Journal of Molecular Graphics and Modelling 27 (2008) 161–169

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Pharao: Pharmacophore alignment and optimization

Jonatan Taminau, Gert Thijs, Hans De Winter *

Silicos NV, Wetenschapspark 7, B-3590 Diepenbeek, Belgium

ARTICLE INFO

Article history:
Received 4 December 2007
Received in revised form 1 April 2008
Accepted 3 April 2008
Available online 11 April 2008

Keywords:
Drug discovery
Pharmacophore
Virtual screening
Clustering

ABSTRACT

Within the context of early drug discovery, a new pharmacophore-based tool to score and align small molecules (Pharao) is described. The tool is built on the idea to model pharmacophoric features by Gaussian 3D volumes instead of the more common point or sphere representations. The smooth nature of these continuous functions has a beneficent effect on the optimization problem introduced during alignment. The usefulness of Pharao is illustrated by means of three examples: a virtual screening of trypsin-binding ligands, a virtual screening of phosphodiesterase 5-binding ligands, and an investigation of the biological relevance of an unsupervised clustering of small ligands based on Pharao.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

The aim of this work is to describe *Pharao*, a pharmacophore-based scoring method that was developed to act as a virtual screening tool to retrieve molecules from different compound libraries and to perform pharmacophore-based clustering. The method is flexible so that it can also be used to guide the search for new molecules in *de novo* design algorithms or in other computer-assisted applications within the field of drug discovery. *Pharao* is an acronym for *Phar*macophore *A*lignment and *O*ptimization.

In general, evaluation or scoring functions can be ligand- and/or target-based, depending on the information available (e.g. crystal structures available, known actives, etc.). Docking methods are categorized as target-based methods, sharing a wide variety of different approaches. However, these methods tend to be time consuming, which can be a problem when evaluating large sets of molecules, typically the case in both virtual screening and *de novo* design. Ligand-based methods, on the other hand, are mostly trying to score small molecules based on their similarity to one or more reference structures. To define 'similarity', a number of concepts have been described. These can be based on molecular topology (fingerprints), molecular shape, molecular field descriptors, pharmacophores, and many others.

In this paper we focus on the concept of 3D pharmacophores in the context of similarity assessments. A pharmacophore is based on the concept that specific interactions are observed in drug–receptor interactions. A pharmacophore is defined as an ensemble of these interactions, or more specific the corresponding chemical features and their relative positions and orientations. It can be seen as a powerful abstraction or representation of small molecule binding to proteins. Essential interactions, corresponding to chemical features, are hydrogen bonding, charge transfer, steric and electrostatic characteristics, and lipophilic interactions. The strength of feature-based pharmacophore models lies in the adequate definition of the pharmacophore points [1].

In literature, a number of algorithms for pharmacophore modeling have been reported, resulting in a variety of computer programs that can be used in drug discovery [2,3]. Amongst the most widely known are Chem-X [4], Catalyst/HipHop [5], GASP [6] and the more recently described LigandScout [7] and PharmID [8]. A general and flexible pharmacophore-based approach should be able to (1) generate an independent pharmacophore model (from a ligand, set of ligands, or target), (2) align two pharmacophore models against each other and (3) score this alignment with a quantitative measure. Not all existing programs can offer this set of functionalities. For example, Wolber and Langer pointed out the lack of flexibility of the pharmacophore model in Catalyst as a main reason to implement their own 'pharmacophore generating' algorithm [7].

Another difference between most programs lies in the way in which conformational flexibility is incorporated in the pharmacophore alignment. Introducing conformational flexibility into the alignment can only be significantly beneficial when it is incorporated 'on-the-fly' during optimization, like it is for example the case in GASP [6]. However, the combinatorial explosion also increases this way, making the actual optimization more difficult and thus less robust. In *Pharao*, a 'rigid alignment' was chosen,

* Corresponding author.
E-mail address: hans.dewinter@silicos.com (H. De Winter).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.04.003

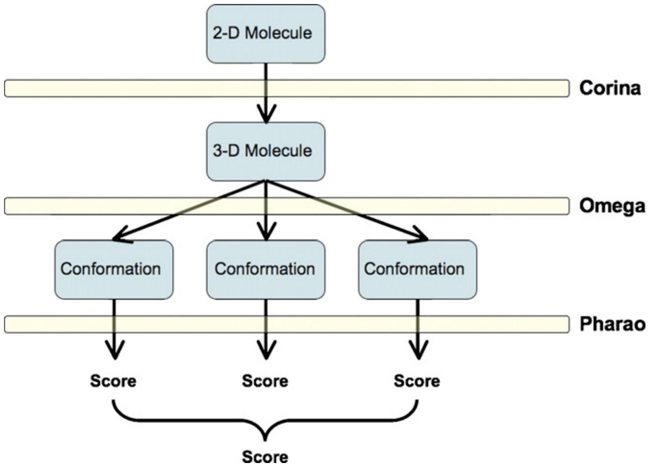


Fig. 1. Possible workflow to combine multiple conformations handling with Pharao. In this example CORINA (Molecular Networks Gmbh, Germany) was used to generate coordinates and OMEGA (OpenEye Scientific Software, USA) to generate a number of conformers. For each generated conformer a score could be obtained using the *Pharao* tool and finally all scores need to be combined, resulting in one score for the initial compound.

meaning that no conformational flexibility of the input structures is assumed during alignment. Conformational flexibility in *Pharao* is handled by introducing a precomputing step, i.e. the generation of a set of conformers for each molecule. A number of different tools capable of generating multiple conformers already exist [9,10]. A possible workflow using the *OMEGA* (OpenEye Scientific Software, USA) and *CORINA* (Molecular Networks GmbH, Germany) tools as precomputing step is illustrated in Fig. 1.

Most programs also differ in the way of similarity between structures or pharmacophores is calculated. GASP, for example, takes the steric overlap of the two structures into account next to the overlap of matching pharmacophore features that all methods use. The GASP-approach might also be mimicked by combining a shape comparison with a pharmacophore feature alignment procedure, the latter being the focus of *Pharao*.

Within the approach taken by *Pharao*, pharmacophore features are modeled by Gaussian 3D volumes instead of the more common point or sphere representations. This way, continuous functions can be introduced into the problem of optimizing the volume overlap and the smooth nature of those functions facilitates the computation of optimal alignments. Grant et al. first pointed out this promising technique [11] and later on it was successfully used in shape-based comparison programs [12], with a Gaussian volume attributed to every atom instead of to every pharmacophore feature.

2. Methodology

The following two sections describe the mechanisms behind *Pharao*. First the perception or detection of pharmacophore points is explained. Next the alignment and scoring of pharmacophore pairs are elucidated.

2.1. Pharmacophore points perception

A pharmacophore model is represented as a collection of different pharmacophore points, each with a corresponding Gaussian 3D volume (m, σ) instead of the more common point or sphere representation. The volume of a pharmacophore point is computed as

$$V = \int p \exp\left(-\frac{|m-r|^2}{\sigma}\right) \mathrm{d}r,$$

with p a scaling constant. Each pharmacophore point is uniquely defined by its position in space m, its spread or sigma σ, and the functional group or chemical feature it is characterizing. In Table 1 all relevant functional groups and chemical features that are implemented in *Pharao* are listed. Some functional groups also need some notion of ‘direction’ in addition to ‘position’. This directionality is included by means of a normal vector, which originates from the center of the pharmacophore point and is given a fixed length of 1 Å. Given this clear and straightforward representation, pharmacophores can be written and stored in a both human- and machine-readable format, resulting in a general and exchangeable abstraction.

Within *Pharao*, the perception of pharmacophore points is a fast and simple process based on the rules as described by Greene et al. [13]. As listed in Table 1, *Pharao* has been implemented to automatically detect and assign the following pharmacophore points:

Aromatic rings: The generation of aromatic ring pharmacophore points includes ring detection and aromaticity detection. Ring detection is performed by calculating the ‘smallest set of smallest rings’ (SSSR) [14]. A ring is labeled as aromatic if it is planar, has no exocyclic double bonds and satisfies Huckel's $4n + 2$ rule [15].

Ring systems consisting of conjugated aromatic rings, for example naphthalene, will count for multiple aromatic ring pharmacophore points.

The normal indicating the orientation of the ring is located perpendicular to the plane of each ring; the angle between the normal vectors of two aromatic ring pharmacophore points will influence the score of the mapping during alignment (vide infra).

Hydrogen bond donors: Hydrogen bond donor pharmacophore points correspond to atoms fulfilling the following conditions:

- atom is nitrogen or oxygen;
- formal charge of atom is not negative;
- atom has at least one covalently attached hydrogen atom.

Hydrogen bond acceptors: The generation of hydrogen bond acceptor points is not as straightforward as the generation of hydrogen bond donor points. The following criteria need to be met:

- atom is nitrogen or oxygen;
- formal charge of atom is not positive;
- atom has at least one available ‘*lone pair*’;
- atom is ‘*accessible*’.

In order to verify that nitrogen atoms have localized lone pair electrons, a number of simple heuristic rules have been implemented: (1) when the nitrogen is part of an aromatic ring, it should have less than three connected bonds, (2) nitrogen should not be a sulfonamide or amide, and (3) when nitrogen has three connections, it should not be adjacent to an aromatic ring.

Condition four, the accessibility of a hydrogen bond acceptor, is slightly more difficult to assess. Accessible means that there is enough space for a hydrogen atom to form a hydrogen bond without any steric hindrance from the rest of the molecule. This is calculated by positioning a sphere around the putative hydrogen bond acceptor with a radius of 1.8 Å, thereby mimicking the possible locations where a hydrogen atom hypothetically can occur [13]. Subsequently a number of dots are uniformly distributed on this sphere and for every dot it is checked if it does collide with any neighboring atoms. In the current implementation approximately 200 dots were sampled on each atom. If at least 2% of the dots are labeled as non-colliding, the hydrogen bond acceptor is labeled as ‘accessible’. The discrete approach of this fourth condition is sufficiently fast and labels

Table 1
List of all possible pharmacophore functional groups or features that can occur in a pharmacophore model generated or processed by Pharao

Code	Description	Normal
AROM	Aromatic ring	Yes
HDON	Hydrogen bond donor	Yes
HACC	Hydrogen bond acceptor	Yes
LIPO	Lipophilic (hydrophobic) region	No
POSC	Positive charge center	No
NEGC	Negative charge center	No
HYBH	Hydrogen bond donor and hydrogen bond acceptor	Yes
HYBL	Aromatic and lipophilic ring	Yes
EXCL	Exclusion sphere	No

approximately 20% of the hydrogen bond acceptors as inaccessible, thereby simplifying the pharmacophore model without losing essential information. Because of condition four, detection of hydrogen bond donors is dependent on the 3D conformation of the structure.

Charge centers: For the generation of charge center pharmacophore points only formal charges are taken into consideration. Atoms with a positive charge will correspond to a positive charge pharmacophore point, while atoms with a negative charge will correspond to a negative charge pharmacophore point. The position of the charge pharmacophore point coincides with the position of the atom with the formal charge.

Lipophilic spots: To generate lipophilic pharmacophore points, a three-step procedure based on the method of Greene et al. [13] is used. First, every atom is assigned a 'lipophilic contribution'. This value is the product of a topology-dependent term t and the accessible surface fraction s. Term t is obtained using some simple heuristic rules that are listed in Table 2, and fraction s is calculated with a similar method as described for the hydrogen bond acceptor pharmacophore points.

Second, when a lipophilic contribution has been assigned to every atom, the next step is to group atoms into lipophilic regions or spots. Grouping atoms into spots is a simple procedure: (1) atoms in a ring of size 7 or less form a group; (2) atoms with three or more bonds, together with their neighbors and not bonded to any other non-hydrogen atom, form a group; (3) the remaining atoms are divided in chains on the basis of their connectivity, and each chain is defined as another group. Rings larger than seven atoms also count as chains.

In the third step the lipophilic contribution for every spot is calculated as the summation of the contributions of the individual atoms belonging to that group or spot. If this value exceeds a predefined threshold, the spot corresponds to a lipophilic pharmacophore point for which the center coincides with the geometric center of this spot. The threshold value is set to 9.87, which is half of the lipophilic contribution of an exposed methyl carbon terminating a carbon chain [13].

Table 2
Topology-dependent lipophilicity factor t

Category	Factor	Description
1	0	N, O or H
2	0	S in SH
3	0	≤2 bonds away from charged atom
4	0	≤2 bonds away from OH or NH with no delocalized electrons
5	0	≤1 bond away from SH with no delocalized electrons
6	0	≤2 bonds away from O with double bond
7	0	≤1 bond away from S with valence >2
8	0	S with double bond
9	0.6	Three bonds away from O with double bond
10	0.6	Two bonds away from S with valence >2
11	0.6	One bond away from S with double bond
12	0	Two or more instances of any of the previous three conditions (cat 9–11)
13	0.25	One neighboring O or N with no delocalized electrons
14	0	>1 neighboring O or N with no delocalized electrons
15	1	Not belonging to any of the previous conditions (cat 1–14)

Table taken from Ref. [13].

Exclusion spheres: Exclusion spheres are pharmacophore points that are completely different from other pharmacophore points because they have a different roles during alignment and they are not extracted from the ligand but from the target to which the ligand binds. However, by taking the same representation, exclusion spheres fit easily into the computational framework of *Pharao*. If no target information is available, exclusion spheres can be placed manually to indicate regions in the pharmacophore model where no pharmacophore points are allowed during alignment.

2.2. Alignment

2.2.1. Problem situation

The quantification of the similarity between two pharmacophores can be computed from the overlap volume of the Gaussian volumes of the respective pharmacophores. The goal is to find the subset of matching functional groups in each pharmacophore that gives the largest overlap. The procedure finds its roots in the work of Grant and Pickup [16], where the volume overlap between two molecules is computed from a Gaussian description of the atom volumes. In *Pharao* this approach is translated to pharmacophore points.

The procedure to compute the volume overlap between two pharmacophores is done in two steps. In the first step, a list of all feasible combinations of overlapping pharmacophore points is generated. In the second step, the corresponding features are aligned with each other using an optimization algorithm. The combination of features that gives the maximal volume overlap is retained to give the resulting score.

2.2.2. Feature mapping

To compute the overlap between two pharmacophores, the first step is to define which points from the first pharmacophore can be mapped onto points from the second pharmacophore. A mapping of two pharmacophores consists of a list of points both pharmacophores where corresponding points have a compatible functional group and the internal distance between points is within a given range requirement. This range, as defined by the parameter ε, controls the feasibility of a combination of pharmacophore points.

The procedure starts by generating a list of all feasible pairs of features. First, two points from the first pharmacophore are selected (a and b) and the distance between them is calculated (d_{ab}). Next, two points with matching features are selected from

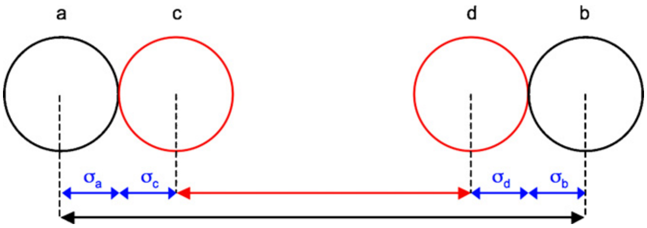


Fig. 2. Illustration of the ε parameter. In this example, the difference between d_{ab} (black line) and d_{cd} (red line) is equal to the sum of the four sigmas (blue lines) of the pharmacophore points. When ε is smaller than 1.0, this implies that $|d_{ab} - d_{cd}|$ should be smaller than the sum of the four sigmas and thus the pharmacophore points should overlap.

the second pharmacophore (c and d) and the distance between these two points is also calculated (d_{cd}). The difference between the two distances is then compared to the sum of the sigma's of the four pharmacophore points. If

$$\varepsilon < \frac{|d_{ab} - d_{cd}|}{\sigma_a + \sigma_b + \sigma_c + \sigma_d}$$

then the combination of the two pairs is said to be feasible. This is also illustrated in Fig. 2. When ε is set to 1.0, this relates to the hard-sphere atom model where the spheres are only touching each other and do not overlap. Smaller values of ε indicate that more overlap is required and becomes as such a more stringent selection criterion. In all the cases demonstrated below, ε is set to 0.5.

Once the list of feasible pairs is constructed, they can be combined into larger feasible combinations. A combination of n pairs can be extended with any other pair if that pair is feasible and compatible with all the n pairs of the combination. This process is combinatorial in nature and the number of possible combinations grows very fast with the number of pharmacophore points. The ε parameter leads to a reduction of the number of feasible combinations.

2.2.3. Alignment phase

Starting from the set of feasible combinations, the combination that gives the largest volume overlap is searched for. For every combination, the procedure starts by orienting the first and second pharmacophore subsets such that their geometric center and their principal axes of inertia coincide. Next, by applying a constrained gradient-ascent to the rigid-body rotation of the second pharmacophore, the maximal volume overlap is determined between the two subsets. The rotational part is implemented using quaternion algebra [17]. The total volume overlap of N matching points is computed as

$$V_{\text{overlap}} = \sum_{i=1}^{N} C_i\, f(\theta(q)) \exp\left(\frac{-1}{\sigma_{i,A} + \sigma_{i,B}} q^{\mathrm{T}} A q\right),$$

where q is a unit quaternion describing the rotation, A is a matrix that only depends on the initial coordinates of the respective pharmacophore points, C_i is a scaling factor, and $\theta(q)$ describes the angle between the pharmacophore normal vectors. $f(\theta(q))$ is one if there is no normal present in two points. $f(\theta(q)) = |\cos(\theta(q))|$ for AROM and HYBL points and $f(\theta(q)) = \cos(\theta(q))$ for HACC, HDON and HYBH points. The use of the Gaussian representation of pharmacophore points offers an elegant way to compute the gradient and Hessian of the volume overlap with respect to q. The gradient-ascent procedure is started from four different initial orientations that correspond to the possible mapping of the principal axes of inertia. The result of the optimization procedure is the rotational angle and axis that gives the optimal overlap, and an alignment score which quantifies this overlap.

The complete alignment procedure starts from the subset with the largest number of matching points and computes the optimal volume overlap of this combination. Next, the smaller combinations are processed until the highest volume overlap so far is higher than the maximum volume overlap any smaller combination hypothetically can achieve. The rationale is that the volume overlap has an upper boundary that depends on the number of features to align. If the current best overlap is larger than this upper bound then there is no need to compute the alignment of smaller subsets since the score will never be larger than the current best.

2.2.4. Alignment score

In the current implementation of *Pharao*, similarity between a pair of pharmacophores is calculated using three different measures:

$$\text{TANIMOTO} = \frac{V_{\text{overlap}}}{V_{\text{ref}} + V_{\text{db}} - V_{\text{overlap}}}$$

$$\text{TVERSKY\_REF} = \frac{V_{\text{overlap}}}{V_{\text{ref}}}$$

$$\text{TVERSKY\_DB} = \frac{V_{\text{overlap}}}{V_{\text{db}}}$$

with V_{overlap} being the volume overlap of the matching subset of pharmacophores points, V_{ref} the volume of the first pharmacophore (reference), and V_{db} the volume of the second pharmacophore (database). The TANIMOTO measure is well known from bit vector comparison and is the default measure to score similarity between pharmacophores. The TVERSKY_REF measure is primarily intended for database searches to identify database compounds having a pharmacophore that is a superset of the reference pharmacophore, while the TVERSKY_DB measure has its use in identifying database compounds having a pharmacophore that is subset of the reference pharmacophore. All three metrics are returning a score between 0 and 1.

3. Results

3.1. Case 1

The first example describes the use of *Pharao* in a virtual screening experiment. A number of active compounds were 'hidden' in a screening database and based on the pharmacophore of one of the actives it was investigated how easily the other actives could be retrieved. This was done by ranking the screened database based on the similarity scores, and looking at the enrichments.

From the PDB database [18] a trypsin-binding ligand 1MTV was arbitrarily selected as the reference structure, i.e. the query of our

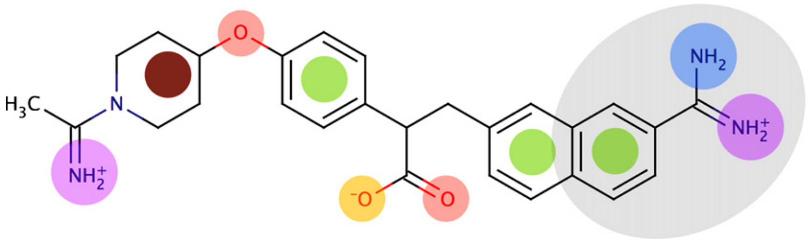


Fig. 3. Schematic representation of the perceived pharmacophore model of compound 1MTV as generated by Pharao. The different pharmacophore types are indicated in different colors (purple, POSC and HDON; blue, HDON; brown, LIPO; yellow, NEGC and HACC; red, HACC; green, HYBL; see Table 1 for an explanation of the pharmacophore codes). The area shaded in gray represents the adapted pharmacophore model encapsulating the critical binding interaction of compound 1MTV.

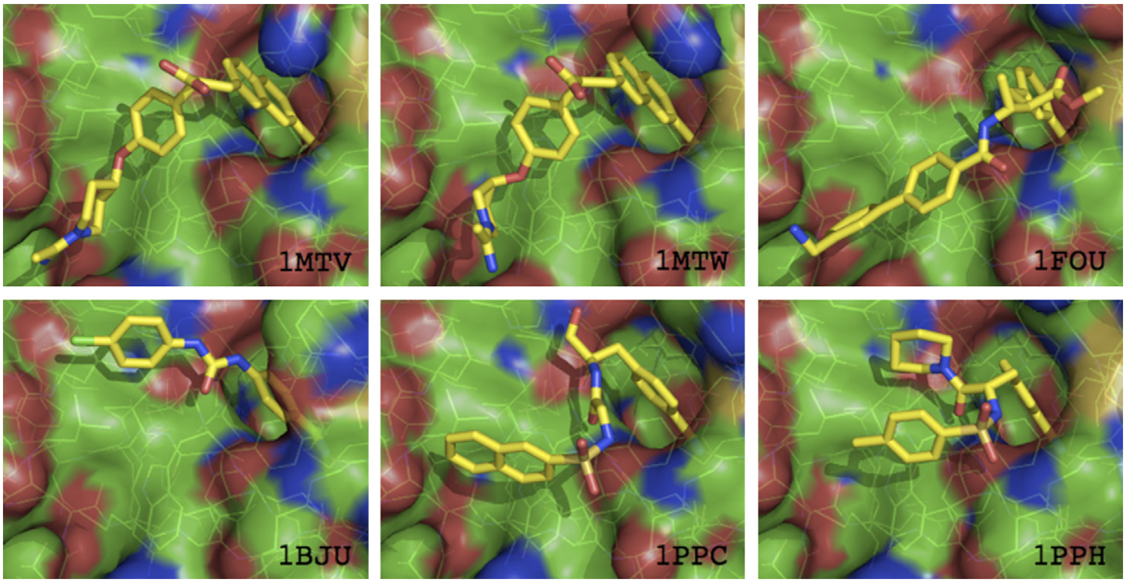


Fig. 4. Representation of the six ligand–target complexes as found in the, respectively PDB entry.

Table 3
The six trypsin-binding references

PDB	Name of ligand
1MTV	(+)-2-[4-[(-1-Acetimidoyl-4-piperidinyl)oxy]-3-(7-amidino-2-naphthyl)]-propionic acid
1MTW	(+)-2-[4-[((*S*)-1-Acetimidoyl-(3*S*)-pyrrolidinyl)oxy]-3-(7-amidino-2-naphthyl)]-propionic acid
1FOU	RPR128515
1BJU	1-(4-Amidinophenyl)-3-(4-chlorophenyl)urea
1PPC	NAPAP
1PPH	3-TAPAP

virtual screening experiment. After automatic pharmacophore perception, which is the first step in the *Pharao* process, a pharmacophore model consisting of 13 pharmacophore points was generated from it (Fig. 3). Looking at the ligand–target complex, the 2-naphtamidine moiety that is located in the cavity as shown in Fig. 4, was identified as critical for binding. Benzamidine is also a known competitive inhibitor for trypsin [25]. Therefore a second and smaller pharmacophore model of four points and based on the 2-naphtamidine moiety was also extracted from the 13-point pharmacophore model (indicated as the gray area in Fig. 3).

Table 4
Case 1: results of the virtual screen with no exclusion spheres

TANIMOTO			TVERSKY_REF		
Rank	Ref.	%	Rank	Ref.	%
Complete pharmacophore model					
1	1MTV	0.05	1	1MTV	0.05
2	1MTW	0.10	2	1MTW	0.10
32	1PPH	1.60	6	1PPC	0.30
50	1PPC	2.49	14	1PPH	0.70
63	1BJU	3.14	186	1BJU	9.27
479	1FOU	23.88	394	1FOU	19.64
Small pharmacophore model					
1	1MTV	0.05	1	1MTV	0.05
18	1FOU	0.90	2	1FOU	0.10
32	1MTW	1.60	3	1MTW	0.15
65	1PPH	3.24	4	1PPH	0.20
67	1PPC	3.34	5	1PPC	0.25
70	1BJU	3.49	7	1BJU	0.35

A screening database was created consisting of 2000 compounds; two sets of 1000 compounds each with an average molecular weight of 360 and 400 Da (the "dl-360" and "dl-400" sets, respectively). These sets were composed by Schrödinger with the aim to be representative of the chemical sample collections of pharmaceutical and biotechnology companies [19,20]. This set was spiked with six active compounds (Table 3): the reference compound and five other trypsin-binding ligands found in the PDB database. This resulted in a database of 2006 compounds.

The first observation that could be made from the obtained rankings in Table 4 is that the *TVERSKY_REF* score performs better than *TANIMOTO* in this particular case for both reference pharmacophore models. With *TVERSKY_REF* no penalty is given to unmatched pharmacophore points and therefore it is not surprising that especially the small pharmacophore model scores so well with this mechanism. Using the complete pharmacophore model this difference is less significant but nevertheless present. The *TANIMOTO* similarity score introduces a bias towards small compounds that map only a local part of the reference pharmacophore and are therefore not penalized for unmatched parts. Larger molecules need a very close map to suppress this indirect advantage and due to the different binding modes there is not always such close map possible between all actives.

In a second part of the experiment, information of the active site of the target protein was incorporated into the reference pharmacophore in the form of exclusion spheres, thereby mimicking the spatial constraints of the active site. The procedure to generate exclusion spheres was straightforward: all protein atoms within a distance of 4.5 Å from the 1MTV ligand were

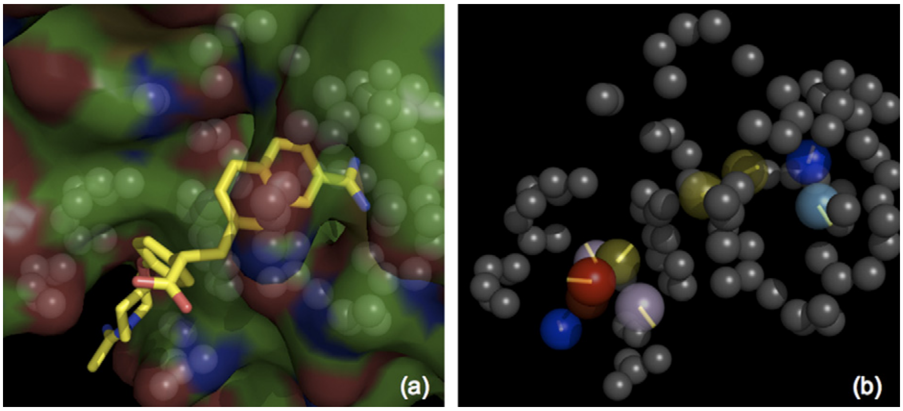


Fig. 5. 3D representation of compound 1MTV. (a) Complexed with its trypsin target and the calculated exclusion spheres. (b) Generated pharmacophore model, extended with EXCL points. The color codes are the same as in the legend of Fig. 3.

Table 5
Case 1: results of the virtual screen with exclusion spheres

TANIMOTO			TVERSKY_REF		
Rank	Ref.	%	Rank	Ref.	%
Complete pharmacophore model					
1	1MTV	0.05	1	1MTV	0.05
2	1MTW	0.10	2	1MTW	0.10
44	1BJU	2.19	12	1PPC	0.60
62	1PPH	3.09	21	1PPH	1.05
85	1PPC	4.24	110	1BJU	5.48
508	1FOU	25.32	366	1FOU	18.25
Small pharmacophore model					
1	1BJU	0.05	1	1MTV	0.05
8	1MTV	0.40	2	1BJU	0.10
9	1PPH	0.45	3	1PPC	0.15
17	1PPC	0.85	4	1MTW	0.20
19	1MTW	0.95	5	1PPH	0.25
27	1FOU	1.35	7	1FOU	0.35

selected and included as exclusion spheres with a sigma value of 0.7. This way, the collection of all exclusion spheres nicely mimics the spatial constraints of the active site of the trypsin (Fig. 5).

Table 5 summarizes the results from the screening with exclusion spheres.

Adding exclusion spheres to the complete pharmacophore model does not improve the ranking, even a greater bias towards small compounds can be observed for both scores (small compounds have again a smaller likelihood of collapsing with exclusion spheres). The higher ranking of compound 1BJU, the smallest active, illustrates this behavior. However, for the small pharmacophore model the addition of exclusion spheres results in a slightly better ranking.

From the examples in Case 1 it can be concluded that *Pharao* is a useful tool for virtual database screening. For all eight settings in Tables 4 and 5, screening of less than 25% of the database compounds retrieves all actives. Using the appropriate settings of model and scoring function results in a need to screen only 1% of the database.

3.2. Case 2

Another virtual screening experiment was performed with phosphodiesterase 5 as reference target. Four PDB entries were selected with phosphodiesterase 5 in complex with the inhibitors sildenafil (1TBF and 1UDT) and vardenafil (1UHO and 1XP0). Both ligands adopt two different positions of their piperazine fragment, indicating two possible different binding modes. All four conformations are illustrated in Fig. 6.

The set of 2000 compounds from Case 1 was again augmented with the four new 'actives'. Sildenafil from PDB entry 1TBF was arbitrary selected as the reference and a ranking of the screening database was made based on the calculated and aligned pharmacophores. Compound 1TBF contains 14 pharmacophore points: two AROM points, three LIPO points, two HDON points and six HACC points (Fig. 7).

Because of the high similarity of the four compounds and the high density of pharmacophore points, it was no problem at all to successfully retrieve them all from the database with 1TBF as query. Both *TANIMOTO* and *TVERSKY_REF* similarity scores lead to a ranking with all four actives at the top as is shown in Table 6.

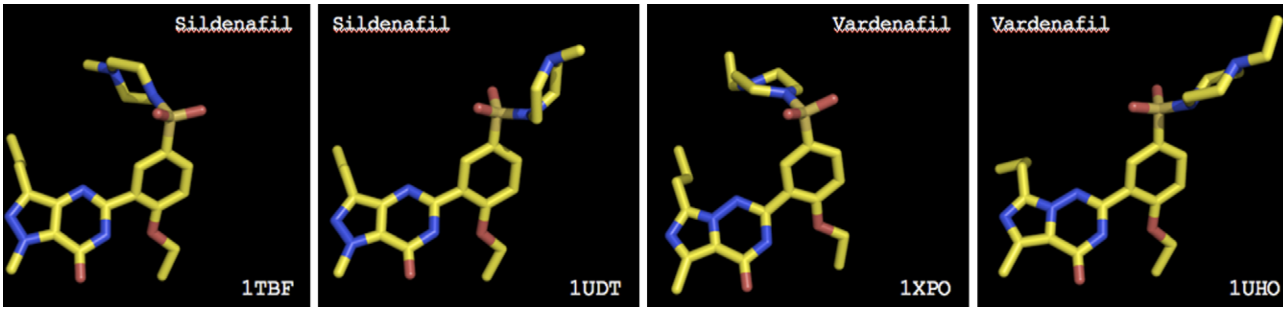


Fig. 6. Different orientations of the piperazine fragment for both sildenafil and vardenafil. All four conformations are obtained from the PDB files and are shown in the alignment obtained by Pharao.

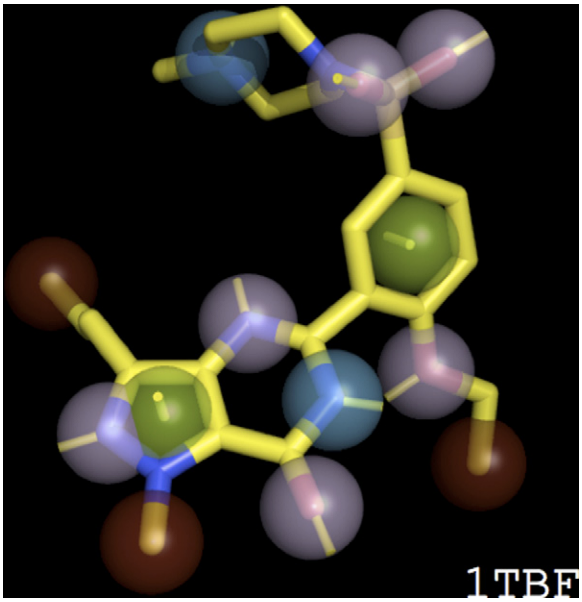


Fig. 7. Compound 1TBF together with its calculated pharmacophore points. Conformation obtained from the PDB file.

Table 6
Case 2: results of the virtual screen

Complete pharmacophore model					
TANIMOTO			TVERSKY_REF		
Rank	Ref.	%	Rank	Ref.	%
1	1TBF	0.05	1	1TBF	0.05
2	1XPO	0.10	2	1XPO	0.10
3	1UDT	0.15	3	1UDT	0.15
4	1UHO	0.20	4	1UHO	0.20

3.3. Case 3

In this last example the application of *Pharao* with respect to the clustering of a collection of molecules and the correlation with biological activity is demonstrated.

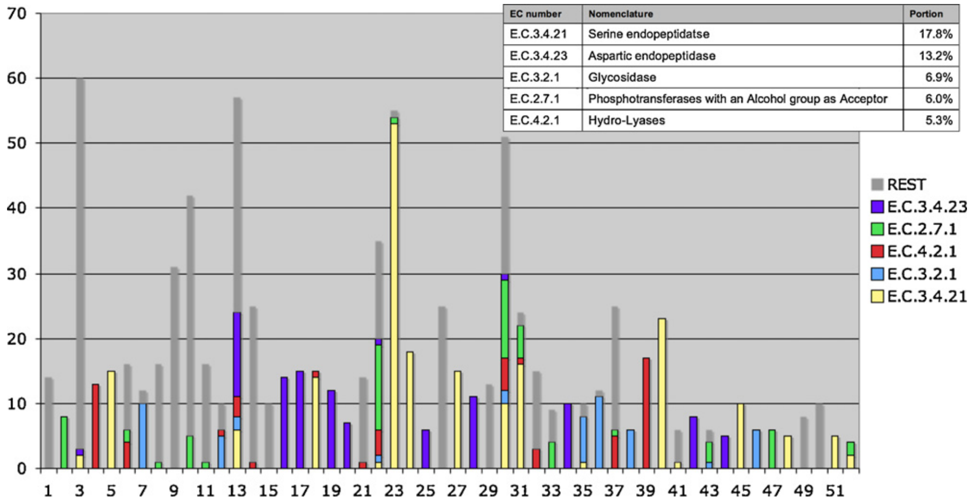
As a starting point, the 'refined data' set from the PDBbind (v.2007) initiative [21,22], consisting of 1300 ligand–target complexes, was taken. All ligands with their 3D structure were processed by *Pharao*. All pharmacophores with more than 10 HACC points or more than five HDON points were excluded from further investigation. This way, a filtered 'drug-like' set consisting of 1121 pharmacophores was obtained.

To investigate the biological relevance of the clustering based on Pharao, each ligand was assigned a code corresponding to the Enzyme Commission number (EC number) of the corresponding protein to which the ligand was bound. EC numbers are a numerical classification scheme for enzymes, based on the chemical reactions they catalyze [23]. EC numbers consist of four numbers, representing a progressively finer classification of the enzyme. If the last number is discarded, 70 different classes were present in our set of 1121 compounds. E.C.2.1.1 is such an example and represents all methyltransferases. All undefined or incomplete EC numbers were discarded, resulting in a final set of 883 compounds to cluster.

The clustering was performed using the k-means clustering algorithm [24], a simple but effective unsupervised learning algorithm. The initial number of clusters was set to 70, but after pruning the singletons only 52 clusters remained using the *TANIMOTO* score as similarity measure for pharmacophore overlap.

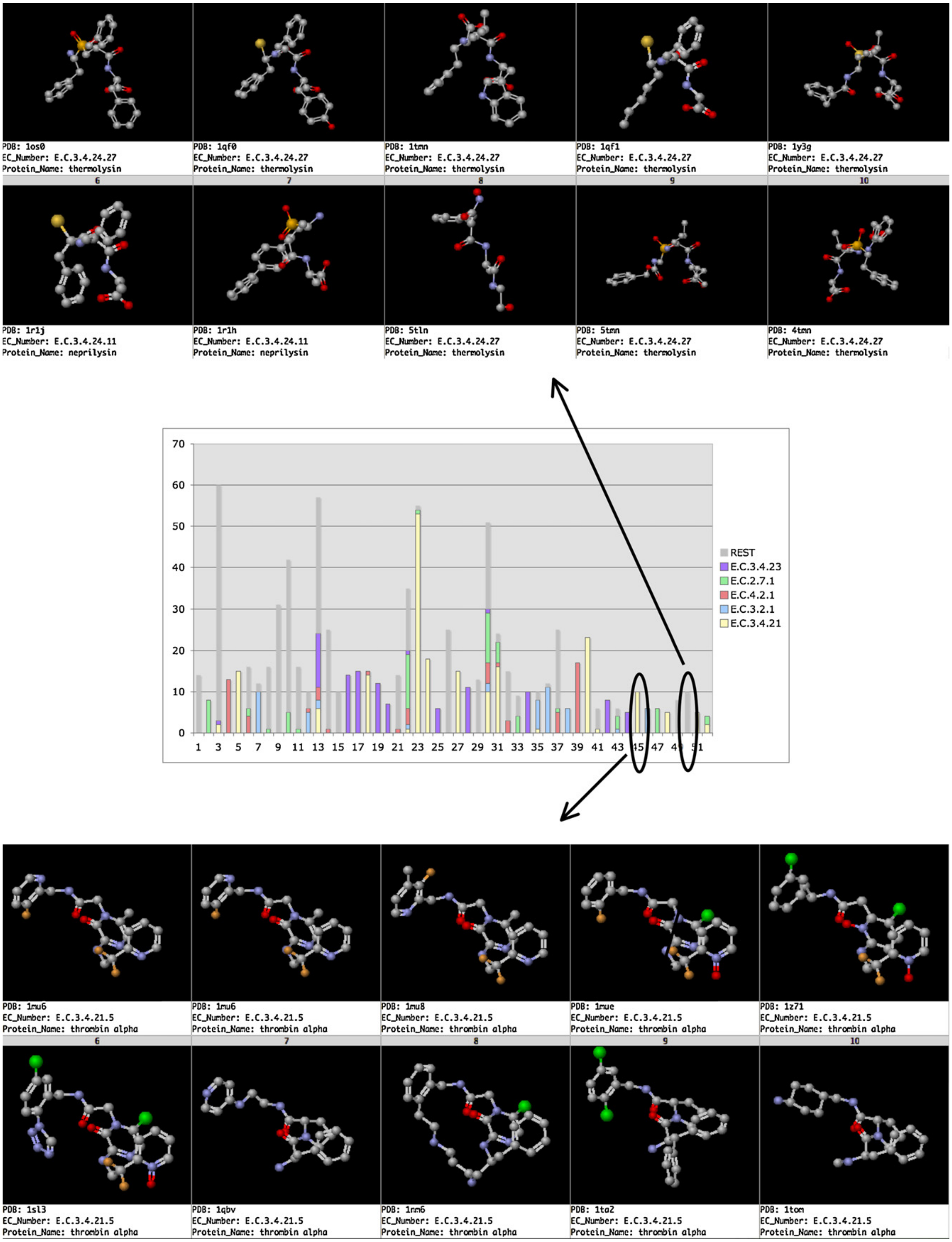
To visualize the biological relevance of the clustering, all clusters were plotted as bars, with their height corresponding to the number of molecules they contain. In Fig. 8, all compounds belonging to the largest EC classes are highlighted, revealing their position in the clustering. The results are encouraging in the sense that molecules belonging to the same EC class are mostly grouped together. For example Clusters 4 and 39 solely consists of 'E.C.4.2.1 compounds' (red). The majority of all 'E.C.3.4.23 compounds' (purple) are divided into 10 clusters while a random sampling of these 110 compounds over 50 clusters would give a completely different results.

In Fig. 9, the content of two clusters, both consisting of 10 compounds, is shown. Cluster 45 contains entirely compounds of the 'E.C.3.4.21' class and cluster 50 contains entirely compounds of the 'E.C.3.4.24' class (metallo-endopeptidases). Despite the observed structural variance of compounds within a single cluster, they were all considered similar enough to be grouped together and representing the same biological function.



EC number	Nomenclature	Portion
E.C.3.4.21	Serine endopeptidatse	17.8%
E.C.3.4.23	Aspartic endopeptidase	13.2%
E.C.3.2.1	Glycosidase	6.9%
E.C.2.7.1	Phosphotransferases with an Alcohol group as Acceptor	6.0%
E.C.4.2.1	Hydro-Lyases	5.3%

Fig. 8. Visualization of the clustering performed in Case 3. All compounds belonging to the five most occurring EC classes are highlighted in different colors. The height of the bars corresponds to the number of compounds in each cluster.



PDB: 1os0
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1qf0
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1tmn
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1qf1
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1y3g
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1r1j
EC_Number: E.C.3.4.24.11
Protein_Name: neprilysin

PDB: 1r1h
EC_Number: E.C.3.4.24.11
Protein_Name: neprilysin

PDB: 5tln
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 5tmn
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 4tmn
EC_Number: E.C.3.4.24.27
Protein_Name: thermolysin

PDB: 1mu6
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1mu6
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1mu8
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1mue
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1z71
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1sl3
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1qbv
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1nm6
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1ta2
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

PDB: 1tom
EC_Number: E.C.3.4.21.5
Protein_Name: thrombin alpha

Fig. 9. Illustration of the content of clusters 45 and 50 from the clustering performed in Case 3 and illustrated in Fig. 8.

4. Conclusions

Pharao, a pharmacophore-based scoring function, is described in this work. With two virtual screening examples and one database clustering experiment, the applicability and usefulness of this tool within the context of drug discovery was demonstrated. Since the representation of pharmacophore points by Gaussian volumes seems to be successful, and given the success of the same approach in shape-based scoring methods [12], a combination of both techniques can be an interesting idea for further improvements.

References

[1] T. Langer, G. Wolber, Pharmacophore definition and 3D searches, Drug Discov. Today: Technol. 1 (2004) 3.

[2] O.F. Guner, Pharmacophore Perception, Development, and Use in Drug Design, International University Line, La Jolla, CA, 2000.

[3] J.H. Van Drie, Pharmacophore discovery: a critical review, in: P. Bultinck, H. De Winter, W. Langenaeker, J.P. Tollenaere (Eds.), Computational Medicinal Chemistry for Drug Discovery, Marcel Dekker, Inc., New York, 2004, pp. 437–461.

[4] N.W. Murrall, E.K. Davies, Conformational freedom in 3D databases. 1. Techniques, J. Chem. Inform. Comput. Sci. 30 (1990) 312–316.

[5] D. Barnum, J. Greene, A. Smellie, P. Sprague, Identification of common functional configurations among molecules, J. Chem. Inform. Comput. Sci. 36 (1996) 563–571.

[6] G. Jones, P. Willett, R.C. Glen, A genetic algorithm for flexible molecular overlay and pharmacophore elucidation, J. Comput.-Aided Mol. Des. 9 (1995) 532–549.

[7] G. Wolbert, T. Langer, LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters, J. Chem. Inform. Model. 45 (2005) 160–169.

[8] J. Feng, A. Sanil, S.S. Young, PharmID: pharmacophore identification using Gibbs sampling, J. Chem. Inform. Model. 46 (2006) 1352–1359.

[9] J. Sadowski, C.H. Schwab, J. Gasteiger, 3D structure generation and conformational searching, in: P. Bultinck, H. De Winter, W. Langenaeker, J.P. Tollenaere (Eds.), Computational Medicinal Chemistry for Drug Discovery, Marcel Dekker, Inc., New York, 2004, pp. 151–213.

[10] M.J. Loferer, I. Kolossvary, A. Aszodi, Analyzing the performance of conformational search programs on compound databases, J. Mol. Graphics Model. 25 (2007) 700–710.

[11] J.A. Grant, M.A. Gallardo, B.T. Pickup, A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape, J. Comput. Chem. 17 (1996) 1653–1666.

[12] T.S. Rush, J.A. Grant, J.A. Mosyak, A. Nicholls, A shape-based 3D scaffold hopping method and its application to a bacterial protein–protein interaction, J. Med. Chem. 48 (2005) 1489–1495.

[13] J. Greene, S. Kahn, H. Savoj, P. Sprague, S. Teig, Chemical function queries for 3D database search, J. Chem. Inform. Comput. Sci. 34 (1994) 1297–1308.

[14] J. Figueras, Ring perception using breadth-first search, J. Chem. Inform. Comput. Sci. 36 (1996) 986–991.

[15] B.L. Roos-Kozel, W.L. Jorgensen, Computer-assisted mechanistic evaluation of organic reactions. 2. Perception of rings, aromaticity, and tautomers, J. Chem. Inform. Comput. Sci. 21 (1981) 101–111.

[16] J.A. Grant, B.T. Pickup, A Gaussian description of molecular shape, J. Phys. Chem. 99 (1995) 3503–3510.

[17] C.F.F. Karney, Quaternions in molecular modeling, J. Mol. Graphics Model. 25 (2006) 595–604.

[18] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[19] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739–1749.

[20] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, J. Med. Chem. 47 (2004) 1750–1759.

[21] R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures, J. Med. Chem. 47 (2004) 2977–2980.

[22] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, The PDBbind database: methodologies and updates, J. Med. Chem. 48 (2005) 4111–4119.

[23] http://www.chem.qmul.ac.uk/iubmb/enzyme/.

[24] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, (1967), pp. 281–297.

[25] O. Guvench, D.J. Price, C.L. Brooks 3rd, Receptor rigidity and ligand mobility in trypsin–ligand complexes, Proteins 58 (2005) 407–417.