Poster Title: Toward fully automated high performance computing drug discovery: A massively parallel virtual screening pipeline for docking and molecular mechanics/generalized born surface area rescoring to improve enrichment

Authors: Xiaohua Zhang, Sergio E. Wong, and Felice C. Lightstone

Affiliation: Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Lab, Livermore, CA 94550.

Abstract:

Fast and accurate screening of drug candidates from large databases is crucial for successful computer-aided drug design efforts. In this poster we present a high throughput virtual screening pipeline for *in-silico* screening of virtual compound databases using high performance computing (HPC).[1] The pipeline includes receptor/target preparation, ligand preparation, docking calculation (VinaLC), and molecular mechanics/generalized Born surface area (MM/GBSA) rescoring using the GB model by Onufriev and co-workers.[2] The pipeline is intended for down-selecting large number of drug candidates by using firstly lower accuracy model (docking) and then higher one (rescoring). Notable features of this pipeline are an automated receptor preparation scheme and unsupervised binding site identification. A mixed parallel scheme that combines message passing interface (MPI) and multithreading was implemented in the VinaLC molecular docking program.[3] To exploit the typical cluster-type supercomputers, thousands of docking calculations were dispatched by the master process to run simultaneously on thousands of slave processes, where each docking calculation takes one slave process on one node, and within the node each docking calculation runs via multithreading on multiple CPU cores and shared memory. Input and output of the program and the data handling within the program were carefully designed to deal with large databases and ultimately achieve HPC on a large number of CPU cores. Parallel performance analysis of the VinaLC program shows that the code scales up to more than 15K CPUs with a very low overhead cost of 3.94%. One million flexible compound docking calculations took only 1.4 h to finish on about 15K CPUs. Furthermore, we leverage HPC resources to perform an unprecedented, comprehensive evaluation of MM/GBSA rescoring when applied to the DUD-E data set (Directory of Useful Decoys: Enhanced), in which we selected 38 protein targets and a total of ~0.7 million actives and decoys. The computer wall time for virtual screening has been reduced drastically on HPC machines, which increases the feasibility of extremely large ligand database screening with more accurate methods. HPC resources allowed us to rescore 20 poses per compound and evaluate the optimal number of poses to rescore. We find that keeping 5-10 poses is a good compromise between accuracy and computational expense. Overall the results demonstrate that MM/GBSA rescoring has higher average receiver operating characteristic (ROC) area under curve (AUC) values and consistently better early recovery of actives than Vina docking alone. Specifically, the enrichment performance is target-dependent. MM/GBSA rescoring significantly out performs Vina docking for the folate enzymes, kinases, and several other enzymes. The more accurate energy function and solvation terms of the MM/GBSA method allow MM/GBSA to

achieve better enrichment, but the rescoring is still limited by the docking method to generate the poses with the correct binding modes.

Acknowledgement

The authors thank Livermore Computing for the computer time and Laboratory Directed Research and Development for funding (12-SI-004). This work was performed under the auspices of the United States Department of Energy by the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Release Number LLNL-ABS-655464.

Reference:

1. Zhang, X.; Wong, S. E.; Lightstone, F. C., Toward Fully Automated High Performance Computing Drug Discovery: A Massively Parallel Virtual Screening Pipeline for Docking and Molecular Mechanics/Generalized Born Surface Area Rescoring to Improve Enrichment. *J. Chem. Inf. Model.* **2014,** *54* (1), 324-337.
2. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A., Generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* **2007,** *3* (1), 156-169.
3. Zhang, X.; Wong, S. E.; Lightstone, F. C., Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines. *J. Comput. Chem.* **2013,** *34* (11), 915-927.