

MORE THAN COLOURED BLOBS – “FUZZY” DESCRIPTIONS OF POTENTIAL PHARMACOPHORE POINTS

**GISBERT SCHNEIDER^{1*}, EWGENIJ PROSCHAK¹,
KRISTINA GRABOWSKI¹, PETRA SCHNEIDER² AND
YUSUF TANRIKULU¹**

¹Johann Wolfgang Goethe-University, Institute of Organic Chemistry and Chemical Biology and Institute of Cell Biology and Neuroscience, Siesmayerstr. 70, 60323 Frankfurt am Main, Germany

²Schneider Consulting GbR, George-C.-Marshall Ring 33, 61440 Oberursel, Germany

E-Mail: * gisbert.schneider@modlab.de

Received: 20th September 2006 / Published: 5th November 2007

ABSTRACT

The design of focused compound libraries aims at enriching bioactive molecules that contain different scaffold structures. Pharmacophore-based similarity searching has been shown to provide a means to achieve this goal. We have developed such a method (LIQUID) that is grounded on the representation of potential pharmacophore points by trivariate Gaussian densities. This “fuzzy” pharmacophore technique is described and discussed in detail, together with a retrospective virtual screening application. LIQUID succeeded in retrieving activity-enriched sets of compound with diverse backbone architecture.

INTRODUCTION

A crucial task in molecular design is the selection of “focused libraries”, that is, a set of compounds exhibiting a desired pharmacological profile. Without any prior knowledge about potential structure-activity relationships or known reference compounds one would have to perform random guessing to find active molecules. The number of possible subsets p of size k from a pool containing N molecules is given by Equation 1. For example, there

are more than 10^{13} possibilities to sample a set of 10 compounds out of 100. A realistic scenario would be to pick 100 or 1000 molecules for testing from a corporate library containing more than one million substances. It is evident that brute-force examination of each possible subset in turn (“full enumeration”) is not feasible. Maximum diversity methods aim at covering the variability of the complete compound pool within a carefully chosen small subset. Cell- and dissimilarity-based clustering and partitioning methods are employed for this purpose [1]. Maximally diverse compound sets often represent reasonable starting points for screening campaigns. Focused libraries, in contrast, typically contain substances only from a certain region (“activity island”) of the chemical space defined by the pool compounds. Generic filtering steps for drug- or leadlike compounds in conjunction with target-specific prediction and selection tools have been shown to be suited for designing activity-enriched focused libraries [2].

$$p = \frac{N!}{k!(N-k)!} \quad (1)$$

$$ef = \frac{actives_{found}}{actives_{expected}}. \quad (2)$$

Successful library design results in a set of compounds containing more actives than a randomly picked subset on average. The enrichment factor ef quantifies this advantage (Equation 2). Figure 1 displays the result of a similarity method, for example a pharmacophore search or any other similarity search, in comparison to the ideal outcome and unbiased (“random”) picking. Such enrichment curves help assess the usefulness of a virtual screening method but it requires several known actives that can be used in *retrospective* virtual screening. Enrichment factors for the simulated experiment shown in Fig. 1 can be obtained as follows: The ef for the first 10% of the screened compounds is $(33/10)=3.3$ (33 actives found, 10 actives expected), for the first 50% of the screened compounds it computes $(85/50)=1.7$ (85 actives found, 50 actives expected). It is generally assumed that a method that performed reasonably well in such an experiment is also likely to succeed in a *prospective* screening study. We wish to point out that this does not necessarily have to be the case since some similarity methods can be strongly biased toward a certain drug target or screening database and thus do not represent generally applicable tools.

The screening compound pool represents the chemical space from which a focused library is compiled by virtual screening approaches such as similarity searching or machine learning techniques. These methods essentially represent predictions that have a certain precision that depend on the molecular descriptors and the similarity measure, among other variables (Fig. 2). Let us assume the accuracy of a prediction tool would be high meaning that most of the compounds picked for the focused library do actually bind to the target with a binding constant below a certain threshold θ (e. g. 1 μ M). This outcome will result in a high enrichment factor. A less accurate method would yield a lower hit rate according to the θ boundary, and as a consequence the calculated enrichment factor would be compar-

ably poor. Does this mean that the second method is unsuited for focused library design? The answer depends on what we actually wish to achieve: Higher activity enrichment does not necessarily mirror great chemotype diversity among the hits. The scenario depicted in Fig. 2b shows that the compound library covers only a small portion of chemical space. According to the similarity principle the architecture of these compounds should not be very different. Broadening the area covered by a library can help increase structural diversity but often comes at the price of decreased activity enrichment. Many retrospective studies have been performed to compare virtual screening methods based on enrichment factors, but only few have considered chemotype diversity of the library as an additional quality criterion.

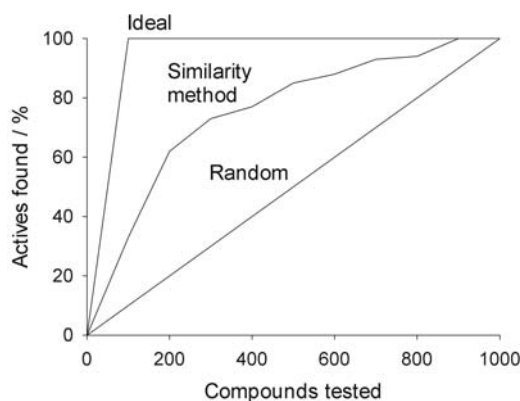


Figure 1. Enrichment curves obtained from simulated similarity searching in a pool of 1000 compounds containing 100 active substances. The ideal method finds all actives on ranks 1 to 100, random picking retrieves one active every 10 compounds, and a successful similarity search lies between these extremes.

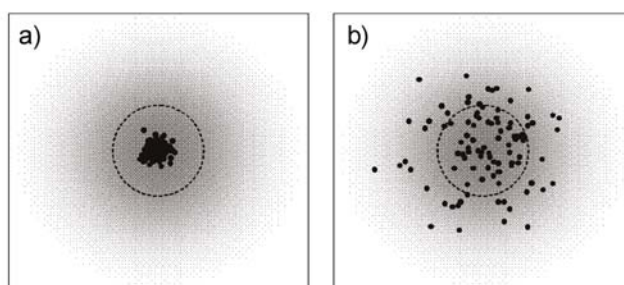


Figure 2. Hitting a target area in chemical space using a virtual screening method with high (a) and low (b) precision. The “fuzzier” method (b) covers a broader area. Shading indicates the biological activity (e.g. target binding) of the compounds, which is highest at the centre of the target area. The dashed circle represents the activity threshold θ .

LIQUID: FUZZY PHARMACOPHORE METHOD BASED ON TRIVARIATE GAUSSIANS

One possibility to increase structural diversity in a focused library is broader sampling as shown in Fig. 2 (for example, by using several prediction methods); a second one is to “fuzzify” the molecular descriptor [3]. Our program LIQUID (*Ligand-based Quantification of Interaction Distributions*) provides such a “fuzzy” pharmacophore method, which can be used for virtual screening (unpublished; a demo version is available at www.modlab.de).

According to the Medicinal Chemistry Section of IUPAC a pharmacophore is the “ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” [4]. According to this concept, ligand-receptor interactions are a function of individual functional group contributions. Since we do not know *a priori* which functional group actually contributes to the interaction they are termed “potential pharmacophore points” (PPPs). Ligand-receptor interactions take place in 3D space. Therefore 3D pharmacophore models represent the most intuitive choice. Noteworthy, in the absence of a receptor-relevant ligand conformation or conformation ensemble, quantitative structure-activity relationship (QSAR) studies that are based on 3D models can still be erroneous. In addition to the problem of conformer generation, an error-prone step in pharmacophore matching methods is the 3D-alignment of molecular features, that is, matching a screening molecule to a given pharmacophore model. To enable rapid database searching the explicit alignment step can be avoided by an *alignment-free* representation of pharmacophore patterns. One idea is to convert the spatial distribution of PPPs to a vector representation. Such vectors are referred to as “fingerprints”, “bitstrings”, “correlation vectors” (CV), or “spectra” depending on the type of information stored. The trick is to compare these reduced molecular representations instead of explicit 3D feature alignment [5], thus formulating a pharmacophore search as a similarity search.

In our software LIQUID, PPPs are modeled as trivariate Gaussian distributions and encoded as CV representations. The statistical spread of every PPP reflects the fuzziness of the pharmacophore model: The probability of a certain interaction decreases with the increasing distance to the PPPs centroid. Three-dimensional visualization tools, e.g. OpenGL [6] embedded in PyMOL [7], enable us to visually analyse generated pharmacophore models. An example of a LIQUID pharmacophore model is shown in Fig. 3, derived from a molecular alignment of the COX-2 inhibitors Rofecoxib, M5, and SC-558.

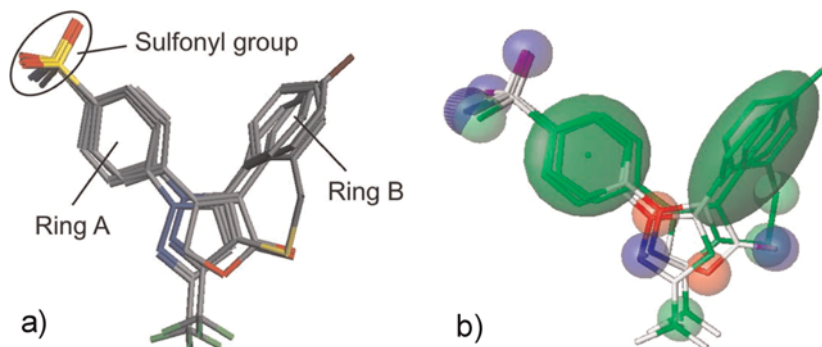


Figure 3: **a)** Molecular alignment of three COX-2 inhibitors (Rofecoxib, M5, SC-558). The common chemical groups, which are essential interactions for specific COX-2 inhibition, are labelled according to ref. [8]. **b)** LIQUID Pharmacophore model of **a)**: the visualization shows the ellipsoidal PPP models of both ring systems, which are members of the maximum common substructure of the respective superposition.

LIQUID can be used to compute a pharmacophore model of either a single molecular conformation or a conformer ensemble. The first fundamental step is the atom type recognition, where potential pharmacophore features are assigned to the query's atoms. We consider three different interaction types: “lipophilic”, “hydrogen-bond donor” and “hydrogen-bond acceptor”. For example, oxygen atoms are always hydrogen-bond acceptor interaction points, whereas the -OH group also possesses a hydrogen-bond donor feature. An atom can represent none, one or maximal two of these features. Atoms lacking a pharmacophore feature are not considered for the model. Atom typing results in a transformation of the query molecule(s) into a three-dimensional disposition of interaction points. To gather a fuzzy approximation of this interaction field via Gaussian functions, single interaction points are clustered into PPPs. Every PPP represents a local maximum of the Gaussian distribution of interaction points it contains.

To determine the maxima of the interaction point distribution the cluster radius dependent local-feature-density (LFD) was introduced [3]. It allows a quantification of common-type atoms in the spatial environment around an atom. Hereby the manually adjustable cluster radius is used to constrain the space around an atom where the maximum is determined. This also enables the user to generate pharmacophore models with varying fuzziness. The LFD of the k^{th} atom of pharmacophore type T is calculated by Equation 3.

$$LFD(atom_k^T) = \sum_{i=1}^n \max\left\{0, 1 - \frac{D_2(atom_k^T, atom_i^T)}{r_c}\right\}, \quad (3)$$

Where n is the total number of atoms of the pharmacophore type T in the query, D_2 is the Euclidean distance of two atoms, and r_c represents the cluster radius of the specific pharmacophore type. The closer a common-typed atom, the bigger is its impact to the

considered atom's LFD. Clustering of the interaction points is done on the basis of the atoms' LFDs via a Union-Find strategy. The following pseudo-code illustrates how the algorithm works:

```

INIT: each atom is a singleton.

FOR each atom  $i$  of type  $T$ 

  FOR each atom  $j$  of type  $T$ 

    calculate Distance( $i, j$ )

    IF Distance  $\leq$  ClusterRadius rc/sub>/emph> THEN

      FIND maxLFD(Cluster $_i$ )

      FIND maxLFD(Cluster $_j$ )

      IF maxLFD(Cluster $_i$ )  $\leq$  maxLFD(Cluster $_j$ ) THEN

        UNION Cluster $_j$  with Cluster $_i$ 

```

Initially, every atom represents a singleton. If a common-typed atom yielding a higher LFD inside the cluster radius of the considered atom is found, both atoms get “united” into a cluster. Finally, each cluster of interactions points (feature-typed atoms) forms a PPP. One can observe that the number of final clusters depends on the adjusted cluster radius. The centroid of a PPP is calculated as geometric centre of its clustered atoms.

We apply the principal component analysis (PCA) [9] in order to compute the size and orientation of a PPP. The covariance matrix is built up from the Cartesian coordinates of the clustered atoms in relation to the PPPs centroid. The orientation of a PPP is given by the resulting principal components, because their directions span the data space according to the highest variances. Eigenvector approximation is done with the NIPALS algorithm [10]. The corresponding Eigenvalues provide the distribution of the PPP in the direction of the Eigenvectors.

After having obtained the position, size and orientation of the PPPs, we encode the pharmacophore model as a correlation vector (Equation 4) [3, 5]. LIQUID computes a correlation-vector from the trivariate Gaussian functions, which are used to model the PPPs. Due to pair-wise PPP correlation, we encounter six PPP pairs: “lipophilic-lipophilic”, “lipophilic-donor”, “lipophilic-acceptor”, “donor-donor”, “donor-acceptor” and “acceptor-acceptor”. Encoding yields an equally partitioned bin vector. Each bin (vector element) contains the correlated frequency of a certain PPP pairing. In the case of PPP instances, the

correlation vector reduces to scaled occurrence frequencies of atom pairs at distance intervals from 1 to 20? [5, 11]. Each bin contains a correlated probability, which indicates the presence of a PPP pair at distance d :

$$CV_d^{A,B} = \frac{1}{\# \text{pairs}(A,B)} \sum_i^A \sum_j^B \cdot \frac{1}{2} \{ \text{trivG}(\sigma_{1,2,3})_i \cdot \text{trivG}(\sigma_{1,2,3})_j \}, \quad (4)$$

for all PPPs i of pharmacophore type A and for all PPPs j of type B . trivG gives the trivariate Gaussian with standard deviations $\sigma_{1,2,3}$, and $\# \text{pairs}(A,B)$ is the number of pairs of PPPs of type A and B . The result is a 120-dimensional correlation vector-based descriptor designed for fast virtual screening.

THE INFLUENCE OF THE GAUSSIAN FUNCTION

We have experimented with different parameters of the bell-shaped curve as an approximation of the PPPs. In addition to the normal Gaussian distribution (Equation 5), we have omitted the linear scaling factor and used an altered function (Equation 6) instead. Finally we expanded the PPPs by multiplying the standard deviation by the factor of two (Equation 7). Essentially, Equation 6 and Equation 7 compute a probability value of one at the centre of a PPP, and Equation 7 corresponds to a wider PPP shape than Equation 6 (Fig. 4). In other words, Equation 7 leads to a “fuzzier” PPP representation than Equation 6 and Equation 5.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \quad (5)$$

$$f(x) = \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \quad (6)$$

$$f(x) = \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{2\sigma}\right)^2\right). \quad (7)$$

In Fig. 5 the effects of Equations 5–7 on CV calculation are illustrated. We observe a general broadening of the CV density (increased number of populated bins) when increasing the “fuzziness” of a PPP. The CV produced by using Equation 5 has the tendency to underestimate broad PPPs (large standard deviation σ) compared to narrow ones (small σ). Figure 5a illustrates two PPPs with an identical standard deviation. The correlation of these two PPPs results in the black correlation vector shown in Fig. 5 d. If we double the standard deviation of one of the PPPs (Fig. 5b), we observe a dramatic reduction of the CV values (white CV in Fig. 5 d). Considering that the trivariate Gaussian distribution is a product of three univariate distributions, this effect increases in the 3D-space. These considerations led to the conclusion that Equation 5 is a poor choice for use as a fuzzy function for PPPs. It penalizes the absence of a small PPP harder than the absence of a large PPP. Fortunately, this problem can be overcome by omitting the linear normalization factor of Equation 5

which leads to Equation 6. Figure 5e illustrates this effect: The dominant CV values have the same height for both cases, and the correlation of the narrow PPP with a wider one leads to slightly expanded distribution of the CV. Employing Equation 7 as the fuzzy function for the PPPs shown in Figs 5a and 5b leads to a CV with even more populated bins (Fig. 5f).

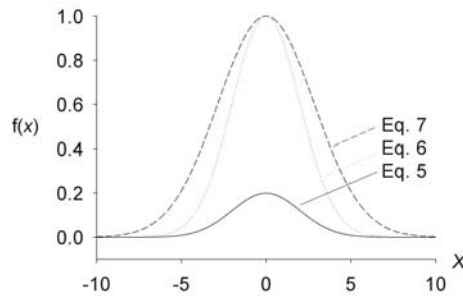


Figure 4. Probability distributions obtained by Equations 5–7 for zero-centred Gaussian data with $\sigma=2$. The curves represent PPP density along an axis.

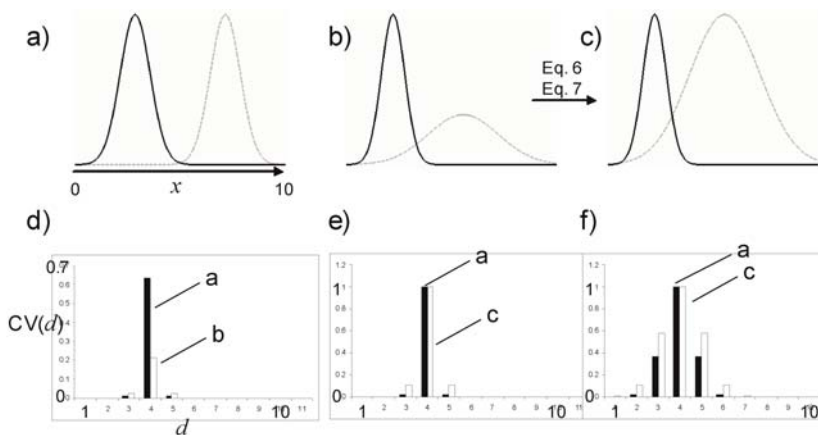


Figure 5. Effect of PPP approximation by Equations 5–7 on correlation vector (CV) calculation. x is a Cartesian coordinate in 3D space; d is the correlation distance.

a) – c) each represent the densities of two different PPPs.

d) Effect of Equation 5: *black*: CV of the PPPs shown in (a), *white*: CV of the PPPs shown in (b).

e) Effect of Equation 6: *black*: CV of two PPPs shown in (a), *white*: CV of the PPPs shown in (c).

f) Effect of Equation 7: *black*: CV of two PPPs shown in (a), *white*: CV of two PPPs shown in (c).

ENRICHMENT OF ACTIVES AND SCAFFOLDS: A RETROSPECTIVE VIRTUAL SCREENING STUDY

Table 1 contains results we obtained using fuzzy functions 5–7 in retrospective virtual screening. For the three ligand classes tested (COX-2, ACE, thrombin; taken from the COBRA collection v.4.6 [12]), we observe a continuous increase in the enrichment of actives from Equation 5 to Equation 7. Obviously, a fuzzier PPP representation is beneficial for retrieving actives in a given library size (here: 60 compounds). Noteworthy, the standard deviations of the average ef values are high. Therefore, this conclusion should be treated with sufficient caution.

Table 1. Enrichment factors ($ef \pm$ standard deviation) yielded for the first percent ($k=60$) of the ranked compounds ($N=6,046$) with different fuzzy functions. COX-2: cyclooxygenase 2, ACE: angiotensin converting enzyme.

Target	Fuzzy function		
	Eq. 5	Eq. 6	Eq. 7
COX-2 (95 actives)	11 \pm 7	14 \pm 10	15 \pm 10
ACE (74 actives)	7 \pm 5	8 \pm 6	10 \pm 7
Thrombin (204 actives)	7 \pm 6	9 \pm 6	10 \pm 7

We then analysed scaffold diversity to investigate the degree of scaffold-hopping in the focused libraries (top-ranking 60 molecules). Two definitions of “scaffold” were employed: i) “Murcko scaffold” [13], that is, the side-chain-depleted atomic scaffold of a molecule, *retaining* all information about atom types and bond order, and ii) “reduced scaffold” [14, 15], the side-chain-depleted molecular graph *ignoring* information about atom types, bond order and ring size. The latter is the more abstract representation of the molecular architecture. While the number of Murcko scaffolds in a library defines chemotype diversity, the number of reduced scaffolds is related to the diversity of molecular shape. For the COX-2 and the ACE example our scaffold analysis reveals that despite high enrichment factors (Table 1) only few (COX-2) or just a single (ACE) scaffold were retrieved, irrespective of the scaffold definition (Table 2, Table 3). Note that the numbers in Table 2 and Table 3 give the numbers of scaffolds that were retrieved by *all* reference compounds. This means that the fuzzy functions facilitated scaffold-hopping in the case of thrombin whereas for the COX-2 and ACE data the approach did not show the desired outcome. Why is that? For both ACE and COX-2, a single scaffold dominates the reference compounds (ACE: amide backbone; COX-2: “mickey mouse” motif). As a consequence, for these two ligand families the number of scaffolds found by all reference compounds among the top-scoring 60 molecules is limited. Clearly, most of the top-ranking molecules (here: in the first percentile) belong to the same chemotype, rendering the “enrichment” listed in Table 1 arguable.

This study demonstrates that i) providing only the enrichment of actives obtained by a virtual screening method is of limited value for the medicinal chemist and for assessing the applicability of a virtual screening tool, and ii) the data sets used for retrospective screening were biased towards some few scaffolds.

Table 2. Murcko scaffolds retrieved by all reference compounds in the first percentile (60 molecules) of the ranked compound lists: #scaffolds_of_actives (#scaffolds_total).

Target	Fuzzy function		
	Eq. 5	Eq. 6	Eq. 7
COX-2 (51 scaffolds)	3 (52)	1 (52)	5 (54)
ACE (40 scaffolds)	1 (47)	1 (49)	1 (50)
Thrombin (143 scaffolds)	13 (49)	19 (50)	21 (51)

Table 3. Reduced scaffolds retrieved by all reference compounds in the first percentile (60 molecules) of the ranked compound lists: #scaffolds_of_actives (#scaffolds_total).

Target	Fuzzy function		
	Eq. 5	Eq. 6	Eq. 7
COX-2 (27 scaffolds)	3 (46)	1 (44)	3 (47)
ACE (33 scaffolds)	1 (44)	1 (45)	1 (44)
Thrombin (111 scaffolds)	11 (44)	16 (49)	21 (51)

Still, our findings actually indicate successful scaffold-hops, in particular for the thrombin ligands. Apparently, this set of reference compounds was less balanced toward some few molecular frameworks, and the higher degree of PPP fuzziness actually facilitated the retrieval of more scaffolds. This provides additional support for Equation 7 as a PPP fuzzy function. The greatest number of scaffolds was compiled using this fuzzy PPP function (up to 21 out of 111 reduced scaffolds present in the thrombin ligands).

Surprisingly, we did not observe differences of the fuzzy functions regarding the total numbers of Murcko scaffolds (Table 2) and reduced scaffolds (Table 3) compiled in a library. This means that the three equations appear to be equally well suited for *overall* scaffold retrieval: between 47 and 54 different Murcko scaffolds and between 44 and 51 different reduced scaffolds were compiled in a library of only 60 molecules. We conclude that the concept of fuzzy pharmacophores by trivariate PPPs seems to be useful in retrieving both active molecules and many different scaffolds in one library – a desired outcome of virtual screening [16, 17]. Extended retrospective analyses and ongoing prospective studies will help assess this preliminary conclusion.

ACKNOWLEDGEMENTS

The Beilstein-Institut zur Förderung der Chemischen Wissenschaften is warmly thanked for support of our research and organization of the Bozen Workshop. We wish to express our gratitude to all participants of the workshop who challenged us by thought-provoking questions; in particular, Drs Carsten Kettner, Martin Hicks, Christof Steinbeck, Tim Clark and Paul Labute contributed to this publication by inspiring after-session discussions (“coloured-blob research”).

REFERENCES

- [1] Leach, A.R., Gillet, V.J.(2003) *An Introduction to Chemoinformatics*. Kluwer, Dordrecht.
 - [2] Roche, O., Guba, W. (2005) Computational chemistry as an integral component of lead generation. *Mini Rev. Med. Chem.* **5**:677–683.
 - [3] Renner, S., Schneider, G. (2004) Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening. *J. Med. Chem.* **47**:4653–4664.
 - [4] Wermuth, C.G., Gannelin, C.R., Lindberg, P., Mitschler, L.A. (1998) Mitscher, Glossary of terms used in medicinal Chemistry. *Pure & Appl. Chem.* **70**:1129-1143.
 - [5] Renner, S., Fechner, U., Schneider, G. (2005) Alignment-free pharmacophore patterns – A correlation-vector approach. In: *Pharmacophores and Pharmacophore Searches*. (Langer, T., Hoffmann, E., Eds), pp.49–79. Wiley-VCH, Weinheim.
 - [6] OpenGL Architecture Review Board (2004) *OpenGL Reference Manual: The Official Reference Document to OpenGL, Version 1.4*, Addison-Wesley, Boston.
 - [7] DeLano, W.L. (2002) The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA.
 - [8] Palomer, A., Cabre, F., Pascual, J., Campos, J., Trujillo, M., Entrena, A., Gallo, M., Garcia, L., Mauleon, D., Espinosa, A. (2002) Identification of novel cyclooxygenase-2 selective inhibitors using pharmacophore models. *J. Med. Chem.* **45**:1402-1411.
 - [9] Jackson, E.(1991) *A User's Guide to Principal Components*. Wiley, New York.
 - [10] Wold, S. (1974) A theoretical foundation of extrathermodynamic relationships (linear free energy relationships). *Chemica Scripta* **5**:97–106.
 - [11] 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.
-

- [12] Schneider, P., Schneider, G. (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.* **22**:713 – 718.
- [13] Bemis, G.W., Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**:2887 – 2893.
- [14] Jenkins, J.L., Glick, M., Davies, J.W. (2004) A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **47**:6144 – 6159.
- [15] Renner, S., Schneider, G. (2006) Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **1**:181 – 185.
- [16] Rush, 3rd T.S., Grant, J.A., Mosyak, L., Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **48**:1489 – 1495.
- [17] Cramer, R.D., Poss, M.A., Hermsmeier, M.A., Caulfield, T.J., Kowala, M.C., Valentine, M.T. (1999) Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* **42**:3919 – 3933.
-