Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes

PEDRO J. BALLESTER, W. GRAHAM RICHARDS
Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, OX1 3QZ, United Kingdom

Received 7 November 2006; Revised 20 December 2006; Accepted 21 December 2006
DOI 10.1002/jcc.20681
Published online 6 March 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Finding a set of molecules, which closely resemble a given lead molecule, from a database containing potentially billions of chemical structures is an important but daunting problem. Similar molecular shapes are particularly important, given that in biology small organic molecules frequently act by binding into a defined and complex site on a macromolecule. Here, we present a new method for molecular shape comparison, named ultrafast shape recognition (USR), capable of screening billions of compounds for similar shapes using a single computer and without the need of aligning the molecules before testing for similarity. Despite its extremely fast comparison rate, USR will be shown to be highly accurate at describing, and hence comparing, molecular shapes.


Key words: molecular shape comparison; chemical similarity search; geometrical shape descriptors; very large molecular databases

Introduction

Databases of three-dimensional (3D) molecular structures have been playing an increasingly important role in modern chemical and pharmaceutical research.1 These molecular databases gather an enormous amount of chemical knowledge such as the spatial location of the atoms in the molecule (whether this is theoretically calculated or experimentally determined by X-rays), its possible availability as a commercial compound or known synthetic route. Information about the flexibility of the molecule via the inclusion of several of its possible conformers, and even details regarding the biological activity of the molecule may also be included. This form of intellectual property is enormously valuable.2

The information within these molecular databases has been exploited in various ways. In molecular docking,3 the process of docking the screened molecule to a macromolecular biological receptor, almost always a protein, is simulated to provide an estimate of binding energy and thus likelihood of being bioactive. These techniques have spurred the generation of massive databases of drug-like molecules. This is the case in our widely publicised screensaver project in which a database of 3.5 billion compounds was screened against a series of proteins of known crystal structure by distributing the necessarily crude calculations over a grid of 3.5 million personal computers, each running a screensaver incorporating the binding energy estimation.

An alternative technique for mining the information contained in these databases is molecular shape comparison, which consists of searching the molecular database for compounds that most closely resemble the shape of a given query molecule. This chemical template can be a known product or inhibitor of a target protein, a natural product, or even a patented compound. Molecular shape has been widely acknowledged as a key factor for biological activity5–7 and thus is regarded as a very important pattern for which to search. There are now very successful applications of molecular shape comparison alone to identify biologically active compounds within molecular databases (e.g. in ref. 5). This success is attributed to the complementarity between the shape of the ligand molecule and its corresponding binding site at the macromolecular receptor, which is a prerequisite for binding. In words of Grant et al.8: “It [the importance of shape] arises because of the intermolecular interactions that stabilize the receptor-ligand complex are enthalpically weak and only become effective if the chemical groups involved can approach each other closely, which is favored by shape complementarity. Entropic contributions advantageous to binding involve the loss of complexation water of both the host and the guest and are also favored by shape complementarity (to avoid empty space being filled with water)”.

An additional advantage of searching a database for molecules with similar shape is that no specification of chemical structure,
such as types of atoms or their bond arrangements, is made and therefore similarly shaped molecules, but with different chemical scaffolds from the template, can be found. Such ability, known as Scaffold Hopping, is very valuable1 and has been used to discover nonintuitive, truly novel biologically active compounds within molecular databases.5,9 In particular, techniques for Scaffold Hopping can be used to break out the protected ‘patent space’ around drugs, or to hop to molecules with different scaffolds when lead compounds have undesirable features such as intractable chemistry or poor pharmacological properties.10

Unfortunately, we cannot currently reach the full potential of molecular shape comparison methods in these applications because of several major problems. Many methods using molecular shape as the pattern to recognize require previous alignment of the molecules being compared, which is an additional source of difficulty10,11 that may lead to suboptimal molecular overlap and thus to a wrong similarity score. Furthermore, appraising the effectiveness of the method (i.e. the extent to which the method assigns a higher similarity score the more similar are molecules being compared) according to molecular shape is a complicated endeavour, which we will discuss later in this article. However, a reliable appraisal of effectiveness is needed to understand how well molecular shapes are being described and thus compared. Note that a small inaccuracy in the description of shape will lead to many similar compounds being undetected, given the very large size of interesting molecular databases. Without such appraisal, it is impossible to estimate what proportion of similar compounds remains undetected in the molecular database.

Most importantly, shape matching is a computationally demanding task,12 and therefore the increasing size of molecular databases poses a serious limitation to the use of these methods. Indeed, shape information is regarded as difficult to encode efficiently and use in database searching.6 By contrast, the importance of having more efficient methods has been vigorously highlighted (e.g. refs. 2, 6, 13). In addition to the relative inefficiency of current techniques, one is interested in using increasingly large molecular databases. Indeed, it is necessary to include more than one conformer per compound in the database, since flexible molecules can adopt different shapes,14 and thus the more of these conformations that are included in the database, the less likely it is to miss molecules with the desired pattern. Furthermore, larger databases can cover a wider region of the chemical space of drug-like molecules and thus increase the likelihood of finding innovative drug candidates.15 Consequently, it is of great importance to develop shape recognition methods, which screen the molecular database as fast as possible.

The main goal of this article is to present a new method for molecular shape comparison, which we call ultrafast shape recognition (USR). USR is devised to overcome the challenges posed by the growing size of molecular databases, which cannot be searched in a reasonable time by current methodologies. We will see that USR is extremely efficient, while accurately describing shape and thus being also highly effective. Such properties will be demonstrated through a rigorous validation procedure. Lastly, it is worth noting that the focus of the study is on how to effectively and efficiently search a very large database for similarly shaped compounds, and not on applying the presented methodology to identify compounds with similar biological activity and/or different chemical scaffolds from the query molecule, two important applications of molecular shape comparison that will be the object of future work.

The rest of the article is organised as follows. First, existing methods for molecular shape comparison and the procedures used to validate them are reviewed and discussed. Second, USR is presented. Thereafter, the experimental setup, including the validation procedure and the used molecular databases, is explained. The results of the performance study are also discussed, with the focus being on studying the efficiency and effectiveness of the method. Lastly, we present our conclusions and future prospects of this work.

Molecular Shape Comparison

Methods

Methods for molecular shape comparison can be roughly divided into two categories: those that rely on finding an optimal superposition of the molecules being compared (superposition methods) and those that, by contrast, are independent of molecular orientation and position (non-superposition methods). Almost always, efficiency and effectiveness are conflicting objectives. In this way, superposition methods are broadly regarded as particularly effective, but not so efficient. Our division into these categories is also motivated by the preference of having rotational and translational invariant methods, which are seen as a major advantage.11

In the non-superposition category, one group of methods measures molecular shape, based on atom triplet distances. Bemis and Kuntz16 devised a method, which considered each molecule as the set of its atom triplets. Molecular shape histograms were calculated with the perimeters of the triangle formed by each atom triplet and used to quantify the shape similarity of molecules. The method has major weaknesses such as not being able to compare molecules with different number of atoms directly and requiring storage in some cases larger than that of the heavy atoms coordinates themselves. Nilakantan et al.17 presented another molecular shape comparison method based on atom triplets. Each molecule is represented by a condensed triplet shape signature. Only those molecules with very similar signatures are compared in detail by generating again all their triplets. This increases the efficiency of the method at the risk of missing similar molecules because of the inaccuracies in the signature representation. Good et al.18 devised a series of molecular descriptors based on triangles of atom triplets. Such descriptors were encoded as bit strings and histograms, while including an extension to compare molecular surfaces. Drawbacks of these descriptors included modest discriminating power and requiring a large amount of disk space to store them. There have also been extensions of this class of descriptors to incorporate additional chemical information (e.g. ref. 19), which are therefore not intended to describe exclusively molecular shape and thus are outside the scope of this article. These techniques are fast (in the region of 500–2 000 comparisons per second19 on a 1995 PC), but they are known to be less effective than superposition methods,20 and thus they are normally used for database prescreening as suggested by their authors (i.e. quickly filtering molecules with very different shapes) instead of stand-alone molecular shape comparison.

In an innovative non-superposition technique, Shape Signatures, due to Zauhar et al.,6 each molecule is described by a histogram of the information derived from the simulation of a ray-trace reflecting within the molecular volume. This technique is rigorously shown to
be quite effective. Regarding its efficiency, it performs about 2,700 shape comparisons per second (exactly 370 μs per molecule) on single 1.5 GHz Pentium IV processor, once the shape signature of each molecule in the database has been calculated. The latter is a very expensive procedure, which takes about 1600 h for a database of just 113,331 molecules on a single 450 MHz Pentium III processor (in the article, this was done in 100 h using 16 of these processors). However, these signatures only need to be calculated once, as they only require a moderate storage space. Therefore, the signatures can be stored on hard disk and be used as many times as necessary without additional computational cost.

The other major family of shape comparison methods is based on the optimal superposition of the compared molecules. An early superposition method by Meyer and Richards,21 calculated the Carbo index22 on a rectilinear grid through a point-counting algorithm. Such calculation requires testing whether each grid point falls inside the van der Waals surfaces of the first, the second, or both superposed molecules. Hahn23 developed a much faster hybrid grid-based method, which used prescreens intended to filter out those molecules with a shape clearly dissimilar to that of the query. At a second stage, the similarity between the query and those molecules passing the prescreen is estimated by comparing their corresponding ratio of molecular surface interior grid points to the grid resolution. The efficiency of the method drops greatly for symmetrical query molecules, as the prescreens are less selective in that case and many more molecules pass to the expensive second stage. In addition, the method needs to define a sufficiently small grid spacing in order to be effective, but lowering this parameter results in an n³ increase of the number of grid point and thus to a large efficiency decrease. This could be an important problem with larger molecules than those used in the numerical tests, since they would require a much higher number of grid points to maintain the same accuracy, which would be much more expensive. The reported efficiency varies greatly depending on factors such as the symmetry of the query or the tolerance of the filters. However, the results for the only reported ranking of hits were obtained at about 1200 comparisons per second on a 1997 UNIX machine.

This strong trade-off between the efficiency and the effectiveness of grid-based shape similarity was considered to be an inherent drawback of this subclass of methods.20 In response, Good and Richards24 introduced the use of Gaussians in molecular shape comparison by describing the shape of each atom through a suitable electron density function and then fitting three Gaussian functions to each of them. This allowed the derivation of an analytical shape similarity index from the Carbo index,22 which ultimately led to an efficiency of two or three orders of magnitude higher than previous grid-based techniques.24

The last strand of superposition methods has also analytical expressions and compares the shape of molecules, according to their optimized volume overlap. Masek et al.25 presented a method to calculate the shape similarity of molecules and to align molecules such that shape similarity is maximized. A molecule is represented by a set of atom-centered intersecting spheres of different radii with the exposed surface of these spheres forming the van der Waals molecular shape, which in turn defines the boundary of the molecular volume. Thereafter, the overlapping volume of the compared molecules is calculated and used as a quantitative measure for shape comparison. As the calculated property depends on the relative position and orientation of the molecules being compared (this is common to all superposition methods), the volume overlap is maximized with respect to the relative molecular translation and rotation of the compared molecules. A local optimiser using the corresponding analytical first and second derivatives is applied from a number of starting points, so as to reduce the likelihood of getting trapped in a local maximum, which represents a particular suboptimal volume overlap.

Grant and Pickup26 argued that the hard-sphere representation had many technical difficulties (e.g. gradient discontinuities negatively affecting the search of local maxima) and was not a very accurate physical representation of a molecule. They defined a Gaussian density for each atom to replace that from the hard-sphere method, while considering the same expression of the molecular volume as a series of integration terms representing intersection volumes between the atoms in the molecule. The elegant resolution of these integrals, based on properties of Gaussians, led to an analytical expression for molecular volume. As the exact evaluation of this expression was found impractical, some reasonable approximations were made such as evaluating the molecular volume expansion up to the sixth term or introducing a control parameter, called Gaussian cutoff, in order to limit the number of atoms contributing to the volume intersection terms. Tests on macromolecules showed that the hard-sphere method was at least an order of magnitude slower than that provided by the Gaussian method, while the agreement between both methods in the volume calculation was generally within 0.1% (this figure does not include the additional error coming from superposition inaccuracies because both calculations were performed on exactly the same molecules with identical orientations and positions). In a follow-up paper,8 this Gaussian description was used to compare the shapes of molecules by optimizing their volume overlap using analytical derivatives with respect to rigid-body rotations and translations. Such optimization can be difficult as it involves six parameters and is likely to contain several separated optima. To alleviate this major technical difficulty, four well-spaced starting points in rotational-translational space are chosen and a local optimizer carried out from each of them. In experiments with a small set of ligands, it was found that the optimization usually converged to the correct orientation, although in some cases this did not occur and results were not so accurate. A significant improvement in efficiency (10–30 times faster) was found with respect to analogous hard-sphere shape comparisons.

A recent implementation of the Gaussian molecular volume comparison method is ROCS (rapid overlay of chemical structures).27 Although ROCS is based on the well founded theory introduced in refs. 8, 26, it relies on a number of approximations to achieve much higher efficiency. In particular, ROCS calculates the volume overlap of a pair of typical drug-sized molecules in about a millisecond28 when compared with timings between 0.6 and 48.2 s reported in the original method,8 using computers available 8 years before (i.e. a two orders of magnitude improvement once corrected with Moore’s Law). Nevertheless, it has been claimed28 that ROCS describes molecular shape in a rigorous and accurate way. An obvious example showing that this might not be the case is that ROCS gives, unlike refs. 8, 26, the same radius value to all heavy atoms in the molecule.28 Furthermore, by only keeping the zero order Gaussians,28 ROCS just calculates the first term of the molecular volume expansion, which simply corresponds to the added volume of all
atomic Gaussians without correcting for any overlap between atoms. By contrast, the original method\textsuperscript{26} required calculating the volume up to the sixth term to provide accurate results (i.e. typically less than 0.1% with respect to the molecular volume calculated with the hard-sphere method). The only reported values of these six terms (ref. 26, p.3509), which are 46,304, 25,481, 6173, 5431,4 and 0.2 Å\textsuperscript{3} (this totals 26,467 Å\textsuperscript{3} because the series alternates sign), were presented as typical values for all volume calculations made. One can see from these data that the ROCS approximation of calculating molecular volume only up to first term introduces an error of 75% with respect to the original method. It would be interesting to see what exactly the magnitudes of these errors are for drug-sized molecules.

More importantly, ROCS does not guarantee that the best superposition between the compared molecules will be found. This major technical problem can be alleviated by increasing the number of starting points at the cost of further optimizations (one per starting point) and thus lowering ROCS efficiency (something that is particularly needed for molecules of high symmetry\textsuperscript{28}). In addition, reduced effectiveness due to suboptimal molecular overlap, i.e. obtaining a lower similarity score than it should, is very hard to detect because only the top ranked molecules are in practice visible, i.e. appear above the selected threshold. Those molecules, which have a sufficiently similar shape to that of the query, but obtain a suboptimal molecular overlap because of superposition errors, will unnoticeably drop below the threshold and be lost among possibly millions of other rejected molecules. Despite this loss of valuable information, the method might still be perceived as effective if the subset above the threshold contains many similarly shaped conformers. Furthermore, whatever is the exact percentage of error introduced by this issue, it is undoubtedly going to become more important with the increasing size of molecular databases, as more and more similarly shaped compounds will remain undetected.

In the light of these facts, it is clear that none of the current shape comparison methods is completely effective.

Validation

In this section, we briefly discuss the problem of validating methods for molecular shape comparison. We focus on how to measure effectiveness, as tests for determining the efficiency (speed of operation) of a method are trivial.

The origin of the problem with measuring the effectiveness of a method is that there is no available ground truth with which to compare. This ground truth could take the form of sets of diverse molecules perfectly ordered by shape similarity or a method capable of doing so (i.e. completely effective). Unfortunately, as we have seen in the previous section, no method for molecular shape comparison has been shown to be sufficiently effective to play this role.

Different studies have addressed this difficulty in various ways. Early studies presented evidence in support of the effectiveness of the introduced method by discussing the consistency of a few (10–100 s) representative examples of pairwise molecular shape comparisons (e.g. refs. 16, 21, 25). This proof of concept approach is insufficient nowadays as the large size of molecular databases requires not only detecting some similarly shaped molecules, but providing an accurate ranking in order to distinguish all the most similar molecules in the database from the rest. Otherwise, given the very small ratio of similar versus dissimilar molecules in a diverse database, there is the danger of obtaining too many dissimilar molecules, but with wrongly assigned high scores, at the subset of top ranked molecules.

Another approach to measure effectiveness has been to compare the new method with a previous technique, so as to establish the degree of agreement between the corresponding similarity score or property (e.g. refs. 18, 24). The main drawback of this approach is that a high correlation between the scores or rankings provided by the compared methods does not necessarily mean that shape is being accurately described, but that is described in a similar way. For instance, one could compare the similarity scores obtained with two different superposition methods and obtain a high correlation among them. However, if both methods had difficulties finding the optimal superposition of say symmetrical molecules, then it would be possible for them to be highly correlated while assigning a wrong low score to all those molecules and thus missing them.

Some superposition methods have evaluated their effectiveness by analyzing the quality of the achieved alignment between the compared molecules while implicitly neglecting the remaining sources of error. This is the case of ref. 8, where the superposition of two identical molecules, initially separated by 2 Å and rotated 30 degrees with respect to the other, was studied for four different molecules. Although the optimal alignment was obtained in all four cases, there exist several reasons why this is not a sufficiently strong validation. First, as the optimal relative orientation and position of two identical molecules is always known, one can automatically check the quality of the obtained superposition. Therefore, a large and diverse molecular database, instead of just four molecules, should be used to provide a reasonable estimate of the superposition error introduced by the used optimization technique. Second, molecules in a database may be in principle registered with any orientation and position. Consequently, tests should use initial random orientations and positions of the compared molecules rather than an initialization that is very close to the optimal superposition, as in ref. 8. This is to avoid having the starting point in rotational–translational space within the basin of attraction of the global optimum representing the best volume overlap, the scenario where a local optimiser, such as the gradient technique used in ref. 8, performs best. The latter would lead to an unrealistically high measure of effectiveness. Finally, analogous tests to measure the quality of the superposition when different molecules are compared would be also required. Such tests are more difficult to design because the optimal superposition of the molecules being compared is not easily known a priori. For example, the approach taken in ref. 8 attempts the alignment of different ligands binding to the same receptor and compare the results with the orientations of the respective bound conformations. One drawback of this procedure is that it cannot be carried out in large numbers (in refs. 8, 15 ligands were used). A more reliable approach would be to calculate the optimal superposition of each pair of compared molecules with a suitable global optimization technique (such as the Real-coded genetic algorithm described in ref. 29) and then use it to estimate the superposition error introduced by the faster, but not so accurate gradient-based superposition embedded in ref. 8. These sound validation practices, which have been neglected by modern developments like ROCS, are important nowadays because superposition errors are not detectable in large molecular databases, as discussed in the previous subsection.
To date, the most reliable way to assess the effectiveness of a method is to analyze the most similar molecules to a set of queries ordered by their corresponding similarity score. In this way, one can estimate the accuracy of the method in describing shape by assessing the quality of the top part of the ranking, as a very effective method will have by definition a very small percentage of wrongly top ranked molecules. Such errors will be more noticeable for larger databases, since the same percentage of errors will result in more errors being made. There are a few methods that have followed this validation procedure to different extents. Nilakantan et al.\textsuperscript{17} presented the 10 molecules with the highest shape similarity score to a query out of 22,498 in the database. The ranking (see Fig. 4 in ref. 17) presents some evident inconsistencies such as that of the molecule with the second highest score (BIWSI), which is very dissimilar from the rest of top ranked molecules, or that of the tenth molecule with the second highest score (BNOVC), which should be higher in the ranking dissimilar from the rest of top ranked molecules, and thus the possible loss of similar molecules. However, only Zauhar et al.\textsuperscript{6} has presented score rankings from several queries using a molecular database of realistic size and diversity, therefore providing the strongest validation of effectiveness for a molecular shape comparison method in the literature. In particular, a reasonably large database (113, 331 compounds, with one conformer per compound) was used, showing the ranking of the five molecules with the highest shape similarity score (see the results for ‘1D Shape Signature’ in ref. 6) in decreasing score order for six query molecules of diverse chemical structure. It can be observed (from the figure in Table 3b in ref. 6) that the provided ranking is largely consistent with human-perceived shape similarity, despite not ranking first the query molecule in most cases.

An Ultrafast Method for Shape Recognition

In this section, a new method for molecular shape comparison, called USR, is presented. USR is based on the observation that the shape of a molecule is uniquely determined by the relative position of its atoms. In this way, the molecule is regarded as a system of bound particles (the atoms), instead of its more conventional treatment as a solid body. The relative position of the atoms in the molecule is in turn completely determined by the set of all interatomic distances. This is a convenient representation, which directly eliminates any need for alignment or translation, as this set of distances is independent of molecular orientation or position.

Nevertheless, the set of all interatomic distances contains more information that is needed to describe accurately the shape of the molecule. This is because the values of these distances are heavily constrained by the forces that hold the atoms together and thus using less information would still provide us with the discriminative power necessary to distinguish between molecules. Indeed, only a limited number of distance values are possible between two bound atoms, which depend on the type of atoms forming the bond. In the case of distances between atoms not linked by a bond, the more flexible is the molecule, the more values will become possible, although still there will be strong restrictions in these distance values because of intermolecular repulsion and attraction forces as well as the bonding arrangement of the molecule.

In this article, we will demonstrate that by selecting a suitable subset of inter-atomic distances, it will still be possible to describe accurately shape while significantly reducing the associated computational cost. In particular, this subset could be chosen as the set of all atomic distances from a reduced number of strategic reference locations, which are uniquely defined in every molecule. As choosing very close molecular locations would result in very similar sets of distances and thus essentially the same information, these locations should be selected to be as separated among them as possible so as to provide the most discriminating power.

In this work, the set of all atomic distances from four molecular locations are considered: the molecular centroid (ctd), the closest atom to ctd (ctf), the farthest atom to ctd (fct), and the farthest atom to ctd (fctf). 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 also considered, since this location is far from ctf for some molecules while having an inexpensive calculation. An example illustrating all these molecular shape representations is in Figure 1.

At this stage, each molecule is described by a number of features, i.e. the four sets of atomic distances, which is proportional to the number of atoms in the molecule. This raises the obvious question of how to compare molecules with different number of atoms. Each set of distances can be regarded as a distribution, which is usually characterized by a histogram. However, histogram calculation has a number of well-known drawbacks, specially for very large databases, such as the difficulty of selecting a bin size suitable for all molecules in the database along with the requirement of relatively large storage and computing power.\textsuperscript{16-18} These difficulties are circumvented by calculating instead the first moments of each distribution, which provides an estimate of the size of the molecule. The second descriptor ($\mu_k^{(N)}$) is the second moment of the same distribution, which is the variance of these atomic distances from the centroid and hence it is related to how compact the molecule is. The third descriptor ($\mu_k^{(N)}$) is the third moment or skewness of the same distribution, which estimates its asymmetry and thus whether the atoms group near or far from the mean atomic position from the centroid. To calculate the remaining nine descriptors, the process is repeated for each of the three remaining distributions (i.e. $d_{k}^{(N)}$, $d_{k}^{(N)}$, and $d_{k}^{(N)}$ for each of the three remaining distributions (i.e. $d_{k}^{(N)}$, $d_{k}^{(N)}$, and $d_{k}^{(N)}$).
where the superscript indicates the location from where the atomic distances are calculated).

Once the shape descriptors for each molecule in the database are available, a normalized score function is needed to quantify the degree of similarity between molecules based on these descriptors. In principle, any monotonic inverse function of the distance between molecules in descriptor space can be used. In this work, we use the inverse of the translated and scaled Manhattan distance between both vectors of shape descriptors, where a value of 1 corresponds to maximum similarity and 0 to minimum similarity. An illustrative example of the comparison operation is in Figure 3, including the expression of the adopted score function and the descriptor values for each of the compared molecules.

The accuracy with which USR describes molecular shape and thus its effectiveness as a molecular shape comparison method will be thoroughly tested in the next section. Nevertheless, we advance now a couple of comparisons to highlight some interesting features of the method. The first example in Figure 3 shows that molecules with similar shape, but different number of atoms, can be found with USR. The top molecule has 33 heavy atoms compared with the 26 heavy atoms, which forms the molecule at the bottom. Despite being calculated with a significantly different number of atomic distances, both vectors of shape descriptors are quite similar because the relative position of the atoms in both molecules are quite similar as well. The second example, in Figure 4, gives an idea of the high discriminating power provided by USR. Both molecules in Figure 4...
Figure 3. Example of molecular shape comparison with USR. USR assigns a high score, $S = 0.812$, to these similarly shaped molecules, despite having a significantly different number of atoms (the top molecule has 33 heavy atoms in front of the 26 heavy atoms which forms the molecule at the bottom). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 4. Example of molecular shape comparison with USR. Both compared molecules differ only in one atom, which in turn introduces a slight difference in the corresponding distance distributions and ultimately in the calculated shape descriptors. Each descriptor has a very similar, but still different value for each molecule, although this is not always appreciable from the values reported due to the truncation at the second decimal. This shows the high discriminating power offered by USR. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
the status of shape as an intrinsic geometrical property of the molecule. This is the key factor to speed up the screening process as cross-calculations between the query and the considered molecule, which typically arise in superposition methods, are avoided. As a consequence, once these shape vectors have been calculated for the whole database, comparing shapes of two molecules consists simply in evaluating the corresponding similarity score, an operation that has the tiny computational cost of calculating the inverse of the distance between two vectors of shape descriptors.

Experimental Results and Discussion

This section presents and discuss the experiments carried out to validate the presented method. There are two types of validation experiments: the first is intended to study the effectiveness of USR as a molecular shape comparison method and the second for testing its efficiency. These experiments will be performed on two different databases of 3D chemical structures. The first database, henceforth refer to as the Vendor Database, contains 2,433,493 commercially available compounds in 3D MDL SD format and it was provided by Inhibox Ltd.. It has a single conformer per compound, with each of them having at least 10 heavy atoms. This is a very large number of diverse compounds in comparison with previously used databases for the validation of molecular shape comparison methods, as previously discussed in this article. Therefore, this database will be useful to try to find particular molecules for which the USR encoding of shape is not sufficiently accurate. Such possibility would manifest as top ranked molecules, which are not visually similar. If there were significant inaccuracies in the USR description of shape, one would expect to detect them after screening almost 2.5 million potentially problematic molecules.

The second database will provide an independent benchmark to check whether the performance of USR depends significantly on the characteristics of the used database or not. This database comes from DrugBank, a publicly available resource from where 3764 small molecule drugs in SMILES format were downloaded. Such set is formed by 708 FDA-approved drugs plus 3056 experimental drugs. More detailed information about DrugBank can be found in ref. 33. Standard preprocessing was made on the original data sets. This ultimately led to a database, which we call simply DrugBank-3D Database, of 666,892 conformers in 3D MDL SD format, i.e. an average of 200 conformations per compound. The DrugBank-3D Database will permit us to analyze the effect of using multiple conformers to represent each flexible molecule on the USR performance, by comparing with the results obtained on the single-conformational database.

Our first experiment is intended to evaluate the efficacy of the proposed descriptors for accurately encoding shape. Such validation of the USR effectiveness will be carried out by following the same approach as Zauhar et al. which was explained in the validation subsection of this study. However, we further increase the strength of our validation by using test databases with a size 6 and 21 times larger than the largest database used in ref. 6. We start by selecting five query molecules from the vendor database. Thereafter, the shape similarity score between the query and every molecule in the database is calculated so as to identify the four molecules with the highest score. This process is repeated for each query molecule. Results are reported in Figure 5. Note that, given the large database size (2,433,493 molecules), a small inaccuracy in the shape description would result in dissimilar molecules within the top ranked subset, which was not observed. These queries were selected because they constitute a diverse sample of the chemical space in terms of number of atoms and type of shape, but we have observed results of similar quality in every additional query we made. These results show that USR is able to identify shapes that closely resemble that of the query, which is also in the database and it is the highest ranked molecule in all cases. By contrast, the latter is not achieved by many other methods such as in refs. 6, 17. In addition, the experiment suggests that the method is particularly good at finding different chemical scaffolds, as it can be observed from the fourth query in Figure 5, which has been pointed out as a very valuable capability.

The second validation test uses the multiconformational database. As our goal is to find different compounds with a shape similar to that of the query, we will direct our attention to the four compounds with the highest score, instead of the four conformers with the highest score. This issue is illustrated in the example in Figure 6. The four conformers with the highest similarity score (top row) to the drug conformer APRD00001-1 from the DrugBank-3D database correspond to four conformations of the same compound, including the query. However, the reason why a database is populated with multiple conformers of each flexible compound is to reduce the possibility of missing compounds with similar shape to the query. Therefore, we are only interested in the most similar conformation of each compound, as a way to highlight the different compounds with similar shape to the query. The bottom row of Figure 6 shows the results of the query in these terms. As expected, discarding additional top ranked conformations of each compound results in less similar conformers in the top four compounds subset. In return, focusing on the top ranked compounds highlights important information such as the fact that the top three compounds have the same biological activity (all are inhibitors of the Histamine H1 Receptor) or the presence of similarly shaped compounds with significantly different chemical scaffolds in the fourth most similar compound. These are important applications of molecular shape comparison, which we plan to study with USR in future work.

Consequently, the effectiveness of USR is also tested on the DrugBank-3D database with the difference that now, after calculating the similarity score of each molecule in the database, we will form the ranking with only the most similar conformer of each compound. Again, we select five queries to have different number of atoms and shapes. For each query, the top four compounds are shown in Figure 7. In all cases, the query molecule is the highest ranked molecule with maximum similarity score 1. As with the vendor database, no dissimilar molecules within the top ranked subset are found after screening 666,892 potentially problematic conformers.
Figure 5. Screened molecules with the highest USR score for five different queries from a single-conformer database. For each row, the four highest ranked molecules out of the 2,433,493 compounds constituting the vendor database are presented. The query molecule is the highest ranked molecule in all cases and thus appears always as the first on the left in each row. This figure shows that the method succeeds in finding very similarly shaped compounds for diverse, in terms of number of atoms and types of shape, query molecules. In addition, USR is good at finding different chemical scaffolds, as it can be observed from the fourth query. The first query molecule has 17 atoms (scores are: $S_1 = 1.000, S_2 = 0.976, S_3 = 0.934, \text{ and } S_4 = 0.908$). The second query molecule has 25 atoms ($S_1 = 1.000, S_2 = 0.912, S_3 = 0.909, \text{ and } S_4 = 0.892$). The third query molecule has 30 atoms ($S_1 = 1.000, S_2 = 0.966, S_3 = 0.960, \text{ and } S_4 = 0.957$). The fourth query molecule has 33 atoms ($S_1 = 1.000, S_2 = 0.812, S_3 = 0.788, \text{ and } S_4 = 0.785$). The fifth query molecule has 38 atoms ($S_1 = 1.000, S_2 = 0.971, S_3 = 0.890, \text{ and } S_4 = 0.884$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
Figure 6. Screened conformers with the highest USR score vs. compounds with the highest USR score for five different queries from the DrugBank-3D Database. For the top row, the four highest ranked conformers out of the 666,892 constituting the DrugBank-3D database are presented. The query is the highest ranked conformer and thus appears always as the first on the left in each row. All four are conformations of the same compound (APRD00001). For the bottom row, the most similar conformer of each top ranked compound are shown instead. This query molecule has 19 atoms. The scores for the conformers at the top row are $S_1 = 1.000$, $S_2 = 0.979$, $S_3 = 0.955$, and $S_4 = 0.952$, while those for the compounds at the bottom row are $S_1 = 1.000$, $S_2 = 0.914$, $S_3 = 0.914$, and $S_4 = 0.866$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

In each query. It is worth noting that the multiconformer database contains remarkably similar compounds, even though its size was four times smaller and has 730 times fewer compounds than the single-conformer database. There are cases where this is less true such as the 5th query in Figure 7. This is a reminder of the importance of the molecular database used: if there are not very similar molecules in the database, obviously no method will find them. On the other hand, highly similar compounds were also consistently found in the single-conformer database, as those shown in Figure 5. This corroborates the common view that both factors, multiple conformers and high diversity, are beneficial for finding similarly shaped compounds in molecular databases. Nevertheless, the latter means that methods sufficiently efficient to search such larger databases are required, if we are to unleash the full potential of this technique.

The other important aspect to study is thus the efficiency of the method. To provide a baseline for efficiency comparison, we will consider the most efficient methods in the literature, Shape Signatures\(^6\) and ROCS.\(^{28}\) Direct comparisons between molecular shape comparisons are difficult to achieve mainly because of the commercial interests involved.\(^2\) Therefore, we will compare to the efficiency of these methods as reported in the literature, while ensuring a fair comparison by using a computer for the numerical experiments with similar or worse calculation power than the used in those studies. In this way, ROCS reports\(^5,28\) a comparison rate approaching 1000 molecules per second on a modern Intel/AMD processor. As pointed out in ref. 5, this is orders of magnitude faster than most 3D similarity search methods, although it is highly unclear from these studies how effective is ROCS when operating at this best comparison rate. In the case of shape signatures, the comparison rate is 2703 molecules per second (exactly $370 \times 10^{-6}$ s per shape comparison) on a 1.5 GHz Pentium IV processor. However, this method requires a very expensive preprocessing before being able to operate at this comparison rate. For example, if we had to use this method to make all the effectiveness tests in this article, the preprocessing of both test databases would have taken almost 4 months running on the 16 Pentium III 450 MHz processors used for this purpose in ref. 6.

We estimated the efficiency of USR as follows. We started by recording the wall clock time required to search the vendor database for similarly shaped molecules with respect to a query molecule. This was repeated for 10 randomly chosen queries to provide an average elapsed time, which in turn was divided by the number of comparisons made (2,433,493) to give the average comparison rate of USR. As a result, USR performs an average of $1.83 \times 10^{-4}$ s molecule when running on a modestly powerful PC (AMD Athlon XP 1800+ CPU at 1.5 GHz). There is a preprocessing coming from the calculation of the vector of shape descriptors, which is an average of $2.23 \times 10^{-4}$ s molecule when performed on the same computer. However, this preprocessing only needs to be carried out once since the descriptors can be stored and used as many times as needed owing to its small storage (12 floating-point numbers per molecule). The rate calculated on the DrugBank-3D database is essentially the same and thus it is omitted. Figure 8

\(^2\)For instance, we applied for an academic license of ROCS in order to use it to carry out the same experiments reported here for USR, which would have provided us with a direct performance comparison between both methods. However, our application was rejected on the grounds of compromising OpenEye’s commercial interests.
Figure 7. Screened compounds with the highest USR score for five different queries. For each row, the four highest ranked compounds, represented by their most similar conformer, out of the 666,892 conformers constituting the DrugBank-3D database are presented. The query molecule is the highest ranked molecule in all cases and thus appears always as the first on the left in each row. This figure shows that the method succeeds in finding very similarly shaped compounds for diverse, in terms of number of atoms and types of shape, query molecules. In addition, USR is good at finding different chemical scaffolds, as it can be observed from both the 2nd and 5th queries. The first query molecule has 15 atoms (scores are: $S_1 = 1.000$, $S_2 = 0.811$, $S_3 = 0.794$, and $S_4 = 0.760$). The second query molecule has 19 atoms ($S_1 = 1.000$, $S_2 = 0.914$, $S_3 = 0.914$, and $S_4 = 0.866$). The third query molecule has 22 atoms ($S_1 = 1.000$, $S_2 = 0.984$, $S_3 = 0.868$, and $S_4 = 0.866$). The fourth query molecule has 26 atoms ($S_1 = 1.000$, $S_2 = 0.916$, $S_3 = 0.882$, and $S_4 = 0.843$). The fifth query molecule has 33 atoms ($S_1 = 1.000$, $S_2 = 0.822$, $S_3 = 0.805$, and $S_4 = 0.798$). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
A new method for molecular shape comparison, USR, has been presented. USR accurately describes the shape of a molecule through its set of interatomic distances, which is ultimately encoded as a vector of geometrical descriptors. The method does not require previous alignment of the molecules being compared. The comparison of molecular shape via these USR descriptors has been rigorously shown to be highly effective while being extremely efficient, as it operates more than two thousand times faster than the previously fastest method in the literature.

Such significant improvement in efficiency will be key for a number of applications of molecular shape comparison. First, USR can identify the most similarly shaped molecules within a multi-billion molecular database in a matter of minutes using a single processor. The use of much larger databases would greatly improve the performance of molecular shape comparison methods in finding biologically active compounds with innovative chemical structure. Specifically, this maybe combined with complementary similarity search methods, such as electrostatic similarity, which cannot directly cope with these large numbers of molecules. On the other hand, USR’s efficiency would be very useful when multiple database queries need to be made simultaneously on a single computer, as it would be the case of an online molecular database being searched concurrently by several users. Lastly, clustering large molecular databases in terms of shape is a valuable technique to generate smaller and diverse databases. However, most clustering algorithms require the evaluation of the whole similarity matrix, which implies calculating the similarity of every possible pair of molecules in the database and thus an impossibility to use large databases. The required $N^2$ similarity calculations could be potentially performed more than four million times faster by using USR as the similarity measure. We intend to study all these applications in future work.

Acknowledgements

The authors thank Paul Finn from Inhibox Ltd. for helpful discussions and feedback. We thank Inhibox Ltd., in particular Dan Butler and Daniel Robinson, for preparing the molecular databases used in the experiments. Early feedback from Lydia Kavraki at Rice University is also thankfully acknowledged.

References