Usage Report for Fiscal Year 2017

Project Title:

Computational structure-based design of protein inhibitors

Name:

Kam Y. J. Zhang
Ashutosh Kumar
Dileep Vijayan
Matej Janezic

Laboratory at RIKEN:
Structural Bioinformatics Team,
Center for Life Science Technologies

---

1. Background and purpose of the project, relationship of the project with other projects

Molecular docking is undoubtedly the most commonly used method for predicting the binding poses of small molecules in protein binding pockets. Despite significant developments, molecular docking is still challenging due to the involvement of several pre- and post-docking steps affecting its performance. Factors influencing the accuracy of prediction include selection and quality of protein structures, preparation of receptor and ligand structures, sampling and scoring methods, solvation effects etc. Among these, the selection of suitable receptor structures is especially challenging as performance of a molecular docking based virtual screening method depends on the choice of protein structures used for molecular docking.

Among the different methods used to account for protein flexibility in molecular docking, ensemble-based methods are quite popular especially for virtual screening. Multiple crystallographic or ensemble of protein structures for a target protein are often used to take in account the protein flexibility and problems associated with single receptor structures. Multiple-receptor docking and selection of poses with the highest score across all docking is not economical when docking a large compound library with millions of small molecules. Moreover, it is challenging to select or generate suitable ensemble of protein structures. Hence in these circumstances, it would be sensible to dock a ligand of type A only to the suitable receptor for type A ligand. Therefore, in this research we have developed a cross-docking based virtual screening (CDVS) pipeline with a goal to improve docking performance by exploiting information from multiple receptor structures. For a target protein system, our method involves the identification of suitable crystallographic protein structures for each of the ligands in small molecule library. Suitable protein structures were identified based on the test ligand three-dimensional (3D) shape similarity with crystallographic ligands. Our CDVS pipeline had not only improved the virtual screening performance by considering protein flexibility but also reduced the computing time required to dock a large library of small molecules.

2. Specific usage status of the system and calculation method

The essence of our CDVS pipeline is the selection of suitable receptor structures for each ligand in the screening library. We have employed 3D shape similarity between the query ligand and crystallographic ligand to identify a suitable receptor structure from all the available co-crystal structures for a target protein. Our CDVS pipeline starts with the identification of all ligand-bound crystal structures of a target protein from Protein Data Bank (PDB). These structures are first aligned with the target protein using "superpose" program from CCP4 suite and crystallographic ligands at the binding site are extracted. Ligand 3D shape similarity calculations are then performed to select the most similar crystal ligand to a particular query ligand in a screening library. The ligand 3D shape similarity between crystallographic ligands and all conformers of test ligands was calculated using ROCS program. These ligand 3D shape similarity scores (ROCS shape similarity measure TanimotoCombo) are then used to identify suitable receptor-ligand pairs for molecular docking. Once receptor-ligand pairs for all compounds in the screening library are identified, docking and scoring are performed following the standard procedure. Our methodology takes advantage of multiple receptor docking, and yet a compound is docked only to a single receptor that is bound to the most similar ligand. The method was tested in Drug Design Data Resource (D3R) grand challenge 2016 exercise which is a platform to prospectively test molecular docking method and protocols.

3. Result

The performance of our CDVS pipeline was first evaluated by calculating the root mean square deviation (RMSD) between the predicted pose and the crystallographic ligand. As shown in Fig. 1, our method performed reasonably well and was able to produce docking poses close to X-ray conformation with a median RMSD of 1.19, 1.07 and 1.00 Å for the top, best of three and the best of five poses respectively. Majority (about 62.8 %) of the predicted top poses were within 2 Å which is a commonly accepted cutoff for successful predictions.

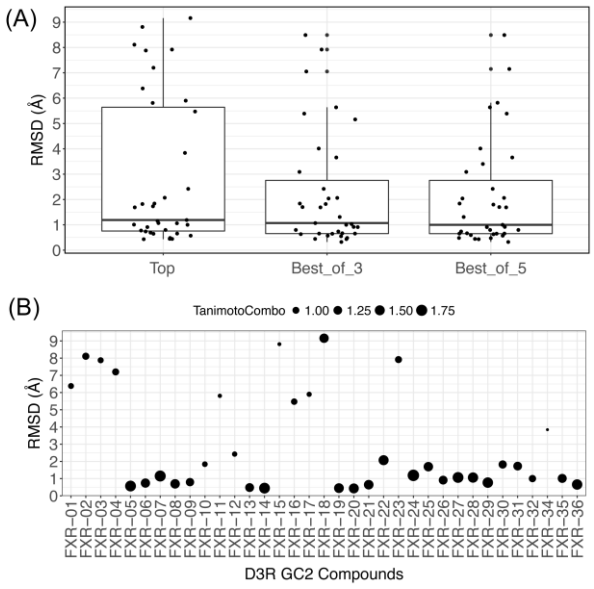

Fig. 1 Boxplots showing the distribution of RMSD between predicted poses and crystal structure ligands.

CDVS pipeline was also compared with standard docking while utilizing single, a subset of receptor structures and all available receptor structures. A boxplot showing the distribution of RMSD for test dataset ligands demonstrates the superiority of CDVS pipeline over single and multiple-receptor docking with the top pose median RMSD of 1.19 Å as compared to 4.11, 3.54 and 1.75 Å for single, a subset of receptors and all available receptors respectively (Fig. 2). We have previously reported that docking to all receptor structures is not always helpful. Hence, the selection of suitable ones among many available crystal structures becomes indispensable and our

method logically approaches this problem.

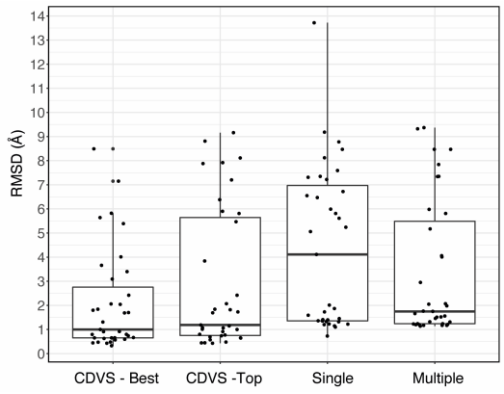

Fig. 2 Comparison of CDVS pipeline with single and multiple-receptor docking methods.

The selection of suitable receptors in CDVS pipeline contributed to the prediction of poses with reasonable accuracy. It is important to see that whether this improved pose prediction performance could be translated into improved virtual screening performance. Hence, the ability of CDVS pipeline in rank-ordering D3R dataset ligands was tested by plotting the sensitivity (true positive rate) and specificity (true negative rate) in a receiver operating characteristic (ROC) curve. As shown in Fig. 3, our CDVS pipeline was able to rank a randomly selected true positive higher than the randomly selected true negative, resulting in a high area under a ROC curve value of 81.3 % with a 95% confidence interval from 72.3 to 90.4%. A scatterplot of docking scores and experimental affinities exhibit respectable Spearman's ρ value of 0.59 with 0.07 standard error value.

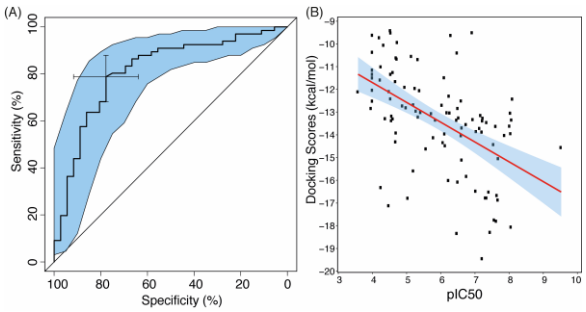

Fig. 3 CDVS pipeline virtual screening performance

Our CDVS pipeline not only demonstrated superiority over single and multiple-receptor docking approaches in predicting accurate poses of test ligands but also in their ranking. As shown in Fig 4, CDVS pipeline demonstrated better or similar performance as multiple-receptor docking with AUC value of 81.3 % for CDVS pipeline as compared to 76.3 % and 68.4% for multiple-receptor docking utilizing all and a subset of receptors respectively. The significant performance improvement was observed for CDVS pipeline when compared with single-receptor docking as AUC value improved from 66.1 to 81.3 % with the ROC test p-value of 0.03.

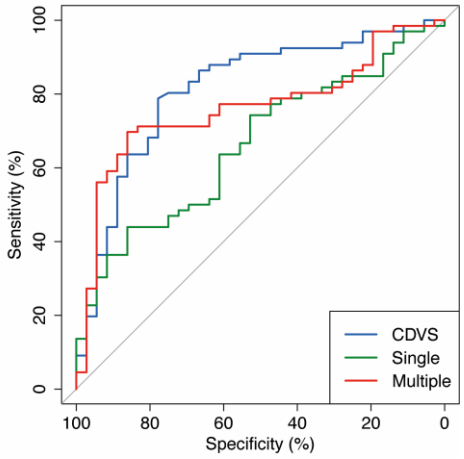

Fig. 4 Comparison of virtual screening performance of CDVS pipeline with single and multiple-receptor docking methods.

4. Conclusion

To improve pose prediction and virtual screening performance, we have developed CDVS pipeline that utilizes multiple-receptor information. However, contrary to docking all small molecules to multiple receptors, our method follows the cross-docking approach. Suitable receptors for each of the ligands in a screening library are selected for docking based on 3D ligand similarity with crystallographic

ligands. We have shown that our CDVS pipeline can achieve similar or better performance as compared to multiple-receptor docking. We have found that the conformation of binding pocket residues is critical in achieving reasonable performance in pose prediction and virtual screening. To identify their optimal conformation, either multiple crystallographic structure or molecular dynamics simulation based receptor ensembles are used. Both of these methods are useful when docking a small library of a few hundred small molecules but become impractical when docking a large small molecule library like ZINC due to computing requirements. Our CDVS pipeline eliminates the need of docking all compounds to all the receptors and thereby saving a significant amount of computing time while maintaining a comparable performance.

5. Schedule and prospect for the future

In future, we plan to extend the applicability of CDVS pipeline. The current implementation of CDVS pipeline is limited to cases where several ligand-bound crystal structures of target protein are available. However, utilizing ligand 3D shape similarity to select suitable receptor structures could be a general approach not limited to target protein. As ligand belonging to the same congeneric series mostly bind to homologous proteins in a similar manner, so conceptually suitable homologous protein structures could also be used as surrogate structures for virtual screening. We plan to use 3D shape similarity of ligands and binding pockets to select these structures and demonstrate its effectiveness in improving pose prediction and virtual screening performance. In addition to extend the utility of CDVS pipeline, we plan to apply CDVS pipeline in various in-house drug discovery projects. We will utilize CDVS pipeline to perform virtual screening of large small molecule libraries to identify small molecule inhibitors of various proteins with potential therapeutic values.

Fiscal Year 2017 List of Publications Resulting from the Use of the supercomputer

[Publication]

1. Kumar, A., Kawamura, T., Kawatani, M., Osada, H. and Zhang K. Y. J. (2017) Identification and structure activity relationship of purine derivatives as novel MTH1 inhibitors. *Chem. Biol. Drug Des.* 89(6):862-869.

2. Matsuoka, M., Kumar, A., Muddassar, M., Matsuyama, A., Yoshida, M., Zhang, K. Y. J. Discovery of Fungal Denitrification Inhibitors by Targeting Copper Nitrite Reductase from *Fusarium oxysporum* (2017). *J Chem Inf Model.* 57(2):203-213.

3. Kumar, A, and Zhang K. Y. J. (2018) A cross docking pipeline for improving pose prediction and virtual screening performance. *J Comput-Aided Mol Design.* 32(1): 163-173.

[Oral presentation at an international symposium]

1. Kumar, A and Zhang, KYJ (2015). A Pose Prediction Approach Based on Ligand 3D Shape Similarity. Structure Based Drug Design Conference, September 05-08, 2017, Lausanne, Switzerland. (Poster presentation)

2. Matsuoka, M., Kumar, A., Muddassar, M., Matsuyama, A., Yoshida, M., Zhang, K. Y. J. Discovery of Fungal Denitrification Inhibitors by Targeting Copper Nitrite Reductase from *Fusarium oxysporum* (2017), JCUPVIII, May 25-27, 2017, Tokyo, Japan. (Poster presentation)

[Others (Press release, Science lecture for the public)]

None