Journal of Molecular Graphics and Modelling 27 (2009) 836–845

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology

Pedro J. Ballester [a,*], Paul W. Finn [b], W. Graham Richards [a]

[a] Physical & Theoretical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, UK
[b] InhibOx Limited, Pembroke House, 36-37 Pembroke Street, Oxford OX1 1BP, UK

ARTICLE INFO

Article history:
Received 31 October 2008
Received in revised form 5 January 2009
Accepted 6 January 2009
Available online 14 January 2009

Keywords:
Molecular shape comparison
Ligand-based virtual screening
Drug Lead identification
Similarity search
Chemoinformatics

ABSTRACT

Large scale database searching to identify molecules that share a common biological activity for a target of interest is widely used in drug discovery. Such an endeavour requires the availability of a method encoding molecular properties that are indicative of biological activity and at least one active molecule to be used as a template. Molecular shape has been shown to be an important indicator of biological activity; however, currently used methods are relatively slow, so faster and more reliable methods are highly desirable. Recently, a new non-superposition based method for molecular shape comparison, called Ultrafast Shape Recognition (USR), has been devised with computational performance at least three orders of magnitude faster than previously existing methods. In this study, we investigate the performance of USR in retrieving biologically active compounds through retrospective Virtual Screening experiments. Results show that USR performs better on average than a commercially available shape similarity method, while screening conformers at a rate that is more than 2500 times faster. This outstanding computational performance is particularly useful for searching much larger portions of chemical space than previously possible, which makes USR a very valuable new tool in the search for new lead molecules for drug discovery programs.

1. Introduction

Computational and medicinal chemists often encounter a situation where a known active molecule does not provide a viable starting point for drug discovery and development, perhaps because of toxicological, potency, selectivity or intellectual property issues. In these circumstances one wants to identify alternative molecular scaffolds that retain the desired biological activity of the initial lead but which are devoid of its disadvantages. Empirical testing of large numbers of physical samples (High Throughput Screening or HTS) has been widely and successfully employed as a source of new leads; however, the huge costs of large scale experimental testing and its relatively slow operation have motivated the development of computational approaches for lead identification.

When the coordinates of the target protein are known, receptor-based virtual screening has probably been the most widely applied technique. However, protein structural data are not available for many therapeutic targets. Furthermore, the computational performance of these methods means that they cannot be applied on the scale of the available databases and a number of difficult technical and scientific challenges remain to be overcome [1]. Therefore, ligand-based approaches continue to represent an attractive alternative when active molecules that can be used as templates are known, which is very often the case. Additionally, very large databases of chemical structures and the growing interest in exhaustive exploration of the chemical universe [2] provides a stimulus to develop the faster and accurate methods needed to mine these databases effectively.

A widely used ligand-based approach is that based on similarity searching, with many different quantitative measures of the similarity between two molecules having been proposed [3]. Most commonly these are based on the 2D structures of the molecules where the number of common, and not in common, substructures in the compared molecules is used to calculate a numerical indication of similarity, e.g. the Tanimoto coefficient, which can then be used to rank the results of a database search. Approaches that consider the 3D structures of molecules include the wide variety of pharmacophore perception techniques (typically requiring several active, ideally structurally diverse, templates to construct the query with which to interrogate the database), 3D-QSAR and shape comparison methods. Many of these 3D methods require previous alignment, or superposition, of the two molecules being compared. This is a vital and often computationally expensive aspect of these methods and many approaches to

* Corresponding author. Current address: Unilever Centre for Molecular Science Informatics, Cambridge University, Lensfield Road, Cambridge CB2 1EW, UK.
E-mail address: pedro.ballester@gmail.com (P.J. Ballester).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.01.001

the alignment problem have been proposed [4]; however important difficulties in this pursuit have been pointed out [5].

This paper focuses on molecular shape comparison as a way to carry out ligand-based Virtual Screening. This approach is attractive given the widely highlighted [8,9,11] central role of shape-complementarity in molecular recognition events as an important indicator of a molecule's biological activity. Indeed, without such complementarity, the ligand and receptor atoms involved in binding would not be sufficiently close to allow favourable interactions, such as hydrogen bonds and ionic interactions. Molecular shape comparison methods are also naturally suited for finding chemically distinct molecules likely to have similar biological activity, as different chemical scaffolds may support similar molecular shapes. Furthermore, it is possible to enrich the search with chemical information relevant to bioactivity beyond that already implicit in molecular shape. One way to construct such hybrid method would be to devise a two-stage hierarchical procedure whereby one would search first for similarly shaped molecules and thereafter feed this much smaller subset into a second virtual screening method (e.g. electrostatic similarity, pharmacophore search or 2D topological methods) in order to improve overall performance.

Most previously reported shape comparison methods rely on optimal molecular superposition. The first generation of such methods calculated molecular overlap by the counting of points on a grid where the two molecules coincided [6]. This approach was hampered by the difficulty of finding optimal solutions, dependence on the grid spacing and it was computationally inefficient. A substantial improvement in performance and robustness was made with the introduction of Gaussian functions to represent shape through electron density functions [7]. Other successful applications of Gaussians to molecular shape comparison were based on the concept of molecular volume overlap [8]. A recent implementation [9] of the latter approach has increased performance further, albeit at the expense of introducing some approximations into the approach [5]. These shape-based methods have shown themselves to be useful in identifying novel chemical scaffolds [10] and to perform well in comparison to docking approaches [11]. Non-superpositional shape comparison methods have also been developed, examples of which include those based on atom triplet distances [12] and shape signatures [13], with the latter having shown remarkable accuracy and efficiency. A critical review of representatives from the various families of molecular shape comparison methods is included in a recent paper [5].

Recently, one of the authors devised a new non-superposition based shape comparison approach, called Ultrafast Shape Recognition (USR) [14]. USR has been shown [5] to describe and compare the shape of molecules accurately using the same database that is used in this paper. In a related paper [15] aimed at a wider audience, the better accuracy in this task in comparison with a commercially available shape similarity method was discussed. However, a distinguishing feature of USR is its efficiency, as it screens conformers at a rate which is at least three orders of magnitude faster than that of pre-existing effective shape similarity methods (see Ref. [5] for further discussion). Thus, USR offers the possibility of application to databases of much greater size using many more active templates than previously possible.

The aim of this paper is to study the ability of USR as a stand-alone method to identify molecules sharing common biological activities through retrospective virtual screening experiments. This will be carried out in comparison to a commercially available shape similarity method and what would be expected by testing randomly selected compounds in the laboratory. The rest of the paper is organized as follows. The next section explains the experimental setup of this validation exercise, including the test database characteristics and performance evaluation measurement. The third section describes the shape comparison methods to be tested and the proposed template selection procedure. The fourth section presents and discusses the obtained results. The paper finishes by summarizing the conclusions of the study.

2. Experimental setup

A validation study of the ability of computational methods to retrieve molecules of common biological activity to a query molecule obviously requires a database that contains both chemical and biological data. The selection of a suitable database is not straightforward because whilst large databases of chemical structures are widely available, the associated biological data are usually lacking. Additionally, for any method based on ligand–receptor shape complementarity, such as shape similarity methods, the relevant level of description of the biological activity is at the level of the receptor (indeed, ideally at the level of the particular binding site on that receptor) rather than at a higher level of description. For example "HIV protease inhibitor" is a more suitable activity indication for virtual screening than "antiviral". Because of these considerations, it has been common to generate a test set of molecules by spiking a chemical database of assumed inactives with known actives. For the studies performed here we have decided to use DrugBank and retrieved all the available structures from the FDA-approved (708) and experimental (3056) drug sets. Although the database is smaller than those used in some other studies, it is sufficient for these validation studies because it represents essentially the same proportion of actives and inactives that one would usually find in a HTS screen. Also, DrugBank is publicly available and freely accessible, so future comparison studies are facilitated. In addition, as also pointed out by McGaughey et al. [16], most molecules in currently used validation databases have only been tested for a few activities and thus they might be actually active against another target if only they had been tested there. This common situation in retrospective studies is undesirable since it clearly affects the validation. However, as the tested methods are of the same nature (shape based) and exactly the same molecular conformations from query and test databases will be made available to each of them, there is no reason to think that this issue will affect the performance of such methods differentially.

We have utilized eight sets of active molecules taken from DrugBank [17]. These sets were chosen in an unbiased manner, before any experiments took place, to span a range of different target proteins and classes. They include two nuclear hormone receptor structures, estrogen receptor and progesterone receptor. AT1 binds a peptide hormone angiotensin. No crystal structure of the receptor, a GPCR, has been determined. Two other targets are members of the 5-HT and histamine receptor families. The enzyme targets are also diverse, with one kinase, a cyclo-oxygenase and neuraminidase. These eight target affinities along with the number of active molecules in each target are specified in Table 1.

The final step for the generation of the test database is to calculate 3D molecular conformations for each of the considered 2D chemical structures. The conformers of a particular molecule are in general geometrically distinct and have low potential energy, as conformers with high internal energy are in principle less likely to occur in Nature. This potential energy is parameterised in what is called a Molecular Mechanics Force Field. In this work, the implementation of MMFF94 [18] provided with MOE [19], a widely used Molecular Modelling software package, is used. MMFF94 has been parameterised for a wide variety of chemical systems of interest to organic and medicinal chemists using computationally derived and experimental data. The high accuracy

Table 1
Target affinities for the considered active molecules.[a].

i	Target affinity	Full name	A_i	P_i (%)
1	NM	Neuraminidase	8	0.24
2	AT1	Type-1 Angiotensin II Receptor	8	0.24
3	PR	Progesterone Receptor	12	0.36
4	TK	Thymidine Kinase	13	0.39
5	5-HT-2A	5-HT-2A Receptor	15	0.45
6	ER	Estrogen Receptor	24	0.72
7	COX-2	Cyclooxygenase-2	28	0.84
8	HH1R	Histamine H1 Receptor	41	1.23

[a] Eight targets were selected from http://www.drugbank.ca on the basis of having been previously used in retrospective virtual screening and being diverse. For each target, the abbreviation, the name and the number of actives is displayed (A_i), which ranges from 0.2% to 1.2% of the total number of compounds (P_i). All the available actives at the time of generating the test database were included, without any pre-selection that could favour a particular virtual screening method being made.

with which these diverse data could be reproduced by MMFF94 [18] suggests its suitability for small organic molecules.

In order to sample the potential energy landscape for a given molecule, a conformational search technique is applied. MOE's stochastic conformational search uses a parallelized fragment-based approach, where molecules are subdivided into overlapping fragments each of which undergoes stochastic search for optimal conformations. The fragment conformations are rapidly assembled by superposing the overlap atoms. A database of fragments is maintained, and augmented as the search proceeds, making conformation generation very fast. MOE's high throughput conformation generator was used with filters allowing only molecules in the range 100–800 Da and with more than 10 heavy atoms, strain limited to 30 kcal/mol and an upper limit of 1000 conformers per molecule. The generated conformers were then filtered such that molecules with energy higher than 0.7 kcal/mol/rotor were removed. This process led to a database with 666,892 conformers containing 3330 chemical structures, that is an average of about 200 conformations per compound.

The next area to look at is how the performance of virtual screening methods is going to be assessed. A good virtual screening method is one that is able to retrieve a subset of molecules containing a significantly higher proportion of actives than would be expected at random. Indeed, one would ideally like to find a number of structurally different active molecules to follow up from the testing of a small number of compounds (the smaller the number of tested compounds needed to provide the desired number of actives, the faster and cheaper the testing, and thus the greater the value of the virtual screening technique). Measuring the performance in such a task has been approached in a number of ways, notably through Enrichment Factor (EF) and ROC/AUC performance measures. We prefer to use early enrichments over ROC curves because the former is a representative measure of the performance of a real-world virtual screening experiment, whereas we consider that the latter is not. As also pointed out by other authors [16], the problem with ROC curves is that it weights evenly the whole accumulation curve, whereas in practice only the very beginning of the curve is relevant to virtual screening performance (the reason for this is that the rest of the database compounds cannot possibly be followed up and thus their activity is not experimentally determined in practice). Therefore, we will measure virtual screening performance by calculating enrichment at the top x% most similar molecules in the database, as the proportion of actives in the selected subset over the proportion of actives in the whole database (note that the active compound used as the query does not count as a found active and therefore zero enrichment is possible at small cutoffs). This cutoff x needs to be sufficiently low to simulate real-world virtual screening, where molecular databases are very large in comparison with the number of compounds that can possibly be tested. With this in mind, calculating enrichments at the top 1% of small test databases has been suggested [16] as a realistic performance measure.

The Enrichment Factor is defined in the usual way as the proportion of actives in the set of molecules retrieved (the top 1% of the sorted database in our case) over the proportion of actives that would be obtained by screening the whole database. Mathematically,

$$\mathrm{EF}_{ij,1\%} = \frac{a_{ij,1\%}/c_{1\%}}{a_{ij,100\%}/c_{100\%}} = \frac{a_{ij,1\%}/c_{1\%}}{A_i/C} \quad (1)$$

where $c_{x\%}$ is the number of compounds in the top x% of the sorted database and $a_{ij,x\%}$ is the number of actives retrieved at the top x% of the run generated by the jth active template from the ith target. By the definition, the denominator of this expression becomes the ratio between the total number of actives for the ith target (A_i; see Table 1) and the total number of compounds in the database (C). This expression can be averaged over all actives for a target to measure the mean performance of a method on that target

$$\overline{\mathrm{EF}}_{i,1\%} = \frac{1}{A_i}\sum_{j=1}^{A_i}\mathrm{EF}_{ij,1\%} \quad (2)$$

and over all the studied targets to provide the average performance of the tested method

$$\overline{\overline{\mathrm{EF}}}_{1\%} = \frac{1}{8}\sum_{i=1}^{8}\overline{\mathrm{EF}}_{i,1\%} \quad (3)$$

Lastly, the diversity of the molecules used in retrospective virtual screening studies has been widely recognised as an important factor to take into account. Because of the way molecular databases are populated (combinatorial chemistry, functional group substitutions from already existing database molecules, etc.), database molecules tend to be more similar in chemical structure than they would be if we could sample the chemical space in an unbiased manner. This has been pointed out [20] as a factor that can lead to an unrealistically high performance of virtual screening methods. It is important to note however that biases in molecular databases affect virtual screening methods differently. Whereas structurally similar molecules can lead to very different 3D shapes, techniques such as 2D structural similarity or atom counts tend to retrieve by definition chemically similar molecules and thus the latter have a larger advantage in databases with close analogues. Indeed, it has been seen elsewhere [10] that the top hits from a shape similarity method are not particularly similar in terms of their chemical structure.

3. Methods

This section describes the molecular shape matching methods that will be compared and the proposed template selection procedure.

3.1. Ultrafast shape recognition (USR)

USR is based on the observation that the shape of a molecule is uniquely determined by the relative position of its atoms. Such positions are in turn determined by the set of all inter-atomic distances. This set contains more information than is needed to describe the shape of the molecule accurately [5], so it is possible to significantly reduce the associated computational cost whilst maintaining accuracy by selecting a suitable subset of inter-atomic distances. In this work, the set of all atomic distances from four molecular locations are considered: the molecular centroid (ctd),

the closest atom to ctd (cst), the farthest atom from ctd (fct) and the farthest atom from fct (ftf). These locations represent the centre of the molecule and its extremes, and thus are well separated. The set of atomic distances from the molecular centroid was included since this location is far from cst for some molecules while being inexpensive to calculate. In this way, each molecular conformation is described by four distributions of atomic distances, where the number of atomic distances is proportional to the number of atoms. This raises the obvious question of how to compare molecules with different number of atoms. That difficulty is circumvented by defining a fixed number of moments of the 1D distributions, whose values characterise the molecule considered. Finally, the shape similarity score of two molecules is calculated through the sum of least absolute differences of their respective moments. Note that the characteristics of the method ensure that a high score will always be assigned to a molecule with similar shape, which would be consequently ranked highly. Further details about the design principles of the method have been published previously [5] (Figs. 1 and 2).

The calculation of the molecular descriptors is as follows, for each molecule in the database. First, the three dimensional position vector for each atom is read. Thereafter, the geometrical centre (centroid) of the molecule is determined from the atomic positions. Next, the set of Euclidean distances of all atoms to the molecular centroid is calculated. These are regarded as the full population of the distribution of all atomic distances from the molecular centroid:

$$\{d_j^{\mathrm{ctd}}\}_{j=1}^{N} \tag{4}$$

where N is the number of atoms of the molecule

The next stage of the process is to calculate the moments of this discrete distribution in order to characterise the geometry of the molecule and thus its shape. The first moment (μ_1^{ctd}) is the average atomic distance to the molecular centroid and thus it provides an estimate of the molecular size. The second moment (μ_2^{ctd}) is the square root of the variance of these atomic distances about μ_1^{ctd}. The third moment (μ_3^{ctd}) is the cube root of the skewness of these atomic distances about μ_1^{ctd} (i.e. a measure of the asymmetry of the distribution). These roots are intended to provide all moments with linear space dimension, typically Å, in order to avoid differences in high order moments overshadowing the contribution to the similarity score of low order moments. Such a balanced set of USR descriptors results in an improvement over the original version of USR [5,15], which used descriptors with different physical units. To calculate the remaining nine descriptors, we repeat the process for each of the three remaining distributions: $\{d_j^{\mathrm{cst}}\}_{j=1}^{N}$, $\{d_j^{\mathrm{fct}}\}_{j=1}^{N}$ and $\{d_j^{\mathrm{ftf}}\}_{j=1}^{N}$, where the superscript indicates the location from where the atomic distances are calculated. Of course, one can include more reference locations or higher order moments leading to more descriptors and thus an even more accurate description of shape. However, we selected the first three moments from each of four considered 1D distributions to describe a molecule $\vec{M} = (\mu_1^{\mathrm{ctd}}, \mu_2^{\mathrm{ctd}}, \mu_3^{\mathrm{ctd}}, \mu_1^{\mathrm{cst}}, \mu_2^{\mathrm{cst}}, \mu_3^{\mathrm{cst}}, \mu_1^{\mathrm{fct}}, \mu_2^{\mathrm{fct}}, \mu_3^{\mathrm{fct}}, \mu_1^{\mathrm{ftf}}, \mu_2^{\mathrm{ftf}}, \mu_3^{\mathrm{ftf}})$, since this choice provides an excellent compromise between the efficiency and the effectiveness of the method. Lastly, it is worth noting that the USR descriptors of a given molecular database only need to be calculated once, as these descriptors can be easily stored due to the very concise characterisation of shape achieved (each conformer is described by 12 real numbers). For instance, the largest test database that we will be using has a size on disk of 1.65GB in 3D MDL SD format, whereas the corresponding file of descriptors is only 78MB including a string identifying the conformer (22.6MB when compressed). This large reduction in file size makes database handling and storage much easier.

Once the USR descriptors are available, a score quantifying the similarity between molecules based on these descriptors is required to rank the conformers in a database according to their shape similarity to a given template. First, the Manhattan distance between the vectors of shape descriptors of the query and the currently screened conformer is calculated, and divided by the number of descriptors. The resulting dissimilarity measure is transformed into a normalised similarity score by translating the dissimilarity by one unit and inverting the resulting value. Other ways to define a normalised similarity score could be of course adopted, as long as the similarity score is inverse-monotonic with respect to the dissimilarity, so as to preserve the ranking order. The similarity score function S_{qi} is therefore:

$$S_{qi} = \left(1 + \frac{1}{12}\sum_{l=1}^{12}\left|M_l^q - M_l^i\right|\right)^{-1} \tag{5}$$

where $\vec{M}^q$ and $\vec{M}^i$ the vector of shape descriptors for the query and ith screened conformer, respectively.

A number of variants of the original version of USR have been already implemented and applied to diverse research areas (e.g. [21,22]). In the area of virtual screening, Cannon et al. [22] implemented a variant of USR with the first four unbalanced moments of each distribution of atomic distances and incorporated additional chemical information through 2D structural similarity.

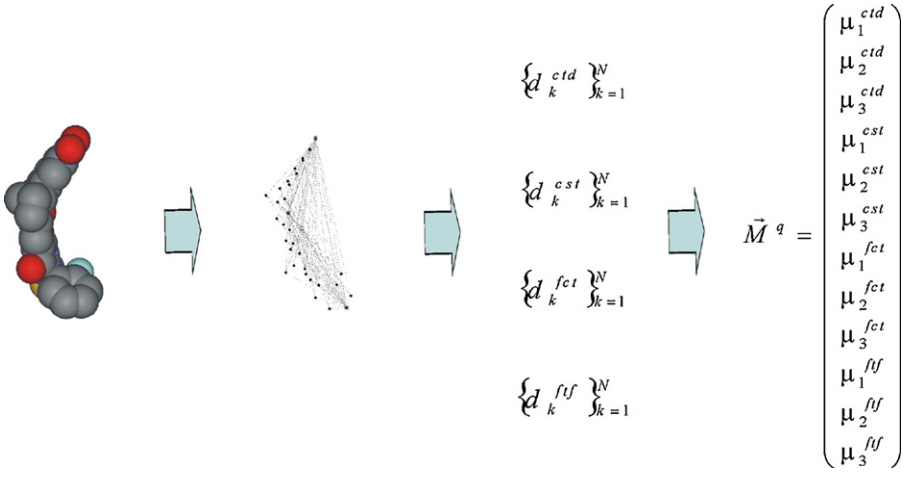

Fig. 1. USR encoding. The shape of the molecule is characterised by the distributions of atomic distances to four strategic reference locations. In turn, each of these distributions is described through its first three moments. In this way, each molecule has associated a vector of 12 shape descriptors.

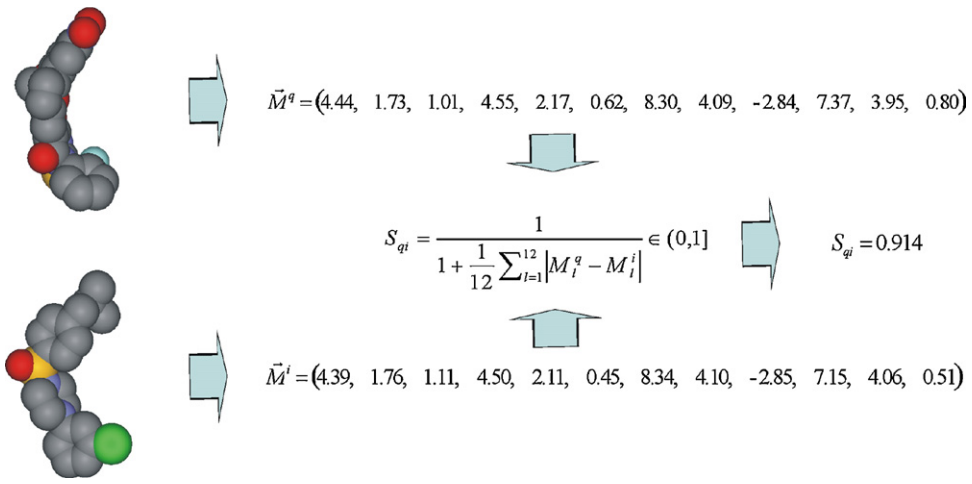

Fig. 2. Comparing the shape of two conformers with USR. Each database conformer has a vector of 12 USR descriptors associated, which are used to compare them through a normalized similarity score.

This hybrid method exploited the information contained in multiple active molecules via a machine learning technique to obtain an outstanding performance in a retrospective virtual screening study.

3.2. Eigenspectrum shape fingerprints (ESshape3D)

ESshape3D is a commercially available technique included in the MOE Molecular Modelling package, which aims at rapidly determining the shape similarity of molecules. ESshape3D starts by calculating the matrix with the Euclidean distances between all heavy atoms in the molecule to thereafter form a spectrum characteristic of its shape with the matrix's eigenvalues. This spectrum consists of 120 descriptors that correspond to the square root of the eigenvalues on a normalised grid of 120 points. Next, the spectrum is encoded as a fingerprint for each molecular conformer. Finally, the similarity between the two molecules is calculated as the inverse of the distance between the corresponding fingerprints [19].

3.3. Selection of molecular conformers as templates

Ligand-based Virtual Screening methods require at least one active molecule to act as the template with which to search the molecular database. This query molecule can be an approved or experimental drug, an HTS hit, a natural product or a failed drug lead. Even if we restrict ourselves to the public domain, there are nowadays very many such active molecules available. Traditionally, all active molecules are in principle regarded as equally suitable to act as templates. However, it is a well known fact that a particular method can obtain very different performance depending on the used query molecule, even when searching the same test database (incidentally, this is the reason why comparing methods only makes sense in the context of the same queries and test database). Therefore, we will use each available active molecule as a search query in turn to provide the average performance of each tested method.

Also, for a given active molecule, there is the question of which conformer represents most accurately its unbound bioactive conformation (i.e. the 3D geometry that fits the binding pocket before the former is possibly distorted by interactions with the target receptor) and hence should be taken as the template. If available, the 3D geometry of a ligand bound to the considered target, as determined by X-ray crystallography or Nuclear Magnetic Resonance (NMR) experiments, is normally used. However, many targets do not have experimentally determined ligands and, even if there is at least one, still there are usually many other active molecules for that target without an available experimentally determined bound conformation. An alternative to bound conformations as templates is using the lowest energy conformation (LEC) of the active molecule as the template. This approximation is particularly good for rigid molecules, as only a few conformers are possible and thus the likelihood that these represent the unbound bioactive conformation is high. For more flexible molecules, a number of studies (e.g. [23]) have investigated whether the set of computationally generated 3D conformations of a molecule include its unbound bioactive conformation by comparing this set with the corresponding bound conformation. However, it is unclear why the latter constitutes a meaningful ground truth, as it has been observed elsewhere [24] that a flexible molecule may undergo profound conformational changes in order to form optimal interactions with its binding partner. By contrast, unbound conformers are generated with molecular force fields that do not take into account possible interactions with the target protein. Due to this inconsistency, the unbound bioactive conformation could be in principle significantly different from the corresponding bound conformation. This fact stresses the importance of investigating alternative approaches to the selection of the bound conformation as the template such as the use of the LEC. This approximation to the unbound bioactive conformation is also expected to introduce a certain level of error leading to suboptimal performance of the tested methods. Nevertheless, as we will be using the same query and database conformers in our validation, any errors in these should affect all tested shape comparison methods equally, which ensures a fair comparison.

In addition to which conformation of a given molecule best represents the unbound bioactive conformation, there is the arguably more important question of which active molecule one should use as the template. In practice, an implicit selection of the query molecule is often carried out by restricting ourselves to those few molecules which happen to have an experimentally determined structure. However, the question arises as to which of the known actives best represents the biologically relevant property, ligand–receptor shape complementarity in this case, and therefore is likely to lead to the best possible results. We propose a new template selection procedure for Virtual Screening that consists in using the most common shape within the known active molecules for the considered target as the template. Thus, for a given target, each active molecule will be represented by its LEC as registered in the database and the resulting group of active

molecules will be clustered. The most common shape is by definition in the cluster with the highest number of active molecules (the main cluster), and is estimated as the closest of these conformers to the centroid of the main cluster. We will henceforth refer to such conformer as C1-CTD. Cluster centroids have been previously used [25] to study the ability of shape, and other molecular descriptors, to discriminate between a part of the known actives and a number of decoys, but this is the first time that their use for Virtual Screening is motivated, proposed and evaluated.

4. Results and discussion

The calculation of enrichments for each method is carried out as follows. In total, there are 149 active molecules spanning across the eight targets specified in Table 1. For each target, the LEC of the active molecule is used to query the test database so that the top 1% most similar molecules to this template is retrieved (as usual, conformers are collapsed into molecules by keeping only the highest ranking conformer of each molecule). This is repeated with each active for the ith target, which leads to A_i rankings of molecules, one for each query. Thereafter, the molecules retrieved in each run are checked against the corresponding list of known actives to determine how many actives, other than the template, have been found and the latter used to evaluate enrichments using Eqs. (1)–(3). The performances of the tested virtual screening methods against each target individually and on average over all targets are presented in Table 2.

These results show that USR performs better on average than ESshape3D ($\overline{EF}_{1\%}^{USR} = 10.4$ versus $\overline{EF}_{1\%}^{ESshape3D} = 6.6$). Both methods perform well above what would be expected on average by random selection of compounds ($\overline{EF}_{1\%}^{random} = 1$), despite the error introduced by using the LEC of each active as the template. However, no shape method dominates when looking at each target individually, as each method is best in half of the targets. A wide range of factors collectively contribute to the large variability in performance observed across targets. First, errors in the generation of 3D conformers may impact differentially on the top ranked molecules depending on the query molecule and thus on the considered target. A second factor is a common limitation of retrospective virtual screening: scarce activity data means that we are forced to assume that those molecules that have not been tested for the studied activity are inactive, when these could be actually active if only they had been tested. The ratio of these false negatives can clearly vary from target to target which contributes to the observed variability. Furthermore, there may be more than a unique type of shape for those molecules that are active against the same target. This could be due to the existence of different binding modes, the flexibility of binding pockets allowing a range of ligand shapes to be complementary to a number of protein conformations or having binding pockets that do not surround the ligand completely (the complementarity in shape would only need to be partial in that case). Lastly, a certain degree of shape complementarity between an unbound conformation of the ligand and the binding pocket of its macromolecular receptor is necessary, but not sufficient, for binding. Consequently, molecules that fit well the binding pocket may not remain bound because of a lack of sufficiently favourable interactions between the atoms at the ligand–receptor interface.

We next analyse the diversity of the actives in terms of shape similarity as follows. For each target, the LEC of each active molecule is considered. Thereafter, centroid-linkage hierarchical agglomerative clustering on the USR similarity matrix of these conformers is performed using $S_T = 0.75$ as the similarity score threshold. This threshold value represents a high level of shape similarity between molecules in the same cluster, but of course choosing other similar values would also be reasonable. Fig. 3 shows the obtained clusters of actives for each of the eight targets. The main cluster is defined as that with the highest number of actives and delimited in each plot by a dark blue box. The closest conformer to the centroid of this cluster, C1-CTD, represents the most common molecular shape in the set of actives.

This clustering analysis evidences that there is not a unique ligand shape for a particular target, but a number of them represented by each of the clusters. The exception is Neuraminidase, as all the actives for this target form a single cluster (the threshold line is not visible because it represents a much lower similarity and thus it is beyond the top of the plot), despite taking all the available experimental and approved drugs for these activities in DrugBank. Some targets such as COX-2 and HH1R present by contrast a high number of shape clusters, which could be due to different binding modes allowed by the binding pocket or even having more than one binding pocket. Another contribution to having dissimilarly shaped actives could come from having an open binding pocket, thus only requiring a partial ligand–receptor shape complementarity. Overall, the presence of several clusters in all but one target evidences the diversity of the actives in terms of shape similarity and hence discards a common source of bias in retrospective virtual screening studies. It is important to note that the diversity in terms of chemical structure, which may artificially enhance the performance of 2D topological methods, is largely irrelevant for this study. This is due to the fact that shape similarity methods are only mildly correlated with chemical structure as it has been already shown elsewhere [10]. In addition, we would like to stress the significance of having consistently found one or more clusters of actives in all the considered targets. The fact that, despite considering only the LEC for each active plus the presence of systematic errors in the generation of conformers, some active molecules in a number of diverse targets are similar according to USR is in itself convincing evidence that accurate description of shape is a robust indicator of biological activity. Also, it is clear that if our assumption that the LEC is in general a sufficiently accurate representation for virtual screening purposes of the unbound bioactive conformation were significantly inaccurate, we would not have found groups of active molecules forming clusters. We cannot stress enough that we are not talking here about bound bioactive conformations represented by experimentally determined structures, which may undergo large conformational changes upon binding to the target receptor and has been in some cases observed to be very different from the unbound LEC in flexible molecules [23].

Following the proposed template selection procedure for each target yields the eight templates shown in Fig. 4. This representative subset illustrates the high degree of diversity in shapes and flexibility that occur in the set of actives. Two nuclear hormone receptors, estrogen receptor and progesterone receptor, are considered as targets. The ligand queries for these receptors

Table 2
Mean enrichment per target and averaged over all targets for each of the tested methods.

Target	NM	AT1	PR	HH1R	TK	5HT2A	ER	COX-2	Average
$\overline{EF}_{i,1\%}^{USR}$	42.3	16.3	8.7	6.4	2.5	3.2	2.2	1.5	10.4
$\overline{EF}_{i,1\%}^{ESshape3D}$	16.3	4.9	13.0	5.9	3.1	2.3	3.8	3.2	6.6

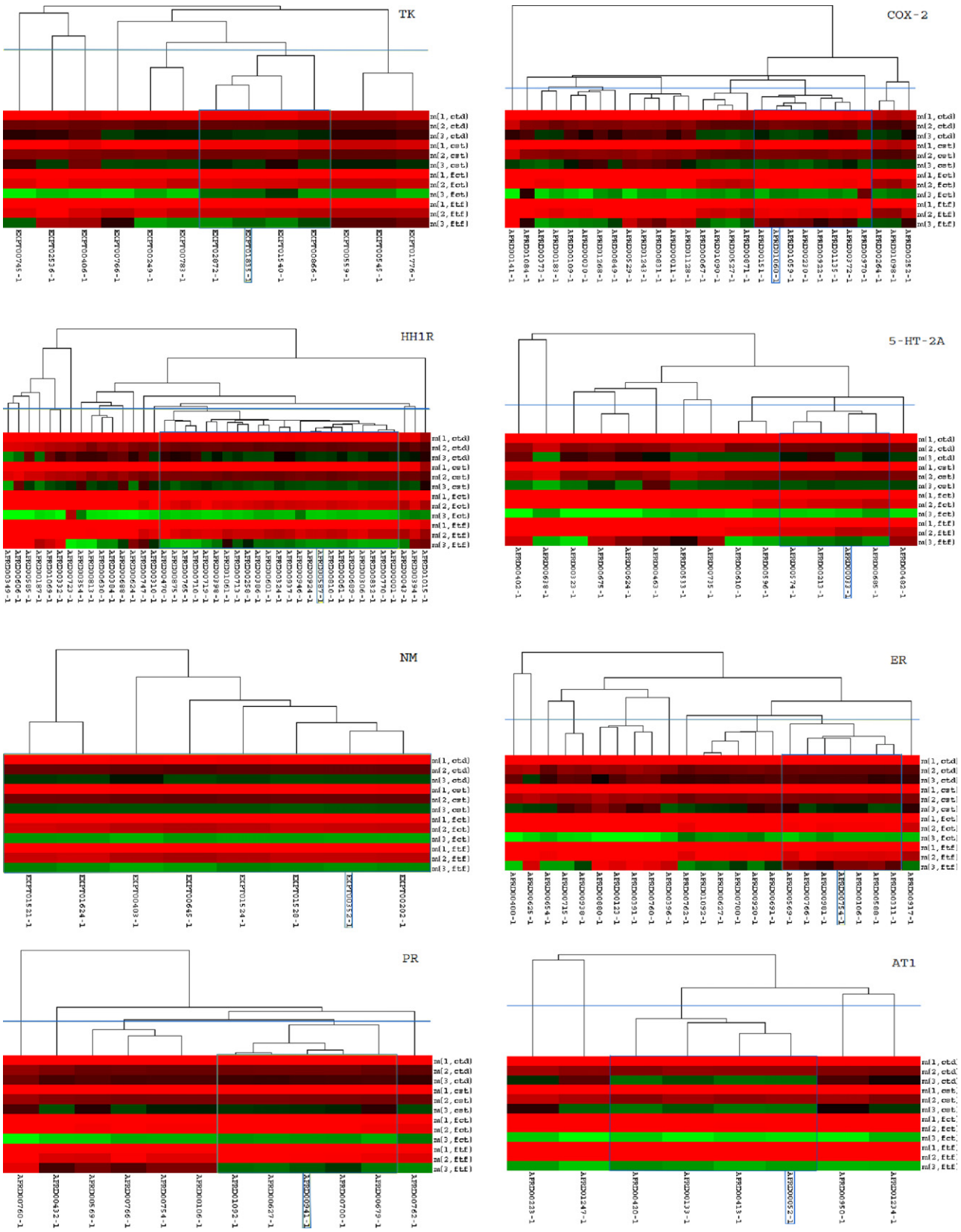

Fig. 3. Clustering of active molecules for each of the considered targets The shape of each active molecule is represented by its LEC. The horizontal dark blue line cuts the dendrogram at the similarity threshold value $S_T = 0.75$ and hence the conformers belonging to a branch below this intersection form a cluster at this level of similarity. The color matrix plots immediately below the dendrogram represent the values of USR descriptors (rows) across actives (columns). For instance, the fourth row shows that the first moment of the distribution of distances from the closest atom to the centroid (cst) has a similar value across actives, which reflects the fact that these actives are all similar in size. It is important to note that different USR descriptors cannot be compared using the plots, as having the same color only implies the same descriptor value within the same row.

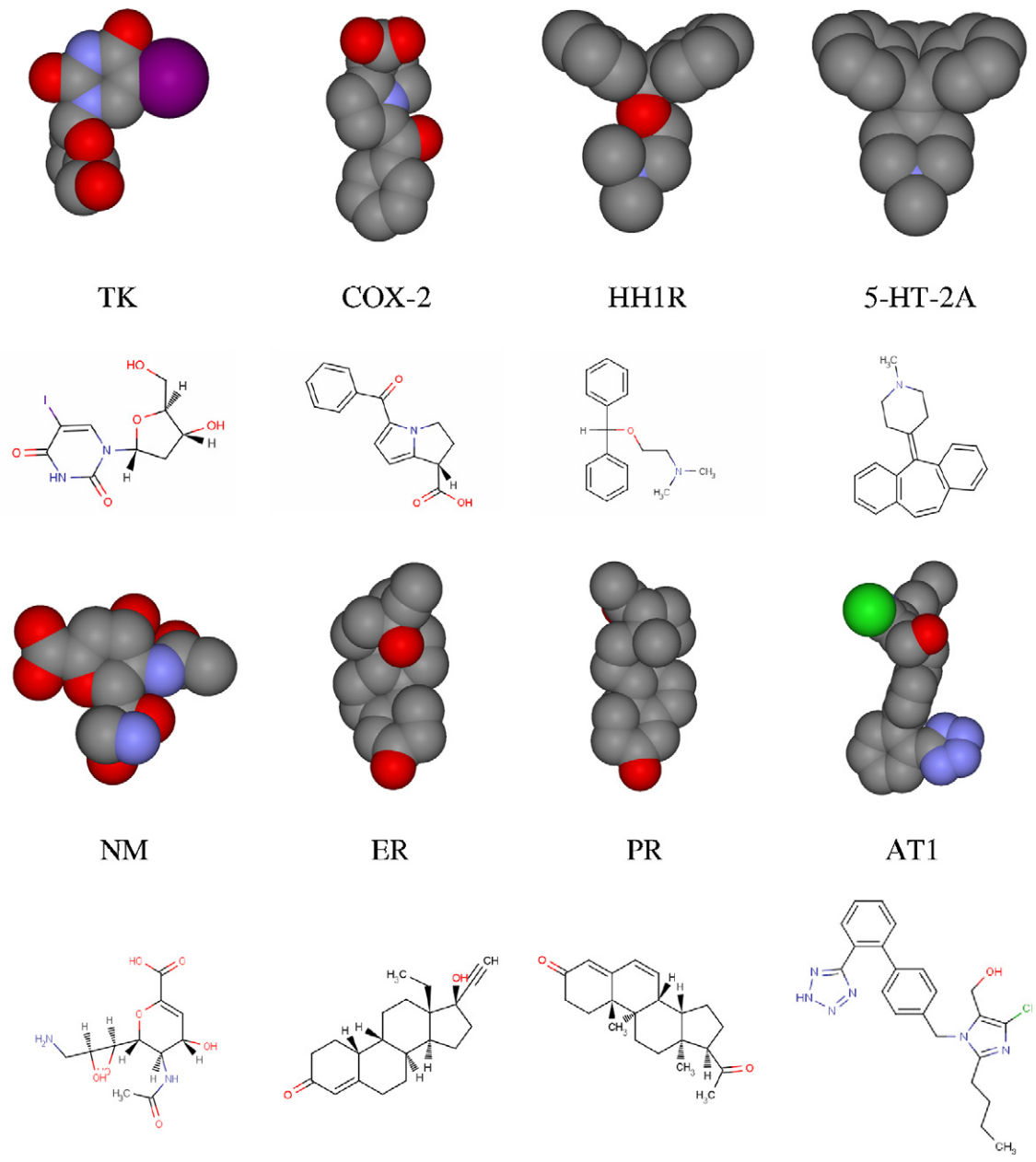

Fig. 4. Representative molecular shapes for each target and their corresponding chemical formula.

include polycyclic molecules with little conformational flexibility. Crystallographic data shows the binding site to be extensive, with the ligands almost completely surrounded by the receptors and mostly hydrophobic in character. The representative AT1 active is relatively flexible and contains a negatively charged carboxylic acid or isostere, such as tetrazole. Two other targets are members of the 5-HT and histamine receptor families. Actives in these targets exhibit various degrees of flexibility. The enzyme targets include challenging ones for a shape-based method, such as neuraminidase, which has a highly polar and relatively solvent exposed binding site.

Now that we have identified the most common shape for each target (shown in Fig. 4), we run USR for each of these C1-CTD queries and calculate enrichments for the top 1% of the sorted

Table 3
Best, C1-CTD, mean and worst enrichment for each target as well as averaged over all targets.

Target	NM	AT1	PR	HH1R	TK	5HT2A	ER	COX-2	Average
$EF^{USR}_{max,i,1\%}$	78.0	39.0	26.0	27.9	8.0	13.9	8.7	7.4	26.1
$EF^{USR}_{C1-CTD,i,1\%}$	52.0	26.0	26.0	25.4	0.0	13.9	4.3	3.7	18.9
$\overline{EF}^{USR}_{i,1\%}$	42.3	16.3	8.7	6.4	2.5	3.2	2.2	1.5	10.4
$EF^{USR}_{min,i,1\%}$	13.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6

Table 4
Enrichment in each target obtained by USR on two databases with the same molecules, but with a distinct average number of conformers per molecule (CPM).

Target	NM	AT1	PR	HH1R	TK	5HT2A	ER	COX-2	Average
$EF^{USR,CPM=200}_{C1-CTD,i,1\%}$	52.0	26.0	26.0	25.4	0.0	13.9	4.3	3.7	18.9
$EF^{USR,CPM=22.5}_{C1-CTD,i,1\%}$	52.0	26.0	26.0	27.9	0.0	0.0	13.0	11.2	19.5

molecular database. Table 3 shows the resulting eight enrichments along with the best, worst and mean enrichment for each target. It is observed that the enrichment obtained with the centroid of the main cluster is on average much better than what would be expected by selecting a template at random ($\overline{EF}^{USR}_{C1-CTD,1\%} = 18.9$ versus $\overline{EF}^{USR}_{1\%} = 10.4$) and not very much worse than the average of the best enrichment obtained in each target ($\overline{EF}^{USR}_{max,1\%} = 26.1$). It is also worth noting that there is a large difference between the best and the worst performance in each target despite applying the same method on the same database, which demonstrates that distinct active templates may lead to very different performance. This evidences the importance of selecting those actives with the most common shapes as templates, rather than using an active molecule simply because it has little flexibility or an available experimentally determined bound conformation.

There is also the question of whether the test database contains an appropriate average number of conformers per molecule (CPM). It is expected that allowing a higher CPM will increase the likelihood of having a close match to the unbound bioactive conformation in the set of computationally generated conformers. Since the active template aims at representing the unbound bioactive conformation, a higher likelihood of including the unbound bioactive conformation for each database active should ultimately result in retrieving a higher number of actives. This is the reason why we used a test database with a high CPM. However, it is important to note that a higher CPM also increases the likelihood of a proportionally much more abundant inactive molecule adopting the unbound bioactive conformation (these molecules would be those that fit the receptor, but do not make sufficiently favorable contacts to remain bound to the target). Consequently, increasing the CPM of a molecular database must be actually affecting the performance of ligand-based virtual screening methods in these two opposite directions, which means that there will be an intermediate CPM value that allows optimal performance. This optimal value should depend on the flexibility of the considered query and database molecules, and thus it will be target and database dependent. Although a detailed study of how performance is affected by the CPM of the test database is beyond the scope of this paper, it is relevant to verify whether a large reduction in the CPM in the used test database would not result in a large performance decrease. In order to test this commonly held view, we generated a second test database from the test database we have used so far by eliminating all the conformations with strain energy higher than 0.3 kcal/mol. The latter kept all 3330 molecules while reducing the total number of conformers to 74,886, i.e. this smaller database has CPM = 22.5 instead of the CPM = 200 of the original test database. The performance of USR on each test database is presented in Table 4. USR is run with each of the eight C1-CTD queries, which provides an enrichment value for each target. Not only was no large decrease in performance averaged across targets observed, but actually average performance becomes slightly better when reducing the CPM of the test database by an order of magnitude. As theoretically anticipated, it is also observed that this reduction in CPM is target dependent, as performance improvements are obtained in HH1R, ER and COX-2, but performance decreases in 5HT2A.

The last area to look at is the speed of operation or efficiency of the tested techniques. Efficiency is key in virtual screening. While the amount of computing power per researcher and target is often very limited, one would like to search as large molecular databases as possible in order to explore wider regions of the chemical space and thus discover higher numbers of structurally dissimilar actives. As a first step to evaluate efficiency, the shape descriptors for a given molecular database must be calculated. USR descriptors are calculated at a rate of 16,230 conformers per second, whereas those for ESshape3D are calculated at 450 conformers per second. Such calculation only needs to be carried out once, as descriptors can be thereafter stored in hard disk to be read and used as many times as desired[1]. However, as many queries are carried out in practice for a particular database, the important efficiency measure is the rate at which conformers are compared. Fig. 5 shows the efficiency of the tested shape similarity methods. USR is found to be more than 2500 times faster than the other tested shape similarity technique (ESshape3D)[2]. Similarly large improvements in efficiency have been also discussed [5] with respect to a wide variety of stand-alone shape comparison methods, both superposition and descriptor based. Such a high efficiency combined with the good enrichment performance obtained on a wide range of biological targets makes USR a very useful addition to the set of currently available virtual screening tools.

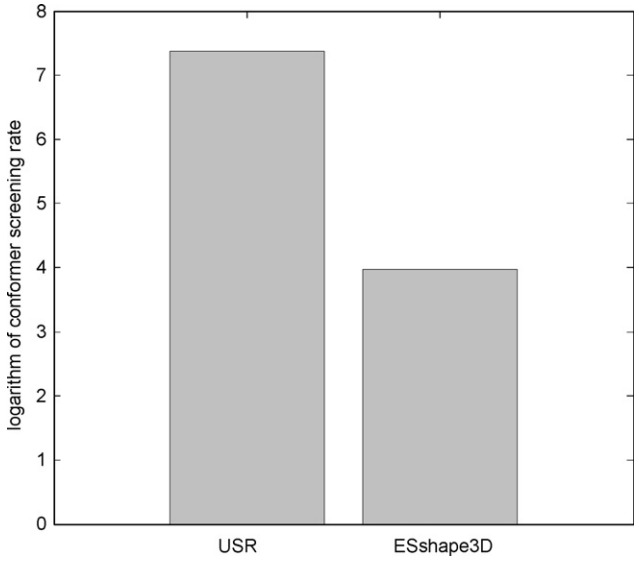

Fig. 5. Efficiency comparison, in logarithm of screened conformers per second, between USR and ESshape3D, the shape comparison method of a widely used molecular modeling software package. Both methods were run on one of the cores of an Intel Xeon 2.0 GHz with 8GB memory, using exactly the same query and test databases. USR obtained a comparison rate of 23,600,000 conformers per second, whereas ESshape3D performed at 9210 conformers per second.

[1] The time needed to read descriptors from hard disk is not reported in ESshape3D, but this is longer than that required for USR descriptors (3.7×10^{-6} s per conformer), due to a larger set of descriptors (120 descriptors for ESshape3D, for only 12 descriptors in USR).

[2] ESshape3D is embedded in the MOE software suite, so it is possible that CCG developers had to sacrifice some of the efficiency that ESshape3D would have if it was as stand-alone method in order to gain additional functionality and integrate it into the package.

5. Conclusions

The first retrospective virtual screening validation of a new shape matching technique (USR) has been presented. This study was performed on a benchmark with almost 150 active molecules spanning eight diverse targets, which were selected in an unbiased manner before any experiment took place. The actives were generally observed to be diverse in terms of shape similarity, with their proportion in the test database being essentially the same that would be found in a typical HTS screen, which ruled out the most common cause of bias in this type of study. USR was compared to the shape similarity method (ESshape3D) from a widely used software package using this benchmark, that is, exactly the same query and database conformers. Results show that USR achieves an overall mean enrichment that is on average better than that obtained by ESshape3D, while being thousands of times faster. This evidences that USR a very useful addition to the set of currently available virtual screening tools. An additional contribution of this paper is to demonstrate that using the lowest energy conformer or LEC of an active compound to query a database is an approximation that works well for virtual screening purposes on a range of diverse biological targets. Furthermore, we have demonstrated that using the most common shape in the set of known actives is a strategy that is expected to result in a much better enrichment than selecting the query molecule at random, as it is usually the case. In addition, we have seen that a much higher number of conformers per molecule does not necessarily lead to better performance. This factor clearly increases the number of conformers that are similar in shape to the template, but these conformers do not necessarily come from an active molecule. The magnitude of this effect is target and database dependent and we will carry out a detailed investigation of this issue in the future. Lastly, although we have seen that effective shape similarity alone is useful to identify molecules with a common biological activity, it is expected that incorporating additional chemical information relevant to the binding process will improve the virtual screening performance of USR further. This is clearly an interesting direction for future research.

Acknowledgement

P.J.B. thanks the US National Foundation for Cancer Research and Oxford University for funding.

References

[1] G.L. Warren, et al., A critical assessment of docking programs and scoring functions, J. Med. Chem. 49 (2006) 5912–5931.
[2] (a) T. Fink, H. Bruggesser, J.-L. Reymond, Virtual exploration of the small-molecule chemical universe below 160 daltons, Angew. Chem. Int. Ed. 44 (2005) 1504–1508; (b) T. Fink, J.-L. Reymond, Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes and drug discovery, J. Chem. Inf. Model. 47 (2007) 342–353.
[3] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[4] F. Melani, P. Gratteri, M. Adamo, C. Bonaccini, Field interaction and geometrical overlap: a new simplex and experimental design based computational procedure for superposing small ligand molecules, J. Med. Chem. 46 (2003) 1359–1371.
[5] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes, J. Comput. Chem. 28 (2007) 1711–1723.
[6] A.Y. Meyer, W.G.J. Richards, Comput. Aided Mol. Des. 5 (5) (1991) 427–439.
[7] A.C. Good, W.G. Richards, Rapid evaluation of shape similarity using Gaussian functions, J. Chem. Inf. Comput. Sci. 33 (1993) 112–116.
[8] A.J. Grant, 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) 1653–1659.
[9] ROCS, Openeye Scientific Software, Santa Fe, NM, USA, http://www/eyesopen.com.
[10] T.S. Rush III, et al., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction, J. Med. Chem. 48 (2005) 1489–1495.
[11] P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of shape-matching and docking as virtual screening tools, J. Med. Chem. 50 (2007) 74–82.
[12] G.W. Bemis, I.D. Kuntz, A fast and efficient method for 2D and 3D molecular shape description, J. Comput. Aided Mol. Design 6 (1992) 607–628.
[13] R.J. Zauhar, G. Moyna, L. Tian, Z. Li, W.J. Welsh, Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design, J. Med. Chem. 46 (2003) 5674–5690.
[14] P.J. Ballester, U.S. Patent no. 12/127559 (2007).
[15] P.J. Ballester, W.G. Richards, Ultrafast shape recognition for similarity search in molecular databases, Proc. R. Soc. A 463 (2007) 1307–1321.
[16] G.B. McGaughey, et al., Comparison of topological, shape, and docking methods in virtual screening, J. Chem. Inf. Model. 47 (4) (2007) 1504–1519.
[17] D.S. Wishart, et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 1 (34) (2006) D668–D672.
[18] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comput. Chem. 17 (5–6) (1996) 490–519.
[19] MOE (The Molecular Operating Environment) Version 2006.08, software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7, http://www.chemcomp.com.
[20] A. Bender, R.C. Glen, A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication, J. Chem. Inf. Model. 45 (2005) 1369–1375.
[21] Q.C. Nguyen, et al., Multiscale approach to explore the potential energy surface of water clusters $(H_2O)nn \leq 8$, J. Phys. Chem. A 112 (2008) 6257–6261.
[22] E.O. Cannon, F. Nigsch, J.B.O. Mitchell, A novel hybrid ultrafast shape descriptor method for use in virtual screening, Chem. Central J. 2 (2008) 3.
[23] J. Boström, et al., Assessing the performance of OMEGA with respect to retrieving bioactive conformations, J. Mol. Graphics Model. 21 (2003) 449–462.
[24] W.L. Jorgensen, Rusting of the lock and key model for protein–ligand binding, Science 254 (1991) 954–955.
[25] A. Nicholls, et al., Variable selection and model validation of 2D and 3D molecular descriptors, J. Comput. Aided Mol. Des. 18 (2004) 451–474.