PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2007) **463**, 1307–1321
doi:10.1098/rspa.2007.1823
Published online 27 February 2007

Ultrafast shape recognition for similarity search in molecular databases

By Pedro J. Ballester* and W. Graham Richards

Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, UK

Molecular databases are routinely screened for compounds that most closely resemble a molecule of known biological activity to provide novel drug leads. It is widely believed that three-dimensional molecular shape is the most discriminating pattern for biological activity as it is directly related to the steep repulsive part of the interaction potential between the drug-like molecule and its macromolecular target. However, efficient comparison of molecular shape is currently a challenge. Here, we show that a new approach based on moments of distance distributions is able to recognize molecular shape at least three orders of magnitude faster than current methodologies. Such an ultrafast method permits the identification of similarly shaped compounds within the largest molecular databases. In addition, the problematic requirement of aligning molecules for comparison is circumvented, as the proposed distributions are independent of molecular orientation. Our methodology could be also adapted to tackle similar hard problems in other fields, such as designing content-based Internet search engines for three-dimensional geometrical objects or performing fast similarity comparisons between proteins. From a broader perspective, we anticipate that ultrafast pattern recognition will soon become not only useful, but also essential to address the data explosion currently experienced in most scientific disciplines.

Keywords: molecular shape comparison; similarity search; pattern recognition; data explosion; virtual screening

1. Introduction

Virtual screening is a key technique in computational drug discovery, aimed at identifying those drug-like molecules that are likely to have beneficial biological properties. It is an obvious way to reduce expensive biological tests and tackle the high failure rate currently faced by the pharmaceutical industry (Böhm *et al.* 2004; Kola & Hazuda 2005). In molecular docking, for instance, the process of docking the screened molecule to a macromolecular biological target (almost always a protein) is simulated to provide an estimate of its binding energy and thus its likelihood of being bioactive. These techniques have spurred the generation of massive databases of drug-like molecules. This is the case in our widely publicized screensaver project (Richards 2002) in which a database of 3.5 billion compounds

* Author for correspondence (pedro.ballester@chem.ox.ac.uk).

Received 11 December 2006
Accepted 18 January 2007

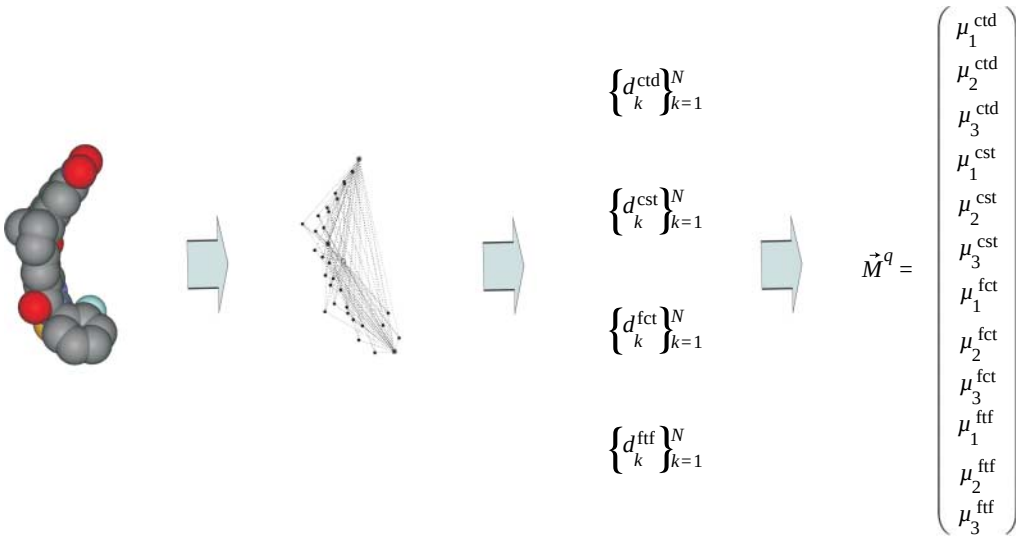

Figure 1. USR encoding. The shape of the molecule is characterized by the distributions of atomic distances to four strategic reference locations. In turn, each of these distributions is described through its first three moments. In this way, each molecule has associated a vector of 12 shape descriptors.

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 virtual screening technique consists of searching a molecular database for compounds that most closely resemble a given query molecule. This chemical template can be (i) a known product or an inhibitor of a target protein, (ii) a natural product, or (iii) even a patented compound. The underlying assumption (Baringhaus & Hessler 2004; Willett 2005*a*,*b*) is that molecules similar to the active query molecule are likely to share similar properties. This similarity can be in terms of molecular shape or a range of molecular descriptors, most of which are in one way or another related to the geometry of the molecule. Here, we focus on three-dimensional shape since it is the most discriminating molecular pattern, being related to the steep repulsive part of the interaction potential between the drug-like molecule and its macromolecular target. A number of previous studies (Kotani & Higashiura 2002; Zauhar *et al.* 2003; Rush *et al.* 2005; Schnecke & Boström 2006) have also highlighted the importance of molecular shape as an indicator of biological activity. An additional advantage of searching a database for molecules with similar shape is that no specification of chemical structure (i.e. types of atoms or their bond arrangements) is made and therefore non-intuitive novel drug candidates can be found (Böhm *et al.* 2004; Rush *et al.* 2005). This alternative application of molecular shape comparison is widely known as chemical scaffold hopping (Böhm *et al.* 2004).

Unfortunately, there are several challenges for methods using molecular shape as the pattern to recognize. Indeed, shape information is regarded as difficult to encode efficiently and use in database searching (Zauhar *et al.* 2003). The comparison of two molecules in terms of shape is often carried out by superposing

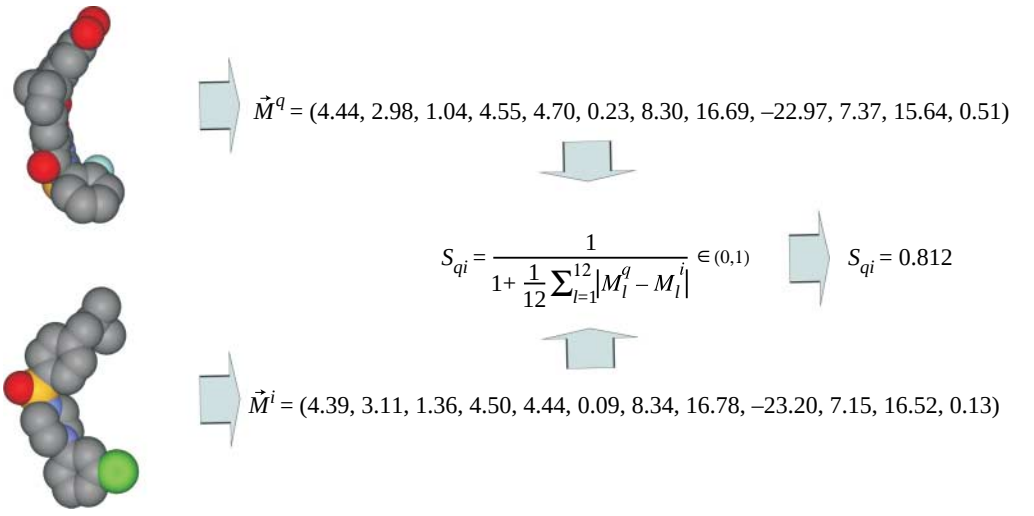

Figure 2. USR comparison. In order to establish the degree of similarity between the shapes of two molecules, the Manhattan distance between the corresponding vectors of shape descriptors is calculated. Thereafter, this dissimilarity is monotonically inverted so as to define a normalized similarity score, where maximum shape similarity is represented by score 1 and the minimum similarity by score 0. USR uses this score to determine which are the most similarly shaped molecules in a database to a given query molecule.

them in order to quantify their overlap (Hahn 1997; Rush *et al.* 2005). These superposition methods require previous alignment of the molecules, which is an additional source of difficulty (Bender & Glen 2004; Jenkins *et al.* 2004) that may lead to suboptimal molecular overlapping and thus to errors in the similarity score. Such errors would result in a decrease in effectiveness (i.e. the extent to which the method assigns a higher similarity score the more similar are the molecules being compared). Most importantly, the increasing size of molecular databases poses a serious limitation to the use of shape recognition methods. Flexible molecules can adopt different shapes (Hahn 1997; Willett 2005*a*) 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. 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. Consequently, it is of great importance to develop shape recognition methods which screen the molecular database as fast as possible.

2. Efficiency in molecular shape comparison

Superposition methods are a popular family of molecular shape comparison methods. A widely used commercially available method in this category is ROCS (Rush *et al.* 2005), which calculates a similarity score from the volume overlap of the molecules being compared. The required alignment is carried out through essentially a local optimization process, where each of the iterations involves the calculation of the volume overlap for the currently tested relative orientation and

position of the molecules. ROCS reports (Rush *et al.* 2005) a comparison rate approaching 1000 molecules per second on a modern Intel/AMD processor (the paper was published in 2005). Such a rate was claimed (Rush *et al.* 2005) to be orders of magnitude faster than other three-dimensional similarity methods, although it is highly unclear from these studies how effective is ROCS when operating at this comparison rate.

Another category of shape comparison methods uses geometrical descriptors to encode the shape of molecule, with the similarity score between molecules calculated by comparing the corresponding descriptors. One group of these descriptor-based shape methods is based on atom triplet distances. Bemis & Kuntz (1992) 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. This method has a processing rate between 13.2 and 31.9 molecules per second on a 1992 UNIX workstation. Nilakantan *et al.* (1993) 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 owing to the inaccuracies in the signature representation. A database with 225 000 compounds was searched in 2–3 h of CPU time (the wall clock time was not stated), which represents a comparison rate between 20.8 and 31.3 molecules per second on the used 1993 VAX computer. Good *et al.* (1995) 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. These methods performed between 500 and 2000 comparisons per second (Good *et al.* 1995) on a 1995 PC. Despite their high efficiency, these descriptor-based shape methods are known to be less effective than superposition methods (Good & Richards 1998) and thus they are normally used for database pre-screening (i.e. quickly filtering molecules with very different shapes), as suggested by Good *et al.* (1995), instead of stand-alone molecular shape comparison. There have also been extensions of this class of descriptors to incorporate additional pharmacophore-like information (Mason *et al.* 1999) while retaining most of this efficiency, which are therefore not describing exclusively molecular shape and thus are outside the scope of this paper.

In an innovative descriptor-based technique, Shape Signatures, due to Zauhar *et al.* (2003), each molecule is described by a histogram of the information derived from the simulation of a ray-trace reflecting within the molecular volume. The high effectiveness of the technique is demonstrated through the ranking of the five molecules with the highest shape similarity score (see the results for '1D Shape Signature' in table 3*b* of Zauhar *et al.* 2003) in decreasing score order for six query molecules of diverse chemical structure. It can be observed that the provided ranking is largely consistent with human-perceived shape similarity, despite not ranking first the query molecule in most cases. Regarding its efficiency, Shape Signatures performs about 2700 shape

comparisons per second (exactly 370 μs per molecule) on a single 1.5 GHz Pentium IV processor, once the shape signature of each molecule in the database has been calculated.

It is very complicated to accurately contrast the performance of all these techniques when a new method is presented. Ideally, we would perform exactly the same effectiveness and efficiency tests for each of the methods being compared. In practice, this is often not possible due to a number of constraints such as not having a publicly available version of the method (e.g. Bemis & Kuntz 1992; Nilakantan *et al.* 1993; Good *et al.* 1995; Hahn 1997; Zauhar *et al.* 2003) or having very restrictive software licence agreements.[1] Although contrasting the effectiveness of a new method with respect to these techniques is not viable under these circumstances, an approximate comparison of their efficiency would still be possible if a computer of similar power to that used in the study was available. The latter is more likely to happen with methods evaluated in recent studies rather than those in older studies, since the used computers are now obsolete and thus mostly unavailable.

Fortunately, we recently had access to a commercially available technique included in the Molecular Operating Environment (MOE 2006) software suite. The method is called EigenSpectrum Shape Fingerprints (ESshape3D) and 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. Next, this spectrum is encoded as a fingerprint and the similarity score calculated as the inverse of the distance between the corresponding fingerprints. This descriptor-based shape comparison method would permit us to directly contrast both the effectiveness and the efficiency of the method presented in §3.

3. Ultrafast shape recognition

Here we propose a new method for ultrafast shape recognition (USR). The approach adopted considers the molecule as a system of bound particles (the atoms), instead of a solid body. USR is based on the observation that the shape of a molecule is uniquely determined by the relative position of its atoms. Thereafter, the molecular shape is characterized by a set of one-dimensional distributions which retains three-dimensional shape information, since such a dimensionality reduction is expected to improve greatly the efficiency of the method. In this work, we use the distributions of all the atomic distances to four different reference locations: the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom to ctd (fct) and the farthest atom to fct (ftf). This directly eliminates any need for alignment or translation, as these distributions are completely independent of molecular orientation or position. Another advantage of the proposed method is that, unlike superposition methods (see for instance ROCS; Rush *et al.* 2005), the shape information for each molecule is independently

[1] We applied for an academic licence 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.

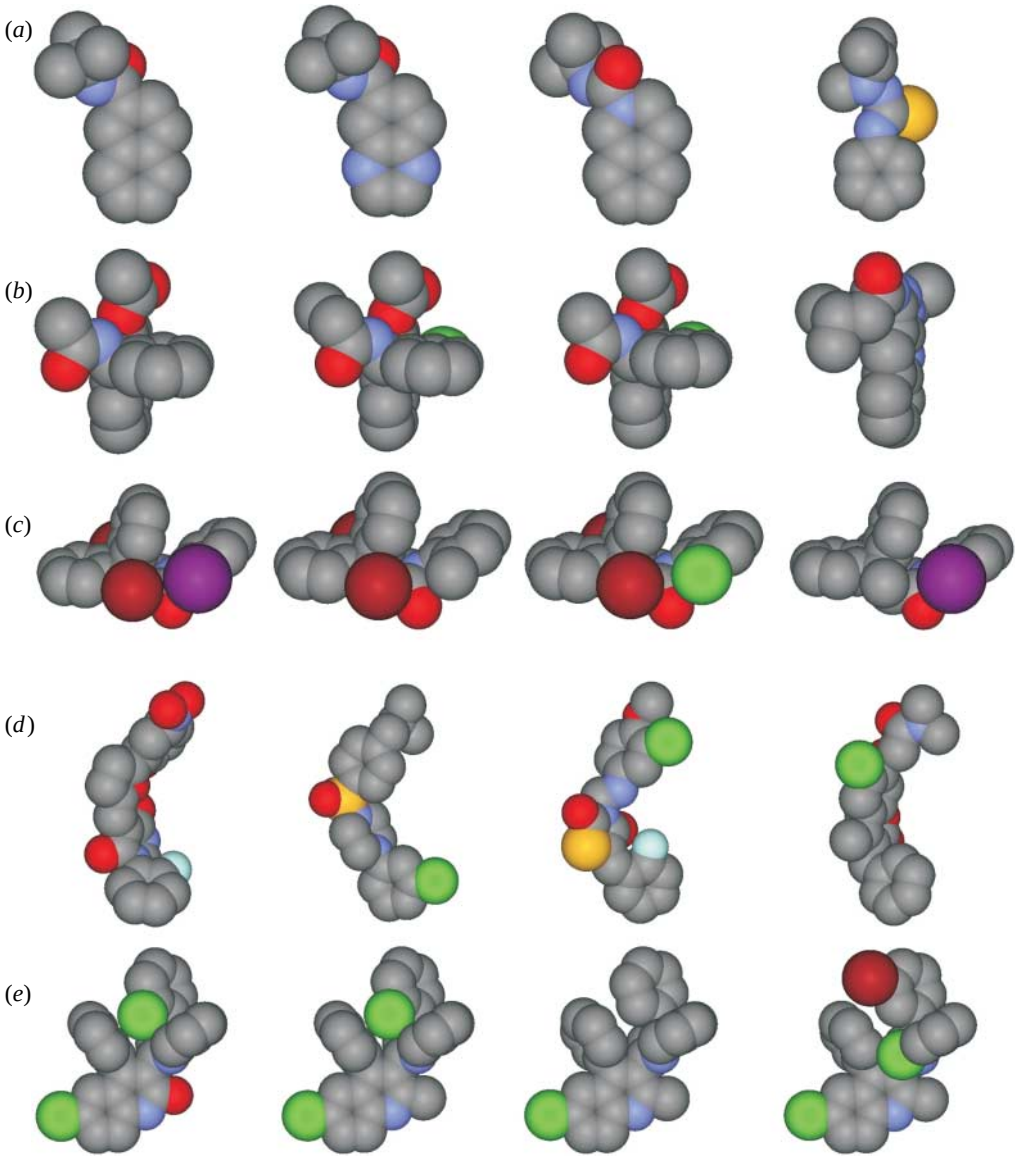

Figure 3. Screened molecules with the highest USR score for five different queries (i.e. each row (*a*–*e*) corresponds to the top four hits for a given query in decreasing similarity score order). For each query, the four highest ranked molecules out of 2 433 493 database compounds are presented. The query molecule is the highest ranked molecule in all cases with similarity score 1 and thus appears always as the first on the left. 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, it is good at finding different chemical scaffolds, as it can be observed from the fourth query (*d*). (*a*) The first query molecule has 17 atoms, (*b*) the second query molecule has 25 atoms, (*c*) the third query molecule has 30 atoms, (*d*) the fourth query molecule has 33 atoms, and (*e*) the fifth query molecule has 38 atoms.

encoded. This is a 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. At this stage, each molecule is described by as many features (the one-dimensional distribution of atomic distances) as the number of atoms in the molecule. This raises the obvious question of how to compare molecules with different number of atoms. This difficulty is circumvented by defining a fixed number of moments of the one-dimensional distributions, whose values characterize the molecule considered. Finally, the shape similarity score of two molecules is calculated by the sum of least absolute differences of their respective moments. Note that the characteristics of the method ensure that a high score will be always assigned to a molecule with similar shape, which would be consequently ranked highly.

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 samples from the distribution of all atomic distances from the molecular centroid (d^{ctd}),

$$\left\{d_j^{\mathrm{ctd}}\right\}_{j=1}^{N}, \tag{3.1}$$

where N is the number of atoms of the molecule considered.

The next stage of the process is to calculate the moments of this discrete distribution in order to characterize 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 variance of these atomic distances about μ_1^{ctd}. The third moment (μ_3^{ctd}) is the skewness of these atomic distances about μ_1^{ctd} (i.e. a measure of the asymmetry of the distribution). To calculate the remaining nine descriptors, we repeat the process for each of the following three remaining distributions:

$$\left\{d_j^{\mathrm{cst}}\right\}_{j=1}^{N}, \left\{d_j^{\mathrm{fct}}\right\}_{j=1}^{N} \text{ and } \left\{d_j^{\mathrm{ftf}}\right\}_{j=1}^{N},$$

where the superscripts 'cst', 'fct' and 'ftf' indicate the location from where the atomic distances are calculated. Of course, one can include more reference locations leading to more descriptors and thus an even more accurate description of shape. However, we selected the first three moments from each of the four one-dimensional distributions considered 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 (figure 1).

Finally, the molecules in the database are ranked according to a similarity score which is defined as follows. First, the Manhattan distance between the vectors of shape descriptors of the query and the currently screened molecule is calculated and divided by the number of descriptors. The resulting dissimilarity measure is transformed into a normalized similarity score by translating the dissimilarity by one unit and inverting the resulting value. Other ways to define a normalized 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

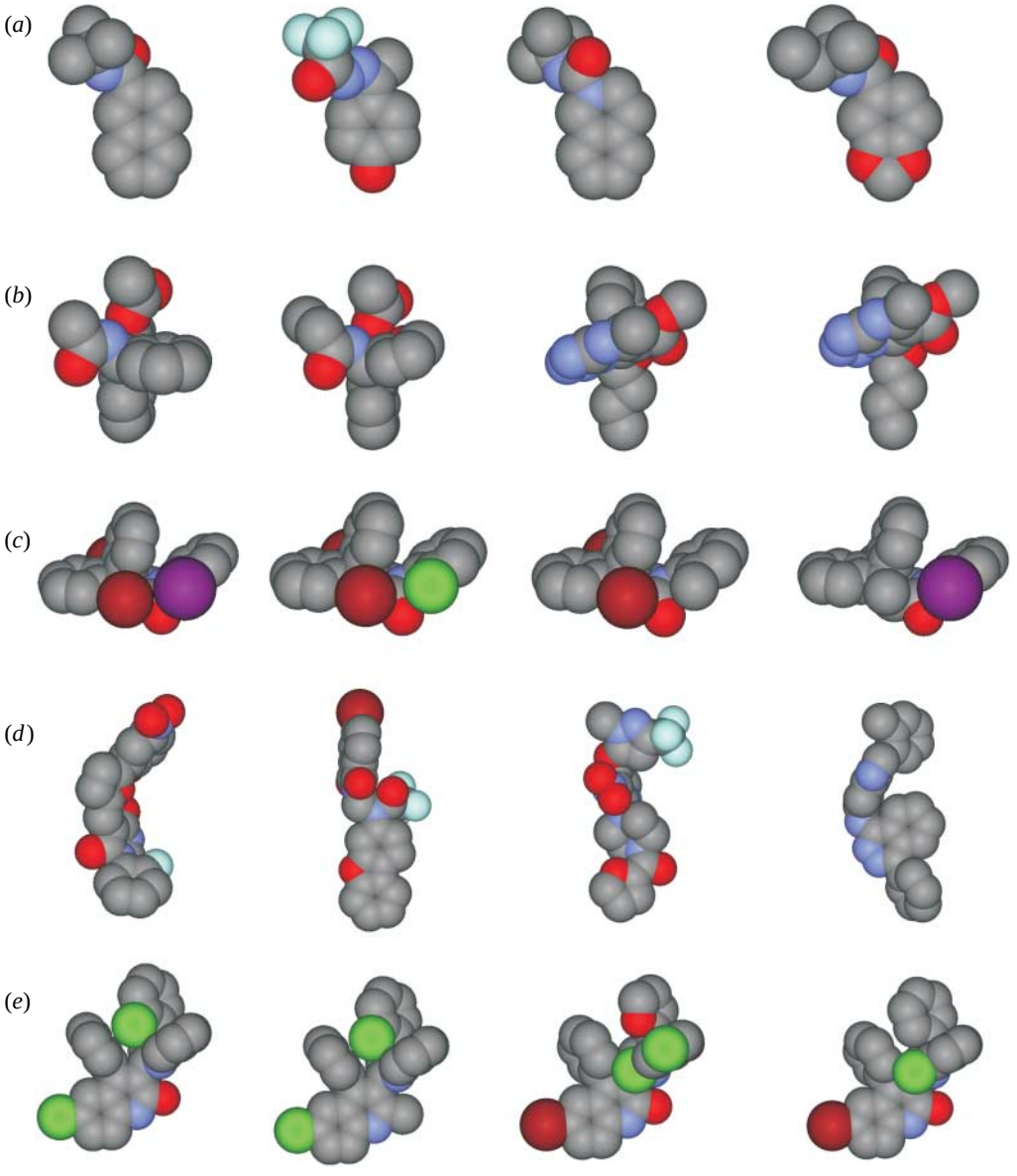

Figure 4. (*a–e*) Screened molecules with the highest ESshape3D score for the same queries as in figure 3. For each query, the four highest ranked molecules out of 2 433 493 database compounds are presented. The query molecule is the highest ranked molecule in all cases with maximum similarity score and thus appears always as the first on the left. The top hits for the third and fifth queries (*c*, *e*) have a consistent ranking and are quite similar to those obtained with USR (compare with figure 3). However, the remaining three queries (*a*, *b*, *d*) have top hits which are not as similar as the USR top hits. This is particularly noticeable in the fourth query (*d*), where the second, third and fourth most similar molecules are visually much more dissimilar to the query than the corresponding USR top hits.

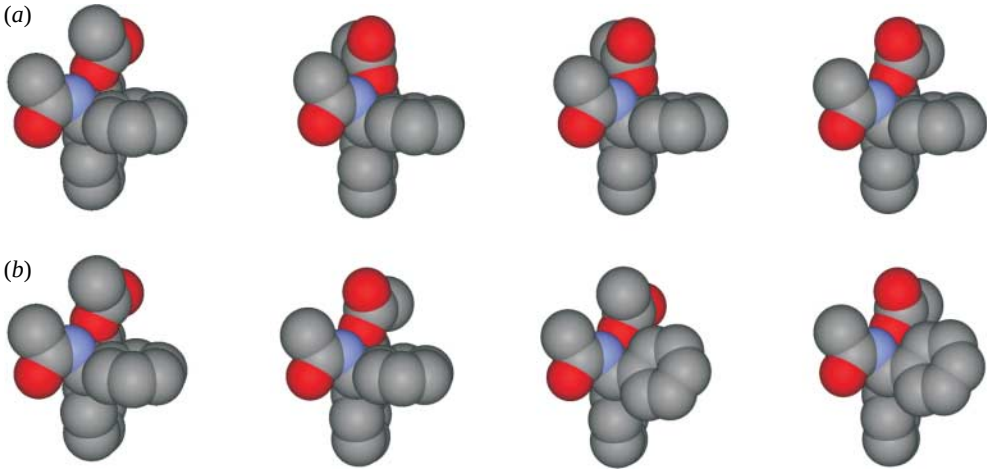

Figure 5. USR robustness to conformations. We took the conformation used for the second query in figure 3, which has four possible extremes, and calculated 292 additional conformations of this molecule. Figure (*a*) shows the conformers with the highest USR scores, while figure (*b*) shows those with the highest ESshape3D scores. Again, it is observed that USR retrieves more similarly shaped conformers than ESshape3D, despite the presence of multiple conformers of the query molecule.

ranking order. The similarity score function S_{qi} is therefore

$$S_{qi} = \left(1 + \frac{1}{12}\sum_{l=1}^{12} |M_l^q - M_l^i|\right)^{-1}, \tag{3.2}$$

where $\vec{M}^q$ and $\vec{M}^i$ are the vectors of shape descriptors for the query and the ith screened molecule, respectively (figure 2).

4. Results and discussion

We have made several experiments to test our methodology. These are carried out with a database that contains 2 433 493 commercially available compounds. Each database entry represents the chemical structure of the compound in three-dimensional MDL SD format (without including hydrogen atoms). The database was generated to contain only one conformer per compound, with each of them having at least 10 heavy atoms.

The first experiment is intended to evaluate the efficacy of the proposed descriptors for accurately encoding shape. This is a complicated endeavour as no shape comparison method has been shown to be completely accurate at describing shape and therefore there is no ground truth to compare with. A number of studies (Nilakantan *et al.* 1993; Hahn 1997; Zauhar *et al.* 2003) have addressed this difficulty by visually comparing the top-ranked molecules provided by the shape comparison method. Figure 3*a–e* shows the screened molecules with the highest USR score for five different queries. 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 is not observed in figure 3. These queries were selected because they represent a diverse subset of the chemical space in terms of the number of atoms and the type of shape, but we observed results of similar quality in every additional query we have made. Our method is able to identify shapes that closely resemble that of the query, which is also in the database and is the highest ranked molecule in all cases. In addition, the experiment shows that the method is particularly good at finding different chemical scaffolds as it can be observed from the fourth query in figure 3*d*, which constitutes a very valuable capability (Böhm *et al.* 2004; Rush *et al.* 2005).

An even stronger validation can be carried out by comparing these hits against those provided by another shape comparison method in order to investigate whether USR misses any molecule with a significantly more similar shape. Figure 4 shows the screened molecules with the highest ESshape3D score for the same queries as in figure 3. ESshape3D also retrieves the query molecule with maximum similarity score in all cases and thus appears on each row (query) as the first molecule on the left. The top hits for the third (*c*) and fifth (*e*) queries have a consistent ranking and are quite similar to those obtained with USR (compare with figure 3). However, the remaining three queries (*a*,*b*,*d*) have top hits which are not as similar to the query molecule as the USR top hits. This is particularly noticeable in the fourth query (*d*), where the second, third and fourth most similar molecules are visually much more dissimilar to the query than the corresponding USR top hits.

It could be argued that the procedure used to locate the reference points in USR might be sensitive to small details of the conformation rather than the overall shape of the molecule. However, similarly shaped conformers share a similar relative position of their respective atoms in the three-dimensional space and therefore the location of the reference points should be similar as well. In order to illustrate this issue, we took the conformation used for the second query in figure 3*b*, which has four possible extremes, and calculated 292 additional conformations of this molecule. Figure 5*a* shows the conformers with the highest USR scores, while figure 5*b* shows those conformers with the highest ESshape3D scores. Again, it is observed that USR retrieves more similarly shaped conformers than ESshape3D, despite the presence of multiple conformers of the query molecule.

The second area to investigate is the efficiency of the method. With this purpose, the molecular shape comparison rate will be calculated for USR and compared with that from three state-of-the-art methods: ESshape3D, Shape Signatures and ROCS. Unlike ESshape3D, a direct efficiency comparison with the last two methods is not possible for the reasons previously discussed in this paper. However, it will still be possible to make an approximate comparison because these methods were recently published and thus we have access to computers with similar power to that used in the studies where their efficiency was reported. In this way, figure 6 presents the comparison rate of USR versus the two descriptor-based shape methods, ESshape3D and Shape Signatures. USR is 1546 and 2038 times faster than ESshape3D and Shape Signatures, respectively.

In figure 7, an approximated comparison with the superposition-based method ROCS is made. USR obtained a comparison rate of 14 238 500 molecules per second, which is hence about 14 238 times faster than that reported by ROCS.

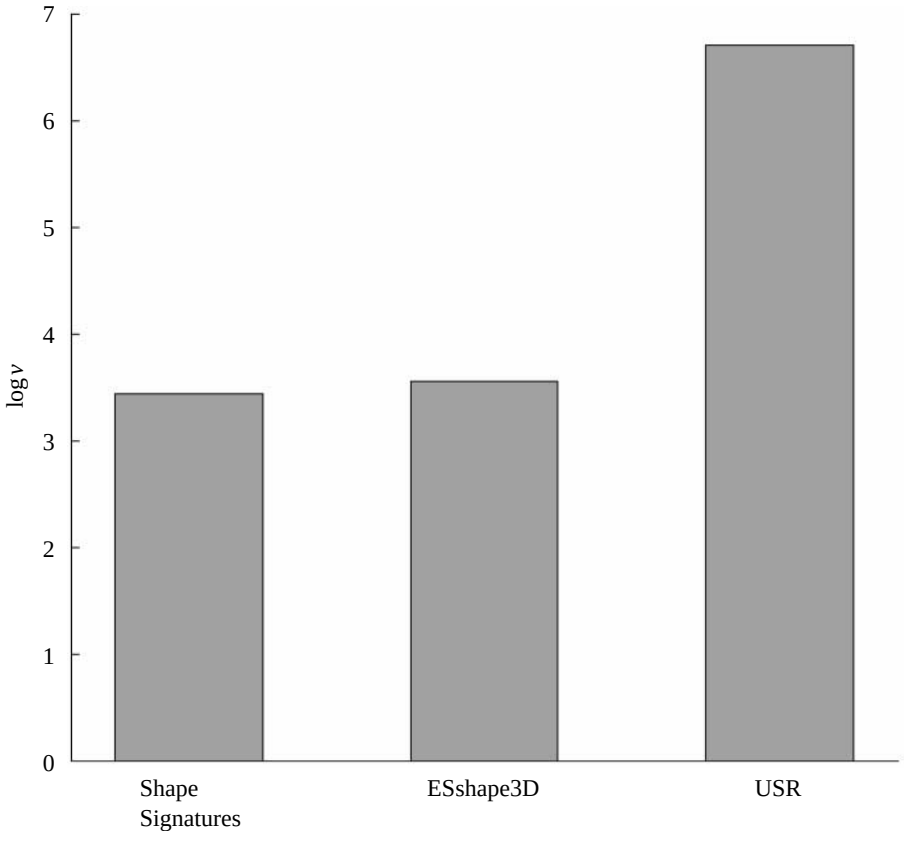

Figure 6. Efficiency comparison (in logarithm of screened molecules per second) between USR and two state-of-the-art descriptor-based shape methods. The rate for USR and ESshape3D was calculated on a modestly powerful PC (AMD Athlon XP 1800+CPU at 1.5 GHz with 512 MB of memory), which is also very similar to that used in Shape Signatures (Zauhar *et al.* 2003). USR is 1546 and 2038 times faster than ESshape3D and Shape Signatures, respectively.

It is worth noting that ROCS is widely regarded as the fastest superposition-based method and it has been claimed (Rush *et al.* 2005) to be order of magnitudes faster than other three-dimensional methods.

USR efficiency makes it sufficiently fast to extract information from the largest molecular databases available, as it would be able to identify the most similar shapes out of the 3.5 billion molecules constituting the screensaver database (Richards 2002) in approximately 4 minutes on a single processor. To illustrate further the significance of USR, let us consider a possible research scenario where one would like to find the most similarly shaped compounds within the screensaver database for each query in a set of 100 interesting molecules. This would take about 7 hours with USR. By contrast, ESshape3D, Shape Signatures and ROCS would take about 1.2, 1.6 and 11.1 years, respectively. This ability to handle larger databases has been pointed out (Claus & Underwood 2002) as a crucial component to addressing the future of the pharmaceutical industry.

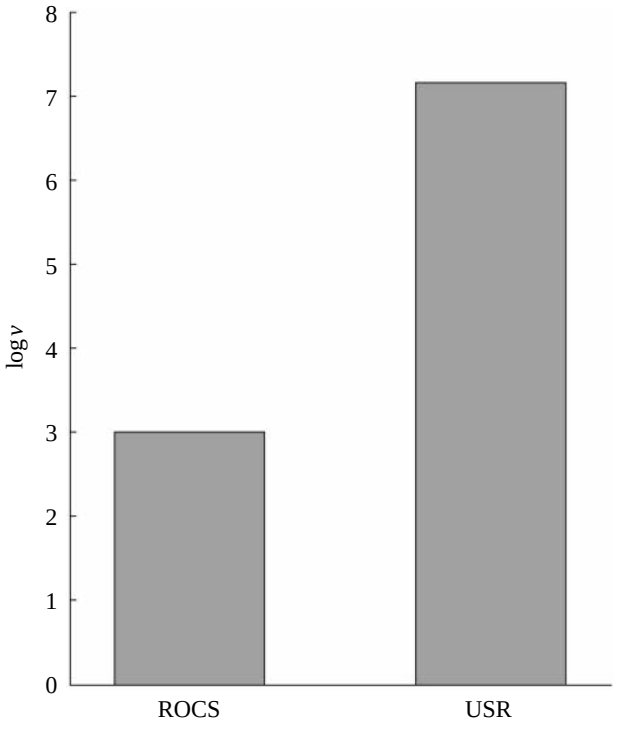

Figure 7. Efficiency comparison (in logarithm of screened molecules per second) between the presented method (USR) and ROCS (Rush *et al.* 2005), a widely used superposition-based shape comparison method. As ROCS is not available for validation studies, an approximate comparison is presented based on its reported (Rush *et al.* 2005) comparison rate on a modern Intel/AMD processor. Therefore, USR was run on one of the cores of an Intel Core2 2.93 GHz processor with 4 GB of memory. USR obtained a comparison rate of 14 238 500 molecules per second, which is hence about 14 238 times (four orders of magnitude) faster than that reported by ROCS.

The last test presents an interesting capability of USR. Unlike superposition methods where shape can only be calculated with respect to the overlapping molecule, USR defines the shape of a molecule independently and uses a fixed set of descriptors for every molecule. The latter ensures that every molecule will have a unique location in the 12-dimensional chemical space spanned by the used descriptors. This is a major advantage when finding and visualizing clusters of molecules with similar shape. There are many applications of such representation. For instance, each of these clusters is a region of the chemical space with similarly shaped molecules and thus it could be regarded as compounds that are likely to share similar biological activity with the query molecule. In addition, such representation shows at a glance where the geometry of the compared molecules differs. In combination with a suitable clustering algorithm, one could find clusters in a molecular database in order to select the most representative molecule of each cluster. The latter could be applied, for example, as a way to avoid repeating expensive biological tests on similar molecules. Figure 8 shows an example of this representation based on the results of five queries presented in figure 3.

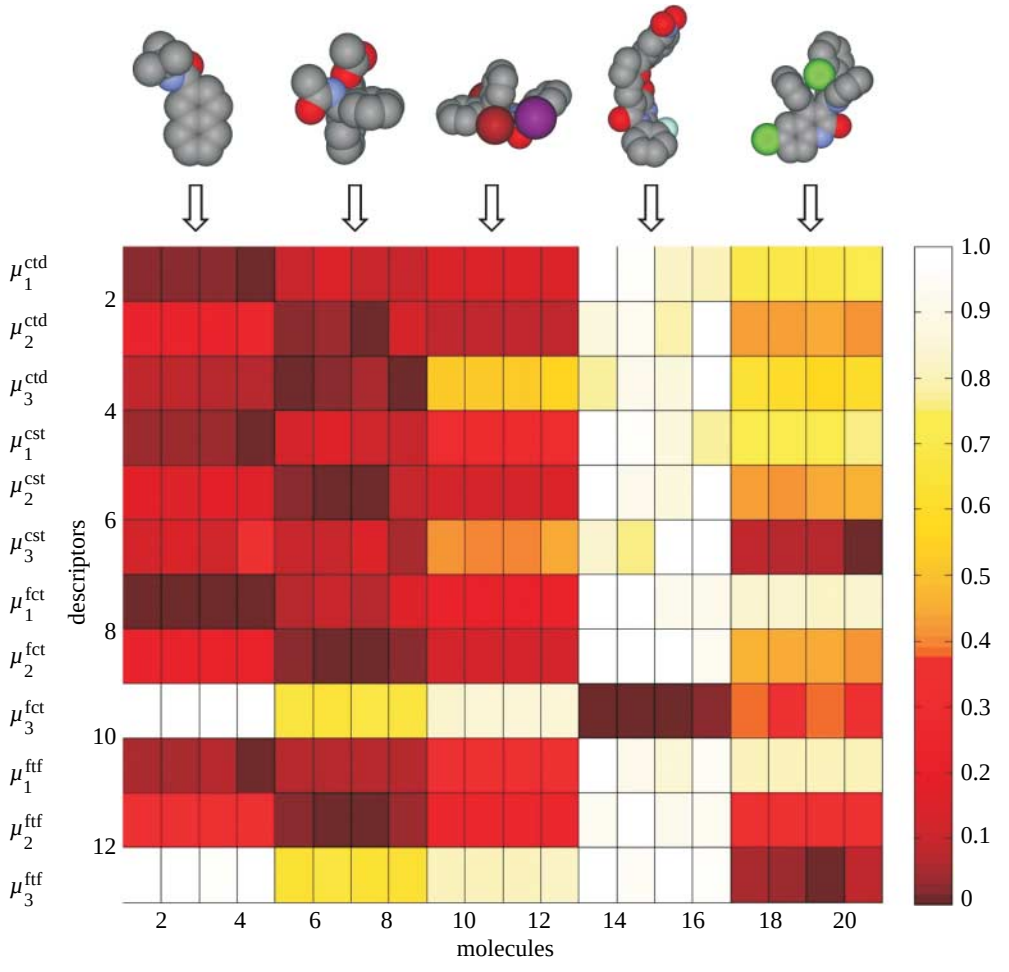

Figure 8. Twelve-dimensional representation of a tiny region of the chemical space according to three-dimensional shape. Each column corresponds to a molecule from figure 3 and each row is one of the molecular descriptors whose normalized values are given by the colour bar on the right (note that such normalization does only allow a direct comparison of molecules across a single molecular descriptor at a time). The molecules are ordered in clusters (a cluster is defined as a group of molecules which are similar among themselves, but dissimilar when compared with molecules from other clusters). Each cluster contains the four most similar molecules to the query molecule (pictured on the top).

5. Conclusions and future prospects

A new method for USR based on moments of interatomic distance distributions has been presented. It was motivated by the relative inefficiency of current shape comparison methods, which are not able to cope with the largest molecular databases available in a reasonable time. USR has been shown to effectively search molecular databases at least 1546 times faster than current methodologies. In addition, the problematic requirement of aligning molecules for comparison is circumvented, as the proposed distributions are independent of the spatial

orientation of database molecules. Lastly, the way USR encodes shape provides every molecule with a unique location in the 12-dimensional chemical space spanned by the used shape descriptors. This opens the door to the application of existing clustering algorithms to find groups of similar molecules as a way to analyse the molecular diversity of a database in terms of molecular shape.

The method could be also adapted to tackle similar shape comparison problems in other fields, such as designing content-based Internet search engines for three-dimensional geometrical objects (Funkhouser *et al.* 2005) or performing fast similarity comparisons between macromolecules (e.g. proteins; Albrecht *et al.* 2004). From a broader perspective, we anticipate that ultrafast pattern recognition will soon become not only useful, but also essential. In most areas of modern science, the amount of archived data is increasing at an explosive rate (Muggleton 2006; Szalay & Gray 2006) and its analysis is becoming more and more complex (Szalay & Gray 2006), a trend that is expected to continue in the foreseeable future (Szalay & Gray 2006). However, this data explosion has not resulted in an information explosion, mainly owing to the difficulties of current methods to cope with massive databases (Szalay & Gray 2006). We believe that the presented approach is one way to tackle the enormous challenge posed by scientific data explosion in pattern recognition in general and molecular shape comparison in particular.

This work was partly supported by the US National Foundation of Cancer Research. The authors would like to thank Paul Finn and Dan Butler from Inhibox Ltd for their helpful discussions and providing the molecular database used in the experiments. The authors are also thankful for feedback from Francis Marriott at University of Oxford on statistical considerations relevant to this work. Early feedback from Lydia Kavraki at Rice University is also thankfully acknowledged.

References

Albrecht, B., Grant, G. H. & Richards, W. G. 2004 Evaluation of structural similarity based on reduced dimensionality representations of protein structure. *Protein Eng. Design Select.* **17**, 425–432. (doi:10.1093/protein/gzh049)

Baringhaus, K.-H. & Hessler, G. 2004 Fast similarity searching and screening hit analysis. *Drug Discov. Today: Technol.* **1**, 197–202. (doi:10.1016/j.ddtec.2004.11.001)

Bemis, G. W. & Kuntz, I. D. 1992 A fast and efficient method for 2D and 3D molecular shape description. *J. Comput. Aid. Mol. Des.* **6**, 607–628. (doi:10.1007/BF00126218)

Bender, A. & Glen, R. C. 2004 Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204–3218. (doi:10.1039/b409813g)

Böhm, H.-J., Flohr, A. & Stahl, M. 2004 Scaffold hopping. *Drug Discov. Today: Technol.* **1**, 217–224. (doi:10.1016/j.ddtec.2004.10.009)

Claus, B. L. & Underwood, D. J. 2002 Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* **7**, 957–966. (doi:10.1016/S1359-6446(02)02433-9)

Funkhouser, T., Kazhdan, M., Min, P. & Shilane, P. 2005 Shape-based retrieval and analysis of 3D models. *Commun. ACM* **48**, 58–64. (doi:10.1145/1064830.1064859)

Good, A. C. & Richards, W. G. 1998 Explicit calculation of 3D molecular similarity. *Perspect. Drug Discov. Des.* **9-11**, 321–338. (doi:10.1023/A:1027280526177)

Good, A. C., Ewing, T. J. A., Gschwend, D. A. & Kuntz, I. D. 1995 New molecular shape descriptors: application in database screening. *J. Comput. Aid. Mol. Des.* **9**, 1–12. (doi:10.1007/BF00117274)

Hahn, M. 1997 Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* **37**, 80–86. (doi:10.1021/ci960108r)

Jenkins, J. L., Glick, M. & Davies, J. W. 2004 A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **47**, 6144–6159. (doi:10.1021/jm049654z)

Kola, I. & Hazuda, D. 2005 Innovation and greater probability of success in drug discovery and development—from target to biomarkers. *Curr. Opin. Biotechnol.* **16**, 644–646. (doi:10.1016/j.copbio.2005.10.014)

Kotani, T. & Higashiura, K. 2002 Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances. *J. Chem. Inf. Comput. Sci.* **42**, 58–63. (doi:10.1021/ci010068d)

Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C. & Labaudiniere, R. F. 1999 New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **42**, 3251–3264. (doi:10.1021/jm9806998)

MOE, 2006.08 release. (http://www.chemcomp.com/)

Muggleton, S. H. 2006 Exceeding human limits. *Nature* **440**, 409–410. (doi:10.1038/440409a)

Nilakantan, R., Bauman, N. & Venkataraghavan, R. 1993 New method for rapid characterisation of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.* **33**, 79–85. (doi:10.1021/ci00011a012)

Richards, W. G. 2002 Virtual screening using GRID computing: the screensaver project. *Nat. Rev. Drug Discov.* **1**, 551–555. (doi:10.1038/nrd841)

Rush III, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. 2005 A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J. Med. Chem.* **48**, 1489–1495. (doi:10.1021/jm040163o)

Schnecke, V. & Boström, J. 2006 Computational chemistry-driven decision making in lead generation. *Drug Discov. Today* **11**, 43–50. (doi:10.1016/S1359-6446(05)03703-7)

Szalay, A. & Gray, J. 2006 Science in an exponential world. *Nature* **440**, 413–414. (doi:10.1038/440413a)

Willett, P. 2005*a* Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **48**, 4183–4199. (doi:10.1021/jm0582165)

Willett, P. 2005*b* Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules. In *Proc. FIS2005*, F69, pp. 1–15. (http://www.mdpi.org/fis2005/proceedings.html)

Zauhar, R. J., Moyna, G., Tian, L., Li, Z. & Welsh, W. J. 2003 Shape signatures, a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **46**, 5674–5690. (doi:10.1021/jm030242k)