Sci Pharm. 2008; 76: 5–18 doi:10.3797/scipharm.0802-05

The Similarity Principle – New Trends and Applications in Ligand-Based Drug Discovery and ADMET Profiling

Rita SCHWAHA, Gerhard F. ECKER *

Emerging Field Pharmacoinformatics, Department of Medicinal Chemistry,
University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

Abstract

Structural similarity is one of the basic underlying principles in drug discovery and development. Numerous algorithms and concepts are known to search compound libraries for analogous compounds assuming that similar compounds show similar biological activity. In recent years the focus shifted towards more complex methods using 3D-shape similarities on one side and highly reductionistic approaches such as smiles substrings on the other side. Furthermore, pharmacological activity profiling becomes increasingly important. Within this review we will highlight selected new concepts and methods applying similarity metrics in the drug discovery and development process.

Keywords

similarity calculation • shape similarity • pharmacological activity profiling • in silico screening

Introduction

Nowadays the drug development process starts with hits obtained in HTS assays. Up to 1.000.000 compounds are biologically screened on a yes/no basis and the resulting hits are prioritised on basis of novelty, patentability, synthetic accessibility and data obtained in early ADMET (Absorption, Distribution, Metabolism, Elimination, Toxicity) profiling programs. A typical HTS library consists of both a so called historical collection (i.e. compounds from previous drug development programs) and commercially obtained substances, which are selected

* Corresponding author: Tel.: +43-1-4277-55110; Fax: +43-1-4277-9551.
E-mail: gerhard.f.ecker@univie.ac.at (G. F. Ecker).

on basis of maximum chemical diversity combined with high drug likeliness. In parallel, *in silico* screening approaches are gaining increasing importance. They are mainly used to select subsets of large virtual combinatorial libraries, which then should show a higher incidence for biological activity (or at least higher drug likeliness) and thus lead to increased hit rates.

The basic underlying principle of both active subset selection and structure-activity relationship studies assumes that compounds similar to biologically active ones should also be active and *vice versa*. This principle basically is also applied in bioinformatics, where proteins sharing similar sequences are considered to have similar structure and exhibit similar function. Thus, the calculation of similarity is one of the most intensively studied area and numerous methods and algorithms have been published. Furthermore, within the past decade also QSAR-approaches based on similarity measures have been reported in the literature. Most of them use N x N similarity matrices as input vector [1]. Very interesting in this context is the chemical global positioning system (ChemGPS) developed by Oprea. Selection of a set of satellite structures with extreme values of standard descriptors and a set of representative drugs (core structures) allowed developing a unique mapping device for the drug-like chemical space [2]. Furthermore, the combination of ChemGPS and VolSurf enabled a pharmacokinetically based mapping of compounds with respect to permeability and solubility [3]. The concept of similarity is also underlying the LASSOO approach of H. Villar [4]. The LASSOO algorithm intends to prioritise compounds that are most similar to a specified set of favourable target molecules (i.e. actives) and, at the same time, most different from compounds that reside in this set.

With this article we do not intend to give a detailed and complete overview on all methods available for calculating similarity values and applying them in the drug discovery and development process. We rather will focus on a few selected, successful examples of applying similarity metrics for virtual screening and pharmacological profiling. With respect to all aspects related with pharmacophores

and pharmacophore searches we would like to refer to the excellent book edited by T. Langer and R. Hoffmann [5].

Similarity values used as Descriptors

In a standard workflow for screening libraries, the concept of similarity is used as a filter criterion for classification purposes. Seri-Levy et al. used shape similarity as a single independent variable in QSAR equations and demonstrated higher predictive abilities for their approach than for multivariate analyses [6]. However, they restricted their method to homologous series of compounds. Ghuloum et al. used molecular hashkeys based on molecular surface similarity [7]. This inspired us to explore the concept of using similarity values as independent variables in QSAR equations. Within SIBAR (Similarity-based SAR), similarity values between training set compounds and a set of reference compounds are calculated and subsequently used as molecular descriptors. The approach for calculation of the SIBAR-descriptors is outlined as follows (Fig. 1):

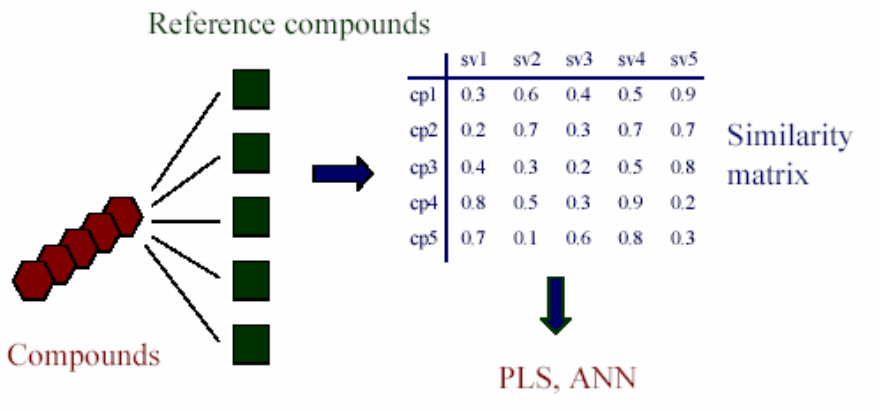

Fig. 1. SIBAR generation process.
1) selection of a reference compound set on basis of maximum diversity; additionally, it may be divided into active and inactive compounds, substrates and non-substrates,
2) calculation of a set of descriptors for both the training set and the reference set
3) calculation of similarity values for each compound of the training set to each compound of the reference set; this leads to a given number of similarity values (equal to the number of reference compounds used) for each compound of the training set, which are assigned as SIBAR-descriptors; similarity may be calculated in various ways, i.e. euclidian distances, city block distances or Tanimoto similarity
4) MLR, PLS or SVM analysis of the training set data matrix
5) Validation of the model using cross validation procedures and external test sets.

So far the approach was successfully applied for a set of 131 propafenone-type inhibitors of P-gp. 100 compounds were used in the training set. The 20 most diverse compounds of the SPECS library were used as reference set. SIBAR-descriptors were calculated using 39 physicochemical and topological descriptors as implemented in TSAR. Subsequent PLS analysis led to models with a predictive power which is significantly higher than those obtained when using the descriptors alone. When using a small data set for gastrointestinal absorption (19 compounds), predictivity equal to those when using polar surface area was obtained [8]. This proves the principal applicability of the approach. It has to be stressed out, that in both cases the reference set was not tailored and optimised for the respective problem (i.e. choosing a set of structurally diverse active and inactive P-gp inhibitors or a set of highly and low absorbed compounds, respectively). Recent results obtained by B. Zdrazil showed that amongst a panel of 4 different reference sets (A: highly diverse, drug like compounds; B: P-gp inhibitors from our in house library; C: P-gp substrates from the literature; D: chemicals) models derived using reference sets related to the target P-gp (B, C) gave highest internal (leave one out cross validation) and external (test set) predictivity [9]. Most strikingly, when applying the SIBAR concept on a small panel of P-gp inhibitors and calculating shape-similarity values instead of using Euclidian distances (3D-SIBAR) the results further improved remarkably [10].

3D-Shape Similarity

Calculating similarity values on basis of 3D-structures imposes the additional problem of conformational sampling. However, although computationally demanding, there are several reports which clearly demonstrate the advantage of considering the molecules as three-dimensional entities. As outlined above, 3D-shape similarity values, as implemented in ***MIMIC***, remarkably improve the performance of the SIBAR approach [10]. MIMIC is a molecular field matching program for quantitatively evaluating the similarity between two molecules on basis of steric and electrostatic field [11]. Considering the fact, that MIMIC only operates

with steric and electrostatic fields and that H-bonds and π-π interactions play a dominant role in drug receptor interactions, a further improvement of using this 3D-SIBAR approach should be obtained when shape similarities are calculated on basis of molecular interaction fields utilising the full panel of probe atoms (H-bond acceptor, H-bond donor, hydrophobic, aromatic,…). This method is utilised in the software package ***MIPSIM***, which compares 3D distribution of both molecular electrostatic potentials derived from quantum chemical calculations and molecular interaction fields of series of biomolecules [12, 13]. Obviously, the computational costs are by far higher than those for MIMIC and require front-end computational tools.

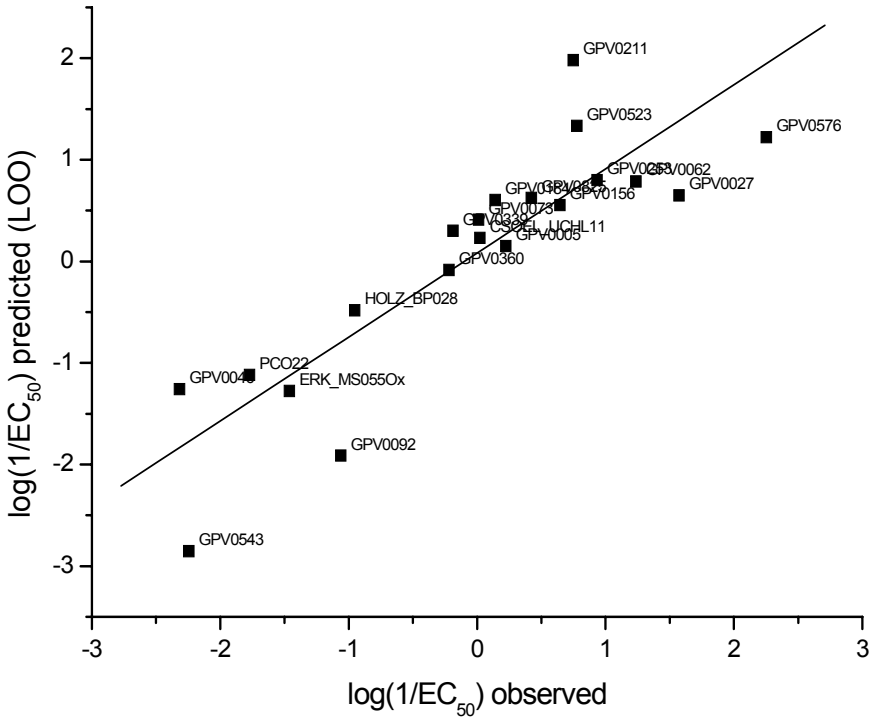

Fig. 2. Plot of observed vs. predicted activity (LOO) of a diverse data set of 20 P-glycoprotein inhibitors derived from a model based on 3D-shape similarities [10].

Recently, the group of A. Jain reported the successful application of the concept of morphological similarity, implemented in ***Surflex-Sim*** [14]. This method optimizes the pose of a query molecule to an object molecule in order to maximize 3D similarity. Morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid, thus yielding a value from 0 to 1. The function is dependent on the relative alignment of two molecules. Results presented assess the utility of the method for ligands of the serotonin, histamine, muscarinic, and $GABA_A$ receptors. In virtual screening runs, Surflex-Sim was able to distinguish true ligands from random compounds just on basis of two or three known ligands. Although the true positive rates are quite moderate (60%), the false positive rates are impressingly low (3%). In selected runs the enrichment rates might reach 150-fold compared to random screening. The method thus is a valuable tool for *in silico* screening and also offer a competitive alternative to structure-based methods.

ROCS *(Rapid Overlay of Chemical Structures)* performs shape-based overlays of conformers of a candidate molecule to a query molecule in one or more conformations. The algorithm is based on the shape comparison method described by Masek et al. [15] and quickly finds and quantifies the maximum overlap of the volume of two molecules. The overlays are based on a description of the molecules as atom-centered Gaussian functions. ROCS maximizes the rigid overlap of these Gaussian functions and thereby maximizes the shared volume between a query molecule and a single conformation of a database molecule. Applying this method to inhibitors of the ZipA-FtsZ protein-protein interaction enabled Rush et al. [16] to identify two new antibacterial lead compounds. In comparison to 2D-similarity search methods, the new lead structures retrieved showed quite diverse chemical structures, which further strengthens the ability of the method to perform scaffold hops

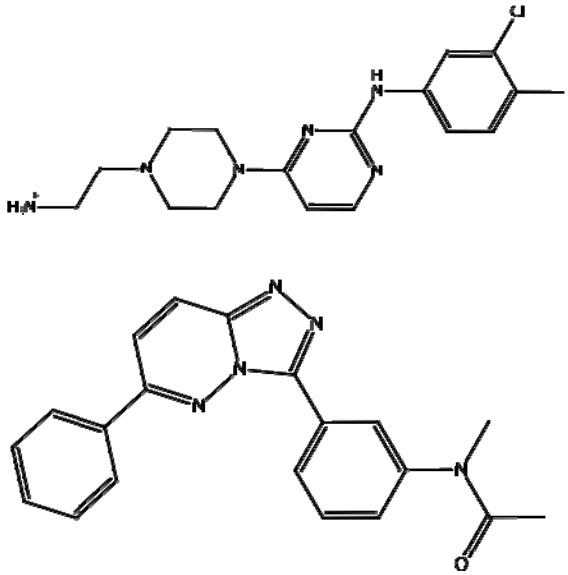

Fig. 3. Above: query structure, below: hit structure found with ROCS by Rush and co-workers [16].

Virtual Pharmacological Profiling

Recently several interesting approaches have been published related to activity profiling. These mainly rely on very fast similarity calculations on basis of descriptors derived from topological information. Gregori-Puigjane and Mestres [17] developed a novel set of descriptors called SHED (Shannon Entropy Descriptors), which are based on the distributions of atom-centered feature pairs extracted directly from the topology of the molecules. The basic assumption behind follows the general line of this review, i.e. molecules having similar features arranged in a similar way should display similar SHED profiles and in consequence also exhibit similar activity profiles. These new descriptors have subsequently been successfully applied for virtual pharmacological profiling for a set of nuclear receptors [18].

Also the group of A. Jain performed a ligand-based modelling of a set of biological targets applying the concept of morphological similarity [19]. Using an annotated data base of roughly 1000 compounds and 270 targets, the authors computed pairwise 3D-similarity values in order to characterise both the ligand and the target space. Drug pairs sharing a target showed significantly higher similarity than the “background” of drug pairs sharing no target. Also, when comparing

targets on basis of their ligands those with no overlap in annotated drugs shared lower similarity. Furthermore, using a sub set of 22 targets, the authors achieved enrichment factors of up to 100 fold when using these similarity derived models for virtual screening of compound libraries. Finally, screening selected compounds over the whole panel of target models (virtual pharmacological profiling), a number of known side effects and drug-drug interactions were identified.

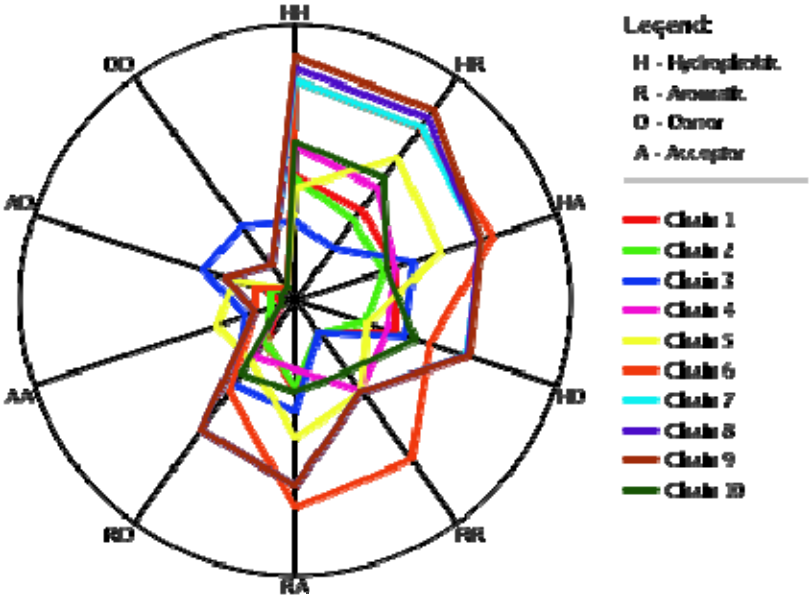

Fig. 4. SHED Profile of 10 compounds of our in-house MDR-Database.

Keiser et al. made an additional step in using ligand similarity to quantitatively group related proteins together [20]. Tanimoto index (TI) based pair wise similarity values of all ligands annotated to a given target were calculated and compared with the respective values of a large drug like chemical space. Thorough statistical analysis and comparison of whole sets of ligands enabled the construction of a minimum spanning tree. This tree, although solely based on ligand similarity, revealed biologically sensible clusters. Furthermore, hitherto unexpected links between targets and compounds were identified. This shows that the rather simple principle of similarity calculations based on Tanimoto indices allows a ligand-based description and clustering of the biological target space. Further attempts of the

group will focus on *in silico* activity- and selectivity profiling of compounds over a large set of targets.

Similarity Metrics and beyond

Besides the sdf-file format smiles strings are one of the most common text based molecular representations. Implicitly a smiles string contains all the information and thus may be directly used for calculation of molecular properties. Vidal et al. developed a method for direct use of smiles strings for calculation of biophysical properties without the necessity of time-consuming conversion of the strings into molecular graphs or 3D structures [21]. First, smiles strings are fragmented into overlapping substrings of a size of 4 digits, called LINGO. This integral set of substrings represents a hologram of the smiles representation of the molecule. This LINGO profile can subsequently be used as input vector for QSAR analysis and similarity calculations. Thus, using a set of 12831 compounds from the PHYSPROP database, a logP model with an r^2 value of 0.93 could be derived. Also the calculation of aqueous solubility worked equally well. Finally, LINGO strings were applied to pair wise similarity calculations and revealed a statistically significant difference between bioisosteric compounds (average TI of 0.36) and random pairs (average TI of 0.07).

Recently, the LINGO concept was expanded to the IUPAC names. Also IUPAC names are a unique string of characters connected to the molecular structure of the compounds. The principle concept stays the same, i.e. the name is fragmented into sub strings of 4 characters which represent a hologram of the chemical structure. This vector was subsequently successfully applied in QSPR studies predicting logP values, solvent accessible surface area, free energy of salvation in water and last but not least also the logarithm of the blood-brain barrier partition coefficient [22].

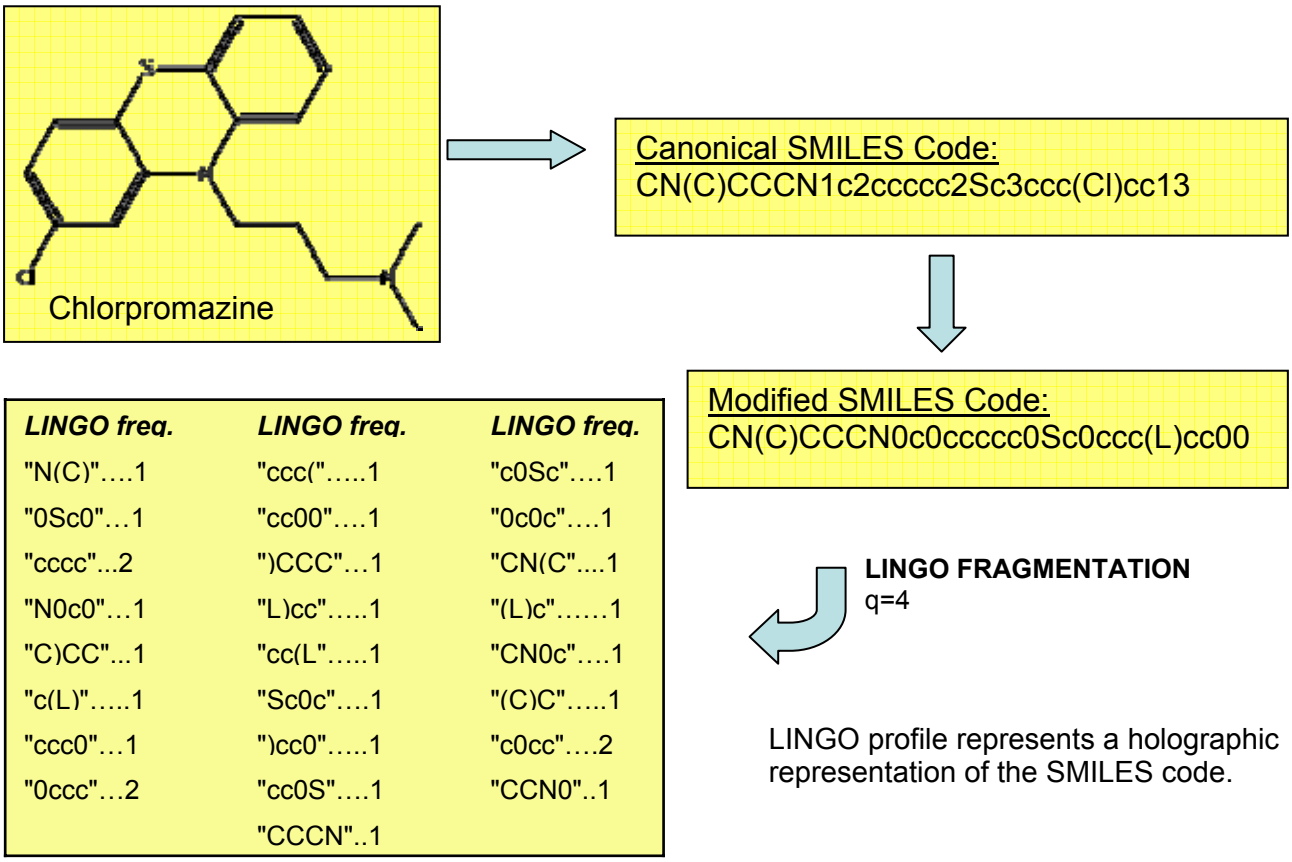

Fig. 5. LINGO generation process, taken from the work of Vidal et al. 2005 [21]:

Outlook

Despite the methods and applications mentioned in this overview several other interesting approaches for utilising similarity metrics in drug discovery have recently been published. Among others, these include a comparison of topological, shape and docking methods in virtual screening [23], a combined approach using molecular shape and electrostatics by Nicholls et al. 2004 [24], a very recent idea of using 2D Pharmacophore Feature Triplet Vectors for classification by Paul Watson [25] and the development of an alignment-recycling method using reference shapes and thereby enhancing speed [26]. This clearly demonstrates the still high importance of similarity calculations. However, the scope shifted from similarity searching to pharmacological activity profiling. Considering the fact that knowledge about protein networks and their importance for drug efficacy and safety is steadily

increasing, the challenge will move towards pharmacological profiling of compounds in the context of dynamic protein networks. This will require front end Pharmacoinformatics tools and algorithms and will definitely include the use of similarity metrics

	$LINGO_i$	f_i
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	(2-h	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	2-he	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	-hex	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	hexa	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	exad	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	xade	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	adec	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	deca	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	ecan	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	cano	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	anoy	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	noyl	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	oylo	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	ylox	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	loxy	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	oxy-	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	xy-3	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	y-3-	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	-3-h	1
(2-hexadecanoyloxy-3-hydroxy-propyl)_hexadecanoate	3-hy	1

Fig. 6. LINGO profile generation, taken from the work of Thormann et al. 2007 [22]. The set of four-character LINGOs are shown in red.

Acknowledgement

We gratefully acknowledge financial support by the Austrian Science Fund, grant # L344-N17.

References

[1] Kubinyi H, Hamprecht FA, Mietzner T.
Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices.
J Med Chem. 1998; 41: 2553–2564.
[doi:10.1021/jm970732a]

[2] Oprea TI, Gottfries J.
Chemography: The Art of Navigating in Chemical Space.
J Comb Chem. 2001; 3: 157–166.
[doi:10.1021/cc0000388]

[3] Oprea TI, Zamora I, Ungell AL.
Pharmacokinetically Based Mapping Device for Chemical Space Navigation.
J Comb Chem. 2002; 4: 258–266.
[doi:10.1021/cc010093w]

[4] Koehler RT, Dixon SL, Villar HO.
LASSOO: A Generalized Directed Diversity Approach to the Design and Enrichment of Chemical Libraries.
J Med Chem. 1999; 42: 4695–4704.
[doi:10.1021/jm990312g]

[5] Langer T, Hoffmann RD.
Pharmacophores and Pharmacophore Searches.
In: Methods and Principles in Medicinal Chemistry. 1st ed. Volume 32.
Weinheim: Wiley-VCH; 2006.

[6] Seri-Levy A, West S, Richards WG.
Molecular Similarity, Quantitative Chirality, and QSAR for Chiral Drugs.
J Med Chem. 1994; 37: 1727–1732.
[doi:10.1021/jm00037a025]

[7] Ghuloum AM, Sage CR, Jain AN.
Molecular Hashkeys: A Novel Method for Molecular Characterization and Its Application for Predicting Important Pharmaceutical Properties of Molecules.
J Med Chem. 1999; 42: 1739–1748.
[doi:10.1021/jm980527a]

[8] Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF.
Similarity based SAR (SIBAR) as tool for early ADME profiling.
J Comput-Aided Mol Des.
2002; 16: 785–793.
[doi:10.1023/A:1023828527638]

[9] Zdrazil B, Kaiser D, Kopp S, Chiba P, Ecker GF.
Similarity-Based Descriptors (SIBAR) as Tool for QSAR Studies on P-Glycoprotein Inhibitors: Influence of the Reference Set.
QSAR Comb Sci. 2007; 26: 669–678.
[doi:10.1002/qsar.200610149]

[10] Gregori E, Zdrazil B, Chiba P, Kopp S, Mestres J, Ecker GF, et al.
Shape similarity values as tool for QSAR-studies on inhibitors of P-glycoprotein.
ACS 231st Spring National Meeting; Atlanta: ACS; 2006.

[11] Mestres J, Rohrer DC, Maggiora GM.
A molecular field-based similarity approach to pharmacophoric pattern recognition.
J Mol Graphics Modell. 1997; 15: 114–121.
[doi:10.1016/S1093-3263(97)00003-X]

[12] De Càceres M, Villà J, Lozano JJ, Sanz F.
MIPSIM: similarity analysis of molecular interaction potentials.
Bioinformatics. 2000; 16: 568–569.
[doi:10.1093/bioinformatics/16.6.568]

[13] Barbany M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J.
Towards a MIP-based alignment and docking in computer-aided drug design.
Proteins: Struct, Funct, Bioinf. 2004; 56: 585–594.
[doi:10.1002/prot.20153]

[14] Jain AN.
Ligand-Based Structural Hypotheses for Virtual Screening.
J Med Chem. 2004; 47: 947–961.
[doi:10.1021/jm030520f]

[15] Masek BB, Merchant A, Matthew JB.
Molecular shape comparison of angiotensin II receptor antagonists.
J Med Chem. 1993; 36: 1230–1238.
[doi:10.1021/jm00061a014]

[16] Rush TS, Grant JA, Mosyak L, Nicholls A.
A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction.
J Med Chem. 2005; 48: 1489–1495.
[doi:10.1021/jm040163o]

[17] Gregori-Puigjane E, Mestres J.
SHED: Shannon Entropy Descriptors from Topological Feature Distributions.
J Chem Inf Model. 2006; 46: 1615–1622.
[doi:10.1021/ci0600509]

[18] Mestres J, Martin-Couce L, Gregori-Puigjane E, Cases M, Boyer S.
Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling.
J Chem Inf Model. 2006; 46: 2725–2736.
[doi:10.1021/ci600300k]

[19] Cleves AE, Jain AN.
Robust Ligand-Based Modeling of the Biological Targets of Known Drugs.
J Med Chem. 2006; 49: 2921–2938.
[doi:10.1021/jm051139t]

[20] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK.
Relating protein pharmacology by ligand chemistry.
Nat Biotechol. 2007; 25: 197–206.
[doi:10.1038/nbt1284]

[21] Vidal D, Thormann M, Pons M.
LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities.
J Chem Inf Model. 2005; 45: 386–393.
[doi:10.1021/ci0496797]

[22] Thormann M, Vidal D, Almstetter M, Pons M.
Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names.
Open Appl Inf J. 2007; 1: 28–32.
[doi:10.2174/1874136300701010028]

[23] McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD.
Comparison of Topological, Shape, and Docking Methods in Virtual Screening.
J Chem Inf Model. 2007; 47: 1504–1519.
[doi:10.1021/ci700052x]

[24] Nicholls A, MacCuish NE, MacCuish JD.
Variable selection and model validation of 2D and 3D molecular descriptors.
J Comput-Aided Mol Des. 2004; 18: 451–474.
[doi: 10.1007/s10822-004-5202-8]

[25] Watson P.
Naive Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors.
J Chem Inf Model. 2008; 48: 166–178.
[doi:10.1021/ci7003253]

[26] Fontaine F, Bolton E, Borodina Y, Bryant S.
Fast 3D shape screening of large chemical databases through alignment-recycling.
Chem Cent J; 2007; 1: 12.
[doi:10.1186/1752-153X-1-12]

Received February 20th, 2008
Accepted February 29th, 2008
Available online at www.scipharm.at March 30th, 2008