Docking computational plan
Terrence Neumann
Marquette University

Introduction

Docking is the computational process of positioning a ligand into the binding site of a protein to identify the preferred pose of the ligand. Using this pose, a binding affinity for the ligand-protein complex can be calculated using a force field.[1] Docking plays a role in the design and discovery of drugs as drug candidates are docked against protein targets to predict their affinity to the target. This process is called virtual screening. [2]

Virtual screening is the process of automated docking of a chemical library into a protein, with the goal of discovering chemicals that bind with high affinity for the protein.[2] It is used in the early phase of drug discovery, since it can substantially reduce the resources dedicated to screening, so resources can be focused on the most promising and relevant compounds.[1] Zanamivir, an influenza therapeutic, is one example of a drug developed in part using the docking process. [3]

The virtual screening process is most efficient if physical compounds are available for rapid experimental verification of computational results. [1] Depending on the throughput of the experimental assay to verify the computational prediction, only a small percentage of ligands are experimentally verified. Concordia University Wisconsin's Center for Structural Drug Design and Development (CSDDD) uses an in-house screening collection of 12,000 compounds that contain drug-like molecules and also includes chemical dyes. The ZINC database maintains a collection of approximately 23 million commercially available compounds for virtual screening and takes 270 GB of storage. [4] Among these chemicals, two important subsets are worth noting. . The first is a fragment-like subset.[5] Fragments are molecules that if they bind proximally to each other in the protein can be chemically linked to produce a more potent inhibitor.[6] This subset is 350,000 chemicals and takes 2 GHz in storage. The second is a set of lead-like chemicals comprising of over 2.6 million compounds needing 27 GHz in storage.[7] The physical properties of these chemicals are typically associated with chemicals that the drug discovery process begins with.

Table 1. Relevant collections of chemicals for virtual screening

Chemical Collection Name	Chemical Collection Size	Storage Size
CSDDD Screening Collection	10,000	45 MB
ZINC fragment-like	350,000	2.6 GB
ZINC lead-like	2.6 million	27 GB
ZINC	23 million	270 GB

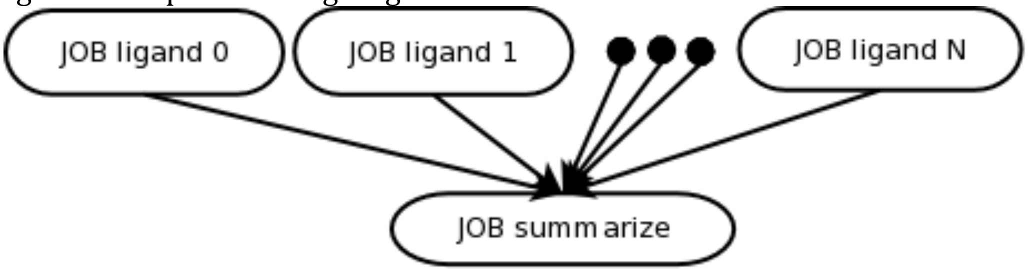
Two popular, large scale, distributed docking projects exist. The FightAIDS@home project uses the World Community Grid. This approach uses Autodock to virtually screen for new inhibitors of HIV protease.[8] The Docking@home project uses the Berkeley Open Infrastructure for Network Computing (BOINC) middleware to manage its volunteer computing grid. [9]

Current state

The Center for Structure-based Drug Design and Development (CSDDD) has designed a virtual screening workflow, Scheme 1, built upon Condor[10, 11] and DAGMan.[12] Using Autodock 4.2,[13] our lab has docked an in-house collection of 10,000 chemicals into six drug targets. Without this system, a virtual screen of this scale would take 277 days (40 minutes per ligand) on a single Intel Xeon 2.67 GHz processor. To facilitate this workflow, we have leveraged our local resources at Marquette University: MUGrid and Père. MUGrid is a distributed Condor pool containing over 500 cores. Père is a 1024-core Intel Nehalem centralized compute cluster split evenly using Condor and Portable Batch System (PBS) as grid middleware. Using the workflow on these resources, this virtual screening experiment can be completed in 6 hours.

This workflow uses a single input file containing all ligand, protein, and paramter files. To further automate the virtual screen, we have written scripts to prepare the Condor submit files, the DAGMan files, and other file management. Each node in the directed-acyclic graph (DAG) shows a specific step in the virtual screen. The docking of a single ligand-protein accounts for one vertex. After all dockings have completed, a daughter job to each of the docking jobs is launched to summarize all the dockings.

Scheme 1. Original Center for Structure-based Drug Design and Development (CSDDD) workflow built on Condor[10, 11] and DAGMan[12] using Autodock 4.2 to dock ligands into protein drug targets.



This document serves to explain how the CSDDD will incorporate larger chemical collections into its virtual screening projects. Challenges to be addressed include: docking simulation speed up, expanding number of dockings, individual file management, data management, and improving the throughput.

Table 2. Times to dock representative chemical collections using Autodock 4.2, a single Intel Xeon 2.67 GHz processor or Père's 500 core Condor pool.

Chemical Collection Size	Single Core	500 Core Condor Pool

1	40 min	40 min
10,000	277 days (est.)	6 hours
350,000	27 years (est.)	51 days (est.)
2.6 million	198 years (est.)	381 days (est.)
23 million	1750 years (est.)	325 years (est.)

Computational Plan

The first obstacle to address is the speed of an individual docking simulation. Using Autodock 4.2 each simulation of a ligand-protein docking takes 40 minutes. Thus, virtual screens with thousands of chemicals quickly eat up large computational resources. Recently, a new docking program, Autodock Vina (ADV), was released. [14] ADV results compares favorably to Autodock 4.2 results, but is roughly an order of magnitude faster.[8] Tables 2 and 3 compare the time to complete a virtual screen using Autodock 4.2 and ADV. While switching the docking simulator drastically reduces the computational time needed, additional resources will be needed.

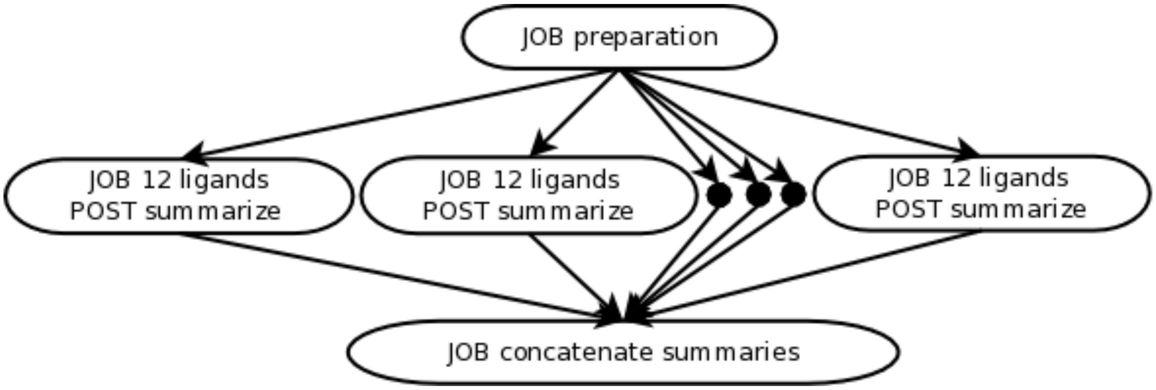
Table 3. Times to dock representative chemical collections using Autodock Vina (ADV), a single Intel Xeon 2.67 GHz processor or Père's 500 core Condor pool.

Chemical Collection Size	Single Core	500 Core Condor Pool
1	5 min	5 min
10,000	35 days (est.)	45 min
350,000	3.5 years (est.)	6.5 days (est.)
2.6 million	25 years (est.)	48 days (est.)
23 million	219 years (est.)	41 years (est.)

Using other resources will constitute additional considerations. Currently, a 10,000 chemical virtual screen uses in input file that is about 45 MB and Père uses a NFS such that each compute node has access to all necessary files without the need to transfer them. Expanding the number of dockings and using additional resources will not have this luxury afforded by the local computational resources.

To adjust the workflow to account for large amounts of storage such that smaller input files can be created. Many opportunistic cores may not have access to the required amount of local storage for such a large input file. Also, sending large files over networks can be time intensive. Scheme 2 shows a proposed method to handle these challenges.

Scheme 2. Virtual screening workflow to alleviate data storage and transferring problems encountered from docking large chemical collections.



The DAG from Scheme 1 has altered to account for a preparation step and the switch to ADV. The preparation step will create input files with the ADV parameter files and 12 ligands. This input file, approximately 100 KB, will then be staged on a web server to allow compute nodes access to the necessary files. In a problem of this scope, granularity becomes an issue. If each job was a single docking, the relative amount of time of file transfer to compute time is somewhat high. This means more work for the matchmaker in the Condor system. Also, if there is some system failure, roughly one hour per failed node would be lost. This allows for a quick recovery after a failure. Summarizing the dockings from each job parallelizes the summary process and leads to a quick process of concatenating the summaries.

The workflow will take advantage of glideins and the Open Science Grid (OSG). Using the submit node on Père, local computational resources will be maximized by keeping the local cluster queue filled as well as pushing jobs to the opportunistic MUGrid cluster as well as submitting jobs across the OSG computational resources.

Conclusion

This new workflow expands the scope of virtual screens undertaken in the CSDDD. Few groups are able to dock extremely large chemical collections and these require extensive grid infrastructure to access the large computational resources needed for such a task.[8] This approach is simple and could be implemented quickly with a small amount of overhead. Since preparation of the ADV parameter files is relatively easy, by automating the preparation, management, and summary of a virtual screen of the entire Zinc chemical collection, a portal could be developed that would benefit a vast number of researchers across many areas of research.

References

1. Shoichet, B. K. Virtual screening of chemical libraries. *Nature reviews. Molecular cell biology* **2004**, *7019,* 862-865.

2. Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem. Int. Ed.* **2002***, 15,* 2644-2676.

3. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993***, 6428,* 418-423.

4. Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, 177-182.

5. Carr, R. A.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: leads by design. *Drug Discov. Today* **2005**, 987-992.

6. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, 1531-1534.

7. Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem. Int. Ed.* **1999**, 3743-3748.

8. Chang, M. W.; Ayeni, C.; Breuer, S.; Torbett, B. E. Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina. *PLoS One* **2010**, e11955.

9. Estrada, T.; Armen, R.; Taufer, M. Automatic selection of near-native protein-ligand conformations using a hierarchical clustering and volunteer computing. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology (BCB '10);* ACM: New York, NY, 2010; .

10. Litzkow, M.; Livney, M.; Mutka, M. Condor - A Hunter of Idle Workstations. *Proceedings of the 8th International Conference of Distributed Computing Systems* **1988**, 104-111.

11. Thain, D.; Tannenbaum, T.; Livny, M. Distributed Computing in Practice: The Condor Experience". *Concurrency Computat. Pract. Exper.* **2005**, 323-356.

12. Couvares, P.; Kosar, T.; Roy, A.; Weber, J.; Wenger, K. Workflow in Condor. In *Workflows for e-Science;* Taylor, I., Deelman, E., Gannon, D. and Shields, M., Eds.; Springer: 2007; .

13. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009***, 16,* 2785-2791.

14. Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **2010**, 455-461.