two PCs model with  $r^2 = 0.90$  and  $q^2 = 0.78$  was obtained. Predictions were quite accurate with the exception of Glu-P-2. The most important residues involved in substrate-enzyme interactions were: Thr115, Asp119, Thr124, Thr223, Asp313, Gly316, Thr321, Leu382, Pro383, Tyr495 and Arg456.

In the GRID/GOLPE analysis, the amines were aligned as in COMBINE. GRID computations were carried out using a phenolic OH probe and a 14×17×16Å box with 1 Å grid spacing. Values greater than +5 kcal/mol were cutoff to this value, and those absolute values lower than 0.1 were zeroed. Variables with SD < 0.1 and these taking only two or three values and having skewed distribution, were eliminated. Smart Region Definition<sup>8</sup> (critical distance = 2Å, collapsing cutoff = 27.2 Å) and two FFDs were used for variable selection. Considering the two first PCs,  $r^2 = 0.96$  and  $q^2 = 0.79$  were obtained. Most important PLS coefficients are grouped in four regions (Figure 1). Dark zones indicate hydrogen bonds or electrostatic interactions in the most active compounds. Light zones reveal the presence of hydrophobic groups in the most active compounds.



**Figure 1.** Contour map of PLS coefficients in the GRID/GOLPE model. MeIQ is shown.

## CONCLUSIONS

A clear coincidence of the results of both methodologies was obtained: the residues nearest to the regions including the largest PLS GRID/GOLPE coefficients are those highlighted by the COMBINE model. Using a docking-guided alignment, GRID/GOLPE yields better fitting (0.96 vs 0.90 for  $r^2$  values), and slightly better predictive indexes (0.79 vs 0.78) than COMBINE. COMBINE has the advantage of giving more detailed insight on which are the residues involved in the ligand-receptor interaction.

## Acknowledgements

This research was supported in part by CICYT (SAF 93-0722-C02-02) and CESCA grants.

## REFERENCES

1. T. Shimada, M. Iwasaki, M.V. Martin and F.P. Guengerich. *Cancer Res.* 49:3218 (1989).
2. C. Pérez, M. Pastor, A.R. Ortiz and F. Gago. *J. Med. Chem.* 41:836 (1998).
3. M. Baroni, G. Constantino, G. Cruciani, D. Riganelli, S. Valigri and S. Clementi. *QSAR* 12:9 (1993).
4. K. Wakabayashi, M. Nagao, H. Esumi and T. Sugimura. *Cancer Res.* 52(suppl):2092s (1992).
5. G.M. Morris, D. Goodsell, R. Huey and A.J. Olson. *J. Comput.-Aided Mol. Des.* 10:293 (1996).
6. J.L. Lozano, E. López de Briñas, N.B. Centeno, R. Guigó and F. Sanz. *J. Comput.-Aided Mol. Des.* 11:39 (1997).
7. S.J. Weiner, P.A. Kollman, D.T. Nguyen and D.A. Case. *J. Comp. Chem.* 7:230 (1986).
8. M. Pastor, G. Cruciani and S. Clementi. *J. Med. Chem.* 40:1455 (1997).

## APPLICATION OF 4D-QSAR ANALYSIS TO A SET OF PROSTAGLANDIN, PGF<sub>2</sub>α, ANALOGS<sup>#</sup>

C. Duraiswami\*, P. J. Madhav<sup>†</sup>, and A. J. Hopfinger<sup>‡</sup>

\* Avon Products, Inc., Skin Care Laboratories, 1 Avon Place, Suffern,  
NY - 10901-5605

<sup>†</sup> Proctor & Gamble Pharmaceuticals, Health Care Research Center,  
Mail code 1128, PO Box 8006, Mason, OH - 45040-8006

<sup>‡</sup> Department of Medicinal Chemistry & Pharmacognosy, The University of Illinois at  
Chicago, College of Pharmacy, (M/C-781), 833 S. Wood Street, Chicago,  
IL - 60612-7231

### ABSTRACT

Four-dimensional Quantitative Structure-Activity Relationship (4D-QSAR) analysis is a method developed recently to determine molecular similarity, diversity, and construct three-dimensional structure-activity relationships (3D-QSARs)<sup>1</sup>. 4D-QSAR analysis incorporates conformational and alignment freedom into the development of 3D-QSAR models for training sets of structure-activity data by performing ensemble averaging, the fourth "dimension". The difference between 4D-QSAR and 3D-QSAR is that instead of examining a single conformation and alignment, an ensemble of conformations and alignments over a short period of time is examined. The descriptors in 4D-QSAR analysis are derived from measures of grid cell (spatial) occupancy of the atoms present in each molecule in the training set, realized from sampling of conformation and alignment spaces. Grid cell occupancy descriptor can be generated for any atom type, group, and/or model pharmacophore. Serial use of partial-least squares, (PLS), regression and a Genetic Algorithm, (GA), is used to perform data reduction and identify the manifold of top 3D-QSAR models for the training set. The unique manifold of 3D-QSAR models is determined by computing the extent of orthogonality in the residuals of error among the most significant 3D-QSAR models generated by the GA. Additionally, a single "active" conformation can be postulated for each compound in the training set, which can be combined with optimal alignment for use in other molecular design applications, including other 3D-QSAR methods. The influence of the conformational entropy on the activity of each compound can also be estimated.

Receptor independent (RI) 4D-QSAR was successfully applied to a set of 42 Prostaglandin, PGF<sub>2</sub>α, analogs, with antinidatory activity.

Two (RI) 4D-QSAR studies were carried out. The first study has been described in reference (1) in great detail, and only the second study has been described here.

### METHODS

The training set comprises of 42 Prostaglandin, PGF<sub>2</sub>α, analogs. Please see reference (1), for the structures of the compounds in the training set and the general method for performing a 4D-QSAR analysis. The methodology parameters and the Interaction Pharmacophore Elements (IPEs) considered in study 2 are shown below in Tables 1 and 2, respectively.

---

<sup>#</sup> Research conducted at the University of Illinois at Chicago.

**Table 1.** Methodology Parameters of 4D-QSAR Analysis for Study 2

Parameter Description	Symbol	Value
• Grid cell size [only cubic cells are allowed]	S (?)	(40, 40, 40) 1.5 ?
• Temperature of the molecular dynamics simulation, MDS	T	300 °K
• Reference molecule	R	None
• Size of ensemble sampling (number of distinct initial starting conformations in the sampling)	E <sub>s</sub> (I)	40,000(1)
• Number of alignments	N <sub>a</sub>	1
• Number of descriptors in the GA initial basis set	N <sub>d</sub>	212

**Table 2.** Interaction Pharmacophore Elements, IPEs, of (RI) 4D-QSAR Analysis

IPE Description	Symbol
All atoms of the molecule	IPE (a)
Polar atoms of the molecule	IPE (p+), IPE(p-)
Non-polar atoms of the molecule	IPE (n)
Hydrogen bond donors	IPE (hbd)
Hydrogen bond acceptors	IPE (hba)

## RESULTS

Best results were obtained when performing GFA analysis with 30,000 crossovers and a smoothing factor of 0.25. The optimum 3D-QSAR model for all 42 analogs for Study 2 is given by equation 1, and the removal of outliers yielded equation 2. In the equations below, GC represents grid cell numbers.

### Equation 1:

$$\log (\text{Rel. ED}_{50}) = 1.52 - 2.93 \text{ GC1(np)} - 1.84 \text{ GC2(a)} + 5.08 \text{ GC3(a)} + 2.42 \text{ GC4(np)} - 1.11 \text{ GC5(a)} + 0.98 \text{ GC6(np)} - 2.91 \text{ GC7(np)}$$

$$N = 42 \quad R^2 = 0.760 \quad xv - R^2 = 0.644 \quad F = 15.6$$

Removal of outliers (MD-021, MD-045, MD-058, and MD-059) resulted in equation 2.

### Equation 2:

$$\log (\text{Rel. ED}_{50}) = 0.97 - 2.86 \text{ GC1(np)} - 1.68 \text{ GC2(a)} + 6.08 \text{ GC3(a)} + 2.31 \text{ GC4(np)} - 1.08 \text{ GC5(a)} + 1.39 \text{ GC6(np)} + 2.38 \text{ GC7(np)}$$

$$N = 38 \quad R^2 = 0.842 \quad F = 22.8$$

## REFERENCES

1. A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, and C. Duraiswami, Construction of 3D-QSAR models using the 4D-QSAR analysis formalism, *J. Am. Chem. Soc.* **119**:10509 (1997).

## DETERMINATION OF THE CHOLECALCIFEROL-LIPID COMPLEX USING A COMBINATION OF COMPARATIVE MODELLING AND NMR SPECTROSCOPY

Mariagrazia Sarpietro<sup>(a)</sup>, Mario Marino<sup>(a)</sup>, Antonio Cambria<sup>(a)</sup>, Gloria Uccello Barretta<sup>(b)</sup>, Federica Balzano<sup>(b)</sup>, Salvatore Guccione<sup>(c)</sup>

<sup>(a)</sup>*Istituto di Scienze Biochimiche e Farmacologiche, Università di Catania, viale Andrea Doria 6, Ed. 12, I-95125 Catania, Italy*

<sup>(b)</sup>*Centro CNR di Studio per le Macromolecole Stereordinate ed Otticamente Attive, Università di Pisa, via Risorgimento 35, I-56126 Pisa, Italy*

<sup>(c)</sup>*Dipartimento di Scienze Farmaceutiche, Università di Catania, viale Andrea Doria 6, Ed. 12, I-95125 Catania, Italy*

Exploration of the systemic disposition of macromolecules in relation to their physicochemical properties, could be a strategy for designing targeting system.

This work deals with the investigation of the Vitamin D<sub>3</sub> conformation/s in the phospholipid bilayer<sup>1,2</sup> in order to define a possible preferred binding site at the C=O- or PO<sub>2</sub>-phospholipid moiety (structure-function studies) to be exploited into drug discovery efforts (*forthcoming paper*).

### NMR analysis

1D-n.O.e. data were compliant to a *s-trans* conformation of the diene moiety with the proton H<sub>6</sub> and Me<sub>18</sub> respectively bent towards H<sub>9</sub> and the diene moiety (**Fig 1**).

The presence of equilibrating conformers,  $\alpha$  and  $\beta$  (**Fig 2**), having the OH group respectively in equatorial and axial orientations, already well documented<sup>3</sup>, is also in agreement with the observation of almost equivalent interproton dipolar interactions between the proton H<sub>19E</sub> and the protons H<sub>1a</sub> and H<sub>1b</sub>, almost equivalent intern.O.e.s H<sub>6</sub>-H<sub>4a</sub> and H<sub>6</sub>-H<sub>4b</sub> were observed. Moreover, the observation of a dipolar interaction between H<sub>3</sub> and H<sub>1a</sub> is an indication of the fact that such protons are in a diaxial arrangement as in conformation  $\alpha$ , whereas the detection of a clear n.O.e. between the methyl protons Me<sub>18</sub> and H<sub>19z</sub>, aside from the complete absence of n.O.e. between H<sub>19z</sub> and H<sub>15</sub> belonging to the D ring, reveals the presence of a conformer in which the unsaturated group is on the same side of the Me<sub>18</sub>, outside the diene plane, as in conformation  $\beta$ . <sup>13</sup>C T<sub>1</sub> measurements revealed that the hypothesis of isotropic overall motion is nearly satisfied, indeed the methylene <sup>13</sup>C T<sub>1</sub>s (0.55 s) of Vitamin D<sub>3</sub>, belonging to the A and CD rings were similar each other and correlated to the <sup>13</sup>C T<sub>1</sub>s of methine carbons (0.99 s). Thus, the same re-orientational time can be attributed to all the molecules and the ratios of the different interproton cross-relaxation rates  $s_{ij}$ , determined by proton selective relaxation rates measurements<sup>2</sup>, can be simply correlated to the ratios of the internuclear distances ( $s_{ij}/s_{ik} = (r_{ik}/r_{ij})^6$ , eq. 1), thus allowing a more precise definition of the stereochemistry. Hence, we determined the cross-relaxation rate  $s_{19Z-19E}$



for the proton pair 19Z-19E, corresponding to the fixed geminal distance  $r_{19Z-19E}$  (1.72 Å) and calculated from equation 1 the distances  $r_{19Z-7}$  (2.31 Å) and  $r_{6-9}$  (1.90 Å). These were consistent with a conformation in which the diene moiety is preferentially coplanar with the C<sub>8</sub>-C<sub>9</sub>-H<sub>9</sub> fragment thus allowing a remarkable spatial proximity between the proton H<sub>6</sub> and the pseudo-equatorial proton H<sub>9</sub> and the unsaturated methylene group belonging to the A ring is outside the above plane. As far as the butadiene bridge is concerned, the cross-relaxation rate for the tightly coupled proton pair 6-7 allowed us to assign the diene moiety a *s-trans* conformation, being their distance over 3 Å (Fig 3). Attempts to carry out the corresponding analysis in the cholecalciferol-lipid complex (containing 40 mol% of vitamin D<sub>3</sub>) are in progress, to overcome the problem coming from the remarkable broadening of vitamin D<sub>3</sub> resonances, making them not distinguishable in the spectra.

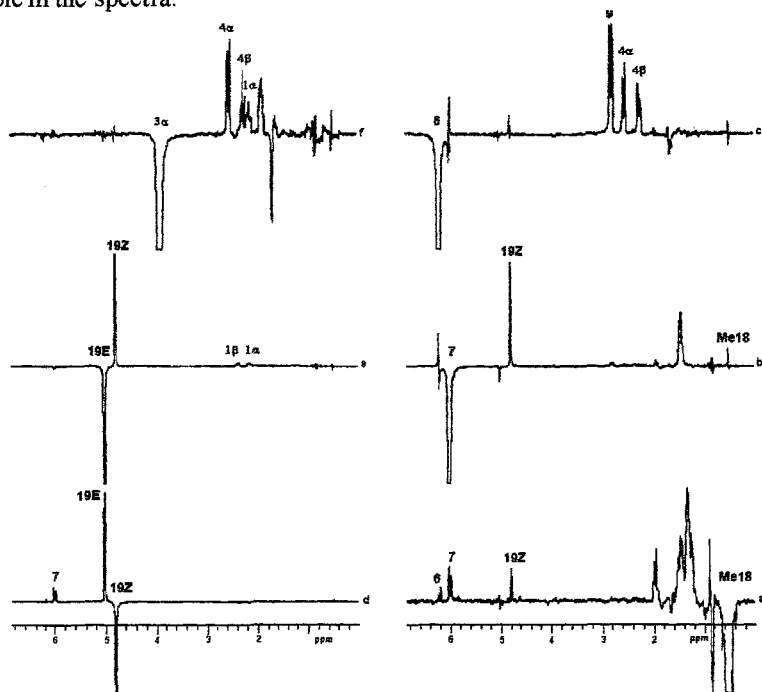


Fig 1 1D n.O.e. difference spectra (300 MHz, CDCl<sub>3</sub>, 25 °C) corresponding to the irradiation of the following resonances of 1: a) Me<sub>18</sub>, b) H<sub>7</sub>, c) H<sub>6</sub>, d) H<sub>19Z</sub>, e) H<sub>19E</sub> and f) H<sub>3a</sub>.

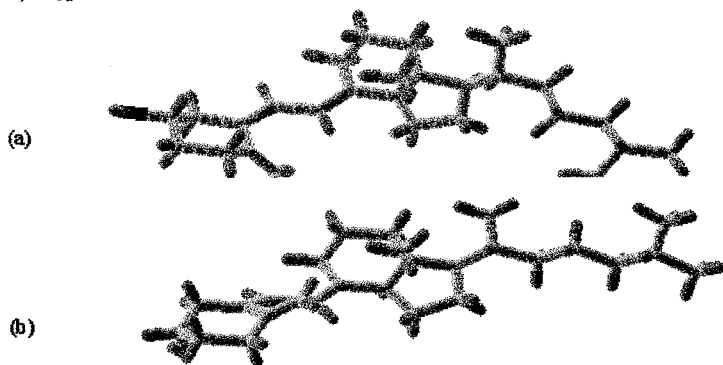
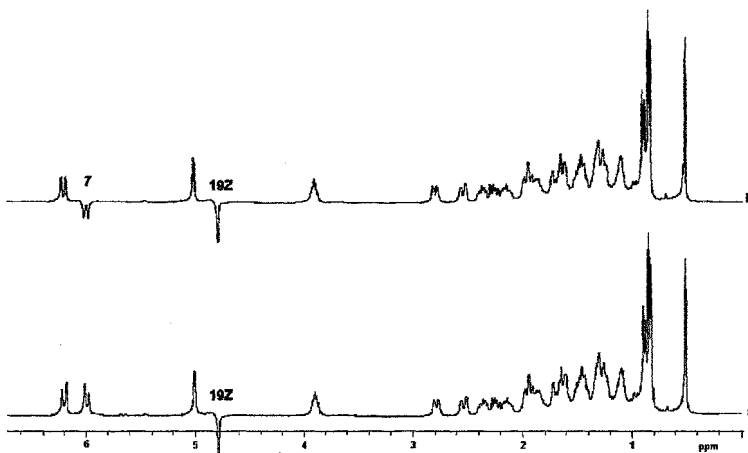


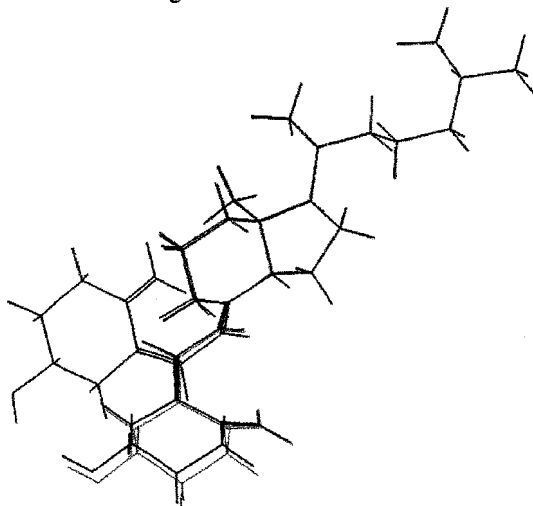
Fig 2 (a) Graphical representation of the  $\alpha$ -conformer of Vitamin D<sub>3</sub>; (b) Graphical representation of the  $\beta$ -conformer of Vitamin D<sub>3</sub>.



**Fig 3**  $^1\text{H}$  NMR (300 MHz,  $\text{CDCl}_3$ , 25  $^\circ\text{C}$ ) spectra of Vitamin  $\text{D}_3$ , showing: a) the monoselective inversion of  $\text{H}_{19\text{Z}}$ , b) the bisselective inversion of  $\text{H}_{19\text{Z}}$  and  $\text{H}_7$ .

### Molecular modelling

A conformational search (0-360 degrees with stepsize of 10 ) was carried out at the torsional angles of the C5-C8 sequence, starting from a Vitamin  $\text{D}_3$  fragment in the Sybyl Standard Library (Version 6.2)<sup>9</sup> as configured on a SGI INDIGO 2 workstation (operating under IRIX 5.2). The energies of the different conformations were considered also computing the electrostatic charges (Pullmann's method). The two lowest energy conformers were selected and furtherly optimized by minimization (gradient conjugate method without simplex with a convergence criterion of 0.05 kcal/mol). The fit of the two optimized structures as represented in black colour (137.5 degrees, 27.480 kcal/mol, rms 0.015; 286.4 degrees, 27.340 kcal/mol, rms 0.013) in comparison with that one coming from the SYBYL library standard fragment<sup>9</sup> (grey) is shown in **Fig 4**. There are two *equiprobable* different but isoenergetic conformations with the same substructure at the torsional angle of the C5-C8 segment.



**Fig 4** Fit of the two optimized structures (137.5 degrees, 27.480 kcal/mol, 0.015 rms; 286.4 degrees, 27.340 kcal/mol, 0.013 rms) as represented in black colour in comparison with that one coming from the SYBYL library standard fragment<sup>9</sup> of Vitamin  $\text{D}_3$  (grey colour).

## References

1. M. G. Sarpietro, Molecular interactions in membran model systems, Dottorato di Ricerca Thesis (*Italian Ph.D.*), University of Catania, (1993).
2. A. Raudino, F. Castelli, M. G. Sarpietro, A. Cambria, Calorimetric analysis of lipid-sterol systems: a comparison between structurally similar cholesterol and vitamin D<sub>3</sub> interacting with phospholipid bilayers of different thickness, *Chem. Phys. Lipids*, **74**: 25 (1994).
3. R. M. Wing, W. H. Okamura, A. Rego, M. R. Pirio, A. W. Norman, Studies on vitamin D and its analogs. VII. Solution conformations of vitamin D<sub>3</sub> and 1 $\alpha$ ,25-dihydroxyvitamin D<sub>3</sub> by high-resolution proton magnetic resonance spectroscopy, *J. Am. Chem. Soc.*, **97**: 4980 (1975).
4. E. Berman, Z. Luz, Y. Mazur, M. Sheves, Conformational analysis of Vitamin D and analogues. <sup>13</sup>C and <sup>1</sup>H nuclear magnetic resonance study, *J. Org. Chem.*, **42**: 3325 (1977).
5. G. Kotowycz, T. T. Nakashima, M. K. Green, G. H. M. Aarts, A proton magnetic resonance relaxation time study of several vitamin D derivatives, *Can. J. Chem.*, **58**: 45 (1980).
6. G. Kotovych, G. H. M. Aarts, K. Bock, A proton magnetic resonance nuclear overhauser enhancement study. Application to vitamin D derivatives D<sub>2</sub> and D<sub>3</sub>, *Can. J. Chem.*, **58**: 1206 (1980).
7. M. D. Mizhiritskii, M. D., L. E. Konstantinovskii, R. Vishkaitsan, 2D NMR study of solution conformations and complete <sup>1</sup>H and <sup>13</sup>C chemical shifts assignments of vitamin D metabolites and analogs, *Tetrahedron*, **52**: 1239 (1996).
8. G. Valensin, T. Kushnir, G. Navon, Selective and nonselective proton spin-lattice relaxation studies of enzyme-substrate interactions, *J. Magn. Reson.*, **46**: 23 (1982).
9. SYBYL Molecular Modelling Software, Tripos Inc., 1699 S. Hanley Raod, Suite 303, St. Louis, MO 63144.

# COMPARATIVE BINDING ENERGY (COMBINE) ANALYSIS ON A SERIES OF GLYCOGEN PHOSPHORYLASE INHIBITORS. COMPARISON WITH GRID/GOLPE MODELS

Manuel Pastor,<sup>1</sup> Federico Gago,<sup>2</sup> and Gabriele Cruciani<sup>1</sup>

<sup>1</sup> Laboratory of Chemometrics  
University of Perugia  
06123 Perugia, Italy

<sup>2</sup> Department of Pharmacology  
University of Alcalá  
28871 Alcalá de Henares, Spain

## INTRODUCTION

When a series of compounds of known biological activity is available and the structure of the receptor is unknown, the only rational drug design approaches applicable are Quantitative Structure-Activity Relationships (QSAR). On the other hand, when the structure of the receptor is known it is possible to use Structure-Based Drug Design (SBDD) techniques in order to gain insight into the ligand interaction. However, an increasingly common situation is that in which a full set of ligand-receptor complex structures is at hand.

The challenge in this case is to rationalize the information contained in the structure of the complexes and use it to advantage for practical purposes, i.e. for the design of more potent or more selective compounds. Indeed, new methodologies are emerging to bridge SBDD and QSAR. Among them, a promising one is the Comparative Binding Energy (COMBINE) approach<sup>1,2</sup>. The essence of this technique is to partition the calculated interaction energy for a series of ligand-receptor complexes into per-residue van der Waals and electrostatic contributions. The energy variables obtained are then correlated with the biological activities of the ligands, using Partial Least Squares (PLS). So far, this technique has been successfully applied to three series of compounds. Unfortunately, in all cases the docking of the ligands into the receptor lacked direct experimental support, since the complexes were obtained by modeling.

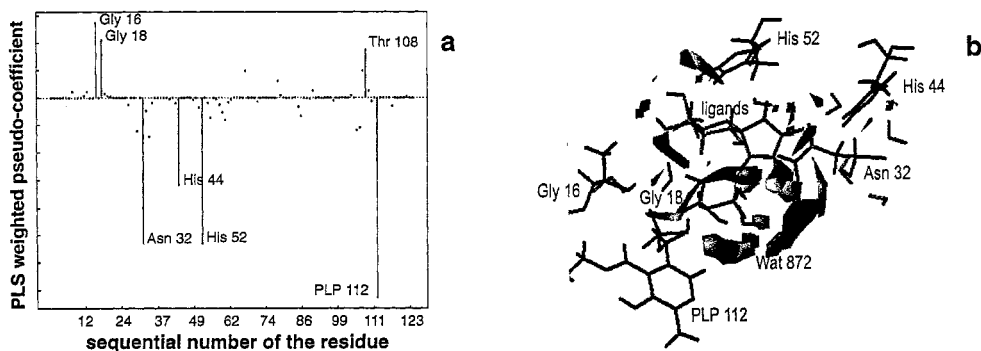
## COMBINE ANALYSIS

The work presented here describes for the first time an application of COMBINE analysis to a series of ligand-receptor complexes solved by X-ray crystallography.

This series of 10 compounds is a carefully designed subset of the series of 54 glycogen phosphorylase inhibitors studied in a previous work<sup>4</sup>. Since the enzyme studied is quite

large all the calculations were carried out on a model of the active site consisting of a 12Å shell of protein residues around the ligands. All the water molecules present in the complexes were removed except for 10 water molecules which were reported as very important for understanding the ligand-receptor interaction in a previous work<sup>4</sup>. Water molecules were considered as part of the ligands. The alternative approach of considering the water molecules as part of the receptor provided worse results and the interpretation of the model proved more difficult. Force field parameters for the ligands were obtained by interpolation, and charges were calculated by using ab-initio quantum mechanical methods at the 6-31\*\* level and MEP fitting. The AMBER 4.1<sup>4</sup> force field was used to mildly refine the structures of the complexes and to compute the ligand-receptor interaction energies, which were partitioned into per-residue contributions.

The matrix of ligand-receptor interaction energies was pretreated by replacing with 0.00 any values smaller than 0.01 and removing any variables with a standard deviation less than 0.01. The PLS analysis of this matrix, after applying GOLPE variable selection, yielded a good model ( $r^2=0.83$ ,  $q^2=0.65$ , cross-validated using 5 randomly formed groups and 20 randomizations). A histogram of the weighted PLS pseudo-coefficients obtained (Figure 1a) reveals the residues that are most important for explaining the differences in activity.



**Figure 1.** (a) PLS coefficients obtained in the COMBINE model, highlighting the most important ligand-residue interactions. (b) A simplified model of the binding site showing the most important residues, according to the COMBINE model and the PLS coefficients produced by the GRID/GOLPE model.

Using the same series, a GRID/GOLPE analysis was carried out under conditions similar to those described in ref 4. Both models showed a general agreement, as seen in figure 1b, but the COMBINE model provided complementary information which simplifies the interpretation and solves some ambiguities found in previous GRID/GOLPE models<sup>4</sup>.

## REFERENCES

1. A.R. Ortiz, M.T. Pisabarro, F. Gago, and R. Wade, Prediction of drug binding affinities by comparative binding energy analysis, *J. Med. Chem.* 38:2681 (1995).
2. C. Pérez, M. Pastor, A.R. Ortiz, and F. Gago, Comparative binding energy (COMBINE) analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design, *J. Med. Chem.* 41:836 (1998).
3. W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* 117:5179 (1995).
4. M. Pastor, G. Cruciani, and K.A. Watson. A strategy for the incorporation of water molecules present in a ligand-binding site into a 3D-QSAR analysis, *J. Med. Chem.* 40:4089 (1997).

## EVA QSAR: DEVELOPMENT OF MODELS WITH ENHANCED PREDICTIVITY (EVA\_GA)

David B. Turner and Peter Willett

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

### INTRODUCTION

QSAR models are of great importance in the rationalisation and prediction of the relative bioactivities of sets of compounds.<sup>1</sup> Over the last decade, field-based 3D-QSAR techniques, such as CoMFA,<sup>2</sup> have proved to be an effective means of correlating shape-related features with bioactivity, provided that a suitable relative alignment of the structures concerned can be found. EVA, which is derived from IR-/Raman-range vibrational frequencies, provides an alignment-free methodology which provides statistically robust QSARs generally comparable to those obtained with CoMFA. The method is sensitive to 3D structure but the descriptor is invariant to the relative rotation and translation of the structures concerned. "Classical EVA" has been extensively validated using many different data sets.<sup>3,4,5</sup> Here we briefly report on work aimed at enhancing both the predictivity and interpretability of an EVA QSAR. This approach, referred to as EVA\_GA, uses a genetic algorithm (GA) to drive the search for better models and has been shown to give models that are statistically superior to or at least as good as those obtained with "classical EVA".

### "CLASSICAL EVA"

The "classical EVA" descriptor<sup>3,4,5</sup> is derived from fundamental vibrational frequencies of which there are  $3N-6$  (or  $3N-5$  for a linear compound such as acetylene) for an  $N$ -atom structure. The frequency values from a classical normal coordinate analysis (the EigenValues) are projected onto a linear bounded frequency scale (BFS) covering the range 1 to  $4,000\text{ cm}^{-1}$  and then smeared out, and therefore overlapped, through the application of Gaussian kernels to each and every frequency value. The BFS is sampled at fixed intervals of  $L\text{ cm}^{-1}$ . The value of the EVA descriptor at a point,  $x$ , on the BFS is the sum of ampli-

tudes of the overlapped kernels at that point. This process is repeated for each dataset structure thus providing a descriptor of fixed dimension for all compounds. The final descriptor is high-dimensional consisting of 4,000/L variables and, therefore, the Partial least squares (PLS) technique<sup>6</sup> is used to provide a robust regression analysis. The aim of the EVA smoothing procedure is not to simulate an experimental IR spectrum (transition dipole data is not used and all kernels are of fixed maximum amplitude) but rather it is to apply a density function such that vibrations at slightly different frequencies in different compounds can be "overlapped" and thus compared with one another. The extent of this overlap is governed by  $\sigma$  and the proximity of vibrations on the BFS. It is therefore the case that a range of different models need be derived using various  $\sigma$  values<sup>3,4</sup> and the best model taken to be that which provides the best crossvalidation and/or test set scores; the optimal value of EVA  $\sigma$  ( $\sigma_{OPT}$ ) is thus dataset-dependent.

## EVA\_GA

In classical EVA the Gaussian kernels have a uniform fixed  $\sigma$  (*i.e.*, equal width, height and shape) for all frequencies in all compounds being analysed. This is important because it means that each frequency (*i.e.*, each part of the spectrum) is equally weighted prior to regression. In EVA\_GA<sup>7</sup> the kernel standard deviation ( $\sigma$ ) is permitted to have localised values at different regions on the BFS. This approach permits the determination of an optimal or near-optimal overlap of kernels across the spectrum, where the quality of this overlap is judged by the scores from subsequent PLS regression with the derived descriptor matrix. Equal weighting of frequencies prior to analysis is ensured by scaling the kernels such that they have unit maximum amplitude; the main difference between the kernels is thus their width and to a lesser extent shape.

For EVA\_GA the BFS is divided up into NBINS bins of equal size and a localised  $\sigma$  linked with each bin. A frequency value falling within a bin range is thus expanded using the associated local  $\sigma$ . A GA is used to drive the search for optimal combinations of localised  $\sigma$  – a GA chromosome consists of a vector of NBINS  $\sigma$  values. A typical value of NBINS is 100 giving a bin width of 40  $\text{cm}^{-1}$ . PLS leave-one-out crossvalidation (LOO CV) regression scores have been used as the fitness function to be optimised by the GA and the final solution(s) validated using an "unseen" test set of compounds. Results with EVA\_GA have thus far been extremely promising<sup>7</sup> with substantial improvements in both  $q^2$  and test set predictive- $r^2$  ( $pr^2$ ) scores with a set of melatonin ligands<sup>8</sup> (Table 1) and improvement in  $q^2$  but no change in  $pr^2$  when applied to a benchmark steroid dataset (not shown).

**Table 1.** Some "classical EVA" and EVA\_GA results: melatonin receptor ligands<sup>8</sup>

Method	Comments	Training Set			Test Set predictive- $r^2$	
		LOO $q^2$	NLatent	$r^2$	With / Without two outliers	
CoMFA	Steric/Electrostatic 1Å grid	0.69	3	0.86	0.72	0.71
EVA	Best Fixed $\sigma = 10 \text{ cm}^{-1}$	0.46	2	0.79	0.67	0.81
EVA_GA	Best of 10 GA runs	0.77	3	0.92	0.74	0.90

Whilst these results are very promising it has been found<sup>6</sup> that a great deal of care is required to prevent training set overfit, even where LOO CV  $q^2$  is used as the GA fitness score. Current work is centred upon evaluating the effectiveness of alternative scoring functions such as, for example, leave- $n$ -out CV, where  $n > 1$ . In addition the GA maybe applied as a variable selection / deletion tool wherein a variable can be deselected when a localised  $\sigma$  of zero is permitted (Note, that in the current version of EVA\_GA a zero valued  $\sigma$  is not permitted). Such variable selection may provide simplified models which in turn may provide greater opportunity to effectively back-track to structure from an EVA QSAR. Model interpretation is one of the most appealing features of the CoMFA method while at present such ready back-transformation is not available within EVA. With this purpose in mind we are also investigating the use of alternative techniques such as continuum regression<sup>8</sup> (CR) and various variable selection procedures that in combination may provide appropriate reduced-variable models.

## REFERENCES

1. C. Hansch and A. Leo. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, ACS Professional Reference Book, American Chemical Society, Washington, DC (1995).
2. H. Kubinyi (Ed.) *3D QSAR in Drug Design. Vol. 1: Theory, Methods and Applications*; ESCOM: Leiden (1993).
3. A.M. Ferguson, T.W. Heritage, S.E. Pack, L. Phillips, J. Rogan, and P.J. Snaith, EVA: A new theoretically-based molecular descriptor for use in QSAR/QSPR analysis, *J. Comput.-Aided Mol. Design* 11:143 (1997)
4. D.B. Turner, P. Willett, A.M. Ferguson, and T.W. Heritage. Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies: 1. General application. *J. Comput.-Aided Mol. Design* 11:409 (1997).
5. D.B. Turner, and P. Willett Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies: 2. model validation using a benchmark steroid dataset, *J. Comput.-Aided Mol. Design* submitted.
6. S. Wold, A. Ruhe, H. Wold, and W.J. Dunn III, The colinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. of Scientific and Stat. Computing* 5:735 (1984).
7. D.B. Turner, and P. Willett Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies: 3. The use of genetic algorithm to find models with enhanced predictive properties. Manuscript in preparation.
8. S. Sicsic, I. Serraz, J. Andrieux, B. Brémont, M. Mathé-Allainmat, A. Poncet, S. Shen, and M. Langlois, Three-dimensional quantitative structure-activity relationship study of melatonin receptor ligands: a comparative molecular field analysis study. *J. Med. Chem.*, 40:739 (1997).



# 3D-QSAR, GRID DESCRIPTORS AND CHEMOMETRIC TOOLS IN THE DEVELOPMENT OF SELECTIVE ANTAGONISTS OF MUSCARINIC RECEPTOR

Paola Gratteri<sup>1</sup>, Gabriele Cruciani<sup>2</sup>, Serena Scapecchi,  
M. Novella Romanelli<sup>1</sup> and Fabrizio Melani<sup>1</sup>

<sup>1</sup>Department of Pharmaceutical Science, University of Florence, Via G. Capponi 9, I-50121 Florence, Italy

<sup>2</sup>Department of Chemistry, University of Perugia, Via Elce di Sotto 10, I-06100 Perugia, Italy

A dataset consisting of 91 M<sub>1</sub>, M<sub>2</sub> and M<sub>3</sub> muscarinic antagonist<sup>1-3</sup> was considered in order to create a 3D-QSAR model able both to predict the activities of compounds not yet synthesized and to evidentiare the structural features of the ligands for a specific receptor subtype selectivity. For example, M<sub>2</sub> selective antagonist could be very useful in the treatment of Alzheimer disease, a pathological state in which central cholinergic activity is reduced. In fact, a possible approach to increase the central cholinergic tone could be the blockade of the presynaptic autoreceptors, that modulate Ach release and probably belong to the M<sub>2</sub> subtype.

Recently we have reported on the synthesis and antimuscarinic properties of a new class of compounds of the general formula reported in Figure 1.

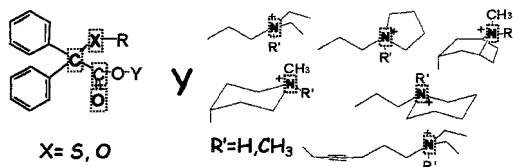


Figure 1. Chemical structure of the molecules of the dataset

The training set was generated firstly describing the overall dataset by the interaction fields derived from the Grid program using two different probes reflecting two different types of interactions. Then the 3D maps obtained were treated with the VolSurf program

and transformed in 36 2D descriptors. The PCA performed on the descriptor matrix (91x36) allowed the selection of the best representative molecules (14) constituting the training set.

The alignment of the fourteen molecules was performed on the basis of pharmacophoric crucial centers (in bold and marked by dotted lines in Figure 1). For the compounds containing the charged N-H group, the direction of the N-H bond was also taken into account. Grid program was then used on the training set for the generation of the molecular descriptors (probe OH) and the multivariate statistical analyses were performed using Golpe program. The mutual position of  $M_1$ ,  $M_2$  and  $M_3$  in the pls partial weights plot of Figure 2 shows that it is not possible to increase one of the three biological activities without simultaneously increase the others.

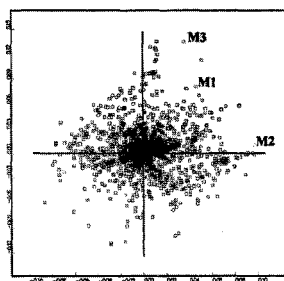


Figure 2. pls partial weights plot of the dataset

However,  $M_2$  and  $M_3$  are almost independent, thus at least two regions exist in the real space of the molecules (Figure 3a and 3b) where a structural modification allows a stronger selective increase in  $M_2$  and  $M_3$ .

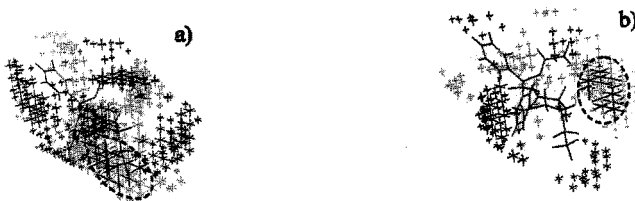


Figure 3. pls partial weights GRID plot relative to a)  $M_2$  response and b)  $M_3$  response

From the analysis of the GRID fields correlated to  $M_2$  and  $M_3$  (bigger crosses variables in Figure 3a and 3b) it is possible to highlight two regions which favour  $M_2$  selectivity located in proximity of the ammonium and in front of the ether oxygen atom. A ramification in N and a stronger attractive interaction with the lone pairs of the oxygen will give back to positive effect in  $M_2$  selective interactions. Other two regions exist which tend to favour  $M_3$  selectivity. These regions are located only close to the ammonium group. For  $M_3$  a N- $\text{CH}_3$  derivative is better than both N-H and N- $\text{C}_2\text{H}_5$ .

1. Scapecchi S., Angeli P., Dei S., Gualtieri F., Marucci G., Moriconi R., Paparelli F., Romanelli M.N., Teodori E., *Bioorg. Med. Chem.* **1994**, *2*, 1061-1074.
2. Romanelli M.N., Hölting H.D., Scapecchi S., *Quant. Struct-Act. Relat.*, **1995**, *14*, 126-143.
3. Scapecchi S., Angeli P., Dei S., Ghelardini C., Gualtieri F., Marucci G., Paparelli F., Romanelli M.N., Teodori E., *Arch. Pharm. Pharm. Med. Chem.*, **1997**, *330*, 122-128.

## SMALL CYCLIC PEPTIDE SAR STUDY USING APEX-3D SYSTEM: SOMATOSTATIN RECEPTOR TYPE 2 (SSTR2) SPECIFIC PHARMACOPHORES

Larisa Golender<sup>1</sup>, Rakefet Rosenfeld<sup>1</sup>, Erich R. Vorpagel<sup>2</sup>

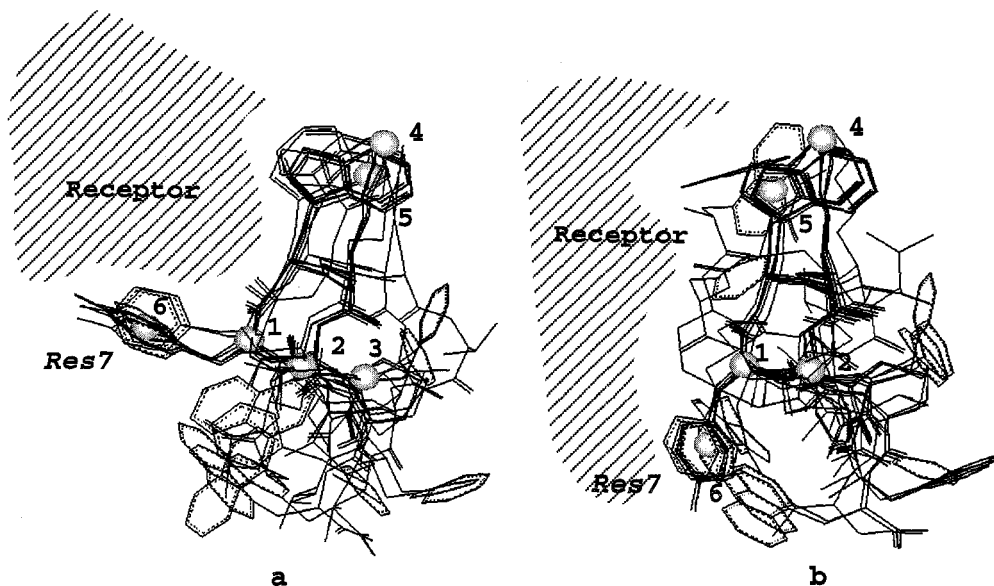
<sup>1</sup>Peptor Ltd., Kiryat Weizmann, Rehovot 76326, Israel

<sup>2</sup>Molecular Simulations Inc., 9685 Scranton Road, San-Diego, CA  
92121-2777

Somatostatin (SST), Gly-Ala-Cys-Lys-Asn-Phe<sup>6</sup>-Phe<sup>7</sup>-Trp<sup>8</sup>-Lys<sup>9</sup>-Thr<sup>10</sup>-Phe<sup>11</sup>-Thr-Ser-Cys, is a pleiotropic regulatory hormone whose functions are mediated by a family of 5 G-protein coupled receptors (SSTR1-5). SST binds non-selectively to all five subtypes, and is unstable under physiological conditions. Numerous small, stable synthetic SST analogs that display varying degree of selectivity have been identified<sup>1</sup>. Detailed knowledge of the 3D-structure of the receptor recognition sites (pharmacophores) is necessary for rational design of new SST based drugs.

We present here results of a computer-aided pharmacophore identification study using a learning Activity Prediction Expert System APEX-3D<sup>2</sup> (MSI). The study was performed on the set of 11 small cyclic peptides whose binding activity to SSTR2 has been well established in different research groups. Pharmacophores were defined as spatial arrangements of Descriptor Centers (DC) common to all active compounds. The following atoms and groups were defined as DCs: C $\alpha$  (any) of the peptide backbone, C $\alpha$  of Trp and C $\alpha$  of Lys (specifically) to represent the peptide backbone and facilitate a proper conformer alignment; N of amine (NH<sub>2</sub>), O of hydroxyl group (OH), C of methyl-group (CH<sub>3</sub>) and center of aromatic ring (CAR). 300 conformers of each compound were generated by a Molecular Dynamics/Energy Minimization (INSIGHTII/DISCOVER) procedure starting from experimental structures. Conformers were clustered, and 35-45 conformers, each representing one cluster, were loaded into the Learning Structure Data Base (LSDB). Pharmacophores were identified and ranked, and the bioactive conformations of each LSDB compound, namely those best displaying the pharmacophore, were extracted.

The **Figure** displays the two highest ranked (based on statistical criteria and molecule shape fit) pharmacophores, which represent two possible shapes of the receptor recognition site - "pocket" and "flat".



**Figure.** LSDB "active" conformers superimposition on *a* - "pocket" shape and *b* - "flat" shape pharmacophores. Light spheres show the Descriptor Centers: 1-3 - C $\alpha$ ; 4 - N(NH $_2$ ); 5,6 - center of aromatic ring.

The geometry of the "active" conformers found in our calculation for L363,301 and Sandostatin (compounds studied experimentally in most detail) is consistent with experimental findings<sup>3-5</sup>. Analysis of pharmacophoric superimpositions of all the LSDB molecules allowed us to formulate a new hypothesis of the receptor recognition site, which is different from the earlier proposed topological scheme<sup>4</sup>. Our recognition site model involves the aromatic residue in position 7 (*Res7*) and is localized on the "back" of the peptide backbone, whereas the published model involves aromatic residue in position 11 and is localized "in the backbone fold".

We tested our pharmacophores for prediction ability on the series of novel compounds with known binding activity to SSTR2. Rules for activity prediction for small cyclic peptides were formulated based on the presence of the three highest ranked pharmacophores and the shape of molecular volume. 3D-search queries were generated from these pharmacophores for MDL IS Data Base mining. MDDR-3D Data base search hits contained nearly all compounds registered in this Data base as somatostatin analogs, GH secretion inhibitors and GH secretion promoters along with many other compounds (most of them GPCR ligands), thus proving the informational value of the pharmacophores and providing new lead candidates for somatostatin based drug design .

## References

1. T.Reisine, G.I. Bell, Molecular biology of somatostatin receptors, *Endocr. Rev.* 16:427(1995)
2. V.Golender et al., Knowledge engineering approach to drug design and its implementation in the APEX-3D Expert System, in: *QSAR and Molecular Modelling*, F.Sanz , J.Giraldo and F.Manaut eds., J.R. Prous Science Publishers, Barcelona (1995).
3. D.F.Veber, Design and discovery in the development of peptide analogs, in: *Peptides. Chemistry and Biology*, J.A.Smith and J.E.Rivier eds., ESCOM, Leiden (1992).
4. Z.Huang, Ya-Bo He, K.Raynor, M.Tallent, T.Reisine, M.Goodman, Main chain and side chain chiral methylated somatostatin analogs: synthesis and conformational analysis, *J.Am.Chem.Soc.* 114:9390(1992).
5. G.Melacini, Qin Zhu, G.Osabay, M.Goodman, A refined model for somatostatin pharmacophore: conformational analysis of lanthionine-sandostatin analogs, *J.Med.Chem.* 40:2252(1997).

### **3D QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (COMFA) STUDY OF HETERO- CYCLIC ARYLPYPERAZINE DERIVATIVES WITH 5-HT<sub>1A</sub> ACTIVITY**

Ildikó Magdó, István Laszlovszky, Tibor Ács, György Domány

Gedeon Richter Ltd., H-1475 Budapest, P.O.B. 27, Hungary  
e-mail: i.magdo@richter.hu

#### **INTRODUCTION**

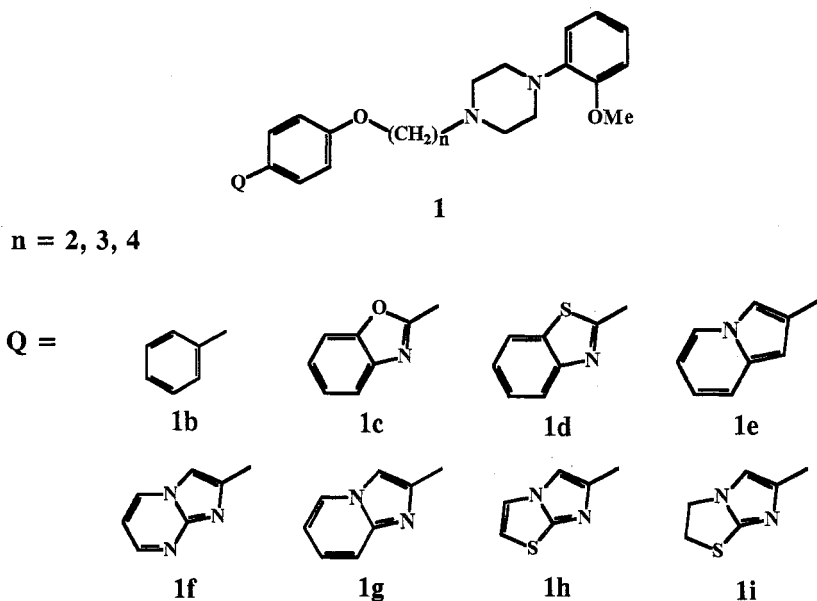
Pharmacological treatment of schizophrenia has been traditionally dominated by dopamine D<sub>2</sub> receptor antagonists, known to cause severe extrapyramidal side effects, which can be attributed to the blockade of D<sub>2</sub> receptors in the striatum. Current antipsychotic research focuses on compounds with multireceptorial activity (different dopamine receptor subtypes, serotonin and adrenergic and muscarinic receptors has been studied). Recent observations indicate, that control of both dopaminergic and serotonergic systems is important for adequate antipsychotic therapy. It has been reported, that 5-HT<sub>1A</sub> receptor agonists reverse antipsychotic induced catalepsy.

Our aim was to achieve compounds with combined 5-HT<sub>1A</sub>/D<sub>2</sub> activity. We have synthesised and evaluated a series of novel arylpiperazine derivatives (see Fig. 1.) several of them showing high affinity for the 5-HT<sub>1A</sub> receptor beside the D<sub>2</sub> activity<sup>1</sup>.

#### **COMFA ANALYSIS**

We have performed a CoMFA analysis using the molecular modelling package SYBYL<sup>2</sup> in order to create a model with which the 5-HT<sub>1A</sub> activity can be predicted. Within the series of compounds with equal n (same chain length) the Q-Phe-OMe fragment of the molecules were aligned. The geometries were optimised and the charges were calculated with AM1 method using the MOPAC module of the package.

The molecules were surrounded by a 3D grid of 2 Å resolution extending 4 Å beyond the union volumes of the superimposed molecules. The electrostatic and steric field was calculated at the gridpoints using the default C(sp<sup>3</sup>) probe atom with a +1 charge. The maximum cutoff values were set to +4 and +1 kcal/mol for the steric and the electrostatic fields respectively.



**Figure 1.** Evaluated heterocyclic arylpiperazine derivatives

PLS analysis was performed using the leave one out cross validation technique to obtain the optimum number of components for the electrostatic field alone and for the electrostatic and steric field together. Addition of the steric field -as expected- did not improved the results significantly.

The final CoMFA model obtained for the series of molecules with  $n=4$  using 2 components had  $r^2 = 0.951$ ,  $s = 0.168$ ,  $F = 48.06$ . The predictive ability of the model was also good ( $R^2_{cv}=0.761$ ).

## CONCLUSIONS

We have succeeded to get a reasonable 3D QSAR (CoMFA) model for the 5-HT<sub>1A</sub> activity of a series of heterocyclic piperazine derivatives. According to the final model there is a strong correlation between the activity and the electrostatic field around the heterocycles (Q). By studying the PLS coefficient contour plots useful information can be gained for the synthesis of new potent compounds.

## REFERENCES

- I. Laszlovszky, Gy. Domány, T. Ács, Gy. Ferenczy, Cs. Szántay, Jr., E. Thúróczyné Kálmán, Preparation of 2-methoxyphenylpiperazine derivatives as antipsychotics, *PCT Int. Appl. WO 9818797* (1998)
- Tripos Associates, SYBYL, version 6.4 (1997)

## Molecular Similarity Analysis and 3D-QSAR of Neonicotinoid Insecticides

Masayuki Sukekawa and Akira Nakayama

Odawara Research Center, Nippon Soda Co., Ltd.  
345 Takada, Odawara, Kanagawa, 250-0280, Japan

A number of neonicotinoid insecticides, such as imidacloprid and acetamiprid (Fig.1), have been developed as agonists of nicotinic acetylcholine receptor (nAChR). In this study, a new method of molecular similarity analysis<sup>1</sup> was applied to the three-dimensional quantitative structure-activity relationship (3D-QSAR) of neonicotinoid insecticides.

A novel electrostatic similarity  $IR_{AB}$  was defined by Eq.1, where  $EA$  and  $EB$  are the vectors of electrostatic potentials ( $\epsilon_{Ai}$  and  $\epsilon_{Bi}$ ) at each grid point  $i$  around the molecules A and B, respectively, when they are superimposed. In the same manner, a novel shape-similarity index  $IS_{AB}$  was defined by Eq.2. The grid values ( $S_{Ai}$  and  $S_{Bi}$ ) which take unity when a grid point  $i$  is inside of the van der Waals surfaces of molecules A and B, respectively, and otherwise zero.

$$IR_{AB} = \vec{EA} \cdot \vec{EB} = \left| \vec{EA} \right| \left| \vec{EB} \right| \cos(\theta) \quad (1)$$

$$\vec{EA} = (\epsilon_{A1}, \epsilon_{A2}, \epsilon_{A3}, \dots, \epsilon_{AN}), \quad \vec{EB} = (\epsilon_{B1}, \epsilon_{B2}, \epsilon_{B3}, \dots, \epsilon_{BN})$$

$$IS_{AB} = \vec{SA} \cdot \vec{SB} = \left| \vec{SA} \right| \left| \vec{SB} \right| \cos(\theta) \quad (2)$$

$$\vec{SA} = (S_{A1}, S_{A2}, S_{A3}, \dots, S_{AN}), \quad \vec{SB} = (S_{B1}, S_{B2}, S_{B3}, \dots, S_{BN})$$

A whole series of  $m$  molecules are compared with each other to give  $m \times 2m$  similarity matrix. The biological activity ( $y$ ; dependent variable) is expressed as a linear combination of the similarity indices in the matrix (Eq.3). The PLS method was applied to analyze the correlation between the similarity indices and the activity.

$$y = a_A IR_{XA} + a_B IR_{XB} + \dots + a_M IR_{XM} + b_A IS_{XA} + b_B IS_{XB} + \dots + b_M IS_{XM} + Const. \quad (3)$$

The above method was applied to the QSAR of the neonicotinoid insecticides (Fig.1). The model molecules, in which the 6-chloro-3-pyridyl group was replaced by hydrogen, were used in the similarity analysis. The Eq.4 was obtained as a quantitative correlation model of the binding activity (pKi) against nAChR and the similarity indices.

$$pKi = 8.862 + [ IR_{AB} \text{ and } IS_{AB} \text{ terms} ] \\ n = 12, A = 3, r = 0.945, R_{pred} = 0.677, s = 0.477, F = 22.340 \quad (4)$$

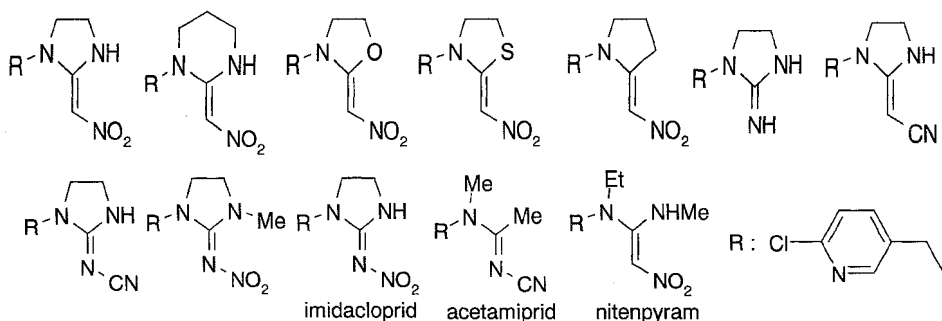


Fig.1 Structure of neonicotinoid insecticides used in the analysis.

The analysis only for shape or electrostatic similarity gave less significant results than Eq.4. This indicates that both the similarities in steric and electrostatic properties are important for the activity. In our previous study<sup>2</sup>, the binding activity was quantitatively correlated with the electrostatic-similarity index<sup>3,4</sup> of each molecule compared with the most active compound. The result shown by Eq.4 coincides with our previous study, and should be more reliable to predict the activity since the obtained QSAR model is based upon the molecular similarity and dissimilarity of the whole series of compounds being compared. The obtained QSAR model is established on the basis of the direction and magnitude of the similarity vectors as shown above. Therefore, the contribution of similarity vectors can be calculated, and the ten grid points which contributed most significantly to the activity were shown by spheres in Fig.2.

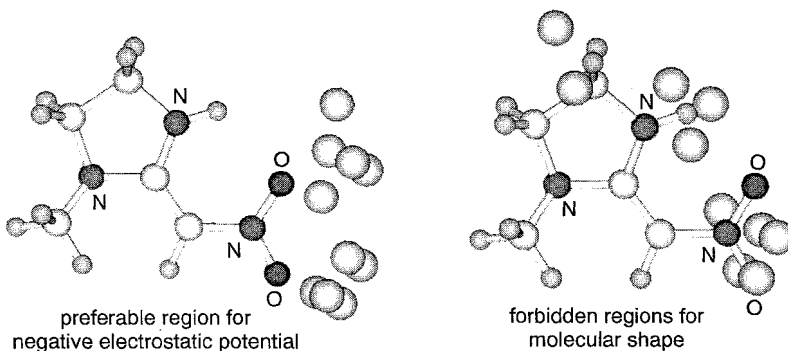


Fig.2 Spherical representation of structural requirements for the receptor-binding activity.

Okazawa *et al.* recently performed the 3D-QSAR of neonicotinoids by CoMFA<sup>5</sup>, and their results looked similar to our results shown above. Lobato *et al.* have recently reported that the technique of quantumchemical and topological similarity indices gives comparable or better results than the one by the current 3D-QSAR procedures such as CoMFA<sup>6</sup>. The method of molecular similarity analysis presented in this study may become one of the useful tools in this aim.

#### Reference

1. M. Sukekawa and A. Nakayama, *24<sup>th</sup> Symp. Struct.-Act. Relat.*, Abstr. No.12S06, Osaka, 1996
2. A. Nakayama and M. Sukekawa, *Pestic. Sci.*, **52**, 104(1998)
3. R. Carbó, L. leyda and M. Arnau, *Int. J. Quantum Chem.*, **17**, 1185(1980)
4. E. E. Hodgkin and W. G. Richards, *Int. J. Quantum Chem. Quantum Biol. Symp.*, **14**, 105(1987)
5. A. Okazawa *et al.*, *Pestic. Sci.*, in press.
6. M. Lobat, L. Amat, E. Besalú and R. Carbó-Dorca, *Quant. Struct.-Act. Relat.*, **16**, 465(1997)



## 3D-SAR STUDIES ON A SERIES OF SULFONATE DYES AS PROTECTION AGENTS AGAINST $\beta$ -AMYLOID INDUCED *IN VITRO* NEUROTOXICITY

M.G. Cima\*, G.Gallo\*, M.Mabilia<sup>o</sup>, M.O.Tinti\*, M.Castorina\*, C. Pisano \*, E. Tassoni\*

\*Direzione Ricerche Sigma-Tau, Via Pontina km 30,400, I-00040 Pomezia (RM) Italy.

<sup>o</sup>S.IN - Soluzioni Informatiche S.a.s., Via Salvemini 9, I-36100 Vicenza, Italy

### INTRODUCTION

Alzheimer's disease (AD), the most common form of dementia in elderly people, is characterized by the extracellular deposition of 39-43 amino acid peptide referred as amyloid  $\beta$ -peptide ( $A\beta$ ). The mechanism by which  $A\beta$  elicits its toxicity is poorly understood, however the aggregation of the peptide into fibrils has emerged as a major factor in  $A\beta$  toxicity.<sup>1</sup>

Congo Red (CR) is a sulfonated diazo dye that stains  $A\beta$  fibrils, inhibits the fibrillation of  $A\beta$ <sup>2</sup> and attenuates the toxic effects of either  $A\beta$  (25-35) and  $A\beta$  (1-40).<sup>3</sup> The protective effects of these compound may results from: **1.** inhibition or reversal of  $A\beta$  aggregation, **2.** inhibition of the peptide binding to cells or **3.** blocking access of bound peptide to cell surface. A proposed model for the interaction of CR with  $A\beta$  involves a salt bridge between two sulfonate groups and positively charged lysine residues on different strands of the antiparallel  $\beta$ -pleated fibrils.<sup>2</sup>

Herein we present 3D-SAR studies on a series of sulfonate dyes, taken from literature<sup>3</sup> and from a selection on commercial catalogues, that were tested *in vitro* as protection agents against  $A\beta$  1-40 neurotoxicity in rat adrenal pheochromocytoma PC12 cells. We identified specific descriptors that can locate active, less active and inactive dyes in separated clusters.

### MATERIALS AND METHODS

**Biological tests** A set of 10 compounds were selected from commercial catalogues to be used for biological testing.

Specific tests were carried out with the aim of studying the effect of test compounds on protection of toxicity induced by  $A\beta$  (1-40)<sup>3</sup> and on inhibitory effects of  $A\beta$  (1-40) aggregation.<sup>4</sup>

**Modeling and 3D-SAR.** A set of 10 compounds was chosen to span a range of distances between sulfonate groups from 5.4 Å to 20.1 Å. In particular, the corresponding distance for CR is 20.1 Å. Another set of molecules was then selected from literature.<sup>3</sup> Minimum-energy conformations were generated (*Discover*®)<sup>5</sup> performing a conformational search (CFF91 force field, conjugate gradient method, 60° rotor, dielectric constant 10). Charges for sulfonate anion were calculated by the quantum mechanics semi-empirical method AM1 (Mopac software<sup>6</sup>). Only conformers within 3 kcal/mol from the global minimum were considered. The structural descriptors that were measured for different conformers were: distance between the two S atoms of sulfonate groups ( $d_1$ ) and between corresponding adjacent C atoms ( $d_2$ ), the torsional angle  $S_1-C_1---C_2-S_2$  ( $\theta$ ), the bond angle  $S_1-C_1---C_2$  ( $\alpha_1$ ) and angle  $S_2-C_2---C_1$  ( $\alpha_2$ ). We also calculated the difference  $|d_1 - d_2|$ , the difference  $|\alpha_1 - \alpha_2|$  and the value of  $|d_1 - NINT(d_1/D) * D|$  ( $\Delta_{4.7}$ ), where:  $D = 4.7$  Å, the inter-strand distance in a Pauling's cross- $\beta$  fibril or  $D = 11$  Å, the distance of Lys residues facing each other on adjacent sheets, NINT is a function that returns the nearest integer. 3D-SAR's were performed with TSAR<sup>TM</sup> software<sup>7</sup>.

## RESULTS AND DISCUSSION

The best 2D- classification map is obtained using  $\theta$  and  $|d_1 - d_2|$  descriptors. All inactive compounds are located on the right side of the plot with  $\theta > 80^\circ$ . Low activity structures are characterized by  $\theta < 60^\circ$  and  $|d_1 - d_2| < 1.5$  Å or  $> 2.5$  Å. The active ones have  $\theta < 40^\circ$  and  $|d_1 - d_2|$  close to 2 Å. Results of our classification analyses support the hypothesis that sulfonate groups, in CR or other molecules, must be on the same side of a given structure in order to face and interact with accessible positively-charged Lys residues. Such ionic interactions would compete and possibly disrupt specific salt bridges involving Lys 28 residues and C-terminal carboxylic functions. The relevance of descriptor  $\theta$  is supported by experimental evidence and can be easily explained in terms of specific conformational requirements of sulfonated ligands. The same cannot be said about the other descriptor here reported,  $|d_1 - d_2|$ . Although this descriptor is able to discriminate between low- and high-activity compounds, its significance is still doubtful. Other classification maps can be obtained using other descriptors based on the distance between sulfonate moieties (e.g.  $\Delta_{4.7}$ ), but in all these cases only two classes are well separated: inactive compounds on one side and high- and low-activity ones on the other. Ongoing studies are trying to address these issues to further investigate and better understand interactions between sulfonated (and/or carboxylated) compounds and A $\beta$  models.

1. L.L. Iversen, R.J. Mortishire-Smith, S.J. Pollack, and M.S. Shearman. *Biochem. J.* 311:1 (1995)
2. W.E. Klunk, M.L. Debnath, and J.W. Pettegrew *Neurobiol. Aging* 15:691 (1994)
3. S. J Pollack, I.I.J. Sadler, S. R. Hawtin, V.J. Taylor, and M.S. Sherman *Neurosciences Letter* 197: 211(1995)
4. A. Lorenzo and B.A. Yankner *Proc. Natl. Acad. Sci USA* 91:12243(1994)
5. *Discover*® 97.0 MSI, San Diego.
6. Mopac v.6.0 QCPE n°455 by J.J.P. Stewart.
7. TSAR<sup>TM</sup> 3.1 Oxford Molecular Group Inc, Beaverton.

**A NEW MOLECULAR STRUCTURE REPRESENTATION:  
SPECTRAL WEIGHTED MOLECULAR (SWM) SIGNALS AND  
SPECTRAL WEIGHTED INVARIANT MOLECULAR (SWIM) DESCRIPTORS**

Roberto Todeschini, Viviana Consonni, David Galvagni and Paola Gramatica

Milano Chemometric Research Group  
Dep. Environmental Sciences  
Milano University  
Via Emanuelli, 15  
I-20126 Milano (Italy)

A new molecular representation based on a semi-invariant decomposition of the 3D molecular structure is presented. The basic approach is the principal component analysis on the (x,y,z) atomic coordinates of a molecule, obtaining the atom projections on the three principal axes (the scores). The direction of each principal axis is uniquely defined, but not the versus.

Whereas WHIM descriptors<sup>1</sup> are invariant statistical indices calculated on the scores, SWM signals are directly obtained by weighting the scores of each axis by the weights defined in the WHIM descriptors framework (mass, polarizability, Mulliken atom charge, van der Waals volumes, electrotopological charges). Thus a molecule can be represented by a sequence of signals obtained from the weighted scores of three principal axes, giving a spectral representation: the signals are the scores along the axes and the signal intensities are the weights.

Similarity analyses based on this representation have been performed on different sets of compounds using the Camberra distance. SWM signals appear a very encouraging approach in assessing similarity among molecules, being a semi-invariant molecular representation containing detailed information about 3D-molecular structures.

New molecular descriptors can be also easily obtained by analyzing spatial autocorrelation of the SWM signals. Spectral Weighted Invariant Molecular (SWIM) autocorrelation descriptors obtained from each principal component can be calculated, together with cross-correlation descriptors between each pair of principal axes. For each WHIM weighting scheme, with a maximum lag of 5, the total number of SWIM correlation descriptors is 90 (15 autocorrelation + 75 cross-correlation descriptors).

The presence of some SWIM descriptors in a QSAR model indicates the molecular regions of interest for the considered activity/property. This allows the possibility of going from the model to the molecular structure, giving insight into the relationships between structure and activity/property.

Due to the 3D local information provided by the SWIM descriptors, the combined use with the WHIM descriptors, containing global molecular information, is recommended. Preliminary applications of these descriptors in QSAR models seem to give very interesting results, not only for the high predictive capabilities, but also for the possibility to come back effectively from descriptors to local structure features, i.e. to perform a reversible decoding.

REFERENCE

- 1) R.Todeschini and P.Gramatica, "3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and Chemical Meaning of WHIM Descriptors", *Quant. Struct.-Act. Relat.* **16**, 113-119 (1997).

## 3D QSAR OF PROLYL 4-HYDROXYLASE INHIBITORS

K.-H. Baringhaus, V. Guenzler-Pukall, G. Schubert and K. Weidmann

Hoechst Marion Roussel  
Chemical Research, Building G 838  
D-65926 Frankfurt am Main, Germany

### Introduction

Prolyl 4-hydroxylase (EC 1.14.11.2) is an important enzyme involved in collagen biosynthesis. This enzyme catalyzes the formation of 4-hydroxyproline in collagens by the hydroxylation of certain proline residues in peptide linkages<sup>1</sup>. Due to the importance of 4-hydroxyproline for the thermal stability of collagenous triple helices, inhibition of this enzyme offers an attractive target for antifibrotic treatment.

For a training set of 26 competitive inhibitors of prolyl 4-hydroxylase with affinities ranging from 55 nM to 4.4 mM, we used the program CATALYST<sup>2</sup> to derive a three-dimensional pharmacophore hypothesis.

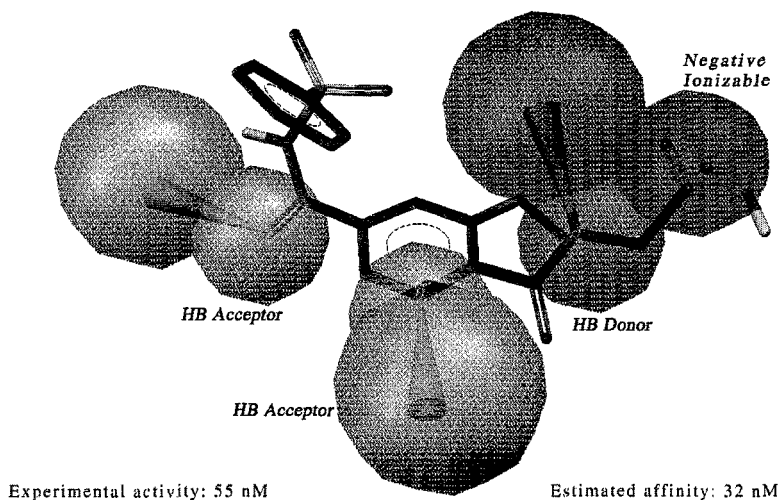
### Methods

All molecules were minimized within CATALYST to the closest local minimum using molecular mechanics. Conformational models were generated which emphasize representative coverage over a 20 kcal energy range above the computed global minimum<sup>3</sup>. Using these conformational models, the training set was submitted to hypothesis generation<sup>3</sup> which aims to identify the best 3D spatial arrangement of chemical functions explaining the activity variations among the training set. The chemical functions used in the hypothesis generation step include hydrogen bond donors and acceptors, hydrophobic groups and negative ionizable functions.

The resulting model was validated with compounds outside the training set and by a subsequent CoMFA<sup>4</sup> study.

### Discussion

The best hypothesis proposed by CATALYST is characterized by 1 Negative Ionizable function, 2 H-bond Acceptor and 1 H-bond Donor feature. Figure 1 shows 0570 (the most active compound of our training set) superimposed on the hypothesis. This compound maps all features of our model and its activity is properly estimated.



**Figure 1.** Alignment of 0570 to the prolyl 4-hydroxylase hypothesis

A set of 20 diverse prolyl 4-hydroxylase inhibitors, different from the members of the training set, was chosen for activity prediction by our current hypothesis. The entire validation set shows a good correlation between the estimated and experimental activities, proving the predictive power of this model.

All compounds of the training set were aligned to our hypothesis for a subsequent Comparative Molecular Field Analysis<sup>4,5</sup> (2 Å grid, steric and electrostatic fields; 30 kcal/mol cutoff) to check the reliability of our model. The PLS analysis (minimum sigma cutoff of 2.0 kcal/mol) revealed a cross-validated  $R^2$  of 0.424 for five components. The CoMFA model explains the variance in the biological data for the 26 compounds within the training set reasonably well, indicating the relevance of the underlying hypothesis for the alignment of the inhibitors. Furthermore, the crossvalidated  $R^2$  of 0.424 suggests that the model should have acceptable predictivity for similar molecules not present in the training set.

## Conclusion

Starting from a set of 26 competitive prolyl 4-hydroxylase inhibitors, we generated a four-feature hypothesis that well explains the affinities of the molecules. This model was validated by an external data set and by a subsequent CoMFA study. Both models were successfully applied in lead optimization of prolyl 4-hydroxylase inhibitors.

## REFERENCES

1. T. Pihlajaniemi, R. Myllylä, K. Kivirikko, *J. Hepatol.* **1991**, *13 Suppl.* 3, S2.
2. Catalyst 3.1, Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121, USA.
3. P.W. Sprague, *Perspect Drug Discovery Des.* **1995**, *3*, 1.
4. R.D. Cramer, D.E. Patterson, J.D. Bunce, *J. Am. Chem. Soc.* **1998**, *110*, 5959.
5. SYBYL Molecular Modeling Package, version 6.4, Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

# AROMATASE INHIBITORS: COMPARISON BETWEEN A COMFA MODEL AND THE ENZYME ACTIVE SITE

Andrea Cavalli,<sup>1</sup> Maurizio Recanatini,<sup>1</sup> Giovanni Greco<sup>2</sup> and Ettore Novellino<sup>2</sup>

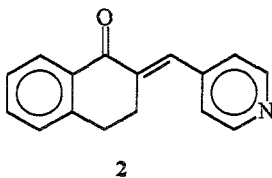
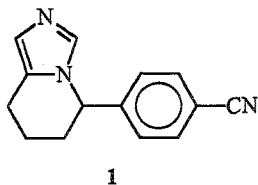
<sup>1</sup>University of Bologna, Dept. of Pharmaceutical Sciences,  
Via Belmeloro 6, I-40126 Bologna, Italy

<sup>2</sup>University of Napoli "Federico II", Dept. of Pharmaceutical and Toxicological  
Chemistry,  
Via D. Montesano 49, I-80131 Napoli, Italy

## INTRODUCTION

Aromatase inhibitors are among the most actively studied compounds in the field of antitumour agents, because of their role in the treatment of breast cancer. Aromatase is a cytochrome P 450 isozyme (P 450 XIX), that can be inhibited either competitively or non competitively by various classes of steroidal and non-steroidal compounds.

Recently, we developed a CoMFA model for the aromatase inhibition by two series of non-steroidal agents (represented by the lead compounds **1** [*S*-fadrozole] and **2**), that allowed us to define on a statistical basis the steric and electrostatic optimal requirements for inhibitors belonging to those classes.<sup>1</sup>



The need of *building bridges* between three-dimensional protein models and 3D-QSAR studies was recently pointed out by Kim, who showed how the two methods can act synergistically in providing useful information towards the goal of ligand design.<sup>2</sup>

In order to investigate this issue in more depth, we built a three-dimensional model of aromatase and compared it with the results of our previous 3D-QSAR analysis. The first step was accomplished by means of a homology building procedure aimed at modeling the main features of the aromatase active site. Then, the steric and electronic characteristics of the space allowed for inhibitors as statistically defined by the CoMFA study were checked against the modeled enzyme active site.

## COMPARISON BETWEEN THE HOMOLOGY BUILT AND COMFA MODELS

Superimpositions of the aromatase active site with the CoMFA steric and electrostatic contours were examined. There is a general agreement between the position of the favorable and unfavorable steric and electrostatic CoMFA regions and the residues forming the active site cavity. The CoMFA sterically allowed area corresponds to an empty region of the active site, while the unfavorable volume partly overlaps with the side chain of Thr310 (helix I). The electrostatic CoMFA red contour surrounds Asp309 and, referring to the CoMFA model, the presence of the carbonyl group of **2** in that zone is unfavorable. This effect might originate from the interaction of the carbonyl functions of the inhibitors with the electron cloud of the COO<sup>-</sup> of Asp309.

One particular aspect that emerged after the docking simulations of **1** and **2** into the aromatase active site is that the inhibitors are mutually oriented in a somewhat different manner from the alignment used in the CoMFA analysis. This confirms that alignments leading to statistically significant CoMFA models do not need to reproduce the results of docking simulations or experimental determinations.

## DISCUSSION

3D-QSAR and homology built protein models provide the drug designer with different kinds of information: it is possible (perhaps desirable) to compare the SAR derived from both the ligand-based and the target-based analyses and to verify the consistency of the conclusions. In the case of the non-steroidal aromatase inhibitors, we found a satisfying correspondence between the quantitative and the qualitative models in terms of the steric and electrostatic properties of both ligands and enzyme.

*Building a bridge* between CoMFA and docking models allows one to take advantage of the strengths of both methods in view of a better comprehension of the enzyme-inhibitor interactions. The CoMFA contours are statistical artifacts which bear no physical meaning, but if they are overlapped onto the active site surface, they may eventually be understood in terms of the presence of aminoacid residues. In turn, a ligand-protein docking model is limited to the explanation of one compound's structure-activity relationships and its integration with a 3D-QSAR model might expand the results of the analysis to a class of congeners.

Checking a CoMFA alignment against a docking model based on a dynamics simulation also points out the issue of how different ligands should be oriented inside the enzyme active site. In the present case, a highly significant 3D-QSAR was obtained, despite an alignment not confirmed by the dynamics simulation. However, a CoMFA performed using the alignment suggested by the molecular dynamics gave comparable statistical results.

## REFERENCES

1. M. Recanatini and A. Cavalli, Comparative molecular field analysis of non-steroidal aromatase inhibitors: an extended model for two different structural classes, *Bioorg. Med. Chem.* 6:377 (1998).
2. H.K. Kim, Building a bridge between G-protein coupled receptor modelling, protein crystallography and 3D-QSAR studies for ligand design, *Persp. Drug. Des. Disc.* 12/13/14:233 (1998).

# IMIDAZOLINE RECEPTOR LIGANDS – MOLECULAR MODELING AND 3D-QSAR CoMFA

C. Marot<sup>a</sup>, N. Baurin<sup>a</sup>, J. Y. Mérour<sup>a</sup>, G. Guillaumet<sup>a</sup>, P. Renard<sup>b</sup>, L. Morin-Allory<sup>a</sup>

<sup>a</sup>Institut de Chimie Organique et Analytique, associé au CNRS, Université d'Orléans BP6759, 45067 ORLEANS. <sup>b</sup>A.D.I.R, 1 rue Carle Hébert, 92415 COURBEVOIE

15 years ago, studies aiming at developing new-line central  $\alpha_2$  adrenergic drugs gave birth to the increasingly recognized concept of non-adrenergic imidazoline receptors [1]. Two major subtypes of imidazoline receptors have been isolated at this time.  $I_1$  receptors, mainly central, whose activation brings about a reduction of elevated blood pressure.  $I_1$  receptors have been recognized as a target of centrally acting antihypertensives devoid of the intense side effects mediated by  $\alpha_2$  receptors. However, conclusive evidence for their existence is still lacking.  $I_2$  receptors, in contrast to the  $I_1$  binding sites, have a much wider tissue distribution and can be subdivided into  $I_2$ -A and  $I_2$ -B sites. No definitive physiological role has yet been determined although their functional role is established, as mediators of neuroprotection in ischemic infarction. Further insights into the imidazoline receptor scope (topology, functionality, localization, distribution, and pharmaco-applications) include the development of more selective compounds. In this connection, a 3D-QSAR study using CoMFA is a powerful tool as it may produce a 3D pharmacophoric model of the ligands defining the spatial region where electrostatic, lipophilic and steric interactions may modulate the binding affinity. A 3D-QSAR CoMFA study was then carried out on *in vitro*  $I_2$  binding affinities of 109 2-substituted imidazoline compounds : an  $I_2$  3D-QSAR model, with good fitting and predictive abilities, is presented.

## Methodology of the 3D-QSAR CoMFA study

**Hardware** - Silicon Graphics Indy (R4600), Indigo2 (R4400) & O2 (R10000)

**Software** - SYBYL v. 6.3 & 6.4 (Tripos Associates, St Louis, MO, USA)

**Ligands** - The structural and biological data, provided by the laboratory and literature, were used to build a database containing 109 molecules. Representing about 10 chemical families (naphthalene, benzene, benzopyran, benzodioxane...), this database presents an essential homogeneity of the binding data ( $pIC_{50}$  range: 4.3 to 9.2) as well as a very interesting molecular diversity for the robustness of the model.

**Conformational analysis** - Ligands were modeled and optimized with SYBYL (6.3 and 6.4) *via* a MOPAC semi-empirical calculation using the AM1 Hamiltonian. Each structure was then submitted to a Monte-Carlo conformational analysis implemented in SYBYL (Random Search): energy minimization, using TRIPOS force field, includes MOPAC partial atomic charges, which better account for the mesomery of the physiological protonated imidazoline ring. All the generated conformers, within a 70 kcal/mol energy range, were then screened through a SPL automatic fitting procedure onto the pharmacophoric elements of the template: the quality of the fit was assessed by RMS (Root Mean Squares).

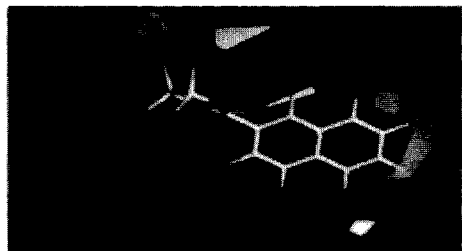
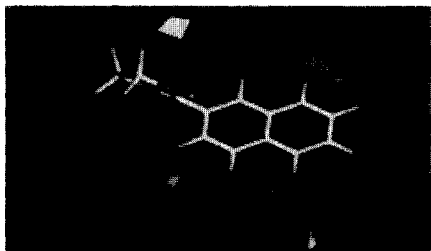
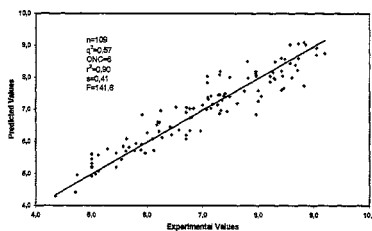
**Determination of the 3D-pharmacophore** - In order to select the local minimum conformer among several available after the Random Search procedure, benzoline [2] was chosen as a template because of its high  $I_2$  affinity and low conformational mobility. As an essential element for the CoMFA alignment step, the choice of the benzoline conformer



was covalidated by Random Search, Systematic Search and Simulated Annealing conformational analysis : among the 2 conformers covalidated, the template ( $\phi=-42.6^\circ$ ) was qualified *via* a RMS fitting procedure on 10 compounds with high affinity for  $I_2$ .

**Molecular alignment** - The different point-by-point alignment rules envisaged were applied with an SPL automatic fitting procedure onto the associated pharmacophoric elements of the template: for each alignment, the CoMFA table was then calculated and the PLS method run.

**A 3D-QSAR CoMFA model of  $I_2$  receptor** - PLS is used as the regression method to develop the relationship between independent variables (steric, lipophilic & electrostatic potentials) and dependent variable ( $pIC_{50}$ ). First, the optimal number of components (ONC) and  $q^2$ , measuring the predictive ability of the model, are determined using the Leave-One-Out cross validation technique. Second, PLS, using the ONC, gives the final model, from which the isocontour map is drawn, and  $r^2$ , measuring the fitting ability of the model. The lipophilic field was calculated by the MLP implemented in the CLIP [3] module of SYBYL. Among the different alignments realized, the model yielding the best statistics, in terms of predictive ability ( $q^2=0.57$ ), is combining lipophilic (52%) and steric (48%) fields: the associated isocontour maps indicate the regions where the variations in lipophilic and steric potentials of the 109 compounds are correlated with the variation of  $pIC_{50}$ .



**Conclusions and prospects** - The CoMFA study on  $I_2$  *in vitro* binding affinity of a large series of 2-substituted imidazoline compounds is yielding an  $I_2$  3D-QSAR model presenting a good predictive ability and explained variance: the associated CoMFA isocontour maps revealed spatial regions where lipophilic and steric interactions may modulate the *in vitro*  $I_2$  binding affinity. Compared to other works [4] using the same template without lipophilic fields, this  $I_2$  3D-QSAR model, based on a much wider range of structurally diverse compounds, presents a slightly lower predictive ability. With the aim at improving this model, the determinant CoMFA alignment step is thoroughly explored with a genetic algorithm-based procedure. The CoMFA methodology is at the moment employed to develop  $I_1$  and  $\alpha_2$  3D QSAR models which, with the  $I_1$  model, could give access to physicochemical and structural requirements for  $I_1/\alpha_2$ ,  $I_2/\alpha_2$  and  $I_1/I_2$  selectivity.

1. A.Yu and W.H.Frushman: 'Imidazoline receptor agonist drugs :a new approach to the treatment of systemic hypertension' . *Journal of Clinical Pharmacology* 1996; 36: 98-111.
2. M.Pigini, *et al.*: 'Imidazoline receptors: qualitative structure-activity relationships and discovery of trazoline and benazoline. Two ligands with high affinity and unprecedented selectivity'. *Bioorganic and medicinal chemistry* 1997, 5, 833-841.
3. Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a toll in 3D-QSAR. Method and Applications. *J. Comput.-Aided Mol. Des.* 1994, 8, 83-96. Institut of medicinal Chemistry, University of Lausanne, BEP CH-1015 Lausanne, Switzerland.
4. A.Carrieri, *et al.*: '2D and 3-D Modeling of Imidazolin Receptor Ligands: Insights into Pharmacophore'. *Bioorganic and Medicinal Chemistry* 1997, 5, 843-856.

**Poster Session III**  
**Prediction of Ligand-**  
**Protein Binding**

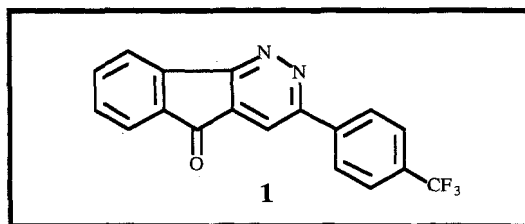
## REVERSIBLE INHIBITION OF MAO-A AND B BY DIAZOHETEROCYCLIC COMPOUNDS: DEVELOPMENT OF QSAR/CoMFA MODELS

Cosimo D. Altomare,<sup>1</sup> Antonio Careri,<sup>1</sup> Saverio Cellamare,<sup>1</sup> Luciana Summo,<sup>1</sup> Angelo Carotti,<sup>1</sup> Pierre-Alain Carrupt,<sup>2</sup> and Bernard Testa<sup>2</sup>

<sup>1</sup>Dipartimento Farmaco-Chimico, Università di Bari  
via E. Orabona 4, I-70125 Bari

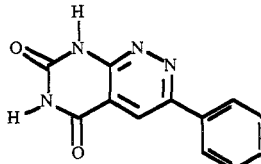
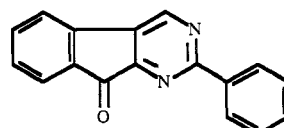
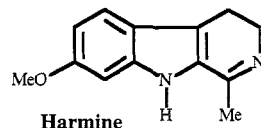
<sup>2</sup>Institut de Chimie Therapeutique, Université de Lausanne  
BEP-Dorigny, CH-1015 Lausanne

Monoamine oxidase (MAO, EC 1.4.3.4) is a FAD-containing enzyme of the outer mitochondrial membrane that catalyzes the oxidative deamination of various neurotransmitters and dietary amines. MAO exists in two forms (A and B), which differ by their amino acid sequence, substrate specificity, and sensitivity to inhibitors. MAO-A and MAO-B inhibitors are useful as antidepressant and adjuvants in the treatment of Parkinson's disease, respectively.<sup>1</sup> In previous studies,<sup>2</sup> we described 5*H*-indeno[1,2-*c*]pyridazin-5-ones as reversible and competitive MAO-B inhibitors. A predictive 3D-QSAR model led to the design of compound **1** with nanomolar inhibition value ( $IC_{50} = 90$  nM). To deepen our understanding of MAO-A/B inhibition and selectivity, we synthesized and tested novel condensed pyridazines, pyrimidines and 1,2,4-triazines showing appreciable MAO-A inhibition.



Various QSAR and CoMFA studies were performed using a set ( $n = 22$ ) of diverse molecules, and the results revealed the physicochemical interactions mainly involved in enzyme-inhibitor complexation. The influence of lipophilicity in increasing inhibition of MAO-B (and not MAO-A) was demonstrated by QSAR Hansch-type analysis and CoMFA including the molecular lipophilicity potential (MLP). As for CoMFA results, lipophilic field alone led to the best one-field model ( $q^2 = 0.585$ ), whereas the best 3D-QSAR model was obtained by combining lipophilic and electrostatic fields ( $q^2 = 0.653$ , O.N.C. = 4,  $r^2 = 0.969$ ,  $s = 0.144$ ). These results agree with and complement a recently published model of reversible MAO-B inhibition.<sup>3</sup> In contrast, both QSAR and CoMFA did not yield reliable models for MAO-A reversible inhibition, where complexation between inhibitor and FAD appears as a critical event. This implies electrostatic interactions and charge transfer bonding as two major contributions to the complex stability.<sup>4</sup> Thus, the molecular electrostatic

potentials (MEPs) of our MAO-A inhibitors were compared with the MEP of *Harmine*, a potent reversible and selective inhibitor ( $pIC_{50} = 7.12$ ), by using the MEPSIM package<sup>5</sup> as computational tool. The electron density distribution included in the MEP calculations was obtained from *ab initio* wave functions (basis set STO-3G), whereas electrostatic similarity was assessed by calculating the Spearman rank correlation coefficient between the MEP values of each pair of molecules (*Harmine* as the template) computed at grid common points. Preliminary results interestingly showed a relation between MAO-A inhibition and MEPSIM index.

			
MEPSIM index	0.57	0.70	1.00
$pIC_{50}$ (A)	4.79	5.63	7.12

Finally, the importance of  $\pi$ - $\pi$  stacking interactions in the modulation of MAO-A inhibition was assessed by measuring retention on a chromatographic stationary phase carrying dinitrobenzoyl group as a  $\pi$ -acceptor system. The above results, even if at a preliminary level, provided information which could aid the design of selective diazine MAO inhibitors.

## Acknowledgments

The authors are indebted to Prof. Ferran Sanz (Institut Municipal d'Investigació Mèdica, IMIM, Barcellona) for his interest, helpful suggestions and support in MEPSIM calculations.

## REFERENCES

1. J. Wouters, Structural aspects of monoamine oxidase and its reversible inhibition, *Current Medicinal Chemistry*, 5:137 (1998)
2. (a) S. Kneubühler, U. Thull, C. Altomare, V. Carta, P. Gaillard, P.-A. Carrupt, A. Carotti, B. Testa, Inhibition of monoamine oxidase-B by 5*H*-indeno[1,2-*c*]pyridazines: biological activities, quantitative structure-activity relationships (QSARs) and 3D-QSARs, *J. Med. Chem.*, 38:3874 (1995). (b) C. Altomare, S. Cellamare, L. Summo, M. Catto, A. Carotti, U. Thull, P.-A. Carrupt, B. Testa, H. Stoeckli-Evans, Inhibition of monoamine oxidase-B by condensed pyridazines and pyrimidines: effects of lipophilicity and structure-activity relationships, *J. Med. Chem.*, in press (1998).
3. J. Wouters, F. Ooms, S. Jegham, J.J. Koenig, P. George, F. Durant, Reversible inhibition of type B monoamine oxidase. Theoretical study of model diazo heterocyclic compounds, *Eur. J. Med. Chem.*, 32:721 (1997).
4. F. Moureau, J. Wouters, M. Depas, D.P. Vercauteren, F. Durant, F. Ducrey, J.J. Koenig, F.X. Jarreau, A reversible monoamine oxidase inhibitor, Toloxatone: comparison of its physicochemical properties with those of other inhibitors including Brofaromine, Harmine, Ro-40519 and Moclobemide, *Eur. J. Med. Chem.*, 30:823 (1995).
5. F. Sanz, F. Manaut, J. Rodriguez, E. Lozoya, E. López-de-Brinas, MEPSIM: a computational package for analysis and comparison of molecular electrostatic potentials, *J. Comput.-Aided Mol. Design*, 7:337 (1993).

## MODELLING OF THE 5-HT<sub>2A</sub> RECEPTOR AND ITS LIGAND COMPLEXES

Estrella Lozoya,<sup>1</sup> Maria Isabel Loza<sup>2</sup> and Ferran Sanz<sup>1,\*</sup>

<sup>1</sup>Research Group on Medical Informatics, Institut Municipal d'Investigació Mèdica (UAB), C/ Dr. Aiguader 80, E-08003 Barcelona (Spain)

<sup>2</sup>Department of Pharmacology, Universidade de Santiago de Compostela, E-15706 Santiago de Compostela (Spain)

### INTRODUCTION

Up to now, modelling of the GPCR is one of the most interesting but most difficult challenges in protein modelling. The difficulties arise from the lack of crystallographic data to be used in a standard homology approach. The first GPCR models that were published were based in the crystallographic data of bacteriorhodopsin, which is an inappropriate template because it is not a GPCR and it shows a very low homology degree with the GPCRs. More recently, a low resolution electron density map of rodhopsin<sup>1</sup> is being used for the packing of the GPCR transmembrane helices (TMH). This history reflects the constant need of ameliorating the existing models by taking into account new experimental data or improved theoretical or computational tools.

### 5-HT<sub>2A</sub> RECEPTOR MODEL BUILDING

The receptor 3D model (Figure 1) was built considering the theoretical and experimental knowledge in 1997. The seven TMHs were packed taking into account: 1) the Baldwin proposal for GPCRs,<sup>2</sup> 2) molecular biology experiments like this showing the proximity of an Asp of Hx II with an Asn of Hx VII<sup>3</sup> (see Figure 1), and, 3) the lipophilicity profile of the helices, in such a way that the major part of the lipophilic residues are exposed to the fosfolipids. The model was geometrically refined by means of MM computations using the Amber force field. The model passed the PROCHECK and WHATCHECK quality tests, and it was stable to MD

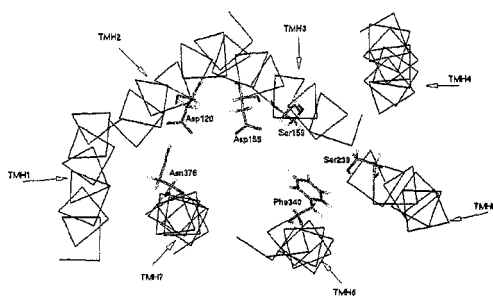


Figure 1. Binding site of the proposed 5-HT<sub>2A</sub> model.

\* To whom correspondence has to be addressed

simulations. Furthermore, the model acceptably fits an improved version of the rhodopsin map recently published.<sup>4</sup>

## 5-HT<sub>2A</sub> RECEPTOR MODEL DOCKING SIMULATIONS

Docking of several 5-HT<sub>2A</sub> ligands (5-HT,  $\alpha$ -Me-5-HT, DOI and ketanserin) was automatically explored with the Affinity module of BIOSYM, taking into account the conformational flexibility of both receptor and ligands. Feasible complexes for all the considered ligands into the receptor binding site were found (Figures 2). These complexes show hydrogen bonds with residues that have been experimentally described as critical for ligand binding<sup>5-6</sup>. Interesting aromatic interactions also appear. Furthermore, the complexes exhibited stability during MD simulations.

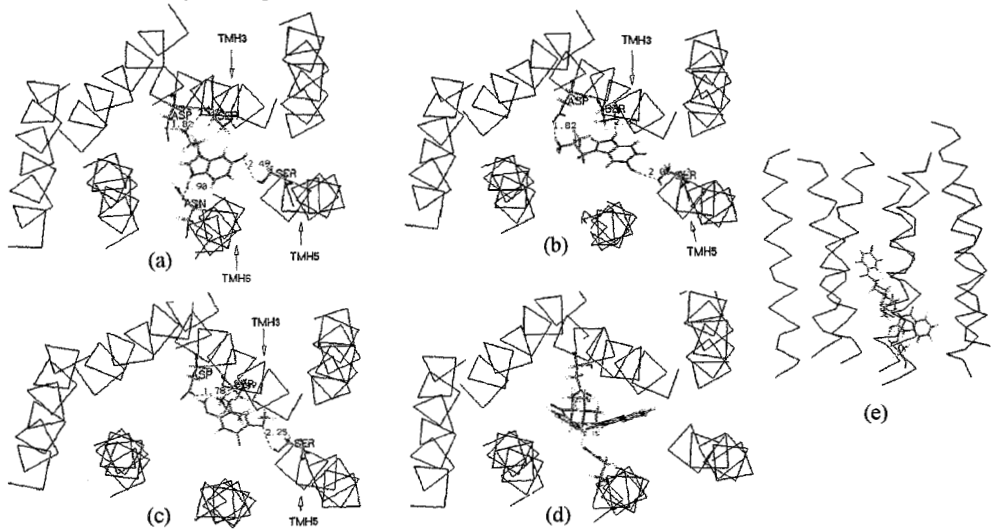


Figure 2. Models of the docking complexes of 5-HT (a),  $\alpha$ -Me-5-HT (b), DOI (c), and ketanserin (d-e).

## Acknowledgements

This research was supported by the CICYT (SAF 94-0593-C04), CIRIT (BQF93-94) and CESCA grants.

## REFERENCES

1. G.F.X. Schertler, C. Villa and R. Henderson. Projection structure of rhodopsin. *Nature* 362:770-772 (1993).
2. J.M. Baldwin. The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* 12:1693-1703 (1993).
3. W. Zhou, C. Flanagan, J.A. Ballesteros, K. Konvicka, J.S. Davidson, H. Weinstein, R.P. Millar and S.C. Sealfon. A reciprocal mutation supports helix 2 and 7 proximity in the gonatropin-releasing hormone receptor. *Mol. Pharmacol.* 45:165 (1989).
4. J.M. Baldwin, G.F.X. Schertler and M.V. Unger. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein coupled receptors. *J. Mol. Biol.* 272:144-164 (1997).
5. C.M. Fraser, C.D. Wang, D.A. Robinson, J.D. Gocayne and J.C. Venter. Site-directed mutagenesis of m<sub>1</sub> muscarinic receptors: conserved aspartic acids play important roles in receptor function. *Mol. Pharmacol.* 36:840-847 (1989).
6. C.D. Strader, I.S. Sigal and R.A.F. Dixon. Structural basis of  $\beta$  adrenergic receptor function. *FASEB J.* 3:1825-1832 (1989).

## TOWARDS THE UNDERSTANDING OF SPECIES SELECTIVITY AND RESISTANCE OF ANTIMALARIAL DHFR INHIBITORS

Thomas Lemcke,<sup>1</sup> Inge Thøger Christensen,<sup>2</sup> and Flemming Steen Jørgensen<sup>2</sup>

<sup>1</sup>Institute of Pharmacy, University of Hamburg, D-20146 Hamburg, Germany

<sup>2</sup>Department of Medicinal Chemistry, Royal Danish School of Pharmacy, DK-2100 Copenhagen, Denmark

### INTRODUCTION

Malaria tropica is caused by *Plasmodium falciparum* and it is most often lethal to the untreated patient. One of the targets of malaria therapy is the dihydrofolate reductase (DHFR) of *P. falciparum*. Several DHFR-inhibitors (e.g. methotrexate, trimethoprim, pyrimethamine), which inhibit the DHFR of different species through selective binding to the enzyme, are known. Pyrimethamine is a selective inhibitor of the plasmodial DHFR, but due to rapid development of resistance against this drug, its use is limited.

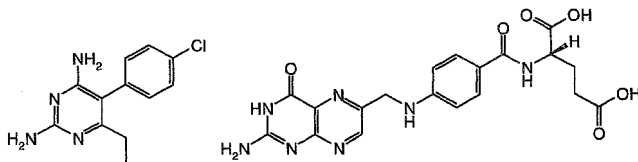


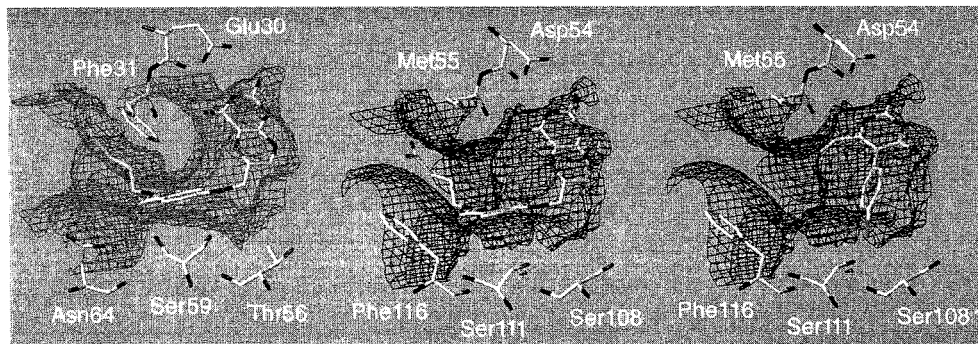
Figure 1. Pyrimethamine and folic acid

### RESULTS AND DISCUSSION

DHFR from vertebrates, bacteria and fungi have a very high structural homology. Therefore, a three-dimensional model of *P. falciparum* DHFR was constructed by homology building using a step-wise procedure (Lemcke et al., 1998). The model was based on a structural alignment of X-ray structures of DHFR from different species (human, chicken, *E. coli*, *L. casei* and *P. carinii*). By superimposing these structures, the structurally conserved regions were identified. The sequence of the pyrimethamine sensitive *P. falciparum* clone 3D7 was aligned to the structurally aligned sequences of the five X-ray structures. The final model was geometry optimized using the AMBER force field and evaluated using the programs PROCHECK and PROSA.

Folic acid could be docked into the active site of the model in the same conformation as in the human enzyme. The pteridine ring is forming a bidentate hydrogen bond to the carboxylate sidechain of Asp54 and Glu30, respectively (Figure 2). The most pronounced difference is the replacement of Asn64 in the human with Phe116 in the plasmodial DHFR. It prevents the donation of a hydrogen bond to the carbonyl oxygen of the benzoic acid moiety of folic acid. However, this hydrogen bond is not essential for substrate binding, as it is not conserved among different species (e.g. yeast, fungi and bacteria).

Pyrimethamine was docked into the active site in a way similar to the binding mode reported for other diamino-pyrimidine inhibitors (Blakley 1995). The phenyl ring is located in the region of the active site, that displays noticeable differences between the human and the plasmodial structure (Figure 2). Thus, according to our model, the chlorine atom in pyrimethamine is in vdW contact with Ser108 and Ser111.



**Figure 2.** Active site of human (left) and model (middle) with folic acid and model with pyrimethamine (right)

Ser108, which is a threonine in most other structures, is reported to be related to building up of drug resistance against pyrimethamine (Sirawaraporn et al., 1997, Peterson et al., 1988). The S108N point mutant of plasmodial DHFR has a considerably reduced sensitivity to pyrimethamine. Consequently, in the mutant there might not be enough space for the inhibitor to bind to the active site. This observation is in consistence with the reported resistance of the S108N mutant.

## CONCLUSIONS

A three-dimensional model of the DHFR domain from *P. falciparum* has been obtained by homology building. The model was based on the X-ray structure of the human DHFR and a structural alignment of five DHFR structures. Based on the model we were able to explain the significance of the S108N point mutation in relation to pyrimethamine resistance.

## REFERENCES

- Blakley, R.L., 1995, *Adv. Enzymol. Relat. Areas Mol. Biol.* 70:23.  
 Lemcke, T., Christensen, I.T., and Jørgensen, F.S., 1998, *Bioorg. & Med. Chem.*, submitted.  
 Peterson, D.S., Walliker, D., Welles, T.E., 1988, *Proc. Natl. Acad. Sci. USA.*, 85: 9114.  
 Sirawaraporn, W., Sathitkul, T., Sirawaraporn, R., Yathavong, Y., Santi, D. V., 1997, *Proc. Natl. Acad. Sci. USA*, 94:1124.



## MODELING OF SURAMIN-TNF $\alpha$ INTERACTIONS

C. Marani Toro<sup>1</sup>, M. Mabilia<sup>2</sup>, F. Mancini<sup>1</sup>, M. Giannangeli<sup>1</sup>, C. Milanese<sup>1</sup>

<sup>1</sup> Angelini Ricerche, P.le Stazione, I-00040 S. Palomba, Rome, Italy

<sup>2</sup> S.IN - Soluzioni Informatiche, Via Salvemini 9, I-36100 Vicenza, Italy

### INTRODUCTION

Suramin, a symmetrical polysulfonated urea derivative (1-2), promotes the dissociation of trimeric Tumor Necrosis Factor  $\alpha$  (TNF $\alpha$ ) into inactive subunits, thus inhibiting the binding of TNF $\alpha$  with its cellular receptor (3).

The purpose of the present study is to investigate location and nature of likely suramin binding site(s) on TNF $\alpha$  by means of computer-aided molecular modeling techniques.

### RESULTS AND DISCUSSION

In order to determine the rotational energy barriers for amidic and ureidic C-N bonds, MonteCarlo/Energy Minimization (MC/EM) (4) searches were carried out, under different conditions, on suramin fragments. Using the conformational preferences suggested by the above results, the whole suramin molecule underwent MC/EM procedure. Being suramin a polyanion and TNF $\alpha$  surface characterized by a high number of positively charged residues, interaction energies between suramin simplified models (1-naphthalene monosulfonic acid, 1,3-naphthalenedisulfonic acid, 1,3,6-naphthalenetrisulfonic acid) and protonated Arg and Lys were first evaluated to establish preferred interaction geometries and corresponding energy contributions.

Subsequently, two different docking modes were examined using MC/EM procedure and the Amber force field (5-6) (Kollman's united atom). First, charged suramin was docked onto the TNF $\alpha$  trimer surface in such a way that sulfonic groups could reach a putative binding region characterized by positively charged residues. Alternatively, suramin was docked inside TNF $\alpha$  trimer, along the three-fold axis, so that one of the aromatic rings of naphthalenetrisulfonic acid could reach the TNF $\alpha$  core region defined by tyrosine

residues 59, 119 and 151. The outcome of these simulations indicates that electrostatic interactions between sulfonated groups and charged residues seem essential for recognition, alignment and initial interaction of suramin, while a relatively long "linear" structure, such as suramin, might then be required to allow penetration in the channel centered on the TNF $\alpha$  trimer symmetry axis.

Other polysulfonated compounds, structurally related to suramin such as trypan blue and Evans blue, were docked inside the trimer according to suramin orientation. These docking studies reveal that suramin, trypan and Evans may interact in a similar manner to TNF $\alpha$ .

A specific immunoenzymatic assay (3) was developed on suramin and related compounds to confirm their capacity to inhibit TNF $\alpha$  /TNF-receptor binding. The results indicate that Evans blue and trypan blue have an activity comparable to suramin, in agreement with our theoretical models.

To determine the minimum size of the pharmacophore for TNF $\alpha$ , also 1,3,6 naphthalenetrisulfonic acid was tested. The fact that the naphthalenetrisulfonic acid does not affect TNF $\alpha$  binding to its receptor seems to indicate that electrostatic interactions alone are not sufficient to induce the trimer dissociation, thus suggesting that other kinds of interactions (e.g. dispersion forces) and molecular size/length might play an important role in this phenomenon.

## ACKNOWLEDGMENTS

The work has been carried out under a research contract with C.Au.T., Pomezia, Italy, within the "Programma Nazionale Farmaci - Seconda Fase" of M.U.R.S.T.

## REFERENCES

- 1 Cheson, B. D. et al. *JAMA, J. Am. Med. Assoc.* **258**, 1347 (1987).
- 2 Stein, C. A et al *J Clin. Oncol.* **7**, 499 (1989).
- 3 Alzani, R. et al *Biochem.* **34**, 6344 (1995).
- 4 *J. Comp. Chem.* **7**, 230 (1986).
- 5 Goodmann, J. et al. *J. Comput. Chem.* **12**, 1110 (1991).
- 6 Halgren's T.A et al *J. Comput. Chem.* **17**, 490 (1996).
- 7 Nakajima, M et al. *J Biol. Chem.* **266**, 9661 (1990).

# DE NOVO DESIGN OF INHIBITORS OF PROTEIN TYROSINE KINASE pp60<sup>c-src</sup>

Thierry Langer<sup>1</sup>, Matthias A. König<sup>1</sup>, Georg Schischkow<sup>1</sup>, Salvatore Guccione<sup>2</sup>

<sup>1</sup> Institute of Pharmaceutical Chemistry, University of Innsbruck  
Innrain 52a, A-6020 Innsbruck, Austria

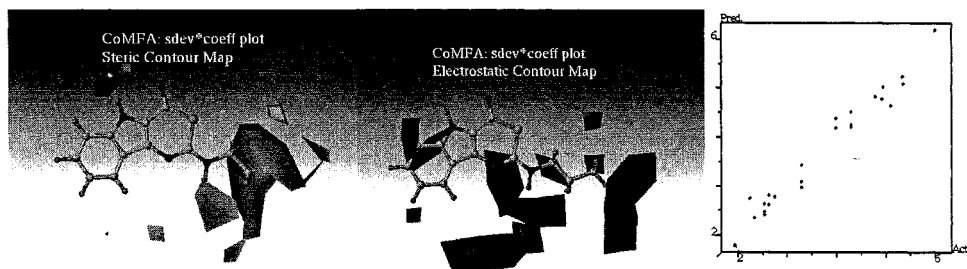
<sup>2</sup> Dipartimento di Scienze Farmaceutiche, Università di Catania,  
viale Andrea Doria 6, I-95125 Catania, Italy

## INTRODUCTION

Protein tyrosine kinase (PTK) pp60<sup>c-src</sup> is a new and promising target for the modulation of cell-proliferation.<sup>1</sup> In order to find new specific inhibitors for this enzyme we performed a two step computer aided ligand design study. First, a 3D QSAR model based on a training set of 25 known ligands was established using the CoMFA approach.<sup>2</sup> Second, a *de novo* approach using the x-ray coordinates<sup>3</sup> of human PTK pp60<sup>c-src</sup> (EC 2.7.1.112) was applied using the LUDI software tool.<sup>4</sup>

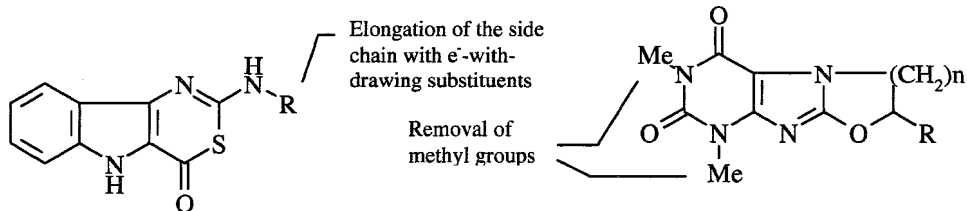
## METHODS AND RESULTS

**CoMFA.** A training set containing the structures of 25 ATP competitive inhibitors of PTK pp60<sup>c-src</sup> covering an activity range of 0.1 - 1000  $\mu$ M was selected. The alignment was generated by using a multistep docking and energy minimization procedure, the x-ray coordinates of the protein structure together with ATP were taken as a template. The model obtained was shown to be predictive as indicated by a  $r_{cv}^2$  of 0.72 ( $s_{press} = 0.66$ ). The results of the final model is given in Fig. 1.

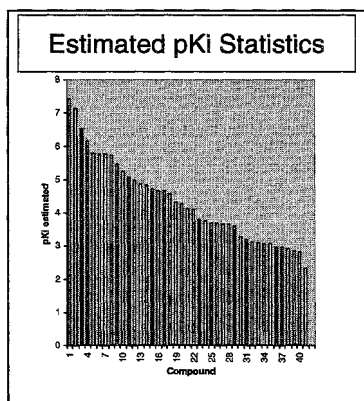


**Figure 1.** Sdev \* coefficient contour plots and predicted against actual  $pIC_{50}$  for inhibitors of PTK pp60<sup>c-src</sup> model as derived by CoMFA

From the interpretation of the CoMFA contour plots, the following conclusions on successful modifications of training set molecules may be drawn:



**LUDI de novo ligand design.** The study was carried out in two consecutive steps: first, molecular structures fitting into the active site were selected from the LUDI standard fragment database. Compounds exhibiting the highest scores (estimated pKi values > 2.5) were selected, and in a second step side chains were added at multiple sites of the starting fragments. This procedure resulted in the determination of approximately 200 compounds. Molecules exhibiting an estimated pKi > 5 were finally docked into the active site in order to study the interaction pattern. An example of compounds designed is given in Table 1.



Compounds Designed	Predicted Affinity (pKi)
	7.4
	7.1
	6.5
	6.2
	5.8
	5.8

## REFERENCES

- 1 König, M.A., Protein-Tyrosin Kinase pp60<sup>c-src</sup>: Molecular modelling studies of newly developed inhibitors, *Diploma Thesis*, University of Innsbruck 1997 and references cited therein.
- 2 Sybyl 6.3, Tripos Ass., St. Louis, MO, 1996.
- 3 Xu, W., Harrison, S.C., Eck, M.J., Crystal Structure Of Human Tyrosine Protein Kinase C-Src, *Nature* 385: 595 (1997).
- 4 Insight95, Molecular Simulations Inc., San Diego, CA, USA (1996)

## ELUCIDATION OF ACTIVE CONFORMATIONS OF DRUGS USING CONFORMER SAMPLING BY MOLECULAR DYNAMICS CALCULATIONS AND MOLECULAR OVERLAY

Shuichi Hirono<sup>1</sup> and Kazuhiko Iwase<sup>2</sup>

<sup>1</sup> *School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan*

<sup>2</sup> *Central Research Laboratories, Kyorin Pharmaceutical Co., Ltd., 2399-1 Mitarai, Nogi-machi, Shimotsuga-gun, Tochigi 329-0114, Japan*

In computer-assisted drug design, it is very important to determine the conformation of the ligand molecules that bind to such proteins as receptors and enzymes, that is, the active conformation. For compounds binding to the same receptor or enzyme, same atomic groups in the compound occupy almost same three-dimensional spaces in the receptor or enzyme. Hence a lenient superposition of atomic groups between two molecules seems to be effective for the extraction of an active conformation. To estimate the active conformations of drugs, we developed a new procedure for superposing two molecules based on the physicochemical properties of the atomic groups in a molecule. The four types of physicochemical properties of the atomic groups within individual molecules -- hydrophobicity, presence of a hydrogen-bonding donor, presence of a hydrogen-bonding acceptor and presence of a hydrogen-bonding donor/acceptor -- were supposed and a score was given to every overlap. Each atomic group belonging to the types of physicochemical properties was approximated by a sphere with an appropriate radius and if any two spheres overlapped by even a little, they were treated as a target of a score.

In order to systematically perform the superposition of two molecules, first, the center of mass of each molecule is translated to the origin of coordinates, and then the circumscribed rectangular boxes are calculated. The molecule with a large box-volume is fixed, and then the center of mass for the molecule with a small box-volume is translated and rotated. The range of translation is the maximum distance that the small box can translate inside the large box. The translational increment is 1 Å and the center of mass is translated on the body-centered cubic lattice points made in the circumscribed

rectangular box of the large volume. The rotation is performed on each of the lattice points. The ranges of three Eulerian angles are  $0 \leq \varphi, \psi \leq 350^\circ$  and  $0 \leq \theta \leq 180^\circ$  and the rotational increment is  $10^\circ$ . When the atomic groups with the same physicochemical properties were overlapped, points were added to the score, while, if the atomic groups with different physicochemical properties were overlapped, points were subtracted. The score is calculated on the orientations of all of superpositions, respectively, and the orientation with the highest value is adopted. If the highest value of the score is redundant, one with the smallest value of the root mean square deviation (rmsd) of the distance between atomic groups between every pair of inhibitors is conveniently selected. As a precaution, however, it might be required to check orientations with the same score. For further improving the score of the adopted orientation, three translations and three Eulerian angles are optimized by the simplex method using the rmsd as an objective function to determine the final orientation of superposition.

We carried out the superposition of conformers sampled by the high temperature molecular dynamics (MD) calculation using the CAMDAS (Conformational Analyzer with Molecular Dynamics And Sampling) program<sup>1</sup> with respect to 12 pairs of 20 enzyme inhibitors in order to check the effectiveness of the procedure. The superposition of each pair was compared with the superposition obtained from the X-ray crystallography of an enzyme-inhibitor complex which is derived by removing only the coordinates of the enzyme molecule after a least-squares fitting between the  $\alpha$ -carbon atomic coordinates of the enzyme molecules in the enzyme-inhibitor complexes. The results showed that the best overlay for each inhibitor pair could successfully reproduce the superposition obtained from the X-ray crystallography.

We next examined whether or not our superposing procedure was able to estimate the active conformation among many conformations. First, the high temperature molecular dynamics calculations for the thrombin inhibitors, MQPA, 4-TAPAP and NAPAP, were executed and 60000 conformers were sampled using the CAMDAS program. As a result, 457 conformers in 4-TAPAP, 113 conformers in NAPAP and 202 conformers in MQPA were selected. Superpositions of conformers sampled by the high temperature MD calculations with respect to the three inhibitors were performed, and 13 sets of conformers having the best common overlay to the three inhibitors were selected. The resulting conformer sets contained the superposition of the active conformations derived from the X-ray crystallography of the thrombin-inhibitor complexes. It is suggested that the method in this work<sup>2</sup> is useful for elucidating a pharmacophore and finding the bioactive conformation among a lot of conformations of a drug obtained from computational calculations.

## REFERENCES

1. Tsujishita, H. and Hirono, S., *J. Compt.-Aided Mol. Design*, **11**, 305 (1997).
2. Iwase, K. and Hirono, S., Submitted to *J. Compt.-Aided Mol. Design*.

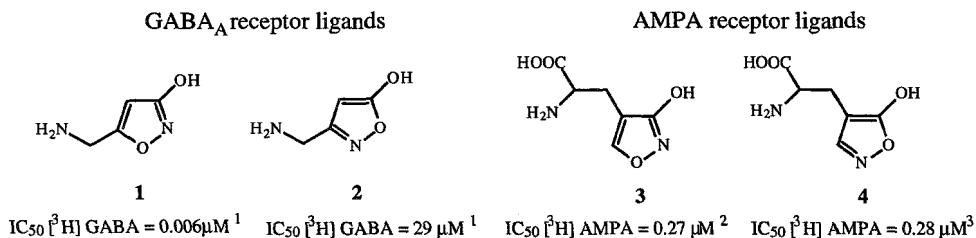
# DIFFERENCES IN AGONIST BINDING PATTERN FOR THE GABA<sub>A</sub> AND THE AMPA RECEPTORS ILLUSTRATED BY HIGH-LEVEL *AB INITIO* CALCULATIONS

Lena Tagmose, Lene Merete Hansen, Per-Ola Norrby and Tommy Liljefors

Department of Medicinal Chemistry, Royal Danish School of Pharmacy  
Universitetsparken 2, DK-2100 Copenhagen, Denmark

## INTRODUCTION

$\gamma$ -Aminobutyric acid (GABA) and (*S*)-glutamic acid are the endogenous receptor ligands for the GABA<sub>A</sub> and the AMPA receptor, respectively. The 3-isoxazolol and the 5-isoxazolol rings have been used as bioisosters for the carboxylic acid group in GABA and the distal carboxylic acid group in (*S*)-glutamic acid leading to a wide range of semi-rigid analogues (Fig. 1).

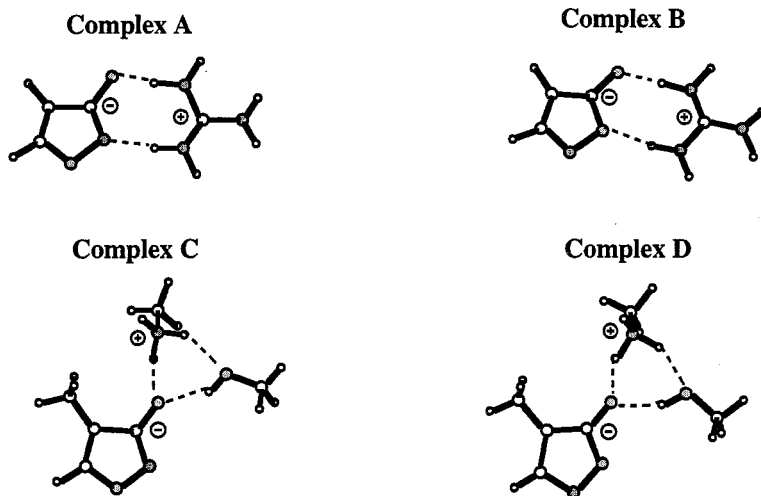


**Fig. 1** GABA<sub>A</sub> and AMPA receptor ligands

In the GABA<sub>A</sub> series a large difference in binding affinity exists for the 3- and 5-isoxazolol compounds, **1** and **2**, whereas the corresponding compounds, **3** and **4**, in the AMPA series exhibit essentially identical binding affinities (see Fig. 1). The different behaviours of the 3- and 5-isoxazolol compounds in the two series have been investigated by high-level *ab initio* calculations.

## RESULTS

For the GABA<sub>A</sub> receptor, bidentate complexes between the 3- and the 5-isoxazolol anions and a guanidinium ion have been studied. For the AMPA receptor, monodentate as well as bidentate complexes between the isoxazolol anions and a methylammonium ion and/or methanol have been studied.



**Fig. 2** Calculated complexes which best rationalize experimental relative affinities

The complexation energies have been calculated according to Eq. 1 and are listed in Table 1.

$$E_{\text{complexation}} = E_{\text{complex}} - (E_{\text{ligand}} + E_{\text{complexing agent(s)}}) - \Delta G_{\text{solv}}(\text{ligand}) \quad (\text{Eq. 1})$$

**Table 1** Calculated complexation and solvation energies

Complex	$\Delta G_{\text{solv}}(\text{ligand})$	$E_{\text{complexation}}$
	AM1/SM2 <sup>4</sup> (kcal/mol)	MP4SDQ/6-31+G**// HF/6-31+G* (kcal/mol)
A	-77.1	-41.0
B	-69.9	-37.1
C	-74.5	-58.6
D	-67.3	-59.3

## CONCLUSION

The calculations indicate that the GABA<sub>A</sub> agonists bind to the receptor forming a bidentate complex. The difference in complexation energy between Complex A and B (3.8 kcal/mol) is in good agreement with the relative binding affinities. In contrast, the binding affinities for the AMPA compounds can be explained by a monodentate binding. The difference in complexation energy between Complex C and D (0.7 kcal/mol) is in good agreement with the relative binding affinities to the AMPA receptor.

## REFERENCES

1. Krosggaard-Larsen, P. and Roldskov-Christiansen, T. *Eur. J. Med. Chem. Chim. Ther.* **1979**, *14*, 157.
2. Sløk, F. A.; Ebert, B.; Lang, Y.; Krosggaard-Larsen, P.; Lenz, S. M. and Madsen, U. *Eur. J. Med. Chem.*, **1997**, *32*, 329.
3. Iwama, I.; Nagai, Y.; Tamura, N.; Harada, S. and Nagaoka, A. *Eur. J. Pharmacol.*, **1991**, *197*, 187.
4. Cramer, C. J. and Truhlar, D. G. *J. Comput.-Aided Mol. Des.*, **1992**, *6*, 629.



# STABILIZATION OF THE AMMONIUM-CARBOXYLATE ION-PAIR BY AN AROMATIC RING

Tommy Liljefors and Per-Ola Norrby  
Department of Medicinal Chemistry  
Royal Danish School of Pharmacy, Copenhagen, Denmark

## INTRODUCTION

A central feature in 7TM models of the binding of monoaminergic neurotransmitters to their receptors is an ammonium/carboxylate ion-pair interaction between the protonated amine and an aspartate residue in helix III.<sup>1</sup>

However, *ab initio* calculations on the amine/carboxylic acid *vs.* the ammonium ion/carboxylate anion complex show that the neutral amine/carboxylic acid complex is more stable *in vacuo* by 11.3 kcal/mol.<sup>2</sup> Thus, the existence of an ion-pair binding requires a significant stabilization of the ion-pair complex. On the basis of the large attractive interactions between an ammonium ion and a benzene ring ( $\Delta H = -19.3$  kcal/mol)<sup>3</sup>, it has been argued that an ion-pair complex may be strongly stabilized by conserved aromatic rings in cationic neurotransmitter receptors.<sup>1</sup> Attractive interactions between aromatic rings and ammonium ions have been shown to be of importance for ligand binding in biological systems and to synthetic receptors.<sup>4</sup> However, it has not previously been studied if such strong attractive interactions also are present in ion-pair complexes with aromatic rings.

In the present study, we have calculated the complexation energy of the ammonium-carboxylate ion-pair/benzene complex **A** and compared it to the corresponding energy for the ammonium/benzene complex **B**.

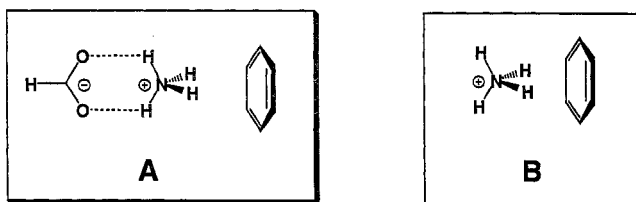


Figure.  $C_{2v}$  symmetric complexes of formate/ammonium/benzene (A) and ammonium/benzene (B).

## RESULTS

The results for the  $C_{2v}$  symmetric complexes **A** and **B** are shown in the Table. Calculations for the corresponding  $C_s$  symmetric complexes have also been performed with very similar results.

Table. Calculated complexation energies for complexes **A** and **B**.

computational level	complex <b>A</b> <sup>a</sup>	complex <b>B</b> <sup>a</sup>
MP2/6-311G**//MP2/6-311G**	-6.0	-19.1
MP4SDQ/6-311G**//MP2/6-311G**	-5.2	-17.3
MP2/6-311+G(2d,2p)//MP2/6-311G**	-6.7	-19.8
“MP4SDQ/6-311+G(2d,2p)”//MP2/6-311G** <sup>b</sup>	-5.9	-18.0

<sup>a</sup> Energies are in kcal/mol. <sup>b</sup> Correlation effects at the MP4SDQ/6-311+G(2d,2p) level were estimated by adding the calculated differences (MP4SDQ/6-311+G\*\* - MP2/6-311+G\*\*) to the calculated MP2/6-311+G(2d,2p) energies.

The calculated results show that the large attractive interaction observed for the ammonium ion - benzene complex is drastically reduced by the presence of a carboxylate anion. The complexation energy for the ion-pair/benzene complex **A** is calculated to be less than one third of the corresponding energy for the ammonium/benzene complex **B**. Furthermore, the complexation energy decreases rapidly with N<sup>+</sup> - aromatic ring distance. Typical closest N<sup>+</sup> - aromatic ring distances in 7TM receptor models are 4.5 - 6.5 Å.<sup>5</sup> At these distances, the stabilization energy of the ion-pair by the benzene ring is less than 2.5 kcal/mol.

## CONCLUSIONS

The complexation energy between the ammonium/formate ion-pair and benzene is only one third of the complexation energy of the ammonium ion - benzene complex.

The attractive interactions between an ammonium/carboxylate complex and an aromatic residue are not sufficiently strong to shift the equilibrium from predominance of the neutral amine/carboxylic acid complex to predominance of the ion-pair complex. Other sources of stabilization, *e.g.* hydrogen bonding to the carboxylate group are required to favor the ion-pair.<sup>6</sup>

## REFERENCES

1. Trumpp-Kallmeyer S., Hoflack J., Bruinvels A., and Hibert M., *J. Med. Chem.* **35** (1992) 3448-62.
2. Heidrich H., van Eikema Hommes N. J. R. and von R. Schleyer P., *J. Comput. Chem.* **14** (1993) 1149-63.
3. Deakyne C. A. and Meot-Ner (Mautner) M. *J. Am. Chem. Soc.* **107** (1985) 474-9.
4. Dougherty, D. A., *Science*, **271** (1996) 163-168.
5. <http://swift.embl-heidelberg.de/7tm/>
6. Liljefors T. and Norrby P.-O., *J. Am. Chem. Soc.* **119** (1997) 1052-5.

# STRUCTURAL REQUIREMENTS FOR BINDING TO CANNABINOID RECEPTORS

Maria Fichera<sup>a</sup>, Alfredo Bianchi<sup>a</sup>, Gabriele Cruciani<sup>b</sup> and Giuseppe Musumarra<sup>c</sup>

<sup>a</sup> Istituto di Farmacologia, Università di Catania, Italy

<sup>b</sup> Dipartimento di Chimica, Università di Perugia, Italy

<sup>c</sup> Dipartimento di Scienze Chimiche, Università di Catania, Italy

## INTRODUCTION

Scientific interest in cannabinoids increased after the isolation of  $\Delta^9$ -tetrahydrocannabinol ( $\Delta^9$ -THC) and prompted a systematic re-evaluation of their use as therapeutical agents. The pharmacological activity of cannabinoids is mediated by two recently identified cannabinoid receptors: the CB<sub>1</sub> receptor localized in specific brain areas and the peripheral receptor CB<sub>2</sub>. In 1992 anandamide, identified as an endogenous ligand for cannabinoid receptors, was shown to share with THC most pharmacological properties in both CNS and peripheral systems<sup>1</sup>.

The striking analogies in the pharmacological activity of structurally different classical and non classical cannabinoids have not yet been rationalized. Available 3D-QSAR studies<sup>2</sup> consider only predominantly rigid compounds but not include anandamide and other derivatives characterized by great rotational freedom.

Aim of this work was to study by 3D-QSAR a set of structurally different molecules in order to obtain general structural information about the CB<sub>1</sub> receptor from the drug-receptor dissociation constants, which are known to be correlated to the potency.

The modelled molecules were selected from literature data<sup>3</sup> reporting the dissociation constants with respect to CB<sub>1</sub> and CB<sub>2</sub> receptors for three series of structurally different compounds: i) THC and derivatives, ii) anandamide and derivatives, iii) indole and derivatives. The dissociation constants for the selected set of 19 molecules, exhibiting a wide variation of both structure and activity, were all determined on the same cell line.

## RESULTS AND DISCUSSION

The structures of all molecules were generated using Sybyl 6.4 molecular modelling package and energy minimized using Tripos force field. The structure of  $\Delta^9$ -THC was chosen as the alignment reference and conformational searches were performed in molecules with

rotational freedom. THC analogues were aligned to  $\Delta^9$ -THC by superimposing common groups, while for indole derivatives different alignments were considered. For anandamide and its analogues, the alignment was operated as proposed by Thomas et al.<sup>4</sup>

The program GRID<sup>5</sup> was used to describe the previously superposed molecular structures. GRID is a computational procedure for detecting energetically favorable binding sites by calculating the interaction energy between small chemical groups (probes) and the target molecule as the sum of Lennard-Jones, electrostatic and hydrogen bond interactions.

A CB<sub>1</sub> pseudoreceptor model<sup>6</sup> proposes that aspartic acid and histidine are involved in the interactions with cannabinoids. Therefore the multi-atom carboxy anion (COO<sup>-</sup>) was chosen as aspartic acid probe and the sp<sup>3</sup> amine NH cation (N1<sup>+</sup>) as histidine probe.

The GRID matrix for the COO<sup>-</sup> probe was correlated with the CB<sub>1</sub> dissociation constants by a PLS model. Removal of the noisy variables from the data set is needed in order to obtain a more stable model and better predictions. From the original 16147 grid variables, a set of 3903 was selected according to the advanced pretreatment in the GOLPE procedure and reduced to 1590 by a further selection based on the recently reported smart region definition (SRD/GOLPE)<sup>7</sup>. The results of the PLS model are reported below.

PLS comp.	SDEC	r <sup>2</sup>	SDEP	q <sup>2</sup>
1	0.2895	0.9340	0.5558	0.7567
2	0.1604	0.9797	0.4586	0.8343
3	0.1009	0.9920	0.4603	0.8331

PLS models with different probes exhibit similar results, showing the validity of the alignment and the stability of the model. The first PCA score parallels the compound activity pointing out the relevance of the selected GRID variables. Accordingly, the first PLS component provides an excellent correlation with the Y values and satisfactory PLS predictions.

The GRID plot of the partial weights (not reported here for the lack of space) allows to identify the regions in the space that contribute most to explain the CB<sub>1</sub> binding constants highlighting areas where a hydrophilic group can increase the dissociation constant and those where hydrophilic interactions decrease it. The above results indicate the pharmacophore structural requirements for binding to cannabinoid receptors for the considered different series of compounds and envisage the design of molecules with higher predicted activities.

## REFERENCES

1. V. Di Marzo, L. De Petrocellis, T. Bisogno and S. Maurelli, *J. Drug Dev. Clin. Pract.*, 7: 199 (1995).
2. B. F. Thomas, D. R. Compton, B. R. Martin and S. F. Semus, *Mol. Pharmacol.*, 40:656 (1991).
3. V. M. Showalter, D. R. Compton, B. R. Martin and M. E. Abood, *J. Pharm. Exper. Ther.*, 278:989 (1996).
4. B.F. Thomas, I. B. Adams, S. W. Mascarella, B. R. Martin and R. K. Razdan, *J. Med. Chem.*, 39:471 (1996).
5. P. J. Goodford, *J. Med. Chem.*, 28:849 (1985).
6. S. Schmetzer, P. Greenidge, K. A. Kovar, M. S. Alexandru and G. Folkers, *J. Comp. Aided Mol. Des.*, 11:278 (1997).
7. M. Pastor, G. Cruciani, S. Clementi, *J. Med. Chem.*, 40:1455 (1997) and references therein.

## DESIGN, SYNTHESIS AND TESTING OF NOVEL INHIBITORS OF CELL ADHESION

David T. Manallack,<sup>1</sup> John G. Montana,<sup>1</sup> Paul V. Murphy,<sup>2</sup>  
Rod E. Hubbard<sup>3</sup> and Richard J. K. Taylor<sup>3</sup>

<sup>1</sup> Chiroscience R&D Ltd.

Cambridge Science Park, Milton Rd, Cambridge CB4 4WE, UK

<sup>2</sup> University College Dublin

Chemistry Department, Belfield, Dublin 4, Rep. of Ireland

<sup>3</sup> University of York

Department of Chemistry, Heslington, York YO1 5DD, UK

### INTRODUCTION

The Selectin family of proteins comprises three carbohydrate binding proteins (E, P and L) involved in cell adhesion events<sup>1</sup>. In response to inflammatory stimuli, these proteins play a crucial role in the recognition of sialyl Lewis X (sLe<sup>x</sup>) and related carbohydrates present on the surface of neutrophils. Following recognition and binding the white blood cells are free to migrate to the sites of injury and infection<sup>2</sup>. In pathogenic states this sequence of events can lead to pain and inflammation. Blocking the binding of sLe<sup>x</sup> could potentially be of benefit in the treatment of inflammatory and autoimmune disorders such as, rheumatoid arthritis, asthma, psoriasis, IBD etc.

A number of SAR studies have demonstrated that the key functional groups for the binding of sLe<sup>x</sup> to E-selectin are the hydroxyl groups of the fucose unit and the carboxylic acid<sup>3</sup> of the sialic acid unit. Clinical trials have been undertaken using some of these sLe<sup>x</sup> analogues, however, the potency of these compounds is poor and they are largely carbohydrate in nature. There is a need, therefore, to develop compounds which are more potent and less susceptible to carbohydrate metabolism.

### PROCEDURE AND RESULTS

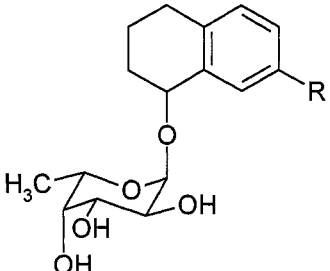
The focus of this study was to design and synthesise small molecule mimics of sLe<sup>x</sup> which were reduced in carbohydrate character, to act as inhibitors of cell adhesion.

To assist the design process, the crystal structure of E-selectin was used (pdbcode 1ESL) along with rat mannose-binding protein (pdbcode 2MSB) which has a carbohydrate molecule bound to a key calcium ion. The similarity of the calcium sites of both proteins enabled a model of fucose bound to E-selectin to be constructed. Information regarding the NMR conformation of sLe<sup>x</sup> bound to E-selectin<sup>4</sup> and suggestions from molecular

biology and molecular modelling studies that the carboxylic acid group on sLe<sup>x</sup> binds to Arg 97<sup>s</sup> were combined to create a model of sLe<sup>x</sup> bound to E-selectin. The model was subsequently minimised using QUANTA (MSI Inc.) (c.f. model of Kogan et al.<sup>5</sup>) and was used to design sLe<sup>x</sup> mimics incorporating a semi-rigid tetralin scaffold (Table 1) to hold the fucose and acid groups in the correct spatial orientations to bind to E-selectin.

Synthetic feasibility was an essential design criteria and the compounds listed in Table 1 were made using palladium catalysed coupling reactions for key synthetic steps<sup>6</sup>.

**Table 1.** Effect of compounds 1 to 6 on the adhesion of resting HL-60 cells to recombinant soluble E-selectin.



No.	R	IC <sub>50</sub> mM
1	CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> COOH	1.7
2	CONHCH(CH <sub>3</sub> )COOH (R)	> 5
3	CONHCH(CH <sub>3</sub> )COOH (S)	> 5
4	CH <sub>2</sub> CH <sub>2</sub> C(CH <sub>3</sub> ) <sub>2</sub> COOH	3.7
5	C≡CC(CH <sub>3</sub> ) <sub>2</sub> COOH	1.7
6	COOH	4.0

Molecular modelling showed that the fucose and acid groups of compound 1 overlapped well with sLe<sup>x</sup> and this compound demonstrated good activity in the cell adhesion assay (Table 1). Indeed, the activity was comparable to sLe<sup>x</sup> (IC<sub>50</sub> 2.6 mM). Compound 5 also showed the same level of potency, however, compounds 2 and 3 were inactive (Table 1). Follow up modelling studies will be required to explain the SAR observed. Interestingly, compound 6 showed moderate activity despite the shorter length between fucose and acid groups. Molecular modelling suggested that this molecule may also bind to Arg 97 by approaching the guanidine group from an alternative perspective.

## CONCLUSIONS

This study has demonstrated the successful application of molecular modelling to the rational design of sLe<sup>x</sup> mimics as inhibitors of E-selectin. The biological activity of a number of these compounds was comparable to the natural ligand, sLe<sup>x</sup>. In addition, this study has also demonstrated a harmonious association between medicinal chemists and molecular modellers as well as providing a platform for future drug development.

## REFERENCES

1. T.A. Springer, Adhesion receptors of the immune system, *Nature* 346:425 (1990).
2. L.A. Lasky, Selectin-carbohydrate interactions and the initiation of the inflammatory response, *Ann. Rev. Biochem.* 64:113 (1995).
3. C.R. Bertozzi, Cracking the carbohydrate code for selectin recognition, *Chem. Biol.* 2:703 (1995).
4. R.M. Cooke, R.S. Hale, S.G. Lister, G. Shah, M.P. Weir, The conformation of the sialyl Lewis X ligand changes upon binding to E-selectin, *Biochemistry* 33:10591 (1994).
5. T.P. Kogan, B.M. Reville, S. Tapp, D. Scott, P.J. Beck, A single amino acid residue can determine the ligand specificity of E-selectin, *J. Biol. Chem.* 270:14047 (1995).
6. P.V. Murphy, R.E. Hubbard, D.T. Manallack, J.G. Montana, R.J.K. Taylor, The synthesis of novel structural analogues of sialyl Lewis X, *Tetrahedron Lett.* 39:3273 (1998).

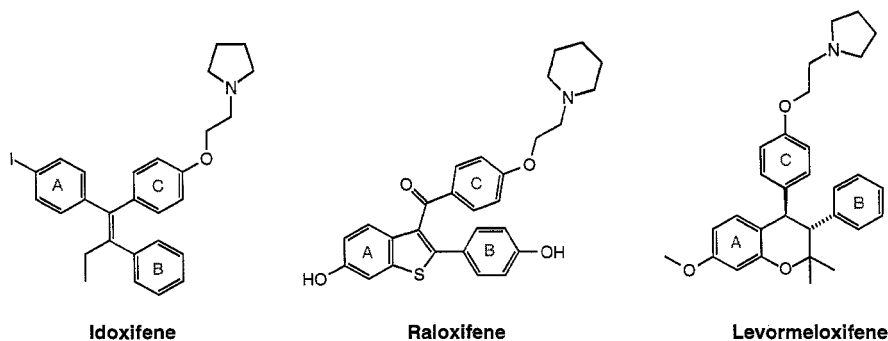
# Conformational Analysis and Pharmacophore Identification of Potential Drugs for Osteoporosis

Jan Høst, Inge Thøger Christensen and Flemming Steen Jørgensen

The Royal Danish School of Pharmacy, Department of Medicinal Chemistry,  
Universitetsparken 2, DK-2100, Copenhagen, Denmark

Lack of stimulation of the estrogen receptors in the bones is the primary reason for postmenopausal osteoporosis. Replacement therapy has been used for years but, unfortunately, it has adverse effects in breast and uterus due to agonistic estrogen receptor effects in these tissues.

NSERT's (Nonsteroid Selective Estrogen Receptor Therapeutics) or SERM's (Selective Estrogen Receptor Modulators) are compounds showing agonistic effect on estrogen receptors in the bones, more or less lacking the adverse effects in breast and uterus [1].



*NSERT's are potential drugs against osteoporosis [1]*

## Conformational analysis

Molecular dynamics simulations on the three NSERT's were performed using the Tripos force field in Sybyl and the MM3\* force field in Macromodel. In Macromodel both the water-solvation option and the no-solvation option were used. Sybyl has no water-solvation option, but a dielectric constant of 4 was chosen to implicitly take solvation into account [2]. The molecular dynamics simulations covered the conformational space for all three NSERT's efficiently, making the conformational analyses reliable.

All Conformations generated in the molecular dynamics simulations were energy minimized and compared. When the low-energy conformations were superimposed, the conformations were very similar for the three NSERT's, apart from the aliphatic sidechain containing the tertiary amine. Raloxifene and idoxifene each had two dominating enantiomeric conformations, while levormeloxifene was limited to just a single conformation apart from at the aliphatic sidechain.

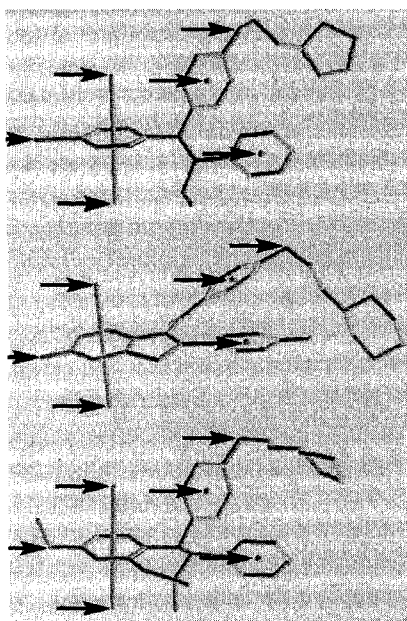
The conformational analysis revealed that all the low-energy conformations are very similar, hence only a negligible fraction adopts other conformations. Due to the high affinities of the three compounds, the receptor binding conformation must be one of the dominating conformations.

## Superimposition

Based on the crystal structure of the ligand binding domain [3], reviews on the subject [4] and on the similarities in chemical structure, the following pharmacophore elements were selected.

- Ring A is probably locked tightly in the receptor. By using two dummy atoms placed on the normal of the ring plane, it was possible to superimpose the ring planes.
- The substituent on ring A, which is a possible hydrogen bond donor, was also used.
- The relative positions of the aromatic rings are important and the ring centers of ring B and C were used for fit.
- The nitrogen atom is supposed to make a hydrogen bond to the receptor. Therefore, the ether oxygen was used in order to control the direction of the aliphatic sidechain.

Conclusively, only one of the two enantiomeric conformations of raloxifene and idoxifene, respectively, fits well with the single dominating conformation of levormeloxifene.

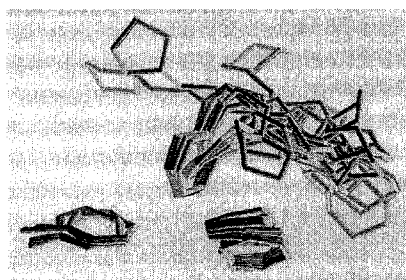


*Atoms included in the superimposition*

## Pharmacophore model

Ring B is located 4.5-6.5 Å from ring A and very close to the plane of ring A. Ring C is 6.2-6.4 Å from ring A and 3.6-4.4 Å from ring B. The anchor points of the aliphatic sidechains are located very close in space. The maximum distance between two of the ether oxygens in the fit is 2.2 Å allowing the nitrogen atom to occupy the same position for all the NSERT's.

The pharmacophore model for the NSERT's may be used for studying SAR and for the design of potential drugs against osteoporosis.



*Superimposition of low-energy conformations of NSERT's*

## Acknowledgements

Financial support from Novo Nordisk and Apoteker Julius Wael's Fond is gratefully acknowledged.

## References

1. B. H. Mitlak & F. J. Cohen, "In Search of Long-Term Female Hormone Replacement: The Potential of Selective Estrogen Receptor Modulators", *Hormone Research*, 1997, **48**, 155-163.
2. I. T. Christensen & F. S. Jørgensen, "Conformational analysis of six- and twelve-membered ring compounds by molecular dynamics", *Journal of Computer-Aided Molecular Design*, 1997, **11**, 385-394.
3. A. M. Brzozowski et al., "Molecular basis of agonism and antagonism in the oestrogen receptor", *Nature*, 1997, **389**, 753-758.
4. G. M. Anstead, K. E. Carlson & J. A. Katzenellenbogen, "The estradiol pharmacophore: Ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site", *Steroids*, 1997, **62**, 268-303.



## MOLECULAR MODELLING OF DNA ADDUCTS OF BBR3464: A NEW PHASE I CLINICAL AGENT

G. De Cillis\*, E. Fioravanzo<sup>§</sup>, M. Mabilia<sup>§</sup>, J. Cox<sup>°</sup>, N. Farrell<sup>°</sup>

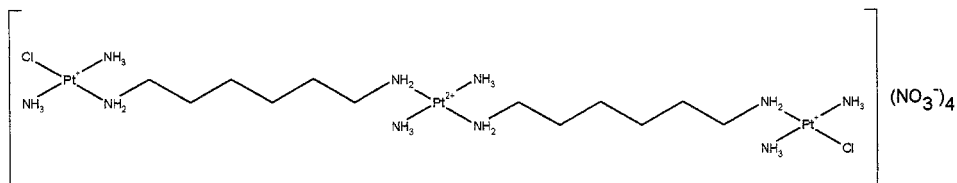
\* Research Center, Boehringer Mannheim Italia S.p.A., Via G.B. Stucchi 101, I-20052 Monza, Italy

<sup>§</sup> S.IN - Soluzioni Informatiche S.a.s., Via Salvemini 9, I-36100 Vicenza, Italy

<sup>°</sup> Department of Chemistry Virginia, Commonwealth University, Richmond, Virginia 23284-2006

### INTRODUCTION

BBR3464 is a novel Phase I clinical agent based on a triplatinum structure (**Figure 1**), trans-[bis{trans-diaminechloroplatinum( $\mu$ -1,6-hexanediamine)}]diamineplatinum tetranitrate salt. Its DNA binding is characterized by a high percentage of interstrand cross-links and the ability to induce the B -- Z conformation in poly(dG.dC).poly(dG.dC). To help characterize these novel DNA adducts further, we have begun a program to model the types of interactions and the ensuing conformational changes induced by covalent binding of BBR3464 to DNA.



**Figure 1** - BBR3464: a novel Phase I clinical agent based on a triplatinum structure

In this preliminary study we present results of molecular mechanics and molecular dynamics calculations of different models of DNA in the Z form<sup>1</sup>. Such form exists in the alternating sequence poly(dC-dG).poly(dC-dG) and is presumed to be the structure formed in solution in high salt (> 2.5 M NaCl) conditions.

Energy minimizations and molecular dynamics simulations are employed to investigate the conformational properties of Z-DNA and, in particular, to compare computer-generated and minimized models with x-ray experimentally-determined structures.

## METHODS

All modeling studies employ the AMBER<sup>ii</sup> force field implemented in Batchmin 6.0, part of MacroModel<sup>iii</sup> 6.0. Since the calculations are performed for the vacuum state, the following procedures are adopted to simulate the effect of counterions and shielding, as well as solvent effects:

1. charges on phosphate groups are reduced so as to give a slightly negative charge on each nucleotide<sup>iv</sup>
2. a distance-dependent dielectric of the form  $\epsilon = cR_{ij}$  is chosen so as to mimic the effects of solvent and shielding<sup>v</sup>.

**Energy Minimizations.** All 3D models are first minimized for 500 cycles with Steepest Descent and then refined with a Polak Ribiere Conjugate Gradient algorithm with the rms derivative convergence criterion set to 0.01 kcal/Å-mol.

**Molecular Dynamics.** Molecular dynamics simulations are started from energy-minimized structures. The Verlet algorithm is used with a time interval of 1 fs per step. Hydrogen atom bond lengths are constrained with the SHAKE algorithm and temperature is maintained by a thermal bath at a value of 300 K. The structures are equilibrated for 50 ps and then the constant temperature dynamics simulation is continued for 500 ps. This portion of the trajectory is then used for subsequent analysis by sampling a structure each 2 ps.

## RESULTS AND DISCUSSION

Energy minimizations were first performed on the crystal structure of the self-complementary 5'-purine start decamer d(GpCpGpCpGpCpGpCpGpC) in the Z-DNA conformation<sup>vi</sup> to establish how well the optimized structure thus obtained compares to the experimentally-determined one. A theoretical model, generated by MacroModel, of the same decamer was then energy minimized to check the consistency of calculated vs. experimental geometries. The minimization procedure led to two different minima, corresponding to Z<sub>II</sub> and Z<sub>I</sub> type of backbone conformation. The rms of the minimized model relative to the original x-ray structure is 1.13 Å. Though the two minimized geometries are very similar, their corresponding energy values are quite different, about 20 kcal/mol.

Molecular dynamics simulations starting from the x-ray and, respectively, the theoretical structures show a higher similarity between these two models, as opposed to the results obtained by energy minimization. The two molecules explore similar conformational space, so that the time-averaged geometries are close to the experimental structure (rms of about 0.8 Å) and very similar to one another. The backbone conformation of both time-averaged structures is Z<sub>II</sub>.

It is also worth noticing that the initial energy difference of 20 kcal/mol for the two minimized structures is now reduced to an average potential energy difference of 3.4 kcal/mol scaled to 300 deg K. Such an agreement is due to the transition - occurred during the equilibration period - of the theoretical structure from the Z<sub>I</sub> to Z<sub>II</sub> conformation. The good overlap of both structures, as observed via dynamics simulations, confirms the validity of 3D molecular models for subsequent platination studies.

<sup>i</sup> Arnott, S., Chandrasekaran, R., Birdsall, D.L., Leslie, A.G.W., and Ratliffe, R.L., *Nature*, **283**, 743 (1980).

<sup>ii</sup> (a) Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S.Jr. and Weiner, P., *J. Am. Chem. Soc.*, **106**, 765 (1984). (b) Weiner, S.J., Kollman, P.A. *J. Comp. Chem.*, **7**, 230 (1986). (c) Weiner, S.J., Kollman, P.A., Nguyen, N.T., Case, D.A., *J. Comp. Chem.*, **7**, 230 (1987).

<sup>iii</sup> (a) MACROMODEL c/o Prof. W.C. Still, Columbia University, New York, NY. (b) Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., Still, W.C., *J. Comp. Chem.*, **11**, 440 (1990).

<sup>iv</sup> Tidor, B., Irikura, K.K., Brooks, R., Karplus, M., *J. Biom. Struct. Dyn.*, **1**, 231 (1983).

<sup>v</sup> Gelin, B.R., Karplus, M., *Proc. Natl. Acad. Sci. USA*, **74**, 801 (1977).

<sup>vi</sup> Ban, C., Ramakrishnan, B., Sundaralingam, M., *Biophys. J.*, **71**, 1215 (1996). PDB ID code : 279d.

## PREDICTION OF ACTIVITY FOR A SET OF FLAVONOIDS AGAINST HIV-1 INTEGRASE

Jarmo J. Huuskonen,<sup>1</sup> Heikki Vuorela,<sup>2</sup> and Raimo Hiltunen<sup>2</sup>

<sup>1</sup>Division of Pharmaceutical Chemistry, <sup>2</sup>Division of Pharmacognosy,  
Department of Pharmacy, POB 56, FIN-00014 University of Helsinki,  
Finland

### INTRODUCTION

The development of potent antiviral drugs against infection by human immunodeficiency virus (HIV), a causative agent of acquired immunodeficiency syndrome (AIDS), still remains an urgent need. An ideal specific target for the chemotherapeutic treatment of HIV is the virus encoded enzyme integrase. Integrase has a key role at the early stage of HIV infection, which is responsible for converting the integrase of the double-strained DNA transcriptase into the host genome. Integrase acts in two steps, i.e. cleavage and integration steps. In *in vitro* assays a method for estimating the inhibition of HIV-1 integrase for both steps has been developed and used to identify classes of compounds with a potent inhibitory activity against HIV-1 integrase.<sup>1</sup> Flavonoids, like quercetin, were found to be one set of these compounds.

Two QSAR studies have been published for a set of 15 active flavonoids against HIV-1 integrase. In a partial least squares (PLS) method with comparative molecular field analysis (CoMFA) parameters, a strong correlation was found between inhibitory activity of these flavonoids, and the steric and electrostatic fields around them.<sup>2</sup> Recently, Kier and Hall<sup>3</sup> introduced electrotopological state (E-state) indices for molecular structure description in which both electronic and topological characteristics are combined together. Bualamwini et al.<sup>4</sup> used 17 skeletal E-state indices common for all flavonoids as structural parameters in a principle component regression (PCR) analysis.

In our previous study E-state indices were found practical in the prediction of water solubility of structurally related drug compounds based on neural network modeling.<sup>5</sup> The present study shows that the same indices can be successfully used to predict the activity for a set of 15 flavonoids.

## METHODS

Activities of the 15 flavonoids against HIV-1 integrase were modified from Fesen et al.<sup>1</sup> and were expressed as negative logarithm values of  $IC_{50}$ ,  $-\log IC_{50}$ . Structural parameters were calculated by Molconn-Z software (Hall Associated Consulting, Quincy, MA). 17 E-state indices calculated for each analyzed compound were analyzed using multilinear regression (MLR) analysis and artificial neural networks (ANNs). The SPSS package was used to run the MLR analysis. The ANNs were conducted by NeuDesk program, and were fully connected, feed-forward back-propagation networks with one hidden layer and bias neurons. The Early Stopping over Ensemble method was used to accomplish the overfitting/overtraining problem and to improve generalization ability of neural networks.<sup>5</sup>

## RESULTS AND DISCUSSION

Stepwise and backward methods were employed in the regression analysis. Satisfactory MLR models were detected for the 15 flavonoids containing 3 parameters ( $R = 0.88$ ,  $q^2 = 0.73$ ,  $s_{LOO} = 0.35$  for the step 1, and  $R = 0.89$ ,  $q^2 = 0.52$ ,  $s_{LOO} = 0.51$  for the step 2), where cross-validated  $q^2$  and the standard deviation  $s_{LOO}$  were calculated by leave-one-out method. The E-state indices for atoms O4, C'4 and C'5 in cleavage step (step 1), and for the atoms C5, C6 and C'5 in integration step (step 2) were found the most significant in MLR analysis.

Neural networks applied to analyze the same sets of 3 E-state indices calculated higher prediction ability for the 15 flavonoids ( $q^2 = 0.81$ ,  $s_{LOO} = 0.30$  for step 1 and  $q^2 = 0.78$ ,  $s_{LOO} = 0.34$  for step 2). No outliers were found in both MLR and ANN models. Their prediction ability for the set of 15 compounds was comparable with those found using other known methods, such as, PLS with CoMFA parameters ( $q^2 = 0.81$  for the step 1 and  $q^2 = 0.78$  for the step 2), and PCR with E-state indices ( $q^2 = 0.51$  for the step 1 and  $q^2 = 0.73$  for the step 2). However, in both of these analyses one compound, 6-methoxyluteolin, was omitted as an outlier and the models were constructed for the remaining 14 flavonoids.

The structure-based design of the new integrase inhibitors relies often on QSAR analysis. The estimation of activity for a set of 15 flavonoids using MLR analysis and E-state indices is accurate and provides reliable  $-\log IC_{50}$  predictions comparable with those obtained by other methods. The use of neural networks provides better predictive ability than the present MLR analysis, and the previous PLS and PCR methods. An advantage of the proposed approach is that the E-state indices can be quickly and easily estimated directly from the chemical structure of the analyzed compounds. Thus, the present approach introduces a fast and accurate method for the estimation of activity of chemical compounds to guide drug design.

## Acknowledgments

This study was partially supported by the Technology Development Center in Finland (TEKES) and Paulig Group, Pimenta Ltd in Finland.

## REFERENCES

1. M.R. Fesen, Y. Pommier, F. Leteurte, S. Hiroguchi, J. Yung, and K.W. Kohn, Inhibition of HIV-1 integrase by flavones, caffeic acid phenethyl ester (CAPE) and related compounds, *Biochem.Pharmacol.* 48:595(1994).
2. K. Raghavan, J.K. Bualamwini, J.K. Fesen, Y. Pommier, K.W. Kohn, and J.N. Weinstein, Three-dimensional quantitative structure-activity relationship (QSAR) of HIV integrase inhibitors: A comparative molecular field analysis (CoMFA) study, *J.Med.Chem.* 38:890(1995).
3. L.B.Kier, and L.H.Hall, An electrotopological-state index for atoms in molecules, *Pharm.Res.* 7:801(1990).
4. J.K. Bualamwini, K. Raghavan, M.R. Fesen, Y. Pommier, K.W. Kohn, J.N. Weinstein, Application of the electrotopological state index to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors, *Pharm.Res.* 13:1892(1996).
5. J.Huuskonen, M. Salo, and J.Taskinen, Neural network modeling for estimation of the aqueous solubility of structurally related drugs, *J.Pharm.Sci.* 86:450(1997).

## STRUCTURE-BASED DISCOVERY OF INHIBITORS OF AN ESSENTIAL PURINE SALVAGE ENZYME IN *TRITRICHOMONAS FOETUS*

Ronald M.A. Knegt<sup>1</sup>, John R. Somoza<sup>2</sup>, A. Geoffrey Skillman Jr.<sup>3</sup>, Narsimha Munagala<sup>3</sup>, Connie M. Oshiro<sup>3</sup>, Solomon Mpoke<sup>3</sup>, Shinichi Katakura<sup>3</sup>, Robert J. Fletterick<sup>2</sup>, Irwin D. Kuntz<sup>3</sup> and Ching C. Wang<sup>3</sup>

<sup>1</sup>Dept. Of Molecular Design and Informatics, N.V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands

<sup>2</sup>Dept. Of Biochemistry and Biophysics, University of California, San Francisco CA94143

<sup>3</sup>Dept. Of Pharmaceutical Chemistry, University of California, San Francisco CA 94143

### INTRODUCTION

Most protozoan parasites, such as Leishmania, Plasmodium, Toxoplasma and Trypanosoma, rely on a salvage pathway for their supply of purine ribonucleotides (Wang, 1984). Inhibition of this pathway therefore presents an interesting approach in the fight against microbial infections. To explore the feasibility of this approach we have attempted to identify inhibitors of the essential purine salvage enzyme hypoxanthine-guanine-xanthine phosphoribosyl transferase (HGXPRTase) of the protozoan parasite *Tritrichomonas foetus*. This sexually transmitted parasite causes bovine trichomoniasis, which can lead to embryonic death and infertility in cows. *T. foetus* relies primarily on a single enzyme, HGXPRTase, to transfer ribose 5-phosphate from  $\alpha$ -D-5-phosphoribosyl-1-pyrophosphate to the N9 nitrogen atom of hypoxanthine, guanine or xanthine (Wang et al., 1983). Selectivity with respect to the mammalian enzyme hypoxanthine-guanine-phosphoribosyl transferase (HGPRTase), that has 27 % sequence identity with the parasite enzyme, is important to avoid serious side effects. Currently available inhibitors are purine analogues with affinities in the millimolar range (Jadhav et al., 1979). Here we report the use of the molecular docking program DOCK 3.5 (Kuntz et al., 1982; Meng et al., 1993) for the discovery of more potent, novel inhibitors of HGXPRTase that are selective with respect to the human enzyme (Somoza et al., 1998).

## METHODOLOGY

Commercially available small molecules listed in the Available Chemicals Directory (MDL, San Leandro USA) were docked with DOCK 3.5 into the enzyme's active site as observed in the 1.9 Å crystal structure of *T. foetus* HGXPRTase (Somoza et al., 1996). Since the active site is large and shallow (10 x 10 x 5 Å), the DOCK 3.5 bump checking routines were modified such that ligand atoms protruding from the box used for grid-based scoring were counted as bumps. This modification forced ligands to fill the binding site region where GMP was observed to bind. Docked inhibitors were observed to be positioned in the guanine binding pocket, suggesting that these inhibitors should be competitive with GMP and guanine. On the basis of the hits found among compounds selected from the initial docking calculations, additional compounds were selected from the ACD by using substructure and similarity searches with Daylight v4.42 (Daylight Chemical Information Systems Inc., Santa Fe, NM) and minimization of docked inhibitors with Sybyl 6.2 (Tripos Associates, St. Louis, MO) in the active site.

## RESULTS AND DISCUSSION

Molecular docking of commercially available compounds yielded two active indol-2-one (isatin) ( $IC_{50}=240\mu M$ ) and phthalic anhydride ( $IC_{50}=300\mu M$ ) derivatives from 18 compounds tested. Further improvement of the affinity was achieved by selecting 22 similar compounds of which 18 (82%) were active and 10 (45%) inhibited the enzyme with potencies equal to or higher than the original lead compounds (up to 22-50  $\mu M$ ). All compounds, except the original isatin derived lead, had  $IC_{50}$ 's for the human enzyme of over 1 mM. One of these compounds (4-[*N*-(3,4-Dichlorophenyl)carbamoyl]phthalic anhydride;  $IC_{50}=50\mu M$ ) is a competitive inhibitor of HGXPRTase with respect to guanine ( $K_i=13\mu M$ ) and GMP ( $K_i=10\mu M$ ). The same compound inhibits *in vitro* growth of *T. foetus* with an  $IC_{50}$  of ~40  $\mu M$ . This inhibition could be reversed by adding hypoxanthine to the growth medium. Our results demonstrate that targeting the HGXPRTase enzyme of protozoan parasites presents a promising approach against microbial infection. Furthermore, it is shown that databases of commercially available compounds can be used to identify and perform a first optimization of selective enzyme inhibitors.

## REFERENCES

- Jadhav A.L., Townsend L.B., Nelson J.A., 1979, Inhibition of hypoxanthine-guanine phosphoribosyl transferase, *Biochem Pharmacol.* 28:1057.
- Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R. and Ferrin T.E., 1982, A geometric approach to macromolecule-ligand interactions, *J. Mol. Biol.* 161:269
- Meng E.C., Gschwend D.A., Blaney J.M., Kuntz I.D., 1993, Orientational sampling and rigid-body minimization in molecular docking, *Proteins* 17:266
- Somoza J.R., Skillman A.G. Jr, Munagala N.R., Oshiro C.M., Knegtel R.M.A., Mpoke S., Fletterick R.J., Kuntz I.D. and Wang C.C., 1998, Rational design of novel antimicrobials: blocking purine salvage in a parasitic protozoan, *Biochemistry* 37:5344.
- Somoza J.R., Chin M.S., Focia P.J., Wang C.C. and Fletterick R.J., 1996, Crystal structure of the hypoxanthine-guanine-xanthine phosphoribosyltransferase from the protozoan parasite *Tritrichomonas foetus*, *Biochemistry* 35:7032
- Wang, C.C., Verham R., Rice A. and Tzeng S., 1983, Purine salvage by *Tritrichomonas foetus*, *Mol. Biochem. Parasitol.* 8:325.
- Wang, C.C., 1984, Parasite enzymes as potential targets for antiparasitic chemotherapy, *J. Med. Chem.* 27:1.

# A 3D-PHARMACOPHORE MODEL FOR DOPAMINE D<sub>4</sub> RECEPTOR ANTAGONISTS

Jonas Boström, Klaus Gundertofte and Tommy Liljefors

Department of Medicinal Chemistry, Royal Danish School of Pharmacy,  
Universitetsparken 2, DK-2100 Copenhagen, Denmark. Research and Development,  
H. Lundbeck A/S, Ottiliavej 9, DK-2500, Copenhagen, Denmark.

## INTRODUCTION

Selective dopamine (DA) D<sub>4</sub> receptor antagonists may be effective antipsychotics, without the extrapyramidal side effects which are well established for DA D<sub>2</sub> antagonists. In order to facilitate the design of new selective DA D<sub>4</sub> receptor antagonists we are currently developing a DA D<sub>4</sub> 3D-pharmacophore model. Previously, a 3D-pharmacophore model for DA D<sub>2</sub> antagonists has been developed<sup>1</sup>. This model rationalizes the high affinity of both enantiomers of octoclothebin (**1**)<sup>2</sup>.

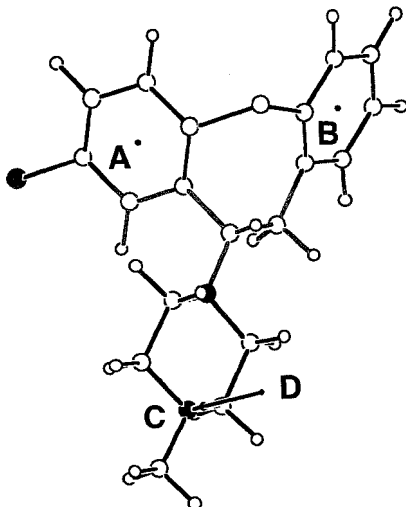
## RESULTS

Both enantiomers of **1** are also high affinity D<sub>4</sub> antagonists. An analysis of the calculated potential energy curves for superimposed (R)-**1** and (S)-**1**<sup>2</sup> indicates that the bioactive D<sub>2</sub> conformation of (S)-**1** may also be used as a template for D<sub>4</sub> receptor antagonists, with respect to the pharmacophore elements A-D defined in Figure 1.

A large number of structurally diverse DA D<sub>4</sub> selective antagonists have been superimposed on the template molecule (S)-**1** in low-energy conformations with low rms deviations. The average rms deviation is 0.27 Å and the average conformational energy is 1.0 kcal/mol. All calculations were carried out with the MacroModel program<sup>3</sup> using the MM3\* force field and the GB/SA continuum model.



Substitutions at various positions in 1-piperazino-3-phenylindans and -1H-indoles give very similar effects on  $D_2$  and  $D_4$  affinities. In these systems,  $D_2/D_4$  selectivity can not be achieved by substitution in the indan or indole rings or the piperazino ring (excluding N-substituents). Our analysis indicates that the major contribution to  $D_2/D_4$  selectivity is to be found in the effects on the affinities of N-alkyl substituents.



**Figure 1.** The suggested DA  $D_4$  bioactive conformation of (S)-1. Four pharmacophore elements are identified, the centre of the two aromatic moieties (A and B), a nitrogen (C) and a site point located 2.8 Å from the nitrogen in the N lone-pair direction (D).

## CONCLUSIONS

(S)-1 in its DA  $D_2$  bioactive conformation may be used as a template for DA  $D_4$  receptor antagonists. A broad selection of structurally diverse DA  $D_4$  antagonists may be superimposed, in a low-energy conformation and with low rms deviation, on the suggested bioactive conformation of (S)-1. The principal difference, and thus the main reason for selectivity between the DA  $D_4$  and  $D_2$  antagonists, is most probably due to the different effects of the N-alkyl substituents. The development of an extended model incorporating the properties of the N-alkyl substituents is in progress.

## REFERENCES

1. Liljefors, T., and Bøgesø, K. P., 1988, *J. Med. Chem.*, 37:306.
2. Bøgesø, K. P., Liljefors, T., Arnt, J., Hyttel, J., and Pedersen, H., 1991 *J. Med. Chem.*, 34:2023.
3. MacroModel V6.0: Mohamadi, F., Richards, N. G. J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrikson, T., and Still, W. C., 1990, *J. Comput. Chem.*, 11:440.

## MOLECULAR MODELING AND STRUCTURE-BASED DESIGN OF DIRECT CALCINEURIN INHIBITORS

Xinjun J. Hou, John H. Tatlock, M. Angelica Linton, Charles R. Kissinger, Laura A. Pelletier, Richard E. Showalter, Anna Tempczyk, and J. Ernest Villafranca

Agouron Pharmaceuticals, Inc.  
3565 General Atomics Court  
San Diego, CA 92121

### INTRODUCTION

Calcineurin (Protein Phosphatase 2B, or PP2B) is a  $\text{Ca}^{2+}$ /calmodulin dependent protein phosphatase which plays critical roles in intracellular signaling processes<sup>1-3</sup>. An important role of calcineurin is its dephosphorylation function of NFAT(nuclear factor-activated T-cells), allowing NFAT to enter the nucleus and activate the transcription of T-cell specific genes. The inhibition of calcineurin by immunosuppressant drug (FK506 or cyclosporin A) disrupts the T-cell activation and leads to immunosuppressant effects.

Calcineurin(CN) is a heterodimer composed of an A subunit (CNA) and a B subunit (CNB)<sup>4</sup>. CNA has four distinct functional domains: a catalytic domain, a CNB binding domain, a calmodulin(CAM) binding domain and an auto-inhibitory(AI) domain. The protein phosphatase activity of calcineurin is stimulated by  $\text{Ca}^{2+}$  binding to CNB and  $\text{Ca}^{2+}$ -induced binding of CAM to CNA. The function of CAM is presumably to remove the AI and CNA interaction, enabling the access to the catalytic active site. The catalytic active site of CNA contains two metal ions (Fe and Zn) and is locally homologous to protein phosphatases PP1 and PP2A, yet calcineurin does not share substrate specificity with them, indicating importance of secondary structure recognition.

Immunophilin and immunosuppressant drug complexes do not bind to the catalytic active but a region of CNA-CNB interface<sup>4,5</sup>, indirectly blocking access to the active site by physiological substrates(NFAT, etc.). Most immunosuppressant drug research efforts were toward the analogs of immunosuppressant FK506 or cyclosporin A where the mechanism of inhibition is mediated or coordinated by immunophilins FKBP or cyclophilin. The determinations of two X-ray crystal structures: a) calcineurin with a portion of the AI peptide and b) its complex with immunosuppressant drug FK506 provided an insight of calcineurin's regulatory mechanism and offered a new opportunity in the design of potential immunosuppressant drugs<sup>4,5</sup>. We summarize here the structure-based design and synthesis of calcineurin specific inhibitors targeted directly toward the catalytic active site, using molecular modeling techniques of binding mode predictions by ligand docking simulation<sup>6</sup>, X-ray crystallography, and binding energetics evaluations<sup>7</sup>.

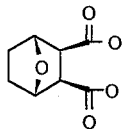
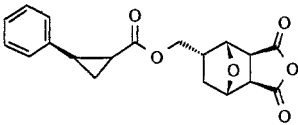
## INHIBITOR DESIGN AND MODELING STRATEGIES

1. Build binding models of initial lead(1) by multiple docking simulation
2. Validate and improve binding prediction method via SAR and X-ray crystallography
3. Enhance binding affinity via ligand-protein hydrophobic interactions
4. Address specificity to calcineurin through AI peptide recognition pocket
5. Analyze and rank designed ligands *a priori* using binding mode prediction and force field energy minimization(MacroModel/Batchmin)

## SUMMARY

1. Novel calcineurin specific inhibitors were design and synthesized with application of molecular model techniques.
2. Binding affinity was increased via hydrophobic interaction and conformational rigidity
3. Specificity was improved by targeting to AI recognition pockets
4. Binding mode prediction method was sensitive to a key protein sidechain conformation, but it could be improved with additional information (SAR and X-ray crystallography)

**Table 1.** Improvement of inhibition and specificity to calcineurin(PP2B)

	Compound	$K_{i,app}$ ( $\mu\text{M}$ )	$K_{i,app}$ ( $\mu\text{M}$ )
		Calcineurin (PP2B)	PP1
1		11.5	4.0
2		0.51	4.0

## REFERENCES

1. Guerini, D. Calcineurin: Not just a simple protein phosphatase. *Biochem. and Biophys. Res. Comm.* **235**: 271 (1997).
2. Klee, C. B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. of Biological Chemistry* **273**:13367 (1998).
3. Kay, J. E. & Gilfoyle, D. J. Immunophilin-mediated inhibition of lymphocyte Signalling in *Lymphocyte Signalling: Mechanisms, Subversion and Manipulation* (ed. Rigley, M. M. H. a. K. P.) (John Wiley & Sons Ltd, 1997).
4. Kissinger, C. R. *et al.* Crystal structures of human calcineurin and the human FKBP12--FK506-calcineurin complex. *Nature* **378**:641 (1995).
5. Griffith, J. P. *et al.* X-ray structure of calcineurin inhibited by the immunophilin-immunosuppressant FKBP12-FK506 complex. *Cell* **82**:507 (1995).
6. Gehlhaar, D. K. *et al.* Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & Biology* **2**:317 (1995).
7. Tatlock, J. H. *et al.* Structure-based design of novel calcineurin (PP2B) inhibitors. *Bioorg. & Med. Chem. Lett.* **7**:1007 (1997).

## CONFORMATIONAL FLEXIBILITY AND RECEPTOR INTERACTION

Lambert H. M. Janssen

Universiteit Utrecht, Faculty of Pharmacy  
Utrecht Institute of Pharmaceutical Sciences  
Department of Medicinal Chemistry, PO BOX 80082  
3508 TB Utrecht, The Netherlands

### CONFORMATION AND RECEPTOR BINDING

Protein-ligand interactions are essential for many biological processes. The ligands are flexible and may assume several conformations in solution. Once bound to the receptor, the ligand is assumed to occur in a so-called receptor bound or bio-active conformation. Sometimes the solution conformation corresponding with this conformation is also given the name of bio-active conformation. The question is: how is the experimentally measured binding constant related to the diversity in (solution) conformations and to the bio-active conformation?

We are going to discuss a number of reactions and equations. These are summarised in Table 1. The notation Eq.1 and Re.1 refers to Equation 1 and Reaction 1 in this table and so on.

The reaction which can be measured in solution is given by Re.1. L represents the ligand in solution (in fact a whole collection of conformations), R the receptor (for simplicity we assume one solution conformation for the receptor), L\*R denotes the drug-receptor complex in which the ligand is bound in a unique conformation. The  $\Delta G^0_{\text{obs}}$  of this reaction can be calculated from  $\Delta G^0_{\text{obs}} = -RT \ln K$  (K is the measured association constant), see Eq.1.

The ligand may assume several conformations in solution, each having its own energy level  $i$ , so  $L_i$  denotes the ligand having a conformational energy  $E(L_i)$ . These energies have increasing values,  $E(L_0)$  indicating the lowest energy of the (lowest energy) conformation  $L_0$ . The reaction between 1 mol of  $L_0$  and the receptor, and the corresponding standard Gibbs free energy  $\Delta G^0_0$  are given by Re.2 and Eq.2. This  $\Delta G^0_0$  includes enthalpy and entropy terms leading to the L\*R complex, but these aspects are not discussed here.

Let us next consider the binding of a higher energy conformation  $L_i$  to R leading to the formation of the L\*R complex (Re.3). The energy involved in this binding process can be calculated by first converting 1 mol of  $L_i$  to the ground state. In this step an amount of energy (or as frequently assumed: enthalpy) is released equal to  $E(L_0) - E(L_i)$ . In the next step the binding of 1 mol of ligand now in the state  $L_0$  takes place. This results in an overall standard free energy change given by Eq.3. Because  $\Delta G^0_i$  is more negative than  $\Delta G^0_0$ , high energy conformations bind more strongly than low energy conformations. It would be advantageous if only high energy conformation became bound to the receptor. However ligands

are distributed over the several energy levels according to the laws of statistical thermodynamics. The Boltzmann distribution law makes it possible to calculate the fraction  $f_i$  of the free ligand molecules which have a conformational energy equal to  $E_i$ . This distribution is not influenced by the presence of receptors. This means that in the experimental process of binding of L to R,  $f_i$  remains constant. If molecules should be selected by the receptor from one single energy level, a redistribution must occur in order to obey the Boltzmann distribution law. Therefore the binding of 1 mol of L as given in Re.1 means that  $f_0$  mol of  $L_0$ ,  $f_1$  mol of  $L_1$ , etc. are bound. The amount of energy involved in the binding of  $f_i$  mol of  $L_i$  to R is given by  $f_i \Delta G_i^0$  (with  $\Delta G_i^0$  defined in Eq.3). The experimental binding of 1 mol of L is there-

**Table 1.** Survey of reactions and equations used in this abstract

Reaction	Reaction no.	$\Delta G^0$ of reaction	Equation no.
$L + R \rightleftharpoons L^*R$	1	$\Delta G_{\text{obs}}^0 = -RT \ln K$	1
$L_0 + R \rightleftharpoons L^*R$	2	$\Delta G_0^0$	2
$L_i + R \rightleftharpoons L^*R$	3	$\Delta G_i^0 = \Delta G_0^0 + E(L_0) - E(L_i)$	3
$L + R \rightleftharpoons L^*R$	1	$\Delta G_{\text{obs}}^0 = \Delta G_0^0 + E(L_0) - E(L)_{\text{av}}$	4
$L_* + R \rightleftharpoons L^*R$	4	$\Delta G_*^0 = \Delta G_0^0 + E(L_0) - E(L_*)$	5
$L + R \rightleftharpoons L^*R$	1	$\Delta G_{\text{obs}}^0 = \Delta G_i^0 + E(L_i) - E(L)_{\text{av}}$	6
$L + R \rightleftharpoons L^*R$	1	$\Delta G_{\text{obs}}^0 = \Delta G_*^0 + E(L_*) - E(L)_{\text{av}}$	7

fore given by  $\Delta G_{\text{obs}}^0 = \sum_i f_i \Delta G_i^0$ . Realising that  $\sum_i f_i = 1$  and that  $\sum_i f_i E(L_i)$  is equal to the average energy of the free ligand molecules, Eq.4 is easily obtained. We note that the observed (standard Gibbs) free energy of binding is more negative (the binding is stronger) than the free energy of binding of the lowest energy conformation.

Eq.5, 6 and 7 can be derived from the Eq.2, 3 and 4.

## DISCUSSION

The conclusion that higher energy conformations have a higher affinity than lower energy conformations is independent of the conformation in which the ligand is bound to the receptor. This is because the term  $E(L_0) - E(L_i)$  is independent of the energy of the receptor bound conformation. This conclusion has no further practical implication because this process can not be realised experimentally. On the other hand it is necessary to state this with some emphasis because remarks can be found in literature which suggest the opposite.

Re.4 is a special case of Re.3.  $L_*$  indicates the solution equivalent of the receptor bound conformation. Note also Eq.7. This kind of relations should be used when calculated interaction energies are to be related with experimentally observed ligand binding constants.

The view developed here is also relevant in considering the activity of a series of related compounds, where it may be assumed that differences in activity are more likely to be caused by differences in  $\Delta G_0^0$  than by differences in  $E(L_0) - E(L)_{\text{av}}$ .

A more detailed account of the views presented here has been published recently<sup>1</sup>.

## REFERENCES

1. L.H.M. Janssen, Conformational flexibility and receptor interaction, *Bioorg. Med.Chem.* 6:785 (1998).

## INVESTIGATING THE MIMETIC POTENTIAL OF $\beta$ -TURN MIMETICS

Susanne Winiwarter, Anders Hallberg, and Anders Karlén

Dept. of Organic Pharmaceutical Chemistry  
Uppsala Biomedical Centre  
Uppsala University  
SE-751 23 Uppsala  
SWEDEN

### INTRODUCTION

The  $\beta$ -turn is a common structural feature in peptides and proteins, consisting of four consecutive amino acids with a distance of less than  $7\text{\AA}$  between the  $C_{\alpha}$  atoms of the first and fourth amino acid ( $d_{\alpha}$ ).<sup>1</sup> Since  $\beta$ -turns are often considered to be important for molecular recognition, they are interesting templates to use in the design and synthesis of peptidomimetics. It is important to know precisely which  $\beta$ -turn conformation(s) the peptidomimetic corresponds to. A classification scheme that is not dependent on the peptide backbone is necessary to classify both peptides and peptidomimetics. We analysed  $\beta$ -turn conformations geometrically and classified them according to the relationships between the directions of the characteristic bonds 1, 2, 3 and 4 (see Figure 1a).

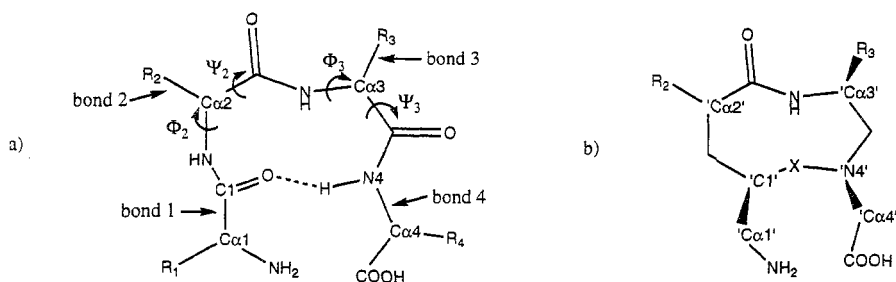


Figure 1. a)  $\beta$ -turn definition.

b)  $\beta$ -turn mimetic structure ( $X = \text{CH}_2$ ,  $\text{CONH}$  or  $\text{CONHC}(\text{CH}_3)_2\text{CH}_2$ ;  $R_2 = \text{H}$ ).

## CONFORMATIONAL CALCULATIONS

The tetrapeptide (Ac-Ala-Ala-Ala-NHMe) and the investigated  $\beta$ -turn-mimetics were built within MacroModel<sup>2</sup> and minimised with the Amber\* force field using the GB/SA solvation model for water. To study the conformational preferences a systematic Monte Carlo search (SPMC) was performed. Only conformations with a  $d_\alpha$  of less than 7Å were kept for the tetrapeptide. All conventional  $\beta$ -turn types<sup>1,3,4</sup> were identified in the resulting conformations, except turn type VI (since cis-proline was not included in the tetrapeptide). The low energy conformations of the  $\beta$ -turn mimetics were investigated by cluster analysis<sup>5</sup> in order to identify conformational families.

## GEOMETRICAL ANALYSIS

The following pseudo torsion angles were found to be interesting to characterise a  $\beta$ -turn (see Figure 1):  $x_2$  (C1-C $\alpha$ 2-C $\alpha$ 3-C $\beta$ 2),  $x_3$  (C $\alpha$ 2-C $\alpha$ 3-N4-C $\beta$ 3),  $\beta$  (C1-C $\alpha$ 2-C $\alpha$ 3-N4)<sup>6</sup> and  $\epsilon$  (C $\alpha$ 1-C1-N4-C $\alpha$ 4). Based on  $x_2$ ,  $x_3$  and  $\beta$  24 different classes were defined.  $\epsilon$  was considered less important because of its correlation to  $\beta$  ( $\beta - 60^\circ < \epsilon < \beta + 60^\circ$ ). Most conformations corresponding to one  $\beta$ -turn type according to the conventional classification scheme were found in only one of the newly defined classes. In addition to  $x_2$ ,  $x_3$  and  $\beta$  also the distances  $d_1$  (C1 - C $\alpha$ 2),  $d_2$  (C $\alpha$ 2 - C $\alpha$ 3),  $d_3$  (C $\alpha$ 3 - N4),  $d_4$  (N4 - C1),  $d_\alpha$  and the torsion angle  $\epsilon$  were considered to estimate the mimetic potential of  $\beta$ -turn mimetics.

## RESULTS AND DISCUSSION

Based on the proposed classification scheme and the measurement of key atomic distances it is possible to estimate which  $\beta$ -turn a certain molecule might mimic. We have studied three different  $\beta$ -turn mimetics<sup>7</sup> (see Figure 1b) which have the possibility to include different sidechains for residues 2 and 3, respectively. The distances were approximately in the range found for the peptide, but for some conformations a non-peptide like angle combination was discerned. Nevertheless, conformations, that could mimic a  $\beta$ -turn, were found for all three compounds. The distances and torsion angles may also be used for a data base search in order to find new  $\beta$ -turn mimetics.

## REFERENCES

1. P.N. Lewis, F.A. Momany, and H.A. Scheraga, Chain reversals in proteins, *Biochim.Biophys.Acta* 303:211 (1973).
2. F. Mohamadi, N.G.J. Richards, W.C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, and W.C. Still, Macro Model - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics, *J.Comp.Chem.* 11:440 (1990).
3. C.M. Venkatachalam, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, *Biopolymers* 6:1425 (1968).
4. C.M. Wilmont, and J.M. Thornton, Analysis and prediction of the different types of  $\beta$ -turn in proteins, *J.Mol.Biol* 203:221 (1988).
5. SYBYL: Molecular Modeling Software, Tripos Associates, Inc. St.Louis, MO063144, 1996.
6. J.B. Ball, R.A. Hughes, P.F. Alewood, and P.R. Andrews,  $\beta$ -Turn topography, *Tetrahedron* 49:3467 (1993).
7. (a) G.L. Olson, M.E. Voss, D.E. Hill, M. Kahn, V.S. Madison, and C.M. Cook, Design and Synthesis of a Protein  $\beta$ -turn mimetic, *J.Am.Chem.Soc.* 112:323 (1990); (b) T. Su, H. Nakanishi, L. Xue, B. Chen, S. Tuladhar, M.E. Johnson, and M. Kahn, Nonpeptide  $\beta$ -turn mimetics of enkephalin, *Bioorg.Med.Chem.Lett.* 3:835 (1993); (c) B. Gardner, H. Nakanishi and M. Kahn, Conformationally constrained nonpeptide  $\beta$ -turn mimetics of enkephalin, *Tetrahedron* 49:3433 (1993).

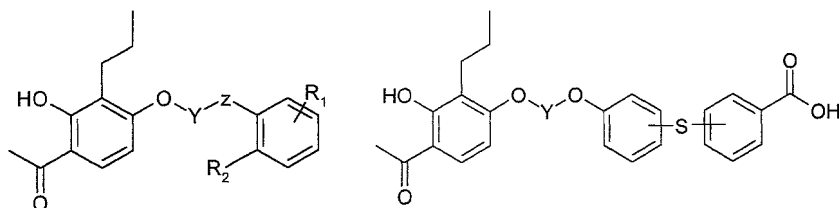
# CONFORMATIONAL ASPECTS OF THE INTERACTION OF NEW 2,4-DIHYDROXYACETOPHENONE DERIVATIVES WITH LEUKOTRIENE RECEPTORS

Miroslav Kuchař, Antonín Jandera, Vojtěch Kmoníček, Bohumila Brůnová,  
Bohdan Schneider

Research Institute for Pharmacy and Biochemistry, Kouřimská 17, 130 60 Praha 3, Czech Republic

## INTRODUCTION

The compounds inhibiting the leukotrienes (LTs) biosynthesis and / or antagonizing their biological functions can be utilized in the antiasthmatic therapy<sup>1,2</sup>. We synthesized<sup>3</sup> the series of 2,4-dihydroxyacetophenone derivatives **1** and **2** bearing carboxyl and their antileukotrienic activities have been determined. The distances *D* between carboxyl and lipophilic part of molecule were calculated for energetically optimized conformations using CHEM-X, Windows 95 programme. The initial geometry with all trans torsions in connecting chain between aromatic nuclei was confirmed in a solid state by the X-ray analysis of two selected derivatives **1**. Lipophilicity of compounds was measured by the use of partitioning tlc and log *P* were calculated from log *P* - *R*<sub>M</sub> relationships. The analysis of quantitative relationships was performed using Statgraphic Programme vers. 4.2.



**1** Y: (CH<sub>2</sub>)<sub>n</sub>, CH<sub>2</sub>CH=CHCH<sub>2</sub>  
Z: O, S  
R<sub>1</sub>: COOH, CH<sub>2</sub>COOH  
R<sub>2</sub>: H, Cl

**2** Y: (CH<sub>2</sub>)<sub>n</sub>

## RESULTS AND DISCUSSION

The following regression equations were calculated for compounds **1**:



LTD<sub>4</sub> receptor binding:

$$\text{Log}(1/IC_{50}) = 3.733D - 0.155 D^2 - 16.790$$

$$n = 13, r = 0.804, s = 0.370, F_{0.005} = 0.173$$

$$\text{Log}(1/IC_{50}) = 4.540 D - 0.193 D^2 + 3.717 \log P - 0.382 (\log P)^2 - 29.93$$

$$n = 13, r = 0.880, s = 0.330, F_{0.01} = 6.880, \log P_{\text{opt}} = 4.87, D_{\text{opt}} = 11.76$$

LTB<sub>4</sub> receptor binding:

$$\text{Log}(1/IC_{50}) = 3.155 D - 0.128 D^2 - 14.398$$

$$n = 11, r = 0.766, s = 0.497, F_{0.03} = 5.623$$

$$\text{Log}(1/IC_{50}) = 5.709 D - 0.239 D^2 + 13.39 \log P - 1.272 (\log P)^2 - 63.50$$

$$n = 11, r = 0.930, s = 0.329, F_{0.09} = 9.533, \log P_{\text{opt}} = 5.26, D_{\text{opt}} = 11.94$$

Inhibition of LTB<sub>4</sub> biosynthesis:

$$\text{Log}(1/IC_{50}) = 2.557D - 0.110 D^2 - 8.313$$

$$n = 11, r = 0.849, s = 0.188, F_{0.006} = 10.340$$

$$\text{Log}(1/IC_{50}) = 3.249 D - 0.141 D^2 + 3.714 \log P - 0.344 (\log P)^2 - 22.07$$

$$n = 11, r = 0.926, s = 0.155, F_{0.01} = 9.020, \log P_{\text{opt}} = 5.40, D_{\text{opt}} = 11.52$$

For compounds **2** the following equations were derived:

LTD<sub>4</sub> receptor binding:

$$\text{Log}(1/IC_{50}) = 6.124 D - 0.214 D^2 - 38.449$$

$$n = 8, r = 0.974, s = 0.134, F_{0.0006} = 46.50, D_{\text{opt}} = 14.31$$

LTB<sub>4</sub> receptor binding:

$$\text{Log}(1/IC_{50}) = 5.991 D - 0.214 D^2 - 36.655$$

$$n = 8, r = 0.897, s = 0.228, F_{0.017} = 10.30, D_{\text{opt}} = 14.00$$

Inhibition of LTB<sub>4</sub> biosynthesis:

$$\text{Log}(1/IC_{50}) = 2.077 D - 0.062 D^2 - 11.132$$

$$n = 8, r = 0.994, s = 0.070, F_{<0.0001} = 226.2, D_{\text{opt}} = 16.75$$

It can be stated from the QSAR analysis that antileukotrienic activities of compounds **1** and **2** depend approximately parabolically on distances *D* between structural elements assumed<sup>4</sup> for their interactions with LT receptors. The nonlinear relationships of antileukotrienic activities to lipophilicity in the series of compounds **1** also occur. In a mentioned series of compounds, the similar optimal values of *D* and *log P* respectively, were found for all antileukotrienic activities under study. The presence of double bond in flexible spacer influences the affinity to LTB<sub>4</sub> receptors and the inhibition of LTB<sub>4</sub> biosynthesis by unexpected manner. The additional aromatic nucleus in compound **2** led to the elevation of optimal values *D*<sub>opt</sub> accompanied with the decrease of corresponding antileukotrienic activities.

## CONCLUSIONS

In the series of compounds **1**, the derivative **1e** (Y=(CH<sub>2</sub>)<sub>3</sub>, Z=S, R<sup>1</sup>=H, R<sup>2</sup>=3-COOH) could be considered as a compound with the integrated antileukotrienic activities. In contrary, the series **2** does not offer such a compound, in accord with the differences among the regression relationships. The influence of distance *D* on antileukotrienic activities is more complex and the constrained and unconstrained energies of compound overlapping over leukotriene must be taken into account.

## ACKNOWLEDGEMENT

This work was financially supported by the Grant Agency of Czech Republic (grant No. 203/97/0212).

## REFERENCES

1. A.W. Ford-Hutchinson, R. Young, and J. Gillard: Leukotriene blockers, novel therapeutic strategies for the treatment of asthma. *Drug News Persp.* 4:264 (1991).
2. C.D.W. Brooks and J.B. Summers: Modulators of leukotriene biosynthesis and receptor activation. *J. Med. Chem.* 39:2629 (1996).
3. M. Kuchar, K. Čulíková, V. Panajotova, B. Brůnová, A. Jandera, and V. Kmoníček: 2,4-Dihydroxyacetophenone derivatives as antileukotrienes with multiple mechanism of action. *Coll. Czech. Chem. Commun.* 63:103 (1998).
4. R.W. Harper, D.K. Herron, N.G. Bollinger, R.F. Baldwin, and J.H. Fleisch: Development of a series of phenyltetrazole leukotriene D<sub>4</sub> receptor antagonists. *J. Med. Chem.* 35: 1191 (1992).

# CONFORMATIONAL STUDIES OF POLY(METHYLIDENE MALONATE 2.1.2)

Eric VANGREVELINGHE<sup>a</sup>, Pascal BRETON<sup>b</sup>, Nicole BRU<sup>b</sup>, Luc MORIN-ALLORY<sup>a</sup>

<sup>a</sup>I.C.O.A. Institut de Chimie Organique et Analytique, CNRS URA 499, Université d'Orléans, BP6759, F-45067 ORLÉANS. Tel. 0238417042 - Fax. 0238417281.

<sup>b</sup>VIRSOL, 46 rue Boissière 75016 PARIS, France. Tel. 0144347414 - Fax. 0144347411.

<http://www.univ-orleans.fr/SCIENCES/ICOA>

## Introduction

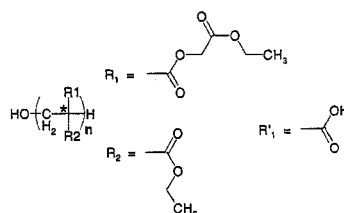
Therapeutic and/or prophylactic efficiency of biologically active molecules depends of the control of drug delivery and targeting. Drug targeting may be performed using biodegradable nanoparticles for delivering therapeutic agents like liposomes or soluble macromolecular carriers. Methylidene malonate 2.1.2, i.e. ethyl-2-ethoxycarbonylmethylenoxycarbonyl acrylate (MM 2.1.2) have recently been studied in order to improve their ability to polymerize and to allow nanoparticles formation in an aqueous media<sup>1</sup>. Erosion and enzymatic bioerosion were suggested to occur at the nanoparticle surface, mainly because of the ester hydrolysis which generates free carboxyl groups and leads to soluble polymers. However this erosion process and its influence on nanoparticle are not well defined. In order to understand, at the molecular level, the influence of various parameters on their conformations, a molecular modeling study was started on Poly(methylidene malonate 2.1.2) (PMM 2.1.2) and derivatives.

## Method

A considerable effort has been made in the past years to devise methods to simulate the behavior of polymer chains in solution<sup>2</sup>. Among various computational methods, molecular dynamics has proven to be a valuable tool for understanding the mechanism and for the evaluation of several time-dependent processes in polymeric systems. Following a preliminary work using an implicit solvation model (GB/SA method), we used here an explicit approach (periodic boundary conditions, NTV ensemble) in order to highlight the influence of the solvent polarity which can simulate polar (aqueous), or non-polar mediums (membranes). All the structures are studied in the molecular form.

## Protocol

Oligomers with 5, 10 and 20 units were studied ( $n = 5, 10$  or  $20$ ). For each one of them, we carried out the study of the native form and the completely eroded form. The eroded forms are obtained by hydrolysis of the longest side chain of the monomer ( $R_1$  to  $R'_1$ ), these eroded forms representing the first stage of the degradation of these compounds<sup>1</sup>. We also studied the influence of the tacticity with the isotactic (all the side chains with the same orientation) and the syndiotactic (regularly alternating) forms.



All the calculations of molecular dynamics were carried out with the following protocol : 3,1 ps heating and 100 ps equilibration-production at 310°K with a 1 fs timestep. During the equilibration-production step, we sampled conformations every picoseconds (ps). Using the integrated programming language SPL of Sybyl, we carried out calculations for various descriptors on all these conformations.

## Results

To define the behavior of these molecules, we have calculated several descriptors. Some of these descriptors allow comparing polymers having same degrees of polymerization but having different degrees of erosion (radius of gyration of the chain, end-to-end distance). Others are specific to a given polymer (radius of gyration of the heavy atoms, surfaces, and volume).

For the 5 units polymers, there are no notable differences between the isotactic and syndiotactic forms. Furthermore, the polarity of solvent does not have a great influence on molecular conformations. That means that these conformations are induced mainly by intramolecular interactions and not by the interactions with the solvent. The limited length of the polymer chain can explain this phenomena.

For the 10 units polymers, it appears an obvious difference between the isotactic and the syndiotactic forms. The averaged values for the radius of gyration of the main chain and the end

Rg polymer chain	N-isotactic	E-isotactic	N-syndiotactic	E-syndiotactic
Water	6,8	6,5	6,6	7,2
Chloroform	6,8	6,2	6,8	7,2

End to end distance	N-isotactic	E-isotactic	N-syndiotactic	E-syndiotactic
Water	20,0	18,9	19,5	22,1
Chloroform	20,8	17,9	20,7	22,9

to end distance during the last 50 ps of the simulation are given in the table. For the native polymer, either isotactic or syndiotactic, the values are similar. They are slightly more important in chloroform than in water, the proof of a conformation slightly lengthen in an apolar solvent. For the eroded polymer the situation is more complex. The isotactic form is compacted, in comparison with the native form, and the value is bigger in water than in chloroform. But, for the syndiotactic form, these values are much more important and they are the proof of an important lengthening of the structure. Furthermore the difference for the two solvents is very small. All these results are good indices of a very important role of the tacticity for the eroded polymers. The absolute configuration of the carbon atoms of the chain has a great influence on the overall conformation.

## Conclusion

Conformational study by molecular dynamics of the PMM 2.1.2. with explicit solvation permits to highlight some characteristics of these compounds. We showed the conformational insensitivity of the native PMM 2.1.2. forms compared to the solvent polarity. On the other hand, we highlighted the importance of the configuration for the completely eroded forms. Complementary studies are currently running for better defining these phenomena.

- 
- 1- Lescure, F., Seguin, C., Breton, P., Bourrinet, P., Roy, D., Couvreur, P., Preparation and characterization of novel poly(methylidene malonate 2.1.2.)-made nanoparticles. *Pharm. Res*, 11 (1994) 1270-1277.
  - 2- Lee, S. J., Park, K., Polymer-solvent interactions studied with computational chemistry *ACS Symp. Ser.*, 545(Polymeric Drugs and Drug Administration), (1994) 221-33.

## A PEPTIDIC BINDING SITE MODEL FOR PDE 4 INHIBITORS

E.E. Polymeropoulos and N. Höfgen

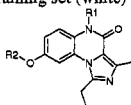
Department of Chemical Research, Corporate R&D, ASTA Medica Group, Weismüllerstr. 45, D-60314 Frankfurt, Germany

### INTRODUCTION

Selective inhibitors of the isoenzyme phosphodiesterase 4 (PDE 4) have attracted increased interest in the last few years as potential drugs for the treatment of allergic diseases such as asthma<sup>1,2</sup>.

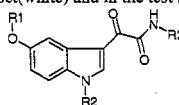
The pharmacophore requirements for inhibiting catalytic activity have been recently analyzed<sup>3</sup>. To further refine this pharmacophore model and define a peptidic model for PDE 4 inhibitors that has the ability to semi-qualitatively predict inhibitory activity we have used the programme PrGen<sup>4</sup>. The structures included in the training set were LAS-31025 (LAS,  $IC_{50}$ =5800 nM<sup>8</sup>), PDA-641 (PDA,  $IC_{50}$ =50 nM<sup>9</sup>), RP-73,401 (RPR,  $IC_{50}$ =0.27 nM<sup>10</sup>), CDP-840 (CDP,  $IC_{50}$ =49 nM<sup>10</sup>), GW3600 (GW3,  $IC_{50}$ =1 nM<sup>11</sup>), Ro 20-1724 (RO1,  $IC_{50}$ =1590 nM<sup>10</sup>), Napp (NAP,  $IC_{50}$ =160 nM<sup>12</sup>), SB 207499 (SBA,  $IC_{50}$ =92 nM<sup>1</sup>), Pfizer (PFA,  $IC_{50}$ =81 nM<sup>13</sup>) as well as the structures shown in Tables 1, 2 (white part). The test set was composed of the structures in Tables 1, 2 (grey part) as well as rolipram (ROL,  $IC_{50}$ =175 nM<sup>10</sup>).

**Table 1.** Structures and *in vivo* activity of pyrido pyrazinones used in the training set (white) and in the test set (grey).



Code name	R1	R2	$IC_{50}$ (nM)
A65	2-F-benzyl	methyl	915
A69	propyl	cyclopentyl	1568
A18	cyclobutyl methyl	methyl	110
D37	H	methyl	1745
A80	isobutyl	methyl	90.5
A85	isobutyl	H	820
D98	ethyl	methyl	414
D70	methyl	methyl	2480
D88	propyl	methyl	232

**Table 2.** Structures and *in vitro* activity of 5-oxyindoles used in the training set (white) and in the test set (grey).



Code name	R1	R3	$IC_{50}$ (nM)
A71	H	4-pyridyl	5060
A8A	methyl	3,5-di-Cl-4-pyridyl	1170
A98	H	3,5-di-Cl-4-pyridyl	17.9
A04	H	2,6-di-Cl-phenyl	92.7
A05	methyl	3,5-di-Cl-4-pyridyl	921
A11	methyl	3,5-di-Cl-4-pyridyl	466
A13	H	3,5-di-Cl-4-pyridyl	6.38
D52	H	4-pyridyl	364
A81	H	3,5-di-Cl-4-pyridyl	6.9
A97	methyl	3,5-di-Cl-4-pyridyl	1010
A07	H	3,5-di-Cl-4-pyridyl	105
A09	methyl	3,5-di-Cl-4-pyridyl	1266

## RESULTS AND DISCUSSION

All structures were fully optimized in their minimum conformation<sup>7</sup> and were aligned to a first approximation by superimposing GRID-contours<sup>3</sup>. The PrGen method was employed as described by Zbinden et al<sup>4</sup>, with the exception of using IC<sub>50</sub> values instead of K<sub>D</sub>'s.

The equilibrated receptor model ( $r_{\text{corr}}=0.994$ ) reveals three hydrogen bond donor binding sites represented by Trp, interacting with alkoxy oxygen (rolipram analogues, pyrido pyrazinones) or nitrogen (LAS) atoms, Tyr, interacting with imidazo nitrogen (pyrido pyrazinones) atoms, and His, interacting with carbonyl oxygen (LAS, RPR, GW3, pyrido pyrazinones, oxyindoles) atoms. PDE 4 requires a divalent zinc ion for catalysis<sup>14</sup>. We have used a histidine coordinated Zn<sup>+2</sup> cation which interacts with carboxy groups (SBA, PFA), hydroxy (NAP) or carbonyl (PDA, RO1) oxygen or pyridyl nitrogen (RPR, CDP, oxyindoles) atoms. The second Trp residue has been added to accommodate lipophilic interactions. The "shape" and "boundaries" of the receptor pocket has been modelled by means of a Van der Waals envelope of charged virtual particles as described in ref. 4.

The resulting pharmacophore hypothesis includes the three potential receptor hydrogen bond donor sites and the lipophilic center described before<sup>3</sup>. In addition, there exists a specific interaction with the Zn<sup>+2</sup> cofactor which appears to enhance the inhibitory activity of ligands. This is exemplified in particular by compounds RPR, CDP, SBA, PFA and the oxyindoles, and to a lesser extent by NAP, PDA and RO1. Furthermore, the hydrogen bond donor binding site represented by Tyr is probably specific to pyrido pyrazinones, and may not in general be necessary for PDE 4 inhibition. The correlation between experimental and calculated free energies of binding for the test set shows that with the exception of A09, A97 and D98 the model is capable of predicting these values to within 0.8 kcal/mol.

## REFERENCES

1. M.N. Palfreyman, Phosphodiesterase type IV inhibitors as antiinflammatory agents, *Drugs Future* 20:793 (1995).
2. J.A. Karlsson, D. Aldous, Phosphodiesterase 4 inhibitors for the treatment of asthma, *Exp. Opin. Ther. Patents* 7:989 (1997).
3. E.E. Polymeropoulos, N. Höfgen, A pharmacophore model for PDE IV inhibitors, *Quant. Struct.-Act. Relat.* 16:231 (1997).
4. P. Zbinden, M. Dobler, G. Folkers, A. Vedani, PrGen: Pseudoreceptor modeling using receptor-mediated ligand alignment and pharmacophore equilibration, *Quant. Struct.-Act. Relat.* 17:122 (1998).
5. N. Höfgen et al, EP 736532.
6. N. Höfgen et al, DE 198 18 964.8 (reg. nr.)
7. AM1; MOPAC v.6.0, QCPE, No. 455, 1989.
8. Third Int. Conf. on Cyclic Nucleotide Phosphodiesterases: From Genes to Therapies, Glasgow.
9. D. Cavalla, R. Frith, Phosphodiesterase IV inhibitors: Structural diversity and therapeutic potential in asthma, *Curr. Med. Chem.* 2:561 (1995).
10. Inhibition of PDE IV in human PMNL. Assay established at ASTA/AWD.
11. J.A. Stafford et al, Introduction of a conformational switching element on a pyrrolidine ring. Synthesis and evaluation of (R,R)-(+/-)-methyl 3-acetyl-4-(3-cyclopentoxy) 4-methoxyphenyl)-3-methyl 1-pyrrolidine carboxylate, a potent and selective inhibitor of cAMP-specific phosphodiesterase, *J. Med. Chem.* 38:4972 (1995).
12. D. Cavalla et al, XIVth International Symposium on Medicinal Chemistry, Maastricht 1996, P-5.11.
13. A.J. Duplantier et al, Biarylcarboxylic acids and -amides: Inhibition of phosphodiesterase type IV versus [<sup>3</sup>H]rolipram binding activity and their relationship to emetic behavior in the ferret, *J. Med. Chem.* 39:120 (1996).
14. S.H. Francis, J.L. Colbran, L.M. McAllister-Lucas, J.D. Corbin, Zinc interactions and conserved motifs of the cGMP-binding cGMP-specific phosphodiesterase suggest that it is a zinc hydrolase, *J. Biol. Chem.* 269:22477 (1994).

# MOLECULAR DYNAMICS SIMULATIONS OF THE BINDING OF A GnRH AGONIST TO A MODEL GnRH RECEPTOR

A.M. ter Laak<sup>1</sup>, R. Kühne<sup>1</sup>, G. Krause<sup>1</sup>, E.E. Polymeropoulos<sup>2</sup>,  
B. Kutscher<sup>2</sup>, E. Günther<sup>2</sup>

<sup>1</sup>Forschungsinstitut für Molekulare Pharmakologie, 10315 Berlin, Germany  
<sup>2</sup>Dept. Medicinal Chemistry, Corporate R&D, ASTA Medica Group,  
60314 Frankfurt-M., Germany

## INTRODUCTION

Gonadotropin-releasing hormone (GnRH) is the naturally occurring agonist for the G-protein-coupled GnRH-receptor. GnRH stimulates the pituitary gland to produce luteinizing hormone (LH) and follicle stimulating hormone (FSH). Both GnRH agonists and antagonists are potentially useful in the treatment of hormone dependent ailments.

The aim of the present modeling study is the generation of a 3D-model for the binding of GnRH agonists to the GnRH receptor in accordance with the available mutation data on the GnRH receptor and on other homologous receptors. For this purpose, the agonist triptorelin is docked into a GnRH receptor model using a molecular dynamics protocol with carefully designed range distance restraint functions. At a second stage in the same molecular dynamics run, the possible conformational changes within the receptor after agonist binding are investigated by simulating hypothetical conformational changes of selected conserved amino acid side chains within the GnRH receptor.

## METHODS

We used an alpha-carbon template for the transmembrane helices in the family A of GPCRs (Baldwin, 1997) as a basis for a model of the GnRH receptor. Starting structures for the intra- and extracellular loops were obtained with the loop search method in SYBL6.4 taking into account the two extracellular disulfide bridges (C14-C200 and C114-C196, Davidson, 1997). To obtain more reliable low energy structures for the intra- and extracellular loop regions of the GnRH receptor we performed a simulated annealing (SA) method using the AMBER 4.1 FF in which the TM domains are restrained and the whole protein is solvated with explicit water molecules.

The MD run presented here consists of two phases: in the first phase (0-80 ps) the GnRH analogue triptorelin is docked into the agonist binding site and in the second phase (80-200 ps) the GnRH receptor rearranges to a hypothetical active conformation. During the whole MD, helical hydrogen bonds were restrained and the center of masses of proline residues

within the transmembrane domains were given position restraints to keep the seven helix bundle intact. The MD simulation is performed *in vacuo* using the AMBER 4.1 FF.

Triptorelin is biased to orient itself in the GnRH receptor pocket by applying carefully designed restraint functions based on experimental information (Flanagan, 1994, Davidson, 1996, Zhou, 1995). Like GnRH, the peptide hormones bombesin and endothelin also begin with pyroGlu1, and, remarkably their receptors also contain an R or K in TM III at the same or nearly the same position as K121 in the GnRH receptor. We therefore assume in our simulation that pyroGlu1 interacts with K121.

In the same molecular dynamics run, we simulated a hypothetical rearrangement of the GnRH receptor interior going from an inactive state to an active receptor state during ligand binding. For this purpose, the following residues were restrained at a close distance: Y323 to R139 and D319 to W280 in the inactive state, and Y323 to D319 and D319 to C279 in the active state. N53 and N87 were kept together in both receptor states.

## RESULTS

From the simulations of triptorelin binding and subsequent GnRH receptor activation it can be concluded that the proposed interactions pyroGlu1<sup>+</sup>K121, Arg8<sup>+</sup>D302 and (Gly10-NH2)<sup>+</sup>N102 are supported by our GnRH receptor model. The model supports the assumption that residue K121, observed in the GnRH receptor, is analogous to R in the bombesin and K (at position i+1) in the endothelin receptor, and that these strong H-donor residues interact with pyroGlu1 which is a common residue in the ligands for these receptors.

A new interaction between T215 in the GnRH receptor and His2 of triptorelin was observed in our binding model. This residue may also play a role in GnRH receptor activation since the T215 position agrees with important residues for receptor activation in aminergic receptors (Wang, 1991, Gantz, 1992).

It is possible to simulate the conformational change of a receptor going from an hypothetical inactive to an active state without losing essential structural properties of the receptor template. The main effects observed in the current activation model are small rotations of TM domains VI and VII and proline kinking in TM domains V and VII. The orientation of TM III in the middle of the seven helix bundle remains largely unaffected upon receptor activation.

## REFERENCES

- Baldwin, J.M., Schertler, G.F., and Unger, V.M., 1997, An alpha- carbon template for the transmembrane helices in the rhodopsin family of G- protein- coupled receptors, *J. Mol. Biol.* 272: 144
- Davidson, J.S., Assefa, D., Pawson, A., Davies, P., Hapgood, J., Becker, I., Flanagan, C., Roeske, R., and Millar, R., 1997, Irreversible activation of the gonadotropin- releasing hormone receptor by photo-affinity cross linking: localization of attachment site to Cys residue in N- terminal segment, *Biochemistry* 36: 12881
- Davidson, J.S., McArdle, C.A., Davies, P., Elario, R., Flanagan, C.A., and Millar, R.P., 1996, Asn102 of the gonadotropin- releasing hormone receptor is a critical determinant of potency for agonists containing C- terminal glycinamide, *J. Biol. Chem.* 271: 15510
- Flanagan, C.A., Becker, I.I., Davidson, J.S., Wakefield, I.K., Zhou, W., Sealton, S.C., and Millar, R.P., 1994, Glutamate 301 of the mouse gonadotropin- releasing hormone receptor confers specificity for arginine 8 of mammalian gonadotropin releasing hormone, *J. Biol. Chem.* 269: 22636
- Gantz, I., DelValle, J., Wang, L.D., Tashiro, T., Munzert, G., and Guo, Y., 1992, Molecular basis for the histamine interaction of histamine with the H2 receptor, *J. Biol. Chem.* 267: 20840
- Oliveira, L., Paiva, A.C., Sander, C., and Vriend, G., 1994, A common step for signal transduction in G- protein- coupled receptors, *TIPS*, 15: 170
- Wang, C.D., Buck, M.A., and Fraser, C.M., 1991, Site- directed mutagenesis of alpha 2A- adrenergic receptors: identification of amino acids involved in ligand binding and receptor activation by agonists, *Mol. Pharmacol.* 40: 168
- Zhou, W., Rodic, V., Kitanovic, S., Flanagan, C.A., Chi, L., Weinstein, H., Maayani, S., Millar, R.P., and Sealton, S.C., 1995, A locus of the gonadotropin- releasing hormone receptor differentiates agonist and antagonist binding sites, *J. Biol. Chem.* 270: 18853



# ANALYSIS OF AFFINITIES OF PENICILLINS FOR A CLASS C $\beta$ -LACTAMASE BY MOLECULAR DYNAMICS SIMULATIONS

Keiichi Tsuchida,<sup>1</sup> Noriyuki Yamaotsu,<sup>2</sup> and Shuichi Hirono<sup>2</sup>

<sup>1</sup>Research Laboratories, Toyama Chemical Co., Ltd.  
4-1 Shimookui 2-chome, Toyama-shi, Toyama 930-8508, Japan  
<sup>2</sup>School of Pharmaceutical Sciences, Kitasato University  
5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

## INTRODUCTION

$\beta$ -Lactamases are widespread bacterial enzymes that effectively cleave and inactivate the  $\beta$ -lactam families of antibiotics by catalyzing the hydrolysis of the amide group of the  $\beta$ -lactam ring via a serine-bound acyl intermediate. Therefore, it is important to elucidate the catalytic mechanism of the  $\beta$ -lactamases in order to design antibiotics with improved activity against  $\beta$ -lactamase-producing bacteria.

Although the structure of the class C  $\beta$ -lactamase has been established by X-ray crystallography, the crystal structure of the complex of the  $\beta$ -lactamase with penicillin has still not been carried out. It is therefore important that the complex structure be built by molecular modeling and studied in detail for the purpose of the design of novel antibiotics and  $\beta$ -lactamase inhibitors.

The inhibition constants ( $K_i$ ) of foramidocillin (FOPC) and piperacillin (PIPC) with the class C  $\beta$ -lactamase from *Enterobacter cloacae* have been measured to be 10.8  $\mu$ M and 1.0  $\mu$ M, respectively. We have built structures of the  $\beta$ -lactamase-FOPC (FOPC complex) and  $\beta$ -lactamase-PIPC complexes (PIPC complex) based on molecular modeling developed from the molecular dynamics (MD) simulation.

We have analyzed the binding affinities of the two ligands, FOPC and PIPC, for the  $\beta$ -lactamase on the structure of the FOPC complex obtained from molecular modeling.<sup>1</sup> To calculate the relative binding affinity, we made use of the thermodynamic integration (TI) method, which involves changing the molecular mechanical parameters during the MD simulation and determining the free energy change for the process. The difference in the binding free energy for two ligands to the enzyme is  $\Delta\Delta G_{\text{bind}} = \Delta G_2 - \Delta G_1$ .  $\Delta G_1$  (FOPC) and  $\Delta G_2$  (PIPC) were determined from the  $K_i$  values and are -7.04 kcal/mol and -8.51 kcal/mol, respectively. As the free energy is a state function, the results of the computer simulations can be related to  $\Delta\Delta G_{\text{bind}} = \Delta G_{\text{int}} - \Delta G_{\text{solv}}$ .  $\Delta G_{\text{solv}}$  is the change in free energy upon mutating FOPC to PIPC in water.  $\Delta G_{\text{int}}$  is the change in free energy for the same mutation with the ligand bound to the enzyme.  $\Delta G_{\text{solv}}$  and  $\Delta G_{\text{int}}$  were calculated by the TI method and the difference,  $\Delta\Delta G_{\text{bind}}$ , was compared with the experimental value of  $\Delta G_2 - \Delta G_1$ .

## METHODS

The starting structures for the simulations involving the enzyme were built using SYBYL v6.1. FOPC was docked into the binding site of the  $\beta$ -lactamase from *Enterobacter cloacae* strain P99 (the Brookhaven Protein Data Bank ref. 2BLT<sup>2</sup>) by reference to the hydrogen-bond information for  $\beta$ -lactam binding sites.<sup>2</sup> The complex obtained by docking was optimized. The PIPC complex was generated substituting the 6 $\alpha$ -NHCHO group and the two OH groups on the phenyl ring with hydrogens, followed by minimization.

The MD/TI calculations were performed with AMBER v4.1. To the starting structures was added a spherical cap of water molecules with a radius of 35 Å at the center of binding site. The systems were equilibrated for 160 ps. In the solution simulations, FOPC was solvated by water molecules in a box, and was equilibrated for 60 ps.

The TI calculations were performed using the window-growth procedure, and the free energies were averaged from the forward ( $\lambda = 1 \rightarrow 0$ ) and backward ( $\lambda = 0 \rightarrow 1$ ) directions. For both the solution and enzyme simulations, the total perturbation time was 202 ps.

## RESULTS AND DISCUSSION

We obtained the model structures for the FOPC and PIPC complexes as averaged structures calculated from the trajectory during the last 30 ps. The binding mode of both of the penicillins seemed to be similar to that of the cephalosporin used as a reference.<sup>2</sup>

We calculated the interaction energies of FOPC and PIPC with their surroundings in the  $\beta$ -lactamase. Since the total interaction energy of FOPC with the enzyme is - 262 kcal/mol, whereas that of PIPC with the enzyme is - 258 kcal/mol, we can conclude that FOPC interacts more favorably with the enzyme. Therefore, the binding free energies of FOPC and PIPC can't be explained by the interaction energies of FOPC and PIPC with the  $\beta$ -lactamase.

The average of the free energy changes (FOPC  $\rightarrow$  PIPC) in solution was 15.7 kcal/mol. It is therefore clear that the desolvation of FOPC is much more difficult than that of PIPC. The average of the free energy changes in the enzyme system was 13.5 kcal/mol. This result shows that FOPC interacts more strongly with the enzyme than PIPC, which agrees with the analysis of the interaction energy.

The binding free energy change between the FOPC complex and the PIPC complex was - 2.2 kcal/mol. The calculated value of the binding free energy change shows that the binding affinity of PIPC is greater than that of FOPC, and is in good agreement with the experimental value of - 1.5 kcal/mol.

The results indicate that the binding affinity of FOPC is lower than that of PIPC because of the greater difficulty of desolvation of FOPC upon binding to the enzyme. An understanding of both interaction energies and the solvation and desolvation of ligands is critical if the relative binding affinities of ligands and proteins are to be described.

Our simulations by molecular modeling and MD/TI methods predicted the structures of the Michaelis complexes of the  $\beta$ -lactamase with both FOPC and PIPC. Using the structures of the complexes built by molecular modeling without the X-ray crystal structures of the complexes, we were able to reproduce the experimental difference in the free energy of binding.

## REFERENCES

1. *Drug Design and Discovery* (1998), submitted.
2. E. Lobkovsky, P.C. Moews, H. Liu, H. Zhao, and J.-M. Frere, Evolution of an enzyme activity: Crystallographic structure at 2-Å resolution of cephalosporinase from the *ampC* gene of *Enterobacter cloacae* P99 and comparison with a class A penicillinase, *Proc. Natl. Acad. Sci. USA*, 90:11257 (1993).

# THEORETICAL APPROACHES FOR RATIONAL DESIGN OF PROTEINS

Jiří Damborský

Laboratory of Structure and  
Dynamics of Biomolecules  
Masaryk University  
Kotlarska 2, 611 37 Brno  
Czech Republic

## INTRODUCTION

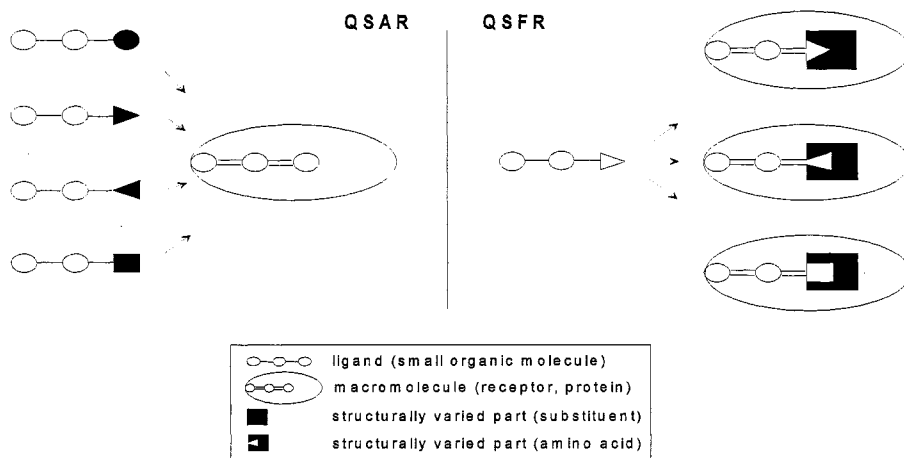
Protein engineering is the field of study involving the creation and modification of proteins (Cleland and Craik, 1996). It has great potential to provide significant advances in science, medicine, and industry. The successful engineering of a protein of interest requires design of protein mutants, their production and evaluation. Rational design of the protein mutants is preferably based on an available 3D structure. However, even with the knowledge of the tertiary structure it can still be very difficult to propose which structural modification of the protein will lead to the desired change in its properties (Atkins and Sligar, 1991). Theoretical approaches can be used in the systematic analysis of structure-function relationships and can assist in the design process.

Two novel theoretical approaches applicable for the rational design of protein variants are discussed in this contribution. Quantitative Structure-Function Relationships (Damborský, 1997) is the statistical approach for systematic analysis of the data from site-directed mutagenesis experiments and prediction of properties of the protein mutants. This analysis can be used in cases where the 3D structure of the protein under investigation is not known. The second approach, called 'computational site-directed mutagenesis' (Damborský et al., 1998) is the molecular modelling procedure suitable for investigation of the catalytic properties of the protein mutants. These mutants are 'constructed' and evaluated using computational chemistry tools. A 3D structure of the protein under study has to be available.

## QUANTITATIVE STRUCTURE-FUNCTION RELATIONSHIPS

Quantitative Structure-Function Relationships (QSFR) investigate and mathematically describes the effect of changes in structure of the protein on its catalytic activity. Trends in molecular properties of the amino acids which are varied, are related to protein activities by means of statistical analysis. Systematic changes in the protein structure, i.e. a number of

point substitutions at a certain position, are required for the statistical analysis. QSFR closely resembles the well known QSAR (Quantitative Structure-Activity Relationships). Figure 1 schematically shows the basic principles of both analyses. Changes in molecular structure are quantitatively expressed using physico-chemical or other molecular properties. Developed QSFR models can be used for the interpretation of data from site-directed mutagenesis experiments and for design of mutants with required properties.



**Figure 1.** Schematic representation of QSAR and QSFR analyses. QSAR is mainly concerned about the activities of small organic molecules (ligands), while QSFR explores the function of macromolecules (enzymes, receptors). Reproduced with permission from Damborský J, *Protein Engineering* 11: 21-30, 1998.

QSFR analysis was applied to a set of 16 mutants in position 172 of haloalkane dehalogenase and a set of 19 mutants in position 222 of subtilisin (Damborský, 1998). The activity data measured for the protein mutants were derived from the literature (Schanstra et al., 1996; Estell et al., 1985). A total of 402 molecular descriptors obtained from AAindex database (Nakai et al., 1988) were used to code the amino acid properties. The multivariate statistical method - partial least squares projection to latent structures, PLS (Höskuldsson, 1988) - was used to correlate descriptors with protein activities. Developed PLS models explained 82% of data variance (77% cross-validated) for haloalkane dehalogenase mutants and 86% of the data variance (81% cross-validated) for subtilisin mutants. Hydrophobic, steric as well as electronic properties of the substituted amino acid were important for the description of mutant activity. Current analysis of the single-point mutants can be extended to analysis of multiple mutants.

## COMPUTATIONAL SITE-DIRECTED MUTAGENESIS

Computational site-directed mutagenesis is a theoretical technique in which a large number of protein variants are constructed and their properties evaluated using computer modelling. Initially an exhaustive set of substitutions is created using the 3D structure of the wild type protein. Calculation of binding energies and/or mapping of reaction coordinates is used for estimation of protein properties - the binding affinities and kinetic properties, respectively. These calculations need to be both reasonably accurate and considerably fast which means that a careful selection of the size of the system (only reacting residues, the active site, or the complete protein) and the computational technique

to be applied (molecular mechanics, semi-empirical quantum-chemical, or *ab-initio* quantum-chemical) has to be made. Only the protein mutants which show the desired properties in the calculation are subsequently experimentally prepared and tested.

Computational site-directed mutagenesis was employed to mutate residue 172 in the haloalkane dehalogenase (Damborský et al., 1998). An exhaustive set of single-point mutants in this position was constructed by homology modelling. The X-ray structure of Verschuere and co-workers (Verschuere et al., 1993) was used as the input structure. Reaction-pathways were mapped with microscopic models of the active sites for each of the mutant. A semi-empirical quantum chemical method was employed in this study (Damborský et al., 1997) and several theoretical parameters (energies, point charges) were extracted for calculation of and comparison with the experimental activities reported by Schanstra and co-workers (Schanstra et al., 1996). Some of these parameters were significantly correlated with the experimental data making it possible to distinguish active mutants from inactives based on these calculations. The whole modelling procedure, including systematic construction of the protein mutants, preparation of the input files for quantum-chemical calculation, mapping of the reaction pathway and data extraction is currently being automated in the program Triton ([www.chemi.muni.cz/lbsd/triton.html](http://www.chemi.muni.cz/lbsd/triton.html)).

## REFERENCES

- Atkins, W.M. and Sligar, S.G., 1991, Protein engineering for studying enzyme catalytic mechanism, *Curr. Opin. Struct. Biol.* 1:611.
- Cleland, J.L. and Craik, C.S., eds., 1996. *Protein Engineering: Principles and Practise*, New York, John Wiley.
- Damborský, J., 1997, Quantitative structure-function relationships of the single-point mutants of haloalkane dehalogenase: A Multivariate approach, *Quant. Struct.-Act. Relat.* 16:126.
- Damborský, J., Kutý, M., Nemeč, M. and Koca, J., 1997, A molecular modeling study of the catalytic mechanism of haloalkane dehalogenase: 1. quantum chemical study of the first reaction step, *J. Chem. Inf. Comp. Sci.* 37:562.
- Damborský, J., 1998, Quantitative structure-function and structure-stability relationships of purposely modified proteins, *Prot. Engng.* 11:21.
- Damborský, J., Bohac, M., Prokop, M., Kutý, M. and Koca, J., 1998, Computational site-directed mutagenesis of haloalkane dehalogenase in position 172, *Prot. Engng.*, in press.
- Estell, D.A., Graycar, T.P. and Wells, J.A., 1985, Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation, *J. Biol. Chem.* 260:6518.
- Höskuldsson, A., 1988, PLS regression methods, *J. Chemometr.* 2: 211.
- Nakai, K., Kidera, A. and Kanehisa, M., 1988, Cluster analysis of amino acid indices for prediction of protein structure and function, *Prot. Engng.* 2:93.
- Schanstra, J.P., Ridder, I.S., Heimeriks, G.J., Rink, R., Poelarends, G.J., Kalk, K.H., Dijkstra, B.W. and Janssen, D.B., 1996, Kinetic characterization and X-ray structure of a mutant of haloalkane dehalogenase with higher catalytic activity and modified substrate range, *Biochemistry* 35:13186.
- Verschuere, K.H.G., Franken, S.M., Rozeboom, H.J., Kalk, K.H. and Dijkstra, B.W., 1993, Refined X-ray structures of haloalkane dehalogenase at pH 6.2 and pH 8.2 and implications for the reaction mechanism, *J. Mol. Biol.* 232:856.

## AMISULPRIDE, SULTOPRIDE AND SULPIRIDE: COMPARISON OF CONFORMATIONAL AND PHYSICO-CHEMICAL PROPERTIES

Audrey Blomme<sup>1</sup>, Laurence Conraux<sup>2</sup>, Philippe Poirier<sup>2</sup>, Anne Olivier<sup>3</sup>, Jean-Jacques Koenig<sup>2</sup>, Mireille Sevrin<sup>3</sup>, François Durant<sup>1</sup> and Pascal George<sup>3</sup>

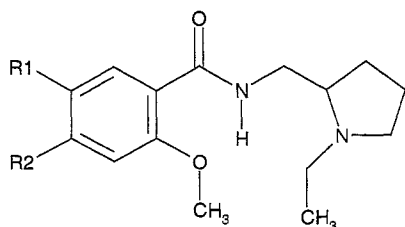
<sup>1</sup> Laboratoire de Chimie Moléculaire Structurale, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium

<sup>2</sup> Synthélabo Recherche, Groupe de Biochimie Moléculaire Structurale, Rueil, France

<sup>3</sup> Synthélabo Recherche, Département de Recherche SNC, Bagneux, France

### INTRODUCTION

Amisulpride, sultopride and sulpiride (Figure 1) are antagonists of the D<sub>2</sub>-like dopamine receptors, which are members of a large family of receptors that interact with specific intracellular signalling pathways through coupling with G proteins. These compounds are substituted benzamides and present a high degree of selectivity for D<sub>2</sub> and D<sub>3</sub> versus D<sub>1</sub> and D<sub>4</sub> dopaminergic receptor subtypes. Amisulpride, sultopride and sulpiride respectively present decreasing *in vitro* affinities for the D<sub>2</sub> receptor (IC<sub>50</sub> = 27, 120 and 181 nM) and the D<sub>3</sub> receptor (IC<sub>50</sub> = 3.6, 4.8 and 17.5 nM).



Compounds	R1	R2
amisulpride	SO <sub>2</sub> C <sub>2</sub> H <sub>5</sub>	NH <sub>2</sub>
sultopride	SO <sub>2</sub> C <sub>2</sub> H <sub>5</sub>	H
sulpiride	SO <sub>2</sub> NH <sub>2</sub>	H

**Figure 1.** Molecular structures of amisulpride, sultopride and sulpiride.

## RESULTS

In the present study, we have compared the conformational and physico-chemical properties of amisulpride, sultopride and sulpiride in order to identify the molecular properties that could explain their *in vitro* binding profile.

Firstly, the conformational space of the S-enantiomers of amisulpride, sultopride and sulpiride was explored by 2D NOESY NMR spectroscopy and molecular mechanics. The resulting conformational families were compared to X-ray structures (Cambridge Structural Database). It was found that the conformational space of the three compounds is quite similar. Therefore it cannot be considered as a relevant property to account for the specific pharmacological profile of amisulpride.

Secondly, we investigated the physico-chemical properties of an optimized common conformation of the drugs. Topology and energy of frontier orbitals (HOMO and LUMO) and molecular electrostatic potential (MEP), were calculated and compared. The topology of the Lowest Unoccupied Molecular Orbital (LUMO) is similar for the three compounds. The Highest Occupied Molecular Orbital (HOMO) of amisulpride is mainly localized on the nitrogen atom of the 4-amino group and on the C<sub>1</sub> carbon atom of the phenyl moiety whilst the HOMO of sultopride and sulpiride is principally localized on the oxygen atom of the 6-methoxy group and on the C<sub>3</sub> carbon atom of the phenyl moiety. The major difference observed between the three compounds is provided by the value of the minimum potential energy, localized on the oxygen atom of the amide function : -67.8, -63.7 and -61.3 kcal/mole for amisulpride, sultopride and sulpiride respectively. The more potent attractive effect of the carbonyl group of amisulpride can be related to the topology of its HOMO and the presence of the 4-amino group on the benzamide moiety, which is conjugated with the amide function.

Moreover, complementary properties, such as pK<sub>a</sub> and logP were measured. The basicity of the nitrogen of the pyrrolidine moiety is characterized by a pK<sub>a</sub> value of around 9 for the three compounds. Amisulpride, sultopride and sulpiride present decreasing experimental values of lipophilicity ( $\log P_{\text{octanol-water}} = 1.6, 1.2 \text{ and } 0.6$  respectively). Switching from a sulfonamide function in sulpiride to an ethylsulfone group in sultopride may be responsible for the greater partition of sultopride in the lipophilic compartment. On the other hand, the total volume of amisulpride is expanded by the presence of the 4-amino group on the benzamide moiety and thus reinforces its hydrophobic character.

## CONCLUSIONS

In the present study, we have shown that the conformational and physico-chemical properties of S-enantiomers of amisulpride, sultopride and sulpiride present some comparable features but that they are not identical. We have identified two factors which could be responsible for the specific pharmacological profile of amisulpride.

- The presence of the 4-amino group on the phenyl moiety of amisulpride could induce a stronger interaction between the oxygen of the carbonyl function and the receptor via hydrogen bonding.
- The pharmacological specificity of amisulpride could be reinforced by the presence of an ethylsulfone group which allows additional interactions with an hydrophobic pocket of the receptor.

As observed for all physico-chemical properties, sultopride behaves like an intermediate compound between amisulpride and sulpiride : this could explain the relative affinity level of each molecule.

## ENTROPIC TRAPPING: ITS POSSIBLE ROLE IN BIOCHEMICAL SYSTEMS

Adolf Miklavc<sup>1</sup> and Darko Kocjan<sup>2</sup>

<sup>1</sup>National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

<sup>2</sup>Lek-Pharmaceutical and Chemical Works, Ljubljana, Slovenia

### INTRODUCTION

Entropy-driven binding which is characterized by  $|\Delta H^\circ| \sim 0$  and  $\Delta S^\circ > 0$  has so far been found in a number of important biochemical systems but explaining the mechanism of it has remained a challenge. An analysis of the experimental results on binding of  $\beta$ -AR antagonists<sup>1</sup>, and on binding in several other systems<sup>2</sup> led to the conclusion that a mechanism must exist, besides, e.g., large scale conformational changes or hydrophobic interactions. More recent experimental work<sup>3</sup> strengthened the above conclusion, by revealing that entropy-driven binding can occur also when hydrophobic interactions are absent. A novel mechanism, entropic trapping, was therefore proposed<sup>1</sup>, consistent with the experimental findings. In computer simulations the existence of the entropic trapping binding mechanism was established<sup>4</sup>. The difference in the entropy increase in binding of simple anesthetics to membrane proteins<sup>3</sup> is interpreted as an example.

### BINDING BY ENTROPIC TRAPPING

Entropy-driven binding characteristically takes place in a hydrophobic, sterically constrained environment, e.g., in a hydrophobic channel or cleft. It can be assumed therefore that the thermodynamic binding constant  $k_D$  then reflects the local equilibrium between the ligand in the binding pocket and in the sterically constrained neighbourhood of it. The assumption of internal nature of the binding constant is consistent with the fact that the temperature effects due to desolvation processes have not been observed<sup>1,2,5</sup>. The binding data of, say,  $\beta$ -AR ligands may be rationalized by assuming that upon reaching the binding pocket deep in the transmembrane channel by a diffusion process, a  $\beta$ -agonist forms a tight 'normal' bond ( $\Delta H^\circ < 0$ ,  $\Delta S^\circ < 0$ ), but a  $\beta$ -antagonist cannot form a tight bond ( $|\Delta H^\circ| \sim 0$ ) because of the structural properties. Due to the looseness of its bond in the binding pocket the phase space of rotations/internal rotations 'opens up', leading to  $\Delta S_r^\circ > 0$ . This entropy increase drives the binding in the systems in question. The changing of the  $\Delta H^\circ$  and  $\Delta S^\circ$  along the diffusion path reaching the binding pocket may be qualitatively presented as in Fig. 1. The existence of the entropic trapping mechanism has been established in the computer experiments<sup>4</sup> on the diffusion of polymers in random environments.



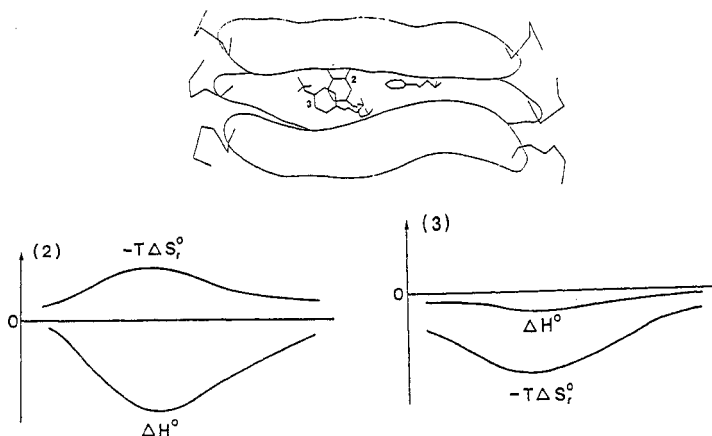


Figure 1. Schematic presentation of  $\Delta H^\circ$  and  $\Delta S^\circ$  along a transmembrane channel in enthalpy-driven (2) and entropy-driven (3) complex formation.

Preliminary qualitative studies show that the structural dependence of the observed  $\Delta S^\circ$  is consistent with the above mechanism. In the case of entropy-driven binding of the anesthetics halothane and propofol (Fig. 2) to two  $\text{Ca}^{2+}$ -ATPases, integral membrane

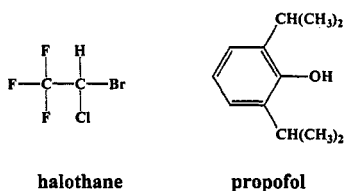


Figure 2. Anesthetics used in the studies of binding to  $\text{Ca}^{2+}$ -ATPases<sup>3</sup>.

proteins PMCA and SERCA1 the present model reproduces quantitatively the differences in the binding entropies. Experimentally it was found<sup>4</sup> for the two ligands dissolved in  $\text{Me}_2\text{SO}$  at  $25^\circ\text{C}$ :  $\Delta\Delta G^\circ \cong T\Delta S^\circ(\text{propofol}) - T\Delta S^\circ(\text{halothane}) = -1.7$  kcal/mol in PMCA, and  $-2.2$  kcal/mol in SERCA1. The entropic trapping model yields, assuming that  $\Delta\Delta S^\circ$  arises primarily from the internal rotations of the two  $-\text{CH}(\text{CH}_3)_2$  groups of propofol:  $\Delta\Delta G^\circ \cong T\Delta\Delta S_i^\circ = 2.1$  kcal/mol. A typical value of 3.5 e.u. was assumed here for one internal rotation.

## REFERENCES

1. A. Miklavc, D. Kocjan, J. Mavri, J. Koller, and D. Hadži, On the fundamental difference in the thermodynamics of agonists and antagonists interactions with  $\beta$ -adrenergic receptors and the mechanism of entropy-driven binding, *Biochem. Pharmacol.* 40:663 (1990).
2. A. Miklavc, Temperature-nearly-independent binding constant in several biochemical systems. The underlying entropy-driven binding mechanism and its practical significance, *Biochem. Pharmacol.* 51:723 (1996).
3. M.M. Lopez and K. Kosk-Kosicka, Entropy-driven interactions of anesthetics with membrane proteins, *Biochemistry* 36:8864 (1997).
4. G.W. Slater and Y.S. Wu, Reptation, entropic-trapping, percolation and rouse dynamics of polymers in "random" environments, *Phys. Rev. Lett.* 75:164 (1995).
5. P.A. Borea, K. Varani, S. Gessi, P. Gilli, and G. Gilli, Binding Thermodynamics at the human neuronal nicotine receptor, *Biochem. Pharmacol.* 55:1189 (1998).

## STRUCTURAL REQUIREMENTS TO OBTAIN POTENT CAXX MIMIC P21-RAS-FARNESYLTRANSFERASE INHIBITORS

Abdelazize Laoui

Medicinal Chemistry Department, Molecular Modelling  
 Rhone-Poulenc Rorer S. A. - Centre de Recherches de Vitry-Alfortville,  
 13, Quai Jules Guesde - B. P. 14 - 94403 Vitry-sur-Seine, France

### INTRODUCTION

Farnesyltransferase (FTase) farnesylates p21ras on the Cys residue of the C-terminal consensus sequence referred to as a CAAX box (where C is cysteine, A is an aliphatic amino acid and X is any amino acid). This modification is required for membrane association and function of both normal and cell transforming ras activity. Transformed ras proteins are implicated in a number of human cancers including colon, pancreatic and lung carcinomas. Therefore selective inhibition of FTase could lead to a new class of potent and specific anticancer agents.

This paper presents in the first section the computer modelling studies of corporate and competitor FTase inhibitors which led to the identification of the structural requirements necessary to obtain potent inhibitors. In the second section we report on the strategy adopted to replace the oxidisable thiol function of our in-house inhibitors with alternative Zinc chelating groups. This should hopefully lead to compounds with improved cellular activity.

The peptidomimetic strategy has allowed the development of a series of inhibitors derived from a known peptidic inhibitor CVFM where the Valine and Phenylalanine were replaced by a naphthyl scaffold which forces an extended conformation (1).

### RESULTS AND DISCUSSION

The FTase bound conformation of a competitive p21 pseudopeptide inhibitor (2), L739787 (NH<sub>2</sub>-C-[YCH<sub>2</sub>NH][YCH<sub>2</sub>O]F-M-CH<sub>2</sub>OH) allowed us to better define the central portion of the FTase substrate binding pocket using the Naphthyl series.

Hx (x=1-3) : Hydrophobic sites

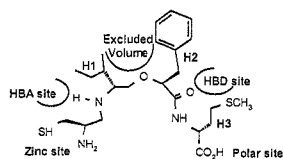


Figure 1. L-739750 template FTase Bound Conformation

Table 1. SAR of Naphtyl series

Naphtyl series	IC <sub>50</sub> (nM)
1,6 series	1.8
1,5 series	5.8
2,6 series	14.0
1,4 series	23.5
1,3 series	31.0
2,4 series	975.0
2,5 series	5400.0
3,5 series	13000.0

Analysis of the Naphthyl series SAR (see table 1) revealed an excluded volume between H1 and H2. In all the active products the Naphthyl scaffold partially occupies the H2 pocket or fits to the peptidic main chain between H1 and H2. In the inactive products the Naphthyl scaffold occupies a position

between the H1 and H2 sites. This observation is important for determining the optimal position to add an extra hydrophobic group on the Naphthyl scaffold in order to better fill the H2 pocket. The template FTase bound conformation allows us to position the appropriate pharmacophores of RPR and competitor series in the correct spatial orientations. Earlier structural comparisons suggest that increased binding energy, specificity and hopefully bioavailability could be gained by increasing the size and hydrophobicity to fully occupy the central portion of the FTase substrate binding pocket. Several in-house and competitor compounds (3,4) have been used to generate a 5 point pharmacophore model with Catalyst. This model contains many of the features of Figure 1 and is shown below, Figure 2. These and similar models have been used to analyse a number of potential new series.

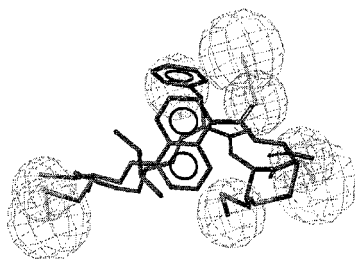
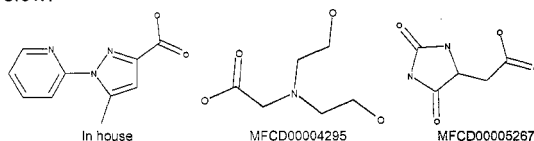


Figure 2. superposition of Naphtyl 1,5 series and the Merck pseudo-peptide inhibitor. The different spheres represent polar and hydrophobic interaction sites.

The bioavailability of the compounds may be improved by replacing the cysteine thiol group by other zinc chelating groups. To assist the chemists in the choice of reagents we have performed searches in several databases for potential complexing groups, taken from an analysis of the in-house Zn ligand database containing 919 Zinc binding groups extracted from the CSD and PDB. In the first step, ligands were sorted into monodentate, bidentate etc. and each list sorted by molecular weight. ISIS 2D searching was used to eliminate heavier groups containing the same core functionality interacting with the zinc. These unique cores were then used to search for acid, acid chloride and aldehyde precursors in commercial and corporate databases, with the results being loaded into a local ISIS database for visualisation by chemists. A number of these have been selected and synthesis and testing of these compounds is underway. Representative reagents selected from these lists are shown below.



## CONCLUSION

We have presented in this paper various strategies which have been or are currently being used in the design of p21ras CAAX mimics. The peptidomimetic approach led to initial lead series. These were optimised in chemistry and pharmacophore mapping strategies have enabled the generation of several models which help to understand the SAR in these series. Such models have also been used in the design of several potential new scaffolds. In-house derived databases of ligand-cation interactions have proved a valuable source of ideas for designing zinc chelating groups which mimic the cysteine residue. It is hoped that this approach will lead to compounds with a better bioavailability.

## REFERENCES

- (1) F.-F. Clerc, J.-D. Guitton, N. Fromage, Y. Lelièvre, M. Duchesne, B. Tocqué, E. James-Surcouf, A. Commerçon and J. Becquart, *Bioorg. Med. Chem. Lett.* 1995, **5**, 1779-1784
- (2) K. Koblan et al. *Prot. Sci.* 1995, **4**, 681-688
- (3) B. Baudoin, C. Burns, A. Commerçon, J.-D. Guitton, WO 95/34535 (12/21/1995); B. Baudoin, C. Burns, A. Commerçon, A. Lebrun, WO 96/22278 (07/25/1996)
- (4) A. Vogt, Y. Qian, M.A. Blaskovich, R.D. Fossum, A.D. Hamilton and S.M. Sebt, *J. Biol. Chem.* 1995, **270**, 660-664

## HYDROGEN-BONDING HOTSPOTS AS AN AID FOR SITE-DIRECTED DRUG DESIGN

James E.J. Mills and Philip M. Dean

Department of Pharmacology  
University of Cambridge  
Tennis Court Road  
Cambridge, UK, CB2 1QJ

### INTRODUCTION

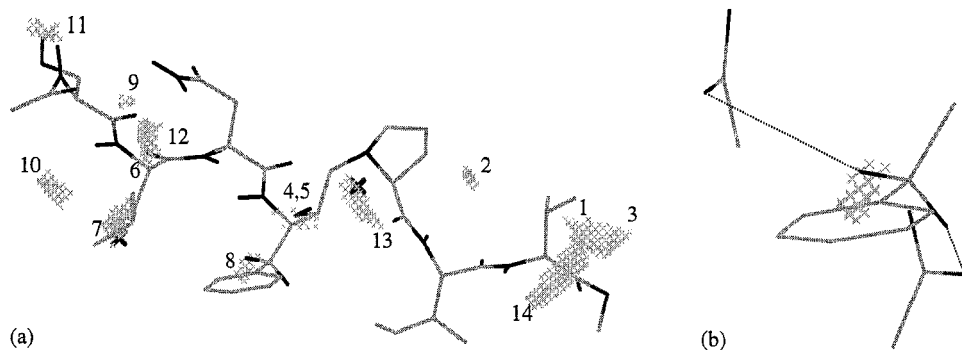
De novo drug design utilises site H-bonding atoms as anchor points to be spanned by a ligand. A problem frequently encountered is the need to select a subset of these points for use by the ligand. For example, there are 44 H-bonding regions in the HIV-protease binding cavity so if 5 points were required, there would be over  $10^6$  possible selections. H-bonding hotspots are positions of ligand atoms that could simultaneously form H-bonds to more than one atom on the receptor. There are fewer hotspots than site atoms, so they could provide a solution to the combinatorial selection problem. A method for hotspot calculation (HOTSPOT) is presented and tested with HIV-protease and a wide variety of other PDB complexes.

### METHODS

The binding cavity is calculated using an implementation of the SURFNET algorithm<sup>1</sup>. For each H-bonding group accessible from this cavity, CSD data<sup>2</sup> are used to plot all possible complementary ligand-atom positions onto a 0.2 Å grid. Gridpoints arising from more than one group are defined as hotspots, provided they satisfy angle criteria determined by a crystal survey of the CSD. These criteria determine whether an atom positioned there could orient its two H-bond valencies in the directions of the receptor atoms. Hotspots where a subset of regions from another hotspot overlap are removed.

### RESULTS

HOTSPOT was run on the HIV-protease PDB<sup>3</sup> complex 9hvp, generating 14 hotspots. The sites of 21 other HIV-protease complexes were superposed onto that of 9hvp, allowing the hotspots to be compared with the positions of heteroatoms of the 22 ligands (Figure 1). Of the 14 hotspots, 4 are on the periphery of the cavity and can be discarded. The remaining hotspots are all occupied by either a ligand or water heteroatom in at least one of the 22 complexes. Interestingly, hotspot 8 is occupied by a methylene carbon atom in 19 complexes.



**Figure 1.** (a) Hotspots calculated for HIV protease complex 9hvp shown with inhibitor JG-365 (from 7hvp). (b) Close-up of hotspot 8, showing how methylene carbon atom could form H-bonds to Gly A 27 O and Asp B 25 OD2 of HIV-protease. Water atoms are shown as large crosses.

HOTSPOT was run on a wide variety of PDB complexes and the hotspots compare very favourably with ligand and water heteroatom positions (Table 1). On average, the sites only make 33% of possible H-bonds but utilise 50% of possible hotspots. Furthermore, 50% of H-bonds made by ligands involve hotspots, showing their potential importance in drug design.

**Table 1.** Number of ligand and water heteroatoms predicted to within 1 Å by HOTSPOT for 10 complexes. Close = number of hotspots close to either ligand or water atom

ID	PDB complex Protein/Ligand	No. site H-bonds		No. hotspots		No. predicted	
		Possible	Made	Total	Close	Ligand	Water
121p	H-ras p21/5'-[ $\beta,\gamma$ -Me] GTP	37	14	27	17	5	5
1abe	L-arabinose-BP/L-arabinose	9	8	12	6	4	
1adl	Adipocyte lipid BP/Arachidonate	26	2	10	4	1	3
1azm	Carbonic anhydrase I/ATS	8	4	5	3	2	
1brn	Barnase/D-(CGAC)	50	15	33	18	9	6
1bvc	Apomyoglobin/Biliverdin	20	4	6	5	3	2
1dhi	DHFR/Methotrexate	34	5	11	7	1	6
1fkb	FK506 binding protein/Rapamycin	21	5	6	2	2	
1tpp	$\beta$ -trypsin/PAPP	28	7	24	14	3	5
3gst	Glutathione S-transferase/GHD	33	10	15	6	4	1

In conclusion, H-bonding hotspots firstly reduce the number of sitepoints for selection, and secondly provide stronger anchor points for ligands than single hydrogen bonds. They therefore provide a means for reducing the complexity of site-directed drug design.

## REFERENCES

1. J.E.J. Mills and P.M. Dean, Three-dimensional hydrogen-bond geometry and probability information from a crystal survey, *J. Comput.-Aided Mol. Design* 10:607 (1996).
2. R.A. Laskowski, SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions, *J. Mol. Graph.* 13:323 (1995).
3. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The protein databank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.* 112:535 (1977).

# **SUPERPOSITION OF FLEXIBLE LIGANDS TO PREDICT POSITIONS OF RECEPTOR HYDROGEN-BONDING ATOMS**

James E.J. Mills and Philip M. Dean

Department of Pharmacology  
University of Cambridge  
Tennis Court Road  
Cambridge, UK, CB2 1QJ

## **INTRODUCTION**

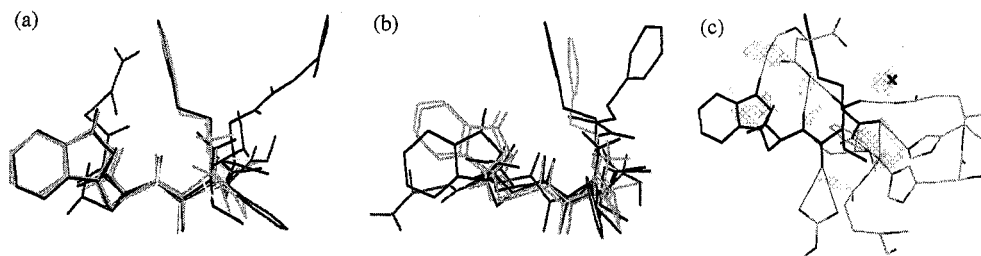
When the structure of a binding site is unknown, information is derived from the ligands known to bind there, which requires accurate ligand superposition, determined by the correct binding conformations. A novel program, SLATE, superposes ligands using a single point to represent each H-bonding group. The method is rapid enough to allow both of the ligands to flex during the superposition. SLATE is tested on thermolysin and  $\alpha_2$ -adrenoceptor ligands.

## **METHODS**

H-bond-donor groups are represented by the optimum position for the complementary receptor atom, projecting the X-H bond to optimum H-bond distance (determined by crystal-survey data<sup>1</sup>). H-bond-acceptor groups are represented by the H-bond acceptor atom because the donor group on the receptor is assumed to be immobile, projecting to the same position for each ligand. Optionally, each aromatic ring is represented by a vector perpendicular to the ring and passing through its centroid. The points are superposed by minimising the sum of the elements of the difference distance matrix with simulated annealing,<sup>2</sup> allowing changes in conformation of each ligand, selection of points for superposition and correspondence between points. MATFIT<sup>3</sup> is used to carry out the superposition. Multiple runs are ranked according to their H-bond (number of overlapping points and the rms between them) and steric (degree of overlap of surface volumes calculated by PLM<sup>4</sup>) properties, the best match having the lowest sum of the ranks. The overlapping H-bond regions<sup>5</sup> of the superposed molecules represent the positions of receptor atoms to which more than one ligand could bind.

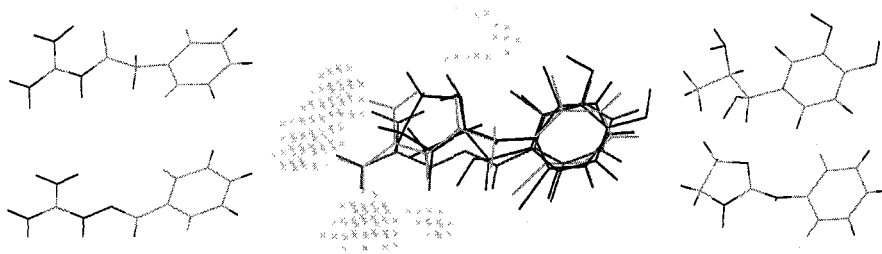
## **RESULTS**

SLATE was used to flex 5 thermolysin ligands (PDB files 1tmn, 2tmn, 3tmn, 5tln and 6tmn) onto the crystal conformation of CCT (1thl), giving a more compact superposition than the crystal superposition. 5 receptor atoms were predicted to within 1 Å by the overlapping H-bond regions (Figure 1).



**Figure 1.** (a) SLATE and (b) crystal superpositions obtained for 6 thermolysin ligands. (c) Overlapping H-bond regions compared with binding-site atoms of thermolysin (CCT is shown in black, making its H-bonds).

SLATE was used to superpose the  $\alpha_2$ -agonists clonidine, guanfacine, guanabenz and  $\alpha$ -methyl noradrenaline. The molecules were superposed pairwise, allowing each to flex freely during the superposition. Only one conformation of guanfacine was found to produce a good match with all the other ligands, so these results were used to generate the superposition shown in Figure 2. Clusters of overlapping H-bonding points defining the possible positions of 5 receptor atoms were generated.



**Figure 2.** Centre shows SLATE superposition obtained for the  $\alpha_2$ -adrenoceptor agonists (from top left, clockwise) guanfacine,  $\alpha$ -methyl noradrenaline, clonidine and guanabenz.

In conclusion, SLATE has been validated as a means for superposing ligands that bind predominantly via H bonds and as such provides a new tool for ligand-based drug design.

## ACKNOWLEDGEMENTS

JEJM is a Rhône-Poulenc Rorer Research Associate and PMD a Wellcome Principal Research Fellow.

## REFERENCES

1. J.E.J. Mills and P.M. Dean, Three-dimensional hydrogen-bond geometry and probability information from a crystal survey, *J. Comput.-Aided Mol. Design* 10:607 (1996).
2. M.T. Barakat and P.M. Dean, Molecular structure matching by simulated annealing. I. A comparison between different cooling schedules, *J. Comput.-Aided Mol. Design* 4:295 (1990).
3. A.D. McLachlan, Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.* 128:49 (1979).
4. T.D.J. Perkins, J.E.J. Mills, and P.M. Dean, Molecular surface-volume and property matching to superpose flexible and dissimilar molecules, *J. Comput.-Aided Mol. Design* 9:479 (1995).
5. J.E.J. Mills, T.D.J. Perkins, and P.M. Dean, An automated method for predicting the positions of hydrogen-bonding atoms in binding sites, *J. Comput.-Aided Mol. Design* 11:229 (1997).

## COMPARATIVE MOLECULAR FIELD ANALYSIS OF MULTIDRUG RESISTANCE MODIFIERS

Ilza K. Pajeva<sup>1</sup> and Michael Wiese<sup>2</sup>

<sup>1</sup>Center of Biomedical Engineering, Bulg. Acad. Sci., BG-1113 Sofia, Bulgaria

<sup>2</sup>Department of Pharmacy, University of Halle, D-06120 Halle, Germany

### INTRODUCTION

Pharmacological modulation of multidrug resistance, MDR, in tumor cells relates to the application of drugs able to block the function of the membrane-integrated P-glycoprotein, P-gp. P-gp is suggested to transport the antitumor agent out of the cells by an ATP-dependent efflux, decreasing in this way its intracellular concentration and the cytotoxic effect. In general, the MDR modulators are considered to interact with the same binding sites as the antitumor agents. Different binding sites as well more than one interaction site are suggested in order to explain the extraordinarily structural variety of the P-gp substrates (antitumor agents) and inhibitors (MDR modulators) [1,2]. The absence of information about the binding site(s) requires identification of common space determinants of structurally different MDR reversing compounds and that is the purpose of the study.

### DATA AND METHODS

The data used are: 21 phenothiazines, 16 thioxanthenes, 2 imipramines, 1 acridine, 22 propafenones and 6 benzofurans. MDR reversal activity in doxorubicin resistant human carcinoma cell line MCF-7/DOX was used for phenothiazines and related drugs [3,4]. MTT assay of daunomycin cytotoxicity and inhibition of rhodamine-123 efflux in vincristine resistant T-lymphoblast cell line were used for the propafenone-type MDR modulators [5,6]. Different sets of compounds were extracted from the data for training and test purposes.

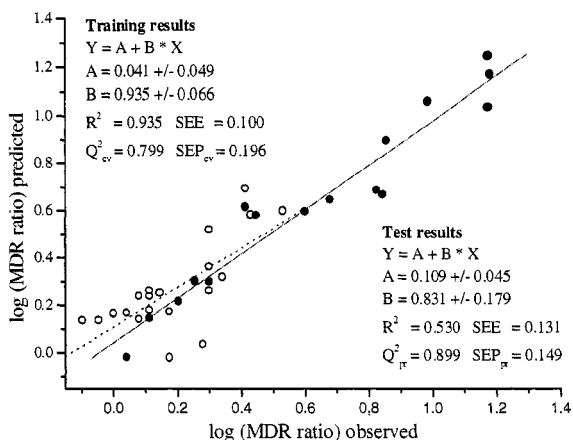
Combined activity data of propafenone-type modulators were calculated by PCA. Molecular modeling was done with SYBYL 6.3 using molecular mechanics (Tripos force field) and quantum chemistry (MOPAC AM1 and PM3). Hydrophobic fields used in CoMFA were calculated with HINT V.2.11.

### RESULTS AND DISCUSSION

The starting conformations of the most active representatives in their classes were taken or built from x-ray structures available in the Cambridge crystallographic database. After the geometries were optimized and the charges calculated the molecules were aligned according to two main criteria: skeleton similarity and shape similarity. The skeleton similar-



ity considers superpositioning according to the largest common substructure of the compounds and the shape similarity considers the pharmacophore concept introduced before [7, 8] about the role of the two aromatic rings and the basic N in the aliphatic chain for the MDR reversal by the compounds studied. It was done by fitting the centroids of the aromatic rings and the N atom using both, rms and field fit alignment techniques. A preliminary investigation of the influence of the CoMFA field type, field fit cutoff and threshold column filtering on the statistical parameters of the models was done and the most relevant to the CoMFA settings established. Each chemical class was trained and tested individually and finally those of them with MDR reversing activity in the same cell line were combined into integrated models. Several hundred CoMFA models were obtained and compared. With exception of the H-bond fields all fields resulted in statistically significant models, most of which were as well highly predictive (Fig.1). The inclusion of hydrophobic fields significantly improved the CoMFA models obtained and they alone and in combination with the steric and both (steric and electrostatic) fields yielded models with the highest cross-validated  $R^2$  and predictivity for the separate test sets.



**Fig. 1** Predicted versus observed activity values of the shape aligned training (16 thioxanthenes: •, —) and test set (21 phenothiazines: ○, ---) using the three-component model with steric and hydrophobic only fields.

The results outline the role of hydrophobicity as a structural characteristic of importance for the activity studied. As no predictive correlations were obtained with the logP values, they point to hydrophobicity as a space directed molecular property for explaining the differences in MDR modulating activity of the investigated compounds.

**Acknowledgment.** The authors express their thanks to Deutsche Forschungsgemeinschaft and Bulgarian Science Fund for the financial support of the presented work.

## REFERENCES

1. Safa A. In: *Multidrug resistance in cancer cells* (eds: S Gupta and T Tsuruo), John Wiley & Sons Ltd, 1996, 231-250.
2. Dey S, Ramachandra M, Pastan I, Gottesman MM, Ambudkar S. *Proc Natl Acad Sci USA* 94, 1997, 10594-10599.
3. Ford JM, Prozialeck WC, Hait WN. *Mol. Pharmacol.* 35, 1989, 105-115.
4. Ford JM, Bruggeman EP, Pastan I, Gottesmann M, Hait WN. *Cancer Res.* 50, 1990, 1748-1756.
5. Ecker G, Chiba P, Hitzler M, Schmid D, Visser K, Cordes H-P, Csöllei J, Seydel JK, Schaper K-J. *J. Med. Chem.* 39, 1996, 4767-4774.
6. Chiba P, Ecker G, Schmid D, Drach B, Tell B, Goldenberg S, Gekeler V. *Mol. Pharmacol.* 49, 1996, 1122-1130.
7. Pajeva IK, Wiese M. *Quant. Struct.-Act. Relat.* 16, 1997, 1-10.
8. Wiese M, Pajeva IK. *Pharmazie* 52, 1997, 679-685.

## PHARMACOPHORE MODEL OF ENDOTHELIN ANTAGONISTS

Mitsuo Takahashi, Kuniya Sakurai, Seji Niwa and Seiji Oono

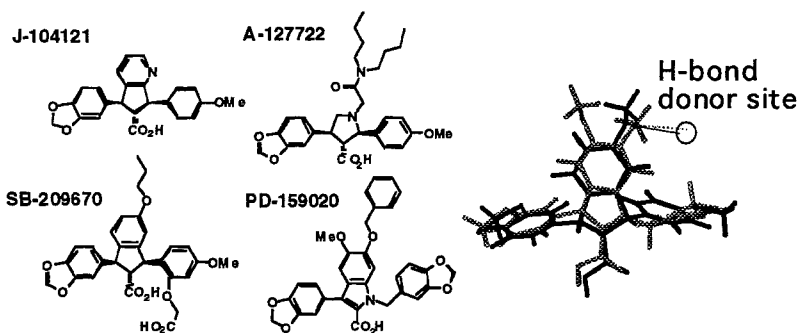
Pharmaceutical Research Laboratories  
Ajinomoto Co., Inc.  
1-1 Suzuki, Kawasaki, Kawasaki, Kanagawa 210, Japan

### Introduction

Endothelin (ET) is a 21-amino acid peptide, and its receptor belongs to a family of G-protein-coupled receptor. ET antagonist is expected to be a therapeutic agent for disease including myocardial infarction, hypertension and restenosis. In an attempt to design new antagonists we have created a pharmacophore model using active antagonists with the consideration of conformational flexibility of molecules in which DISCO was applied. A new hydrogen bond donor site was identified in the model, and was utilized successfully in the design of highly active antagonists.

### Results and Discussion

In modelling pharmacophore the compounds used are selected carefully so that they can have some degree of structural similarity rather than a variety of structure and can be expected to bind to the ET receptor in a common binding mode. Selected antagonists<sup>1,2,3,4</sup> are shown in Fig.1, and they have in common a carboxylic acid attached to a five membered ring in the middle, methoxy and methylenedioxy groups attached to the phenyl rings on both sides.

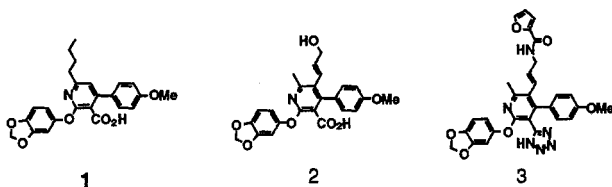


**Figure 1.** Antagonists used in search of pharmacophore model (left and middle), and molecular alignment of SB-209670 (dark) and PD-159020 (gray) in the pharmacophore model (right).

It is rather trivial that these groups are expected to be pharmacophoric features in common. Our main focus is to find a new non-trivial pharmacophoric feature. We have assumed that a hydrogen bond donor site in the receptor side could make a hydrogen bond with an acceptor atom in the upper part of these antagonists. The important part in setup for pharmacophore search is the selection of conformers and candidate features. The numbers of low energy conformers generated by systematic search using Sybyl<sup>5</sup> are 50 for J-104121, SB-209670 and PD-159020, and 100 for A-127722 in which both of alternate sp<sup>3</sup> geometries of nitrogen in pyrrolidine ring are adopted. In the conformation search less important chemical groups were pruned for computational simplicity: for example butyl groups were changed to methyl groups in A-127722 and so on.

In the default setup of features in DISCO there are a lot of candidate features in both ligands and receptor sides. However, we reduced features to focus on our assumption since the number of features is crucial for efficient pharmacophore search. The pharmacophore search using DISCO resulted with 65 candidate models, each of which has a different set of features and molecular alignment. With the inspection of molecular overlap in each model on graphic terminal a plausible model was finally selected, in which the distance tolerance was 2.0 Å and a new hydrogen bond donor site was detected. The molecular alignment of SB-209670 and PD-159020 in this model is shown in Fig. 1.

The existence of this new feature could be verified also in the structure-activity relationship study, undertaken concurrently in the medicinal chemistry approach: the lead compound (1) shown in Fig.3 was modified to the compound (2) satisfying the pharmacophore requirement and having a higher activity. Based on this model further synthetic approach resulted with a highly active antagonist (3) which was synthesized through both the chemical optimization of linker, keeping the hydrogen bond acceptor property, and also the bioisosteric replacement for a carboxylic acid to a tetrazole.



**Figure 2.** Optimization from the lead compound (1) to the compounds (2) and (3): PIC50's for porcine heart ET<sub>A</sub> receptor are 5.9, 7.3 and 8.6, respectively.

## References

1. T. Nagase et al., *AFMC International Medicinal Chemistry Symposium*, 181 (1995).
2. M. Winn, T.W. von Geldern et al., *J. Med. Chem.*, 39:1039 (1996).
3. J.D. Elliott, M.A. Lago et al., *J. Med. Chem.*, 37:1553 (1994).
4. A.M. Dohert, W.C. Patt et al., *J. Med. Chem.*, 38:1259 (1995).
5. Sybyl molecular modelling package, Tripos Associates Inc., St. Louis, MO, USA.

## THE ELECTRON-TOPOLOGICAL METHOD (ETM): ITS FURTHER DEVELOPMENT AND USE IN THE PROBLEMS OF SAR STUDY

Nataly M. Shvets, Anatoly S. Dimoglo

Institute of Chemistry, Academy of Sciences, Kishinev, *MOLDOVA*  
Gebze Institute of Technology, Gebze/Kocaeli, *TURKEY*  
e-mail: [dimoglo@yahoo.com](mailto:dimoglo@yahoo.com)

Among the *distinguished features of the ETM* there are the absence of the dependence on the compounds' structural skeletons, the most detailed (atomic) level of the compound description, mathematical backgrounds underlying the compound description language, high predictive ability, etc. The ETM-system uses as its input data a series to be investigated that includes both active and inactive molecules and is supplied with the corresponding activity values for every molecule. The core algorithm takes as its input the results of conformational analysis and quantum-chemical computations applied to the series selected.

It is well known that the choice of an appropriate CSDL is the primary source of success in any QSAR method. In contrast to other QSAR methods, the CSDL used in the ETM has the following useful properties.

- The notion of molecular structure is given strict formal background.
- The numerical values that are used as elements of the corresponding mathematical structures are exact and theoretically justified.
- Well known mathematical techniques can be used for the CSDL processing, as the result of the said above.

Every molecular structure has its mathematical counterpart, namely, three-dimensional matrix. Each layer of the matrix is an ordinary,  $n \times n$  matrix called electron-topological matrix of contiguity (ETMC). Its triangular form is due to the symmetric nature of the chemical bonds and atomic distances. The values of its elements are defined by a definite atomic characteristic, if they are diagonal elements; for non-diagonal elements a property of chemical bonds (if the two atoms are chemically bonded), otherwise the distance are used.

*The ETM-system* is a menu-based application with the ETM as an item of the main menu. To apply the method, two preliminary steps are to be done, namely, forming the ETMCs and setting parameters that control the active fragments selection.

*The core algorithm* solves the same problem of the pattern recognition, as expert systems do. But it does not belong to the class of software; as soon as the objects of the investigation and their properties have been given a strict formal sense, the feature selection either gets a

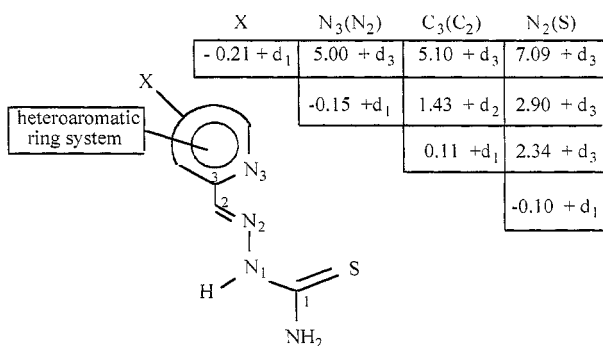
mathematical (not logical) background. To select the molecular features causing the activity presence, we have to solve the task of searching the intersection of two complete graphs. When studying new compounds to predict their activity, we have to solve the task of their isomorphic inclusion. The most important is that one algorithm can manage both tasks plus to filter the matrix elements.

The parameters setting *pre-conditions* for the activity features selection are the values of  $d_1$ ,  $d_2$  and  $d_3$  used for determining the relations of equivalence for atoms, bonds and distances, the threshold value of activity,  $A_{thr}$ , used to divide the compounds into two groups (active/inactive or the most active and the rest) and a target compound. It is the most active one, when searching the features of activity, or inactive, when searching 'the breaks of activity'. The *post-conditions* that take decision on the procedure completion are minimal values of the probabilistic estimations  $P_a$  and  $\alpha_a$  aiming in classifying the fragments obtained into 'good' and 'bad' ones. The procedure can be repeated under different selections of the electronic characteristics, so as the pre- and post-conditions.

*Databases* are important components of the software dealing with the QSAR problems. In contrast to the databases used normally in expert system, the local database (LDB) in the ETM is extendable and opened. Its tables possess different levels of the data access. To reduce the LDB size, the ETMCs for the series studied are not kept in memory. Instead, a special program manages the matrices formation. The services provided for the LDB management are to input, to extract and reformat the data that has arrived from different sources, to process users' queries and to communicate with remote databases.

After processing ETMCs, we get the features of activity that belong to active compounds only. (Analogously, the features of inactivity, or "the breaks of activity", can be found"). The features of activities searched are submatrices of an ETMC of a compound taken as a target for comparison.

As the example of a typical SAP, the task concerning the antitumor activity studied in the series of thiosemicarbazones can be demonstrated. The values of  $d_1$ ,  $d_2$ ,  $d_3$  found are 0.05, 0.10, 0.20, respectively. The level of prediction is 94%.



For the fragment shown at the picture the following rules are to be obeyed: 1) the fragment enters both the heteroaromatic ring and thiosemicarbazone being a part of the molecule. 2) negatively charged atom X is bonded immediately with pyridine or isoquinoline. 3) chemically bonded N and C (0.05 Å distanced), may belong to the heteroaromatic ring or thiosemicarbazone, to which belongs the 4th atom (N or S).

To complete the SAP, some examples of the rules violation are given, causing the activity loss. If needed, the quantitative SAR model can be added, giving more exact activity estimations.

**Acknowledgments.** This study was partially supported by INTAS-Ukraine grant 95-0060.

**Poster Session IV**  
**Computational Aspects of**  
**Molecular Diversity and**  
**Combinatorial Libraries**

## **MOLDIVS - A NEW PROGRAM FOR MOLECULAR SIMILARITY AND DIVERSITY CALCULATIONS**

Vadim A. Gerasimenko, Sergei V. Trepalin, Oleg A. Raevsky

Institute of Physiologically Active Compounds of Russian Academy of Sciences, 142432, Chernogolovka, Moscow region, Russia

At present molecular similarity and diversity calculations are the important tools for lead generation and optimization, especially in the fields of high-throughput screening and combinatorial chemistry.

There are many approaches to this problem, which differ in descriptors used, similarity and diversity measures and compounds selection algorithms.

Descriptors of different types (topological indexes, physical property descriptors, 2D and 3D structural keys) can be used for this purpose. It was shown,<sup>1</sup> that structural 2D descriptors perform better than others in their ability to distinguish between biologically active and inactive compounds. The discriminating power of these descriptors depends on the degree to which they encode information relevant to ligand-receptor binding (hydrophobic, dispersion, electrostatic, steric and hydrogen bonding interactions).<sup>2</sup>

One of the approaches to this problem is to produce composite descriptors from structural and global physical property descriptors by means of principal component analysis and multidimensional scaling.<sup>3</sup> However this method is unable to handle sets of compounds of real sizes (10.000 - 1.000.000 compounds) because of computational limitations of multidimensional scaling required for transforming discrete structural descriptors to continuous variables.

In this report we propose an alternative approach based on combination of structural fragments and local physicochemical property descriptors.

On the basis of this approach the new program MOLDIVS (MOlecular DIversity and Similarity) for Microsoft Windows 95/NT was created. MOLDIVS has friendly graphic user interface and it permits to perform the whole range of similarity and diversity calculation tasks on large sets of compounds.

In this program it is possible to use the structural descriptor of two types: plain structural fragment and combined structural-physicochemical fragments. Both fragments are defined as atom-centered concentric environments.<sup>4</sup> Fragment consists of a central atom and neighboring atoms connected to it within the predefined sphere size (number of bonds between the central and edge atoms). For each fragment the complete connection table is stored. For each atom in a fragment the information on the atom and bond type, charge, valency, cycle type and size is coded into fixed-length variables, which are subsequently

used to define a pseudo-random hash value for this fragment. The complete set of fragments with selected sphere size is created automatically and forms a fragments library. For each fragment in the library the frequency of occurrence is calculated. An unlimited number of fragments and sphere of any size can be used.

In structural-physicochemical fragments each atom is characterized by three parameters: partial atomic charge,<sup>5</sup> polarizability<sup>6</sup> and H-bond donor/acceptor factor<sup>7</sup> instead of atomic element type as in plain structural fragments. Adjustable ranges of these properties are used as atomic types. There are many examples when similarity based on plain structural fragments substantially differs from similarity based on structural-physicochemical fragments. In many cases structural-physicochemical fragments produced better separation of biologically active compounds because they explicitly encode information relevant to ligand-receptor interactions.

The program permits an estimation of similarity of each molecule in the database with all other molecules sorting them on the value of similarity with the initial molecule. It is possible to use different molecular similarity coefficients: Tanimoto, Euclidean and Cosine.

Different measures of diversity of the whole database  $A$  are available in this program:

$$DIVERSITY (A) = \sum_{I, J} (DISSIMILARITY (I, J)) / N^2 \quad (1)$$

$$DIVERSITY (A) = \sum_I (MIN_J (DISSIMILARITY (I, J))) / N \quad (2)$$

The program allows rapid estimation of diversity of the whole database according to equation (1) using the cosine similarity coefficient on the basis of the centroid algorithm.<sup>8</sup>

Different compound selection algorithms for diverse subset formation (stepwise elimination and cluster sampling,<sup>9</sup> number of maximum dissimilarity selection algorithms<sup>10</sup>) are used in this program.

The program was successfully tested on databases with biological and medicinal activity data and in real drug design work. The comparison of results obtained by MOLDIVS and other commercially available programs is carried out.

1. R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, *J. Chem. Inf. Comput. Sci.* 36:572 (1996).
2. R.D. Brown, Y.C. Martin, The informational content of 2D and 3D structural descriptors relevant to ligand-receptor binding, *J. Chem. Inf. Comput. Sci.* 37:1 (1997).
3. E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos, Measuring diversity: experimental design of combinatorial libraries for drug discovery, *J. Med. Chem.* 38:1431 (1995).
4. S.V. Trepalin, A.V. Yarkov, L.M. Dolmatova, N.S. Zefirov, WinDat: an NMR database compilation tool, user interface and spectrum libraries for personal computers, *J. Chem. Inf. Comput. Sci.* 35:405 (1995).
5. D.B. Kireev, V.I. Fetisov, N.S. Zefirov, Approximate molecular electrostatic potential computations: applications to quantitative structure-activity relationships, *J. Mol. Struct. (Theochem)* 304:143 (1994).
6. K.J. Miller, Additivity methods in molecular polarizability, *J. Am. Chem. Soc.* 112:8533 (1990).
7. O.A. Raevsky, Hydrogen bond strength estimation by means of the HYBOT program package, in: *Computer-Assisted Lead Finding and Optimization*, H. van de Waterbeemd, B. Testa, G. Folkers, eds., Wiley-VHC, Basel (1997).
8. J.D. Holliday, S.S. Ranade, P. Willett, A fast algorithm for selecting sets of dissimilar molecules from large chemical databases, *Quant. Struct.-Act. Relat.* 14:501 (1995).
9. R. Taylor, Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals, *J. Chem. Inf. Comput. Sci.* 35:59 (1995).
10. D. Chapman, The measurement of molecular diversity: a three-dimensional approach, *J. Comput.-Aid. Mol. Design* 10:501 (1996).



## EASY DOES IT: REDUCING COMPLEXITY IN LIGAND-PROTEIN DOCKING

Djamal Bouzida, Daniel K. Gehlhaar, and Paul A. Rejto

Agouron Pharmaceuticals, Inc.  
3301 North Torrey Pines Court  
La Jolla, CA 92037

### INTRODUCTION

Computational methods in structure-based drug design are used in a number of applications, including prediction of the structure of ligand-protein complexes also known as the docking problem, estimation of ligand-protein binding affinity, and in *de novo* design. Depending on the level of detail incorporated into the model, as well as the number of times the calculation is performed, the computational demands of these studies range from a few seconds on a small workstation to months on dedicated supercomputers. In the pharmaceutical industry, the criterion for a useful computational technique is simple: it must provide information of sufficient quality to impact the discovery or optimization of lead compounds, and it must do so in a timely manner.

In our work, we have found that a critical decision that governs the successful application of computational methods in structure-based drug design is the choice of the model used to represent the problem. Traditionally, computational chemists have developed highly detailed force fields to describe atomic interactions. While in principle such efforts provide accurate representations of chemical systems, there are two significant practical problems that arise in their application. First, it is difficult to obtain high-quality parameters for the force field in a rapid manner, and second, it is not possible to adequately sample the enormous conformational space of ligand-protein systems. As a consequence, the computational requirements of detailed atomic-level simulations are not compatible with the large number of molecules that are now available in commercial databases or in typical combinatorial libraries. As such, there is a need for methods that efficiently reduce the size and complexity of the problem, while still providing useful information.

Previously, we have developed a method for the prediction of bound ligand-protein complexes based on a simplistic, short-ranged potential.<sup>1</sup> Because structure prediction is a much easier problem than free energy calculation, this potential, while not sufficiently accurate to estimate ligand-protein binding affinities, correctly predicts the bound conformation for a variety of ligand-protein complexes. Unlike detailed force fields, this potential yields a smoother energy landscape and is more compatible with high throughput computational database screening. More recently, we have extended this method to two

types of docking simulations where some features of the resulting ligand-protein complex are known *a priori*.<sup>2</sup> In complexes where a covalent bond is formed between a nucleophilic cysteine or serine residue and an electrophilic ligand atom, constraints are placed on the location of the ligand. Likewise, when combinatorial libraries are developed that include a substructure whose bound conformation is known,<sup>3</sup> the size of the available conformational space is reduced.

## RESULTS

To validate the ability to identify leads from a database, ligands containing ketones and esters were screened against the reactive enzyme porcine pancreatic elastase. A known inhibitor was ranked in the top one percent of all compounds that satisfied the screening criteria, which included a generalization of the LUDI scoring function to estimate binding affinity.<sup>4</sup> In addition, the correct stereoisomer and binding mode for this compound were selected. Compounds unrelated to this inhibitor were also found, some of which form favorable hydrogen bonds in the active site, though none have been tested for activity.

A virtual library was generated by direct alkylation of the pteridine ring in methotrexate with 7,677 compounds, each of which had a molecular weight less than 250 and an amine group with at least one hydrogen and one neighbor in an aromatic group. From this virtual library, only 516 satisfied the screening criteria, 7% of the original library. As anticipated, methotrexate was predicted to have the best binding energy, but a number of other compounds were generated that also form good hydrogen bond interactions within the active site.

## CONCLUSIONS

In order to successfully apply computational tools in structure-based drug design, it is important to use all information about the system of interest prior to beginning the computational study. We have developed a simplified representation of ligand-protein interactions that provides a balance between accuracy and speed, and software that takes advantage of knowledge about the structure of certain types of ligand-protein complexes in order to reduce computational complexity. We have shown that the predicted structure of known inhibitors of dihydrofolate reductase and porcine pancreatic elastase correspond to the experimentally observed structure with increased probability compared to an unrestricted simulation. When combined with a simple estimate of binding affinity, these inhibitors were ranked favorably, thus enriching the hit rate of the targeted library.

## REFERENCES

1. D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, *Chem. Biol.* 2:317 (1995).
2. D. K. Gehlhaar, D. Bouzida, and P.A. Rejto, Reduced dimensionality in ligand-protein structure prediction: covalent inhibitors of serine proteases and design of site-directed combinatorial libraries in: *ACS Symposium Series on Rational Drug Design* (in press).
3. E.K. Kick, D.C. Roe, A.G. Skillman, G. Liu, G., T.J. Ewing, Y. Sun, I.D. Kuntz, and J.A. Ellman, Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D, *Chem. Biol.* 4:297 (1997).
4. P.W. Rose, Scoring methods in ligand design, in *2nd UCSF Course in Computer-Aided Molecular Design*, San Francisco, (1997).

# STUDY OF THE MOLECULAR SIMILARITY AMONG THREE HIV REVERSE TRANSCRIPTASE INHIBITORS IN ORDER TO VALIDATE GAGS, A GENETIC ALGORITHM FOR GRAPH SIMILARITY SEARCH

Nathalie MEURICE<sup>1</sup>, Gerald M. MAGGIORA<sup>2</sup>, Daniel P. VERCAUTEREN<sup>3</sup>

<sup>1</sup> F.R.I.A. PhD Fellowship

<sup>1,3</sup> Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles, 61, B-5000 NAMUR (Belgium)

<sup>2</sup> Computer-Aided Drug Discovery, Pharmacia & Upjohn, 301 Henrietta Street, Kalamazoo, MI 49007-4940

## INTRODUCTION

The conception of potent therapeutical agents relies on the knowledge of the interaction mode between the ligands and their receptor sites. However, very often, the direct study of these interactions is difficult as the three-dimensional (3D) structure of the receptor sites is not completely known. Consequently, an indirect approach resides in the comparison of the ligands of interest on the basis of their physico-chemical properties, in such a way to deduce the nature of their common molecular sites involved in the binding to the macromolecule and/or responsible for their particular activity.

In this general framework, we have focused our efforts on the elaboration and improvement of an original genetic algorithm method, named GAGS (Meurice et al., 1997), for computing the similarity between ligands of biopharmacological interest, especially those whose receptor crystal structures have not been determined.

In order to validate our GAGS approach, we study a system of ligands whose receptor structures are available and compare the molecular alignements to the available experimental (XRAY) and theoretical (MIMIC) models.

## STUDIED SYSTEM

We have selected a set of three HIV Reverse Transcriptase Inhibitors (HIV RTI's), namely Nevirapine,  $\alpha$ -APA, and TIBO. The crystal structures of these ligands bound to HIV Reverse Transcriptase (RT) are available. An « experimental » model is thus obtained by superimposing the crystal structures of HIV RT with the bound inhibitor, and then removing the protein.

## TOPOLOGICAL ANALYSIS OF 3D SMOOTHED ELECTRON DENSITY MAPS

*Ab initio* 3D electron density maps (EDM) of the three selected ligands have been obtained using RHF/SCF/6-31G\* calculations. Removal of the details contained in these maps using wavelet multiresolution analysis (Daubechies filter, 20 coefficients, 3 levels of smoothing) produces smoothed 3D grids. The information contained in such 3D smoothed

EDM can be further simplified into molecular graphs using topological analysis, which allows to locate the critical points of the electron density function, *i.e.*, peaks and passes in our study. Punctual values of the density and distances between critical points are set as the diagonal and non-diagonal elements of property matrices, respectively. As a result, the molecular graphs of TIBO, Nevirapine, and  $\alpha$ -APA contain 6, 12, and 14 critical points, respectively. The three molecular graphs are then compared using our GAGS method.

## GAGS, GA FOR GRAPH SIMILARITY SEARCH

GAs are optimization techniques inspired by the natural concepts of the Darwinian evolution. Within GAGS, the chromosomes are defined as 2D integer arrays where the first dimension is the number of ligands to be compared, and the second one is determined by the number of fitting points. In such a way, each chromosome is a hypothesis of subgraph match between the initial set of molecular graphs. The evaluation function measures an RMS value between the property matrices built from the evaluated subgraph match, and is thus minimized during the GA generations. An automated decoding process has been implemented in order to create molecular overlays corresponding to each of the solution chromosomes.

The GAGS comparison leads to overlays in agreement with the « experimental model ». When optimized in the MIMIC steric and electrostatic fields, the GAGS overlay converges towards the superimposition of the RTI's that was obtained by the MIMIC model, with a similarity of 61% (Mestres *et al.*, 1997).

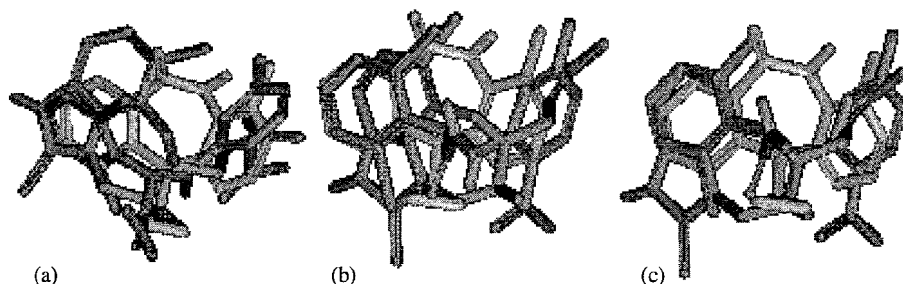


Figure 1- (a) GAGS, (b) experimental, and (c) MIMIC superimposition models.

## CONCLUSIONS AND OUTLOOK

This work allows to assess the GAGS approach as a valuable tool for the discovery of good ligand alignments. As a consequence, GAGS might be used as a powerful search engine as well and the resulting molecular overall superimpositions might then be quickly optimized in MIMIC fields in order to produce precise overlays and yield quantitative similarity indices.

## REFERENCES

- Mestres, J., Rohrer, D.C., Maggiora, G.M., 1997, MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches, *J. Comput. Chem.*, 18:934.
- Meurice, N., Leherte, L., Vercauteren, D.P., Bourguignon, J.-J., Wermuth, C.G., 1997, Development of a genetic algorithm method especially designed for the elucidation of the benzodiazepine receptor pharmacophore, in: *Computer-Assisted Lead Finding and Optimization*, H. van de Waterbeemd, B. Testa, G. Folkers, eds., Verlag Helvetica Chimica Acta (VHCA), Basel, 497.

# A DECISION TREE LEARNING APPROACH FOR THE CLASSIFICATION AND ANALYSIS OF HIGH-THROUGHPUT SCREENING DATA

Michael F.M. Engels, Hans De Winter, Jan P. Tollenaere

Dept. Theoretical Medicinal Chemistry, Janssen Research Foundation, Beerse, Belgium

High-throughput screening (HTS) of large libraries of compounds is applied by drug companies to pick up active molecules. Besides this "fishing" part, HTS can also play an important role as a source for structure-activity analysis, which relates physicochemical or structural features to biological activity. The latter aspect, although of great importance, has hardly been explored over the last years due to i) the size and complexity of the data sets, ii) the lack of, or rather ignorance about, suitable mathematical tools, iii) the quality of the biological data - biological activity is frequently described by just two (active - not active) or three (active - medium active - not active) categories - and iv) the lack of appropriate molecule descriptors.

In this study, decision tree learning (DTL) and rule induction<sup>1</sup> (RI) have been used for the classification and structure-activity analysis of an in-house set of data on 27000 compounds tested for dopamine D2 binding activity (biological activity indicated as either "active" or "not active"; around 1300 compounds were found to be active). Both, DTL and RI are machine learning methods for finding complex interactions between many variables which try to explain a distinct set of responses. Both methods are able to deal with large numbers of data (observations) and show good performance in analysing noisy data,<sup>2</sup> as HTS data may be. Compounds have been represented either by a set of topological keys created by a customized Daylight program (Daylight Chemical Information Systems Inc.) which calculates all possible substructures consisting of up to four atoms, or by a set of 3D keys as implemented in the ChemX software.

DTL identifies the descriptor with the strongest association with the biological activity. Using that descriptor, the set of data is split into two sets, one in which all compounds possess that feature, and one in which all compounds lack that feature. This procedure is repeated with all resultant subsets as long as the degree of association is above a given threshold criterion. Since sets of data are always split into

---

<sup>1</sup> Quinlan, J.R. "C4.5: Programs For Machine Learning", Morgan Kaufmann Publishers, 1993.



**Poster Session V**  
**Affinity and Efficacy**  
**Models of G-Protein**  
**Coupled Receptors**

## APPLICATION OF PARM TO CONSTRUCTING AND COMPARING 5-HT<sub>1A</sub> AND $\alpha_1$ RECEPTOR MODELS

Maria Santagati<sup>(a)</sup>, Hongming Chen <sup>(b§§)</sup>, Andrea Santagati<sup>(a)</sup>, Maria Modica<sup>(a)</sup>, Salvatore Guccione<sup>(a)</sup>, Gloria Uccello Barretta<sup>(c)</sup>, Federica Balzano<sup>(c)</sup>

<sup>(a)</sup> *Dipartimento di Scienze Farmaceutiche, Università di Catania, viale Andrea Doria 6, Ed. 12, I-95125 Catania, Italy*

<sup>(b)</sup> *Laboratory of Computer Chemistry, Institute of Chemical Metallurgy, Chinese Academy of Sciences, P.O. Box 353 Beijing 100080, P. R. China*

<sup>(c)</sup> *Centro CNR di Studio per le Macromolecole Stereordinate ed Otticamente Attive, Università di Pisa, via Risorgimento 35, I-56126 Pisa, Italy*

---

Based on the Walters' s GERM (Genetic Evolved Receptor Model), PARM (PseudoAtomic Receptor Model) uses a combination of genetic algorithms and a cross-validation technique to produce atomic-level pseudo-receptor models starting from a set of known ligands.

These putative pseudo-receptor models can be used to predict bioactivity of *virtual* molecules by aligning these molecules with the training set molecules, computing the interaction energy between each molecule and interpolating the computed interaction energy in the QSAR regression equation to obtain a predicted bioactivity, so reducing the trial-error procedure in the synthesis of new chemical entities.

---

Serotonin modulates many processes in mammalian peripheral and central nervous system through its interactions with at least 14 receptor subtypes, all but one (5-HT<sub>3</sub> subtype) of which are G protein (heterotrimeric GTP-binding protein)-coupled.

The 5-HT<sub>3</sub> subtype is a ligand-gated ion channel that shares functional and structural similarities with nicotinic acetylcholine receptors.

Aim of the present investigation is to create a 5-HT<sub>1A</sub> model capable of aiding the synthesis of new compounds with improved activities elucidating the possible role of heteroaromatic interactions<sup>1,2</sup> in the receptor binding, and to compare the predictive ability of the new paradigm PARM<sup>3,4</sup> with two *traditional* 3D QSAR techniques such as

---

<sup>§§</sup> **Present address:** Bayer AG, Pharma-Forschung, PH-R Strukturforchung, D-42096, Wuppertal, Germany.



CoMFA<sup>5</sup>(Comparative Molecular Field Analysis) and HASL<sup>6</sup> (Hypothetical Active Site Lattice), as reported in chapter: APPLICATION OF PARM TO CONSTRUCTING AND COMPARING 5-HT<sub>1A</sub> AND  $\alpha_1$  RECEPTOR MODELS. In addition, worth of interest was mapping possible features underlying the 5-HT<sub>1A</sub> or *alpha* selectivity, as shown by some ligands in the investigated thienopyrimidinone series<sup>7</sup>.

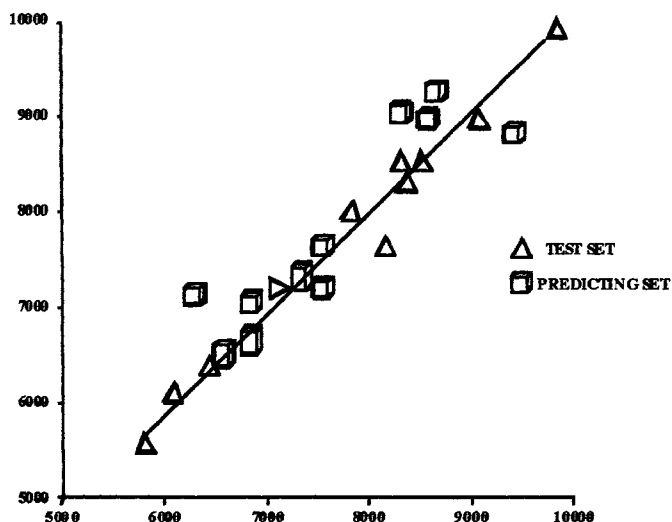
In the PARM<sup>3</sup> computation, 15 kinds of pseudo-receptor atoms are defined first. Then, the molecules in the training set are superimposed on a specific pharmacophore model and a set of grid points is generated around the common surface of the superimposed ligands. Receptor models are made by placing atoms at these points in 3D space, to simulate a receptor active site. These atoms interact with the ligands and the interaction energy between each ligand and the receptor model is computed. By using a genetic algorithm and a cross-validation technique, a number of atomic-level pseudo-receptor models which have a high correlation between intermolecular energy and bioactivity can be built. A QSAR equation is constructed for each model in the linear form of **Bioactivity** = A + B\*E<sub>inter</sub>. Energetic computation in PARM<sup>3</sup> makes use of the TRIPOS 5.0 force field.

PARM<sup>3</sup> generates the receptor models in the MOL2 file, so that we can check the characteristics of the receptor model within the SYBYL software<sup>8</sup>.

In this study, (*forthcoming paper*) the initial population of pseudo-receptors was set to 1500, the maximum generation to 2000, the number of grid points was set to 49 and the cushion distance (the distance between grid point and the closest ligand atom) was 0.5 Å. PARM<sup>3</sup> is allowed to run until a series of receptor site models with high conventional correlation coefficients and cross-validated R<sup>2</sup> are obtained. Usually, the top 20 models are used to predict bioactivity and compared with a test set.

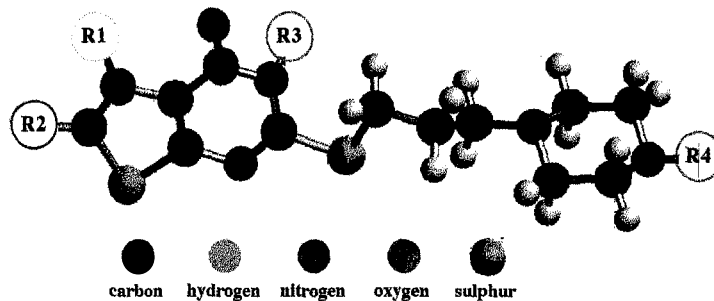
Models fifteen and four (*Table I and II*) were found to have the best predictions for the 5-HT<sub>1A</sub> and  $\alpha_1$ -AR data sets, respectively.

These two models are analysed in **Figs 1 and 2**. See also **Fig 6 and 7** of chapter 5-HT<sub>1A</sub> RECEPTORS MAPPING BY CONFORMATIONAL ANALYSIS (2D NOESY/MM) AND "THREE WAY MODELLING" (HASL, CoMFA, PARM).



**Fig 1** Analysis of the best predictive 5-HT<sub>1A</sub> model (model fifteen)

Table I PARM computation results of the  $HT_{1A}$ -receptor model



Compd	R1	R2	R3	R4	Exp -log IC50	Calc -log IC50	Residual	$E_{inter}$ (kcal/mol)
2 (44)	Me	Me	3-CIPh	H	6.005	6.304	0.299	9.267
3 (46)	Me	Me	H	2-OMe-Ph	7.620	7.891	0.271	-3.306
4 (48)	Me	Me	H	1-naphtyl	6.450	6.312	-0.138	9.199
5 (49)	Me	Me	H	2-pyrimidinyl	6.646	6.424	-0.222	8.314
7 (53)	$-(CH_2)_4-$	H	H	2-OMe-Ph	7.229	7.681	0.452	-1.640
11 (61)	H	Ph	H	2-OMe-Ph	6.413	6.898	0.485	4.564
12 (63)	H	Ph	H	1-naphtyl	5.697	5.609	-0.088	14.773
13 (64)	$-(CH=CH)_2$	H	H	2-OMe-Ph	7.337	7.539	0.202	-0.519
14 (65)	H	H	NH <sub>2</sub>	2-OMe-Ph	8.921	8.325	-0.596	-6.740
17 (68)	Me	Me	NH <sub>2</sub>	Ph	8.481	8.616	0.135	-9.047
19 (69)	Me	Me	Me	2-OMe-Ph	9.523	9.119	-0.404	-13.031

Table I continued

Compd	R1	R2	R3	R4	Exp -log IC50	Calc -log IC50	Residual	E <sub>inter</sub> (kcal/mol)
20 (70)	Me	Me	NH <sub>2</sub>	2-OMe-Ph	8.523	8.618	0.095	-9.068
21 (71)	Me	Me	NHPh	2-OMe-Ph	6.304	6.044	-0.260	11.323
22 (72)	Me	Me	Me	2-pyrimidinyl	8.167	7.697	-0.470	-1.770
23 (73)	Me	Me	NH <sub>2</sub>	2-pyrimidinyl	9.301	9.540	0.239	-16.371
<b>-logIC50=7.474-0.126*E<sub>inter</sub> r<sup>2</sup>=0.962, R<sup>2</sup>cv=0.906 SD=0.353</b>								
1 (43)*	Me	Me	H	2-OMe-Ph	6.337	7.620	1.283	-1.156
6 (50)*	-(CH <sub>2</sub> ) <sub>4</sub> -	H	H	2-Cl-Ph	6.074	7.823	1.749	-2.770
9 (56)*	-(CH <sub>2</sub> ) <sub>4</sub> -	H	H	1-naphtyl	6.431	6.939	0.508	4.236
10 (57)*	-(CH <sub>2</sub> ) <sub>4</sub> -	H	H	2-pyrimidinyl	6.297	6.888	0.591	4.641
15 (66)*	-(CH <sub>2</sub> ) <sub>4</sub> -	Me	H	2-OMe-Ph	8.155	8.596	0.441	-8.889
16 (67)*	-(CH <sub>2</sub> ) <sub>4</sub> -	NH <sub>2</sub>	H	2-OMe-Ph	8.886	8.760	-0.126	-10.193
24 (74)* <sup>a</sup>	Me	Me	NH <sub>2</sub>	2-OMe-Ph	7.187	7.743	0.556	-2.137
25 (78)* <sup>b</sup>	-	-	NH <sub>2</sub>	2-OMe-Ph	9.097	9.468	0.371	-15.801

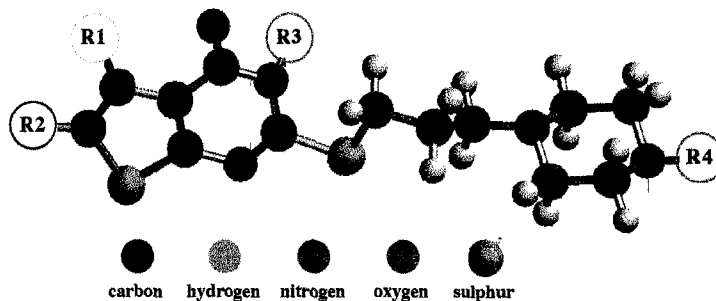
SD\*=0.86

\*In brackets the number in the paper (see ref 10).

\*Test set compounds

<sup>a</sup>The piperazine ring has been replaced by a piperidine nucleus.<sup>b</sup>The thiophene ring has been replaced by a benzene nucleus.

Table II PARM computation results of the  $\alpha_1$ -AR model



Compd	R1	R2	R3	R4	Exp -log IC50	Calc -log IC50	Residual	E <sub>inter</sub> (kcal/mol)
2 (44)	Me	Me	3-ClPh	H	6.524	6.652	0.128	16.030
3 (46)	Me	Me	H	2-OMe-Ph	7.389	7.594	0.205	7.128
4 (48)	Me	Me	H	1-naphtyl	6.053	6.136	0.083	20.900
5 (49)	Me	Me	H	2-pyrimidinyl	5.959	5.982	0.023	22.349
7 (53)	-(CH <sub>2</sub> ) <sub>4</sub> -	H		2-OMe-Ph	7.420	7.566	0.146	7.3930
11 (61)	H	Ph	H	2-OMe-Ph	6.650	6.697	0.047	15.601
12 (63)	H	Ph	H	1-naphtyl	5.610	5.602	-0.008	25.945
13 (64)	-(CH=CH) <sub>2</sub>	H		2-OMe-Ph	7.041	7.258	0.217	10.306
14 (65)	H	H	NH <sub>2</sub>	2-OMe-Ph	8.538	8.538	0.000	-1.785
17 (68)	Me	Me	NH <sub>2</sub>	Ph	7.367	6.796	-0.571	14.669
19 (69)	Me	Me	Me	2-OMe-Ph	7.569	7.565	-0.004	7.407

Compd	R1	R2	R3	R4	Exp -log IC50	Calc -log IC50	Residual	E <sub>inter</sub> (kcal/mol)
20 (70)	Me	Me	NH <sub>2</sub>	2-OMe-Ph	8.137	7.962	-0.175	3.652
21 (71)	Me	Me	NHPh	2-OMe-Ph	7.495	7.403	-0.092	8.932
22 (72)	Me	Me	Me	2-pyrimidinyl	5.693	5.745	0.0522	4.588
23 (73)	Me	Me	NH <sub>2</sub>	2-pyrimidinyl	6.296	6.245	-0.051	19.865

**-logIC50=8.349-0.106\*E<sub>inter</sub> r=0.975, R<sub>cross</sub><sup>2</sup>=0.941 SD=0.197**

Table II continued (Test set molecules)

1 (43)*	Me	Me	H	2-OMe-Ph	6.793	6.886	0.093	13.811
6 (50)*	-(CH <sub>2</sub> ) <sub>4</sub> -		H	2-Cl-Ph	6.775	6.581	-0.194	16.696
9 (56)*	-(CH <sub>2</sub> ) <sub>4</sub> -		H	1-naphtyl	6.352	6.593	0.241	16.578
10 (57)*	-(CH <sub>2</sub> ) <sub>4</sub> -		H	2-pyrimidinyl	5.741	6.830	1.089	14.341
15 (66)*	-(CH <sub>2</sub> ) <sub>4</sub> -	Me		2-OMe-Ph	7.194	7.919	0.725	4.063
16 (67)*	-(CH <sub>2</sub> ) <sub>4</sub> -		NH <sub>2</sub>	2-OMe-Ph	7.409	7.924	0.515	4.014
24 (74)* <sup>a</sup>	Me	Me	NH <sub>2</sub>	2-OMe-Ph	7.444	8.217	0.773	1.251
25 (78)* <sup>b</sup>	-	-	NH <sub>2</sub>	2-OMe-Ph	8.398	7.893	-0.505	4.307

**SD\*=0.61**

\*In brackets the number in the paper (see ref 10).

\*Test set compounds

<sup>a</sup>The piperazine ring has been replaced by a piperidine nucleus.

<sup>b</sup>The thiophene ring has been replaced by a benzene nucleus.

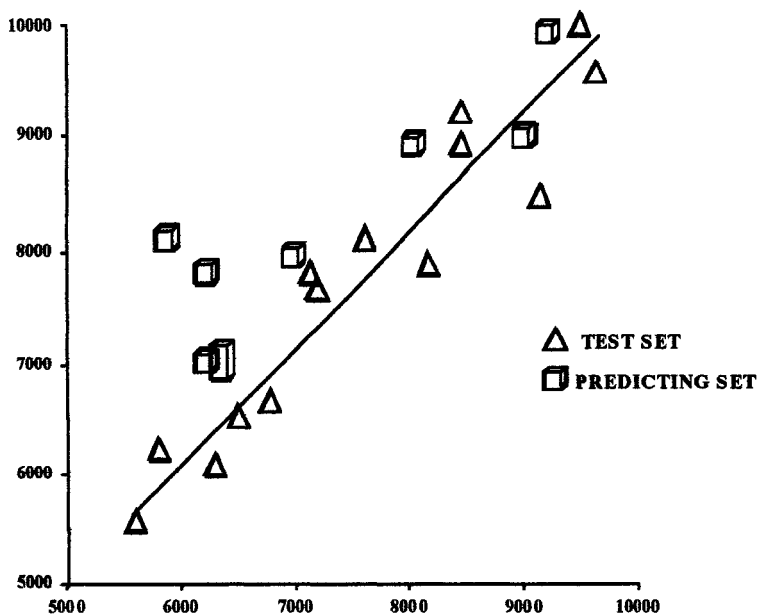


Fig 2 Analysis of the best predictive  $\alpha_1$ -AR model (model four)

**Acknowledgements.** Financial support (40%) from Italian MURST and the kind technical support from TECHNOSOFT (via Galliano, 25, I-95125 Catania, Italy) are gratefully acknowledged.

S. Guccione thanks Prof. Eric Walters for the helpful discussion and directions.

Hongming Chen thanks Prof. J. J. Zhou for the helpful support and the high scientific contribution to the ongoing PARM investigations.

#### References <sup>(1)</sup>

1. T. M. Fong, H. Yu, R. R. C. Huang, M. A. Cascieri, and C. J. Swain, Relative contribution of polar interactions and conformational compatibility to the binding of neurokinin-1 receptor antagonists, *Mol. Pharmacol.*, **50**: 1605 (1996) and enclosed references.
  2. M. Modica, Synthesis of thieno[2,3-d]pyrimidine derivatives. Ligands to the 5-HT<sub>1A</sub> serotonergic receptor, Tesi di Dottorato di Ricerca (*Italian Ph.D.*), University of Catania (1994).
  3. H. M. Chen, J. J. Zhou, G. R. Xie, PARM: A genetic evolved algorithm to predict bioactivity, *J. Chem. Inf. Comput. Sci.*, **38**: 243 (1998).
  4. D. E. Walters and T. D. Muhammad, Genetically evolved receptor models (GERM): a procedure for construction of atomic-level receptor site models in the absence of a receptor crystal structure, in: *Genetic Algorithms in Molecular Modelling*, J. Devillers, ed., Academic Press, London (1996).
- <sup>(1)</sup> Refs. 5.- 8. -see chapter: 5-HT<sub>1A</sub> RECEPTORS MAPPING BY CONFORMATIONAL ANALYSIS (2D NOESY/MM) AND "THREE WAY MODELING" (HASL, CoMFA, PARM), by S. Guccione et al. See refs 13., 7.,10., 11.

# A NOVEL COMPUTATIONAL METHOD FOR PREDICTING THE TRANSMEMBRANAL STRUCTURE OF G-PROTEIN COUPLED ANAPHYLATOXIN RECEPTORS, C5aR and C3aR

Naomi Siew, Anwar Rayan, Wilfried Bautsch<sup>1</sup> and Amiram Goldblum

Department of Medicinal Chemistry, School of Pharmacy, Hebrew University of Jerusalem, Jerusalem, ISRAEL 91120, and <sup>1</sup>Institut für Medizinische Mikrobiologie, Medizinische Hochschule, Carl-Neuberg-Str.1, D-30625 Hannover, Germany

**Introduction:** The receptor C5aR (350 residues) is found in the membranes of polymorphonuclear leukocytes. When activated by its ligand, C5a, a very potent chemoattractant, an amplification of the inflammatory process occurs. C3aR (482 residues) is similarly associated with such events, although to a lesser extent. High levels of C5a (74 aa) and C3a (77 aa) were connected to inflammatory and autoimmune diseases, such as Rheumatoid Arthritis and Adult Respiratory Disease Syndrome, that can even lead to death. The design and construction of potent antagonists to each of the two receptors is a major avenue that could lead to control of such conditions. C5aR and C3aR belongs to the superfamily of G Protein-Coupled Receptors (GPCR), which includes over 700 members, involved in many important biological activities. The structure of these proteins has not been determined yet and attempts to rationally design drugs for them are still limited.

One of the very few membranal proteins whose structure was solved is bacteriorhodopsin, a membranal proton pump. It consists of seven transmembranal helices, connected by extra- and intra-cellular hydrophilic loops, an extra-cellular N-terminal and an intra-cellular C-terminal. Bacteriorhodopsin is not a GPCR and has no significant homology with this family, yet there is experimental evidence that demonstrates a similar topology. The structure of bacteriorhodopsin has been initially determined by electron microscopy at low resolutions parallel and perpendicular to the membrane (1BAD). More recently, X-ray structure of bacteriorhodopsin was determined at 3.5Å resolution (2BRD). Due to the fact that the three dimensional structure of the GPCRs was not solved yet, constructing theoretical models for these receptors, in order to investigate their interactions with their ligands and their activation mechanism, has become very common.

**Method:** We view the process of receptor assembly as a result of two different mechanisms: An equilibrium of helices between water and the membrane, governed by their hydrophobicity, followed by an association of helices which may be close to interactions in globular proteins. We employed a knowledge-based force field constructed from the Protein Data Bank (globular proteins), where all the interactions between pairs of amino acid residues have been evaluated according to their occurrence and the appropriate statistical weights (Miyazawa and Jernigan<sup>1</sup>). Seven regions along the sequence, which are assumed to contain the seven transmembranal helices, were found by means of hydrophobicity profiles and multiple sequence alignment with other GPCRs, with the program HOMOLOGY. These regions are input to our program THREAD. Each region is longer than the sequence that is expected to reside in the membrane in a helical structure. The program suggests the limits for each helix. It threads the seven sequences simultaneously on the coordinates of bacteriorhodopsin, combining all the possible options for each helix.

THREAD employs the template structure of 1bad.pdb or 2brd.pdb (or any other template) and "threads" a GPCR in order to find the best GPCR structure by using two methods:

1) Calculating the overall contact energy of the structure. Two residues, whose C $\alpha$ -C $\alpha$  distance is less or equal to 7Å (for Gly - 6Å) and whose C $\beta$ -C $\beta$  distance is less than their C $\alpha$ -C $\alpha$  distance, are considered to be in contact. The contact energy value for every pair is summed up for the whole protein. The lowest energy structures are retained for further processing. The detailed structure of side chains of residues are not taken into account at this stage.

2) Summing up the hydrophobicity values in the membrane and outside. For every structure threaded, the hydrophobicity values of each residue in the membrane (i.e. in a helix) are summed. The program searches for the most hydrophobic structure.

Side chains were added by two methods that employ a rotamer library. HOMOLGY uses a backbone independent library of rotamers, and the side chains are added depending on the sequence of addition. SCWRL<sup>2</sup> adds side chains from a backbone-dependent library, and optimizes the results by identifying clashes and combining all clashing side-chains into a group, for which all combinatorics for the rotamers are tested.

**Results:** THREAD was first tested on the theoretical set of coordinates for bacteriorhodopsin, 1bad.  $9.3 \cdot 10^5$  structures were threaded. The best result was obtained (table 1), but for some helices other results had very close weights. The hydrophobicity method is least accurate in the case of helix B ( $\Delta$ =two turns), which is more hydrophilic than other helices. Contact energy gave accurate results for most helices, with helix F being about one turn distant from experimental.

**Table 1. The beginnings of the helices of bacteriorhodopsin**

Helix	A	B	C	D	E	F	G
3D Structure	22	51	87	119	150	179	215
Contact Energy	23	51	88	120	148	176	216
Hydrophobicity	22	57	88	119	148	180	214

For C5aR,  $1.7 \cdot 10^7$  structures were checked. The two methods gave fairly close results (table 2). For helix C we got two possibilities for the beginning in the hydrophobicity method: residue 104 or residue 111. Helix G could begin at residue 281 or residue 284. In the contact energy method, helix C fluctuates between 107 and 109, helix F between 245 and 241, and helix G between 281 and 284. The two best solutions for each method are depicted in table 2. However, quite a few other results with close energies exist. The results for C3aR based on 1bad coordinates gives as helix starts: A, 24; B, 57; C, 98; D, 141; E, 342; F, 379; G, 410 (contact energy only).

**Table 2. The beginnings of the helices of C5aR**

Helix	A	B	C	D	E	F	G
Contact Energy	38	71	107	153	207	245	281
						241	
Hydrophobicity	38	72	104	152	207	243	281
		73	111				284

## REFERENCES

1. Miyazawa, S. and Jernigan, R. (1985). *Macromolecules* 18: 534-552.
2. R. L. Dunbrack, Jr. and M. Karplus (1993) *J. Mol. Biol.* 230: 543-571



## RECEPTOR-BASED MOLECULAR DIVERSITY: ANALYSIS OF HIV PROTEASE INHIBITORS

Tim D.J. Perkins, Nasfim Haque, and Philip M. Dean

Drug Design Group  
Department of Pharmacology  
University of Cambridge  
Tennis Court Road  
Cambridge UK CB2 1QJ

### INTRODUCTION

Focused combinatorial libraries are a useful way of approaching structure-based drug design, but they may show unexpected bias in exploring the receptor site. One way to monitor this coverage is by assessing which hydrogen-bonding groups at the receptor site are used by each ligand in the library. In this communication, we present an analysis of the hydrogen bonds formed between inhibitor and enzyme in a set of HIV protease complexes. These data are a model for a larger combinatorial library, and have allowed us to develop methods for receptor-based diversity analysis.

### ANALYSIS OF HIV PROTEASE INHIBITORS

The Brookhaven Protein Databank<sup>1</sup> was searched for X-ray coordinates of HIV protease-inhibitor complexes with resolution better than 2.5 Å, and 31 non-mutant entries were selected. The ligands were extracted, and hydrogen atoms were added semi-automatically. The active-site water molecule (sometimes labelled residue HOH 301) was considered as part of the protein site and relabelled consistently. Hydrogen bonds were identified between each inhibitor and its enzyme using X-ray crystal criteria.<sup>2</sup> These data were then indexed by the 29 site atoms used by at least one ligand. In the cases where two orientations of the inhibitor are present in the complex, a hydrogen bond from either orientation was sufficient for the site atom to be marked as occupied.

Each pair of ligands was then compared, in terms of the site atoms occupied, using two separate metrics: Tanimoto similarity coefficient and Euclidean distance. A similarity or distance matrix was constructed, and input to cluster analysis using Ward's minimum variance method (see Figure 1). The number of significantly different clusters was determined with Mojena's stopping rule,<sup>3</sup> at a significance level of  $P < 0.05$ .



# APPLICATION OF SELF-ORGANIZING NEURAL NETWORKS WITH ACTIVE NEURONS FOR QSAR STUDIES

Vasyl V. Kovalishyn,<sup>1</sup> Igor V. Tetko,<sup>1,2</sup> Alexander I. Luik,<sup>1</sup>  
Alexey G. Ivakhnenko,<sup>3</sup> and David J. Livingstone<sup>4</sup>

<sup>1</sup>Institute of Bioorganic & Petroleum Chemistry  
Murmanskaya 1, Kyiv, 253660 Ukraine

<sup>2</sup>Institut de Physiologie, Rue du Bugnon 7  
Lausanne, CH-1005 Switzerland

<sup>3</sup>Glushkov Institute of Cybernetics  
Acad. Glushkov Avenue, 20, Kyiv 252207, Ukraine

<sup>4</sup>ChemQuest, 19-21 Cheyney Street, Herts, SG8 0LP and Centre for  
Molecular design, University of Portsmouth, Hants, PO1 2EG U.K.

## INTRODUCTION

The interest for development of rational methods for investigation of relations between structure and activity of chemical compounds has essentially increased in the last years. Artificial neural networks have become one of the leading methods in this field.<sup>1</sup> However, there are some difficulties (such as limitation in speed, local minima, overfitting/overtraining problems,<sup>1</sup> etc.) with an application of these methods for analysis of data sets with a large number of input parameters and, particularly, three-dimensional electronic parameters of compounds generated by 3D QSAR approaches, such as CoMFA. The current study analyses a new method, i.e. neural networks with active neurons,<sup>2,3</sup> that could be used in such QSAR studies. We also propose to combine this method with Kohonen's Self-Organizing Maps (SOM) used for preprocessing of 3D QSAR data sets. The performance of new method is compared with that of fixed size neural networks.

## METHODS

The neural network with active neurons consists of certain number of layers, each of which is composed of several computing modules. These modules are refer to as the active neurons.<sup>2</sup> The neurons at the same layer can differ one from another both in a set of input and output variables. The process of learning (self-organizing) of an active neuron consists in estimation of importance of inputs used to minimize the given objective function of the neuron. The choice of the optimal set of variables is realized by reduced sorting of possible sets of variables and the variables increasing the objective functions are eliminated. The choice of links by active neurons defines the structure of the whole network. It is important to note that the set of output variables in addition to the analyzed activity also includes input variables. Thus the number of active neurons in each layer is equal to number of variables given in initial data sampling.

Each layer of active neurons acts similarly to the Kalman filter, i.e. the output set of variables repeats the input set but with filtration of noise. The output variables of previous

layers are used as secondary inputs for the neurons of next layer. The computing modules are united in a multi layered structure with the purpose to increase the algorithm accuracy by a more complete processing of the input information.

The self-organization of active neurons was done using the analogues complexing (AC) algorithm,<sup>3</sup> that is one of the approaches developed within framework of the Group Method of Data Handling (GMDH) methods.<sup>4</sup> This algorithm detects an analog of each analyzed molecule (i.e., the molecule that is the nearest neighbor of the analyzed molecule in the Euclidean space) and considers the activity of analog as the predicted value of the molecule.

The total number of layers in neural network was restricted to 10. The analysis of CoMFA dataset included a preprocessing of input variables using SOM. We used the regularity criterion of minimum variance of the prediction error<sup>4</sup> to calculate the optimal partitioning of the input parameter space.

## RESULTS

The antifilarial activity of 53 antimycin analogues<sup>5</sup> and charge-transfer properties of 35 monosubstituted benzenes<sup>1</sup> were analyzed. The both sets have a large number of input parameters. An optimization of inputs, e.g. by pruning algorithms, improved prediction ability of the fixed size neural networks applied to these data.<sup>1,5</sup> The CoMFA dataset included 82 benzylpiperidine derivatives with AchE inhibitory activity.<sup>6</sup>

**Table 1.** The leave-one-out results calculated for analyzed QSAR examples.

data set	total params	neuronet with active neurons			fixed-size neural network	
		no <sup>1</sup>	first layer	best layer	all params	pruned params
Antimycin analogues	53	6	0.74 <sup>2</sup> (0.51) <sup>3</sup>	0.91 (0.81)	0.66 (0.43)	0.91 (0.67)
Benzenes	31	2	0.74 (0.51)	0.95 (0.89)	0.89(0.78)	0.97 (0.95)
Benzylpiperidines	188 (28224) <sup>4</sup>	4	0.86(0.71)	0.89(0.78)	0.56(0.55)	0.72(0.73)

<sup>1</sup>cardinal number of the layer (the best layer) with the lowest error of the network; <sup>2</sup>correlation coefficient  $R$ ; <sup>3</sup>cross-validated  $q^2$ ; <sup>4</sup>number of CoMFA parameters before preprocessing with SOM.

The calculated results for antimycins and benzylpiperidines by neural networks with active neurons were better than results of back-propagation neural networks, while the opposite was true for benzenes. The further studies are required to provide a more objective comparison of the methods.

## Acknowledgments

This study was partially supported by INTAS-Ukraine grant 95-0060. The authors thank Prof. Jacques R. Chretien (University of Orleans) for providing us the CoMFA data.

## REFERENCES

1. D.J. Livingstone, D. T. Manallack, and I. V. Tetko, Data Modelling with Neural Networks - Advantages and limitations, *J. Comp. Aid. Mol. Design.* 11:135-142.(1997).
2. A.G. Ivakhnenko, G.A. Ivakhnenko, and J.-A. Muller, Self-organization of neural networks with active neurons, *Pattern Recognition and Image Analysis* 2:185-196 (1994).
3. A.G. Ivakhnenko, V.V. Kovalishyn, I.V. Tetko, A.I. Luik, G.A. Ivakhnenko, and N.A. Ivakhnenko, Application of self-organizing neural networks with active neurons for prediction of bioactivity of chemical compounds by the analogues search algorithm, *Problems of control and information* in press.
4. H.R. Madala, A.G. Ivakhnenko. *Inductive Learning Algorithms for Complex Systems Modeling*, CRC Press Inc., Boca Raton (1994).
5. V.V. Kovalishyn, I.V. Tetko, A.I. Luik, V.V. Kholodovych, A.E.P. Villa, and D.J. Livingstone, Neural network studies. 3. Variable selection in the cascade-correlation learning architecture, *J. Chem. Inf. Comput. Sci.* 38:651-659 (1998).
6. P. Bernard, D.B. Kireev, J.R. Cretien, P.-L. Fortier, and L. Coppet, Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with "natural" alignment, *J. Comp. Aid. Mol. Design.* in press (1998).

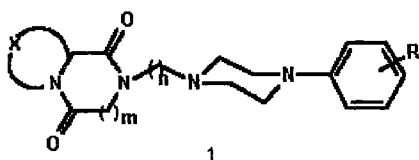
# APPLICATION OF ARTIFICIAL NEURAL NETWORKS IN QSAR OF A NEW MODEL OF PHENYLPYPERAZINE DERIVATIVES<sup>1</sup> WITH AFFINITY FOR 5-HT<sub>1A</sub> AND $\alpha_1$ RECEPTORS: A COMPARISON OF ANN MODELS

María L. López-Rodríguez,<sup>a</sup> M. Luisa Rosado,<sup>a, b</sup> M. José Morcillo,<sup>a</sup> Esther Fernández,<sup>a</sup> and Klaus-Jürgen Schaper<sup>c</sup>

<sup>a</sup>Departamento de Química Orgánica I, Facultad de Químicas, Universidad Complutense 28040 Madrid, Spain; <sup>b</sup>Saint Louis University, Avenida del Valle 34, Madrid Campus; <sup>c</sup>Research Center, D-23845 Borstel, Germany

During the last years artificial neural networks (ANN) have been applied successfully in the QSAR field. It has been demonstrated that this new technique is often superior to the traditional Hansch approach, providing more accurate predictions. The advantage of ANN is that with the presence of hidden layers, neural networks are able to perform nonlinear mapping of the physicochemical parameters and of the corresponding biological activity.

Here, a test series of 32 phenylpiperazines **1** with affinity for 5-HT<sub>1A</sub> and  $\alpha_1$  receptors was subjected to QSAR analysis using artificial neural networks (ANNs). Our aim is to get insight into the structural requirements that are responsible for 5-HT<sub>1A</sub>/ $\alpha_1$  selectivity in order to design new ligands with high selectivity for the 5-HT<sub>1A</sub> receptor.



X =  $-(\text{CH}_2)_3-$ ,  $-(\text{CH}_2)_4-$ ; m = 0, 1; n = 3, 4  
R = *o*-CH<sub>3</sub>, *o*-OCH<sub>3</sub>, *o*-OBu, *o*-COOPr,  
*o*-OCONHPr, *o*-CN, *m*-CF<sub>3</sub>, *m*-NH<sub>2</sub>,  
*m*-NHCOPr<sup>1</sup>, *m*-Br (selected by  
EDISFAR program)

The data set used was the *in vitro* 5-HT<sub>1A</sub> and  $\alpha_1$  receptor affinities (expressed as pK<sub>i</sub> values). Each compound was parametrized with six physicochemical descriptors (*F*, *R*, *V<sub>o</sub>*, *V<sub>m</sub>*,  $\pi_o$ ,  $\pi_m$ ) and three indicator variables (*I<sub>A</sub>* = 1 or 0 for X =  $-(\text{CH}_2)_4-$  or  $-(\text{CH}_2)_3-$ , *I<sub>B</sub>* = 1 or 0 for m = 1 or 0, *I<sub>n</sub>* = 1 or 0 for n = 4 or n = 3).

The neural network employed for this modeling was a fully connected three layer network (input, hidden, output) trained by back-propagation of error. Initially the number of neurons in the input layer was equal to the number of molecular descriptors and indicator

variables, whereas the output layer had only one neuron. The number of neurons in the hidden layer was determined by trial and error. The best ANN models are shown in Table I.

Table I. ANN Models

Receptor	Non significant Parameters	Architecture	r	r <sup>2</sup>	s
5-HT <sub>1A</sub>	R, $\pi_m$	7-2-1	0.983	0.966	0.149
$\alpha_1$	R	8-2-1	0.991	0.982	0.136

The dependence of biological activity on the physicochemical parameters was illustrated in 3-D diagrams. On the basis of the obtained plots, the 5-HT<sub>1A</sub> affinity has a non linear dependence with  $F$ ,  $V_o$ ,  $V_m$  and  $\pi_o$ , nevertheless the nonlinear relationship is not far from the planar one. The  $\alpha_1$  affinity has a clear nonlinear dependence with  $F$ ,  $V_o$ ,  $V_m$ ,  $\pi_o$  and  $\pi_m$ .

A comparison of both analyses gives an additional understanding for 5-HT<sub>1A</sub>/ $\alpha_1$  selectivity: (a) High  $F$  values increase the binding affinity for 5-HT<sub>1A</sub> receptors and decrease the affinity for  $\alpha_1$  sites; (b) The lipophilicity at the *meta* position has only influence for the  $\alpha_1$  receptor; (c) The *meta* position seems to be implicated in the 5-HT<sub>1A</sub>/ $\alpha_1$  selectivity.<sup>2</sup> While the 5-HT<sub>1A</sub> receptor is able to accommodate bulky substituents (about 60 Å<sup>3</sup>) in the region of its active site, the steric requirements of the  $\alpha_1$  receptor at this position are more restricted (between 0-22 Å<sup>3</sup>). A way to improve 5-HT<sub>1A</sub>/ $\alpha_1$  selectivity would be the synthesis of long chain derivatives bearing bulky substituents with high  $F$  values and low  $\pi$  values at the *meta* position. Among the different groups that fulfill these requirements the *m*-NHSO<sub>2</sub>Et was chosen ( $F = 0.419$ ,  $\pi_m = -0.64$ ,  $V_m = 65.31$ ). On this basis, the new ligand EF-7412 (X=-(CH<sub>2</sub>)<sub>3</sub>-, m=0, n=4, R=*m*-NHSO<sub>2</sub>Et) was designed and synthesized. This analog bound at 5-HT<sub>1A</sub> sites [ $K_{i \text{ obsd}}$  (nM)=27.3±5.9;  $K_{i \text{ calcd}}$  (nM)=36.7] and showed high selectivity over the  $\alpha_1$  receptor ( $K_{i \text{ obsd}}$  (nM)>1000;  $K_{i \text{ calcd}}$  (nM)=2745). These results clearly reveal the predictive power of the ANN model and the importance of the nonlinear relationships mapped by the neural networks.

This work was supported by DGICYT PB940289 and CICYT 960360.

- (a) López-Rodríguez *et al.*, 2-[4-(*o*-Methoxyphenyl)piperazin-1-ylmethyl]-1,3-dioxoperhydroimidazo[1,5-*a*]pyridine as a new selective 5-HT<sub>1A</sub> receptor ligand, *Bioorg. Med. Chem. Lett.* 6:689 (1996). (b) López-Rodríguez *et al.*, Synthesis and structure-activity relationships of a new model of arylpiperazines. 1. 2-[[4-(*o*-Methoxyphenyl)piperazin-1-yl]methyl]-1,3-dioxoperhydroimidazo[1,5-*a*]pyridine: a selective 5-HT<sub>1A</sub> receptor agonist, *J. Med. Chem.* 39:4439 (1996). (c) López-Rodríguez *et al.*, Synthesis and structure-activity relationships of a new model of arylpiperazines. 3. 2-[ $\omega$ -(4-Arylpiperazin-1-yl)alkyl]perhydropyrrolo[1,2-*c*]imidazoles and -perhydroimidazo[1,5-*a*]pyridines: study of the influence of the terminal amide fragment on 5-HT<sub>1A</sub> affinity/selectivity, *J. Med. Chem.* 40:2653 (1997). (d) López-Rodríguez *et al.*, 1-[ $\omega$ -(4-Arylpiperazin-1-yl)alkyl]-3-diphenylmethylene-2,5-pyrrolidinediones as 5-HT<sub>1A</sub> receptor ligands: study of the steric requirements of the terminal amide fragment on 5-HT<sub>1A</sub> affinity/selectivity, *Bioorg. Med. Chem. Lett.* 8:581 (1998).
- López-Rodríguez *et al.*, Synthesis and structure-activity relationships of a new model of arylpiperazines. 2. Three-dimensional quantitative structure-activity relationships of hydantoin-phenylpiperazine derivatives with affinity for 5-HT<sub>1A</sub> and  $\alpha_1$  receptors. A comparison of CoMFA models, *J. Med. Chem.* 40:1648 (1997).

## Atypical Antipsychotics: Modelling and QSAR

Benjamin G Tehan<sup>1</sup>, Margaret G Wong<sup>1</sup>, Graeme J Cross<sup>2</sup>, Edward J Lloyd<sup>3</sup>

<sup>1</sup>Chemistry Department

Swinburne University of Technology, Australia 3122

<sup>2</sup>Water Centre, Monash University, Australia 3145

<sup>3</sup>Medicinal Chemistry Department

Victorian College of Pharmacy, Monash University, Australia 3052

### Introduction

Schizophrenia is a debilitating disease that effects approximately 1% of the population with the onset of the disease occurring usually in the mid 20's and persisting in many cases for the lifetime of the patient. Researchers have hypothesised that schizophrenia is due to excessive limbic dopaminergic function within the brain (Jaber et al 1996).

Antipsychotic drugs may be defined as medications that alleviate delusions, hallucinations and some aspects of formal thought disorder that occur in a variety of illnesses, most notably schizophrenia. The mechanism of action of these drugs has focused on their interaction with the central nervous system (CNS) neurotransmitter dopamine (DA). However recent work strongly implicates the neurotransmitter serotonin (5HT) as a further target of action (Schmidt et al 1995). Antipsychotic drugs are further loosely classified into typical or atypical, initially based on animal model tests. Nowadays it is on their reduced liability to produce extrapyramidal side effects (EPS) (Waddington and O'Callaghan 1997), and this has lead to hypotheses in terms of limbic selectivity and 5HT<sub>2</sub>/D<sub>2</sub> ratios.

### Method

A set of ligands ((R)- and (S)-octoclothepein, clozapine, Org5222, seroquel, olanzapine, sertindole, risperidone, ziprasidone, zotepine, remoxipride, loxapine) with high affinity for D<sub>2</sub> and 5HT<sub>2A</sub> and classified as atypical and typical antipsychotics were selected for pharmacophore mapping. Further studies were carried out on sertindole, risperidone, zotepine, ziprasidone and haloperidol once a pharmacophore model had been established. The binding affinity data were gathered from Schotte et al. (1996)

A conformational analysis using a systematic search method was performed for each compound in the set in order to identify low energy conformations for each active molecule. *Sybyl 6.4* from Tripos with Tripos force field and charges assigned to each atom according to the method of Gasteiger and Marsili, was used in all calculations of initial conformations.

### Results

Low energy conformations of the ligands were determined. Molecular superimposition techniques were used to identify which low-energy conformation from each set of molecular conformations was to be used in construction of the pharmacophore. The main criteria for

each conformation were that the selected pharmacophore elements superimpose within the set.

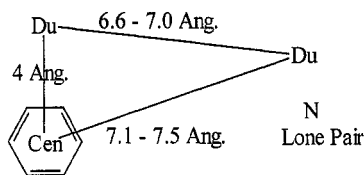
The resulting conformations were then run through the program GRID using a variety of probes to predict the points with the best interactions (Goodford 1984). This produces an array of energy values that can be used to generate three-dimensional contour surfaces at selected energy levels which gives additional information as to other possible areas of interaction.

## Discussion

All initial modelling was undertaken using the lone pair off the distal nitrogen and a region of hydrophobic interaction, such as a phenyl ring, as essential criteria (Petcher). The electron rich substituent off the phenyl ring is an optional substituent that increases activity as a D<sub>2</sub> antagonist. In the case of clozapine *versus* iso-clozapine where the electron rich substituent is moved from the non-interacting position 2 on clozapine to the interactive position 8 in iso-clozapine, the D<sub>2</sub> antagonist binding affinity changes from 330nM to 13nM respectively (Liao et al.). The position of the nitrogen lone pair from the atypical antipsychotic Org5222 also fits with the proposed pharmacophore. The GRID probe contours helped establish other areas of possible interaction. The mirror image of this proposed pharmacophore is also a viable pharmacophore.

## Future Directions

The quantitative aspect of electrostatic potential of the hydrophobic region in question, in relation to D<sub>2</sub> antagonist binding affinity will be investigated. A Neural Network approach to mixed receptor interaction of atypical antipsychotics is currently being examined, with a resultant atypical receptor ratio profile to be established.



Hydrophobic Region

## Bibliography

- Goodford, P.J., 1984, A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* 28, 849-57.
- Jaber, S.W. Robinson, Missale, C., Caron M.G., 1996, Review: Dopamine receptors and brain function. *Neuropharmacology*, 35: 1503-1519
- Liao, Y.; DeBoer, P.; Meirer, E.; Wikstrom, H., 1997, Synthesis and pharmacological evaluation of triflate-substituted analogues of clozapine: identification of a novel atypical neuroleptic. *J. Med. Chem.* 40, 4146-53.
- Liljefors, T., Pettersson, I., 1997, "Computer aided development of three-dimensional pharmacophore models". *A Textbook of Drug Design and Development*, Larson, P.K.; Liljefors, T.; Madsen, U. (editors), Publisher, Harwood Academic.
- Petcher, T.J., 1995, "Topology of dopamine receptors" in *The Role of Brain Dopamine* Fluckiger, E., Muller, E.E., Thorner, M.O., (Editors), Publisher, Springer Sandoz.
- Schmidt, C.J., Sorensen, S.M., Kehne, J.H., Carr, A.A., Palfreyman M.G., 1995, Minireview: The role of 5HT<sub>2A</sub> receptors in antipsychotic activity. *Life Sciences*, 56: 2209-2222
- Waddington, J.L., O'Callaghan, E., 1997, What makes an antipsychotic 'atypical'? Conserving the definition. *CNS Drugs*, 7: 341-6
- Schotte, P.F.M., Janssen, W., Gommeren, W.H.M.L., Luyten, P., Van Gompel, A.S., Lesage, De Loore K., Leysen J.E., 1996, Risperidone compared with new and reference antipsychotic drugs: *in vitro* and *in vivo* receptor binding. *Psychopharmacology*, 124: 57-73.



**Poster Session VI**  
**New Methods in Drug**  
**Discovery**

## GENETIC ALGORITHMS: RESULTS TOO GOOD TO BE TRUE?

M.G.B. Drew<sup>1</sup>, J.A. Lumley<sup>1</sup>, N.R. Price<sup>2</sup>, R.W. Watkins<sup>2</sup>

<sup>1</sup> Department of Chemistry, Reading University, Reading, RG6 6AD, UK

<sup>2</sup> Central Science Laboratory, MAFF, Sand Hutton, York, YO4 1LZ, UK

### Introduction

We are currently studying several distinct families of molecules with pesticide properties including cinnamic acids and anthranilates, which are used as bird repellents, and organophosphate insecticides. Experimental activity measurements were accurately obtained and have been used for QSAR/MFA studies. The genetic function algorithm (GFA) method was used and QSAR equations derived for each dataset that accurately reproduced the activity of the compounds. In each case the correlation and cross correlation coefficients were higher ( $r^2 > 0.90$ ) than those from conventional techniques, but problems of overfitting and statistical validity remain. Here, the optimum methodology is developed for using the GFA in developing successful QSARs.

### Technique, Results and Discussion

For each set of compounds, conformational analyses were done by comparison to known crystal structures, and by the use of conformational analysis in Quanta97<sup>1</sup>. Charge calculation was done with Gaussian94<sup>2</sup>. Without knowing the active conformation, lowest energy conformations were used. Geometry optimization at the *ab initio* level was done in Gaussian94 in some cases to ensure maximum accuracy.

Over 100 molecular, substituent and electronic descriptors were calculated via Cerius2<sup>1</sup>, TSAR, and single point Gaussian94 calculations in the 6-311G\* basis set. For descriptors cross-correlating over 70%, the one correlating least with the dependant variable was removed. This removed 75% of the descriptors to ensure minimal co-linearity.

For MFA, manual matching of key structural features was done to overlay all molecules, then some 2000 field points calculated for H<sup>+</sup>/OH/CH<sub>3</sub><sup>+</sup> probes, distributed randomly at approximately 1Å intervals. Analysis of the correlation matrix removed descriptors containing a high degree of co-linearity. The GFA/PLS method was used to select the optimum set of descriptors for use in multiple linear regression, but even when high  $r^2$  and  $r^2(\text{cv})$  values were obtained, there is still the possibility of chance correlation occurring.

It is not the number of terms in the derived QSAR equation that is the concern in

validating an equation, but more the number of descriptors originally screened to derive the equation<sup>3</sup>. The more descriptors used, and the fewer the observations in the training set, then the higher the mean  $r^2$  value, i.e.: larger probability of a chance correlation occurring. When deriving equations with powerful genetic algorithms, over-fitting of data becomes a large concern. Traditional statistical tests such as  $r^2(cv)$  are very useful<sup>4</sup>, but do not always pick up poor equations, and further validation must be done.

**Table 1.**  $r^2$  and  $r^2(cv)$  for best equations found in organophosphate QSAR/MFA studies

Data Set	GFA $r^2 / r^2(cv)$	GFA/PLS $r^2 / r^2(cv)$
QSAR data	0.943 / 0.905	0.943 / 0.905
Scrambled activity, and QSAR data	0.944 / 0.888	0.907 / 0.826
MFA data	0.939 / 0.896	0.992 / 0.869
Scrambled activity, and MFA data	0.953 / 0.933	0.918 / 0.852

In our first study we successfully derived a GFA/QSAR for the repellent activity of 14 cinnamic acids. The resulting four-term linear regression equation was based on electronic descriptors, confirming a Hammett-style relationship previously seen to work. We also report here (Table 1) the results from an QSAR/MFA study of 20 organophosphates, taken from a previous study<sup>6</sup>, which had been unable to predict their activity using one equation. Using the GFA/PLS method we have successfully derived a spline function equation using 4 descriptors with an  $r^2=0.943$  and  $r^2(cv)=0.905$ . No equation has yet been derived which predicts organophosphate activity so well, but we originally screened some 30 descriptors for only 20 observations so how do we know this relationship is not due to chance. It is suggested<sup>3</sup> that for 20 observations you only need 14 variables to get a chance correlation of  $>0.9$ , though this may be a little lower in practice. This probability statistic is for linear regression, but a spline function may increase the probability of a chance correlation. As seen from our results, confidence in our final equation is only gained by considering the highest derived chance equation, and by using the GFA/PLS technique.

Thus we have found activity data scrambling, and using external data sets to validate training set derived QSARs, extremely important verification techniques. Our equation is really only valid if  $r^2$  and  $r^2(cv)$  values are distinctly higher than the mean calculated values. We also stress the importance of considering the descriptors used, to assess why they are working as predictors. We must consider what information we can glean from them as chemists, as opposed to using them as a blind prediction tool. Confidence increases if the descriptors used are describing properties expected to control activity. In some cases it may be more appropriate to use several smaller predictive equations to explain large variation in a training-set rather than turning to more powerful QSAR analysis and searching techniques. This is seen for the organophosphates, which are successfully described with partitioning of descriptors rather than trying to find more complex descriptors or techniques, which has so far been unsuccessful.

JAL thanks the EPSRC (UK) for funding, and the UoR MMRG for constant input.

1. Quanta97, Cerius2 3.5, July 1997, Molecular Simulations Inc., San Diego, Calif. U.S.A.
2. Gaussian, Inc. Carnegie Office Park, Building 6, Pittsburgh, PA 15106 U.S.A.
3. J.G.Topliss and R.P.Edwards, Chance factors in studies of QSAR, *J.Med.Chem.* 22, 10, 1238, (1979).
4. R.D.Cramer III, J.D.Bunce and D.E.Patterson, Crossvalidation, bootstrapping and PLS compared with multiple linear regression in conventional QSAR studies, *Quant. Struct.-Act. Relat.* 7, 18 (1988).
5. G.Schuurmann, Do hammett constants model electronic properties in QSARs?, *Sci.Total Env.* 109/110, 221-235, (1991).
6. J.D.Bruijn and J.Hermens, Inhibition of acetylcholinesterase and acute toxicity of organophosphorous compounds to fish: a preliminary QSAR study, *Aquat. toxicol.* 24, 257, (1993).

## PROPERTY PATCHES IN GPCRS: A MULTIVARIATE STUDY

Per Källblad and Philip M. Dean

Drug Design Group, University of Cambridge, Department of Pharmacology  
Tennis Court Road, Cambridge CB2 1QJ, United Kingdom

The human genome project has accumulated primary sequence data for several hundred members of the G-protein-coupled receptor (GPCR) superfamily. The lack of structural information forces the development of new ways to obtain functional and structural knowledge directly from protein sequences. An attempt is described here to convert multiple sequence alignments into more easily interpretable property data. The main aim of the study was to find residues on adjacent secondary structures that share principal properties and might therefore be involved in inter-helical contacts.

194 GPCR sequences from 10 rhodopsin-like receptor sub-families including amine, peptide and nucleotide receptors were retrieved from Release 35.0 of the SWISS-PROT database<sup>1</sup> (Table 1). Rhodopsin is the closest relative for which structural data are available and is included because of its importance as a reference for homology modellers. The inter-helical loop regions are unlikely to be part of the common 3D architecture shared by rhodopsin-like receptors and were therefore excluded from the analysis. Multiple sequence alignments were produced using CLUSTALW<sup>3</sup> and thereafter edited manually to exclude non-helical regions, eliminate gaps in transmembrane regions and ensure the correct positioning of all conserved residues. The helical regions selected for each helix are adapted from Baldwin *et al.*<sup>2</sup>.

**Table 1:** Families included in the study

	5ht	aar	acm	ade	bar	dr	hh	nyr	opr	ssr	rho
No. sequences	43	22	16	17	15	17	9	15	16	16	8
No. sub-types	13	6	5	4	3	6	2	5	4	5	1

26 physico-chemical variables were used to describe each amino acid<sup>4</sup>. The descriptors were scaled to unit variance and compressed through Principal Components Analysis (PCA) to give 5 principal properties. The principal properties roughly correspond to hydrophobicity (1), steric (2) and electronic properties (3–5) and explain 91% of the variation in the original

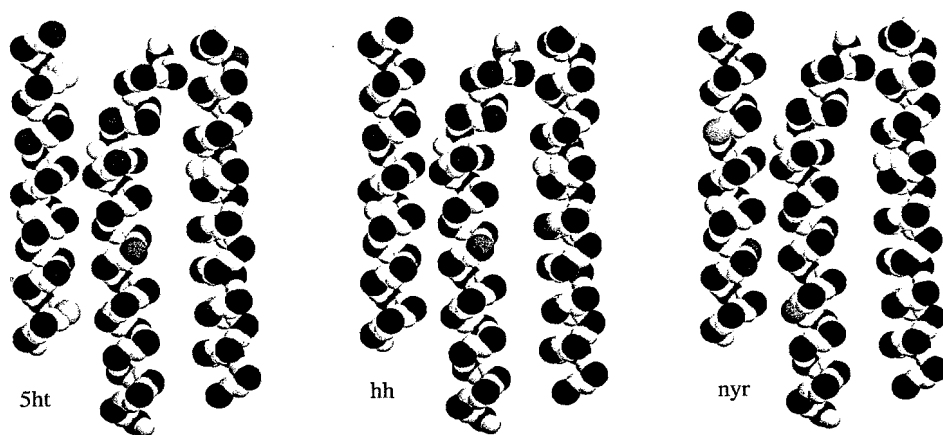
physico-chemical descriptor set. A property matrix was generated for each alignment by substituting every amino acid with its descriptors. Each matrix was compressed through PCA to give the final matrix to be used for the alignment position classification (Table 2).

**Table 2.** PCA results

	5ht	aar	acm	ade	bar	dr	hh	nyr	opr	ssr	rho
No. PCs	13	8	6	8	7	8	6	9	6	7	6
R <sup>2</sup> X(cum) <sup>1</sup>	0.87	0.93	0.94	0.87	0.94	0.95	0.91	0.84	0.90	0.88	0.98

<sup>1</sup>R<sup>2</sup>X(cum) is the cumulative sum of squares of all the variables in the original data matrix (X) explained by the principal components (PCs).

Clustering using the fuzzy *c*-means algorithm<sup>5</sup> was performed to find groups of alignment positions that share similar properties. In fuzzy clustering the different objects are assigned membership values between 0 and 1 for each cluster and can hence belong to more than one cluster. This type of clustering was chosen for its advantage in clustering hybrid objects which is the case for multiple sequence alignment positions. The calibration of fuzziness and distance measure was made through Partial Least-Squares projection to Latent Structures (PLS) between the original sequence property matrix and the cluster membership values of the different alignment positions. A fuzziness factor of 1.10 together with the Mahalanobis distance was chosen to obtain maximal chemical significance. The use of 8 clusters gives a separation of biochemically conserved positions into relevant groups and enables the identification of positions with a high level of property variation as non-members. Non-members are defined as objects with membership values below 0.50 in all of the clusters. Using 8 clusters, the fraction of non-members varies from 1% to 20% depending on the size of the alignment. Property class membership values of alignment positions were projected onto the suggested 3D structure for examination of spatial distribution (Figure 1).



**Figure 1.** Examples of cluster membership values displayed on the C $\alpha$  atoms of the proposed 3D structure of rhodopsin<sup>2</sup>. The members of the main hydrophobic cluster are shown for helices 5, 6 and 7 (right to left) for six of the alignments. The colours range from dark grey (membership value = 0) to light grey (membership value = 1). Backbone atoms are included in very light grey. The view is from inside the helical bundle.

The quality of the clustering was validated through PLS between the original sequence property matrices and the cluster membership matrices of the different alignments (Table 3). The PLS also enabled identification of alignment positions with poor fit to the statistical model. A further biochemical validation was performed by extracting the alignment positions of each cluster and comparing them with those of the other alignments. It was observed that the clusters from the different alignments share properties and conserved members.

**Table 3.** PLS results

	5ht	aar	acm	ade	bar	dr	hh	nyr	opr	ssr	rho
No. PCs	7	7	7	5	6	7	6	7	7	7	5
R <sup>2</sup> X(cum)	0.68	0.87	0.90	0.84	0.89	0.81	0.82	0.77	0.92	0.88	0.98
R <sup>2</sup> Y(cum)	0.66	0.64	0.55	0.66	0.59	0.62	0.57	0.66	0.69	0.65	0.56
Q <sup>2</sup> (cum) <sup>1</sup>	0.57	0.57	0.51	0.59	0.52	0.54	0.51	0.58	0.59	0.55	0.50

<sup>1</sup>Q<sup>2</sup>(cum) is the fraction of the total variation in the two data sets that can be predicted by the principal components (PCs).

The method developed enables identification of groups of multiple sequence alignment positions with a fine level of property variation that is difficult or impossible to detect through "sequence-gazing". Projected onto the proposed 3D structure, the cluster membership values provide a way of displaying biochemical properties conserved throughout a multiple sequence alignment and may help in the identification of contact points between different sub-units.

## Acknowledgements

Per Källblad is grateful to Rhône-Poulenc Rorer for a studentship. Philip M. Dean is a Wellcome Principal Research Fellow.

## References

1. A. Bairoch and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TREMBL, *Nucleic Acids Res.*, 25:31–36 (1997).
2. J.M. Baldwin, G.F.X. Schertler and V.M. Unger, An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors, *J. Mol. Biol.*, 272:144–164 (1997).
3. J.D. Thompson, D.G. Higgins, and T.J. Gibson, ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22:4673–4680 (1994).
4. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström and S. Wold, New chemical descriptors for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *J. Med. Chem.*, 41:2481–2491 (1997).
5. J.C. Bezdek, 1981, *Pattern Recognition With Fuzzy Objective Function Algorithms*, Plenum Press, New York.

# A STOCHASTIC METHOD FOR THE POSITIONING OF PROTONS IN X-RAY STRUCTURES OF BIOMOLECULES

M. Glick and Amiram Goldblum

Department of Medicinal Chemistry, School of Pharmacy, Hebrew University of Jerusalem, Jerusalem, ISRAEL 91120

**Introduction:** Inclusion of all hydrogen atoms in protein and nucleic acid models is necessary for a more accurate representation of biological systems during molecular dynamics simulations, and for understanding molecular recognition. This is especially important for polar hydrogens that play a critical role in determining secondary structure through hydrogen bonds. At present, X-ray crystallography is still not efficient for locating the proton positions. Neutron diffraction studies that can locate the protons are quite rare and only a few combined x-rays/neutron diffraction studies have been deposited in the protein data bank (PDB).

Most molecular modeling packages places hydrogens in a non specific manner and any subsequent step suffers from the multiple minima problem. Another method (Brunger and Karplus) employs an iterative process of energy minimizations applied to the torsion angle of each polar hydrogen in its environment. This method is suitable for systems in which close contacts between hydrogens are absent. A third method suggested by Ornstein et al. divides the system into groups of interacting hydrogen bond donors and acceptors called networks. The algorithm maximizes the number of hydrogen bonds in each network and minimizes the total distance between donors and acceptors. The number of comparisons scales with the factorial of the number of elements in the network-A fact that limits the calculation to small networks. In addition, this approach is not based on energy evaluation criteria. As a result, the output might contain high energy interactions between the located hydrogens and their environment.

We present a novel energy based method for the location of polar hydrogens. It requires the division of the full system into networks, but has the following advantages: Each network is evaluated by energy criteria, and the code can handle large biological systems defined as one huge network. The code was designed so it can easily be modified to handle any force field.

**Method:** Coordinates are read from a PDB file. Hydrogens and lone pairs which are to be added are divided into two categories: (1) Trivial hydrogens/lone pairs-those that may be located using the hybridization of heavy atoms. (2) Non trivial hydrogens/lone pairs: those that have rotational degrees of freedom, such as Ser, Thr and Tyr hydroxyls, water, etc. Non trivial hydrogens and lone pairs are divided into ensembles: groups which interact among themselves. Each ensemble is treated separately. The energy criterion used to evaluate the quality of each combination is a non bonded energy function with Lennard-Jones (6-12) and coulombic interactions. With a large ensemble cutoff, the user can force the program to handle the system as one huge ensemble.

It is obvious that in case of a large biological system constituting a single ensemble, we face a very large combinatorial problem. In trypsin (1NTP), for example, there are  $5.84 \cdot 10^{39}$  alternative combinations. To reduce the size of the problem, we developed a unique stochastic approach. For each non trivial hydrogen or lone pair there are usually a few alternative locations, but only one would give the lowest energy.

Let  $X=(X_1, X_2, \dots, X_{d_0})$  be a configuration of  $d_0$  segments in one ensemble. For each configuration  $X$ , the energy  $E=E(X)$  is calculated. The objective is to find the configuration which minimizes  $E$ . Since it is impossible to evaluate all the alternative configurations due to the large number of combinations, we follow the steps: 1. Sample at random  $n$  configurations out of the large population of combinations  $X_1=(x_{11}, x_{12}, \dots, x_{1d_0}), \dots, X_n=(x_{n1}, x_{n2}, \dots, x_{nd_0})$ . Compute the corresponding energy values:  $E_1 = \sum (e_{1j})$  ( $j = 1, d_0$ ) for configuration  $X_1$ ,  $E_n = \sum (e_{nj})$  ( $j = 1, d_0$ ) for configuration  $X_n$ ; 2. Construct the

distribution  $F_E^n$ .  $F_E^n$  is an assembly of energies that corresponds to  $n$  sampled configurations. Define cutoff points  $H$  and  $L$  in  $F_E^n$ .  $H$  contains all the configurations that satisfy the condition:  $E_i \geq F_E^n(1-\alpha)$  where  $F_E^n(\alpha)$  is the  $\alpha$ -th percentile of  $F_E^n$ , and  $L$  contains all configurations satisfying  $E_i \leq F_E^n(\alpha)$ . The number of configurations in each of  $H$  and  $L$  is  $n_0 = n * \alpha$ ; **3.** Construct vectors  $h$  and  $l$  for the positions in configurations corresponding to the energies in  $H$  and  $L$ , respectively. The vector  $h$  is the element-wise intersection of all the configurations in  $H$ : if all configurations in  $H$  share the same value, say 5->1, at component  $j$ , (corresponding to  $x_{nj}$  of configuration  $X_n$ ) then  $h_j = 5 \rightarrow 1$ ; otherwise,  $h_j = 0$  (no common position for segment  $j$  in all high energy configurations) The vector  $l$  is constructed similarly from  $L$ . Using values of  $n = 1000$  and  $\alpha = 0.004$  was chosen as a reasonable compromise that satisfies the probability of obtaining excellent results with relatively short computation times; **4.** Compare  $h$  and  $l$ . If both  $h_j$  and  $l_j$  have a similar vector component,  $j$ , it will remain as a viable configuration for that segment, because it contributes also to low energy values. However, if  $h_j \neq l_j$ , then the corresponding segment component  $h_j$  will be evicted from subsequent iterations; **5.** Repeat steps 1 to 4 for the reduced location-space until the number of possible configurations is smaller than a user defined threshold; **6.** Compute exhaustively all the remaining configurations to find the best one.

**Results and discussion:** We tested our algorithm on four high resolution crystal structures: Bovine Pancreatic Trypsin Inhibitor (Brookhaven Protein Data bank file 5PTI), RNase-A (file 5RSA), Trypsin (file 1NTP), and Insulin (file 3INS), for which the neutron diffraction structures are available. We tested our program both as a minimization and polar hydrogen addition tool. We removed all the hydrogens (and deuterium atoms) from the PDB file and activated the algorithm to reconstruct their location. Each system was treated by two variations of the method and was compared to a "self consistent" approach. Energy criteria were applied in all three variations. **1.** Combined "Ensemble-stochastic approach": All possible combinations in an ensemble are evaluated, and the one with the lowest energy is the result. In ensembles with a very large combinatorial demand the "stochastic approach" was activated to reduce the number of combinations. The calculation on 3INS by this method is interactive on a Silicon Graphics R10000 machine and takes about 30 seconds. However, this approach requires an approximation of distances between non trivial hydrogens lone pairs in different ensembles and the accuracy is thus somewhat reduced. **2.** Pure "stochastic approach": This approach suffers from a large CPU demand: The calculation on 3INS takes about 15 minutes on a Silicon Graphics R10000 machine. Results are however better than with the other methods (lower minimal energy values). **3.** Self consistency: Rotations of consecutive separate segments to the minimum of each. This approach has the lowest CPU demand of the three methods. The calculation is then reiterated from beginning to end until self consistency is achieved. It may start with another segment rather than the first. The results are higher in energy than the other two methods.

RMS values (theoretical vs. experimental) are low (0.3-0.65 for the different proteins). The overlaying of predicted on experimental structures reveals that most of the inconsistent results stem from rotatable hydroxyls on the surface of the proteins, where water molecules plays a role in determining their positions, but these waters were not included in the PDB structure.

## REFERENCES

1. Axel T. Brunger and Karplus M. *Polar Hydrogen Positions in Proteins: Empirical Energy Placement and Neutron Diffraction Comparison.* *Proteins* 4:148-156 (1988).
2. Michael B. Bass, Derek F. Hopkins, W. Andrew N. Jaquysh and Rick L. Ornstein. *A Method for Determining the Positions of Polar Hydrogens Added to a Protein Structure That Maximizes Protein Hydrogen Bonding.* *Proteins* 12:266-277 (1992).



## MOLECULAR FIELD TOPOLOGY ANALYSIS (MFTA) AS THE BASIS FOR MOLECULAR DESIGN

Eugene V. Radchenko, Vladimir A. Palyulin, Nikolai S. Zefirov

Department of Chemistry, Moscow State University  
Moscow 119899 Russia

Modern studies of quantitative structure-activity relationships (QSARs) for organic compounds seek to reveal the structural features responsible for the interaction of molecules of an active compound (ligand) with a biological target. From the viewpoint of both the ligand/target fit and the possible design of new structures, it is clear that location of these features with respect to the molecule is as important as their character. A number of approaches taking into account the location of such features based on the topological as well as the spatial representation of structures was suggested in literature. However, all of them have some drawbacks with respect to generality and/or applicability to conformationally flexible compounds.

Recently we proposed<sup>1,2</sup> as the generalization of the previous approaches the *Molecular Field Topology Analysis* technique which may be considered as a 'topological CoMFA analogue'. Our previous investigation of some datasets suggests that the full 3D set of parameters is often redundant and might introduce additional noise. Thus, the topological alignment is employed in MFTA method which leads to models that are often comparable or even superior in quality to those based on other widely used QSAR approaches. The method could be regarded as complementing the existing techniques such as CoMFA. It is especially suitable for solving the problems where the analysis of 3D structure is either unnecessary or complicated.

In the framework of MFTA for the structures of the training set the molecular supergraph (a not necessarily minimal graph such that any structure of the set can be represented as its subgraph) is constructed. Crucial structural features can be quantitatively represented as the local physico-chemical parameters of the compound, that is, various characteristics of atoms and bonds. One might expect that the distribution pattern of these parameters for active compounds would reflect the complementary features of a target. Generally, the interaction cannot necessarily be attributed to a few key positions within a ligand molecule due to the correlation between the parameters in the neighbouring positions, possible involvement of the entire regions and the ability of the system to accommodate to certain variations. The technique allows the use of the open descriptor set. Currently implemented descriptors include electrostatic and steric parameters as well as the

characteristics of lipophilicity, hydrogen bonding, atom and bond presence.

The mapping procedure employed in a supergraph generation and descriptor vector formation is very flexible and allows taking into account the atom type, valence state, stereochemistry, bond type and order and special user requirements as well as the similarity of local properties distribution over the structure. Thus the mechanistically sound rather than formal mapping can be achieved. One of the algorithms combines the features of vertex-by-vertex expansion approach and the algorithm of searching for maximum cliques (complete subgraphs) of the module graph product and efficiently finds the maximum connected graph intersections. As another option, the non-deterministic (genetic) algorithm is also used to search for the graph intersections. During the construction of the MSG, the structures from the training set are processed sequentially. At each step, the intersection between the MSG constructed by this time (originally empty) and the next structure of the set is determined. Then, the MSG is augmented by the atoms and bonds that do not occur in the intersection, and the values of local properties are updated for all the MSG vertices.

The descriptor vector for the compound is then formed by taking, for each position in the supergraph, the values of the local descriptors on the corresponding atom in the compound. It is possible to select the best of several mappings or use the averaged descriptor vector on the basis of the descriptor difference between the current and reference (most potent) structure. Since the number of descriptors is rather large (though much smaller than in CoMFA), the partial least squares (PLS) regression is used to analyze the descriptor-activity relationships. As a result, the quantitative characteristic of the influence of each descriptor in each position, including common structural fragments, on activity can be determined. Subsequent selection of variables based on their impact on model output or predictivity is possible, enabling the rational identification of key structural features for the design of potentially more active structures and for use as anchor points in 3D alignment. The application of the method to a number of well-known 'benchmark' cases often leads to the models comparable or superior in terms of fitting and prediction quality to both conventional and 3D QSAR techniques.

Several approaches to the design of novel potentially active structures based on the MFTA models can be formulated. First, we can perform an exhaustive structure generation from a common fragment and the substituents built from a number of elementary fragments that are present in the training set and/or can be easily introduced synthetically. Then, the structures with the desired activity values are selected using the predictive model. Alternatively, it is possible to construct the prospective structures directly from the MFTA molecular supergraph and substructural templates taking into account the effect of local descriptors on the activity. The activity value for the complete structure is subsequently verified by predicting it from the model. The third option is based on the fact that any structure for which the reliable prediction can be expected may be represented as a subgraph of the MSG. Thus, we can use the genetic algorithm to propose the optimal structures by searching for the optimal labeling of MSG vertices and edges.

The authors gratefully acknowledge support from Russian Foundation for Basic Research and Russian Federal Program "Development of new drugs by means of chemical and biological synthesis".

## References

1. N.S. Zefirov, V.A. Palyulin, and E.V. Radchenko, Molecular field topology analysis in studies of quantitative structure-activity relationships for organic compounds, *Doklady Chemistry* 352:23-26 (1997).
2. V.A. Palyulin, E.V. Radchenko, and N.S. Zefirov, Molecular field topology analysis (MFTA) method in QSAR studies of organic compounds, *J. Chem. Inf. Comput. Sci.* (1998), submitted for publication.

## **RANK DISTANCE CLUSTERING - A NEW METHOD FOR THE ANALYSIS OF EMBEDDED ACTIVITY DATA**

**John Wood<sup>1</sup> and Valerie S. Rose<sup>2</sup>**

**<sup>1</sup>Department of Health Sciences & Clinical Evaluation  
University of York, Heslington, York YO1 5DD, UK**

**<sup>2</sup>BioFocus plc, Sittingbourne Research Centre  
Sittingbourne, Kent ME9 8AZ, UK**

### **INTRODUCTION**

'Embedded' activity data describes the situation where active compounds cluster together, with inactives dispersed. There is thus a centre of activity and moving away from this centre results in a decrease in activity. This may be observed, for example, in a plot of molecular weight against log P where, to retain activity, compounds must fall in a specific size and hydrophobicity window. From our experience, embedded relationships tend to occur in complex biological test systems such as cellular or *in vivo* assays.

2D embedded relationships are readily detected using 2D plots for all pairs of descriptors. The situation becomes more complex in wide multivariate data sets and those sets requiring more than 2 properties to define the active cluster. Methods such as Cluster Significance Analysis (CSA)<sup>1</sup> or SIMCA<sup>2</sup> have been applied to such problems where the activity data is classified as active or inactive.

More recently, we have described several novel methods developed to extend the variety of situations in which embedded relationships can be detected and which allow the activity data to be a quantified measure rather than an active/inactive classification. One class of methods, Single Class Discrimination,<sup>3,4,5</sup> identifies informative latent axes in the data set, while the other class, consisting of extensions to CSA,<sup>6</sup> identifies individual descriptors or low dimensional descriptor combinations which result in clustering of active compounds relative to the data set as a whole.

### **RANK DISTANCE CLUSTERING**

A potential drawback of the existing CSA algorithms for large datasets is the amount of computing required, because of the necessity of generating the permutation distribution to approximate probability values for the various models. This problem is

overcome in Rank Distance Clustering (RDC) by using a different way of testing for statistical significance, based on the ranks of distances rather than their actual values. Like CSA, RDC is used to identify low dimensional descriptor sub-spaces in which active compounds cluster. Although the distances are replaced by their ranks in RDC, the properties of Euclidean distance - in particular Pythagoras' theorem - lie behind the method as it stands. Thus considerations of scaling and the like affect RDC in much the same way as they do CSA. RDC is currently restricted to the analysis of classified activity data.

## BRIEF OUTLINE OF THE RDC METHOD

Both an all-combinations and a forward stepwise algorithm have been developed and programmed in SAS. The algorithm proceeds as follows :

1. *Optional scaling (e.g. autoscaling) of descriptors*
2. *Determine the means of the descriptors for the active set*
3. *Centre the data matrix to this mean*
4. *Construct the inter-sample squared distance matrix*
5. *Rank distances in ascending order*
6. *Sum the ranks of the active class*
7. *Use a Mann-Whitney test to see whether the sum of ranks of the actives is significantly smaller than expected*
8. *Determine the marginal significance of adding a new term to the model - the Z-score (forward stepwise method only)*

The new method is significantly faster than CSA as it is not dependent on generating random permutations to test for significance. It is also better suited to handling large datasets and may prove more robust to outliers than CSA as it uses ranks rather than Euclidean distances directly.

## ACKNOWLEDGEMENTS

This work was undertaken as part of a BBSRC project for *Improved Mathematical Methods in Drug Design*. The authors are grateful to the BBSRC, the industrial partners (Glaxo Wellcome, Unilever, Molecular Simulations Inc, Astra Charnwood and Smithkline Beecham) and the academic partners (The University of Portsmouth and The Institute of Food Research, Reading) of the project for funding, support and valuable discussion.

## REFERENCES

1. J.W. McFarland and D.J. Gans, On the significance of clusters in the graphical display of structure-activity data, *J. Med. Chem.* 29:505 (1986).
2. S. Wold, Pattern-recognition by means of disjoint principal components models, *Pattern Recognition* 8:127 (1976).
3. V.S. Rose, J. Wood and H.J.H. MacFie, Single class discrimination using principal component analysis (SCD-PCA), *Quant. Struct.-Act. Relat.* 10:359 (1991).
4. V.S. Rose, J. Wood and H.J.H. MacFie, Generalized single class discrimination (GSCD). A new method for the analysis of embedded structure-activity relationships, *Quant. Struct.-Act. Relat.* 11:492 (1992).
5. J. Wood, V.S. Rose and H.J.H. MacFie, Significance testing of single class discrimination models, *Chemometrics & Intell. Lab. Systems*, 23:205 (1994).
6. V.S. Rose and J. Wood, Generalized cluster significance analysis and stepwise cluster significance analysis with conditional probabilities, *Quant. Struct.-Act. Relat.* (1998) in press.

# THE APPLICATION OF MACHINE LEARNING ALGORITHMS TO DETECT CHEMICAL PROPERTIES RESPONSIBLE FOR CARCINOGENICITY

C. Helma<sup>1,2</sup>, E. Gottmann<sup>2</sup>, S. Kramer<sup>3</sup> and B. Pfahringer<sup>3</sup>

<sup>1</sup>Institute for Tumor Biology–Cancer Research, Borschkegasse 8a, A-1090 Vienna

<sup>2</sup>Institute for Environmental Hygiene, Kinderspitalgasse 15, A-1090 Vienna

<sup>3</sup>Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna

## Introduction

The prediction of carcinogenicity based on chemical structures is one of the most challenging problems in predictive toxicology. This task is extremely difficult, because carcinogens act by many different mechanisms and they do not share common structural attributes.

Information about the chemical structures can be provided as structural fragments (e.g. functional groups, structural alerts) or as physical or chemical properties (e.g. volume, mass, logP, HOMO, LUMO ...). Previous approaches have focussed on one of these representations only. The purpose of this work was to evaluate the suitability of machine learning programs for the prediction of complex toxicological effects and to perform a systematic comparison of different sets of descriptors within a single framework. Structural Regression Trees (SRT), an algorithm from the field of Machine Learning, can handle both representations easily and is therefore especially suited for this comparison.

## Methods

Carcinogenicity classifications for rodents (rats and mice) were obtained from the NCI/NTP part of the Carcinogenic Potency Database (CPDB) compiled by Gold et al.[1]. Compounds without defined chemical structures (e.g. mixtures) were excluded from this study.

The structural information (connectivity) for the CPDB compounds was derived from SMILES strings and encoded as Prolog facts. Physical/chemical properties and shape indices were calculated with MOPAC and TSAR (Oxford Molecular).

Structural Regression Trees SRT [2] was used to induce general theories (about factors affecting the carcinogenicity of compounds) from the given CPDB examples. The SRT theories were quantitatively validated by 10-fold cross-validation, and summarized in terms of predictive accuracy.

## Results

Table 1 summarizes the predictive accuracies for both descriptor sets after 10-fold cross validation. The combination of physical/chemical and structural fragment descriptors led to a significant improvement of the predictive accuracy of the SAR model. The performance of this model was better than the performance of other carcinogenicity structure-activity relationships for noncongeneric compounds reported in the literature (typical predictivity: 60-70%). This is an indication, that both types of descriptors should be used in SAR models for noncongeneric compounds. The Inductive Logic Programming algorithm SRT provides a flexible framework with the ability to use relational information (e.g. chemical structures). It generates regression trees which are, in terms of predictive accuracy, competitive to other types of SAR models. In contrast to other regression or neural network models, SRT provides rules for which are easily interpretable by toxicological experts, and may therefore lead towards a better understanding of the mechanisms of rodent carcinogenicity.

Table 1. Predictive Accuracy for Rodent Carcinogenicity Classifications after 10-Fold Cross Validation

structural fragments	Descriptors	
	physical/chemical	combined
65.4%	67.3%	75.2%

## Acknowledgements

This work was supported by the Austrian Ministry of Science and Transport and the "Jubiläumsfond der Österreichischen Nationalbank".

## References

- [1] L.S. Gold et al. Sixth plot of the carcinogenicity potency in the general literature 1989 to 1990 and by the National Toxicology Program 1990 to 1993. *Environ. Health Perspect.*, 103 (Suppl7):1-122, 1995.
- [2] S. Kramer. Structural regression trees. In *Proc. Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, Menlo Park, 1996. AAAI Press.
- [3] D. Weininger. SMILES, a chemical language and information system 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31-36, 1988.

# STUDY OF GEOMETRICAL/ELECTRONIC STRUCTURES - CARCINOGENIC POTENCY RELATIONSHIP WITH COUNTERPROPAGATION NEURAL NETWORKS

Marjan Vračko

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia  
Phone: +386 61 1760315, Fax: +386 61 1259244, E-mail: marjan.vracko@ki.si

## INTRODUCTION

In the last years the artificial neural networks (ANN) are becoming an important tool in chemistry-related non-linear modelling. In a presented contribution we applied ANN with counter-propagation (CP) learning strategy. CP ANN learning strategy represents an extension from unsupervised (Kohonen) to supervised learning strategy. Details are given by *Hecht-Nielsen*<sup>1</sup>. CP ANN seem to be a proper tool in quantitative structure-property relationship (QSPR) studies particularly when this relationship is weak. This is certainly true for the structure-carcinogenic potency relationship. The 'carcinogenic potency', which is given as a dose where a particular compound causes a cancer, defines just a biological endpoint, but barely describes the mechanism of cancer on biochemical level. The data sets taken for modelling usually consist of diverse compounds that cause a cancer by different mechanisms. It seems, that for modelling with compounds acting over different mechanisms the ANN are superior to linear methods.

## DATA DESCRIPTION

Set of 58 compounds was treated in this study. All of compounds are amino derivatives, most of them have a benzene ring as a common substructure. (The compounds are: 2-chloro-p-phenyldiamine; 2,6-dichloro-p-phenyldiamine; 2-nitro-p-phenyldiamine; 2,4-xylidine.HCl; 2,5-xylidine.HCl; 2,4,6-trimethylaniline.HCl; 2-acetylaminofluorene; 2-aminoanthraquinone; 2-aminodiphenylene oxide; af2; 2-amino-4-(5-nitro-2-furyl)thiazole; aniline.HCl; methotrexate; azobenzene; benzidine.2HCl; chloramben; chlorambucil; 4-chloro-m-phenyldiamine; p-chloroaniline; 4-chloro-o-phenyldiamine; m-cresidine; 3-chloro-p-toluidine; 2,4-diaminoanisole sulfate; pyrimethamine; 3-(3,4-dichlorophenyl)-1,1-dimethylurea; p-nitrosodiphenylamine; 2,4-dimethoxyaniline.HCl; C.I.Disperse yellow 3; formic acid 2-[4-(5-nitro-2-furyl)-2-

thiazolyl]hydrazide; fluometuron; 5-nitro-2-furaldehyde semicarbazone; furosemide; 2-hydrazino-4-(p-aminophenyl)thiazole; 2-hydrazino-4-(5-nitro-2-furyl)thiazole; melphalan; melamine; 1-amino-2-methylanthraquinone; 4,4'-methylenebis(2-chloroaniline).2HCl; 4,4'-methylene- dianiline.2HCl; metronidazole; 4,4'-methylenbis(N,N-dimethyl)benzenamine; 5-nitro- acenaphthene; 5-nitro-o-anisidine; N-[4-(5-nitro-2-furyl)-2-thiazolyl]formamide; nitrophen; nithiazide; o-anisidine.HCl; o-aminoazotoluene; o-phenylendiamine.2HCl; p-cresidine; p-phenylendiamine.2HCl; proflavine.HCl hemihydrate; N-nitrosodiphenylamine; 2,2'-5,5'-tetrachlorobenzidine; 2,4-diaminotoluene.2HCl; m-toluidine.HCl; o-toluidine.HCl; mexacarbate.) Carcinogenic potency values given as TD<sub>50</sub> dose for mice were taken from CPDB<sup>2</sup>.

## MODELS AND RESULTS

Three models (A, B, C) built with different descriptors are compared in this study.

A) "Spectrum like representations of 3D structures<sup>3</sup> with atomic charges" were taken as descriptors.

B) "Spectrum like representations of 3D structures with atomic charges" plus calculated log D values were taken as descriptors. Log D values were calculated with HazardExpert<sup>4</sup> program.

C) Model built with different physico-chemical descriptors included log D values.

Models were tested with the one-leave-out cross validation method. The correlation coefficients (r) and the parameters of line predicted versus experimental values (b<sub>0</sub>, b<sub>1</sub>) are shown in Table 1. The results of all three models are similar, however, the best results were obtained with the model B. This is mostly due the fact that the model B gives good prediction values for some of compounds which are outliers in the models A and C. Such examples are 2-aminoanthraquinone (outlier in model A) and furosemide (outlier in model C). It was shown that the quantities calculated with expert system (HazardExpert) can improve the quality of QSPR models<sup>5</sup>.

**Table 1.** Statistical parameters for models A, B and C.

A	B	C
r=0.71, b <sub>0</sub> =1.05, b <sub>1</sub> =0.64	r=0.76, b <sub>0</sub> =0.95, b <sub>1</sub> =0.65	r=0.63, b <sub>0</sub> =1.39, b <sub>1</sub> =0.59

## ACKNOWLEDGMENT

This work was supported by EU project COPERNICUS CP94 1029 and additionally by Ministry of Science and Technology of R Slovenia under contract J1-5014-104. Author also thanks organizers of QSAR'98 conference for the grant.

## REFERENCES

- 1.R. Hecht-Nielsen, Counter propagation networks, *Appl. Optics* 26:4979 (1987).
- 2.L. S. Gold et al, Sixth plot of the carcinogenic potency database: results of animal bioassays published in the general literature from 1989 to 1990 and by the National toxicology program 1990 to 1993. *Environ. Health Persp*, 103, Suppl.8: 3 (1995).
- 3.J. Zupan, M. Novič , General type of a uniform and reversible representation of chemical structures, *Anal. Chim. Acta* 348:409 (1997).
- 4.HazardExpert, © 1995 CompuDrug Chemistry Ltd.
- 5.G. Gini, E. Benfenati, F. Darvas, J. Zupan, M. Tichy COPERNICUS CP94 1029, *Development of Second Generation Expert Systems for Environmental Toxicology*, Final Technical Report, 1998.



# COMBINING MOLECULAR MODELLING WITH THE USE OF ARTIFICIAL NEURAL NETWORKS AS AN APPROACH TO PREDICTING SUBSTITUENT CONSTANTS AND BIOACTIVITY

Igor I. Baskin, Svetlana V. Keschtova, Vladimir A. Palyulin,  
Nikolai S. Zefirov

Department of Chemistry, Moscow State University, Moscow 119899 Russia

## INTRODUCTION

Nowadays neither molecular model, no matter how elaborate it may be, is able to encompass all possible interactions, in which a real chemical/biological system is involved, as well as to take them properly into account. In this connection the problem of relating theoretically derived molecular characteristics with experimentally observed properties becomes very important. As a way of solving this problem, we see the use of a technique that would allow one to reveal nonlinear relationships of any complexity between theoretically derived characteristics of molecules and observed experimental properties. As the most promising candidate for that, we consider the use of artificial neural networks, since only this approach allows to find relationships of any complexity between parameters without the need to know in advance or guess its generic form.

## MAIN RESULTS

In this study, artificial neural networks were used to correlate parameters derived from semiempirical quantum-chemical treatment of specially designed model compounds (which consist of some common fragments with substituents attached to them) with the values of substituent constants ( $\sigma^m$ ,  $\sigma^p$ , F, R,  $E_s$ ) over a wide range of diverse substituents. Model compounds were formed by attaching a substituent to some common molecular fragment (hydrogen, methyl, phenyl, para-nitrophenyl, para-oxyphenyl, ortho-dialkylphenyls). All model compounds were treated with the PM3 method with full optimization of geometry. Computed heats of formation, HOMO and LUMO energies as well as charges on certain atoms were used as descriptors. Neural networks were simulated using the NASA WIN program developed at Moscow State University.

In order to control predictive performance of neural network, a database consisting of 160 substituents was splitted at random into two parts: training and validation sets. The

RMS error of predicting values of the  $\sigma^m$  constants is 0.06 on the training set (correlation coefficient 0.969) and 0.13 on the validation set. The RMS error of predicting values of the  $\sigma^p$  constants is 0.10 on the training set (correlation coefficient 0.959) and 0.16 on the validation set. The RMS error of predicting values of the F constants is 0.07 on the training set (correlation coefficient 0.940) and 0.14 on the validation set. The RMS error of predicting values of the R constants is 0.13 on the training set and 0.15 on the validation set. And, finally, the RMS error of predicting values of the  $E_s$  constants is 0.66 (correlation coefficient 0.980) on the training set and 0.40 on the validation set. The results of the study outperform results of analogous studies aimed at predicting substituent constants reported in literature so far.

Preliminary studies also show that both the use of substructural or topological descriptors instead of quantum-chemical ones as well as the use of multiple linear regression instead of the artificial neural networks results in a sharp deterioration of the predictive performance.

## CONCLUSIONS

The results of the study show that the combined use of molecular modelling and artificial neural networks may constitute a reliable basis for predicting various parameters of substituents and through them the biological activity of organic compounds.

## ACKNOWLEDGEMENT

We gratefully acknowledge support from the Russian Foundation for Basic Research and the Federal Program "Development of New Drugs by means of Chemical and Biological Synthesis" for the support of this work.

## REFERENCES

- Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., 1995, NASA. A computer program for performing QSAR/QSPR studies using artificial neural networks. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona.
- Halberstam, N.M., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., NASAWIN - Emulator of neural networks for QSAR. In *Proceedings of the IV Russian National Congress 'Man and Drug'*, PharMedInfo: Moscow.
- Baskin, I.I., Ait, A.O., Halberstam, N.M., Palyulin, V.A., Alfimov, M.V., and Zefirov, N.S., 1997, The use of artificial neural networks for predicting properties of complex molecular systems. Prognosis of the position of long-wave absorption band of symmetrical cyane dyes. *Dokl. Akad. Nauk.* 357:57.
- Kvasnicka, V., Sklenak, S., and Pospichal, J., 1993, Neural network classification of inductive and resonance effects of substituents. *J. Am. Chem. Soc.* 115:1495.

# APPLICATION OF NEURAL NETWORKS FOR ESTIMATING PARTITION COEFFICIENT BASED ON ATOM-TYPE ELECTROTOPOLOGICAL STATE INDICES

Jarmo J. Huuskonen<sup>1</sup> and Igor V. Tetko<sup>2</sup>

<sup>1</sup>Department of Pharmacy  
POB 56, FIN-00014 University of Helsinki, Finland

<sup>2</sup>Institute of Bioorganic & Petroleum Chemistry  
Murmanskaya 1, Kyiv, 253660, Ukraine

## INTRODUCTION

The logarithm of the partition coefficient between octanol and water,  $\log P$ , is a useful parameter to correlate transport properties of drug molecules, interactions between drugs and receptors, and changes in the structure of drugs with various biochemical or toxic effects of these compounds.<sup>1</sup> The measurement of  $\log P$  through synthesis of the compound and its subsequent experimental determination is time consuming and expensive. Hence, there is a strong interest in the structure-based prediction of  $\log P$  for rational drug design.

Among several approaches for computing  $\log P$  there are two essentially empirical methods for the estimation of  $\log P$ : Rekker's  $f$  constant method,<sup>2</sup> and Leo and Hansch's fragment approach.<sup>3</sup> Both methods divide a compound into basic fragments and calculate its  $\log P$  by the summation of the hydrophobic contributions of each fragment. However, the difficulty of these methods is how to fragment a molecule, especially large drug molecules, into basic fragments. Usually these methods use some correction factors for complex structures to compensate the interactions between functional groups.

Recently, Kier and Hall<sup>4</sup> introduced electrotopological state (E-state) indices for molecular structure description in which both electronic and topological characteristics are combined together. The E-state can be used in a group contribution manner and has been found to be useful in structure-property relationship studies, i.e. to predict the boiling points and critical temperatures for a set of heterogeneous organic compounds,<sup>5</sup> estimations of aqueous solubility,  $\log S$ , of drug compounds.<sup>6,7</sup> The present study shows that the same indices can be successfully used to estimate  $\log P$ .

## METHODS

326 compounds from different structural classes were randomly selected from the Hansch-Leo compilation.<sup>8</sup> The partition coefficients of these compounds were represented as logarithm values,  $\log P$ , and were in the range -2.11 to 5.90, corresponding to urea and thioridazine respectively. This data set was divided into a training set of 300 compounds and a test set of 26 compounds (selected at random). An additional test set of 19 compounds<sup>9</sup> was included in the present study to compare our approach with currently available ones.

Structural parameters were calculated by Molconn-Z software (Hall Associated

for each analyzed compound were analyzed using multilinear regression (MLR) analysis and artificial neural networks (ANNs). The SPSS package was used to run the MLR analysis. The ANNs were fully connected, feed-forward back-propagation networks with one hidden layer. The Early Stopping over Ensemble method was used to accomplish the overfitting/overtraining problem and to improve generalization ability of neural networks.<sup>10</sup>

## RESULTS AND DISCUSSION

Stepwise and backward methods were employed in the regression analysis. A satisfactory statistical model was detected for the training set containing 19 parameters ( $R = 0.93$ ,  $q^2 = 0.83$ ,  $s_{LOO} = 0.71$ ), where cross-validated  $q^2$  and the standard deviation  $s_{LOO}$  were calculated by leave-one-out method. The prediction ability of these parameters for the test sets was  $R = 0.93$ ,  $s = 0.73$  ( $n = 26$ ) and  $R = 0.91$ ,  $s = 0.69$  ( $n = 19$ ).

Neural networks applied to analyze all descriptors calculated similar prediction ability for the training test ( $q^2 = 0.83$ ,  $s_{LOO} = 0.70$ ) but higher for the test sets  $R = 0.95$ ,  $s = 0.60$  ( $n = 26$ ) and  $R = 0.93$ ,  $s = 0.58$  ( $n = 19$ ). Their prediction ability for the set of 19 compounds was comparable with that found using other known methods, such as CLOGP ( $R = 0.97$ ,  $s = 0.42$ ), XLOGP ( $R = 0.94$ ,  $s = 0.52$ ), Moriguchi's method ( $R = 0.93$ ,  $s = 0.53$ ) and was better than that of the Rekker's method ( $R = 0.92$ ,  $s = 0.77$ ). The analysis of residuals showed that in the test sets some compounds, i.e. loratidine and flufenemic acid, had particularly large errors for ANN regression. Both these compounds have logP values near the highest value (5.90) in the training set. These findings indicate that the training set should be extended by including more compounds with high logP values.

The most important advantage of the present approach is that only 33 parameters and no correction factors were used for coding each molecule and calculation of logP, while other methods required hundreds of parameters. The prediction of partition coefficients using neural networks and atom-type E-state indices is accurate and provides reliable logP estimations comparable with those obtained by other methods. An advantage of the proposed approach is that the atom-type E-state indices can be quickly and easily estimated directly from chemical structure of analyzed compounds. Thus, the present approach introduces a fast method for estimation of logP of chemical compounds.

## Acknowledgments

This study was partially supported by the Technology Development Center in Finland (TEKES) and INTAS-Ukraine grant 95-0060.

## REFERENCES

1. C. Hansch and A. Leo. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York (1979).
2. R.E. Rekker. *Hydrophobic Fragment Constant*, Elsevier, New York (1977).
3. A. Leo, P. Jow, C. Silipo, and C. Hansch, Correlation of hydrophobic constant (logP) from  $\pi$  and  $f$  constants, *J. Med. Chem.* 18:865-868 (1975).
4. L.B. Kier and L.H. Hall, An electrotopological-state index for atoms in molecules, *Pharm. Res.* 7:801-807 (1990).
5. L.H. Hall and C.T. Story, Boiling point and critical temperature of a heterogeneous data set: qsar with atom type electrotopological state indices using artificial neural networks, *J. Chem. Inf. Comput. Sci.* 36:1004-1014 (1996).
6. J. Huuskonen, M. Salo, and J. Taskinen, Neural network modeling for estimation of the aqueous solubility of structurally related drugs, *J. Pharm. Sci.* 86:450-454 (1997).
7. J. Huuskonen, M. Salo, and J. Taskinen, Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, *J. Chem. Inf. Comput. Sci.* 38:450-456 (1998).
8. C. Hansch, A. Leo, and D. Hoekman. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, American Chemistry Society, Washington, (1995).
9. I. Moriguchi, S. Hirono, I. Nakagome, and H. Hirano, Comparison of reliability of logP values for drugs calculated by several methods, *Chem. Pharm. Bull.* 42:976-978 (1994).
10. I.V. Tetko and A.E.P. Villa, Efficient partition of learning data sets for neural network training, *Neural Networks* 10:1361-1374 (1997).

## VARIABLE SELECTION IN THE CASCADE-CORRELATION LEARNING ARCHITECTURE

Igor V. Tetko,<sup>1,2,\*</sup> Vasyly V. Kovalishyn,<sup>1</sup> Alexander I. Luik,<sup>1</sup>  
Tamara N. Kasheva,<sup>1</sup> Alessandro E.P. Villa,<sup>2</sup> and David J. Livingstone<sup>3</sup>

<sup>1</sup>Institute of Bioorganic & Petroleum Chemistry  
Murmanskaya 1, Kyiv, 253660 Ukraine

<sup>2</sup>Institut de Physiologie, Rue du Bugnon 7  
Lausanne, CH-1005 Switzerland

<sup>3</sup>ChemQuest, Cheyney House, 19-21 Cheyney Street, Steeple Morden,  
Herts, SG8 0LP U.K. and Centre for Molecular design,  
University of Portsmouth, Hants, PO1 2EG U.K.

### INTRODUCTION

Recently there has been a growing interest in the application of neural networks in the field of QSAR. It was demonstrated that this method is often superior to the traditional approaches.<sup>1</sup> Other studies have shown that prediction ability of such methods can be substantially improved if the number of input variables for neural networks is optimized.<sup>2,3</sup>

The neural networks used in the previous studies usually were characterized by fixed-size architectures (FNN), i.e. the number of hidden neurons, the number of connection weights and the connectivity amid layers were all fixed. The capacity and accuracy of a network mapping is determined by the number of free parameters (typically weights) in the network. Neural networks that are too small (underfitting) cannot accurately approximate the desired input-to-output mapping, while too large networks can have a lower generalization ability because of the overfitting/overtraining problem.<sup>4,5</sup> Thus, an incorrect selection of the topology can decrease the performance of this method.

Some other algorithms, so called topology-modifying<sup>6</sup> neural network algorithms, are able to automatically determine an optimal neural network architecture that is pertinent to the analyzed problem. These algorithms start with a small network and add weights or/and nodes until the problem has been solved. These methods are of considerable interest for practical applications because of their ability to solve some tasks (e.g., the two-spirals problem<sup>7</sup>) which represent substantial difficulties for the training of fixed-size neural networks. The current study introduces pruning algorithms for one of the most popular topology-modifying algorithms -- the Cascade Correlation neural network (CCN).<sup>7</sup>

### METHODS AND DATASETS

The original version of the CCN<sup>7</sup> was incorporated in Neural Network Ensemble software (C++). Several pruning methods that were found to be most efficient for FNN<sup>2</sup> were programmed for this algorithm as described elsewhere.<sup>8</sup>

---

\* <http://www.lnh.unil.ch>

The performance of the developed algorithms was verified using a set with linear and two sets with non-linear dependencies between inputs and the output. The QSAR examples used to better access generalization ability of the pruning methods included 51 benzodiazepine derivatives with anti-pentylentetrazole activity, 37 2,5-bis(1-aziridinyl)-p-carboquinones with antileukemic activity, 74 2,4-diamino-5-(substituted benzyl)-pyrimidines (inhibitors of dihydrofolate reductase from MB1428 *E. coli*) and 31 antimycin analogues with antifilarial activity.<sup>2,8</sup>

## RESULTS

An application of the pruning algorithm for simulated datasets calculated results that were in perfect agreement with theoretical expectations as well as with previous results calculated by fixed-size neural networks. All methods were able to correctly estimate the order of sensitivity for analyzed input parameters with and without noise in the input data.

**Table 1.** Leave-one-out cross-validated  $q^2$  coefficients calculated for the QSAR examples

analyzed dataset	n	all inputs		optimized inputs	
		CCN	FNN	CCN	FNN
Benzodiazepines	51	0.64±0.02	0.64±0.03	0.66±0.01	0.67±0.02
Carboquinones	37	0.76±0.02	0.79±0.03	0.78±0.02	0.85±0.02
Pyrimidines	74	0.37±0.05	0.40±0.03	0.68±0.01	0.65±0.02
Antimycin analogues	31	0.32±0.03	0.32±0.03	0.67±0.01	0.67±0.01

The calculated results suggest that the elaborated pruning methods can be successfully used to optimize the set of variables for QSAR studies. The use of variables selected by the elaborated methods improves neural network prediction ability compared to that calculated using the unpruned sets of variables. The results calculated by FNN and CCN are similar on average, however the CCN is considerably faster particularly since no optimization of topology is required for this method.

## Acknowledgments

This study was partially supported by INTAS-Ukraine grant 95-0060.

## REFERENCES

1. D. Maddalena, Applications of artificial neural networks to quantitative structure-activity relationships, *Exp. Opin. Ther. Patents*. 6:239-251 (1996).
2. I.V. Tetko, A.E.P. Villa, and D.J. Livingstone, Neural network studies. 2. Variable selection, *J. Chem. Inf. Comput. Sci.* 36:794-803 (1996).
3. D.J. Maddalena and J.A.R. Johnston, Prediction of receptor properties and binding affinity of ligands to benzodiazepines/gabaa receptors using artificial neural networks, *J. Med. Chem.* 38:715-724 (1995).
4. S. Geman, E. Bienenstock, and R. Dourstat, Neural networks and the bias/variance dilemma, *Neural Computation* 4:1-58 (1992).
5. I.V. Tetko, D.J. Livingstone, and A.I. Luik, Neural network studies. 1. Comparison of overfitting and overtraining, *J. Chem. Inf. Comput. Sci.* 35:826-833 (1995).
6. T. Ash and G. Cottrell, Topology-modifying neural network algorithms, in: *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib, ed., MIT, Cambridge, (1995).
7. S.E. Fahlman and C. Lebiere, The cascade-correlation learning architecture, in: *NIPS\*2*, D.S. Touretzky, ed., Morgan-Kaufmann, San Mateo (1990).
8. V.V. Kovalishyn, I.V. Tetko, A.I. Luik, V.V. Kholodovych, A.E.P. Villa, and D.J. Livingstone, Neural network studies. 3. Variable selection in the cascade-correlation learning architecture, *J. Chem. Inf. Comput. Sci.* 38:651-659 (1998).

## CHEMICAL FINGERPRINTS CONTAINING BIOLOGICAL AND OTHER NON-STRUCTURAL DATA

Fergus Lippi<sup>a</sup>, David Salt<sup>ab</sup>, Martyn Ford<sup>a</sup> and John Bradshaw<sup>c</sup>

<sup>a</sup> Centre for Molecular Design

<sup>b</sup> School of Computer Science and Mathematics

University of Portsmouth

Portsmouth, PO1 2EG, UK

<sup>c</sup> Glaxo-Wellcome

Stevenage, SG1 2NY UK

### INTRODUCTION

The pharmaceutical industry is concerned with identifying drugs for safe treatment for human disease. Searching for these bio-active compounds is like looking for a needle in a haystack. The aim of this research is to develop a methodology for searching large databases for lead compounds using similarity, and chemical libraries for high throughput screening (HTS) using diversity. In this study, similarity has been calculated using the Tanimoto Index (TI) and diversity has been derived as (1-TI).

Compounds can be defined in terms of the three property classes which may be termed structural, chemical and biological. These classes contain information on which to base similarity and diversity measurements, but only a fraction of the available information needs to be used to search for lead compounds. Much of the available information is irrelevant and is useless for identifying hits, and some of the information is redundant as it is shared by more than one variable or source of data. Furthermore, information can be described as nominal data (categorical), ordinal data (e.g. ranks) and continuous data. Any of these data types can be used to describe the properties of molecules and hence form a basis for determining molecular similarity and/or diversity.

Bitstrings are appropriate for representing all of these types of properties and, because they are exact and unique, they provide an efficient means of coding data for use in computer database searches. Bitstrings based on structural data, e.g. Daylight fingerprints, are commonly used for searching for molecular similarity or diversity. However, these fingerprints are abstract descriptions of chemical structure that are multidimensional in character. The information that they contain is often obscure and highly redundant. As a result, interpretation of the nature of the sets of molecules obtained from database searches using this type of bitstring can be difficult, even confusing. What is required is a more explicit form of representation. Furthermore, it is often useful to take account of other descriptors, e.g. chemical and biological properties that lead to more focussed searches in

lower dimensional space than can be achieved with fingerprints, or properties, e.g. compound availability and cost, that might influence the cost-effectiveness of the drug design process.

For use in database searching for lead compounds, bitstrings should contain enough bits to encode diversity of information to an acceptable precision, yet ensure that the density of bits on is appropriate. For use with the Tanimoto Index, too high a density results in too many false positives (non-hit lead compounds categorised as hits) being identified, whereas too low a density leads to too many false negatives (true lead compounds categorised as non-hits). We report a coding system that meets these requirements and maintains an appropriate balance between the diversity and the density of the bitstring.

## **BAND CODING**

The code adopted was first described by Albus (1975), is derived for control theory and provides a mapping from decimal variable to binary equivalents. Studies undertaken at Portsmouth have demonstrated the utility of this coding system for accurate and rapid manipulation of continuous, ordinal and categorical data. An advantage of this coding procedure is that it allows the precision of the mapping to be adjusted to meet user requirements. Each source or type of data is encoded by the procedure of Albus to create a field of bits of fixed length and density. These fields may then be appended to produce bitstrings of appropriate length and containing the required diversity of information. In this way, structural data (e.g. chemical fingerprints) may be combined with molecular or biological properties, compound cost and availability.

Band codes have been constructed for an MSI database containing 75 active compounds covering 14 activity classes (containing at most five active compounds per class) and characterised by 4 descriptors (number of rotatable bonds, MW, logP and molecular refractivity). The Tanimoto Index has been formulated in terms of the Band Coding parameters -  $N$  (the length of the bitstring) and  $b$  (the number of bits set on) – and used to search for hits based on similarity to target molecules representing the 14 activity classes. The results showed that, almost without exception, bitstrings comprising structural, chemical and biological information, outperformed those based on chemical fingerprints alone. The number of compounds that had to be sampled during a search (the run length) in order to extract five lead compounds (where possible) was reduced by increasing the diversity of information contained in a string. Thus, the average run length fell from 42 for the MSI property set and 36 for the Daylight chemical fingerprints alone, to 27 when these sources of information were combined in a single binary Band Coded bitstring. The mean run lengths obtained for the MSI, Daylight and pooled MSI/Daylight bitstrings, diverged as the number of hits found increased. This suggests that there could be considerable benefit in using extended Band Coded bitstrings with very large databases.

## **REFERENCE**

Albus, J. S., 1975, A new approach to manipulator control: the cerebellar model articulation controller (CMAC). *Journal of Dynamic Systems, Measurement and Control*, 97:220-227.



## **RODENT TUMOR PROFILES INDUCED BY 536 CHEMICALS CARCINOGENS: AN INFORMATION INTENSIVE ANALYSIS**

R. Benigni, A. Pino and A. Giuliani

Lab. Comparative Toxicology and Ecotoxicology  
Istituto Superiore di Sanità  
Rome - Italy

The rodent carcinogenicity bioassay, generally considered the most reliable predictor of human cancer hazard, provide a wealth of data and information collected in large databases. The use of computerized data analysis techniques suitable for the exploration of these databases makes its investigation much more fruitful, and its results more reliable. In the past years, interrogation of such databases has focused mainly on: a) the relationship between mutagenicity and carcinogenicity; b) the role of toxicity and cellular proliferation on the chemical carcinogenesis; c) the organ specificity for the tumors induced by mutagens and nonmutagens, respectively; d) the relationship between chemical structure and biological activity.

The aim of this work was to consider the carcinogenicity database and to find a formalization of the carcinogenicity activity - usually simplified in a +/- - appropriate for SAR and QSAR studies. We also present a preliminary analysis of the relationship between type of tumor profile and chemical class/mode of action.

The total number of rodent carcinogens was 536 derived from the NCI/NTP<sup>1</sup> (185 chemicals) and the Gold et al.<sup>2</sup> database (351 chemicals). Each carcinogen was associated with the information on the induction of 44 tumor types (target organs) for a total of 176 variables (44 tumor types x 4 experimental groups: male and female rat, male and female mouse). The final data matrix contained 536 rows (chemicals) and 176 variables (tumor types); the values of the variables were 1 (induction of tumor) and 0 (non induction). The analysis of the data was performed with the Kohonen Self-Organizing Map<sup>3</sup> (SOM) artificial Neural Network, that constructs maps of similarities among statistical units.

From a preliminary comparison of the tumor distribution in the two subsets under study (chemicals bioassayed from NTP and the whole dataset) we can observe that the ratio of the tumors induced to number of chemicals was identical in the sets of the database, thus pointing to similar average characteristics. In addition the relative distribution of tumors in the four experimental groups remained unchanged, so that the major characteristics are preserved. The main types of tumors are the same in the two sections (for example L -Liver, LU-Lung, UB-Urinary Bladder, ZG-Zymbal's Gland). This evidence of similarity is important to show that the NTP chemical carcinogenicity database

available at present is quite representative of the general trends of organ, species and sex specificity of chemical carcinogenicity. For most of these tumor there is a clear species specificity with L, LA (Liver Adenomas), and LU more frequent in mouse, whereas LN (Liver Nodules, MG (Mammary Gland), K (Kidney), are more frequent in rat.

To visualize and quantify the relationships among patterns and tumor types, the 536 x 176 data matrix was analyzed with SOM. First, the subset of 185 chemicals with complete NTP experimentation was analyzed; then, with a further SOM application, the remaining 351 chemicals were allocated in the map based on their similarity to the above carcinogens. In this way, each carcinogen (tumor pattern) was assigned two quantitative parameters ( $k_1$  and  $k_2$ ). To highlight the structural features of the map, representatives of different zones ( $n=9$ ) were sampled. Then the 176 types of tumor were considered as statistical units, and the 185 NTP chemicals with complete experimentation as variables. We can observe that most often species specificity overcomes organ specificity.

The possibility of practically applying the above results to QSAR studies has been examined in two preliminary analyses. In the first analyses we checked the hypothesis that similar structures would induce similar tumors profiles in the experimental animals. We selected from the 536 compounds, the chemicals belonging to three chemical/mode of action classes (among the most studied, numerically well represented and mechanistically well understood): a) aromatic amines; b) natural electrophilic/alkylating agents; c) nitroarenes. From the observation of the distribution of the carcinogens within the three classes, based on the induction tumor profiles, appears that no obvious association exists between chemical/mode of action class and tumor profiles, while the three classes homogeneously span the entire space of tumor profiles.

In summary this work produced a quantitative classification of tumor profiles, suitable for further QSAR studies and pointed to complicated relationships between chemicals class/mode of action and tumor profiles.

1. J.K. Selkirk and S.M. Soward. Compendium of abstracts from long-term cancer studies reported by the National Toxicology Program from 1976 to 1992, *Environ. Health Perspect.* 101 (1993)
2. L.S. Gold, T.H. Slone, N.B. Manley, G. B. Garfinkel, E.S. Hudes, L. Rohrbach and B.N. Ames, The carcinogenic potency database: analyses of 4000 chronic animal cancer experiments published in the general literature and by the U.S. National Cancer Institute/National Toxicology Program, *Environ. Health Perspect.* 96:11-15 (1991).
3. T. Kohonen, *Self-organization and Associative Memory*, Springer-Verlag, New York, (1988)

# COMPARISON OF SEVERAL LIGANDS FOR THE 5-HT<sub>1D</sub> RECEPTOR USING THE KOHONEN SELF-ORGANIZING-MAPS TECHNIQUE

Joachim PETIT, Daniel P. VERCAUTEREN

Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles, 61, B-5000 NAMUR (Belgium)

## INTRODUCTION

The purpose of this present work is to improve our knowledge of the 5-HT<sub>1D</sub> receptor, an important therapeutic target for the treatment of migraine. Actually, we wish to determine the required structural features to understand the behaviour of the ligand/receptor system, *i.e.*, why some ligands present a high activity and a great selectivity.

As in numerous biochemical systems, the three-dimensional (3D) structure of the receptor is still not well known and consequently Structure/Activity Relationships (SAR) studies of ligand candidates constitute a necessary and amenable approach to the problem. In order to characterize our set of ligands, we have chosen the molecular electrostatic potential (MEP) evaluated on the van der Waals (VDW) surface, which is well-known to be one of the responsible factor for the binding of a substrate molecule at the active site of the biological receptor. However, the 3D nature of the MEP contours makes difficult to visualize simultaneously their spatial distribution and values; hence, we can easily understand the advantage of the Kohonen self-organizing-maps (SOM) technique, which permits to tackle all the information at the same time.

The presented results consist in the calculation of the Kohonen's maps for our set of ligands, which will be used further in order to perform SAR studies.

## SET OF STUDIED MOLECULES

We collected, from the literature, a set of 13 ligand candidates for the 5-HT<sub>1D</sub> receptor, for which we know the activity and selectivity (in comparison with the 5-HT<sub>2A</sub> receptor) values. These molecules are as different as:

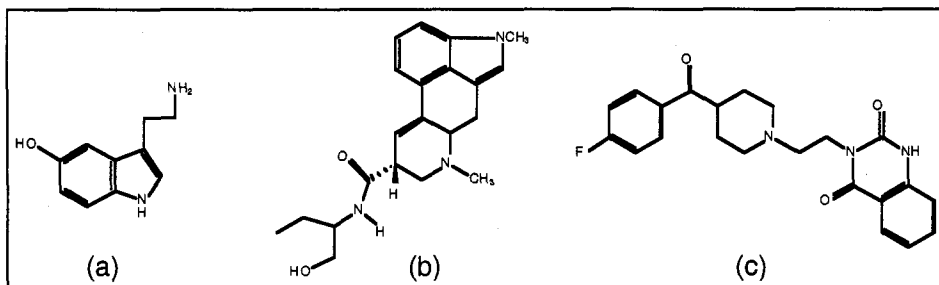


Figure 1- Structural formulae of (a) 5-hydroxytryptamine, (b) methysergide, and (c) ketanserine.

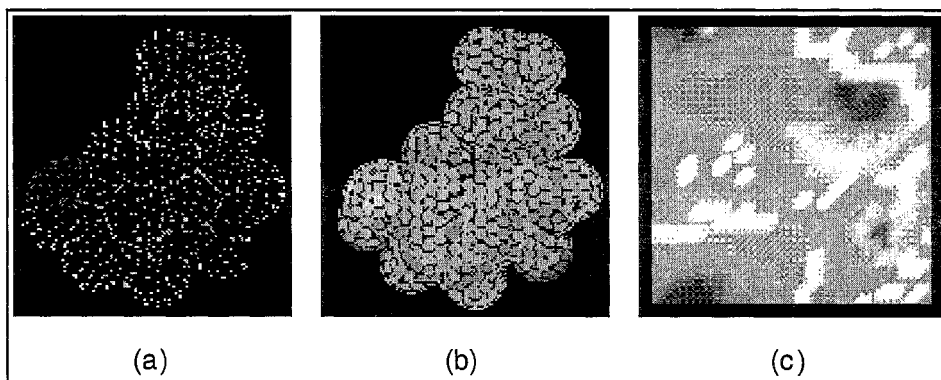
## COMPUTATION AND MAPPING OF THE MEP

The conformations of the molecules were optimized with the TRIPOS force field (they are not necessarily the most stable ones but are superimposable on the reference rigid template: methysergide).

We generated points on the VDW envelope (with a density of 5 points per Angström<sup>2</sup>) using the in-house KEMIT software, as shown for the 5-hydroxytryptamine (Fig. 2a).

The MEP values were calculated at the RHF/SCF/6-31G\*\* level of theory with the Gaussian94 software (tight SCF convergence criteria). The obtained results are visualized with KEMIT (Fig. 2b).

The 2D maps of the 3D MEP contours were obtained using an adapted version of the K-ctr (K=Kohonen, ctr=counter-propagation) program of Prof. J. Zupan of the Nat. Inst. of Chem. of Ljubljana. Based on the Kohonen's SOM technique, this program can perform a non-linear projection (mapping) of a data set of high dimension (3D, *i.e.*, the three cartesian coordinates of points in our case) to a 2D space, conserving the topology of the information. We have opted for a toroidal mapping space, because of its continuous character. The visualization and color-coding, according to the MEP values, are obtained using Data Explorer (IBM) (Fig. 2c).



**Figure 2-** (a) Selected points at the VDW surface, (b) MEP values, and (c) mapping of the MEP, for 5-hydroxytryptamine.

## COMPARISON OF MAPS

Applying the procedure described above to the entire set of molecules and tiling the obtained maps, we can easily notice the real interest of that technique. Structurally different molecules, presenting similar behaviours regarding to the 5-HT<sub>1D</sub> and 5-HT<sub>2A</sub> receptors, lead to maps reproducing similar features.

## CONCLUSIONS AND OUTLOOK

Here, we explained the procedure that we have settled in order to obtain MEP 2D-maps of several 5-HT<sub>1D</sub> candidate molecules. We have emphasized the real capabilities of the Kohonen SOM theory to facilitate the visualization and comparison of 3D-properties.

Our future plans consist in developing a systematic and automated comparison method in order to bring to the fore the similarities and dissimilarities of the generated maps without preliminary superimposition of the molecules.

## ACKNOWLEDGEMENTS

The authors wish to thank Prof. J. Zupan and Dr. A. Michel for fruitful discussion and the FUNDP and IBM Belgium for the use of the SCF center.

## **BINDING ENERGY STUDIES ON THE INTERACTION BETWEEN BERENIL DERIVATIVES AND THROMBIN AND THE B-DNA DODECAMER D(CGCGAATTCGCG)<sub>2</sub>**

Júlio C. D. Lopes, Ramon K. da Rocha, Andrelly M. José, and Carlos A. Montanari

NEQUIM - Departamento de Química - ICEx/UFMG  
Av. Antônio Carlos, 6627 - 31.270-901 Belo Horizonte, MG - Brazil  
E-mail: jlopes@dedalus.lcc.ufmg.br

### **INTRODUCTION**

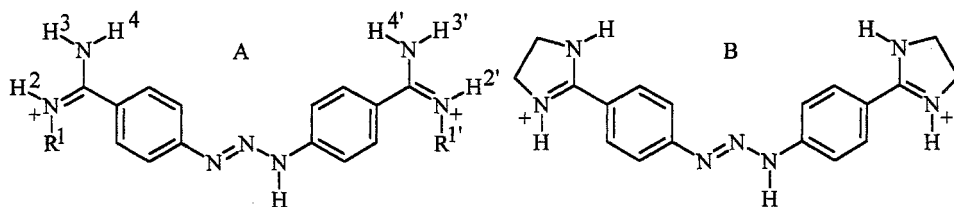
An important mechanism for some drugs action is by their interaction with genetic material of infecting agent. The formation of drug-DNA complex obstructs the transcription as well DNA replication, inhibiting the multiplication of cell and production of fundamental proteins for its survival. One way for interaction is the drug insertion inside B-DNA's minor groove and for this a complementary isohelic conformation to DNA is necessary.<sup>1</sup> The diamidines like berenil show high affinity for AT rich sequences offering, thus, special interest in antiviruses, antitumor and antiprotozoal drug development. Recent studies show that pharmacological activity of these substances is correlated to their DNA-binding affinity.<sup>2,3</sup> However, diamidines like pentamidine have an expressive effect on the blood coagulation system by thrombin inhibition<sup>4,5</sup>, and berenil is a parabolic competitive thrombin inhibitor.<sup>6</sup> Our goal is to propose new synthetic drugs with higher affinity towards DNA and lower one for thrombin.

### **METHODOLOGY**

We have an on-going project on the study of drug-receptor interaction to suggest the synthesis of new drugs with higher potency against *Leishmania sp.*<sup>1,7</sup> In order to obtain new derivatives of berenil that present higher affinity for its receptor (DNA) we have performed binding energy calculations for some berenil analogues with DNA. In order to reduce the side effect of berenil the binding energy calculations for the same berenil analogues with thrombin have also been carried out.

In the present communication our calculations followed a molecular mechanics approach making use of the AMBER force field within INSIGHTII/DISCOVER (MSI). The crystallographic structure of berenil complex with the dodecanucleotide

d(CGCGAATTCGCG)<sub>2</sub> (PDB: 2DBE) and the crystallographic structure of a complex with  $\alpha$ -thrombin and benzamidine (PDB: 1DWB) were used as starting point for calculations. Berenil and analogues studied are showed in Figure 1.



**Figure 1.** Structure of the study compounds. A: Berenil R = H **1**, ethynilic derivative R = CCH **2**, ethylic derivative R = CH<sub>2</sub>CH<sub>3</sub>, **3**, perfluoroethylic derivative R = CF<sub>2</sub>CF<sub>3</sub>, **4** and acetynic derivative R = CH<sub>2</sub>COCH<sub>3</sub>, **5**. B: Dihydroimidazolic derivative **6**.

There are sixteen planar conformations for each analogue **2-5** that have been constructed and docked into DNA and thrombin. After minimization of the complexes, DNA (or thrombin) and drug were minimized separately. The interaction and binding energies were calculated using the method showed below.

$$E_{\text{interaction}} = E_{\text{complex}} - (E_{\text{receptor-complex conformation}} + E_{\text{ligand-complex conformation}})$$

$$E_{\text{binding}} = E_{\text{complex}} - (E_{\text{receptor-global minimum}} + E_{\text{ligand-global minimum}})$$

## RESULTS

All derivatives displayed more negative interaction energy than berenil itself, toward DNA. The derivatives that have the more negative binding energy were **5** < **6** < **4** < **1** < **3** < **2**. All derivatives have more positive interaction energy than berenil itself, toward thrombin. The binding energy in thrombin complexes were positives for all derivatives. The derivatives that showed the more positive binding energy were **3** > **2** > **5** > **6** > **1** > **4**.

The derivatives **5** and **6** showed higher affinity to DNA in relation to the berenil. On other hand the same derivatives showed poorer affinity for thrombin in relation to the berenil. We concluded that the derivatives **5** and **6** are the more suitable to encompass our objectives. Nowadays, we are working on the synthesis of these new berenil derivatives.

## REFERENCES

- 1 - Montanari, C. A.; Tute, M. S.; Beezer, A. E. and Mitchell, J. C.; *J. Comput.-Aided Comput. Sci.* **1996**, 10, 67.
- 2 - Bell, C. A.; Cory, M.; Fairlay, T. A.; Hall, J. E. and Tidwell, R. R.; *Antimicrob. Agents Chemother.* **1991**, 35(6), 1099.
- 3 - Boykin, D. W.; Kumar, A.; Xiao, G.; Wilson, W. D.; Bender, B. C.; McCurdy, D. R.; Hall, J. E. and Tidwell, R. R.; *J. Med. Chem.* **1998**, 41, 124.
- 4 - Geratz, J. D. & Whitmore, A. C., *J. Med. Chem.* **1973**, 16, 970.
- 5 - Vieira, L. M., Barbosa de Deus, Nicolato, R. L. C. & Andrade, M. H. G. *Rev. Bras. Anal. Clin.* **1992**, 24, 43.
- 6 - Junqueira, R. G. Silva, E. & Mares-Guia, M. *Arq. Biol. Tecnol.* **1989**, 32, 210 (Abstract L-5).
- 7 - Bell, C. A.; Hall, J. E.; Kyle, D. E.; Grogl, M.; Ohemeng, K. A.; Allen, M. A. and Tidwell, R. R.; *Antimicrob. Agents Chemother.* **1990**, 34(7), 1381.

# A COMPARISON OF *AB INITIO*, SEMI-EMPIRICAL, AND MOLECULAR MECHANICS APPROACHES TO COMPUTE MOLECULAR GEOMETRIES AND ELECTROSTATIC DESCRIPTORS OF HETEROATOMIC RING FRAGMENTS OBSERVED IN DRUGS MOLECULES

G. Longfils<sup>1</sup>, F. Ooms<sup>1</sup>, J. Wouters<sup>1</sup>, A. Olivier<sup>2</sup>, M. Sevrin<sup>2</sup>, P. George<sup>2</sup>, F. Durant<sup>1</sup>

<sup>1</sup>Laboratoire de Chimie Moléculaire Structurale, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium

<sup>2</sup>CNS Research Department, Synthélabo Recherche, 31, Av. Paul Vaillant Couturier, 92200 Bagneux, France

## INTRODUCTION

In order to compare physico-chemical properties of a set of heterocycles (figure 1.), four parameters have been calculated for each heterocyclic rings. These variables were selected on the basis of their potential involvement in the molecular drug-receptor recognition. Dipole moment, atomic charges, orbital energies (especially the HOMO and LUMO energies) and molecular electrostatic potential (PEM) were selected to determine the most appropriate calculation methods acceptable for this purpose. Moreover, the geometry optimization procedure was also investigated.

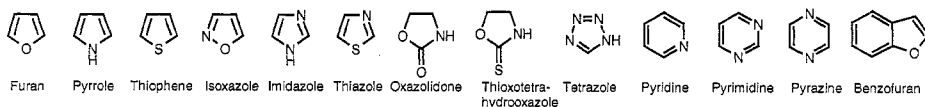


Figure 1. Representation of the heterocycles used in this study.

## RESULTS

### Optimization geometries procedures

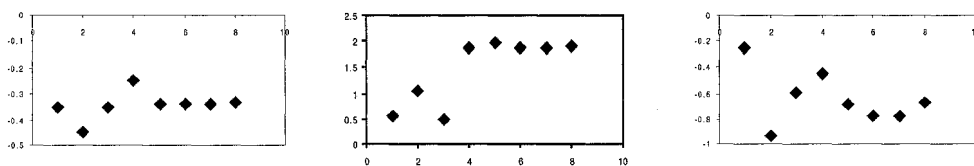
The optimization procedure has been investigated by *ab initio* (STO-3G, 3-21G, 6-31G and 6-31G\* basis set), semi-empirical (CNDO, MNDO and AM1) and molecular mechanic (CFF95 forcefield) methods. In order to determine the method giving the best compromise between the accuracy of the results and the calculation time, we used 2-oxazolidone for which X-ray results were available for comparison. Bond lengths and angles of oxazolidone were determined by these three methods and compared to the crystallographic structure.

The analysis of the bond lengths (Å) and angles (°) shows that semi-empirical methods as well as the STO-3G basis set are less appropriate to optimize the geometry correctly. This fact is clearly shown by the hybridization of the nitrogen atom of the oxazolidone. This atom adopts an  $sp^2$  hybridization in the X-ray structure as given by the sum of the bond angles ( $359,4^\circ$ ) while with the semi-empirical and *ab initio* STO-3G ( $335,8^\circ$ ) basis set, the N atom adopts an  $sp^3$  hybridization. On the other hand, the bond lengths and angles obtained by the *ab initio* 3-21G, 6-31G and 6-31G\* basis set are very close to those obtained experimentally. Results obtained by molecular mechanics are also very close to the X-ray structure.

Similar results are observed for the other heterocycles studied.

## Properties calculation procedure

The aim of this part of the work is to choose a suitable basis for the calculations of the properties like HOMO or LUMO energies, dipole moment and charges (ChelpG). The input geometries were obtained by molecular mechanic with CFF95 forcefield. Properties were calculated with semi-empirical methods (MNDO, CNDO, AM1) and *ab initio* procedures (STO-3G, 3-21G, 4-31G, 6-31G and 6-31G\*).



**Figure 2.** Representation of molecular property versus basis set: MNDO (1), CNDO (2), AM1 (3), STO-3G (4), 3-21G (5), 4-31G (6), 6-31G (7), 6-31G\* (8), (left) of the HOMO energies of thiophene, (middle) the dipole moment of pyrrole and (right) the ChelpG charge on the N of the pyridine

The HOMO energies (left) and dipole moment (middle) are similar from 3-21G until 6-31G\*. The semi-empirical methods as well as STO-3G are less suitable for the calculation of those properties (figure2). This is demonstrated by the fact that the properties obtained with these methods are quite different when compared with more sophisticated basis set (6-31G\*). For the charges (right), the *ab initio* methods present a slight minimum for the 4-31G and 6-31G basis set. Similar results are obtained for the other heterocycles studied here.

That observation suggests that the 6-31G\* basis is a good method for the calculation of electrostatic properties. The choice of this basis set is also necessary for compounds containing a heavy atom like thiophene do. Moreover; these results are also convenient for the determination of properties such as the topology of the MEP.

For bigger systems the 3-21G(\*) basis set is suitable to derive such properties.

## CONCLUSIONS

This study has shown that:

- Geometry optimization with molecular mechanic (CFF95 forcefield) and the 3-21G(\*) or 6-31G(\*) *ab initio* methods give results similar to those observed in the X-ray structure.

- 6-31G\* basis is the most appropriate method to calculate electronic properties like HOMO -LUMO energies, MEP, dipole moment and charges.

Further studies will be performed on larger sets of molecules using molecular mechanics (CFF95) which is the most rapid and accurate method for optimization and 6-31G\* for properties calculation.



## ELABORATION OF AN INTERACTION MODEL BETWEEN ZOLPIDEM AND THE $\omega_1$ MODULATORY SITE OF GABA<sub>A</sub> RECEPTOR USING SITE-DIRECTED MUTAGENESIS

Olivier A.<sup>1</sup>, Renard S.<sup>2</sup>, Even Y.<sup>2</sup>, Besnard F.<sup>2</sup>, Graham D.<sup>2</sup>, Sevrin M.<sup>1</sup>, George P.<sup>1</sup>

Synthélabo Recherche, 31, Avenue Paul Vaillant-Couturier 92220 Bagneux

<sup>1</sup> CNS Research Department

<sup>2</sup> Genomic Biology Department

### INTRODUCTION

Molecular cloning experiments have revealed the existence of five different families ( $\alpha_{1-6}$ ,  $\beta_{1-3}$ ,  $\gamma_{1-3}$ ,  $\delta_1$  and  $\rho_{1-2}$ ) of subunits which constitute the GABA<sub>A</sub> receptor complex. The functional brain receptor is an oligomer composed of a combination of  $\alpha$ ,  $\beta$  and  $\gamma$  subunits. The pharmacology of GABA<sub>A</sub> receptor subtypes critically depends on the particular  $\alpha$  subunit isoform that is present in the complex. The aim of this work was to elaborate a model of interaction between zolpidem, an  $\omega_1$  selective ligand (high affinity for the  $\alpha_1\beta_2\gamma_2$  subunit combination *versus*  $\alpha_5\beta_2\gamma_2$ ) and the  $\alpha_1$  subunit of the  $\omega$  modulatory site present on the GABA<sub>A</sub> receptor complex.

Two kind of approaches were used to elaborate this model :

- 1) Evaluation of physico-chemical properties of zolpidem implicated in the interactions with its target
- 2) Sequence analysis of the  $\alpha$  subunits and point mutations on the  $\alpha_5$  subunit.

### STEREOELECTRONIC AND CONFORMATIONAL PROPERTIES OF ZOLPIDEM

In addition to NMR conformational studies of zolpidem, <sup>13</sup>C NMR shifts of 4-phenyl substituted carbon atoms were identified as an index of the ability of the 2-phenyl to be involved in a  $\pi$ -H interaction. These studies led to a better understanding of the physico-chemical properties of zolpidem and allow us to propose a pharmacophoric model for zolpidem. This is composed of four zones : two hydrogen acceptor sites, one localized on N<sub>1</sub> of zolpidem and the other on the carbonyl of the acetamide side chain. This latter is localized at 2Å above the plane of the heterocycle, close to the pyridine ring and is implicated in the selectivity of zolpidem for the  $\omega_1$  site ; an hydrophobic interaction zone on the pyridine ring and a charge transfer interaction as  $\pi$ - $\pi$  or  $\pi$ -H localized on the phenyle moiety in position 2.

## ELABORATION OF AN INTERACTION MODEL

In order to identify those amino acids of the  $\alpha_1$  subunit that interact with zolpidem, sequence alignment of GABA<sub>A</sub> receptor  $\alpha$ -subunits was realized. This analysis suggested two regions localised between the Cys-Cys loop and the first transmembrane segment that varied from one subunit to another, in particular between the  $\alpha_1$  and  $\alpha_5$  subunits and which could account for the selectivity of zolpidem for the  $\alpha_1$  subunit. To evaluate this hypothesis, chimaeric receptors were constructed with  $\alpha_2/\alpha_1$  subunits coexpressed with  $\beta_2$  and  $\gamma_2$  subunits and the affinity of zolpidem was evaluated. From the binding profile of zolpidem to chimaeric receptors, it was observed that mutation of at least two amino acid residues of the  $\alpha_2$  subunit are necessary to endow the mutated receptor with a high-affinity for zolpidem. These studies allow us to propose a hypothetical interaction model between zolpidem and the  $\omega_1$  modulatory binding site (Figure 1). The interaction model for zolpidem and  $\omega_1$  site is based on the following hypotheses :  $\alpha_1$  histidine 101 and  $\alpha_1$  serine 204 interact respectively with the N<sub>1</sub> of imidazole ring and the carbonyl of the acetamide side chain ; hydrophobic aminoacids in the region around  $\alpha_1$  threonine 162 could interact with pyrimidine ring of zolpidem ; and finally aminoacids of  $\gamma_2$  could interact with the phenyl in position 2 of the heterocycle of zolpidem.

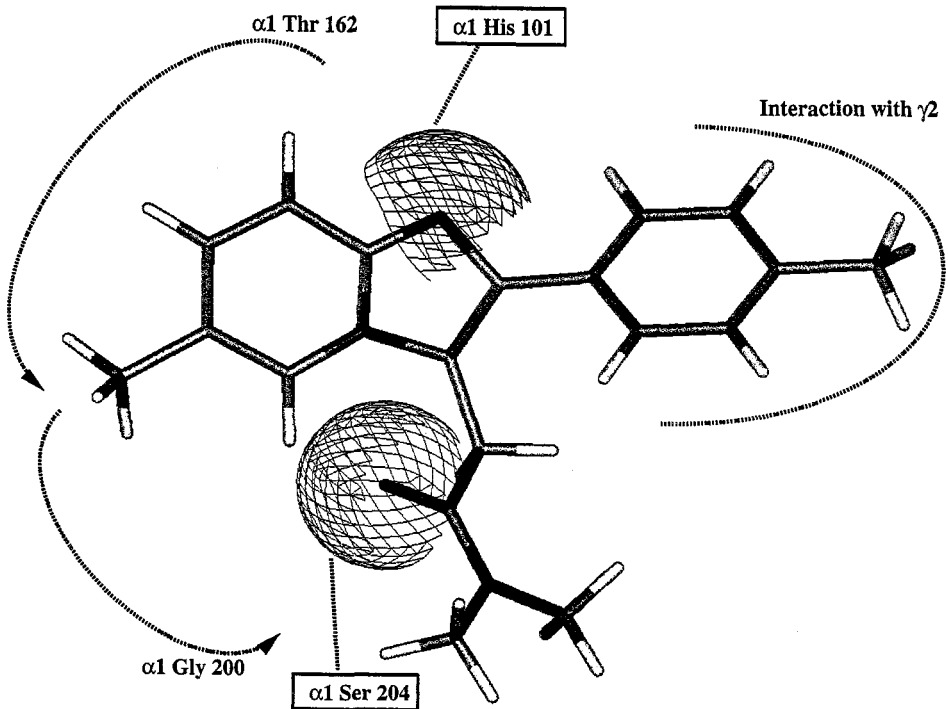


Figure 1. Interaction model between zolpidem and the  $\omega_1$  modulatory site.

**Poster Session VII**  
**Modelling of Membrane**  
**Penetration**

## SLIPPER — A NEW PROGRAM FOR WATER SOLUBILITY, LIPOPHILICITY AND PERMEABILITY PREDICTION

O. A. Raevsky, E. P. Trepalina, and S. V. Trepalin

Institute of Physiologically Active Compounds of Russian Academy of Sciences  
142432, Chernogolovka, Moscow Region, Russia

It is well-known that chemicals absorption, pharmacokinetics, protein binding, uptake in the brain and to certain extent hydrophobic drug-receptor interactions depends on lipophilicity, aqueous solubility and liposome permeability of compounds. That is why there are many approaches and commercially available programs for prediction these values. The major part of such approaches is based on fragmental or atom-based procedures.

It has been proposed <sup>1</sup> that lipophilicity encodes two major structural contributions: a volume-related term (describing steric bulk effects) and a term reflecting such interactions as dipole-dipole and hydrogen bonding. This approach has been laid by us in the basis of quantitative description of water solubility, octanol-water partition and permeability. First our researches in this field have been published in <sup>2-4</sup>.

The distribution coefficient octanol-water logP is predicted on the basis of the following formula:

$$\log P_{\text{oct}} = 0.266 \alpha - 1.00 \sum C_a^0 \quad (1)$$

where  $\sum C_a^0$  is the sum of overall free energy H-bond factors for all acceptor atoms in molecule,  $\alpha$  is a molecular polarizability, calculated in accordance to <sup>5</sup>.

Prediction of solubility is carrying out by using the equation (joint research with Dr. K.-J. Schaper, Borstel Research Institute, FRG):

$$\log S_w = -0.36 - 0.205 \alpha + 0.43 \sum C_a^0 - 0.26 \sum C_d \quad (2)$$

where  $\sum C_d$  is the sum of free energy H-bond factors for all donor atoms in molecule.

A new program SLIPPER (Solubility, LIPOphilicity, PERmeability) may be used for calculation aqueous solubility, lipophilicity and permeability. These properties depend on

pH of solvents and so in addition to the prediction all of these properties for neutral structures SLIPPER calculates these parameters for ionized structures participating in equilibria and complete pH-dependent profiles of solubility and lipophilicity (by using corresponding formula for acid-bases equilibria <sup>6</sup>).

Here we present the main features of the program SLIPPER briefly. For calculation pH-dependent octanol-water partition coefficient and water solubility profiles user should create a chemical structure of interest in the Structure Editor of the program or import it to a designated library using \*.sdf file. The logP and logS<sub>w</sub> for neutral forms are calculated upon closing the Structure Editor and then exiting the Data window or upon completion the Import procedure (if in \*.sdf file was only neutral form). In the Data window you can also add the other information: e.g., name, pK<sub>a</sub> values (when it is known or easily estimated) then after saving this information and closing the window SLIPPER will also calculate both values of lipophilicity and solubility for ionized forms. User may also get this information as a plot of logD-pH dependence (see fig.). By sliding the cursor along the profile curves the corresponding values of logP or logS<sub>w</sub> at any pH will be obtained.

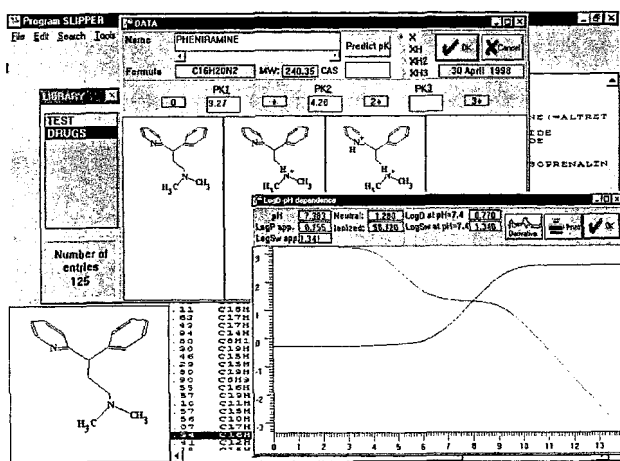


Fig. pH-dependent profiles of lipophilicity and water solubility for pheniramine.

Next version of the program SLIPPER will also predict pK<sub>a</sub> values and liposome permeability.

## REFERENCES

1. H.van de Waterbeemd, M.Kansy, B.Wagner, H.Fischer, Lipophilicity measurement by reversed-phase high performance liquid chromatography (RP-HPLC), in: Lipophilicity in Drug Action and Toxicology, V.Pliska, B.Testa and H.van de Waterbeemd, eds., VCH, Weinheim, 1996
2. O.A.Raevsky, K.-J.Schaper, J.K.Seydel, H-bond contribution to octanol-water partition coefficients of polar compounds, Quant. Struct.-Act. Relat., 14, pp.433-436 (1995)
3. O.A.Raevsky, Quantification of non-covalent interactions on the basis of the thermodynamic hydrogen-bond parameters, J.Phys.Org.Chem., v.10, pp.405-413 (1997)
4. O.A.Raevsky, in Computer-Assisted Lead Finding and Optimization, Eds. H van de Waterbeemd, B.Testa, G.Folkers, Wiley-VCH, Weinheim, 1997, 367-378
5. K.J.Miller, Additivity methods in molecular polarizability, J.Am.Chem.Soc., v.112, pp.8533-8538 (1990)
6. A.Avdeef, Assesment of distribution-pH profiles, in: Lipophilicity in Drug Action and Toxicology, V.Pliska, B.Testa and H.van de Waterbeemd, eds., VCH, Weinheim, 1996

## CORRELATION OF INTESTINAL DRUG PERMEABILITY IN HUMANS (*IN VIVO*) WITH EXPERIMENTALLY AND THEORETICALLY DERIVED PARAMETERS

Anders Karlén<sup>1</sup>, Susanne Winiwarter<sup>1</sup>, Nicholas Bonham<sup>1</sup>, Hans Lennernäs<sup>2</sup>, Anders Hallberg<sup>1</sup>

<sup>1</sup> Dept of Organic Pharmaceutical Chemistry and

<sup>2</sup> Dept of Pharmacy, Uppsala Biomedical Centre, Uppsala University, SE-751 23 Uppsala, Sweden

### INTRODUCTION

The extent of intestinal drug absorption, often described by the fraction of drug absorbed ( $F_a$ ), is governed by several different processes: (a) dose/dissolution ratio, (b) chemical degradation and/or metabolism in the lumen, (c) complex binding in the lumen, intestinal transit, and (d) effective permeability ( $P_{eff}$ ) across the intestinal mucosa. In many cases  $P_{eff}$  is considered to be the rate-limiting step in the overall absorption process and is therefore an interesting parameter in bioavailability studies.

However, due to experimental difficulties, very few correlation studies have been performed using  $P_{eff}$  values of drugs and nutrients determined *in vivo* in the human intestine. As part of constructing a Biopharmaceutical Classification System for oral immediate-release products<sup>1</sup> the human jejunal  $P_{eff}$  values for 22 compounds have been determined using a recently introduced experimental technique which enables direct estimation of the local absorption rate in humans.

The aim of the present investigation was to derive a QSAR equation by use of multivariate modelling which, based on these human *in vivo*  $P_{eff}$  values and relevant physicochemical descriptors of the above set of compounds, will allow for the prediction of passive absorption of drugs in the human intestine.

### METHODS

Two compound data sets were used in this study: *Data set 1* consists of 22 compounds for which human  $P_{eff}$  values have been determined. At least three different routes of transportation exists for these drugs. Fifteen of the compounds are passively absorbed and these form the basis for this study. *Data set 2* consists of the 22 drugs from data set 1 combined with a set of 136 drugs derived from an internet database of the Pomona College Medicinal Chemistry Project (<http://clogP.pomona.edu/medchem/chem/clogp/>) giving altogether 158 compounds.

*Data set 2* was used in the molecular diversity study in order to ensure that the molecules in data set 1 are representative of drugs in general.

*Lipophilicity measurements.* Determinations of  $pK_a$ ,  $\log P$  and  $\log P_{ion}$  values for the compounds in data set 1 were performed by use of the Sirius PCA101 potentiometric system<sup>2</sup>. Based on these experiments  $\log D$  values were calculated at pH 5.5, 6.5 and 7.4.

*Theoretical molecular descriptors.* The 22 drugs in data set 1 were built in their neutral form in an extended conformation using SYBYL<sup>3</sup>. All structures were minimized with the AM1 method<sup>4</sup> using the keywords PRECISE, XYZ and NOMM. Fourteen theoretical descriptors were used in this study: molecular weight (MW), molecular volume (V), molecular surface area (S), ovality (O), NATOM (number of atoms), E\_HOMO, E\_LUMO, hardness (H), dipole moment (DM), polar surface area (PSA), hydrogen bond donors (HBD, number of hydrogens connected to N- and O-atoms) and acceptors (HBA, number of O- and N-atoms in an appropriate functional group). The sum of HBD and HBA was denoted HB. ClogP values for the molecules in data set 1 were obtained from the drug compendium in Comprehensive Medicinal Chemistry (eds Hansch, Sammes and Taylor, 1990).

## STRATEGY

The following strategy was used to obtain statistically sound models that can be used to predict passive absorption of drugs in human from physicochemical data:

1. Characterization of the physicochemical properties of the compounds in data set 1 with experimentally determined  $\log P$  and  $\log D$  values and theoretically calculated molecular descriptors.
2. Calculation of the theoretical molecular descriptors also for the compounds in data set 2 and performance of a Principal Component Analysis (PCA) using SIMCA<sup>5</sup> on all theoretical data in order to check the molecular diversity of the 22 compounds of data set 1.
3. Selection of a training and a test set of compounds from the passively absorbed compounds in data set 1 according to statistical design principles based on the PCA above.
4. Investigation of the relationship between physicochemical variables and human *in vivo* permeability data of the training set compounds by PLS analysis.
5. Evaluation of the resulting PLS models by use of the test set of compounds.
6. Calculating final models based on both test and training set compounds.

## RESULTS

We were able to determine the pKa values for 18 and  $\log P$  values for 15 of the 22 compounds by use of the potentiometric method. In addition to these experimentally determined values 14 theoretical descriptors were calculated. Based on the score plot obtained from the PCA it could be shown that the 22 compounds of data set 1 are reasonably well separated implying that they are representative of drugs in general (step 2). Based on statistical design principle a training ( $n=5$ ) and a test ( $n=8$ ) set of passively absorbed compounds were selected (step 3). Several PLS models with good  $R^2$  and  $Q^2$  values could be developed by use of the training set compounds (step 4). These models were also evaluated by predicting  $\log P_{eff}$  for the test set compounds and determining the mean residuals for each model (step 5). Three models were selected as especially interesting and final models were calculated based on the 13 passively absorbed compounds for which all data existed (step 6).

## REFERENCES

1. Amidon, G. L., Lennernäs, H., Shah, V. P., Crison, J. R., 1995, A theoretical basis for a biopharmaceutical drug classification: The correlation of *in vitro* product dissolution and *in vivo* bioavailability, *Pharm. Res.*, 12, 413.
2. Avdeef, A., 1992, pH-Metric  $\log P$ . Part 1. Difference plots for determining ion-pair octanol-water partition coefficients of multiprotic substances, *Quant. Struct.-Act. Relat.* 11, 510.
3. SYBYL: Molecular Modeling Software, Tripos Associates, Inc.: St.Louis, MO 63144, 1996.
4. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., Stewart, J. J. P., 1985, AM1: A new general purpose quantum mechanical molecular model, *J. Am. Chem. Soc.*, 107, 3902.
5. SIMCA, Umetri AB, Box 7960: SE-90719 Umeå, Sweden, 1996

# A CRITICAL APPRAISAL OF LOGP CALCULATION PROCEDURES USING EXPERIMENTAL OCTANOL-WATER AND CYCLOHEXANE-WATER PARTITION COEFFICIENTS AND HPLC CAPACITY FACTORS FOR A SERIES OF INDOLE CONTAINING DERIVATIVES OF 1,3,4-THIADIAZOLE AND 1,2,4-TRIAZOLE

Athanasia Varvaresou, Anna Tsantili-Kakoulidou, Theodora Siatra-Papastaikoudi

Department of Pharmacy, Division of Pharmaceutical Chemistry, University of Athens, Panepistimiopoli, Zografou, 157 71, Athens, Greece

## INTRODUCTION

The accumulation of several heteroatoms in hybrid molecules may affect the safe prediction of lipophilicity, while such compounds may differentiate in their hydrogen bonding capability, also important in the manifestation of drug action. The title compounds, which belong to the general types **1,2,3,4** (Figure 1) have shown CNS and antimicrobial activities.<sup>1,2</sup> In this study their lipophilicity was investigated and compared to the values obtained by different calculative procedures. Their hydrogen bonding capability was also assessed through the  $\Delta\log P$  approach.<sup>3</sup>

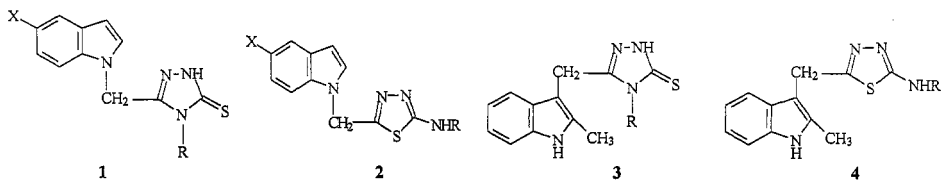


Figure 1. Structures of the investigated compounds

## MATERIAL AND METHODS

High Performance Liquid Chromatography was applied for the determination of extrapolated  $\log k_w$  values as lipophilicity indices.<sup>3</sup> Partition coefficients in octanol-water



( $\log P_{\text{oct}}$ ) and cyclohexane water ( $\log P_{\text{cyc}}$ ) were directly measured by the shaking flask method. Calculations of octanol-water  $\log P$  ( $\log P_{\text{calc}}$ ) were performed according to: modified Rekker's ( $\log P_{\text{cdr}}$ ), modified Ghose-Crippen ( $\log P_{\text{GC}}$ ) and Broto's ( $\log P_{\text{B}}$ , only for triazole derivatives) systems, implemented in the program PrologP, Suzuki-Kudo system ( $\log P_{\text{SK}}$ ) using Chemicalc-2 and ClogP (only for thiadiazole derivatives).

## RESULTS AND DISCUSSION

Extrapolated  $\log k_w$  values are found practically to coincide with octanol-water  $\log P$  values. In both sets of data a lower than expected lipophilicity was observed for triazoles when R is naphthalene, due to conformation effects. This effect cannot be considered by any of the calculation systems. Compounds of type 3 and 4 show the same or slightly lower lipophilicity than compounds of type 1 and 2. This observation is correctly reflected in  $\log P_{\text{GC}}$  and  $\log P_{\text{SK}}$ . In Rekker's, ClopP and Broto's systems the presence of the extra  $\text{CH}_3$  group and the hydrogen on the indole nitrogen considerably raise the lipophilicity. In Suzuki-Kudo system the thiadiazole derivatives are underestimated. Introduction of appropriate indicator variables leads to very good correlations between  $\log k_w$  (or  $\log P_{\text{exp}}$ ) and  $\log P_{\text{calc}}$  with  $r > 0.96$  for all calculation systems. Omitting the naphthalene derivatives of the triazoles, the regressor coefficients of  $\log P_{\text{calc}}$  shift towards 1 for all calculation systems, the intercept however remains relatively large in most cases.

Partially calculated  $\log P_{\text{cyc}}$  according to Rekker's available fragmental constants are generally higher than the experimental values.  $\Delta \log P$  values are  $\sim 0.5$  for the triazole derivatives. However, when X is  $-\text{NO}_2$ ,  $\Delta \log P$  increases reaching the value of 2. For the thiadiazole derivatives  $\Delta \log P$  is higher than for the corresponding triazoles, with values  $\sim 1$ , due to the presence of the aromatic  $-\text{NH}$  group.

## REFERENCES

1. A.Tsotinis, A.Varvaresou, Th.Calogeropoulou, Th.Siatra-Papastaikoudi, A.Tiligada, Synthesis and antimicrobial evaluation of indole containing derivatives of 1,3,4-thiadiazole and 1,2,4-triazole and their open-chain counterparts *Arzneim. Forschung* 47: 307 (1997)
2. A.Varvaresou, Th.Siatra-Papastaikoudi, A.Tsotinis, A.Tsantili-Kakoulidou, A.Vamvakides, Synthesis, lipophilicity and biological evaluation of indole containing derivatives of 1,3,4-thiadiazole and 1,2,4-triazole "*Il Farmaco* 53:320 (1998)
3. El Tayar, Testa B., Carrupt P.-A. Polar intermolecular interactions encoded in partition coefficients: a indirect estimation of hydrogen-bond parameters of polyfunctional solutes. *J.Phys. Chem.* 96:1455 (1992)
4. A.Tsantili-Kakoulidou, E.Filippatos, A.Papadaki-Valiraki, Use of reversed phase high performance liquid chromatography in lipophilicity studies of 9H-xanthene and 9H-thioxanthene derivatives containing an aminoalkanamide or a nitrosureido group. Comparison between capacity factors and calculated octanol-water partition Coefficients' *J.Chromatogr.A* 654: 43 (1993)

## DETERMINATION OF ACCURATE THERMODYNAMICS OF BINDING FOR PROTEINASE-INHIBITOR INTERACTIONS

Frank Dullweber, Franz W. Sevenich and Gerhard Klebe

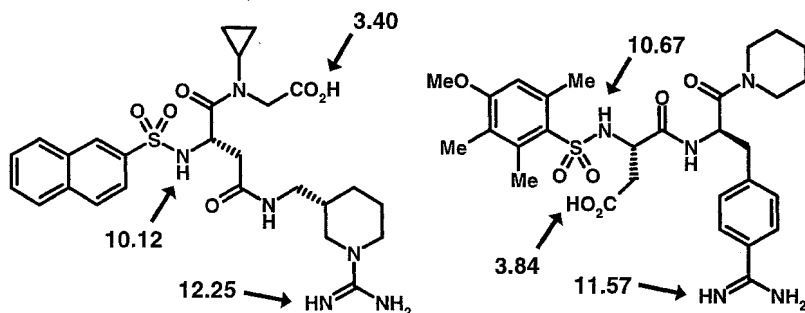
Philipps-Universität Marburg  
Department of Pharmaceutical Chemistry  
Marbacher Weg 6, 35032 Marburg/Germany

The affinity of a low-molecular weight ligand to a macromolecular target protein is usually described by the binding constant  $K_i$  that typically corresponds to a negative free energy of binding of 10-80 kJ/mol in aqueous solution. It comprises enthalpic and entropic contributions that arise from several underlying phenomena. To better understand and subsequently describe the binding process detailed measurements of these quantities are required.

The temperature-dependent measurement of  $K_i$  allows one to elucidate thermodynamic properties via van't Hoff plots, however since heat capacity is likely to change with temperature also  $\Delta H$  and  $\Delta S$  will be temperature-dependent.<sup>1</sup> As an alternative, isothermal titration calorimetry (ITC) provides direct access of the heat produced during the binding process.<sup>2</sup> The shape of the titration curve unravels the dissociation constant  $K_D$ .<sup>3</sup> We performed several measurements of  $K_D$  with various ligand binding either to thrombin, trypsin or thermolysin. In all cases we could demonstrate that  $K_D$ 's obtained by ITC correspond within the experimental errors to  $K_i$  values in literature resulting from photometric assays. We altered buffer and salt conditions, however no effect of affinity could be detected.

The integrated heat measured during an ITC experiment comprises all changes in enthalpy, among them the enthalpy of binding. The binding of napsagatran (1) to trypsin and thrombin shows considerable differences in  $\Delta H$  depending on the buffer conditions used. Three different buffers, tris, hepes and pyrophosphate have been applied. They show decreasing heat of protonation. Buffer dependence points to the release or capture of protons upon ligand binding. Potentiometric titrations of the three protonatable groups reveal three different pKa values (Fig. 1). Most likely the carboxy group uptakes a proton during binding. To verify this assumption, the ethyl ester of napsagatran has been studied and obviously no protonation step occurs during binding. The related thrombin inhibitor CRC 220 (2) also comprises three functional groups likely to be involved in protonation steps. Similar pKa values have been detected. However, no buffer dependence is observed

for this ligand. This surprising difference in behavior of (1) and (2) can be explained with respect to their distinct binding modes to thrombin. According to the crystal structure of napsagatran, the carboxy group is binding towards Ser 195 and the oxyanion hole.<sup>4</sup> Thus, it is fully buried into the binding site and hydrogen-bonded to His 57, Ser 195 and a neighboring water molecule. The captured proton is used in this H-bonding network. In contrast, the aspartate of CRC 220 orients to the rim of the binding pocket and remains largely solvent exposed only forming a hydrogen bond to the NH of Gly 219.<sup>5</sup> The local dielectric conditions experienced by the carboxy groups in the two inhibitors induce in the case of napsagatran such strong pKa shifts that protonation occurs. This shift spans several orders of magnitude since under aqueous conditions with a pKa of 3.40 napsagatran will be clearly deprotonated at a buffer pH of 7.8.



**Figure 1.** Potentiometric titration of napsagatran (1, left) and CRC 220 (2, right) reveal three different pKa values for the protonable groups

The present results demonstrate that ITC ligand binding studies require measurements from different buffer conditions in order to detect protonation/deprotonation along with ligand binding. This is a first step to decompose the measured integral heat into different contributions comprising among others the enthalpy of binding.

## REFERENCES

1. H. Naghibi, A. Tamura, J.M. Sturtevant, Significant discrepancies between van't Hoff and calorimetric enthalpies, *Proc. Natl. Acad. Sci. USA* 92:5597 (1995).
2. T. Wisemann, S. Williston, J.F. Brandts, L.N. Lin, Rapid measurement of binding constants and heat of binding using a new titration calorimeter, *Anal. Biochem.* 179:131 (1989).
3. D.R. Bundle, B.W. Sikurskjold, Determination of accurate thermodynamics of binding by titration calorimetry, *Methods Enzym.* 247:288 (1994).
4. K. Hilpert, J. Ackermann, D.W. Banner, A. Gust, K. Gubernator, P. Hadváry, L. Labler, K. Müller, G. Schmid, T.B. Tschoop, H. van de Waterbeemd, Design and synthesis of potent and highly selective thrombin inhibitors, *J. Med. Chem.* 37:3889 (1994).
5. M. Reers, R. Koschinsky, G. Dickneite, D. Hoffmann, J. Czech, W. Stüber, Synthesis and characterisation of novel thrombin inhibitors based on 4-aminidophenylalanine, *J. Enzyme Inhib.* 9:61 (1995).

## AUTHOR INDEX

- Ács, T., 338  
 Ahmed, S.A., 273  
 Akamatsu, M., 263, 286  
 Altomare, C.D., 353  
 Andersson, P., 65  
 Andersson, P.M., 27  
  
 Balzano, F., 183, 325, 433  
 Baringhaus, K.-H., 345  
 Barretta, G.U., 183, 325, 433  
 Barril, X., 129  
 Baskin, I.I., 468  
 Baurin, N., 349  
 Bautsch, W., 440  
 Beezer, A., 297  
 Beleta, J., 295  
 Benigni, R., 476  
 Berglund, A., 231  
 Besnard, F., 484  
 Bianchi, A., 369  
 Blomme, A., 404  
 Böhm, M., 103  
 Bonham, N., 491  
 Boström, J., 382  
 Bouzida, D., 425  
 Bradley, M.P., 282  
 Bradshaw, J., 474  
 Breton, P., 393  
 Bru, N., 393  
 Brünová, B., 390  
 Brusati, M., 95  
 Buelow, R., 111  
 Burden, F.R., 175  
 Bursi, R., 215  
  
 Cambria, A., 325  
 Carotti, A., 353  
 Carrieri, A., 353  
 Carrupt, P.-A., 353  
 Castorina, M., 342  
 Cavalli, A., 347  
 Cellamare, S., 353  
 Centeno, N.B., 141, 321  
 Chen, H., 433  
  
 Chen, H.-T., 47  
 Chiodi, P., 275  
 Christensen, I. Thøger, 231, 357, 373  
 Christensen, S. Brøgger, 316  
 Cima, M.G., 342  
 Clark, R.D., 95  
 Clementi, M., 207  
 Clementi, Sara, 207  
 Clementi, Sergio, 73, 207  
 Collantes, E.R., 201  
 Colominas, C., 129  
 Conraux, L., 404  
 Consolaro, F., 292  
 Consonni, V., 344  
 Contreras, J.-M., 53  
 Cox, J., 375  
 Cramer, C.J., 245  
 Cramer, R.D., 95  
 Crespo, M.I., 295  
 Cronin, M.T.D., 273  
 Cross, G.J., 448  
 Cruciani, G., 73, 89, 207, 265,  
 321, 329, 334, 369  
  
 da Rocha, R.K., 480  
 Damborský, J., 401  
 De Cillis, G., 375  
 de la Torre, R., 141  
 De Winter, H., 429  
 Dean, P.M., 410, 412, 442, 455  
 Dearden, J.C., 273  
 Dimoglo, A.S., 418  
 do-Amaral, A.T., 290  
 Dohalsky, V.B., 311  
 Domány, G., 338  
 Doménech, T., 295  
 Doweiko, A., 183  
 Drew, M.G.B., 284, 453  
 Dullweber, F., 103, 495  
 Duraiswami, C., 323  
 Durant, F., 404, 482  
  
 Edman, M., 27  
 Engels, M., 429  
  
 Eriksson, L., 65, 271  
 Ertl, P., 267  
 Even, Y., 484  
  
 Fängmark, I., 293  
 Farrell, N., 375  
 Faust, M., 292  
 Feltl, L., 311  
 Fernandez, E., 446  
 Fichera, M., 369  
 Filipek, S., 195  
 Finizio, A., 292  
 Fioravanzo, E., 375  
 Fletterick, R.J., 380  
 Ford, M., 474  
 Ford, M.G., 301, 303  
 Frokjaer, S., 231  
  
 Gago, F., 321, 329  
 Gallo, G., 275, 342  
 Galvagni, D., 344  
 Gasteiger, J., 157  
 Gehlhaar, D.K., 425  
 George, P., 404, 482, 484  
 Gerasimenko, V.A., 423  
 Giannangeli, M., 359  
 Giesbrecht, A., 290  
 Giuliani, A., 476  
 Glick, M., 458  
 Głowka, M.L., 299  
 Gohlke, H., 103  
 Goldblum, A., 440, 458  
 Golender, L., 336  
 Gomes, S.L., 290  
 González, M., 141  
 Gottmann, E., 464  
 Gràcia, J., 295  
 Grädler, U., 103  
 Graham, D., 484  
 Gramatica, P., 292, 344  
 Grassy, G., 111  
 Gratteri, P., 334  
 Greco, G., 347  
 Guba, W., 89

- Guccione, S., 183, 325, 361, 433  
 Guenzler-Pukall, V., 345  
 Guillaumet, G., 349  
 Gundertofte, K., 382  
 Günther, E., 397  
  
 Hallberg, A., 388, 491  
 Hammarström, L.-G., 293  
 Handschuh, S., 157  
 Hansen, L.M., 365  
 Haque, N., 442  
 Helma, C., 464  
 Hemmer, M.C., 157  
 Hendlich, M., 103  
 Heritage, T., 95  
 Hermens, J.L.M., 245  
 Herndon, W.C., 47  
 Higata, T., 263  
 Hiltunen, R., 377  
 Hirono, S., 363, 399  
 Hoare, N.E., 303  
 Hoffmann, R.D., 318  
 Höfgen, N., 395  
 Höltje, H.-D., 135  
 Hongming, C., 183  
 Hopfinger, A.J., 323  
 Høst, J., 373  
 Hou, X.J., 384  
 Hovgaard, L., 231  
 Howlett, A.C., 201  
 Hubbard, R.E., 371  
 Hudson, B.D., 303  
 Huuskonen, J.J., 377, 470  
  
 Ikeda, I., 263  
 Ivakhnenko, A.G., 444  
 Ivanov, A.A., 307  
 Iwase, K., 363  
  
 Jandera, A., 390  
 Janssen, L.H.M., 386  
 Javier Luque, F., 129  
 Jilek, R., 95  
 Johansson, E., 65, 271  
 Jönsson, P.G., 293  
 Jørgensen, F.S., 357, 373  
 José, A.M., 480  
 Jurs, P.C., 249  
  
 Kaczorek, M., 111  
 Källblad, P., 455  
 Kansy, M., 237  
 Karlén, A., 388, 491  
 Kasheva, T.N., 472  
 Katakura, S., 380  
 Kharazmi, A., 316  
 Kissinger, C.R., 384  
 Klebe, G., 103, 495  
 Kleinöder, T., 157  
 Kmojëek, V., 390  
  
 Knegtel, R.M.A., 380  
 Kocjan, D., 406  
 Koenig, J.-J., 404  
 Koike, K., 263  
 König, M.A., 361  
 Kovalishyn, V.V., 444, 472  
 Kramer, S., 464  
 Krarup, L.H., 231  
 Kratzat, K., 237  
 Krause, G., 397  
 Kuchar, M., 390  
 Kühne, R., 397  
 Kuntz, I.D., 380  
 Kutscher, B., 397  
  
 Lahana, R., 111  
 Langer, T., 318, 361  
 Laoui, A., 408  
 Laszlovszky, I., 338  
 Lemcke, T., 357  
 Lemmen, C., 169  
 Lengauer, T., 169  
 Lennernäs, H., 491  
 Liljefors, T., 316, 365, 367, 382  
 Linton, M.A., 384  
 Linusson, A., 27  
 Lippi, F., 474  
 Livingstone, D.J., 444, 472  
 Lloyd, E.J., 448  
 Longfils, G., 482  
 Lopes, J.C.D., 480  
 López, M., 295  
 López-de-Briñas, E., 141  
 López-Rodríguez, M.L., 446  
 Loza, M.I., 355  
 Lozano, J.J., 141, 321  
 Lozoya, E., 355  
 Lucic, B., 288  
 Luik, A.I., 444, 472  
 Lukavsky, P., 318  
 Lumley, J.A., 453  
 Lundstedt, T., 27  
  
 Mabilia, M., 275, 342, 359, 375  
 Madhav, P.J., 323  
 Magdó, I., 338  
 Maggiora, G.M., 83, 427  
 Malpass, J., 301  
 Manallack, D.T., 371  
 Mancini, F., 359  
 Mannhold, R., 265  
 Marino, M., 325  
 Marot, C., 349  
 Martynowski, D., 299  
 Matter, H., 123  
 McFarland, J.W., 221, 280  
 McFarlane, S.L., 293  
 Melani, F., 334  
 Mérour, J.Y., 349  
 Mestres, J., 83  
 Meurice, N., 427  
  
 Miklavc, A., 406  
 Milanese, C., 359  
 Mills, J.E.J., 410, 412  
 Mochida, K., 263  
 Modica, M., 183, 433  
 Montana, J.G., 371  
 Montanari, C.A., 297, 314, 446, 480  
 Montanari, M.L.C., 297  
 Morin-Allory, L., 349, 393  
 Motohashi, N., 286  
 Mpoke, S., 380  
 Mungala, N., 380  
 Murphy, P.V., 371  
 Muskal, S.M., 249  
 Musumarra, G., 369  
  
 Nakayama, A., 340  
 Ness, A.L., 293  
 Nevell, T.G., 303  
 Nielsen, S.F., 316  
 Nikaido, T., 263  
 Nilsson, J.E., 207  
 Nilsson, L., 269  
 Niwa, S., 416  
 Nordén, B., 27  
 Norman, P.R., 293  
 Norrby, P.-O., 365, 367  
 Novellino, E., 347  
 Novic, M., 59, 305  
  
 Ohmoto, T., 263  
 Olczak, A., 299  
 Olivier, A., 404, 482, 484  
 Ooms, F., 482  
 Oono, S., 416  
 Orozco, M., 129  
 Oshiro, C.M., 380  
 Osmond, N.M., 293  
 Ozoe, Y., 263  
  
 Pajeva, I., 414  
 Palacios, J.M., 295  
 Palyulin, V.A., 460, 468  
 Parrilla, I., 237  
 Pastor, M., 73, 207, 321, 329  
 Pawlak, D., 195  
 Pelletier, L.A., 384  
 Perkins, T.D.J., 442  
 Petit, J., 478  
 Pfahringer, B., 464  
 Pino, A., 476  
 Pires, J.R., 290  
 Pisano, C., 342  
 Poirier, P., 404  
 Polymeropoulos, E.E., 395, 397  
 Pompe, M., 59, 305  
 Price, N.R., 284, 453  
  
 Radchenko, E.V., 460  
 Raevsky, O.A., 221, 280, 423, 489

- Ramos, E.U., 245  
 Rayan, A., 440  
 Recanatini, M., 347  
 Rejto, P.A., 425  
 Renard, P., 349  
 Renard, S., 484  
 Rival, Y., 53  
 Rohrer, D.C., 83  
 Romanelli, M. Novella, 334  
 Rosado, M.L., 446  
 Rose, V.S., 462  
 Rosenfeld, R., 336  
 Rucki, M., 311  
 Rum, G., 47  
 Ryder, H., 295
- Sacks, J., 149  
 Sadowski, J., 157  
 Sakurai, K., 416  
 Salt, D., 474  
 Salt, D.W., 301  
 Sandberg, M., 27, 65, 231, 271  
 Santagati, A., 183, 433  
 Santagati, M., 183, 433  
 Santaniello, M., 275  
 Sanz, F., 141, 321, 355  
 Sarpietro, M., 325  
 Scapecchi, S., 334  
 Schaper, K.-J., 221, 261, 446  
 Schischkow, G., 361  
 Schleifer, K.-J., 135  
 Schneider, B., 390  
 Schubert, G., 345  
 Schwab, C.H., 157  
 Schwab, W., 123  
 Segarra, V., 295  
 Segura, J., 141  
 Senner, S., 237  
 Sevenich, F.W., 495  
 Sevrin, M., 404, 482, 484  
 Shapiro, S., 277  
 Sharra, J.A., 273  
 Shim, J.-Y., 201  
 Showalter, R.E., 384  
 Shvets, N.M., 418
- Siatra-Papastaikoudi, T., 493  
 Siew, N., 440  
 Sippl, W., 53  
 Sjöström, M., 27  
 Skillman, A.G. Jr., 380  
 Snyder, F.D., 3  
 Snyder, J.P., 3  
 Somoza, J.R., 380  
 Staszewska, A., 299  
 Sukekawa, M., 340  
 Summo, L., 353
- Tagmose, L., 365  
 Takahashi, M., 416  
 Tassoni, E., 342  
 Tatlock, J.H., 384  
 Taylor, R.J.K., 371  
 Teckentrup, A., 157  
 Tehan, B.G., 448  
 Tempczyk, A., 384  
 ter Laak, A.M., 397  
 Testa, B., 353  
 Tetko, I.V., 444, 470, 472  
 Tichy, M., 311  
 Tinti, M.O., 275, 342  
 Todeschini, R., 292, 344  
 Tolan, J.W., 249  
 Tollenaere, J.P., 429  
 Tomic, S., 269  
 Toro, C.M., 359  
 Tot, E., 135  
 Trepalin, S.V., 423, 489  
 Trepalina, E.P., 489  
 Trinajstic, N., 288  
 Tsantili-Kakoulidou, A., 493  
 Tsuchida, K., 399  
 Turner, D., 277  
 Turner, D.B., 331  
 Tysklind, M., 65
- Ueno, T., 263  
 Uppgård, L.-L., 27
- Vaes, W.H.J., 245  
 van de Waterbeemd, H., 221
- van Geerestein, V.J., 215  
 Vangrevelinghe, E., 393  
 Varvaressou, A., 493  
 Veber, M., 305  
 Vercauteren, D.P., 427, 478  
 Verhaar, H.J.M., 245  
 Vighi, M., 292  
 Villa, A.E.P., 472  
 Villafranca, J.E., 384  
 Vorpagel, E.R., 336  
 Vracko, M., 466  
 Vuorela, H., 377
- Wade, R.C., 269  
 Wagener, M., 157  
 Wagner, B., 237  
 Waller, C.L., 282  
 Wang, C.C., 380  
 Watkins, R.W., 453  
 Weidmann, K., 345  
 Welsh, W.J., 201  
 Wermuth, C.G., 53  
 Wessel, M.D., 249  
 Wiese, M., 414  
 Wilkerson, W.W., 280  
 Willett, P., 331  
 Winger, M., 318  
 Winiwarter, S., 388, 491  
 Winkler, D.A., 175  
 Wold, S., 27, 65, 271  
 Wong, M.G., 448  
 Wood, H.J., 284  
 Wood, J., 462  
 Wouters, J., 482  
 Wyatt, J.A., 303
- Yamagami, C., 286  
 Yamaotsu, N., 399  
 Yasri, A., 111  
 Young, S. Stanley, 149
- Zefirov, N.S., 460, 468  
 Zhang, Y., 47  
 Zupan, J., 59, 305

## SUBJECT INDEX

- Absorption, 249  
Active site, 347  
Activity, Estimation, 111, 195, 377  
ADME, 13  
Affinities, 123, 399  
Agonists, 7, 365, 388, 397  
Alignment, 318  
Antagonists, 7, 334, 382, 404, 416  
Antimutagenic activity, 286  
APEX-3D, 336
- Beta-turn mimetics, 388  
Binding  
    affinities, 107, 365, 369, 397, 495  
    cavity, 410  
    constants, 406  
    energy, 480  
    sites, 135, 207, 263, 395  
Bioactivity, 305  
Bioavailability, 13, 238  
Bioinformatics, 27
- CATALYST, 318, 345, 409  
Chemometrics, 207  
Classification, 429, 477  
Combinatorial chemistry, 27  
COMBINE, 269, 321, 329  
CoMFA (Comparative Molecular Field Analysis), 183  
    analysis, 314  
    applications, 216, 286, 303, 338, 349, 361  
    prediction, 318, 377, 414  
    receptor mapping, 183  
    target-based, 53, 124, 347  
Comparative modelling, 325  
Complexation energies, 366, 367  
Computational site-directed mutagenesis, 401  
CoMSIA, 124  
Conformational analysis, 183, 373  
Conformational studies, 393  
Conformer sampling, 363  
Continuum regression, 301
- De novo design, 361, 410  
Descriptors, 95, 157, 267, 277, 482
- DISCO, 203, 416  
Distance clustering, 462  
Diversity, 95, 423, 442  
DNA, 480  
DNA adducts, 375  
Docking, 129, 425  
D-optimal design, 232
- Electron Topology, ETM, 418  
Entropic trapping, 406  
EVA, 278, 331
- Fingerprints, 474  
Flexibility, 162, 386  
Flexible fitting, 171  
Flexible ligands, 412  
FlexS, 170  
4D-QSAR, 323  
Free-Wilson analysis, 261, 269
- Genetic algorithms, 288, 427, 453  
GERM, 433  
GOLPE, 53, 317  
GPCR, 5, 113, 207, 355, 455  
GRID, 54, 74, 89, 316, 334, 370  
GRID/GOLPE, 124, 321, 329
- HASL, 183  
Henry's law, 273  
High-Throughput Screening, 149, 175, 237, 429  
Hydrogen bonding, 221, 280, 410, 412, 458
- Inhibitor, Interactions, 390, 495  
Inhibitors  
    AChE, 53  
    calcineurin, 384  
    cell adhesion, 371  
    CYP1, 141, 347  
    DHFR, 305, 357  
    DNA-gyrase, 299  
    Ftase, 408  
    glycogen phosphorylase, 329  
    HIV protease, 442  
    kinases, 361

- Inhibitors (*cont.*)
  - MAO-B, 353
  - metalloproteinase, 123
  - PDE 4, 295, 395
  - platelet aggregation, 318
  - prolyl 4-hydroxylase, 345
  - purine salvage enzyme, 380
  - reverse transcriptase, 427
- Interactions
  - drug-DNA, 480
  - protein-ligand, 103, 355, 359, 367, 386, 390, 484, 495
- Kohonen maps, 158, 444, 478
- Kohonen network, 158
- Lipophilicity, 223, 265, 489
- LUDI, 362
- Machine learning algorithms, 464
- Microcalorimetry, 297
- Model building, 355
- Model validation, 271
- MOLDIVS, 423
- Molecular descriptors, Specmat, 215
- Molecular design, 33
- Molecular dynamics simulations, 399
- Molecular Field Analysis (MFA), 196
- Molecular representations, 175
- Multivariate design, 27, 65
- Neural networks
  - artificial, 446, 466, 468, 470
  - baysian, 177
  - genetic algorithm, 251, 288
  - Kohonen network, 158, 444
- Nonlinear mapping, 307
- Opioid peptides, 195
- PARM (Pseudoatomic Receptor Model), 183, 433
- Partition coefficients, 245, 311, 470, 493
- PCBs, 284
- Peptide absorption, 231
- Peptides, 111, 232, 336, 388, 416
- Peptidomimetics, 408
- Permeability, 223, 237, 489, 491
- Pharmacophore
  - alignment, 196, 349
  - development, 136, 141, 201, 382, 416, 448
  - identification, 303, 336, 373
- Pharmacophores, in general, 7
- Protein engineering, 401
- Pseudoreceptors, 136
- QSAR/CoMFA, 353
- QSPR, 249, 273, 466
- Receptor maps, 204
- Receptor models, 183, 433
- Receptor Surface Analysis (RSA), 196
- Receptors, 3, 440, 446, 455, 478, 484
- Recursive partitioning, 149
- Resistance, 357, 414
- RigFit, 169
- SAR by NMR, 6
- Screening of databases, 169
- Selectivity, 107, 123, 357, 382
- SERM, 373
- Similarity, 47, 83, 340, 423, 427
- Site-directed drug design, 410
- Site-directed mutagenesis, 484
- Solubility, 223, 237, 489
- Solvation, contributions to, 129
- SRD/GOLPE, 370
- Stabilization, 367
- Statistical design, 293, 316
- Structure-based design, 329, 380, 384, 425
- Substrates, 141, 275, 321
- 3D representation
  - SWIM, 344
  - SWM, 344
  - influence of, 59
- 3D-QSAR
  - alignment, 318
  - CoMFA, 286, 338, 349
  - methodology, 73, 340, 461
  - models, 316, 334, 345
  - studies, 135, 321, 369
- 3D-SAR, 342
- Toxicity, 292
- Variable selection
  - by neural networks, 472
  - validation, 282
- Virtual Receptor, 178
- VolSurf, 74, 90
- Water accessible surface area, 232
- World Wide Web, Descriptors on, 267