Research on Identity Mechanism of Large-scale Drug Virtual Screening in Heterogeneous Multicore Environment and SMP Environment

Xiaohu Li¹, Zhiyu Zhang¹, Xinru Wang¹, Hao Liu¹ and Zhiqiang Wei¹*

¹ College of Information Science and Engineering, Ocean University of China, Qingdao, ShanDong, 266100, China
*Corresponding author’s e-mail: weizhiqiang@ouc.edu.cn

Abstract. This paper studies the mechanism of ultra-large-scale drug virtual screening in the heterogeneous multicore environment and SMP (Symmetric Multi-Processor) environment. Given the differences caused by different CPUs, we solved the incompatibility of the open-source molecular docking software Autodock Vina on two supercomputer systems in different environments and realized the automatic and intelligent large-scale virtual screening process for small-molecule ligands. At the same time, we propose a processing method of large-scale virtual screening using Autodock Vina in two environments, which makes full use of the dominant computing power of supercomputers to improve the efficiency of large-scale molecular docking and reduce the load of different supercomputer systems facing large-scale molecular docking tasks. Finally, we conduct a virtual screening experiment on the X86 platform supercomputer and the Sunway TaihuLight Supercomputer based on the heterogeneous multicore architecture CPU—SW26010. The experimental results show that our large-scale drug virtual screening method for different CPU environments is effective.

1. Introduction

Virtual screening is one of the critical technologies in the field of computer-aided drug design. Screening drugs with supercomputers is an essential means of discovering new drugs. In this process, the molecular docking program simulates the interaction between the target macromolecule and the small-molecule ligands by combining site recognition, conformation search and scoring function, and predicts the binding mode and affinity between them, thereby improving discovery efficiency of lead compounds [1-3].

With the increasing demand for computing resources in many fields, such as atmospheric science and biomedicine, many different types of supercomputer systems are beginning to emerge. In addition to those supercomputers built with traditional CPUs, heterogeneous multicore CPUs have started to be widely used in the construction of supercomputer systems. H. Fu et al. used the heterogeneous multicore architecture CPU—SW26010[4] to successfully build the world's first supercomputer system with a peak performance of more than 100 PFlops—the Sunway TaihuLight [5-6], and they have made significant progress in the research of atomic simulation, phase-field simulation and other fields using this system.

Faced with the supercomputer systems built by different types of CPUs, we integrate tools and data in the virtual screening process effectively on the X86 platform supercomputer and the Sunway TaihuLight based on heterogeneous multicore architecture CPU, aiming at the super large-scale
molecular docking scenario, which makes the combination of molecular docking program and computing resources of supercomputer systems effective. At the same time, according to the difference between them, we have explored the practical method of using molecular docking program to perform large-scale virtual screening in heterogeneous multicore environment and SMP environment, and proposed the identity mechanism of mass drug virtual screening, which makes the virtual screening process of molecular docking program more automated and intelligent in the two processor environments.

2. Related work
Supercomputer systems play an essential role in the virtual screening of drugs. More and more different types of supercomputers are used for virtual screening of drugs. How to use computing resources more effectively has attracted more and more attention. In this research, Sun Y X et al. developed the software platform IVSPlat 1.0 for the virtual screening process, which integrates the tools needed for virtual screening process into an operating environment and reduces the complexity of virtual screening [7]. Peng S et al. proposed a coordinated parallel framework called mD3DOCKxb for dynamic task scheduling and load balancing between nodes, which reduces communication and I/O latency [8]. These studies focus on more efficient and convenient use of computing resources to make virtual screenings more intelligent.

Because different types of supercomputer systems are highly independent of each other, there are many differences between them. We have studied the identity of large-scale virtual screening in different environments. On the heterogeneous multicore platform, if we want to make the software that belongs to the SMP environment run successfully, we must first make changes to the software that adapts to the heterogeneous multicore environment and port it successfully. We used the open-source molecular docking program Autodock Vina [9] in the virtual screening and successfully ported it by modifying the compilation parameters, porting the boost library and rewriting the program module, so that Autodock Vina can run in Sunway TaihuLight Supercomputer. At the same time, on the X86 platform supercomputer in the SMP environment and the Sunway TaihuLight Supercomputer in the heterogeneous multicore environment, we segment the data of small-molecule ligands so that virtual screening can meet the needs of large-scale molecular docking more quickly. Besides, according to the open-source characteristics of Autodock Vina, we have improved its I/O process and added MPI function to enable computing nodes of supercomputer systems to handle large-scale molecular docking tasks in parallel. Finally, we combine the job management system on the supercomputer with a series of automated scripting programs to make the entire virtual screening process more automated and intelligent.

3. Method

3.1 Improve the large-scale drug virtual screening process in the heterogeneous multicore environment and SMP environment.
Aiming at the characteristics of supercomputer systems in different architectures, we have established a virtual screening identity mechanism to make the virtual screening of drugs in both environments more intelligent. Throughout the virtual screening process, we have improved the four aspects of segmenting large-scale docking data, changing the I/O processes of ligands and receptors, increasing MPI parallel functions, and automating and intelligentizing processing processes. These changes are shown in Figure 1. The Heterogeneous Multicore environment is based on the Heterogeneous Multicore architecture CPU. It is widely used in the Sunway TaihuLight Supercomputer. The CPU contains four core groups, each of which contains one master core and sixty-four slave cores, the computational resource interaction between the core groups requires our manual code implementation. Figure 2 shows the heterogeneous multicore architecture CPU—SW26010. By eliminating the incompatibility of Autodock Vina in both environments and adding the corresponding MPI functions according to the characteristics of different compilation environments, we have achieved the identity
of the molecular docking process in the two environments and solved the differences caused by different systems.

![Virtual screening flow chart of drugs in two different environments](image)

**Figure 1.** Virtual screening flow chart of drugs in two different environments

### 3.2 Virtual Screening Process for Heterogeneous Multicore Environments

Figure 3 shows the modification process of AutoDock Vina on the platform of Sunway TaihuLight Supercomputer. By modifying the compilation parameters, porting the boost library and rewriting the program module, we successfully ported this molecular docking program. Then we encapsulate the central processing logic of AutoDock Vina and add MPI parallel module according to the characteristics of SW26010. As a result, the docking tasks can be executed in parallel between the core groups of SW26010. Figure 4 shows the change. Considering the large-scale docking computing scenario, we modified the I/O process of AutoDock Vina and added txt files as input parameters to store the list of ligands. Besides, we modified the command line parsing process in the source code, added the ability of a single receptor to dock ligands circularly, and changed the original one-time docking to all ligands in the docking file. These changes reduce the scheduling time of the initial docking tasks, and significantly improve the efficiency of molecular docking tasks.

![Symmetric Multiprocessor Environment](image)

**Figure 2.** SW26010 CPU

![Modification of Vina in Heterogeneous Multicore Environment](image)

**Figure 3.** Modification of Vina in Heterogeneous Multicore Environment
3.3 Virtual Screening Process for SMP Environment
In the compiler environment of SW26010 CPU, the collaborative computing between the master core and the slave cores, core groups and core groups needs to be implemented by adding code manually. In an SMP environment, CPU can automatically schedule computing resources. We only need to add MPI function to the main docking logic of AutoDock Vina. The collaborative computing between core and core will be processed automatically by CPU. At the same time, there is no need to port the molecular docking program in this environment.

Figure 4. Schematic diagram of Vina I/O changes in heterogeneous multicore environment

4. Experiments

4.1 Experimental Environment and Dataset
We conducted experiments on Sunway TaihuLight Supercomputer and the X86 platform supercomputer. The former is mainly composed of heterogeneous multicore processor SW26010, while the latter is primarily composed of Intel Xeno series CPUs. Due to differences in processors, the
compilation environment of these two supercomputer systems is also very different. The same application needs to conform to the compilation rules of the corresponding system in order to run normally.

In the experiment, we used the PDBbind core set v. 2016 to do batch molecular docking according to the virtual screening process. The dataset contains 285 compounds, each of which is split into separate receptor and ligand. Receptor molecules are stored in PDB format and SYBYL Mol2 format. Ligand molecules are stored in SYBYL Mol2 format and MDL SDF format. Finally, we use MGLTools to preprocess the dataset.

4.2 Batch Molecular Docking on Supercomputer Systems in Two Environments

4.2.1 Comparing the accuracy of AutoDock Vina's docking results in two supercomputer systems
Because the chip architectures of the two supercomputer systems are different, their compilation environment is also quite different. We have ported the molecular docking program AutoDock Vina for this difference, enabling it to be used for virtual drug screening in different supercomputer systems. At the same time, due to the nature of AutoDock Vina itself, when it repeatedly docks the same ligand and receptor, there may be a slight difference in results each time. With this in mind, we compared the critical parameters in the docking result files generated on the two supercomputer systems and found that the important parameters "active torsions", "status" and "ATOM" in the file are almost identical. The changes in these parameters are within reasonable limits. Therefore, we have found that the changes made in the process of porting the software did not affect the accuracy of the docking results of the docking program.

4.2.2 Comparison of Docking Efficiency between Two supercomputer system.
By adding a timestamp to the task submission script and the source code of AutoDock Vina, we can get three important time points: the time of task submission, the time when the task starts, and the time when the task ends. The interval between the submission time and the start time is the scheduling time of the supercomputer system, and the interval between the task start time and the task end time is the docking calculation time of AutoDock Vina.

Figure 6, Figure 7, Figure 8, and Figure 9 are the results of our batch docking of the dataset on a core group of SW26010 using the ported AutoDock Vina on Sunway TaihuLight supercomputer. Figure 6 is a trend diagram of docking time of unchanged Vina. The data in the graph is sorted in ascending order based on the docking calculation time of the compounds in the dataset. In order to achieve the same docking task as the improved vina, we use unimproved Vina repeatedly docked the dataset three times. Figure 7 is a trend diagram of the docking time after the MPI function and the cyclic docking function are added to the vina after the porting, and the data in the figure is arranged in ascending order according to the docking calculation time of the compounds in the dataset. Each task submission instruction will have one receptor and the same three ligands cyclically docked. The purpose of this is to ensure that the improved vina has the same docking workload as the unimproved vina. Figure 8 is a comparison of the docking time trends of the improved vina and the unimproved vina, and Figure 9 is a total docking time histogram.
Figure 6. Time graph for dataset to repeat three docking times using Unimproved Vina

Figure 7. Time diagram of the dataset using Improved Vina loop docking

Figure 8. Trend charts of docking time between Improved Vina and Unimproved Vina

Figure 9. Total docking time histogram of Improved Vina and Unimproved Vina

Figure 10, Figure 11, Figure 12, and Figure 13 are our experimental results on a CPU using a supercomputer based on the X86 platform. The experimental process and description of this part are consistent with our experiments mentioned above.

Figure 10. Time graph for dataset to repeat three docking times using Unimproved Vina

Figure 11. Time diagram of the dataset using Improved Vina loop docking
5. Conclusion

Faced with the differences between heterogeneous multicore environment and SMP environment, we effectively integrated the tools and data in the virtual screening process according to the characteristics of the ultra-large-scale drug virtual screening scenario, and explored an effective way to perform large-scale virtual screening using docking software in heterogeneous multicore environment and SMP environment. Finally, we propose a drug virtual screening mechanism that tends to be identical in heterogeneous multicore environment and SMP environment, making the virtual screening process more intelligent in different processor architecture environments.

Acknowledgments

This paper is supported by Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (No.2018SDKJ0402).

References

[1] Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. [J]. Journal of Medicinal Chemistry, 2001, 44(7):1035-1042.
[2] Kitchen D B, Decornez, Hélène, Furr J R, et al. Docking and scoring in virtual screening for drug discovery: methods and applications[J]. Nature Reviews Drug Discovery, 2004, 3(11):935-949.
[3] Irwin J J, Shoichet B K. ZINC? A Free Database of Commercially Available Compounds for Virtual Screening[J]. Journal of Chemical Information and Modeling, 2005, 45(1):177-182.
[4] Xu Z, Lin J, Matsuoka S. Benchmarking SW26010 Many-Core Processor[C]// 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE, 2017.
[5] Fu H, Liao J, Yang J, et al. The Sunway TaihuLight supercomputer: system and applications[J]. Science China Information Sciences, 2016, 59(7).
[6] Dongarra J. Report on the Sunway TaihuLight System[J]. UT EECS Technical Reports, 2016:1-24.
[7] Sun Y X, Huang Y X, Li F L, et al. IVSPlat 1.0: an integrated virtual screening platform with a molecular graphical interface[J]. Chemistry Central Journal, 2012, 6(1):2-2.
[8] Peng S, Zhang X, Yang S, et al. mD3DOCKxb: An Ultra-Scalable CPU-MIC Coordinated Virtual Screening Framework[C]// IEEE/ACM International Symposium on Cluster. IEEE Press, 2017.
[9] T. Oleg and O. A. J. (2009) “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading,” J. Comput. Chem.