Challenging different scoring schemes in structure-based virtual screening

Malgorzata N. Drwal and Esther Kellenberger

Laboratoire d'innovation thérapeutique, Université de Strasbourg, 67400 Illkirch, France

Context:

The time-saving and cost-effective virtual screening approach has become an important tool in drug design. The literature of the last decade reports an important number of novel bioactive compounds discovered by structure-based approaches. With the rapid growth of structural information on the proteome, molecular docking has naturally emerged as one of the standard methods to evaluate large compounds collections on a potential protein target [1,2].

While the method efficiently identifies lowly potent hits from crystal structure of the protein target, the hit rate highly depends on the computational strategy for input preparation and data post-processing. Generally, a good knowledge of the protein properties (protonation state, local flexibility, hot spots) drastically increases the chance to prioritize true hits.

Objective and methods:

The tutorial focuses on the issue of scoring, which after 20 years of development, still remains the Achilles heel of the method [3]. To that aim, we will evaluate the ability of the program *PLANTS* to discriminate the active compounds from decoys in the *DUD-e* datasets for two protein kinases. Both the programs and benchmark are freely available [4,5].

We will analyze screening data comparing compound ranking based on different scoring schemes: (a) docking score, (b) docking score weighted by compound molecular weight, (c) similarity to the reference binding mode -in the crystallographic structure-, as evaluated by interaction fingerprint (IFP) and (d) docking score after filtering of poses based on IFP [6,7]. Next we will compare results to a control screening, where the *DUD-e* compounds are docked into an orthogonal protein pocket, with no sequence, structure or functional similarities to the true target. Such "decoy" pockets have been recently suggested to help docking score correction [8].

Conclusion:

The screening performance is target dependent. Filtering hit lists based on binding mode can limit the number of false positives in hit list. Surprisingly, docking in one of the "decoy" pocket can also yield to enrichment in kinase ligands in hit list.

Bibliography:

1. Ferreira L, dos Santos R, Oliva G, Andricopulo A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*. 20(7), 13384–13421 (2015).
2. Yuriev E, Holien J, Ramsland PA. Improvements, trends, and new ideas in molecular docking: 2012-2013 in review: Improvements, Trends, and New Ideas in Molecular Docking. *J. Mol. Recognit.* 28(10), 581–604 (2015).
3. Sotriffer C, Matter H. The Challenge of Affinity Prediction: Scoring Functions for Structure-Based Virtual Screening [Internet]. In: *Methods and Principles in Medicinal Chemistry*. Sotriffer C (Ed.). . Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 177–221 (2011) [cited 2016 Jun 8]. Available from: http://doi.wiley.com/10.1002/9783527633326.ch7.
4. Korb O, Stützle T, Exner TE. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* 49(1), 84–96 (2009).
5. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55(14), 6582–6594 (2012).
6. Marcou, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* 47(1), 195–207 (2007).
7. Tan L, Batista J, Bajorath J. Computational Methodologies for Compound Database Searching that Utilize Experimental Protein-Ligand Interaction Information: Methods Using Protein-Ligand Interaction Information. *Chem. Biol. Drug Des.* 76(3), 191–200 (2010).
8. Schmidt D, Rickmeyer T, Krotzky T, Kolb P. DUPED: A concept for protein decoy pockets. GLISTEN COST meeting, Erlangen, Germany (2016).

Material and protocol:

The general flowchart is given in Fig1.

1. Protein and ligand source files

The *DUD-e* datasets can be downloaded from http://dude.docking.org/. The representation of active compounds and decoys is standardized. The MOL2 files describe 3D-coordinates of well-typed atoms, and are acceptable input for *PLANTS* docking.

The PDB files of targets (3EOC, 4FKL, 3CQW, 3OW4) and the orthogonal protein (1A42, 2VNI) can be downloaded from RCSB PDB or PDB-e. Please check the quality report of each of the crystal structures to evaluate the overall precision (resolution) as well as possible issues and ambiguities in the studied binding pockets (such as missing atoms, alternate positions, tautomers or isomers).

2. File preparation

For docking purpose, the minimal preparation of PDB files involves the protonation of the protein. For IFP calculation, proper typing of ligand atoms is required, too. To display/edit structure, it is possible to use the UCSF Chimera or a modelling suite. Here, we have prepared complexes as it has been described for sc-PDB entries (http://cheminfo.u-strasbg.fr/scPDB/).The following MOL2 files, extracted from original PDB complex, are required for the calculation:

- the protein chain (protein.mol2)
- the ligand in the binding pocket of the protein (ligand.mol2)
- the binding site (i.e., all residues at less than 6.5 Å from any ligand heavy atoms) (site.mol2)

3. Software

- To visualize molecular structure files, use your favorite program (Pymol, UCSF Chimera, etc…)
- The *PLANTS* docking program is freely available for academic use (http://www.tcd.uni-konstanz.de/research/plants.php).
- IFP calculations require the IChem program (available upon request from D Rognan)

4. Control docking

Before starting screening experiment, it is worth checking that the program successfully docks the ligand, which was extracted from the crystallographic structure of a ligand/protein complex, back into the protein binding site prepared from the same crystallographic structure. Here, we can conclude that *PLANTS* is able to reproduce the experimentally observed intermolecular interactions if we consider the three top-ranked poses (Fig 2).

5. Virtual screening and analysis

One screening of the *DUD-e* dataset lasts several days on a single CPU. For the sake of time, the tutorial only deals with the analysis of docking / IFP scores. The following files will be provided to the participants:

- mol2 files of prepared proteins and binding sites
- csv files of docking scores and IFP scores

The analysis of docking scores will be performed using the Knime Analytics Platform (Knime.com AG). Knime is a data analysis and modelling tool and can be downloaded from https://www.knime.org/downloads/overview. The Knime workflow will also be provided to the participants.

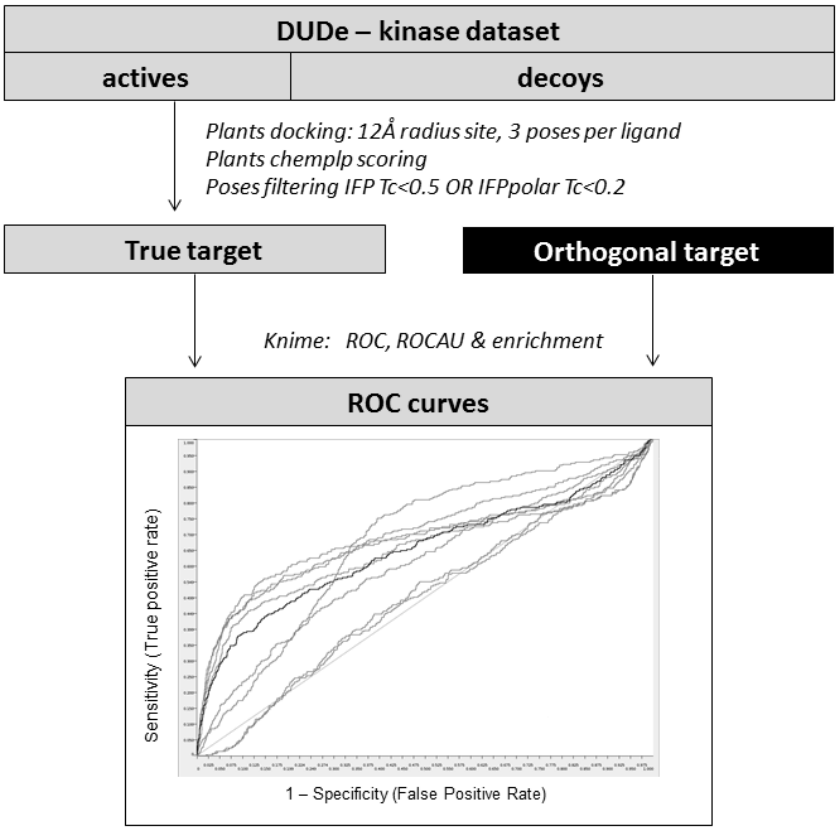

Figure 1. Bechmarking *PLANTS* using CDK-2 *DUD-e* dataset: overall flowchart and main results

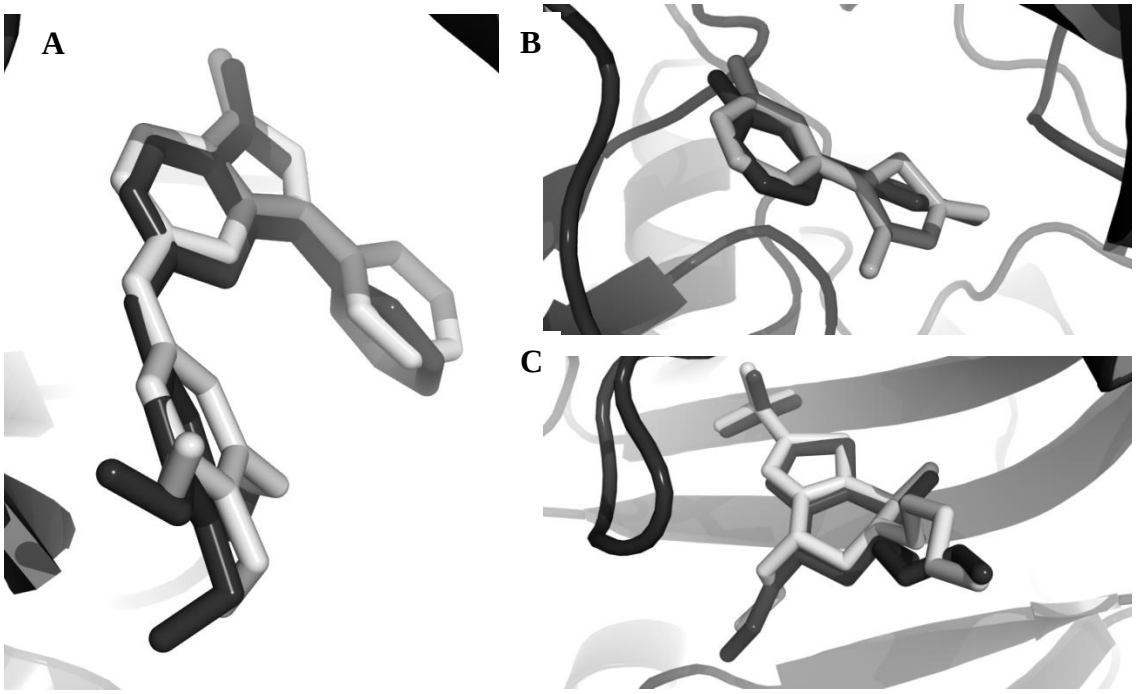

Figure 2. Example of comparison of the experimentally-observed binding mode (dark grey) of ligand in the crystal structure of protein, and docking pose generated using *PLANTS* (light grey). (A) Complex between CDK-2 and imidazo triazi-2-amine –T2A (3EOC) (B) Complex between CDK-2 and thiazolpyrimidine inhibitor -CK2 (4FKL), (C) Complex between carbonic anhydrase II and brinzolamide –BZU (1A42).