Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Molecular similarity including chirality

M. Stuart Armstrong [a,*], Garrett M. Morris [b], Paul W. Finn [b], Raman Sharma [b], W. Graham Richards [c]

[a] InhibOx, Pembroke House, 36-37 Pembroke Street, Oxford OX1 1BP, UK
[b] InhibOx, Pembroke House, 36-37 Pembroke Street, Oxford OX1 1BP, UK
[c] University of Oxford, Department of Chemistry, InhibOx Laboratory, Pembroke House, 36-37 Pembroke Street, Oxford OX1 1BP, UK

ARTICLE INFO

Article history:
Received 4 August 2009
Accepted 8 September 2009
Available online 16 September 2009

Keywords:
Molecular similarity
Ligand-based virtual screening
Drug design
Molecular shape
Enantiomers
Chirality
Ultra-fast Shape Recognition

ABSTRACT

This paper presents CSR, or Chiral Shape Recognition, a novel method to compute molecular similarity that builds on the Ultra-fast Shape Recognition (USR) method, but distinguishes enantiomers. It has great potential for generalisation, and was tested on the DUD dataset, where it was found a significant improvement in enrichment over USR having screened and ranked the top 0.25 %, 0.5 % and 1% of the database.

1. Introduction

Quantitative measures of molecular similarity have proved to be a useful tool in computer-aided drug discovery. They have been used to explore large databases of molecular structures to find compounds which resemble a given natural product. They have also been used to find compounds of similar activity to an active molecule that is not itself suitable, either because it is patented or has side effects. Most recent molecular similarity research has focused on non-superpositional approaches, motivated by the fact that the superposition methods are too slow for routine use. However, these fast methods cannot distinguish enantiomers, thus missing a very important facet of biological interactions. Here we solve this problem by building on ultra-fast shape recognition (USR) [1,2]; the novel feature is to position the centroids in such a way that it clearly distinguishes between enantiomers.

Furthermore, the weight afforded to chirality can be tuned as needed. We anticipate that this approach will be generalised to a great variety of fast shape comparison methods. This is because the crucial component of the approach – the use of the cross product – is suitable for use in almost all of these methods, and can be added in ‘artificially’ even if it is not natural for the method.

2. The method

When engaged in the search for new chemical entities that possess a particular useful property, such as biological activity, it has frequently been found useful to test compounds that are structurally similar to a known active molecule. This is based on the idea that structurally similar compounds possess similar biological properties, also known as neighbourhood behaviour [5]. A wide variety of algorithms have been constructed [6] including approaches comparing molecules at the 2D (chemical connectivity) and 3D (shape) levels and these have been shown to be useful in a variety of studies [3].

Our interest is in those methods that compare molecules based on their shape. Many of these require that the two molecules being compared be structurally aligned and superposed in a common coordinate framework. This poses difficulties in finding the global optimum in the search space and also tends to have very high computational requirements, hence being quite slow. An alternative strategy, which avoids these difficulties, is to use a non-superpositional approach, such as the recently developed USR [1,2].

The standard USR method can be summarised as follows: for a given molecule, a collection of four centroids are computed (the geometric centre, the closest atom to this centre, the furthest atom from the second centroid, and the furthest atom from the third

[*] Corresponding author. Tel.: +44 0 1865 262 000; fax: +44 0 1865 243 072.
E-mail addresses: garrett.morris@inhibox.com (G.M. Morris), paul.finn@inhibox.com (P.W. Finn), ramansharma78@hotmail.com (R. Sharma), stuart.armstrong@inhibox.com, graham.richards@chem.ox.ac.uk (W.G. Richards).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.09.002

centroid). Then for each centroid, the distances to each atom in the molecule is calculated, giving four distance distributions. The first three moments of each distribution are then calculated, generating a structural descriptor that consists of twelve numbers. Two molecules are compared by determining the distance between these twelve-dimensional vectors.

The USC method however does not distinguish between enantiomers. Schematically, USR can be decomposed as follows:

1. Assigning centroids.
2. Computing distance distributions from the centroids.
3. Computing moment information from the distributions.

There is no possible scope for distinguishing between enantiomers in the second step – distances are preserved by reflections – and hence none in the third. The first step offers possibilities, however, as long as we can isolate some operation that transforms equivariantly under rotations and translations, but not under reflections. The cross product is such an operation: if $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are vectors, and ρ is any reflection, then

$$\rho(\vec{\mathbf{a}}) \times \rho(\vec{\mathbf{b}}) = -\rho(\vec{\mathbf{a}} \times \vec{\mathbf{b}}). \tag{1}$$

The minus sign makes all the difference. In our modification of the method, we assign three centroids cen_1, cen_2 and cen_3 traditionally as the geometric centre, the atom furthest from the geometric centre, and the atom furthest from that one. We then define vectors $\vec{\mathbf{a}} = cen_2 - cen_1$ and $\vec{\mathbf{b}} = cen_3 - cen_1$. Unless the centroids are colinear (a very rare situation), then the cross product $\vec{\mathbf{a}} \times \vec{\mathbf{b}}$ will be non-zero. To keep everything in the same units and of comparable magnitudes, we normalise this vector to have half the norm of the vector $\vec{\mathbf{a}}$, thus:

$$\vec{\mathbf{c}} = \left(\frac{||\vec{\mathbf{a}}||}{2}\right) \frac{\vec{\mathbf{a}} \times \vec{\mathbf{b}}}{||\vec{\mathbf{a}} \times \vec{\mathbf{b}}||}. \tag{2}$$

The fourth centroid is assigned to be the point $cen_4 = cen_1 + \vec{\mathbf{c}}$ (in the case where $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are colinear, we take $cen_4 = cen_1$). From the way cen_1, cen_2 and cen_3 have been assigned, $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are invariant under translations of the molecule and equivariant under rotations, implying the same thing about $\vec{\mathbf{c}}$.

Now replace the molecule with a mirror image, via the central symmetry, taking cen_1 as the origin (any reflection can be made equivalent to this symmetry by the action of translations and rotations). Then each atom position $at\,pos_i$ goes to $at\,pos'_i = -at\,pos_i$, while $cen'_1 = -cen_1$, $cen'_2 = -cen_2$ and $cen'_3 = -cen_3$. However, because of the properties of the cross product, $cen'_4 = cen_4$, not $-cen_4$. This means – because all the atom positions have been inverted, but the position of the fourth centroid has not – that the fourth distance distribution is different for a given molecule and its enantiomer.

Using these four centroids, we can construct the distance distributions and the moments, generating the twelve descriptors as in USR. We will call this method Chiral Shape Recognition, or CSR for short.

If needed, we could make the method even more sensitive to chirality by constructing more centroids that have the above properties. Starting from five temporary centroids cen_1 to cen_5, we could construct four vectors $\vec{\mathbf{a}}_i = cen_i - cen_1$, and six vectors $\vec{\mathbf{a}}_i \times \vec{\mathbf{a}}_j$, giving us up to six enantiomer-distinguishing centroids, for instance; but we found that the current implementation was sufficiently capable of distinguishing enantiomers for our purposes.

3. Performance

Both USR and CSR are most useful for ranking and filtering molecules, rather than taking the twelve descriptors as having an absolute meaning. There are a variety of reasons for this. The procedure of computing centroids in not a continuous procedure, as concepts such as 'closest' and 'furthest' atom do not vary continuously as the molecule changes shape. Furthermore, there is no sensible way of weighting the contributions of the different moments, nor any evident renormalisation procedure for molecules with greatly different numbers of atoms. Thus early enrichment studies seem the best ways of comparing such methods.

Table 1
Enrichment values for USR and Chiral USR compared.

Method used:	0.25 % enrichment	0.5 % enrichment	1 % enrichment
USR	10.87	8.92	6.94
CSR	13.10	10.10	7.36

USR and CSR were tested using the DUD datasets [4]. The DUD dataset consist of forty different targets (from angiotensin-converting enzyme through to vascular endothelial growth factor receptor kinase). Each dataset consists of two collections of molecules: the ligands (known actives that bind to the target) and the decoys (molecules with similar chemical and physical properties to the ligands, assumed to be non-binding). On average, the ratio of ligands to decoys is just under 35. Since the goal is to test chirality, the enantiomers of every ligand and decoy were generated and added to the appropriate decoy set for a given target (the chiral nature of biological interactions makes it a reasonable assumption that these would be non-active). This increased the average ratio of ligands to decoys to just over 70.

Then our implementation of the USR method and CSR were tested in turn on each molecule in the ligand set. The Manhattan distance between the vectors of moments was used to rank all the other molecules in the target set according to their distance from the given molecule. Then the top 0.25 % of this ranked set were analysed, and the enrichment calculated: this is the ratio between the number of molecules in this top set that were ligands, divided by the number that would be expected by chance alone. The 0.25% enrichment for each molecule was then averaged across the ligand set, to give the average enrichment for the target set. Finally, this number was averaged across the forty ligand sets, to give the average enrichment. This was repeated for the 0.5% and 1% enrichments. The results are presented in Table 1.

As can be seen, CSR outperformed USR in all cases, boasting a 20 % improvement for 0.25 % enrichment. The standard deviation of the differences between the USR and CSR enrichments are 4.2, 2.5 and 1.2 respectively. With a total of 3961 ligand molecules, this improvement was statistically significant at the 99% level for all enrichments; the one-tailed confidence intervals were of width 0.15, 0.091 and 0.04 respectively.

4. Computational complexity

Compared with USR, CSR only requires the extra step of computation and renormalisation of a single cross-product—a computationally trivial step compared with the multiple distance calculations needed to generate the distance distributions from each centroid. Hence USR and CSR should be considered to be of roughly comparable speed.

5. Generalisation

The use of the cross product generalises in many possible ways to different distance comparison methods. If a method uses a pair of three-dimensional vectors derived from the molecule, then we can construct a suitably normalised cross product of these two, and use it instead of one of the original vectors.

Even if the method only uses a single vector, we can simply add a correction term to that vector (some multiple of the vector defined above), a correction term that is not reflection equivariant. The norm of this correction term can be scaled, depending on how important it is to distinguish enantiomers. For instance, if any method makes use of the electrical dipole, then the dipole vector can be modified as described.

In general, there will be various ways of including a cross-product term, which can be adapted to the particular method. The previous paragraph is a proof of principle that the concept extends to a wide variety of distance-based molecular comparison methods.

Acknowledgements

This research was partly financed by a grant of the 6th Framework Programme of the European Union (DeZnIT project).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.09.002.

References

[1] P.J. Ballester, P.W. Finn, W.G. Richards, Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology, J. Mol. Graph. Model. 27 (2009) 836–845.

[2] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes, J. Comput. Chem. 28 (2007) 1711–1723.

[3] P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of shape-matching and docking as virtual screening tools, J. Med. Chem. 50 (1) (2007) 74–82.

[4] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking sets for molecular docking, J. Med. Chem. 49 (23) (2006) 6789–6801.

[5] D.E. Patterson, R.D. Cramer, A.M. Ferguson, R.D. Clark, L.E. Weinberger, Neighborhood behavior: a useful concept for validation of molecular diversity descriptors, J. Med. Chem. 39 (1996) 3049–3059.

[6] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.