Computational Biology and Chemistry 62 (2016) 133–144

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

Machine learning optimization of cross docking accuracy

Esben J. Bjerrum

Wildcard Pharmaceutical Consulting, Frødings Alle 41, 2860 Søborg, Denmark

ARTICLE INFO

Article history:
Received 11 January 2016
Received in revised form 8 April 2016
Accepted 9 April 2016
Available online 4 May 2016

Keywords:
Molecular docking
Docking power
Scoring function
Machine learning optimization
Smina
Autodock Vina
Drug discovery
Cross docking

ABSTRACT

Performance of small molecule automated docking programs has conceptually been divided into docking -, scoring -, ranking - and screening power, which focuses on the crystal pose prediction, affinity prediction, ligand ranking and database screening capabilities of the docking program, respectively. Benchmarks show that different docking programs can excel in individual benchmarks which suggests that the scoring function employed by the programs can be optimized for a particular task. Here the scoring function of Smina is re-optimized towards enhancing the docking power using a supervised machine learning approach and a manually curated database of ligands and cross docking receptor pairs. The optimization method does not need associated binding data for the receptor-ligand examples used in the data set and works with small train sets. The re-optimization of the weights for the scoring function results in a similar docking performance with regard to docking power towards a cross docking test set. A ligand decoy based benchmark indicates a better discrimination between poses with high and low RMSD. The reported parameters for Smina are compatible with Autodock Vina and represent ready-to-use alternative parameters for researchers who aim at pose prediction rather than affinity prediction.

1. Introduction

Automated ligand docking to receptor models is an important method in structure based molecular design and structural bioinformatics. Docking algorithms aim to predict the preferred conformations and binding affinities of small molecule ligands given a receptor model. Docking of small molecule ligands for receptor targets has been used in a variety of studies, ranging from *in silico* screening of large compound databases to predicting affinity of novel compound designs in lead optimization. Usually the docking algorithms use an optimization search strategy that generates proposed conformations of the ligand in the receptor in combination with a scoring function that evaluates how the ligand fits in the receptor binding pocket (Meng et al., 2011).

Performance of docking programs has conceptually been divided into docking power, scoring power, ranking power and screening power (Li et al., 2014a) to probe the usefulness of the docking program for different applications such as crystal pose prediction, affinity prediction, ligand ranking and virtual screening, respectively (Wang et al., 2004a).

Docking power is focused on the programs ability to predict the correct conformation of the ligand and the placement in the receptor molecules (the pose). It can be evaluated by docking ligands to receptors with known binding mode of the ligands and comparing the result with the already known experimentally determined binding mode. The docked poses are usually evaluated by computing the root mean square deviation (RMSD) between the heavy atoms of the docked ligand and native crystal pose. The average RMSD can be used, but often a threshold of 2 Å RMSD is used to calculate the fraction of correct predictions (Cheng et al., 2009; Xu et al., 2015).

The scoring, ranking and screening power focuses on the affinity prediction, ligand series ranking and database screening performance, respectively. Scoring power is usually benchmarked by calculating the correlation between predicted and measured affinities for large databases such as the PDBbind (Wang et al., 2004b; Wang et al., 2005) database (Li et al., 2014a). Ranking power represents a variant of the scoring power where the focus is on the ranking of the ligand series (Plewczynski et al., 2011) and screening power is measured by the ability to identify known binders seeded in large databases of decoys by calculating the area under the receiver operator curves (ROC) (Li et al., 2014a; Christofferson and Huang, 2011; Bauer et al., 2013).

The differences in benchmarks reflect the versatility of the *in silico* docking approach, but also they emphasize that evaluation of the docking programs is linked to the intended use. For example, the medicinal chemist who wants to use bio-structural inspiration for lead optimization, may not be interested in the affinity prediction as the experimental data are already available, but

E-mail address: esbenbjerrum@gmail.com (E.J. Bjerrum).

http://dx.doi.org/10.1016/j.compbiolchem.2016.04.005
1476-9271/

instead care very much about the accuracy of the predicted pose. In an *in silico* compound library screening program, the ability of the docking algorithm to distinguish binders from non-binders is important, whereas the affinity prediction will be the important aspect for a medicinal chemist who wants to examine the possible effect of a small structural change to a lead compound with known binding mode.

With regard to evaluation of docking power there is an additional aspect to be taken into account. Either the ligand can be re-docked into the receptor X-ray structure where the ligand binding mode was determined or docked into a receptor model from an X-ray structure determined with a different co-crystallized ligand (Cross docking). The last procedure reflects the real world use and has been used as a benchmark for docking accuracy (Morris et al., 2009; Liu et al., 2012; Forli and Olson, 2012; Stigliani et al., 2012).

A widely used docking program is Autodock Vina (Trott and Olson, 2010) programmed as an alternative to Autodock (Morris et al., 2009). Autodock Vina employs a docking algorithm and scoring function different than Autodock, but is compatible with the file formats used in Autodock and uses the same tools to prepare receptor and ligand files. In comparison with the original Autodock, Autodock Vina was shown to be much faster and with a large improvement of docking accuracy for compounds with large conformational freedom in the form of torsions (Trott and Olson, 2010). It is unclear precisely how the weights were determined for the original scoring function in Autodock Vina, except that it was tuned against PDBind (Wang et al., 2004b; Wang et al., 2005). Autodock Vina uses an empirical scoring function to dock and to predict the affinity of the ligand. Five terms, gauss1, gauss2, repulsion, hydrophobic and hydrogen bond, analyzing pairwise atom interactions are scaled and linear combined, whereas a last term, Nrot, is combined in a non-linear fashion (Trott and Olson, 2010). The source code for Autodock Vina was released in 2010, and has been forked and modified by others. QuickVina2 (Alhossary et al., 2015) focuses on improving computational efficiency, whereas Smina (Koes et al., 2013) has exposed already existing experimental score terms and added convenience functions such as automatic calculation of the docking box from a reference ligand. VinaLC (Zhang et al., 2013) added multi-threading and message passing interface (MPI) support for use at high performance clusters.

When ranking docking programs and the scoring functions evaluating the different "powers", no single program is a clear winner amongst all benchmark types. For example GOLD (Verdonk et al., 2003) has been found to be one of the best in benchmarks for pose prediction (Li et al., 2014a), whereas Schrödingers Glide (Friesner et al., 2004; Halgren et al., 2004) has performed better in database enrichment studies and affinity prediction (Li et al., 2014a). It has been noted that no single scoring function outperforms others in the benchmarking of the different aspects of docking performance (Cheng et al., 2009). Recent efforts in optimizing binding affinity prediction with the use of non-linear machine learning models such as random forests (Zilian and Sotriffer, 2013; Li et al., 2015; Fourches et al., 2015) or support vector machines (Kinnings et al., 2011a), have led to improvements in affinity prediction. As an example, the RF-Score-v3 (Li et al., 2015) nearly doubles the Pearsson's correlation coefficient in comparison with GlideScore-XP (Friesner et al., 2006). However, the improvements in affinity prediction seem unrelated to improvements in the docking performance, an effect which has also been noted in previous benchmarks of docking and scoring performance (Warren et al., 2006). This seemingly counter intuitive divergence between the various "powers" is the reason why the machine learning approach has been criticized (Gabel et al., 2014). However, machine learning approaches have on several occasions outperformed more classical scoring functions in a variety of benchmark experiments (Kinnings et al., 2011b; Li et al., 2011; Ballester et al., 2012; Ding et al., 2013; Durrant et al., 2013). Recently these differences in scoring performance have been reviewed in depth (Ain et al., 2015).

Taken together, these observations indicate that docking performance can be tuned for different tasks. The different approaches and the scoring functions optimal for the different tasks may have some overlap as illustrated in Fig. 1, but the extent of this overlap is not known in detail.

Generally speaking, machine learning is the process of getting computers to perform actions without being explicitly programmed. In supervised machine learning, the computer algorithm is presented with a task (here docking) and the results compared with the provided correct results. The error is formulated mathematically in a loss or cost function which calculates how wrong the outcome is. The terms and parameters of the algorithm are then stepwise optimized to minimize the loss function (Mohri et al., 2012). In this study a machine learning approach is used to re-optimize the weights for the scoring function of Smina (Koes et al., 2013) using a loss function calculated from the pose prediction accuracy of a small curated cross docking data set.

2. Computational methods

2.1. Train and test sets

A database of receptor-ligand complexes for cross docking was manually curated using the protein data bank (PDB) id's from the Astex diverse data set (Hartshorn et al., 2007) with known binding affinities determined as Ki/Kd in the Binding MOAD (Hu et al., 2005). All the PDB files matching the UniProtID of the receptor from the Astex set were downloaded from the protein data bank and structurally aligned. Hydrogen atoms were added and the X-ray model split into receptor, water molecules, co-factors and

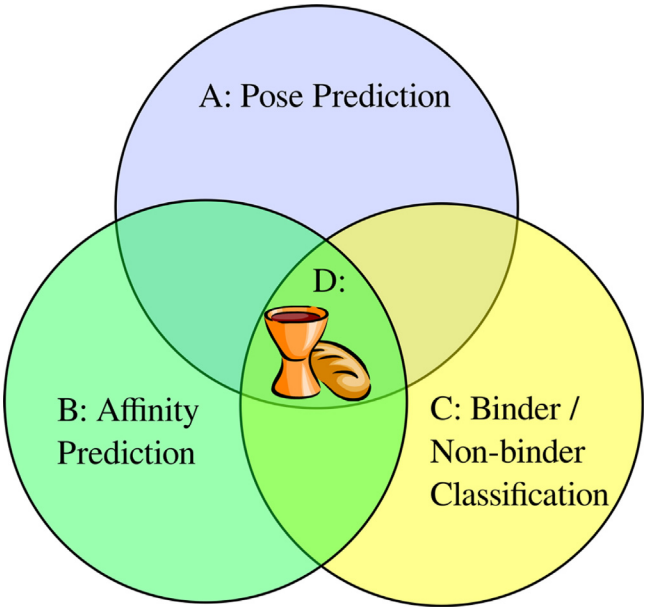

Fig. 1. Aims of docking and scoring. A: Binding mode prediction as it will be used in docking and lead optimization studies. B: Affinity Prediction as it will be used in the assessment of proposed novel compounds. C: Binding classification for the use in *in silico* based virtual library screening and development of focused libraries for high throughput screening. D: The Holy Grail of scoring functions: A scoring function that can do all three things optimal.