A Comprehensive Comparison of Ligand-Based Virtual Screening Methods Against the DUD Dataset

Vishwesh Venkatraman, Violeta I. Pérez-Nueno, Lazaros Mavridis, David W. Ritchie
INRIA Nancy - Grand Est, LORIA, 615 rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France

INTRODUCTION

The goal of ligand-based virtual screening (VS) is to search chemical databases to find compounds which best match a given query. Different VS tools are usually compared by assessing their ability to distinguish known active molecules from a large number of inactive compounds or " decoys " in a database. In recent years, many different VS tools have been developed, and these often employ different representations of molecular properties and have different speed and accuracy characteristics. Hence there is a need to perform an objective comparison of currently available VS tools.

Here, we compare ten popular ligand-based VS tools, five of which are based on 2D chemical fingerprint representations, and five of which use 3D molecular shape-based representations. These methods were evaluated using the publicly available Database of Useful Decoys (DUD) dataset 1 comprising over 100,000 compounds distributed across 40 pharmaceutically relevant targets. In our evaluation, we measured the ability of each scoring function to identify the known active molecules for each target when using the bond crystallographic structure of the corresponding ligand as the query.

Although it is conventional wisdom that 3D molecular shape is an important determinant of biological activity, our results using a number of metrics show that the 2D fingerprint-based methods out-perform the 3D shape-based approaches for surprisingly many of the DUD targets. In order to help understand this unexpected finding, we present an analysis of the nature of the scoring functions tested and of the composition of the DUD dataset itself.

METHODS

The following VS tools were compared:

- 2D – OPENBABEL, DAYLIGHT ™, MACCS. BCI, MOLPRINT2D ™
- 3D – ROC5 ™, PARAFIT ™, USP ™, SHAEP ™, ESHAPE3D ™

These methods were evaluated using the publicly available DUD dataset downloaded from http://dud.docking.org/r2. In this study, the known actives and the target-specific decoys were used to compare the selected 2D and 3D ligand-based methods. For each target, the crystallographic ligand conformation was used as the query.

The VS results have been evaluated using commonly used metrics that include: the area under the ROC curve (AUC) as a measure of the overall performance and those that emphasize early recognition such as BEDROC ™, BAROC ™, early AUC (5%, 10%) and the Normalised sum of rank log (NSLR). The VS performance of all methods for this target is considered significantly different if the calculated p-value is less than 0.05 (95% confidence level). Figure 1 shows the AUC values for each scoring method for each of the 40 DUD targets in which A) the AUC is calculated for each target-specific set of actives and decoys, and B) the AUC is calculated using approximately 120,000 non target-specific decoys for each target.

In order to evaluate how similar the decoys are to the actives for each target, the decoy set for each target was assembled from the decoys of the remaining 39 targets (~120,000 decoys). The relative performance of the VS methods can also be appreciated from the AUC " heat map " shown in Figure 3. It can be seen that the overall VS performance of both the 2D and 3D methods improves for essentially all of the targets except comt (which has only 11 actives) and trypsin. This confirms that the target-specific DUD decoy sets are in fact very well constructed.

RESULTS

Figure 1. Aggregate ROC plots for the 40 DUD targets (LEFT). Bar chart of aggregate AUC values for the overall curve (grey), the top 10% (dark grey) and 5% (black) (RIGHT).

Figure 2. Heat map plots showing the AUC for each scoring method for each of the 40 DUD targets in which A) the AUC is calculated for each target-specific set of actives and decoys, and B) the AUC is calculated using approximately 120,000 non target-specific decoys for each target.

ANALYSIS OF SELECTED TARGETS

Here, we present an analysis of the nature of the scoring functions tested and of the composition of the DUD dataset itself.

a) COX2

The physical properties of the query are generally similar to those of the actives, presumably because all the actives are derived from a central scaffold. Analogue bias in this case is beneficial with all methods except ESHAPE3D achieving high retrieval rates (see Figure 3A).

b) TRYPsin

The physicochemical properties of the crystallographic query differ considerably from those of the actives and decoys. These differences are mainly due to the absence of a sulphonamide group in the query but present in many actives and decoys. The screening performance of all methods for this target is therefore very low (see Figure 3B).

OPERATIONAL CORRELATION ANALYSIS

A permutation test was used to compare the different ranking methods. Two ranking methods are considered significantly different if the calculated p-value is less than 0.05 (95% confidence level). Figure 4 shows the ROC curves comparing the performance of the 2D and 3D methods. Each spoke or radial line represents a method. Each colour-coded curve also corresponds to a ranking method. The intersection of a curve and a spoke shows the number of targets for which the spoke method gives significantly better VS performance than the curve method, as defined by the permutation test.

Figure 3. PARAFIT superpositions for the top-ranking cox2 active and trypsin ligands with and without the sulphonamide group. The covalent structure and surface of the query is shown using a blue-red gradient and the superposed structures are coloured according to the atom type: nitrogen-blue, oxygen-red and carbon-green. Bar charts summarising performances of the methods on the two targets are shown for four early recognition metrics.

TRYPsin The physicochemical properties of the crystallographic query differ considerably from those of the actives and decoys. These differences are mainly due to the absence of a sulphonamide group in the query but present in many actives and decoys. The screening performance of all methods for this target is therefore very low (see Figure 3B).

CONCLUSIONS

To our knowledge, this is the first comprehensive evaluation of ligand-based tools using this dataset. Although the validity of using the DUD as a VS benchmark has been questioned, operational correlation analysis results show that the DUD is in fact well suited for ligand-based VS. It is found that 2D fingerprint-based methods give better VS performance than the 3D shape-based approaches for many of the DUD targets. Using multiple database conformations does not improve the results appreciably (data not shown). We propose that in order to improve the VS performance of current 3D methods, it will be necessary to devise screening queries which can represent multiple possible conformations.

ACKNOWLEDGEMENTS

We thank OpenEye Scientific Software Inc., Cepos Insilico Ltd., Chemical Computing Group, DAYLIGHT, Chemical Information Systems and Digital Chemistry for providing academic, licenses for ROC5 ™, ROC5 ™, MACCS, BCI, MOLPRINT2D ™, 3D – ROC5 ™, PARAFIT ™, USP ™, SHAEP ™, ESHAPE3D ™. We acknowledge financial support from grants from the programme " Redes " of the " Consell de l'lnvesigador " del " CSIC " and "infoerca " of the " FAM " of the " UAB " (2006-0177).