Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT)*

Andreas Bender, Hamse Y. Mussa, Gurprem S. Gill, Robert C. Glen
Unilever Centre for Molecular Science Informatics
Department of Chemistry, University of Cambridge
Lensfield Road, Cambridge CB2 1EW, United Kingdom
rcg28@cam.ac.uk

Abstract - *A novel method (MOLPRINT) for virtual screening and the elucidation of ligand-receptor binding patterns is introduced which is based on environments of points on the molecular surface.*

In combination with the Tanimoto coefficient and applied to virtual screening, it achieves retrieval rates which are comparable to 2D fingerprints. In combination with information-gain based feature selection and a Naïve Bayesian Classifier, information from multiple molecules can be combined and classification performance can be improved. The descriptor uses points relative to the coordinates of the molecule which are uniformly binned, thus it is translationally and rotationally invariant. Due to its local nature the descriptor is conformationally tolerant. The identification of active structures with minimal 2D similarity is facilitated, commonly referred to as "scaffold hopping".

Features which are selected by the information-gain based feature selection step can be projected back on the molecular surface. They are shown to be consistent with experimentally determined binding patterns.

Keywords: Similarity searching, Virtual Screening, Drug Discovery, Binding Pattern Analysis, Molecular Fields, Surfaces.

1 Introduction

Molecular similarity searching relates differences between experimentally determined properties of a set of molecules to their differences in descriptor space, also known as "chemical space". The representation of compounds in chemical space can then be used to make predictions about properties of untested molecules.

The determination of similarity between two molecules generally involves generation of representative features for each molecule (which have to conserve as much relevant information as possible). Selection of those features deemed to be important may be performed (this step is optional). Finally the actual similarity metric is applied to define the proximity (similarity) or distance (dissimilarity) of molecules in descriptor space.

A variety of descriptors for molecular structures exist, which are commonly classified according to the dimensionality of data used to calculate them. One-dimensional descriptors use overall molecular properties such as logP[1], two-dimensional descriptors are derived from the connectivity table (such as topological indices[2] and fragment-based descriptors). Three-dimensional descriptors use spatial information such as three-point pharmacophores[3] and the well-known comparative field analysis (CoMFA)[4].

Descriptors which are invariant to both rotation and translation are known as TRI (Translationally and Rotationally Invariant) descriptors. Translational invariance can be achieved by using a coordinate system relative to the molecule and by centering the molecule with respect to it. Rotational invariance can be achieved by using distances between features instead of measuring coordinates in absolute space. This is the basis of autocorrelation approaches, which are well-known in both two dimensions[5] and three dimensions[6].

"Surface point environments", the descriptor introduced in this paper, are constructed in a three-step process. First, points on a "molecular surface" are computed. Second, interaction energies at surface points are calculated using hypothetical probes with varying parameters corresponding to different interaction types. Third, interaction energies are encoded into descriptors, encoding only local information about interaction profiles in binary presence/absence features. Note that the surface point environment descriptor is the surface point analogous of atom environments[7], using surface point interaction energies instead of different types of heavy atoms.

* **0-7803-8566-7/04/$20.00 © 2004 IEEE.**

Following the idea that most of the features calculated are (for our purposes) noise, a feature selection method is advisable. Here, we employ information-gain based feature selection as introduced by Quinlan[8]

If information from more than one molecule is given, the problem of merging information needs to be addressed. Similarity coefficients have a shortcoming in that they, by nature, are only able to deal with single fingerprints. A simple method to combine information from multiple molecules is to define minimum cutoff frequencies for a feature to enter the merged fingerprint.

Here we follow a different route in calculating similarity. In a fashion similar to binary kernel discrimination a type of data fusion is performed prior to scoring by using the Naïve Bayesian Classifier[9]. For comparison, a Tanimoto Coefficient is employed and used for database screening with single active molecules.

The method is validated using different sets of biologically active compounds from the MDDR[10] database. Features with high information gain are projected back on the molecular surface and shown to correlate with experimental binding patterns.

2 Material and Methods

2.1 Descriptor

A descriptor based on "surface point environments" has been developed. A unique surface environment descriptor of each individual point of the molecular surface is calculated (Figure 1), where the presence/absence of interaction energies up to a given layer depth are considered to construct the surface point environment vector (Figure 2).

Sets of surface point environments define the fingerprint of a molecule. The molecule is thus represented in terms of binary presence/absence interaction energy vectors an example of which is given in Figure 2.

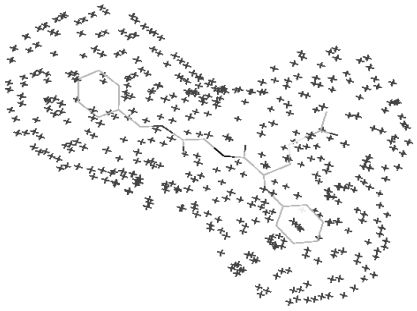

Figure 1 – Feature generation: creation of points on the molecular surface

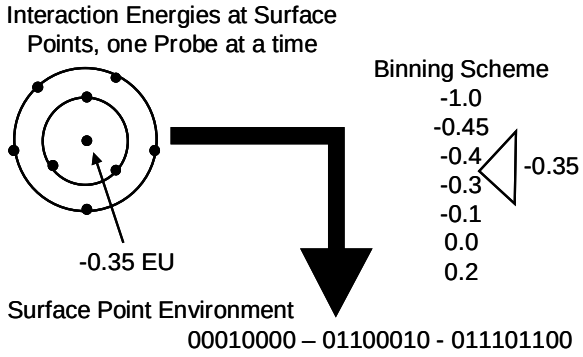

Figure 2 – Feature generation: determination of interaction energies and binning for descriptor generation

2.2 Feature Selection

The resulting fingerprints are subject to information-gain based feature selection[8], based on the entropy of the subsets that are partitioned with respect to the feature under consideration (where S is entropy, |S| is the total number of data sets, Sv is the entropy in data subset v, |Sv| is the number of data sets in subset v and p is the probability for a compound to belong to the class under consideration):

$$I = S - \sum_v \frac{|S_v|}{|S|} S_v \qquad (1)$$

$$S = -\sum p \log_2 p \qquad (2)$$

2.3 Classification

Scoring of compounds is performed using a Naïve Bayesian Classifier[9], which has shown to give results comparable to more sophisticated machine learning approaches in text classification (h_{act}/h_{inact} is the hypothesis for active and inactive compounds respectively, d_i is descriptor number i and P states a probability).

$$\frac{P(h_{act} \mid D)}{P(h_{inact} \mid D)} = \frac{P(h_{act}) \prod_i P(d_i \mid h_{act})}{P(h_{inact}) \prod_i P(d_i \mid h_{inact})} \qquad (3)$$

2.4 Compilation of Dataset and Preprocessing

The dataset used comprises 957 ligands[11] extracted from the MDDR[10] database. The set contains 49 5HT3 Receptor antagonists, 40 Angiotensin Converting Enzyme inhibitors, 111 3-Hydroxy-3-Methyl-Glutaryl-Coenzyme A Reductase inhibitors, 134 Platelet Activating Factor

antagonists and 49 Thromboxane A2 antagonists. An additional 574 compounds were selected randomly from the MDDR database and did not belong to any of these activity classes. A number of methods have been applied to this dataset[11] which enables us to compare similarity searching performance of our method to established algorithms.

2.5 Computational Details

10-fold random sets of 5 active molecules have been selected and the set of inactive molecules was used in a 50/50 split. Feature selection was set to select 200, 500 or 1000 features for each set of molecules. The hit rate among the ten highest ranked hits of the sorted library was calculated. The average hit rate for each class of active molecules (= the number of molecules among the top 10 structures belonging to the same activity class as the query structure) was then calculated from the ranked list of structures. Performance, defined as average hit rates, was compared for surface point environments calculated with a triangulation density of $0.5/Å^2$ and $2.0/Å^2$. For comparison, a Tanimoto coefficient was employed in combination with features derived from single molecules.

Three-dimensional descriptors always depend (to a varying degree) on the particular conformation of the molecule to be described, hence tolerance of the descriptor described here with respect to conformational changes was examined. 10 molecules from each of the five sets of active compounds were chosen randomly. Using the genetic algorithm conformational search in Sybyl[12] a set of 10 random conformations of each molecule was created. The window size was set to 10° in case of rigid 5HT3 ligands and 100° in case of all other datasets (ACE, HMG, PAF, TXA2), giving highly diverse conformations. Structures were optimized using the Tripos force field for 100 iterations to remove steric strain. All 10 conformations were put into the database containing "inactive" structures as well as all active structures from the five active datasets, excluding the query structure. The query was generated using Concord[13] and optimised using the Tripos force field for 100 iterations. All structures of the database were ranked according to Tanimoto similarity to the query structure. For a truly conformationally invariant descriptor, all ten conformations should occur on top of the sorted list because all descriptors were calculated for different conformations of the same structure. For a very sensitive descriptor, considerable spread throughout the database is expected. The number of different conformations of the query structure among the top 10, 20, 30, 40 and 50 positions of the sorted library was calculated to gauge conformational tolerance of the descriptor.

Finally it is likely that a method captures sensible features for classification (as opposed to randomly finding active compounds) if it performs well on several different data sets. By selecting features which are identified as being important for activity by the algorithm and projecting them back on the molecular surface it can be verified that they do not constitute incomprehensible sets of features which are only accidentally correlated with activity. These features are examined with respect to whether they correspond to experimental binding patterns. Surface fingerprint descriptors were calculated at point densities of $2.0/Å^2$ for all six interaction fields and using layers 0 – 4 for descriptor generation. Information gain feature selection was performed to select those features possessing highest information gain, which were more frequent in the set of active molecules. A number of selected features showing highest information gain were projected back onto the molecular surface. Examples of ligands of 3-hydroxy-3-methylglutaryl coenzyme A reductase and antagonists of Thromboxane A2 are given.

3 Results

Overall performance is compared to other methods in figure 3. Compared are atom environments with the Tanimoto coefficient, Feature Trees, surface fingerprints with the Bayesian classifier (as described in this work), atom environments with the Naïve Bayesian Classifier, ISIS MOLSKEYS, surface point environments with the Tanimoto coefficient (as described here), Daylight fingerprints, SYBYL Hologram QSAR, and three virtual affinity fingerprint methods: Flexsim-X, Flexsim-S and DOCKSIM. Performance of methods other than atom environments and surface point environments are taken from ref. 11 (where also references to all methods mentioned here are given).

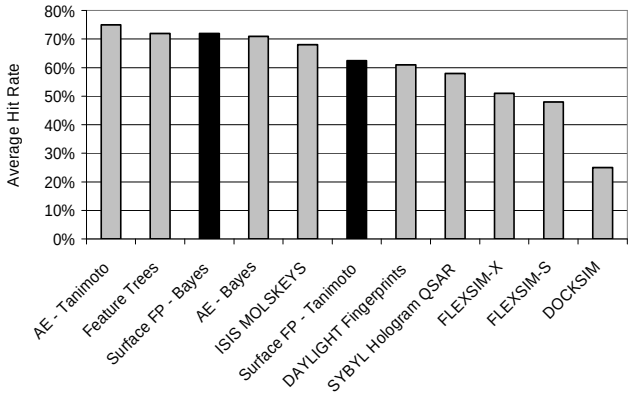

Figure 3 - Comparison of performance of surface point environments (black bars) with other commonly employed methods (grey bars)

Using surface point environments in combination with the Tanimoto coefficient, performance is comparable to 2D fingerprint methods such as Daylight fingerprints (if 0 – 4 layers are used for descriptor generation at a point density

of 2.0/Å^2). Combining information from multiple molecules using the Naïve Bayesian Classifier increases performance further, comparable to the best 2D methods (if layers 0 – 4 are used for descriptor generation and 200 features are selected at a point density of 0.5/Å^2). The influence of conformational variance on descriptor generation is given in table 1. Nearly two thirds (64%) of all conformations of the same molecule are identified as most similar by the Tanimoto coefficient (placed at the top 10 positions of the sorted list). 94% of all conformations are found in the top 50 positions (roughly 5%) of the sorted library.

Table 1 - Percentage of conformations found in the top n positions of the sorted database

Percentage of Conformations of the Same Structure Found at Top n Positions of the Whole Database						
n	5HT3	ACE	HMG	PAF	TXA2	Mean
10	70	69	75	56	50	64
20	85	87	91	81	70	82.8
30	89	94	94	90	78	89
40	90	96	96	93	88	92.6
50	90	97	96	95	92	94

Features identifying the putative pharmacophore of a Thromboxane A2 antagonist are shown in figure 4. Polar interactions of the carboxylic acid group on the left hand side, hydrogen bond acceptor potential of the sulfonamide moiety and the lipophilic interaction of the fluorobenzyl ring match binding patterns derived from homology models of the binding site. The bound conformation of the ligand is likely to be bent [14] at an angle of about 90 degrees so that the lipophilic rings points downwards.

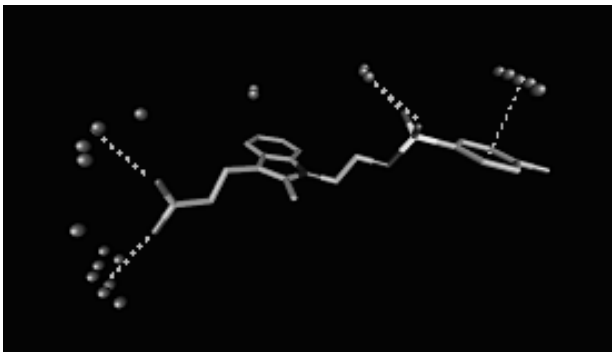

Figure 4 - Putative pharmacophore of a Thromboxane A2 antagonist

Features identifying the putative pharmacophore of a 3-hydroxyl-3-methylglutaryl-CoenzymeA reductase inhibitor are shown in figure 5. The polar interactions in the upper left corner and the lipophilic interaction of the fluorobenzyl moiety match binding patterns observed in crystal structures of other HMG-CoA inhibitors [15].

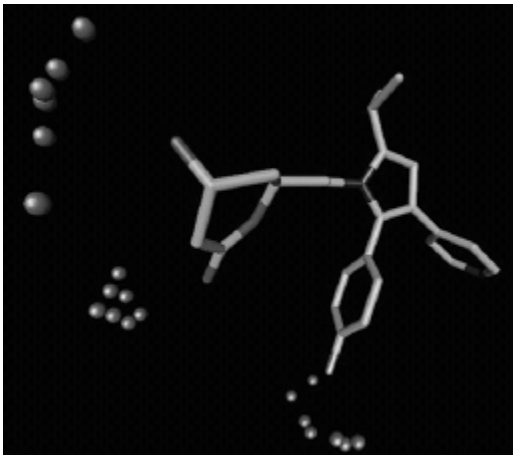

Figure 5 - Putative pharmacophore of a HMG-CoA reductase inhibitor

4 Discussion

The superiority of the two-dimensional descriptor compared to its three-dimensional analogue purely with respect to hit rates is in agreement with earlier findings[16]. Still, in relative numbers, surface point environments retrieve only 17% and 24% less active compounds than atom environments, which still compares favorably with a number of 2D methods, as given in Fig. 3. One of the reasons for that may be conformational tolerance of this descriptor, as discussed in detail below. Other 3D descriptors, which employ overall distance information between pharmacophores (be it surface points or atom centered pharmacophores) change considerably if the descriptor is calculated for multiple conformations, while this descriptor is reasonably tolerant to conformational changes. The performance of surface point environments in combination with the Tanimoto coefficient is slightly higher if finer surface point densities of 2.0/Å^2 are employed, with hit rates of on average 6.2 structures among the ten highest ranked compounds vs. on average 6.1 hits if a point density of 0.5/Å^2 is used. Finer point spacing may be better at capturing local properties, although differences are minimal. Overall, significant enrichment is observed for each of the point densities chosen above.

The influence of conformational variance on descriptor generation is given in table 1. Nearly two thirds (64%) of all conformations of the same molecule are identified as identical by the Tanimoto coefficient (placed at the top 10 positions of the sorted list) while 94% of all conformations are found in the top 50 positions (roughly 5%) of the sorted library. If a molecule that is similar to the query molecule is present in the database, it is likely to be ranked at the top of the sorted database. This leads to the tentative conclusion (based on the five different datasets and diverse sets of conformations employed here) that the descriptor is unlikely to miss an active molecule when it is not present in the "correct" conformation (e.g. the binding conformation or any other pharmacophoric conformation) in the database.

The binding pattern of a TXA2 antagonist (figure 4) is compared to a ligand-target complex employing homology modelling. An arginine residue of Thromboxane A2 is thought to form a charge interaction with a carboxylic acid group of the ligand. A serine residue from the target in this model forms a hydrogen-bond interaction, where a hydroxyl group of the ligand acts as an acceptor. In addition, a large lipophilic pocket is present perpendicular to the arginine-serine axis. All three features, carboxylic acid hydrogen bond acceptor (in this case a sulfonamide group) and the fluorobenze which points in the lipophilic pocket, are identified correctly by the algorithm presented here. This is achieved without having the binding conformation of the ligand available, which is more likely to be bent (the C-C bond between the sulfonamide and indole moiety can be rotated by 180 degrees at both carbon atoms to achieve a bent conformation)[14].

Figure 5 shows features selected to be characteristic for binding as illustrated on a 3-hydroxyl-3-methylglutaryl-CoenzymeA reductase inhibitor. Crystal structures of HMG-CoA reductase complexed with statins[15] show a common binding pattern between the carboxylic acid and the hydroxyl groups of the HMG moiety and polar sidechains of the protein. In addition a lipophilic cleft perpendicular to the axis of polar interactions is present, which is surrounded by a flexible a helix that is able to accommodate lipophilic groups of different shapes and sizes. Both features, the oxygen atoms corresponding to the polar interactions of the HMG moiety and the lipophilic fluorobenzene, are identified by the algorithm.

5 Conclusions

We present a novel similarity searching algorithm based on surface point environment descriptors in combination with the Tanimoto coefficient and the Naïve Bayesian Classifier. The descriptor is shown to be tolerant to conformational variations of the ligand structure. It exhibits high retrieval rates, the identification of active structures with different scaffolds and back-projectability of features which can be correlated with experimentally determined binding patterns. Used in combination with Tanimoto coefficients, its performance is comparable to that of commonly used 2D fingerprints. If the Tanimoto coefficient is replaced by a Naïve Bayesian Classifier, information from multiple structures can be combined which is shown to improve classification performance.

Acknowledgements

We thank Unilever, The Gates Cambridge Trust and Tripos Inc. for support. Uta Lessel is thanked for providing us with the MDDR data set. A number of participants at the Joint Sheffield Conference on Chemoinformatics 2004 are thanked for input on the method.

References

[1] G.M. Downs, P. Willett, and W. Fisanick, "Similarity searching and clustering of chemical-structure databases using molecular property data", *J. Chem. Inf. Comput. Sci.*, Vol. 34, pp. 1094-1102, 1994.

[2] E. Estrada, and E. Uriarte, "Recent Advances on the Role of Topological Indices in Drug Discovery Research" *Curr. Med. Chem.*, Vol. 8, pp. 1573-1588, 2001.

[3] J.S. Mason, A.C. Good, and E.J. Martin, "3D-Pharmacophores in Drug Discovery", *Curr. Pharm. Des.*, Vol. 7, pp. 567-597, 2001.

[4] R.D. Cramer, D.R. Patterson, and J.D. Bunce, "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins", *J. Am. Chem. Soc.*, Vol. 110, pp. 5959-5967, 1988.

[5] G. Moreau, and P. Broto, "Autocorrelation of a topological structure: A new molecular descriptor", *Nouv. J. Chim.*, Vol. 4, pp. 359 – 360, 1980.

[6] M. Wagener, J. Sadowski, and J. Gasteiger, "Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks", *J. Am. Chem. Soc.*, Vol. 117, pp. 7769-7775, 1995.

[7] A. Bender, H.Y. Mussa, R.C. Glen, and S. Reiling, "Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier", *J. Chem. Inf. Comput. Sci.*, Vol. 44, pp. 170 – 178, 2004.

[8] J.R. Quinlan, "Induction of Decision Trees", *Machine Learning*, Vol. 1, pp. 81-106, 1986.

[9] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[10] MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.

[11] H. Briem, and U.F. Lessel, "In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes", *Perspect. Drug Discov. Des.*, Vol. 20, 231-244, 2000.

[12] SYBYL, Version 6.5.3, Tripos Inc., St. Louis, Minnesota, USA.

[13] R.S. Pearlman, "CONCORD: Rapid Generation of High Quality Approximate 3D Molecular Structures.", *Chem. Design Automation News*, Vol. 2, pp. 5-7, 1987.

[14] Y. Yamamoto, K. Kamiya, and S. Terao, "Modeling of human thromboxane A2 receptor and analysis of the receptor-ligand interaction", *J. Med. Chem.*, Vol. 36, pp. 820 – 825, 1993.

[15] E.S. Istvan, "Structural mechanism for statin inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase", *American Heart Journal*, Vol. 6, pp. S27 – S32, 2002.

[16] R.D. Brown, and Y.C. Martin, "The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding", *J. Chem. Inf. Comput. Sci.*, Vol. 37, pp. 1 – 9, 1997.