Beilstein-Institut The Chemical Theatre of Biological Systems, May 24th - 28th, 2004, Bozen, Italy

NAVIGATION IN CHEMICAL SPACE BASED ON CORRELATION-VECTOR REPRESENTATION OF MOLECULES

GISBERT SCHNEIDER*, STEFFEN RENNER AND ULI FECHNER

Johann Wolfgang Goethe-Universität, Beilstein Endowed Chair for Cheminformatics,
Institute of Organic Chemistry and Chemical Biology,
Marie-Curie-Str. 11, D-60439 Frankfurt am Main, Germany

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

Received: 20th September 2004 / Published: 22nd July 2005

ABSTRACT

Correlation-vector representation (CVR) of molecular structure and properties results in an alignment-free descriptor. This facilitates rapid virtual screening of large virtual compound libraries and chemical databases. The approach has a tradition in chemoinformatics and has already led to the identification of several new lead structures. Its foremost application is ligand-based design of activity-enriched, focused compound libraries. Before applying CVR it is essential to consider appropriate descriptor scaling, select a suitable similarity metric and choose meaningful reference molecules. It was demonstrated that there exists no cure-all recipe for this task. Both three-dimensional and two-dimensional CVR and different similarity metrics complement each other yielding an improved hit rate of the combined approach.

INTRODUCTION

The quest for novel drugs might be considered as a "journey through chemical space", and in order to be successful, we need a navigation system - unless we are satisfied with a random walk. Navigation can be defined as "The process of determining and maintaining a course or trajectory to a goal location" [1].

This means that we need:

- a coordinate system that defines the search space. A convenient way to do this is to employ molecular descriptors, which can be used to generate molecular encoding schemes reaching from general properties (e.g. lipophilicity, molecular weight, total charge, volume in solution, etc.) to very specific structural and pharmacophoric attributes (e.g. multi-point pharmacophores, field-based descriptors).
- a target location. Sets of reference molecules, or "seed structures", exhibiting a desired biological activity (ligand-based approach), or a model of the binding pocket of the macromolecular target receptor (structure-based approach) serve this purpose. The aim is to find "activity islands" in chemical space that are populated by molecules that are similar to the reference compounds (i.e., they are found in a neighbourhood of the reference structures), or can be predicted to bind to the target receptor.
- a method of moving in search space. Typically, sampling methods ("cherry picking") are employed for focused library design, or *de novo* design approaches to identify novel candidate molecules.
- a map. The map separates the search space into regions of high and low quality. This provides the basis for a directed movement toward the target location. Quantitative Structure-Activity (QSAR) models, property prediction methods, and scoring functions - just to name some possibilities - can be used. It should be stressed that each map has a certain resolution and meaning, and depending on the definition of search space and the aim of our journey, we will have to use different maps and navigation systems. Systematic navigation with appropriate maps can not only help us find potential new lead structures but also tell us something about the expected activity profile of a candidate molecule, which is desirable to make an informed selection and prioritization of candidate leads with a reduced attrition liability.

Most critical are the choice of the search space coordinates, and the map resolution. Molecules must be represented in a suitable fashion for reliable prediction of molecular properties [2]. In other words, the appropriate level of abstraction must be defined to perform rational virtual screening.

"Filtering" tools can then be constructed using a simplistic model relating the descriptors to some kind of bioactivity or molecular property. However, the selection of appropriate descriptors for a given task is not trivial and careful statistical analysis is required. Besides an appropriate representation of the molecules under investigation, any useful feature extraction system must be structured in such a way that meaningful analysis and pattern recognition is possible. Technical systems for information processing are intuitively considered as mimicking some aspects of human capabilities in the fields of perception and cognition. Despite great achievements in artificial intelligence research during the past decades and an increasing application of machine learning methods in virtual screening such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [3], we are still far from understanding complex biological information processing systems in detail. This means that a feature extraction or "navigation" task that appears simple to a human expert can be extremely hard or even impossible to solve for a technical system, e.g. particular virtual screening software. As we have learned from many years of "artificial intelligence" research, it is extremely difficult (if not impossible) to develop virtual screening algorithms mimicking the medicinal chemists' intuition. Furthermore, there is no common "gut feeling" because different medicinal chemists have different educational backgrounds, skills and experience. Despite such limitations there is, however, substantial evidence that it is possible to support drug discovery in various ways with the help of computer-assisted library design and selection strategies. There are two specific properties of computer-based approaches which make them very attractive for exploratory navigation in chemical space, namely their speed of execution which can be significantly faster than in vitro experiments, and the ability to access virtual compound libraries. By this means "unexplored" regions of chemical space can be entered and analysed.

Due to its ease of implementation, chemical similarity searching has a long tradition in this area, and many different similarity metrics have been proposed to analyse rapidly very large virtual libraries [4-7]. Again, similarity searching can only be successful if molecules are represented by a suitable description of the chemical space. The definition of "important" attributes heavily depends on the query structure and therefore on its associated binding partner. Descriptors of chemical space can be categorized, e.g., according to their data representation and according to the dimensionality of molecular attributes (1D, 2D or 3D) they describe. Binary fingerprints are a typical data representation for similarity searching [8].

They describe the presence or absence of a feature, e.g. a substructure, or a certain pharmacophore, in a linear bit string format. Fingerprints vary in length from 57 bits for mini-fingerprints (a collection of 1D and 2D molecular descriptions [9]) up to millions of bits of 4D-pharmacophore fingerprints (all combinations of four-point pharmacophores) [10]. For an extensive review of issues related to conformer generation in the process of property calculations, see elsewhere [11].

Pharmacophore models seem to be specifically suited for "scaffold hopping" and library design [12]. If we wish to pick members of a compound library from a very large virtual chemistry space, the calculation of 3D-conformers and subsequent structural or potential pharmacophore point-based alignment of molecules can be a time-limiting factor. Therefore, alignment-free models have a value particularly during the early phases of library design [13-15]. To demonstrate the idea, one representative of these methods shall be discussed in more detail: correlation-vector representations (CVR). The correlation vector approach was introduced to the field of cheminformatics by Broto and Moreau over two decades ago [16], and brought to wider attention through studies by Gasteiger and coworkers [17-19]. The basic idea of CVR is to map molecular features, e.g. pharmacophore points or properties, to a numerical vector of fixed length which is similar to fingerprint generation. As a consequence, such a vector of a given dimension encodes each molecule, and pair-wise comparison of vectors (e.g., by similarity calculation) can be executed very quickly without having to explicitly align the molecular structures. CVR belongs to the class of alignment-free descriptors. Our group and others have reported several CVR applications to similarity searching previously, exploiting the possibility for very fast virtual screening of large compound collections or in *de novo* design [20]. Here we give an overview of our contributions to this field.

SIMILARITY SEARCHING USING CORRELATION-VECTOR REPRESENTATIONS

Chemical similarity searching can serve as a guide through chemical space with the goal of identifying novel molecules that reveal similar biological activity as a query structure. It is often employed in early-phase virtual screening for the selection of activity-enriched subsets [21]. Ligand-based similarity measures facilitate similarity searching in the absence of receptor structure information and are frequently and successfully used for this purpose [22-24].

Basically, these methods rely on:

- one or more representative reference structures ('query structures'),
- molecular descriptors that capture biologically relevant attributes,
- a suitable similarity metric.

The foundation of chemical similarity searching is the pair-wise compound comparison between the query structure and the compounds of a screening library. This reflects the underlying supposition that structurally similar molecules have similar biological activities [25]. The result of a similarity search is a similarity-ranked list. High-ranking compounds in such a list are assumed to be more similar to a query structure than low-ranking compounds. The similarity metric is responsible for considering molecules as different, which do not share important attributes. The definition of 'important' attributes crucially depends on the query structure and thus on the respective biological target.

A similarity search can be performed either retrospective (retrospective screening) or prospective (prospective screening). Retrospective screening is carried out with a set of molecules that are active against a certain biological target (query structures) and a screening library that is compiled of compounds that are inactive against the same target. (Most datasets suffer from *assumptive inactivity*: because of the non-inexistence of measured data it is often unknown whether the inactive compounds are really inactive against the particular biological target). For each of the *n* known actives of a given dataset a pair-wise similarity search is performed. Hence, there are *n* similarity searches where each of the known actives in turn, is the query structure. The pair-wise similarity of the respective query structure is calculated against all the remaining known actives (*n*-1) and all compounds of the screening library (inactive compounds). This procedure yields *n* similarity ranked lists that are ultimately fused into a final similarity ranked list that incorporates the rankings of the n individual lists. There exist different ideas of how to combine the individual lists [26]. Retrospective screening starts with the a priori knowledge of which compounds are active (query structures) and which are not (screening library). This knowledge is then consulted to assess the quality of the final similarity ranked list. Thus, retrospective screening does not aim to come up with novel molecular structures. Its aim is rather to evaluate the quality of the applied parameters, i.e., the molecular descriptor and the similarity metric. It is noteworthy that the dataset (query structures and screening library) also has a prominent influence on the screening results. This limits the explanatory power of comparisons of virtual screenings with different datasets.

The descriptors and similarity metrics that were tested in a retrospective screening can then be applied to prospective screening. Here, one starts with one or more known active molecules (query structures) as well. However, the screening library is distinct from the one employed with retrospective screening. The intended purpose of prospective screening is to identify molecules within the screening library that exhibit activity against the same biological target as the query structures. Therefore, the screening library should contain compounds that are likely to exhibit that similarity.

The quality of a retrospective screening is often quantified by the enrichment factor (*ef*):

$$ef = \left(\frac{S_{act}}{S_{all}} \right) \Big/ \left(\frac{D_{act}}{D_{all}} \right) \tag{1}$$

The enrichment factor is calculated for defined fractions (subsets) of the similarity ranked list, e.g. the first percentage, the first two percentages etc. D_{all} is the total number of compounds in the dataset (query structures and screening library), and S_{all} is the number of molecules in the subset. D_{act} is the number of known active molecules (query structures) in the dataset, and S_{act} is the number of actives found in the subset. The plotting of S_{all}/D_{all} on the abscissa and S_{act}/D_{act} on the ordinate leads to the visualization of the enrichment factor, i.e. the enrichment curve. A method that is superior to a random selection of compounds returns an $ef > 1$ and an enrichment curve above the diagonal line. It should be emphasized that the enrichment factor has an upper limit that is contingent on the fraction of active compounds in the dataset. Again, this stresses the fact that a comparison of several retrospective-screening runs is only meaningful if they are carried out with the same dataset.

In a recent study we investigated the influence of individual parameters on ligand-based virtual screening [27]. This influence was examined on three levels: On the level of the dataset, the descriptor, and the similarity metric. For this purpose, we employed twelve different datasets, three different molecular representations and three different similarity metrics (the Manhattan Distance, the Euclidian Distance and the Tanimoto Coefficient). Special focus was on the evaluation of CVRs of molecular features [16]. The basic idea of CVR is to map molecular features, e.g. pharmacophore points, to a numerical vector of fixed length. Since CVRs have the convenience of being alignment-free descriptors, pair-wise comparisons of vectors can be executed very quickly.

The twelve datasets were compiled from the COBRA database [28] and consisted of a set of active compounds (query structures) and the respective remainder of the COBRA database as 'inactive compounds' (screening library). All molecules of one set of active compounds bind to the same interaction partner, but the definition of these interaction partners differs between the individual sets. These distinct levels of specificity range from receptor classes that comprise a rather diverse set of molecules (e.g. GPCR) to particular receptor subtypes (e.g. COX2). The twelve sets of active compounds contain ligands that bind to angiotensin converting enzyme (ACE, 44 compounds), cyclooxygenase 2 (COX2, 93 compounds), corticotropin releasing factor (CRF antagonists, 63 compounds), dipeptidyl-peptidase IV (DPP, 25 compounds), G-protein coupled receptors (GPCR, 1642 compounds), human immunodeficiency virus protease (HIVP, 58 compounds), nuclear receptors (NUC, 211 compounds), matrix metalloproteinase (MMP, 77 compounds), neurokinin receptors (NK, 188 compounds), peroxisome proliferator-activated receptor (PPAR, 35 compounds), beta-amyloid converting enzyme (BACE, 44 compounds) and thrombin (THR, 188 compounds).

All datasets were encoded by three different descriptors: CATS2D [29], CATS3D and CHARGE3D. The two CATS descriptors belong to the category of atom-pair descriptors. The centres of the atom-pairs are not characterized by their chemical element type but by their membership to a potential pharmacophore point (PPP) group. CATS2D considers five PPP groups: hydrogen-bond donor, hydrogen-bond acceptor, positively charged or ionized, negatively charged or ionized, and lipophilic. The PPPs of CATS3D were associated with the PATTY-Type function of MOE [30] that closely follows the assignment scheme of the PATTY publication by Bush and Sheridan [31]. This assignment scheme comprises seven PPPs: cationic, anionic, polar, acceptor, donor, hydrophobic and other. All possible pairs of PPPs were then interrelated with the distance between the corresponding atoms. Whereas the CATS2D descriptor regards topological distances measured in bond lengths, the CATS3D descriptor incorporates the spatial Euclidian distance of atom pairs. Both CATS descriptors were scaled to diminish the influence of the molecular size. CHARGE3D is based on the correlation vector approach of Gasteiger and co-workers [17]. Calculated partial atom charges are assigned to all atoms. To yield a single charge value for each possible atom pair of a molecule the charge values of the two respective atoms are multiplied.

Finally, the single charge value of each atom pair is associated with the spatial Euclidian distance between the two atoms to obtain a high-dimensional correlation vector. In case of CATS3D and CHARGE3D descriptor a single conformation for each molecule of the dataset was calculated with the program CORINA [32].

Enrichment factors between 2 (GPCR dataset) and 26 (CRF antagonists) were acquired for the first percentage of the datasets (a complete list of the enrichment factors can be found in the original publication). Aside from the GPCR dataset considerable enrichment factors were achieved with all three descriptors. In most cases the influence of the similarity metric was marginal, albeit there were a few combinations of datasets and descriptors that showed a significant discrepancy in terms of the enrichment factor with different similarity metrics. Figure 1 illustrates the enrichment curves for the COX2, the HIV, and the MMP subsets of the COBRA database with the Manhattan Distance as a distance metric. Many active compounds received top ranks with CATS2D and CATS3D for all three datasets depicted in Fig. 1. Whilst COX2 ligands were successfully enriched with the CHARGE3D, after approximately 20% of the screened dataset the enrichment curves for the HIVP and MMP ligand datasets drop below the curve that represents a random distribution of active compounds in the dataset. Nevertheless, HIVP and MMP ligands were enriched by CHARGE3D in the first percentiles of the datasets. Figure 1 clearly demonstrates that none of the descriptors is superior for all three datasets, but there is a preferred one for a given dataset. The suitability of the descriptors depends on the underlying dataset, i.e. the binding patterns of a specific ligand-receptor pair. Distinct performances of the descriptors were expected, as the CATS2D encodes topological information of PPPs, the CHARGE3D three-dimensional information of partial atom charges and the CATS3D spatial information of PPPs.

The separation of active and inactive compounds performed variably contingent on the dataset. Irrespective of the descriptor and similarity metric the approximate classification accuracy seems to be determined by the dataset. Some target classes yielded better enrichment factors than others. We deduced two possible reasons for this behaviour. First, the three descriptors may cover the essential binding pattern of particular datasets to a different extent. Second, the individual datasets are defined at different levels of specificity. The latter hypothesis is substantiated by the fact that GPCR ligands could not be considerably enriched but significant enrichments were achieved for stricter defined subsets of GPCRs. For example, enrichment factors for CRF range from 9 to 26 in the first percentile.

Hence, the separation of actives and inactives may even be possible in difficult cases provided the definition of these two classes is specific enough. Again, this emphasizes that the dataset with its inherent properties has a major influence on the outcome of a virtual screening campaign.

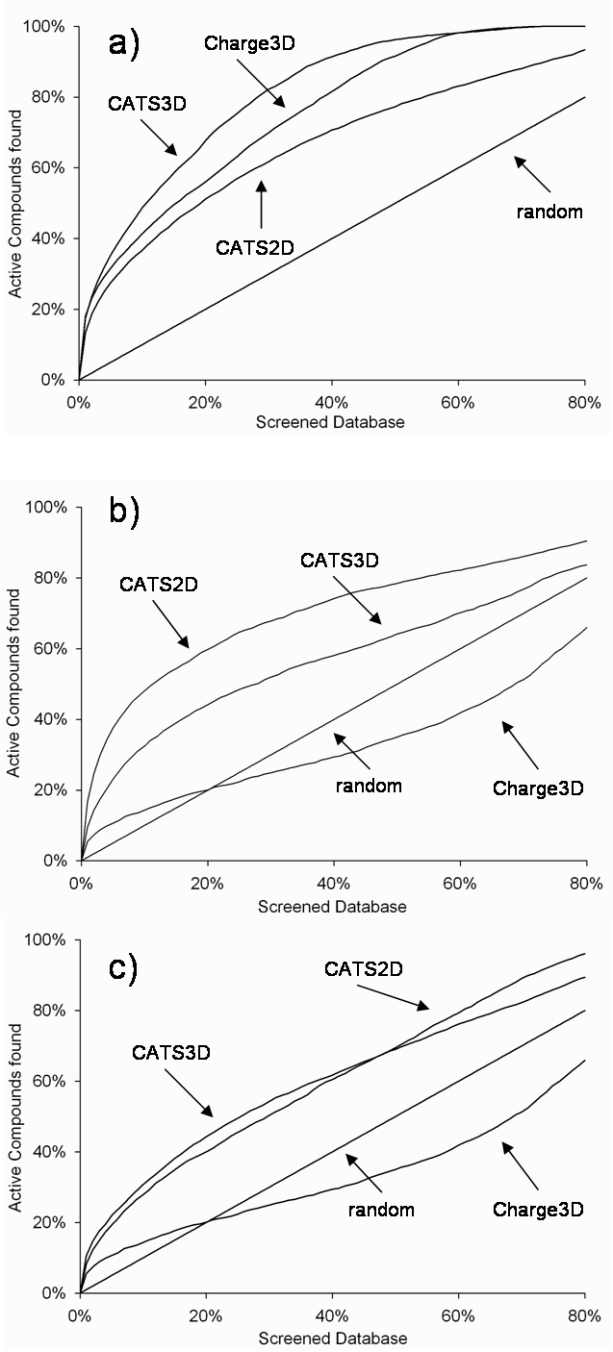

Figure 1. Enrichment curves with the **(a)** COX2, **(b)** HIV protease and **(c)** MMP subset of the COBRA database as active molecules based on the Manhattan distance. The curve indicated with 'random' results from a random distribution of the actives (query structures) among the inactive molecules (screening library).

Whereas the enrichment factor discriminates only between active and inactive compounds, we then investigated which active compounds were retrieved by the three descriptors among the top-ranking ones. Figure 2 depicts this for the first five percentiles of the COX2, MMP and HIVP dataset by means of Euler-Venn diagrams. It is noteworthy that although the enrichment factors with different descriptors were approximately the same, the active compounds among the top-ranking ones varied. Figure 2 shows that a large number of compounds were exclusively retrieved with a single descriptor and that the intersection sizes of all three descriptors were rather small. These two observations sustained the hypothesis that each descriptor covers a certain, and to a varying extent different, aspect of the ligand-receptor binding pattern. Moreover, the information contents of the three descriptors complement each other. The extent of completion can be measured by calculation of the "cumulative percentages" for a given dataset: for a given dataset the active compounds among the first 5% of the similarity-ranked lists of the three descriptors are extracted to obtain three sets of active compounds. The sets are then united and the number of elements of the united set is related to the total number of active compounds of the particular dataset. Cumulative percentages facilitated the retrieval of additional 7% to 52% of active compounds compared to the exclusive employment of the CATS2D. Thus, it may be appropriate to unite the information encoded by different descriptors if a similarity search is performed to cover more facets of the ligand-receptor binding pattern under investigation.

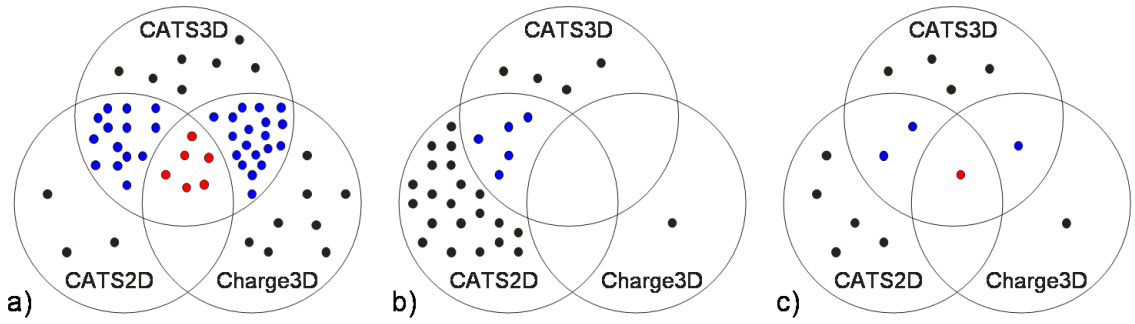

Figure 2. Elements of the Euler-Venn diagrams correspond to compounds that are retrieved among the first 5% of the similarity ranked list that results from retrospective screening with the **(a)** COX2, **(b)** HIV protease and **(c)** MMP subsets of the COBRA database as actives. The Manhattan distance was employed as a distance measure. Membership indicates that the respective compound was retrieved by retrospective screening with the corresponding descriptor.

Another study focused on the comparison of seven similarity metrics for ligand-based similarity searching [33]. The same twelve datasets compiled of the COBRA database as in the aforementioned experiments were employed.

All datasets were encoded with the CATS2D descriptor and retrospective screenings were carried out with seven similarity metrics: Manhattan Distance, Euclidian Distance, Tanimoto Coefficient, Soergel Distance, Dice Coefficient, Cosine Coefficient, and Spherical Distance. Again, apart from the GPCR dataset considerable enrichments were achieved. Enrichment factors for the same datasets but different similarity metrics differed only slightly. For almost all datasets, the Manhattan and the Soergel Distance yielded the overall highest enrichment factors. One might conclude that if only a single distance metric is applied the Manhattan Distance should be preferred due to its computational simplicity and altogether above-average performance.

The study also addressed the question to which extent the active compounds among the top-ranking ones were identical if different similarity metrics were applied. Each of the twelve datasets yielded seven similarity ranked lists obtained with the seven similarity metrics. For each dataset the cumulative percentages were calculated for the first 5% of these lists. This procedure led to the retrieval of significantly more hits than found by any single similarity metric. The increase of the cumulative percentages for all seven metrics compared to the employment of only the Manhattan Distance ranged from additional 5% (COX2) to 28% (NUC and MMP) with an average of 19% over all twelve datasets.

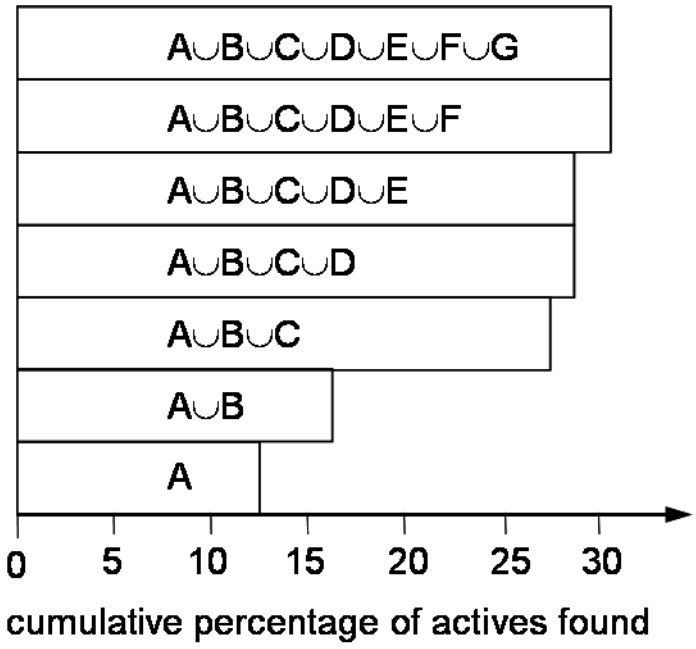

Figure 3. Bars indicate the "cumulative percentage" of active compounds (ligands of nuclear receptors, NUC, from the COBRA data collection) found among the top-ranking 5% of the similarity-ranked list that results from a retrospective screening. **A)** Manhattan distance, **B)** Euclidian distance, **C)** Tanimoto coefficient, **D)** Soergel distance, **E)** Dice coefficient, **F)** Cosine coefficient, **G)** Spherical distance.

Figure 3 illustrates this gradual rise for the NUC dataset. This study suggests that not only different descriptors complement each other but also different similarity metrics. Therefore, it might be advantageous to employ several molecular descriptors and similarity metrics in parallel and benefit from a unification of the various definitions of "chemical similarity".

CODING RECEPTOR INFORMATION INTO CORRELATION-VECTOR REPRESENTATION

Different types of reference points can guide navigation through chemical space. So far, we have pinpointed approaches that employed ligand information to reach our goal location. But if the receptor structure of a biological target is available it can also direct our search for novel ligands. Classical approaches in this field are molecular docking tools such as FlexX or GOLD [34,35]. One drawback with these methods is their prolonged calculation time compared to topological ligand-based methods, due to conformer calculation and spatial feature alignment. Our research group developed a method that combines the speed advantages of ligand-based similarity searching with the ability to exploit binding pocket information (A. P. Schüller, U. Fechner, S. Renner, L. Franke, L. Weber, G. Schneider, unpublished). This is accomplished by introducing the "virtual ligand" (VL). The VL represents a super-structure of potentially all conceivable ligands of a binding pocket. A three-dimensional binding pocket model serves as a starting point to define three types of receptor-based PPPs (hydrogen-bond acceptor, hydrogen-bond donor and lipophilic). These receptor-based PPPs are then used to create potential interaction sites in the cavity of the binding pocket to form a VL. The interaction types of the interaction site are complementary compared to those of the receptor-based PPPs (hydrogen-bond donor is converted to hydrogen-bond acceptor and vice versa, the lipophilic type is unchanged). Encoding the VL as a correlation vector descriptor circumvents computationally expensive 3D alignment of molecular features of the VL and candidate structures. The three interaction types and 20 equidistant bins yielded a 120-dimensional descriptor. For such a calculated VL three parameters were optimized with ten-fold cross-validation prior to virtual screening (scaling of the descriptor, a weighting scheme and a distance metric). As the parameter optimization was performed on a set of reference compounds the method is not solely receptor-based but may be regarded as a hybrid-approach that includes both receptor and ligand information. Up to 50 conformations were calculated for each compound in the screening

library with the Molecular Operating Environment (MOE) conformer generator [30]. The MOE atom typing (PATTY types) was used to assign PPPs to the structures in the screening library [31]. Subsequent application of the same correlation function as used for the calculation of the VL, generated a 120-dimensional descriptor.

Several retrospective and one prospective screening study were performed to provide a proof-of-concept for this approach. All screenings included a randomized VL as a negative control to determine the meaningfulness of the receptor-derived VLs. Co-crystal structures of the PDB served as a starting point for the calculation of the VL. The first series of experiments was considered as retrospective screening because the reference compounds and the screening library were part of the same dataset. Two retrospective screenings with the Factor Xa and COX-2 subset of the COBRA database as actives and the remainder of the database as inactives resulted in significant enrichments. Comparable results were achieved with the Factor Xa inhibitors of a compilation of 15,840 products of the three-component Ugi-reaction, that were tested for inhibition of five serine proteases {courtesy of Dr L. Weber, Morphochem AG, Munich, Germany, unpublished}. Moreover, the receptor-based VL perspicuously outperformed the randomized VL in all three screenings. A final prospective screening was carried out with the Factor Xa dataset of the COBRA database for parameter optimization and the Factor Xa ligands in the Ugi dataset for similarity searching. Here, measured activity values of the Ugi dataset were only employed to assess the quality of the similarity-ranked lists. Among the 50 top-ranking compounds four molecules exhibited a K_i below 10 μM.

The outcome of these experiments demonstrates that the VL approach is suited to retrieve active chemotypes from a library of reference compounds by means of chemical similarity searching. The encompassment of information from binding pockets and their known ligands, bridges the gap between structure-based and ligand-based virtual screening methods.

FUZZY PHARMACOPHORES: THE SQUID APPROACH

Using information from descriptors representing single active molecules is one possible way to define reference points in chemical space. An alternative approach is to use pharmacophore models which comprise information from multiple active molecules within one model. This approach can be characterized as an "ideal ligand" approach [36]. In comparison to many

machine learning-, clustering- or consensus scoring-techniques a pharmacophore model is typically calculated from an explicit alignment of molecules and not from the feature space of multiple single molecule descriptors.

A pharmacophore model represents the spatial configuration of generalized interaction sites, which are essential for biological activity [36,37].

New molecules, which comprise these features, are assumed to be active. Usually these interaction sites represent the most conserved features of a set of known active molecules, which are not present in inactive molecules.

Based on the alignment of active molecules, tolerances for the features are usually estimated to compensate for ligand and receptor flexibility. SQUID fuzzy pharmacophore models extend this approach by approximation of an alignment of known active ligands by a set of spatially distributed Gaussian probability densities for the presence of pharmacophoric features [38]. Features that are present in many of the reference molecules result in a high probability and features which are rarely present in the underlying molecules result in a low probability. Tolerances of the features, which are considered by this approach, might be better represented by Gaussian densities than by rigid spheres. For the resulting fuzzy pharmacophore models different degrees of fuzziness can be defined, e.g. the model can be very generalized or more restricted to the underlying distribution of atoms from the alignment.

For virtual screening the three-dimensional spatial distribution of Gaussian densities is transformed into a two-point correlation vector representation which describes the same probability density for the presence of atom pairs, comprising defined pharmacophoric features. This representation is independent from translation and rotation which makes rapid database screening possible without the necessity to explicitly align the molecules, which can be a limiting step for the screening of large databases. This renders the fuzzy pharmacophore CVR useful for ranking 3D pharmacophore-based CVR representations of molecules, namely CATS3D descriptors of molecules [27]. Consequently SQUID can be characterized as a hybrid approach between conventional pharmacophore searching, similarity searching and fuzzy modelling.

An overview over the calculation of the pharmacophore model and the CVR is given in Fig. 4. The starting point is an alignment of known active ligands. Each atom of the molecules is assigned to a general pharmacophoric atom type like hydrogen-bond donor, hydrogen-bond acceptor or hydrophobic, which results in a field of features.

Maxima in the local feature densities (LFD) are used as cluster seeds to cluster the features into potential pharmacophore points (PPPs) for a more general representation of the underlying alignment. The degree of abstraction, generality or fuzziness of the resulting model can be defined by the cluster radius, a variable which affects the calculation of the LFD and the radius within which maxima are determined.

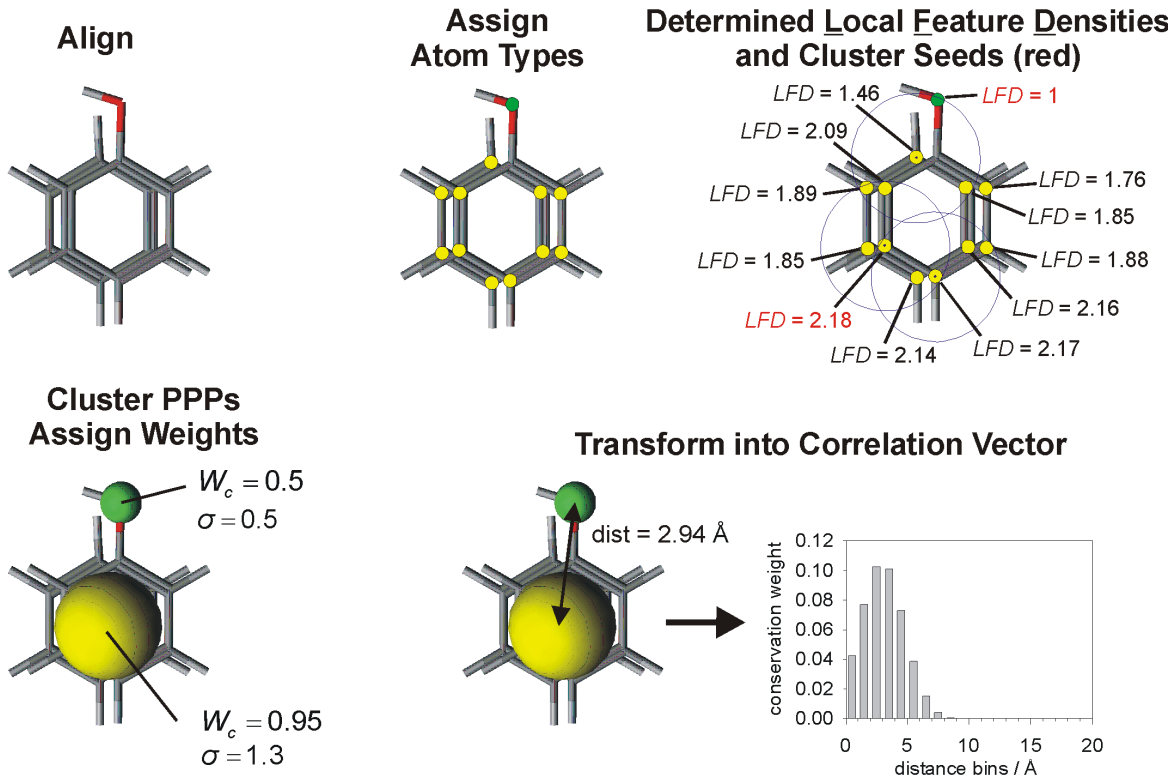

Figure 4. Overview over the calculation of the SQUID fuzzy pharmacophore descriptor.

The radius or standard deviation of the PPPs is dependent on the distribution of the atoms, which are clustered into each particular PPP. Conservation weights (Wc) of the PPPs quantify the conservation of the pharmacophoric features of each of the PPPs in the underlying alignment. As the last step, the three-dimensional representation of the fuzzy pharmacophore model is transformed into a correlation vector for virtual screening. The distance dependent probability for all pairs of PPPs was calculated and subdivided by distance bins and pharmacophoric feature types, resulting in the SQUID CVR.

Figure 5 shows an alignment of the COX-2 inhibitors M5, SC-558 and Rofecoxib, which was adapted from Palomer and coworkers [39]. According to Palomer et al. essential interactions for specific COX-2 inhibition, are mediated by the aromatic rings A and B and the sulfonyl group.

A set of pharmacophore model representations were calculated with different cluster radii resulting in models with different degrees of fuzziness. The model with 1 Å cluster radius resulted in the most detailed representation of the underlying alignment, accordingly with the lowest abstraction from the scaffolds of the molecules in the alignment.

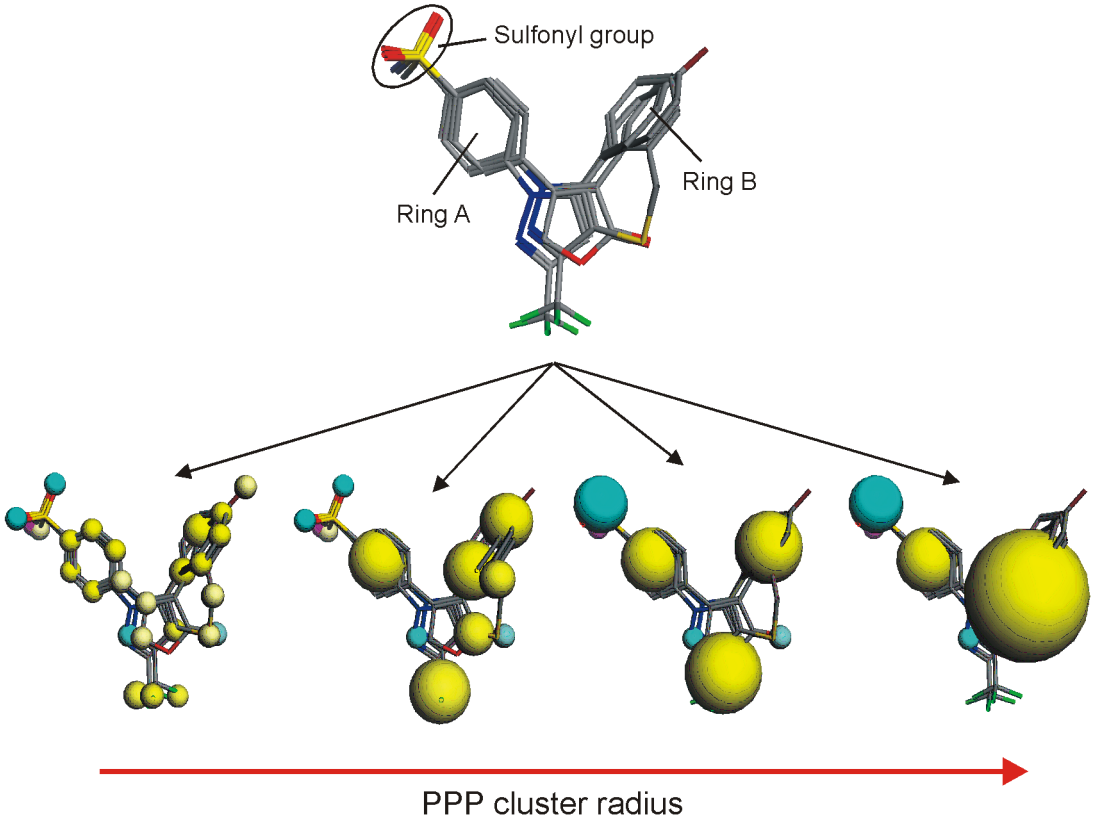

Figure 5. Alignment of three COX-2 inhibitors and SQUID fuzzy pharmacophore models with different resolutions. From left to right, cluster radii of 1.0 Å, 1.5 Å, 2.5 Å and 3.5 Å were used for the model calculation.

Using larger cluster radii results in pharmacophore models with higher degrees of generalization until, like the model resulting from 3.5 Å, the underlying alignment is only marginally visible. For virtual screening it has to be tested for each target and each set of molecules in an alignment from which the degree of fuzziness results in molecules which are most likely to be active, or possess some desired characteristics. Retrospective screening for known active molecules with models with different resolutions provides one possible rationale for that goal.

In Figure 6 the CVR of the best-found COX-2 fuzzy pharmacophore model [38] with a cluster radius of 1.4 Å is shown in comparison to the scaled CATS3D vectors of the underlying molecules from the alignment.

As one can see the fuzzy and thus "generalizing" representation of the underlying molecules from the alignment is retained in the CVR. It becomes clear that the SQUID CVR and the CATS3D CVRs differ significantly in the meaning of their content.

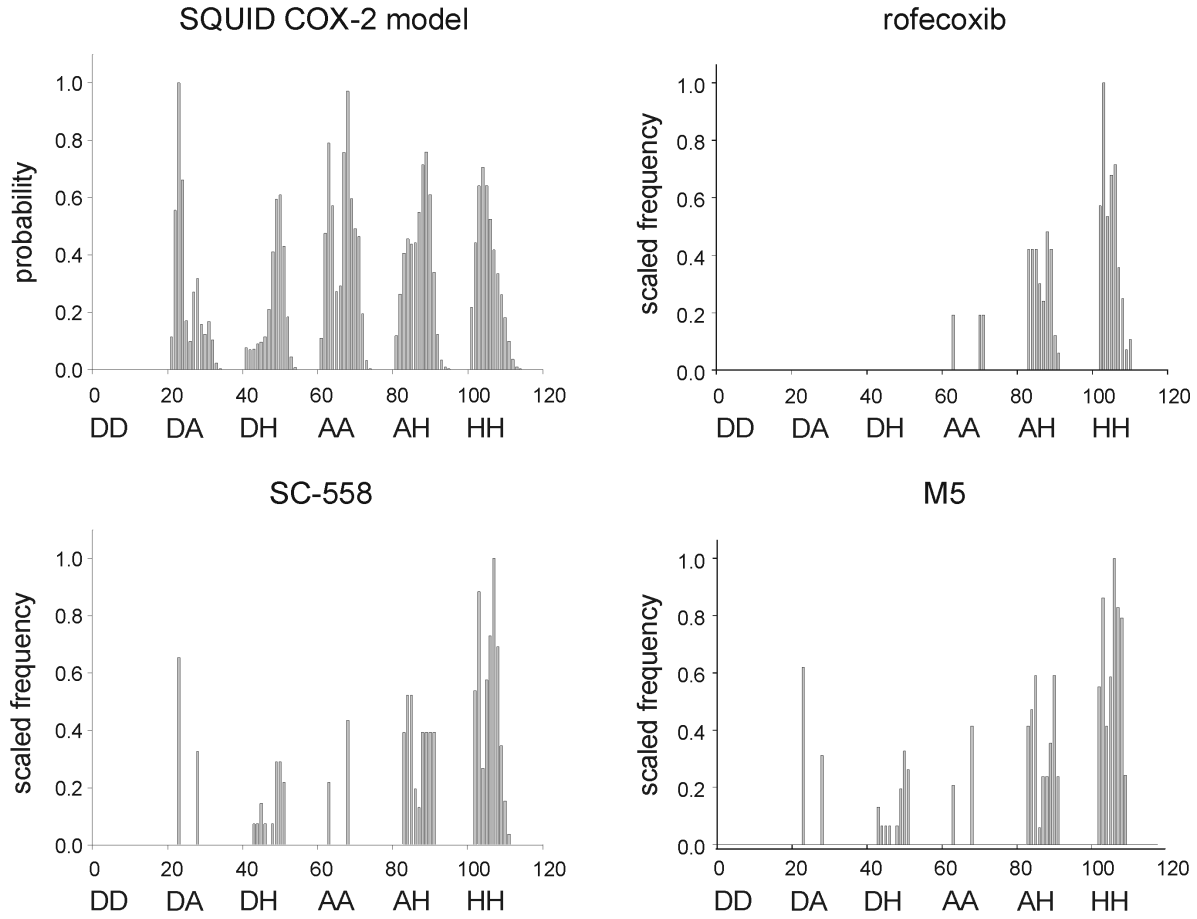

Figure 6. Correlation vector representations of the SQUID COX-2 model and CATS3D vectors of the molecules from the alignment used for the calculation of the pharmacophore model.

The SQUID CVR describes a broad range of descriptor areas which are favourable for the desired biological activity, while the CATS3D descriptor contains only a smaller subset of the actual occurrences of atom-pairs in a specific ligand. Consequently, commonly used similarity metrics like the Euclidean distance or the Tanimoto index, which are based on the assumption that both descriptors, which are to be compared, represent objects in the same way, cannot be used to assess the activity of the molecules under consideration. To overcome this problem a SQUID similarity score was developed (Equation 2):

$$S(a,b) = \frac{\sum_{i=1}^{n}(a_i * b_i)}{1 + \sum_{i=1}^{n}((1-a_i) * b_i)} \qquad (2)$$

where a_i is the value of the i-th element of the SQUID CVR, b_i is the value of the i-th element of a molecule CVR and n is the total number of dimensions. The value a_i may be considered as the idealized probability of the presence of features in b_i.

This results in high scores for molecules with many features in regions of the query descriptor which have a high probability. To penalize the presence of such atom pairs in regions with a low probability, the denominator weights the presence of atom pairs with the inverted probabilities of the descriptor of the pharmacophore model (a value of 1 was added to the denominator to avoid division by zero and high scores resulting from a very low value in the denominator of the term). Accordingly the SQUID scores for the CVRs from the COX-2 inhibitors in Fig. 6 decrease in the order of M5 > SC-558 > Rofecoxib.

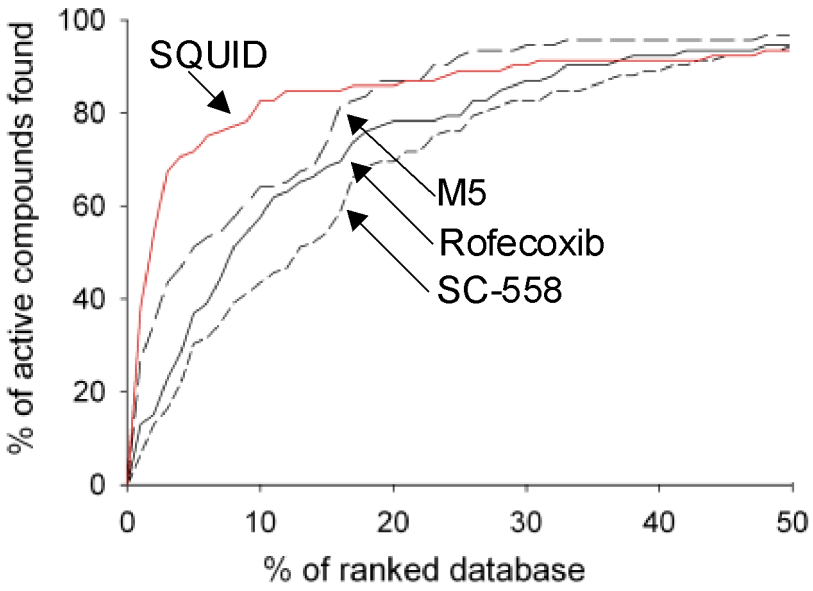

Figure 7. Enrichment curves of the retrospective screening with the SQUID model in comparison to the CATS3D descriptors of the COX-2 inhibitors used for the model calculation.

In Fig. 7 enrichment curves are shown from a comparison of SQUID fuzzy pharmacophores and CATS3D similarity searching for COX-2 ligands in the COBRA dataset of bioactive reference molecules [28]. Similarity searching was performed with the three COX-2 inhibitors from the alignment.

In this example SQUID fuzzy pharmacophore models outperformed similarity searching. While SQUID retrieved 75 % of the active molecules in the first 6 % of the ranked result-database, the best similarity search with rofecoxib resulted in 75 % of the actives ranked into the top 16 % of the database. This result demonstrates that it can be beneficial to integrate information from multiple molecules with the desired activity into a pharmacophore query rather than perform multiple individual database searches.

The combination of fuzziness and conservation of important features among the molecules provides a promising means for hopping from one "activity island" to another in chemical space.

ACKNOWLEDGEMENTS

The authors are most grateful to Dr Petra Schneider and Andreas Schüller for helpful discussions and their research contributions. Dr Lutz Weber is warmly thanked for leaving us the Ugi-data set to work with. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

REFERENCES

[1] Franz, M.O., Mallot, H.A. (2000) Biomimetic robot navigation. *Robot. Autonom. Syst.* **30**:133-153.

[2] Todeschini, R., Consonni, V. (2000) *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim.

[3] Byvatov, E., Fechner, U., Sadowski, J., Schneider, G. (2003) Comparison of support vector machines and artificial neural networks for drug-likeness prediction. *J. Chem. Inf. Comput. Sci.* **43**:1882-1889.

[4] Willett, P. (2000) Chemoinformatics - similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **11**:85-88.

[5] Dean, P.M., Lewis; R.A. Eds. (1999) *Molecular Diversity in Drug Design*. Kluwer Academic, Dordrecht.

[6] Rarey, M., Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* **15**:497-520.

[7] Raymond, J.W., Willett, P. (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput. Aided Mol. Des.* **16**:59-71.

[8] Xue, L., Godden, J.W., Stahura, F.L., Bajorath, J. (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **43**:1151-1157.

[9] Xue, L., Godden, J.W., Bajorath, J. (1999) Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **39**:881-886.

[10] Mason, J.S., Cheney, D.L. (1999) Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores. *Pac. Symp. Biocomput.* pp. 456-467.

[11] Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **40**:195-209.

[12] Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J. (2001) Fundamentals of pharmacophore modeling in combinatorial chemistry. In: *Combinatorial Library Design and Evaluation*. (Ghose, A.K., Viswanadhan, V.N., Eds), pp. 51-71.Marcel Dekker, New York.

[13] Schnitker, J., Gopalaswamy, R., Crippen, G.M. (1997) Objective models for steroid binding sites of human globulins. *J. Comput. Aided Mol. Des.* **11**:93-110.

[14] Cui, S., Wang, X., Liu, S., Wang, L. (2003) Predicting toxicity of benzene derivatives by molecular hologram derived quantitative structure-activity relationships (QSARS). *SAR QSAR Env. Res.* **14**:223-231.

[15] Jewell, N.E., Turner, D.B., Willett, P., Sexton, G.J. (2001) Automatic generation of alignments for 3D QSAR analyses. *J. Mol. Graph. Model.* **20**:111-121.

[16] Broto, P., Moreau, G., Vandyke, C. (1984) Molecular structures: Perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.* **19**:66-70.

[17] Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J., Gasteiger, J. (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **36**:1205-1213.

[18] Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J., Polanski, J. (1996) 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. Des.* **10**:521-534.

[19] Zupan, J., Gasteiger, J. (1999) *Neural Networks in Chemistry and Drug Design*. Wiley-VCH, Weinheim.

[20] Schneider, G., Chomienne-Clement, O., Hilfiger, L., Kirsch, S., Böhm, H.J., Schneider, P., Neidhart, W. (2000) Virtual screening for bioactive molecules by evolutionary *de novo* design. *Angew. Chemie Int. Ed.* **39**:4130-4133.

[21] Barnard, J.M., Downs, G.M., Willett, P. Descriptor-based similarity measures for screening chemical databases. In: *Virtual Screening for Bioactive Molecules.* (Böhm, H.J., Schneider, G., Eds), pp.59-80. Wiley-VCH, Weinheim.

[22] Schneider, G., Nettekoven, M. (2003) Ligand-based combinatorial design of selective purinergic receptor A(2A) antagonists using self-organizing maps. *J. Comb. Chem.* **5**:233-237.

[23] Schuffenhauer, A., Floersheim, P., Acklin, P., Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **43**:391-405.

[24] Stahl, M., Rarey, M., Klebe, G. (2001) Screening of drug databases. In: *Bioinformatics: From Genomes to Drugs*. (Lengauer, T., Ed.), Vol. 2, pp.137-170. Wiley-VCH, Weinheim.

[25] Johnson, M., Maggiora, G.M. (1990) *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, New York.

[26] Ginn, C.M.R., Willett, P., Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Perspec. Drug Discov. Des.* **20**:1-16.

[27] Fechner, U., Franke, L., Renner, S., Schneider, P., Schneider, G. (2003) Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des.* **17**:687-698.

[28] Schneider, P., Schneider, G. (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.* **22**:713-718.

[29] Schneider, G., Neidhart, W., Giller, T., Schmid, G. (1999) "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**:2894-2896.

[30] MOE, Molecular Operating Environment. Distributor: Chemical Computing Group, 1010 Sherbrooke St. West, #910, Montreal, Canada H3A, http://www.chemcomp.com.

[31] Bush, B.L., Sheridan, R.P. (1993) PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **33**:756-762.

[32] Gasteiger, J., Rudolph, C., Sadowski, J. (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **3**:537-547.

[33] Fechner, U., Schneider, G. (2004) Evaluation of distance metrics for ligand based similarity searching. *Chembiochem* **5**:538-540.

[34] Rarey, M., Kramer, B., Lengauer, T., Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**:470-489.

[35] Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**:727-748.

[36] Guner, O. (Ed.) (2000) *Pharmacophore Perception, Development and Use in Drug Design*. International University Line, La Jolla.

[37] Pickett, S. (2003) The biophore concept. In: *Protein-Ligand Interactions*. (Böhm, H.J., Schneider, G., Eds), pp.72-105.Wiley-VHC, Weinheim.

[38] Renner, S., Schneider, G. (2004) Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening. *J. Med. Chem.* **47**:4653-4664.

[39] 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.