

Structural bioinformatics

CRDS: Consensus Reverse Docking System for target fishing

Aeri Lee and Dongsup Kim*

Department of Bio and Brain Engineering, KAIST, Daejeon 34141, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on May 1, 2019; revised on August 10, 2019; editorial decision on August 15, 2019; accepted on August 17, 2019

Abstract

Motivation: Identification of putative drug targets is a critical step for explaining the mechanism of drug action against multiple targets, finding new therapeutic indications for existing drugs and unveiling the adverse drug reactions. One important approach is to use the molecular docking. However, its widespread utilization has been hindered by the lack of easy-to-use public servers. Therefore, it is vital to develop a streamlined computational tool for target prediction by molecular docking on a large scale.

Results: We present a fully automated web tool named Consensus Reverse Docking System (CRDS), which predicts potential interaction sites for a given drug. To improve hit rates, we developed a strategy of consensus scoring. CRDS carries out reverse docking against 5254 candidate protein structures using three different scoring functions (GoldScore, Vina and LeDock from GOLD version 5.7.1, AutoDock Vina version 1.1.2 and LeDock version 1.0, respectively), and those scores are combined into a single score named Consensus Docking Score (CDS). The web server provides the list of top 50 predicted interaction sites, docking conformations, 10 most significant pathways and the distribution of consensus scores.

Availability and implementation: The web server is available at <http://pbil.kaist.ac.kr/CRDS>.

Contact: melugina@kaist.ac.kr

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Target identification is a key early step for discovering clinically relevant targets of chemical compounds in the field of drug discovery and development (Chan *et al.*, 2010; Schenone *et al.*, 2013). Although high-throughput experimental techniques are becoming available, an experimental procedure is time-consuming and expensive endeavor. Accordingly, there has been an urgent need for developing a practical computational tool to investigate a small molecule by identifying its interaction sites and some web tools are available (Peon *et al.*, 2019).

Inverse or reverse docking is a powerful technique for *in silico* target fishing against ligands in a database of target proteins (Lee *et al.*, 2016). The objective of reverse docking is to predict true targets among many clinically relevant protein targets. However, it has been known that the scoring functions of current docking programs have scoring bias toward the proteins with certain properties, which hinders accurate retrieval of target structures in reverse docking (Luo *et al.*, 2017).

One way to unravel this problem is to employ machine-learning scoring functions (Wojcikowski *et al.*, 2017; Yasuo and Sekijima, 2019). Another approach is to exploit consensus scoring method (Luo *et al.*, 2017). Consensus scoring evaluates poses of the docked

ligand with multiple scoring functions and combines the docking scores to improve the success rates. It has been reported that applying consensus scoring scheme which is incorporating with dissimilar types of scoring functions has proven to perform better than using a single scoring function (Cheng *et al.*, 2009). Hence, an increased probability of the ratio of true targets can be expected by using multiple scoring functions if one wants to identify targets for a compound of interest by applying docking.

Consequently, we have constructed a web-based server named Consensus Reverse Docking System (CRDS), which conducts quantitative screening of ligand interaction sites by reverse docking using consensus scoring and provides ranks with docked ligand–receptor structures, ranks of three of each algorithms, pathway analysis results and the complete set of consensus scores (see [Supplementary Fig. S1](#)).

2 Materials and methods

2.1 Consensus Docking Score

We adopted three types of scoring functions, which are GoldScore from GOLD version 5.7.1 (a force field-based) (Verdonk *et al.*, 2003), Vina from AutoDock Vina version 1.1.2 (a combination of

empirical and knowledge-based) (Trott and Olson, 2010) and LeDock from LeDock version 1.0 (a combination of physics and knowledge-based) (Wang *et al.*, 2016). To combine three docking values into a single score named Consensus Docking Score (CDS), we first normalized the docking scores derived from each scoring methods using min-max scaling approach, and the sum of the normalized three docking values were arranged in descending order (see Supplementary Fig. S2).

2.2 Target database

It is desirable to execute reverse docking in a large number of diverse target space. We were able to build a human protein target database resulting in a total of 5254 druggable binding sites from the sc-PDB (resolution < 2.5 Å) (Desaphy *et al.*, 2015). The analysis on the frequency of unique UniProt IDs showed that these 5254 protein structures consisted of 869 different UniProt IDs. For more detailed results, see Supplementary Figs S9 and S10.

3 Validation results

Performances of our server were validated in two different aspects, target fishing and virtual screening. We first demonstrated that consensus scoring scheme was able to retrieve more number of known target proteins within top 10 highest scoring proteins than each individual scoring functions [CDSs ($n = 242$), GOLD ($n = 119$), Vina ($n = 123$) and LeDock ($n = 186$)] when tested on 122 ligands with 6365 known targets compiled from DrugBank (<http://www.drugbank.ca>) and BindingDB (<http://www.bindingdb.org>) (see Supplementary Fig. S3 and Table S1). Another experiment to evaluate the reliability of the consensus scores to perform virtual screening using DUD-E dataset showed that the CDS achieved the highest area's under the curve scores (0.77) when compared to three existing scoring functions (see Supplementary Fig. S4). Furthermore, docking-based target prediction approach is most useful for targets with little ligand information because similarity-based methods such as quantitative structural activity relationship cannot be applied to those cases. Therefore, we looked for such cases and demonstrated that our docking-based consensus scoring method was effective for those targets with little ligand information (see Supplementary Material).

4 Web server

4.1 Input

The input window in our job submission page requires a job name, an email address and an ID from public chemical compound databases. A Tripos Mol2 file (mol2) format or a Structure Data File (sdf) format of a newly synthesized small molecule or a natural compound can be uploaded. Currently, the amount of time necessary to complete a job varies from 7 to up to 20 h depending on the molecular size and the loading of the server. Users can monitor the progress of their job on 'Queue' page.

4.2 Output

The web link to the results is reported to the user via email or through 'Queue' page. The first result section delineates the top 50 predicted interaction sites along with their corresponding PDB IDs, the CDSs, the ranks of Gold, Vina and LeDock, UniProt IDs, gene symbols and description of PDBs. The visualization buttons for binding pose of the ligand are provided. In addition, all complex structures are downloadable. The second section presents the top 50

predicted interaction sites of each algorithms along with their docking types, docking scores, PDB IDs, UniProt IDs, gene symbols and description of PDBs. The third section displays the pathway frequencies that are based on the mapping analysis of UniProt IDs of top 50 structures to pathway data in Reactome (<http://reactome.org/>) (Fabregat *et al.*, 2018). The 10 most meaningful pathways that the predicted 50 gene sets are involved in are illustrated on a pie chart. The fourth result section shows a total distribution of consensus scores.

5 Conclusion

We developed a large scale of predictive modeling tool named CRDS through the implementation of reverse docking with consensus scoring which can help finding probable interaction sites of small molecules such as existing drugs and natural products. We expect that the predicted drug interaction sites can be prioritized for identification of novel binding sites or used in extended applications for drug repurposing or adverse drug effect investigation.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grants (2017M3A9C4065952, 2019R1A2C1007951) funded by the Korea Government (MSIT).

Conflict of Interest: none declared.

References

- Chan, J.N.Y. *et al.* (2010) Recent advances and method development for drug target identification. *Trends Pharmacol. Sci.*, **31**, 82–88.
- Cheng, T. *et al.* (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, **49**, 1079–1093.
- Desaphy, J. *et al.* (2015) sc-PDB: a 3D-database of ligandable binding sites-10 years on. *Nucl. Acids Res.*, **43**, D399–D404.
- Fabregat, A. *et al.* (2018) The Reactome pathway knowledgebase. *Nucl. Acids Res.*, **46**, D649–D655.
- Lee, A. *et al.* (2016) Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Dis.*, **11**, 707–715.
- Luo, Q.Y. *et al.* (2017) The scoring bias in reverse docking and the score normalization strategy to improve success rate of target fishing. *PLoS One*, **12**, e0171433.
- Peon, A. *et al.* (2019) MolTarPred: a web tool for comprehensive target prediction with reliability estimation. *Chem. Biol. Drug Des.*, **94**, 1390.
- Schenone, M. *et al.* (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.*, **9**, 232–240.
- Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.
- Verdonk, M.L. *et al.* (2003) Improved protein-ligand docking using GOLD. *Proteins*, **52**, 609–623.
- Wang, Z. *et al.* (2016) Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.*, **18**, 12964–12975.
- Wojcikowski, M. *et al.* (2017) Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.*, **7**, 46710.
- Yasuo, N. and Sekijima, M. (2019) Improved method of structure-based virtual screening via interaction-energy-based learning. *J. Chem. Inf. Model.*, **59**, 1050–1061.