J Comput Aided Mol Des (2011) 25:785–790
DOI 10.1007/s10822-011-9463-8

Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension

M. Stuart Armstrong · Paul W. Finn · Garrett M. Morris · W. Graham. Richards

Received: 20 May 2011 / Accepted: 26 July 2011 / Published online: 6 August 2011

Abstract In a previous paper, we presented the ElectroShape method, which we used to achieve successful ligand-based virtual screening. It extended classical shape-based methods by applying them to the four-dimensional shape of the molecule where partial charge was used as the fourth dimension to capture electrostatic information. This paper extends the approach by using atomic lipophilicity (alogP) as an additional molecular property and validates it using the improved release 2 of the Directory of Useful Decoys (DUD). When alogP replaced partial charge, the enrichment results were slightly below those of ElectroShape, though still far better than purely shape-based methods. However, when alogP was added as a complement to partial charge, the resulting five-dimensional enrichments shows a clear improvement in performance. This demonstrates the utility of extending the ElectroShape virtual screening method by adding other atom-based descriptors.

Keywords Molecular similarity · Molecular descriptors · Ligand-based virtual screening · Drug design · Lipophilicity · Electrostatics

This research was financed in part by the 6th Framework Program of the European Union (DeZnIT, Project Number 037303).

M. S. Armstrong (✉) · P. W. Finn · G. M. Morris
InhibOx, Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, UK
e-mail: stuart.armstrong@inhibox.com

W. Graham. Richards
Department of Chemistry, University of Oxford, InhibOx Laboratory, Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, UK

Introduction

In a previous work [1], we introduced ElectroShape, an ultra-fast, descriptor-based tool for ligand-based virtual screening. This built on the original ultra-fast shape recognition (USR) method [2, 3] by incorporating the ability to distinguish chiral molecules [4] and adding atomic electrostatic information (via atomic partial charge) as a fourth dimension. The ElectroShape descriptors are thus calculated for each molecule by treating it as collection of points in four-dimensional Euclidean space (three spatial dimensions plus one for the partial charges). This method showed superior performance to USR in benchmarking studies with the widely used DUD dataset.

This approach, of adding molecular properties to the comparison through the incorporation of extra dimensions, is quite general and, at the end of that paper, we suggested that further improved methods might be possible, for example by replacement of the electrostatics dimension with another property, or through the addition of further properties, thus generating a comparison in a higher dimensional space. In this note, we report the results of doing so, using atomic lipophilicity. This is more than a simple extension of the method, however; by opening up further dimensions, it demonstrates that what could have been a one-trick pony (adding charge makes shape-based enrichments better) is a general result (adding extra independent atomic properties makes shape-based enrichments better).

Lipophilicity is one of the key properties used in the development of quantitative structure-activity relationships and in medicinal chemistry in general. It plays an important role in ligand affinity, where the mutual removal of lipophilic groups on the protein and ligand from the aqueous environment is a strong driving force favouring molecular

interaction, and in the ability of drugs to reach their site of action, which usually involves passage through many lipid membranes. Thus, we expected that lipophilicity would be a useful property to add to the ElectroShape paradigm.

Methodology

Many methods have been described for the computational estimation of lipophilicity, as logP. In order to incorporate lipophilicity into ElectroShape an atomic representation is required. Thus, we use the atomic lipophilicities of the alogP method [5], as implemented in the modeling package MOE [6].

Following our previous work, the effect on performance of inclusion of atomic lipophilicity values will be studied in enrichment experiments that will compare use of partial charges as a fourth dimension (as in [1], atomic lipophilicity as a fourth dimension, and their combination into a five dimension method.

We used the DUD datasets [7] for the enrichment studies to provide the best comparison with the results from the previous work. DUD is a collection of datasets for forty diverse targets, ranging from ACE (angiotensin-converting enzyme) to VEGFR2 (vascular endothelial growth factor receptor kinase). For each target, DUD supplies two sets of molecules: one containing known active ligands for the target, and one containing decoy molecules. The decoys are presumed not to bind with the given target, but have similar molecular properties to the actives for that target (charge, chemical formula, molecular weight, etc). These sets were assembled so that any screening program differentiating actives from decoys could not do so based on simple molecular properties alone. The partial charges were re-assigned using the modified MMFF94 [8] charges implemented in CCG MOE [6] (the difference between standard MMFF94 and MOE's MMFF94x is that in the later, hydrogen atoms on alkane carbons have a non-zero partial charge). These were the partial charges that resulted in the highest enrichment factor at 1% in the previous studies [1].

Enrichment

The enrichment ratio at $n\%$ of the library screened is a simple numerical measure, hereafter designated by $E_{n\%}$. It is calculated by looking at the top n % closest molecules to a given active—closest being according to the similarity method being tested, in this case ElectroShape. The enrichment score is the number of actives in this top $n\%$, divided by the number of actives that would be expected by chance alone. The enrichment score is then averaged across the active set, to get the $E_{n\%}$ for the entire target set.

We choose $E_{1\%}$ to simulate a typical virtual screening experiment.

ElectroShape implementation

The implementation of ElectroShape (calculation of the centroids, and then of the corresponding descriptors) is described in detail in Armstrong et al. [1]. Briefly, given a collection of atom positions, the procedure defines the creation of four or five distinct "centroids". Then, for each centroid, the (Euclidean) distances from that point to every atom position is computed, giving a distribution of distances. Then the mean, standard deviation, and third root of the third central moment of the distribution are calculated, making up the descriptors. There are thus three times as many descriptors as there are centroids.

In order to define a point in n-dimensional space, the distance from that point to $n + 1$ centroids is needed; see Fig. 1 for an illustration of this in three dimensions. Hence the three-dimensional USR and CSR methods used four centroids, standard ElectroShape made use of five, and this five-dimensional method needs to add a sixth centroid (the previous five are the same as in [1]).

It can be defined quite simply: it is the geometric centre of the molecule (hence equal to the first centroid), with an offset in the fifth dimension. This offset is defined to be equal to the norm of the distance between the first centroid and the second (which is the atom furthest from the first centroid). Using c_n to designate the n-th centroid, this is:

$$c_6 = c_1 + (0, 0, 0, 0, ||c_1 - c_2||).$$

This means that in five dimensions there will be 18 descriptors, rather than the 15 in four dimensions. Note that the fourth and fifth centroids were designed so as to be different from their equivalents in a mirror image molecule, and hence the six descriptors corresponding to

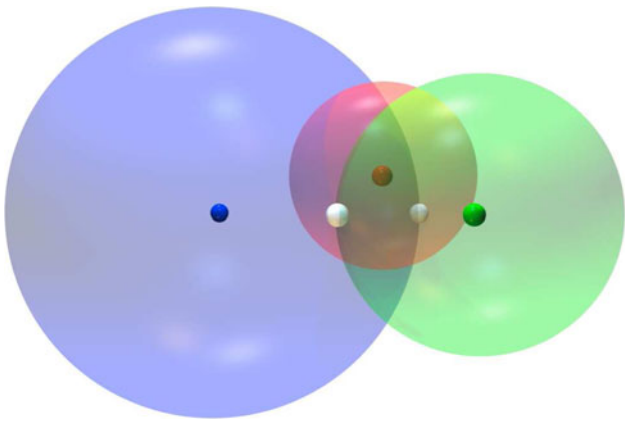

Fig. 1 Three spheres around each centroid intersect at two points (*shown in white*). To unambiguously determine the position of a point in three dimensions, a fourth centroid is required

them differ between enantiomers. This not the case for this six centroid, and hence the last three descriptors do not contribute to distinguishing enantiomers. However the enantiomers are still distinguished by the method overall.

Distance scaling

Molecular distances are given in Ångstroms, while charge is in electron charge and alogP in its own unit. Hence we

Table 1 Average enrichment at 1% for four-dimensional ElectroShape (using partial charge—MMFF94 (modified)—at a ratio of 25Å per electron charge, and alogP, also at a ratio of 25Å per alogP unit) and five-dimensional ElectroShape (at a ratio of 25 Å per electron charge and 5 Å per alogP unit)

Dataset:	4-dim MMFF94 (modified):	4-dim alogP:	5-dim:	Improvement over MMFF94 (modified):	Improvement over alogP:
ACE	9.2	12.2	10.6	1.4	−1.6
ACHE	7.1	6.5	6.2	−0.9	−0.3
ADA	15.2	6.2	15.0	−0.2	8.8
ALR2	9.1	10.5	10.8	1.7	0.3
AMPC	17.5	6.7	13.9	−3.6	7.2
AR	14.9	15.0	15.0	0.1	0.0
CDK2	5.1	4.8	5.8	0.7	1.0
COMT	3.3	9.8	15.2	11.9	5.4
COX1	3.8	4.7	4.8	1.0	0.1
COX2	17.9	12.6	20.8	2.9	8.2
DHFR	15.5	13.8	18.4	2.9	4.6
EGFR	13.7	9.9	9.4	−4.3	−0.5
ER agonist	16.3	13.9	16.3	0.0	2.4
ER antagonist	20.8	17.7	21.4	0.6	3.7
FGFR1	8.2	12.1	13.2	5.0	1.1
FXA	19.1	9.3	16.0	−3.1	6.7
GART	14.7	16.2	15.4	0.7	−0.8
GPB	8.2	5.5	6.6	−1.6	1.1
GR	16.6	17.6	21.2	4.6	3.6
HIVPR	11.6	15.8	15.7	4.1	−0.1
HIVRT	2.9	6.4	4.6	1.7	−1.8
HMGR	33.3	32.6	34.4	1.1	1.8
HSP90	16.5	11.1	14.6	−1.9	3.5
INHA	15.2	15.6	17.2	2.0	1.6
MR	15.0	10.8	15.0	0.0	4.2
NA	17.4	7.4	17.1	−0.3	9.7
P38	10.9	12.5	12.1	1.2	−0.4
PARP	10.7	10.9	17.5	6.8	6.6
PDE5	8.7	8.5	9.0	0.3	0.5
PDGFRB	14.8	14.3	18.4	3.6	4.1
PNP	12.5	5.2	8.3	−4.2	3.1
$PPAR_{\gamma}$	16.4	18.4	18.0	1.6	−0.4
PR	19.6	18.1	17.2	−2.4	−0.9
RXR_{α}	17.9	19.1	21.4	3.5	2.3
SAHH	13.3	11.9	17.6	4.3	5.7
SRC	8.8	8.1	9.5	0.7	1.4
THROMBIN	18.5	15.9	21.3	2.8	5.4
TK	6.8	5.9	7.5	0.7	1.6
TRYPSIN	20.4	16.3	22.5	2.1	6.2
VEGFR2	6.1	5.6	7.8	1.7	2.2

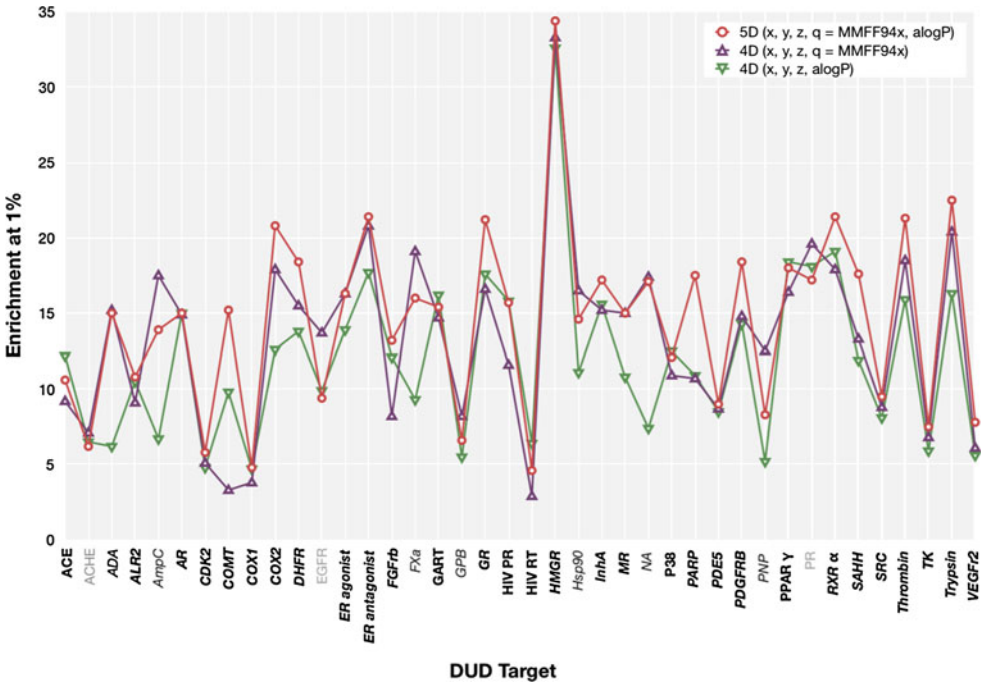

Fig. 2 Average enrichment at 1% for four-dimensional ElectroShape (using partial charge—MMFF94 (modified) and AM1—at a ratio of 25Å per electron charge, and alogP, also at a ratio of 25 Å per alogP unit) and five-dimensional ElectroShape (at a ratio of 25Å per electron charge and 5Å per alogP unit)

need a scale factor to convert everything into Ångstroms. For standard four-dimensional ElectroShape, the best ratio was found to be 25 Å per electron charge. For the four-dimensional version with alogP instead of partial charge, the best ratio was similarly of 25 Å per alogP unit.

For five-dimensional ElectroShape, which uses both partial charge and alogP, the best ratios were found to be 25 Å per electron charge (as before) and 5 Å per alogP unit.

Results

The enrichments at 1% are presented in Table 1, for each of the forty datasets in DUD, along with improvement (positive or negative) of five-dimensional ElectroShape over both four-dimensional versions. The graphical version of this is given in Fig. 2. Note that when the five-dimensional method performs (non-strictly) better than the four-dimensional charge-based method, the dataset name is in bold; when it performs better than the four-dimensional alogP method, the database name is in italics. Thus it can be seen that there are 30 databases in the first situation, 31 in the second, and 24 where the five-dimensional outperforms both four-dimensional methods. In only three databases—ACHE, EGFR and PR—did the five-dimensional method do worse that both four-dimensional methods.

The average of these enrichment results are presented in Table 2. For comparison, the enrichment results for the partial charge assignment of AM1 [9] (see [1]) are also presented.

As can be seen, overall performance of the atomic lipophilicity descriptor on its own is better than that of some partial charge allocation schemes studied previously (such as AM1 and Gasteiger, see [1]), but does not lead to improvement enrichment over the best performing electrostatics scheme (MMFF94 modified charges), with the alogP method showing an average enrichment at 1% of 11.9, compared to 13.3 for the MMFF94 modified charge scheme. Regarding performance on individual datasets, it performs better in 15 of the 40 datasets. The best performing datasets relative to electrostatic include ACE, COMT, FGFR1, HIVPR and HIVRT. The ADA, AMPC, COX2 and PNP are among the relatively poorest performing. The reasons behind this variation in performance may be worth pursuing. They do not appear to be simply related to a particular target class, for example.

However, when electrostatics and lipophilicity are combined into a five-dimensional similarity comparison, there is an impressive and significant increase in performance, from

Table 2 Average enrichment at 1% for four-dimensional ElectroShape (using partial charge—MMFF94 (modified) and AM1—at a ratio of 25 Å per electron charge, and alogP, also at a ratio of 25 Å per alogP unit) and five-dimensional ElectroShape (at a ratio of 25 Å per electron charge and 5 Å per alogP unit)

Method:	Extra dimensions:	Average enrichment at 1%:
ElectroShape 4-dim	MMFF94 (modified)	13.3
	alogP	11.9
	AM1 (partial charges)	10.8
ElectroShape 5-dim	MMFF94 (modified) and alogP	14.6

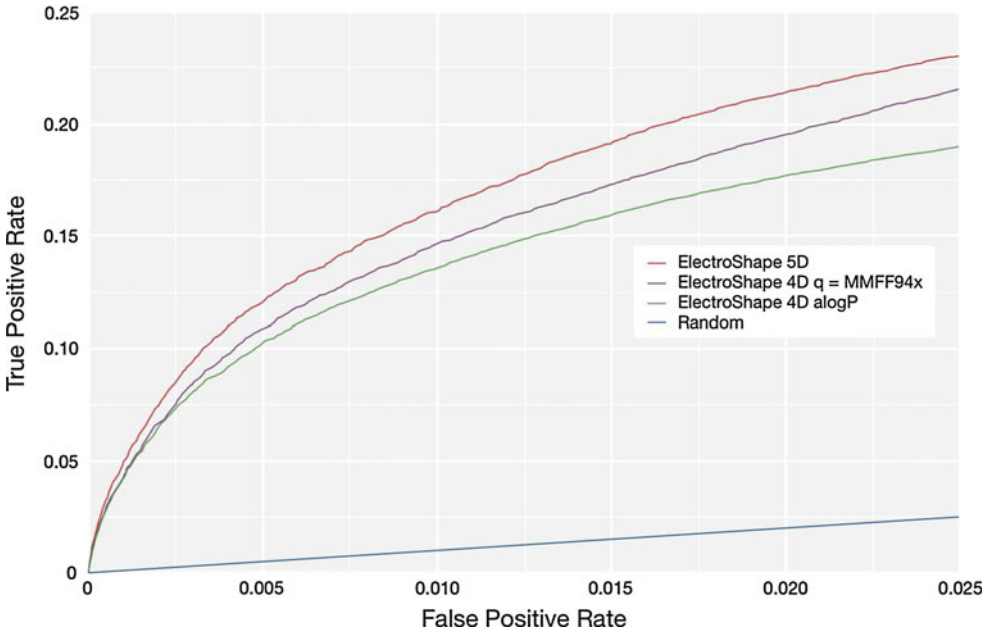

Fig. 3 Average ROC curves for four-dimensional ElectroShape (using partial charge—MMFF94 (modified) and AM1—at a ratio of 25Å per electron charge, and alogP, also at a ratio of 25Å per alogP unit) and five-dimensional ElectroShape (at a ratio of 25Å per electron charge and 5Å per alogP unit)

13.3 (the best previous enrichment) to 14.6. Interesting examples can be seen in the individual datasets. Datasets that perform well for both descriptors individually generally also perform well with the combined measure, but it is hard to predict the effect of the combination on specific cases, as there are examples where the combined enrichment is higher than, between, or lower than the individual scores. The COMT dataset, for example, shows a significant synergistic effect where a poor electrostatic enrichment (3.3) and modest lipophilicity enrichment (9.9) combine to give a very respectable combined value of 15.2.

The improved performance of the extended method is not simply a consequence of adding an additional dimension. Experiments done with redundant information—for instance, with two types of partial charge as fourth and fifth dimensions—did not result in a significant improvement in enrichment (data not shown). Partial charge and atomic lipophilicity are therefore to some extent complementary, and their enrichment results add rather than interfere with each other. This is because the alogP values and MMFF94 modified partial charges are relatively independent of each other—their correlation is 0.26. In comparison, the correlation between the MMFF94 modified charges and the AM1 and Gasteiger charges are 0.75 and 0.84 respectively.

Also of interest is the ROC curve for the three different methods. The three ROC curves are shown in Fig. 3; the improvements over the random line (shown in blue) are substantial. These graphs focus on the early part of the curve, as ElectroShape is of most interest in these early enrichments (later performance, after about a quarter of the database has been screened, becomes very close to random). This focus is justified because typical enrichment experiments are interested in the top few percentage results, not in the top quarter of the returned hits.

The improvement in performance is achieved with a minimal increase in execution times, either for calculation of the ElectroShape descriptors or for the similarity calculation itself.

Conclusion

The generality and wider applicability of ElectroShape, asserted in our previous work, has been substantiated through the demonstration of the ability of the alogP value to perform well as a replacement for partial charge in the fourth dimension, and the utility of combining these values into a five-dimensional method.

Further extension, including for example the addition of atom-based vectorial information, may be expected to lead to additional incremental gains in performance.

References

1. Armstrong MS, Morris GM, Finn PW, Sharma R, Moretti L, Cooper RI, Richards WG (2010) Electroshape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. J Comput Aided Mol Des 24(9):789–801
2. Ballester PJ, Finn PW, Richards WG (2009) Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. J Mol Graph Model 27:836–845
3. Ballester PJ, Richards WG (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28:1711–1723

4. Armstrong MS, Morris GM, Finn PW, Sharma R, Richards WG (2009) Molecular similarity including chirality. J Mol Graph Model 28:368–370
5. Wildman SA, Crippen GM (1999) Prediction of physiochemical parameters by atomic contributions. J Chem Inf Comput Sci 39:868–873
6. Chemical Computing Group, Montreal, Canada, MOE 2008.10, http://www.chemcomp.com
7. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801
8. Halgren TA (1996) The merck force field. J Comp Chem 17: 490–641
9. Dewar MJS, Zoebisch EG, Healy F, Stewart JJP (1985) AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107(13):3902–3909. doi:10.1021/ja00299a024