Journal of Computer-Aided Molecular Design **17:** 687–698, 2003.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Comparison of correlation vector methods for ligand-based similarity searching

Uli Fechner[a], Lutz Franke[a], Steffen Renner[a], Petra Schneider[b] & Gisbert Schneider[a,*]
[a]*Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie, Marie-Curie-Str. 11, D-60439 Frankfurt, Germany;* [b]*Schneider Consulting GbR, George-C.-Marshall Ring 33, D-61440 Oberursel, Germany*

Received 13 August 2003; accepted in revised form 1 October 2003

Key words: chemical space, drug design, library design, pharmacophore, virtual screening

Summary

Correlation vector methods were tested for their usefulness in ligand-based virtual screening. Three molecular descriptors – two based on potential pharmacophore points and one on partial atom charges – and three similarity measures – the Manhattan distance, the Euclidian distance and the Tanimoto coefficient – were compared. The alignment-free descriptors seem to be particularly applicable when a course-grain filtering of data sets is required in combination with a high execution speed. Significant enrichment of actives was obtained by retrospective analysis. The cumulative percentages for all three descriptors allow for the retrieval of up to 78% of the active molecules in the first five percent of the reference database. Different descriptors retrieved only weakly overlapping sets of active molecules among the top-ranking compounds. If a single similarity index is to be used, the Manhattan distance seems to be particularly applicable. Generally, none of the three different descriptors tested in this study clearly outperformed the others. The suitability of a descriptor critically depends on the ligand-receptor interaction under investigation. For ligand-based similarity searching it is recommended to exploit several descriptors in parallel.

Introduction

Early-phase virtual screening and compound library design often employ similarity searching routines for diversity analysis and the selection of activity-enriched subsets [1]. Ligand-based similarity measures are frequently and successfully used for this purpose [2–4]. Many different approaches have been described, and it is not trivial to select the most appropriate concept for a given task. Basically, these techniques rely on (i) representative reference structures (also termed ‘query’ or ‘seed’ structures), (ii) molecular descriptors that are correlated with biological activity, and (iii) an appropriate similarity measure. Chemical similarity searching is a popular approach to identify novel molecules revealing similar biological activity to a query structure by pair-wise compound comparison. The result of a similarity search in a compound database is a ranked list. High-ranking compounds in such a list are assumed to be more similar to the query structure than low-ranking compounds. The similarity measure should consider molecules as different which do not share important attributes. The definition of ‘important’ attributes heavily depends on the query structure and therefore on its associated binding partner.

The aim of a similarity search can be characterized in one of the following two ways. First, it can be applied with a set of *n* known active molecules. Then, one can evaluate the employed parameters (query structures, descriptor, distance measure) by means of the enrichment factor. This application of a similarity search is called ‘retrospective screening’. In contrast, ‘prospective screening’ can be performed to find molecules that potentially exhibit activity for the same target as the query structure. The decision which specific parameters should be employed for a prospective

[*] To whom correspondence should be addressed. E-mail: gisbert.schneider@modlab.de

screen has to be made on prior gathered experience, and retrospective screening provides a useful means for this purpose.

In this study we compared two different molecular representations – potential pharmacophore points (PPP) and an atom charge model – and three different similarity measures for their suitability for ligand-based design of activity-enriched libraries. Special focus was on the evaluation of correlation vector representations (CVR) of molecular features. The correlation vector approach was introduced to the field of cheminformatics by Broto and Moreau over two decades ago [5], and brought to a wider attention through studies by Gasteiger and co-workers [6–8]. The basic idea of CVR is to map molecular features, e.g. pharmacophore points or properties, to a numerical vector of fixed length. As a consequence, each molecule is encoded by such a vector of a given dimension, and pair-wise comparison of vectors ('similarity' calculation) can be executed very quickly without having to explicitly align the molecular structures. CVR belongs to the class of alignment-free descriptors. Several CVR applications to similarity searching have been reported by our group and others previously, exploiting the possibility for very fast virtual screening of large compound collections or in de novo design [6, 3, 9, 10]. A particularly attractive application of CVR in combination with pharmacophore representations is 'scaffold hopping', i.e. identification of isofunctional molecules with significantly different backbone architecture [9]. Typically, one starts with a single known active, and screens a database for the most similar compounds in terms of CVR similarity. Here we compared three correlation vector methods for two-dimensional (2D) and three-dimensional (3D) pharmacophore similarity searching in a quantitative and qualitative manner. Additionally, three distance measures – the Manhattan distance, the Euclidian distance and the Tanimoto coefficient – were employed to allow for a comparison between them.

Data and methods

Data sets

All molecules for the virtual screening exercise were taken from the COBRA database, a collection of reference molecules for ligand-based library design compiled from recent scientific literature [11]. The database is non-redundant and is annotated by target receptor information and activity data. For this work version 2.1 of the database was used. It consists of 4,705 compounds which were divided into non-overlapping subsets. Each subset contains compounds binding to the same receptor-class. Among these subsets are ligands binding to G-protein coupled receptors (GPCR), proteases, other enzymes, ion channels, hormones, and to molecules not belonging to one of the classes mentioned before. The detailed annotations of the COBRA database allow for the subdivision of these subsets, e.g. according to the receptor type or even the receptor subtype where applicable.

Twelve subsets of the COBRA database were examined in detail in this work. These subsets contained ligands that bind to angiotensin converting enzyme (ACE, 44 compounds), corticotropin releasing factor (CRF antagonists, 63 compounds), cycloxygenase 2 (COX2, 94 compounds), dipeptidyl-peptidase IV (DPP, 25 compounds), G-protein coupled receptors (GPCR, 1642 compounds), human immunodeficiency virus protease (HIVP, 58 compounds), hormone receptors (HOR, 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). The letters in brackets indicate the corresponding abbreviations of the subsets that are used throughout the following text.

Two-dimensional pharmacophore model (CATS2D)

The 'CATS' descriptor was introduced by Schneider and co-workers to provide a concept for 'scaffold hopping' [9]. It is based on the two-dimensional structure of a molecule and therefore avoids the issue of conformational flexibility. It belongs to the category of atom-pair descriptors and encodes topological information of a molecule [12]. The centers of the atom-pairs are not characterized by their chemical element type, but by their membership to a potential pharmacophore point (PPP) group (generalized atom-types). Five PPP groups were considered: hydrogen-bond donor (D), hydrogen-bond acceptor (A), positively charged or ionizable (P), negatively charged or ionizable (N), and lipophilic (L). The upper-case letters in parentheses are the abbreviation of each group. These five groups were assumed to represent potential pharmacophore points of a molecular structure. In the following text these groups will be referred to as 'CATS types'. If an atom does not belong to one of the five CATS types it was not considered. In the present study,

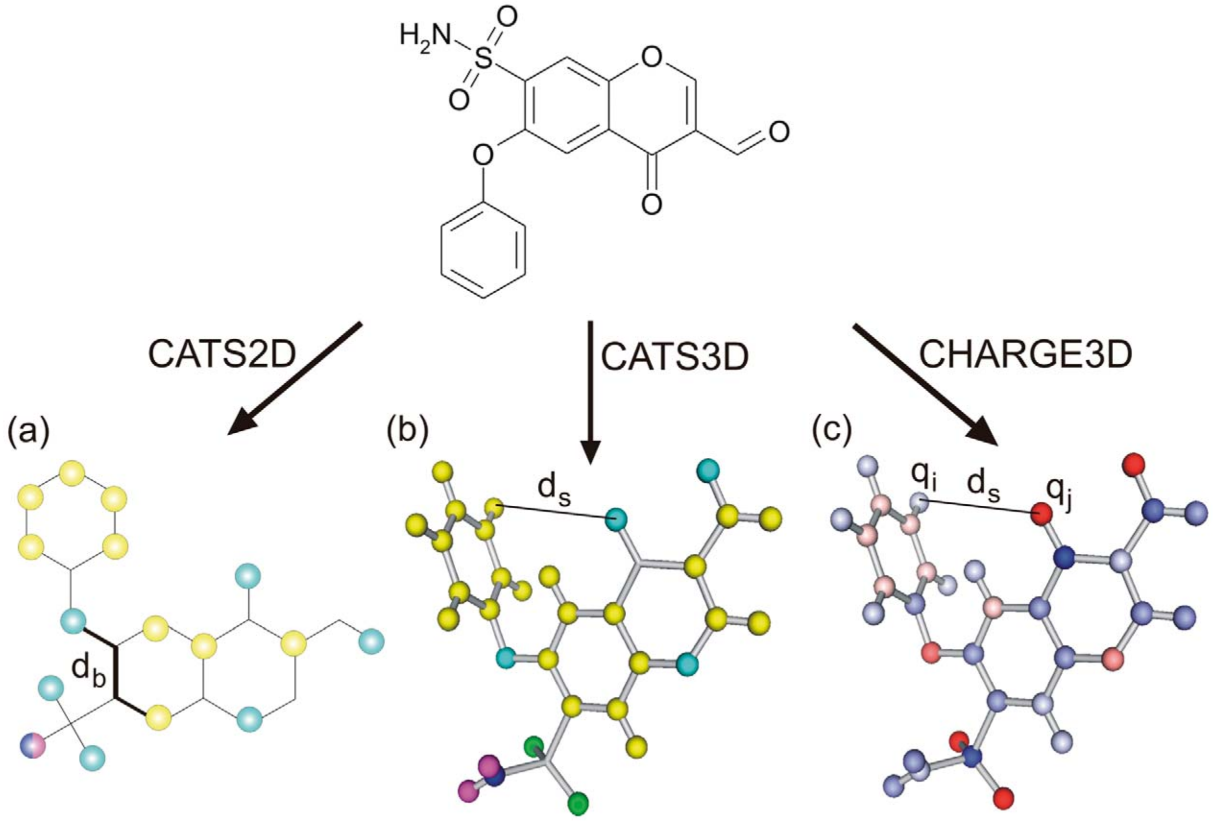

Figure 1. The basic principle of the calculation of the CATS2D, CHARGE3D and CATS3D descriptors by means of a COX2 inhibitor. (a) During calculation of the CATS2D descriptor the two-dimensional structure of the molecule is converted to the molecular graph representation and generic atom types are assigned to the vertices of this graph. Finally, all possible pairs of atom types are counted having regard to the intervening bonds whereby only the shortest path between two vertices is taken into account. Yellow balls represent lipophilic centers, cyan balls hydrogen-acceptors, blue balls positive centers and magenta balls hydrogen-donors. An example of an acceptor-lipophilic pair spaced three bonds apart is depicted by bold edges. d_b is the distance in bonds. (b) For the calculation of the CATS3D descriptor the explicit three-dimensional conformation of a molecule is converted into a three-dimensional distribution of potential pharmacophore points. The descriptor encodes the number of pairs of generalized atom types which fall into predefined distance bins. The color scheme of the atom types is consistent with (a). Additionally, green balls represent polar atoms. An example of an acceptor-lipophilic pair is depicted with a black line. d_s indicates the spatial distance between the two atoms. (c) The CHARGE3D descriptor maps the partial atom charges of a molecule to predefined distance bins. Blue color represents positive charge, red color represents negative charge. Color intensity indicates the charge value. The black line shows the spatial distance d_s between a partially positive atom with charge q_i and a partially negative atom with charge q_j.

atom types were assigned according to the following definition. *Lipophilic*: {C(C)(C)(C)(C), Cl}; *Positive*: {[+], NH_2}; *Negative*: {[−], COOH, SOOH, POOH}; *Hydrogen-bond Donor*: {OH, NH, NH_2}; *Hydrogen-bond Acceptor*: {O, N[!H]}. Thus, every atom of a molecule was assigned to no, one or two CATS types.

The occurrences of all 15 possible pairs of CATS types (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) were counted and the resulting histograms were divided by the number of non-hydrogen atoms in the molecule to obtain a scaled vector. All 15 possible pairs of CATS types were then associated with the number of intervening bonds between the two corresponding atoms, whereby the shortest path length was used. The minimum distance between a pair of CATS types was zero bonds, the maximum distance was nine bonds. Thus, the result of the calculation of the CATS descriptor was a 150-dimensional correlation vector representation (CVR). The general procedure to obtain the CATS two-dimensional pharmacophore model was:

(1) Extract the unweighted, hydrogen-depleted molecular graph,

(2) Assign CATS atom types to the nodes of the molecular graph,

(3) Calculate the distance matrix,

(4) Calculate the correlation vector representation CVR:

$$CVR_d^T = \frac{1}{A}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij,d}^T, \qquad (1)$$

where d is the path length, T is the atom type-pair, A is the number of non-hydrogen atoms and δ^T is the Kronecker delta that evaluates to 1 if a pair T exists and 0 otherwise.

Calculations were performed with the program *speedcats*. It was written in ANSI compatible C. The program takes an MDL SD file with one or more molecules as an input and generates a tab-separated ASCII file containing the high-dimensional CATS descriptor as an output. The program *speedcats* is able to calculate the CATS descriptor of approximately 1,000 compounds per second on a 2.4 Gigahertz Pentium IV machine running Linux. Thereby, the program qualifies for being used in applications that deal with very large numbers of compounds like early-phase virtual screening applications. A light version of the program *speedcats* is accessible over the URL: http://www.modlab.de.

Three-dimensional pharmacophore model (CATS3D)

The 'CATS3D' descriptor is an extension of the CATS2D descriptor to three-dimensional space. For this approach, the spatial pair-wise Euclidian distances between potential pharmacophore points were considered instead of the topological graph-based distances of the CATS2D descriptor. Generalized atom types were assigned with the PATTY_Type function of MOE [13]. Each atom, including hydrogens, was set to one of seven features: cationic, anionic, polar, acceptor, donor, hydrophobic, or other. In contrast to CATS2D, multiple potential pharmacophore point assignments to one atom were not allowed. Detailed information about the assignment scheme of the PATTY function is given by Bush and Sheridan [14].

Calculation of the descriptor resulted in a histogram of the frequencies of all possible pairs of generalized atom types. The frequencies of the 28 possible atom type pairs were partitioned into 20 equal distance bins from 0 to 20 Å. Thus, the resulting correlation vector was 560-dimensional. Finally, the value of each feature-pair defined bin was normalized by the added incidences of the two respective features. The correlation vector was calculated following Equation 1, whereby $i \neq j$, i.e. distances between identical atoms were not considered.

The software for the calculation of the descriptor was implemented in SVL, the programming language of MOE. The COBRA database was preprocessed as follows: Removal of salts and addition of hydrogens was performed with CLIFF [15]. Then, a single 3D conformation per molecule was calculated with CORINA [15, 16].

Atom charge model – CHARGE3D

'CHARGE3D' is based on the correlation vector approach of Gasteiger and co-workers [6]. It is based on the estimated three-dimensional structure of molecules and calculated partial atom charges. The three-dimensional Euclidian distances of all atom pair combinations in one molecule were calculated. Distances within a certain range (0.1 Å) were allocated to the same bin. The charges of the two atoms that form a pair were multiplied to yield a single charge value per pair. Charge values that were assigned to the same bin were added. We used 101 bins in equal steps of 0.1 Å, so that the distance between 0 and 10 Å was covered. All distances above 10 Å were associated with the last bin. The output was a 100-dimensional correlation vector which characterizes the molecules by means of their partial atom charge distribution. The correlation vector representation was calculated using Equation 2:

$$CVR_d = \sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}(q_i q_j)_d, \qquad (2)$$

where d is the atom-atom distance, q_i and q_j are partial atom charges, A is the number of atoms and δ is the Kronecker delta that evaluates to 1 if a pair exists and 0 otherwise.

Removal of salts and addition of hydrogens was performed with CLIFF [15]. The spatial structures for the COBRA database were calculated with the program CORINA [15, 16], and partial atom charges, including hydrogens, were assigned using the PETRA software [15].

Retrospective screening and compound ranking

The Manhattan distance, the Euclidian distance and the Tanimoto coefficient were chosen since they are probably most commonly used in similarity searching and the majority of other distance measures are more or less correlated to them [1]. As the values of the three applied descriptors are non-binary, the formulas for continuous variables were employed for all three distance measures. Mathematical equations are

Table 1. Equations of distance metrics for continuous variables that were used in this work. *A* and *B* are objects (here: molecules), *i* and *j* are attributes of these objects, *n* is the total number of attributes of an object, x_{jA} the value of the *j*th attribute of object A, $S_{A,B}$ denotes the similarity between objects A and B, and $D_{A,B}$ the distance between objects A and B. Note that the range of the Tanimoto coefficient is 0 to 1 if all attributes of A and B are restricted to non-negative values.

Distance metric	Equation	Range
Manhattan distance	$D_{A,B} = \sum_{j=1}^{j=n} \lvert x_{jA} - x_{jB} \rvert$	0 to ∞
Euclidian distance	$D_{A,B} = \sqrt{\sum_{j=1}^{j=n} (x_{jA} - x_{jB})^2}$	0 to ∞
Tanimoto coefficient	$S_{A,B} = \frac{\sum_{j=1}^{j=n} x_{jA}x_{jB}}{\sum_{j=1}^{j=n}(x_{jA})^2 + \sum_{j=1}^{j=n}(x_{jB})^2 - \sum_{j=1}^{j=n} x_{jA}x_{jB}}$	−0.333 to +1

given in Table 1. Detailed information related to these similarity measures can be found in [1, 17].

The enrichment factor *ef* allows for assessing the result of a similarity search [18]: Given a data set containing D_{all} compounds, where a proportion D_{act} of the D_{all} compounds exhibits desired biological activity. Then, the complete data set is ordered according to a similarity measure, and a certain fraction, e.g. the top 10%, is selected from this ranked list. The selection contains F_{all} compounds, among which F_{act} experimentally validated actives are found. The expected number of actives ('background frequency') in the selection F_{all} is

$$F_{act,BG} = F_{all}\frac{D_{act}}{D_{all}}. \tag{3}$$

Thus, a similarity search can be qualified by calculating the enrichment of active molecules within F_{all} over a random distribution of the active molecules:

$$ef = \left(\frac{F_{act}}{F_{all}}\right) \Bigg/ \left(\frac{D_{act}}{D_{all}}\right), \tag{4}$$

where ef is the enrichment factor. An enrichment factor above 1 is returned by a method that is superior to a random selection of compounds within F_{all}. The enrichment factor was visualized by plotting F_{all}/D_{all} on the x-axis and F_{act}/D_{act} on the y-axis ('enrichment curve'). A well-performing similarity search should result in a curve above the diagonal line.

If the query of a similarity search does not consist of a single molecule but of a set of *n* molecules (known actives), *n* similarity searches were performed. Each active molecule was the query structure in turn. Pairwise comparison was performed against the *m* compounds forming the database and the *n*−1 actives. Consequently, one obtains *n* ranked lists. The enrichment factor was calculated for each of the *n* list. The final enrichment factor of such a similarity search was the average of the enrichment factors of the individual lists. To obtain a final ranked list the compounds were sorted according to their average positions within the individual ranked lists that resulted from the *n* similarity searches.

Calculations were performed with the program *rankIt*. It was written in ANSI compatible C. The heapsort algorithm was implemented for the sorting procedure. Thus, sorting could be performed in O(NlogN) [19]. A Linux executable of the program is available for download over the URL http://www.modlab.de.

Results and discussion

The aim of this work was the detailed exploration of the influence of individual parameters on pharmacophore-based similarity searching using CVR. This influence was examined on three different levels. First, on the level of the distance measure; second, on the descriptor level and third, on the data set level. For this purpose, three distance measures (the Manhattan distance, the Euclidian distance and the Tanimoto coefficient), three descriptors represented by correlation vectors (the CATS2D descriptor, the CHARGE3D descriptor and the CATS3D descriptor) and 12 different datasets were employed. The datasets were subsets of the COBRA database, a non-redundant collection of reference molecules for ligand-based library design compiled from recent scientific literature [11]. All compounds of one subset bind to the same interaction partner, but the definition of these interaction partners takes place at different levels of specificity. These characterizations range from receptor classes that comprise a rather diverse set of molecules (e.g., GPCR) to particular receptor subtypes (e.g., COX2). Twelve retrospective screens were performed with one of the subsets considered as 'active' at a time and the respective remainder of the COBRA database considered as 'inactive'.

Enrichment factors obtained from the combination of the three descriptors with the three distance measures for the 12 COBRA subsets are listed in Table 2. We yielded enrichment factors between 2 (GPCR data set) and 26 (CRF antagonists) for the first percentage of the database. Aside from the GPCR data set we were able to considerably enrich the active compounds with our method. Table 2 clearly indicates that in most cases the influence of the distance measure on the enrichment factor was marginal. The differences between the three distance measures are negligible if standard deviations are taken into account. The standard deviation of the *ef*-values was between 3% and 84%, where values of approximately 40% occurred most frequently. Still, there are exceptions. For example, for both the CHARGE3D and the CATS3D descriptor the HIV protease subset showed a significant improvement in terms of the enrichment factor with the Tanimoto coefficient instead of the Manhattan distance. We concluded from these facts that if one distance measure had to be chosen, the Manhattan distance might be preferred because almost always all three measures used in this study led to comparable enrichment factors, and the Manhattan distance is computationally least expensive. Nevertheless, there exist combinations of data sets and descriptors that are not compliant with this rule. Therefore, it might be advisable to calculate enrichment factors with all three distance measures and select the one yielding the best result in a retrospective analysis. In the case of the CATS2D descriptor the Manhattan distance seems to be particularly applicable, as in all 12 subsets the Manhattan distance was as least as good as the other two measures. It is noteworthy that we observed significant differences among the active compounds that were found as the most similar ones in the final ranked list depending on the employed distance measure (not shown here). This emphasizes the counsel to incorporate all three distance measures, as their focus differs with respect to which compounds are considered being similar.

A comparison of the three descriptors reveals that none of the descriptors is generally superior for all data sets, but often a preferred one for a given data set. This reflects the suitability of a specific descriptor for a specific ligand-receptor pair. As the interaction patterns of ligand-receptor pairs differ, distinct performances of the descriptors were expected. The CATS2D encodes topological information of potential pharmacophore-points, the CHARGE3D three-dimensional information of partial atom charges and the CATS3D spatial information of potential pharmacophore-points.

Figure 2 illustrates the enrichment curves for the COX2, HIV protease and MMP subsets of the COBRA database. The Manhattan distance was chosen as distance measure. Curves above the diagonal indicate an enrichment of actives ($ef > 1$). Many active compounds receive top ranks with CATS2D and CATS3D for all three data sets shown in Figure 2. With CHARGE3D we were able to enrich COX2 ligands while the enrichment curves for the HIVP and MMP receptor classes even drop below the 'expected' curve after approximately 20% of the database. Nonetheless, HIVP and MMP ligands were enriched by CHARGE3D in the first percentiles of the database. The three plots clearly confirm that the usefulness of the descriptors depends on the underlying data set. In the case of the COX2 data set, CATS3D seems to be superior to CATS2D, whereas the opposite scenario holds for the HIV protease data set. One possible reason might be that the calculated three-dimensional conformers of the HIV protease ligands are less similar to the receptor-relevant binding conformation than the ones of the COX2 data set (*vide infra*).

The efficacy of the descriptors in terms of the enrichment factor sometimes changes with the applied distance measure. For example, the enrichment factors for the HIVP data set obtained by the CATS2D, CHARGE3D and CATS3D descriptors are 17, 6 and 16 (Manhattan distance), 15, 9 and 20 (Euclidian distance), and 15, 13 and 24 (Tanimoto coefficient). It is evident that the distribution of molecules in a chemical space is defined by the respective descriptor. We assume that the proper choice of a distance measure critically depends on the descriptor that defines

Table 2. Enrichment factors for 12 selected subsets of the COBRA database. The number of compounds in each subset is given in brackets. Enrichment factors were rounded to the nearest integer. Note that particularly at the 50% level of the screened database enrichment factors of one and two are in the range of the expected background frequency of actives.

	Enrichment factor								
	Manhattan distance			Euclidian distance			Tanimoto coefficient		
Screened database /%	CATS2D	CHARGE3D	CATS3D	CATS2D	CHARGE3D	CATS3D	CATS2D	CHARGE3D	CATS3D
ACE (44)									
1	23	16	20	18	15	20	19	17	21
5	7	6	6	7	8	6	7	8	7
10	4	4	4	4	5	4	5	6	4
50	2	1	1	2	1	1	2	2	2
COX2 (94)									
1	13	18	22	12	17	21	12	16	21
5	5	6	8	5	6	8	5	6	8
10	4	4	5	3	4	5	3	4	5
50	2	2	2	1	2	2	2	2	2
CRF (63)									
1	12	12	26	9	15	24	9	15	25
5	6	5	10	4	6	10	5	6	10
10	4	4	6	3	4	6	3	4	6
50	2	2	2	1	2	2	2	2	2
DPP (25)									
1	14	18	21	12	14	21	10	11	22
5	4	5	7	4	5	7	4	4	7
10	3	3	4	3	3	4	2	2	4
50	1	1	2	1	1	1	1	1	1
GPCR (1642)									
1	2	2	2	2	2	2	2	2	2
5	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1
50	1	1	1	1	1	1	1	1	1
HIVP (58)									
1	17	6	16	15	9	20	15	13	24
5	8	2	6	7	4	8	7	6	10
10	5	1	4	5	3	5	5	4	7
50	2	1	1	2	1	2	2	2	2
HOR (211)									
1	9	11	10	8	11	9	8	9	10
5	5	5	4	5	5	4	6	5	4
10	4	3	3	4	3	3	4	3	3
50	1	2	2	1	2	2	1	2	2
MMP (77)									
1	8	7	11	7	9	11	7	11	12
5	4	3	4	3	4	4	4	6	5
10	3	2	3	3	3	3	3	4	4
50	1	1	1	1	1	1	1	2	2
NK (118)									
1	10	6	8	9	7	8	8	9	10
5	5	2	4	5	3	4	4	4	5
10	4	2	3	3	2	3	3	3	4
50	2	1	2	2	1	2	2	1	2

PPAR (35)									
1	11	18	15	10	19	16	9	19	17
5	5	5	6	4	6	6	4	6	6
10	4	3	4	3	4	4	3	4	4
50	2	1	2	1	1	2	1	2	2
BACE (44)									
1	8	14	11	7	15	10	6	12	10
5	3	4	3	2	4	3	2	4	3
10	2	3	2	2	3	2	1	2	2
50	1	1	1	1	1	1	1	1	1
THR (188)									
1	11	6	10	10	8	11	9	10	12
5	5	2	5	5	4	5	5	6	6
10	4	2	3	4	3	4	4	4	5
50	2	1	1	2	1	2	2	2	2

chemical space. Moreover, it appears reasonable to simultaneously employ several distance measures.

The separation of the active and inactive compounds performed variably contingent on the data sets. Irrespective of the influence of the descriptors and the distance measures on the classification results, the approximate classification accuracy seems to be determined by the underlying data set. We observed that some target classes yielded better results than others (Table 2). There are at least two possible reasons for this. First, the descriptors used in the study may cover the essential pharmacophore features of different data sets to a different extent. Second, the individual data sets are defined at different levels of specificity. For example, the GPCR data set is defined on the level of a receptor class whereas the NK data represent a subset of the GPCR set, and are defined on the level of a receptor type. But even within a stricter defined data set heterogeneity is present. The COBRA database provides detailed annotations for each data set. If applicable and available, these annotations include discrimination between agonists and antagonists and between different receptor subtypes. As some data sets comprise a rather small number of compounds, these annotations were not used to subdivide these sets any further. Hence, some data sets were rather loosely defined and diverse, i.e. they likely contained compounds with different binding locations and mechanisms. This might contribute to the varying quality in terms of separating the actives from the inactives as well as to the observed high standard deviations.

Using the three correlation vector approaches we were not able to enrich GPCR ligands, i.e. we got an enrichment factor of 2 for the first percent of the database. The methods perform better for stricter defined subsets of GPCR; for example, enrichment factors for CRF range from 9 to 26 for the first percentile. The same observation holds, to a lesser extent, for the hormone receptor class and PPAR. We conclude that the separation of actives and inactives may even be possible in difficult cases provided the definition of '*active*' is specific enough. Again, this stresses the fact that the quality of the query structures and the heterogeneity of a reference set crucially influence the separation of 'actives' and 'inactives'.

We then examined the active compounds which were retrieved with the three correlation vector descriptors more closely. While the enrichment factor discriminates only between active and inactive, we analyzed which active molecules were retrieved by each descriptor individually. Retrospective screening with the three different correlation vector descriptors and one similarity measure yielded three final ranked lists (see section Data and methods for a definition of the 'final ranked list'). The active compounds that were within the first five percent of the final ranked list of the CATS2D, CHARGE3D and CATS3D descriptors were extracted. These active compounds are referred to as $\text{actives}_{\text{CATS2D}}$, $\text{actives}_{\text{CHARGE3D}}$ and $\text{actives}_{\text{CATS3D}}$ in the following. The percentage of actives of a particular subset of the COBRA database that is in $\text{actives}_{\text{CATS2D}}$ is labeled CATS2D in Table 3. We then established the union of $\text{actives}_{\text{CATS2D}}$ and

Table 3. Cumulative percentages of active molecules in the first five percent of the ranked list that results from a retrospective screen. Calculations were performed with 12 selected subsets of the COBRA database as active molecules. The number of active molecules of each subset is given in brackets.

	Cumulative percentage actives		
	Manhattan	Euclidian	Tanimoto
ACE (44)			
CATS2D	43	41	39
CATS2D∪CHARGE3D	59	64	59
CATS2D∪CHARGE3D∪CATS3D	66	75	73
COX2 (94)			
CATS2D	24	27	26
CATS2D∪CHARGE3D	53	62	57
CATS2D∪CHARGE3D∪CATS3D	61	73	78
CRF (63)			
CATS2D	33	24	27
CATS2D∪CHARGE3D	44	46	41
CATS2D∪CHARGE3D∪CATS3D	68	76	78
DPP (25)			
CATS2D	12	12	4
CATS2D∪CHARGE3D	36	40	16
CATS2D∪CHARGE3D∪CATS3D	52	64	48
GPCR (1642)			
CATS2D	8	7	5
CATS2D∪CHARGE3D	14	13	11
CATS2D∪CHARGE3D∪CATS3D	19	19	18
HIVP (58)			
CATS2D	45	43	40
CATS2D∪CHARGE3D	47	48	53
CATS2D∪CHARGE3D∪CATS3D	53	64	79
HOR (211)			
CATS2D	9	9	11
CATS2D∪CHARGE3D	31	41	30
CATS2D∪CHARGE3D∪CATS3D	34	44	48
MMP (77)			
CATS2D	10	12	16
CATS2D∪CHARGE3D	13	26	43
CATS2D∪CHARGE3D∪CATS3D	18	36	57
NK (118)			
CATS2D	35	31	26
CATS2D∪CHARGE3D	36	39	31
CATS2D∪CHARGE3D∪CATS3D	42	44	49
PPAR (35)			
CATS2D	43	31	23
CATS2D∪CHARGE3D	49	54	46
CATS2D∪CHARGE3D∪CATS3D	57	63	49
BACE (44)			
CATS2D	7	5	9
CATS2D∪CHARGE3D	14	20	14
CATS2D∪CHARGE3D∪CATS3D	25	23	23
THR (188)			
CATS2D	34	36	26
CATS2D∪CHARGE3D	34	46	51
CATS2D∪CHARGE3D∪CATS3D	39	60	69

actives$_{\mathrm{CHARGE3D}}$ (actives$_{\mathrm{CATS2D}}$∪actives$_{\mathrm{CHARGE3D}}$ = actives$_{\mathrm{CATS2D\cup CHARGE3D}}$). The proportion of actives that is in actives$_{\mathrm{CATS2D\cup CHARGE3D}}$ is labeled CATS2D∪CHARGE3D in Table 3. Finally, the union of actives$_{\mathrm{CATS2D\cup CHARGE3D}}$ with actives$_{\mathrm{CATS3D}}$ was composed (actives$_{\mathrm{CATS2D\cup CHARGE3D\cup CATS3D}}$). The percentage of actives that is in actives$_{\mathrm{CATS2D\cup CHARGE3D\cup CATS3D}}$ is labeled CATS2D∪CHARGE3D∪CATS3 in Table 3. These cumulative percentages for all three descriptors allow for the retrieval of up to 78% of the actives in the first five percent of the database (COX2 with Tanimoto coefficient). The increase of the cumulative percentages for all descriptors used in this study compared to the employment of only CATS2D ranges from additional 7% to 52%. It is noteworthy that the three descriptors retrieved different active compounds even though the enrichment factors are approximately the same (e.g., thrombin set with Euclidian distance or Tanimoto coefficient, Table 3). As each descriptor itself is able to enrich active compounds, all three descriptors cover a certain aspect of information about the interaction of a ligand with its binding partner. However, the comprised information of each descriptor seems to differ. This is illustrated in Figure 3 by the number of compounds that were exclusively identified by a single descriptor. The size of the intersection is six, zero and one for the COX2, HIV protease and MMP subset, respectively. These rather small intersection sizes suggest that the information contents of the individual descriptors complement each other. The exclusive selection of active compounds that are recognized by all three descriptors results in a loss of information. Such a selection scheme implicates the assumption that all aspects of molecular features covered by the three descriptors are necessary for a ligand to bind successfully. Thus, it may be appropriate to unite the information encoded by different descriptors if a similarity search is performed [20, 21]. This can be done by merging the sets of the highest-ranking compounds each descriptor returns.

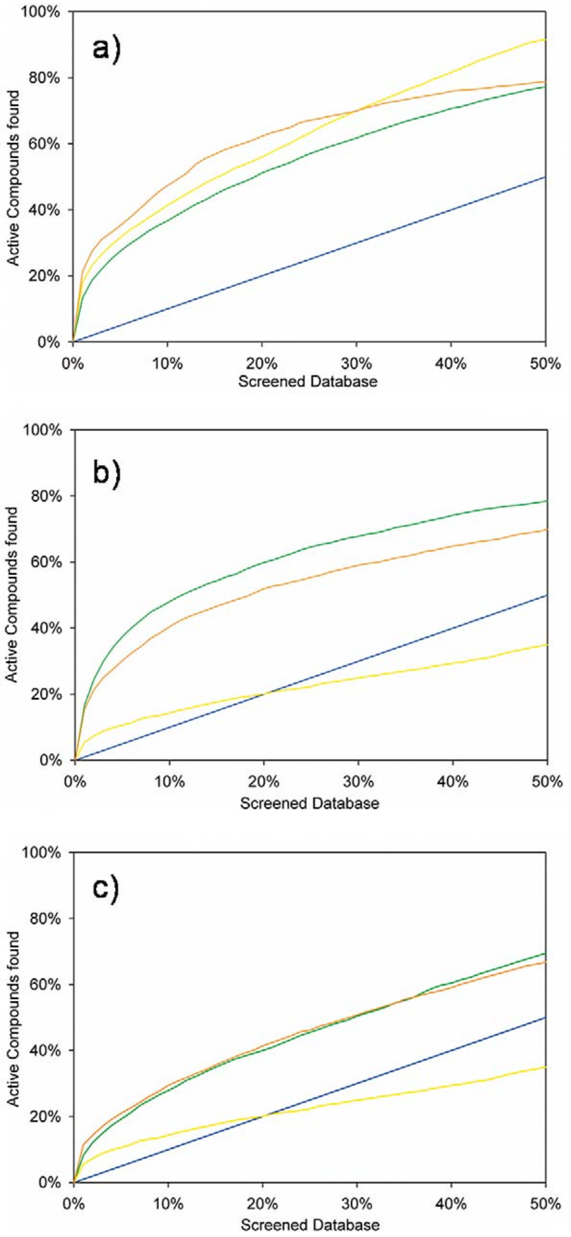

Figure 2. Enrichment curves with the (a) COX2, (b) HIV protease and (c) MMP subset of the COBRA database as active molecules based on the Manhattan distance. The green curve corresponds to CATS2D, the orange curve to CATS3D and the yellow curve to CHARGE3D. The blue curve results from a random distribution of the actives among the inactive molecules.

Figure 3 also shows how differently the descriptors perform according to the data set. The binding pocket of COX2 is buried and narrow, and large parts of the ligands participate in binding. This is reflected by considerable performance of all three correlation vector descriptors for the COX2 data set. The HIVP binding pocket is deep, long and tunnel-like. Here, the topological correlation vector descriptor achieved better results than the two descriptors that are based on three-dimensional distances. A closer look at the

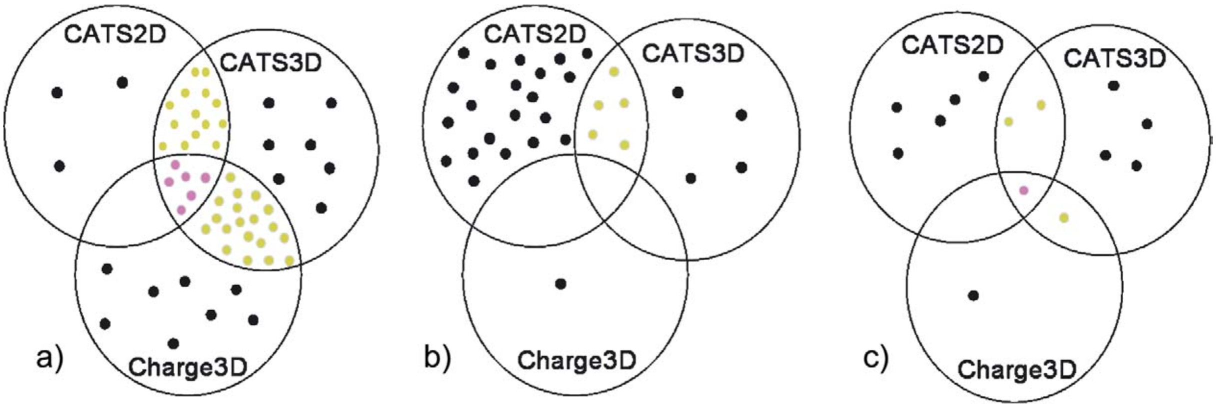

Figure 3. Elements of the Euler-Venn diagrams represent compounds that can be found among the first five percent of the ranked list that results from retrospective screening with the (a) COX2, (b) HIV protease and (c) MMP subsets of the COBRA database as actives. The Manhattan distance was employed as a distance measure. Membership indicates that the respective compound was retrieved by retrospective screening with the corresponding descriptor. The diagrams reveal that the three descriptors complement one another to a different extent depending on the underlying dataset.

HIVP ligands revealed that they contain more rotatable bonds (on average 19 compared to 6 and 15 for COX2 and MMP ligands, respectively) and a higher molecular weight (on average 607 compared to 360 and 439 for the COX2 and MMP ligands, respectively) than the COX2 and MMP ligands. Thus, it is more difficult to generate conformations resembling the receptor-bound ligand conformation. Computationally more expensive multiple conformation searching might be a possibility to overcome this limitation [22, 23]. We fixedly aligned a CORINA-generated conformation of the HIVP ligand L-735,524 with a conformation that was taken from an X-ray structure (PDB file: 1HSG). The significant difference of these two conformations (rigid fit RMSD = 5 Å) explains the low performance of the two three-dimensional correlation vector descriptors. In contrast to COX2 and HIVP, the binding pocket of MMP is rather shallow. Hence, a great portion of the ligand surface is accessible to the solvent. As the descriptors take into account PPP and partial charge information of the entire molecule, much of the encoded information is needless for the binding pattern in case of the MMP ligands. This may contribute to noise in the descriptor data. Generally, the Euclidian distance and the Tanimoto coefficient yielded the highest cumulative percentages of active molecules (Table 3). Different definitions of distance in the chemical space lead to a varying degree of diversity among the highest-ranking active compounds. These differences are not observable in the value of the enrichment factor. Equation 4 considers all active compounds as equivalent, i.e. it only differentiates between active and inactive compounds. Therefore, it is not advisable to compare the performance of different descriptors or similarity searching methods exclusively by means of the enrichment factor. We deduce that the enrichment factor is a measure to quantify a similarity search, but it does not comprehend qualitative aspects. Typically, one is interested in those molecules which are ranked in between known actives, and these have to be tested for activity. We wish to stress that the Manhattan and Euclidean are in the same family of measures (Minkowski metric), and the comparison might be consolidated by adding other coefficients, e.g., the Cosine coefficient, which are in a different family (in preparation). Also, in future studies the effect of similarity measures on a potential bias regarding general compound properties might be investigated in more detail [24].

Conclusions

Correlation vector methods were shown to be suited for ligand-based similarity searching, i.e. considerable enrichment of actives was obtained by retrospective analysis. These alignment-free descriptors seem to be applicable to early-phase virtual screening campaigns, where a course-grain filtering of data sets is required in combination with a high execution speed. If a single similarity measure is to be used, the Manhattan distance seems to be particularly applicable because of its computational simplicity and still produces significant enrichment of actives. This might be specific to the particular descriptors and CVR studied here [25]. Generally, none of the three different descriptors tested in this study clearly outperformed the others. The usefulness of a descriptor clearly depends on the ligand-receptor interaction under investigation and the quality of the query structures. Moreover, appropriate tuning of additional method parameters is essential, e.g. binning options or the definition of pharmacophore types, as well as meaningful data pre-processing, e.g., descriptor scaling. This aspect is of utmost importance for applications of similarity concepts in chemogenomics and large-scale HTS data analysis. For straightforward ligand-based similarity searching we recommend to exploit several descriptors in parallel based on the observation that different descriptors retrieved only weakly overlapping sets of active molecules among the top-ranking compounds. This strategy might be particularly useful in combinatorial library design and building-block selection aiming at activity-enriched subsets, complementing computationally more demanding techniques [26, 27].

Acknowledgements

The authors are grateful to Norbert Dichter for setting up the LSF Linux cluster. Alireza Givehchi and Evgeny Byvatov are thanked for many stimulating discussions. This work was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt, Germany.

Note added in proof

The authors wish to stress that the 1% enrichment factors might be limited in their statistical robustness due to small compound numbers. Also, the term ‘similarity measure’ or ‘distance measure’ is not used in a strict mathematical sense in the present study.

References

1. Barnard, J.M., Downs, G.M. and Willett, P. Descriptor-Based Similarity Measures for Screening Chemical Databases, In: Böhm, H.J. and Schneider, G. (Eds.) Virtual Screening of Bioactive Molecules, Wiley-VCH 2000, Weinheim, New York, pp. 59–80.
2. Schneider, G. and Nettekoven, M., J. Comb. Chem., 5 (2003) 233.
3. Schuffenhauer A., Floersheim P., Acklin P. and Jacoby E., J. Chem. Inf. Comput. Sci., 43 (2003) 391.
4. Stahl, M., Rarey, M. and Klebe, G., Screening of drug databases, In: Lengauer, T. (Ed.) Bioinformatics: From Genomes to Drugs Vol. 2, Wiley-VCH 2001, Weinheim, New York, pp. 137–170.
5. Broto, P., Moreau, G. and Vandyke, C., Eur. J. Med. Chem., 19 (1984) 66.
6. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 36 (1996) 1205.
7. Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J. and Polanski, J., J. Comput.-Aided Mol. Des., 10 (1996) 521.
8. Zupan, J. and Gasteiger, J., Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, New York, 1999.
9. Schneider, G., Neidhart, W., Giller, T. and Schmid, G., Angew. Chem. Int. Ed. Engl., 38 (1999) 2894.
10. Schneider, G., Chomienne-Clement, O., Hilfiger, L., Kirsch, S., Böhm, H.-J., Schneider, P. and Neidhart, W., Angew. Chem. Int. Ed. Engl. 39 (2000) 4130.
11. Schneider, P. and Schneider, G., QSAR Comb. Sci., 22 (2004) in press.
12. Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
13. MOE, Molecular Operating Environment. Distributor: Chemical Computing Group, Montreal, Canada.
14. Bush, B.L. and Sheridan, R.P., J. Chem. Inf. Comput. Sci., 33 (1993) 756.
15. CLIFF, CORINA, PETRA. Distributor: Molecular Networks GmbH, Computerchemie, Erlangen, Germany, http://www.mol-net.de/.
16. Gasteiger, J., Rudolph, C. and Sadowski, J., Tetrahedron Comp. Method., 3 (1990) 537.
17. Willett, P, Barnard, J.M. and Downs, G., J. Chem. Inf. Comput. Sci., 38 (1998) 983.
18. Xu, H. and Agrafiotis, D.K., Curr. Top. Med. Chem., 2 (2002) 1305.
19. Corman, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C., Introduction to Algorithms, 2nd Edition, MIT Press, Cambridge, MA, 2001.
20. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 37 (1997) 1.
21. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
22. Sadowski, J., J. Comput.-Aided Mol. Des., 11 (1997) 53.
23. Xu, H., Izrailev, S. and Agrafiotis, D., J. Chem. Inf. Comput. Sci., 43 (2003) 1186.
24. Holliday, J.D., Salim, N., Whittle, M. and Willett, P., J. Chem. Inf. Comput. Sci., 43 (2003) 819.
25. Chen, X. and Reynolds, C.H., J. Chem. Inf. Comput. Sci., 42 (2002) 1407.
26. Schneider, G., Curr. Med. Chem., 9 (2002) 2095.
27. Tropsha, A. and Zheng, W., Comb. Chem. High Throughput Screen., 5 (2002) 111.