Variable selection and model validation of 2D and 3D molecular descriptors

Anthony Nicholls\textsuperscript{a,}\textsuperscript{*}, Norah E. MacCuish\textsuperscript{b} \& John D. MacCuish\textsuperscript{b}
\textsuperscript{a}OpenEye Scientific Software, Inc., 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507, USA; \textsuperscript{b}Mesa Analytics \& Computing, LLC, 212 Corona St., Santa Fe, NM 87501, USA

Received 4 May 2004; accepted in revised form 26 August 2004
© Springer 2005

\textbf{Key words:} cluster analysis, electrostatics, field overlaps, hypothesis generation, \textit{k}-fold cross-validation, molecular shape, QSAR, structural fingerprints, Tversky

\textbf{Summary}

We have found that molecular shape and electrostatics, in conjunction with 2D structural fingerprints, are important variables in discriminating classes of active and inactive compounds. The subject of this paper is how to explore the selection of these variables and identify their relative importance in quantitative structure–activity relationships (QSAR) analysis. We show the use of these variables in a form of similarity searching with respect to a crystal structure of a known bound ligand. This analysis is then validated through \textit{k}-fold cross-validation of enrichments via several common classifiers. Additionally, we show an effective methodology using the variables in hypothesis generation; namely, when the crystal structure of a bound ligand is not known.

\textbf{Introduction}

Quantitative structure–activity relationships (QSAR) analysis assumes a relationship between the chemical properties of a molecule and its physical behavior, such as binding to a target protein. As a result, researchers have tried to encode physicochemical properties and structure in so-called molecular descriptors. Finding a combination of such descriptors or variables that best reveals activity relationships across a wide range of ligand/target complexes is an area of ongoing research. This problem is confounded by the fact that ligand/target complexes can have numerous binding modes. It is well known that 2D chemical structure, molecular shape, and electrostatics all play a significant role in the binding of ligands to targets for most complexes, and should therefore correlate with the activity response found in screening compounds. Hence shape, electrostatic, and 2D structural descriptors that are robust and efficient should in principle provide the necessary ease of use and discriminatory power to fully reveal activity relationships.

Two-dimensional structural fingerprint methods [1–3] and chemical properties [4,5] are manifold, but researchers have also created many forms of 3D descriptors containing shape and electrostatic information and many methods of using such variables [6–8]. All have been used in QSAR analysis. Difficulties with these methods have been various. For instance, CoMFA does not generalize well beyond single chemical series and the alignment of active molecules can affect the accuracy. Pharmacophore and pharmacophore-like methods have no formal theory of comparison; i.e., there is no good way of understanding the statistical significance of a given model. In addition, pharmacophoric elements require ‘weightings’ that have no physical basis and hence are not easily transferable. Other methods (BCUTS, surface SOMs) have much the same reduction of 3D information found in 2D molecular fingerprints, and are thus incomplete in their molecular
description. In general, 3D methods are slow and limited in the scale of problems they can address. 3D methods also have particular problems, most notably the multiple conformer issue. Aligning shapes efficiently for comparison or electrostatic analysis is problematic. Additional problems can arise due to incorrect tautomers, ionization states, or mistakes in atom typing due to incomplete or incorrect chemistry perception. The latter can be particularly acute in the perception of aromaticity. The authors of the prediction method xLog P [9] considered furan non-aromatic and had a significant discrepancy with the Log P from experiment. This discrepancy disappears if furan is more appropriately considered aromatic. However, formal descriptions of the shape of molecular conformers that have been refined with modern force-fields, set with high-quality charges, utilizing a method of accurately calculating the potentials in solution, should have a decided advantage over previous approaches because these are the variables that are intimate to molecular interaction. Using these descriptors and their respective comparative measures with a drug-like, feature-based, 2D fingerprint such as the Molecular Design Limited 320 keys based fingerprint [10] presents a formidable combination of three relatively orthogonal variables for discrimination, or regress, on activity.

A crucial aspect that separates this work from others is that it is based on formal theory of molecular shape comparison. This theory is general, extensible and parameter free. It needs no tuning for each new application, as it requires no training. It shows that shape, as defined in this paper, is a fundamental molecular property and that shape difference forms a metric space. Lack of a theory of this nature need not negate the usefulness of a 3D method – heuristics and hypothesis generation are common enough in chemistry – but it limits the transferability, the reliability, and the generalization of the method, for instance, to other physical phenomena. In particular, when a heuristic fails there is no mechanism of assigning value to this failure, i.e., is this just a singleton not caught by the ad hoc collection of rules, or does the failure say something about the model and how it is being applied? What separates a theory from a heuristic is the value of negative information. Failures need to mean something with respect to the theory. We show several examples of ‘good’ failures that provided useful biophysical information in the Results section.

There are several important issues to settle in developing useful models with shape and electrostatics. The first is to reliably construct the conformer search space with a robust measure of internal energy: high-energy conformers are unlikely to contribute meaningful information and can swamp a signal from low-energy shapes. The second is to quantify and ameliorate the sensitivity of the electrostatic similarity to shape alignment. The third is to design an optimal combination of shape and electrostatic comparisons. With effective methods for these issues in hand, shape and electrostatic descriptors can address two forms of traditional QSAR analysis. The first is similarity searching to a bound crystal ligand so as to produce a predictive model; the second is hypothesis generation without structural data, e.g. assay data alone.

In the Theoretical basis section, the shape, electrostatic, and 2D MDLI descriptors, and their respective measures, will be discussed in detail. An overview of the forms of unsupervised and supervised learning techniques, used in the similarity searching and hypothesis generation, will be addressed. Methods for conformer search space, conformation and electrostatic matching, and the optimal association between conformers and electrostatic comparisons will be presented in the Experimental methods section, as are overall protocols for the use of shape and electrostatic descriptors and measures used in similarity searching and hypothesis generation. Finally, similarity searching results for Cox2 and Progesterone, and the hypothesis generation for the assay data of Dopamine and Calcium ion channel are presented in the Results section.

Theoretical basis

Shape

The theoretical basis for molecular shape analysis in this work derives from the concept of volume overlap, as first applied by Masek et al. [11]. Molecules are typically represented visually and hence conceptually as a set of N overlapping hard spheres.