Optimisation Of Shape Fingerprints for protein-ligand systems

Joanna Zarnecka, Andrew Leach*, Steven Enoch, Mark Cronin, Al Dossetter*

School Of Pharmacy And Biomolecular Sciences, Liverpool John Moores University
*MedChemica Limited, Ebenezer House, Ryecroft, Newcastle-Under-Lyme, Staffs, ST5 2BE, England

Introduction And Aim

Molecules similar in size and shape are more likely to show similar activity towards the same target macromolecule, as implied by the concept often described as bioisosterism. The pharmaceutical industry uses this approach to improve potential drugs, seeking compounds similar in shape and size, in the hope that they show similar activity, but are different enough to have physicochemical properties that may improve drug action and reduce toxicity.

Shape fingerprints are binary bit strings that encode the shape of compounds. Their main advantage is the speed of the calculations and low storage needs. The shape is measured indirectly by alignment to a database of standard shapes, which consists of diverse shapes of molecules. The link between shape fingerprints and biological activity has never been demonstrated.

The aim of the study was to generate a database of reference shapes in order to link shape fingerprints with biological activity and to make it available online.

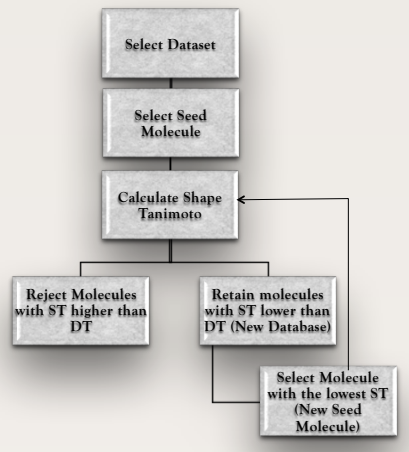
1000000100000000000000000000
0000001010

Methods

Generation of Database of reference shapes

The algorithm described by Haigh *et al.* [1] was implemented for the Ligand Expo (http://ligand-expo.rcsb.org) dataset to derive database of reference shapes. It consists of 812 031 experimental coordinates for non-polymer molecules and non-standard amino acids and nucleotides.

- Similarity between a randomly chosen seed molecule and all the molecules in the dataset was computed using Openeye Toolkit [3]. Shape Tanimoto was used as similarity coefficient (ST, ranging from 0 -1).



- Each molecule was either rejected if the calculated ST was higher than Design Tanimoto (DT, a user-defined threshold similarity) or retained if ST < DT.
- The molecule with the lowest ST was chosen as the next seed molecule and shape comparison was performed with the remaining molecules in the database.
- All the seed molecules were stored as our database of reference shapes.

Shape Fingerprint Generation And Comparison

The Tanimoto between each molecule and each reference shape in the database is computed. The bit corresponding to each reference shape is set On (1), if the Tanimoto is above a user-defined cut-off (the bit on value), or Off (0) if it is not.

Bit Strings were compared using Tanimoto as a similarity measure:

$$Sim_{Tanimoto} = \frac{bothAB}{onlyA + onlyB + bothAB}$$

Where onlyA and onlyB are the numbers of unique bits On in the bits strings for A and B respectively, while bothAB is the number of bits On in common to A and B.

Statistical Analysis

We binned the similarity Tanimoto values (with division for values inside and outside of the clusters) obtained in shape fingerprints comparison stage and performed logistic regression.

Conformations

Conformations were generated using OMEGA software with 5 as a maximum number of generated conformers. For all the conformations, shape fingerprints were generated and then all of them were compared with each other.

Testset

Protein Name	Abbreviation	No. of complexes
Protein kinase 5	PK5	2
Fatty acid binding protein	FABP	3
Neprilysin	NEP	4
Dihydrofolate reductase	DHFR	6
Checkpoint kinase	CHk1	16
Neuraminidase	NEU	11
Carbonic anhydrase	CA	13
Adenosine deaminase	ADA	11
Heat shock protein 90	HSP	10
Acetylcholinesterase	AChE	11

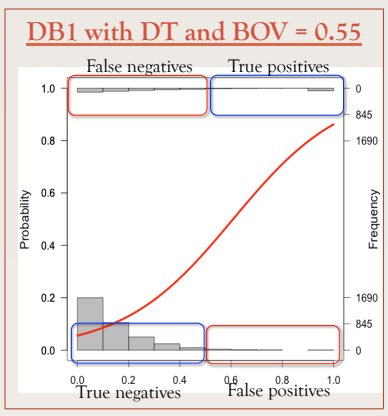
A test set described by Taylor *et al.* [2] (devised to test pharmacophore models) was used to evaluate the performance of different settings of shape fingerprints.
The ability to correctly group these sets is a simple test that a method can make useful connections between the shape of molecules and their biological activity.

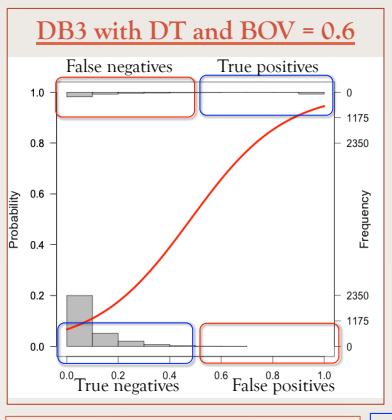
Results

Grouping crystal structures

We analysed the data by applying logistic regression, which links computed Tanimoto values with the likelihood of being inside a cluster (i.e. the probability of having shared biological activity).

The plot shows that shape can be used to group molecules that share biological activity if the bioactive conformation of a compound is known.

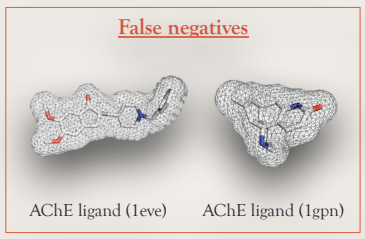


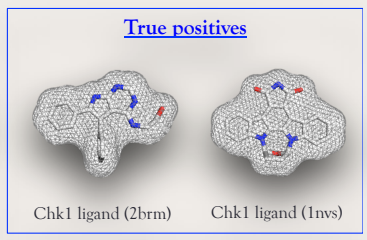


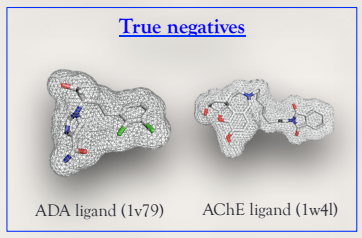
DT,BOV	AUC	
	DB1	DB3
0.55	0.714	0.672
0.60	0.708	0.706
0.65	0.685	0.688

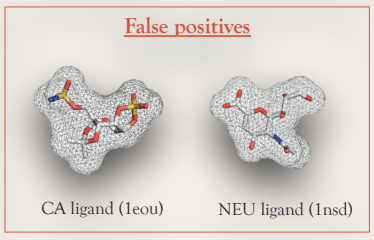
DB1_055 is the most accurate database, but DB3_06 is the fastest and almost as accurate.

Database (DT,BOV = 0.6)	AUC
DB1	0.708
DB2	0.705
DB3	0.706
DB4	0.696
DB5	0.685
DB6	0.702







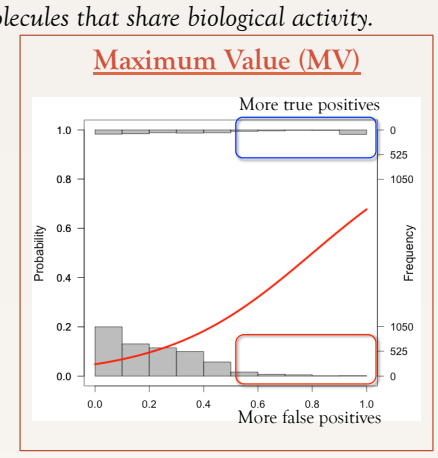


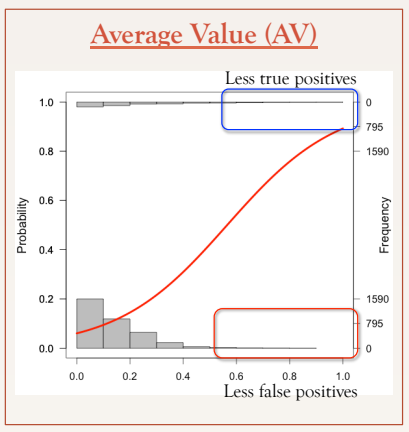
Grouping Conformations Generated From SMILES

In order to investigate the applicability of the method beyond crystal structures extracted from protein-ligand complexes, we generated conformations from SMILES.

Two approaches for evaluating the comparison of two molecules were investigated: 1) the highest value of similarity Tanimoto amongst the array arising from comparisons of all conformations of one molecule with all conformations of the other or 2) the average of those values (example plots shown for DB1_055)

If the bioactive conformation of a compound is not known, taking the average value is the best method to group molecules that share biological activity.





DB	Summary Method	AUC
DB1_0.55	MV	0.681
	AV	0.672
DB3_0.60	MV	0.665
	AV	0.674

While the MV method might lead to a higher number of false positives, using the AV method leads to reduction of the number of false positives, but might also lower the FT values of true positives.

Conclusions

The results show that DB1_055 is the most accurate database, DB3_06 is the fastest and almost as accurate (1380 vs. 232 shapes)

In future:

- Investigate the applicability of both databases,
- Use different dataset to maximize the shape fingerprints performance,
- Use more test sets
- Make the best Shape Database freely available

References

1. Haigh, J. A.; Pickup B. T.; Grant J.A.; Nicholls A. Small Molecule Shape-Fingerprints, J. Chem. Inf. Model., 2005, 45, 673–84.
2. Taylor, R.; Cole, J.C.; Cosgrove, D.A.; Gardiner E.J.; Gillet V.J.; Korb, O. Development and Validation of an Improved Algorithm for Overlaying Flexible Molecules, J. Comput-Aided Mol Des, 2012, 26, 451–72.
3. OpenEye Toolkits 2015. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

Acknowledgments

The project is conducted in collaboration with MedChemica Limited, Ebenezer House, Ryecroft, Newcastle-Under-Lyme, Staffs, ST5 2BE, England