Grid-enabled Hight Throughput *in-silico* Screening Against Influenza A Neuraminidase

Hurng-Chun Lee[1], Jean Salzemann[2], Nicolas Jacq[2], Li-Yung Ho[1], Hsin-Yen Chen[1], Vincent Breton[2], Ivan Merelli[3], Luciano Milanesi[3], Simon C. Lin[1] and Ying-Ta Wu[1]

[1] Academia Sinica, Taipei 115, Taiwan
[2] Laboratoire de Physique Corpusculaire, Université Blaise Pascal/IN2P3-CNRS UMR 6533, France
[3] Institute of Biomedical Technologies, National Research Council/CNR-ITB, Italy

Abstract. Encouraged by the success of first EGEE biomedical data challenge against malaria [1], the second data challenge was kicked off in April, 2006, fighting against avian flu. In the paper, we demonstrated how to adopt a world-wide deployed Grid infrastructure to efficiently produce a large scale virtual screening to speed up the drug design process. The 6-weeks activity of molecular docking on the Grid has covered over 100 years of computing power required for discovering new drug for avian flu. Around 600 Gigabytes of output has also been produced and archived on the Grid for further biological analysis and test.

1 Introduction

The potential for re-emergence of influenza pandemics has been a great threat since the report that avian influenza A virus (H5N1) could acquire the ability to be transmitted to humans [2]. An increase of transmission incidents suggests the risk of human-to-human transmission, and the report of development of drug resistance variants [3] is another potential concern. Two of present drugs (oseltamivir and zanamivir) were discovered through structure-based drug design targeting influenza neuraminidase (NA), a viral enzyme that cleaves terminal sialic acid residue from glycoconjugates. The action of NA is essential for virus proliferation and infectivity; therefore, blocking its activity generates antivirus effects. To minimize non-productive trial-and-error approaches and to accelerate the discovery of novel potent inhibitors, medical chemists take advantage of modeled NA variant structures and structure-based design.

A key work in structure-based design is to model complexes of candidate compounds to structures of receptor binding sites. The computational tools for the work are based on docking tools, such as AutoDock [4], to carry out quick conformation search of small compounds in the binding sites, fast calculation of binding energies of possible binding poses, prompt selection for the probable binding modes, and precise ranking and filtering for good binders. Although docking tools can be run automatically, one should control the dynamic conformation of the macromolecular binding site (rigid or flexible) and the spectrum of

the screening small organics. This process is characterized by computational and storage loads which pose a great challenge to resources that a single institute can afford.

In April 2006, the second biomedical data challenge of EGEE project led by Academia Sinica in Taiwan, CNRS-IN2P3 in France and CNR-ITB in Italy was kicked off to tackle the computational challenge of screening about 300,000 compounds selected from ZINC [5] and a chemical combinatorial library against 8 variants of neuraminidases predicted by homology method. Using AutoDock as the docking engine, the computation requires over 100 years if running on an average PC. In order to compress the overhead so that biomedical chemists can have best response to instant threads while the mutation of the virus happens, more than 2000 CPUs in the EGEE Grid infrastructure have been mobilized to perform large scale distributed virtual screening during 6 weeks. About 600 Gigabytes of output data have been produced and archived on the Grid with one additional backup.

Apart from the biological goal of reducing the time and cost of the initial investment of structure-based drug design, there are two Grid technology objectives for this activity: one is to improve the performance of the *in-silico* high-throughput screening (HTS) environment based on what has been learned in the previous challenge against Malaria (WISDOM) [6]; the other is to test another environment which enables users to have efficient and interactive control of the massive molecular dockings on the Grid. Therefore, two Grid tools were used in parallel in the second data challenge. An enhanced version of WISDOM high-throughput workflow was designed to achieve the first goal and a light-weight framework called DIANE [7] was introduced to carry a significant fraction of the deployment for implementing and testing the new scenario.

The paper is organized as follows. The next section introduces briefly the Grid environments on which the data challenge was executed, the Auvergrid, EGEE and TWGrid infrastructures. In section 3, the data challenge activity is discussed in detail in terms of achieved deployment, general statistics, efficiency and issues. The last section draws the final conclusion.

2 The Grid infrastructures

Three infrastructures were used to achieve the deployment: Auvergrid, TWGrid and EGEE. In this section, we are describing them briefly.

Auvergrid is regional grid deployed in the French region Auvergne. Its goal is to explore how a grid can provide the resources needed for public and private research at a regional level. With more than 800 CPUs available at 12 sites, Auvergrid hosts a variety of scientific applications from particle physics to life sciences, environment and chemistry.

TWGrid is responsible of operating a Grid Operation Center in Asia-Pacific region. Apart from supporting the world-wide Grid collaboration in high-energy physics, TWGrid is also in charge of federating and coordinating regional Grid

resources to promote the Grid technology to the e-Science activities (e.g. life science, atmospheric science, digital archive, etc.) in Asia.

The Enabling Grids for E-sciencE project (EGEE) [8] brings together scientists and engineers from more than 90 institutions in over 30 countries worldwide to provide a seamless Grid infrastructure for e-Science that is available for scientists 24 hours-a-day. The EGEE Grid consists of over 30,000 CPU available to users 24 hours a day, 7 days a week, in addition to about 5 Petabytes of storage, and maintains 10,000 concurrent jobs on average. Expanding from originally two scientific fields, high energy physics and life sciences, EGEE now integrates applications from many other scientific fields, ranging from geology to computational chemistry.

To efficiently operate the distributed resources as a whole system, the EGEE Grid middleware [9] provides a User Interface (UI), a Workload Management System (WMS), a Data Management System (DMS), an Information System (IS), several monitoring and application deployment tools based on the Grid Security Infrastructure (GSI). All the Grid activities and resource sharing within EGEE are operated and coordinated within the scope of Virtual Organization (VO) [10], a virtual community across laboratories and institutes around the world.

The data challenge against avian flu was officially supported by the biomedical VO of the EGEE project. Concerning that the EGEE Grid infrastructure is used by other VOs and the available resources for biomedical VO may variant during the data challenge, resources from AuverGrid and TWGrid were explicitly allocated to complement the EGEE resources for biomedical VO with dedicated computing power.

3 The Data Challenge

The development of the WISDOM workflow has been done in the previous data challenge. Concerning the execution efficiency, the workflow was slightly modified to address the issues observed in previous data challenge [6]. Since the DIANE framework takes care of the control of the communication and the workflow on behalf of the application, implementing an AutoDock adaptor for DIANE costs approximately 3 days and the effort is less than 500 lines of Python codes.

The input of the data challenge consists of 8 protein targets predicted from the neuraminidases subtype 1 to simulate the possible mutations of the H5N1 virus and 308,585 chemical compounds selected from ZINC and a chemical combinatorial library. By dividing 308,585 chemical compounds into 2 subsets, the whole data challenge activity consisted of 16 instances, the execution of each took care of the dockings between a NA variant and the compounds in one of the 2 subsets. Concerning that the concurrent executions of all the instances will overload the Grid system and reduce the Grid efficiency, the initialization time of each instance was also well scheduled.

Since the scalability had been tested in the first data challenge, the majority of the data challenge was executed by using the WISDOM platform. Due

to the fact that the CPU wall time in most of the Grid computing elements are restricted to 24 hour, the Grid jobs submitted by WISDOM were carefully partitioned to prevent from running over the limitation. Taking into account the approximation that the computing time of each single docking is about 30 minutes[4], each Grid job of WISDOM was prepared to run on 40 dockings. Thus each instance consisted of about 7715 Grid jobs. In order to balance the load of the Grid Workload Management System, WISDOM submitted the jobs to 18 Workload Management Systems in a round-robin order.

In parallel with the WISDOM activity, DIANE was used to run as many dockings as it can handle in the given 4 weeks with limited concurrent resources. Unlike WISDOM, the overall distribution efficiency of a DIANE job is essential to how the job is split into the independent DIANE tasks which will be pulled and executed by the DIANE workers. In this case, each DIANE task is trivially defined to run on a single docking. During the data challenge, a DIANE master was maintained on the UI to hold a queue of the waiting dockings and a separate process for submitting DIANE worker agents on the Grid was manually triggered while more CPU power was needed to improve the throughput. The result of each docking was interactively returned back to the Grid UI once the task was successfully completed. All the results were also summarized and archived into the Grid.

Table 1 and Table 2 summarize the data challenge activities done by WISDOM and DIANE, respectively. For WISDOM, 2000 Grid worker nodes distributed in 17 countries were successfully integrated to cover over 100 years of CPU power in 6 weeks. It has produced over 2 millions of the docking complexes on the Grid with the size of about 600 Gigabyte. Figure 1 shows how those Grid resources were distributed in different regions around the world. However, the long job waiting time mainly due to the Grid scheduling overhead and the Grid Resource Broker's unawareness of the fair-share control defined on the local queuing system restricted WISDOM to speed up the dockings by a factor of 767, or 38% in terms of the distributing efficiency[5]. In the WISDOM activity, we found that around 30% of the jobs were failed during the Grid scheduling phase or not successfully completed (i.e. the result cannot be obtained on the Grid) and those failed jobs require a resubmission procedure in WISDOM.

Although a similar job failure rate was also observed in the DIANE activity, the failure recovery mechanism in DIANE automated the resubmission and guaranteed a fully completed job. The pull-mode task scheduling and the interactive feedback from the DIANE worker agent also compressed the overhead of the Grid job scheduling thus the distribution efficiency was pushed forward to higher than 80% and a steady throughput was achieved (Fig 2). The improved efficiency was mainly resulted from a good resource utilization shown in Fig 3. Even though DIANE improves the efficiency, it remains a scalability issue. Re-

[4] The measurement was done on a PC with one Xeon 2.8 GHz CPU and 2 Gigabytes physical memory.

[5] Distribution efficiency here is approximated as the ratio between the overall seedup and the maximum number of concurrently CPUs.

sent performance evaluation observed that the current implementation of the DIANE master is restricted to handle few hundred DIANE workers at the same time due to the limitation of the concurrent stateful communication channels that a master manages to maintain.

Concerning that the follow-up biological analysis will be contributed by several research laboratories, for sharing the output, the docking results consisting of 123,440 files were archived in Taiwan with one replication in France. The centralized LCG File Catalog (LFC) system was used to index all the Grid files produced by the data challenge.

4 Conclusion

We have performed a large-scale *in-silico* screening on the Grid in search of the potential drugs for the predicted variants of the avian flu virus, H5N1. Using the world-wide deployed Grid infrastructure, we have successfully compressed the duration from over 100 years to 6 weeks. The results are now under analysis and the outcome will help biomedical chemists to reduce the cost of the first investment of the structure-based drug design.

The two activities adopting different Grid tools to execute the data challenge has been summarized and slightly compared. In the WISDOM activity, we proved again that the platform is feasible to control a hight through-put screening with a reduced preparation effort. In the DIANE activity, we demonstrated that using the light-weight framework, an improved distribution efficiency as well as a steady throughput of the distributed molecular dockings on the Grid can be effortlessly achieved.

Several issues concerning the tools and the Grid middleware has been obtained. Based on the results, investigations and discussions with the developers have been taking place in the preparation of the next data challenge fighting against several targets of neglected diseases.

References

1. WISDOM: Wide In-Silico Docking On Malaria, http://wisdom.eu-egee.fr
2. K. S. Li, Y. Guan, J. Wang, G. J. D. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. S. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. H. Hanh, R. J. Webby, L. L. M. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster and J. S. M. Peiris, Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia, Nature 430:209-213, 2004
3. M. D. de Jong, T. T. Tran, H. K. Truong, M. H. Vo, G. J. Smith, V. C. Nguyen, V. C. Bach, T. Q. Phan, Q. H. Do, Y. Guan, J. S. Peiris, T. H. Tran, J. Farrar, Oseltamivir Resistance during Treatment of Influenza A (H5N1) Infection, N. Engl. J. Med., 353(25):2667-72, 2005.
4. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function, J. Computational Chemistry, 19:1639-1662, 1998.

5. Irwin and Shoichet, J. Chem. Inf. Model., 45(1):177-82, 2005.
6. V. Breton, N. Jacq, M. Hofmann, Grid added value to address malaria, Proceedings of the 6-th IEEE/ACM CCGrid conference (2006)
7. DIANE: Distributed Analysis Environment, http://cern.ch/diane
8. EGEE: Enabling Grids for E-science in Europe, http://public.eu-egee.org
9. LCG-2 Middleware Overview, https://edms.cern.ch/file/498079/0.1/LCG-mw.pdf
10. I. Foster, C. Kesselman, S. Tuecke, The Anatomy of the Grid: Enabling Scalable Virtual Organizations, Int. J. Supercomputer Applications, 15(3), 2001.
11. W. P. Walters, M. T. Stahl and M. A. Murcko, Virtual Screening - an Overview, Drug Discovery Today, 3:160-178, 1998.

Table 1. Statistical summary of the WISDOM activity

Total number of completed dockings	$2 * 10^6$
Estimated duration on 1 CPU	88.3 years
Duration of the experience	6 weeks
Cumulative number of Grid jobs	54,000
Maximum number of concurrent CPUs	2000
Number of used Computing Elements	60
Overall speedup	767.37
Distribution efficiency	38.4%

Table 2. Statistical summary of the DIANE activity

Total number of completed dockings	308,585
Estimated duration on 1 CPU	16.7 years
Duration of the experience	30 days
Cumulative number of Grid jobs	2580
Maximum number of concurrent CPUs	240
Number of used Computing Elements	36
Overall speedup	203
Distribution efficiency	84%

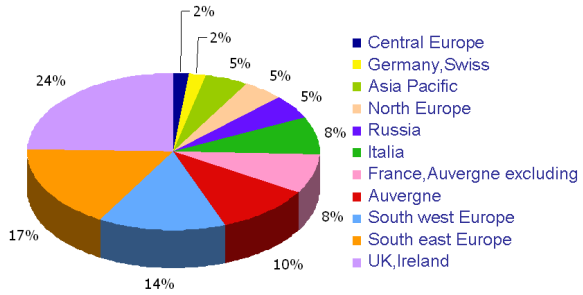

Fig. 1. The distribution of the Grid jobs in different region.

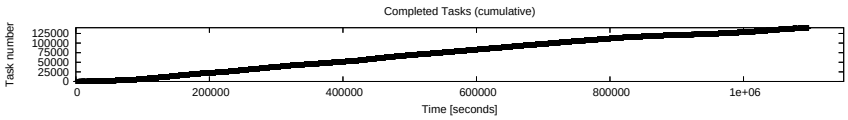

Fig. 2. The docking throughput of a DIANE job. The curve shows the cumulative number of the completed dockings during the DIANE job runtime.

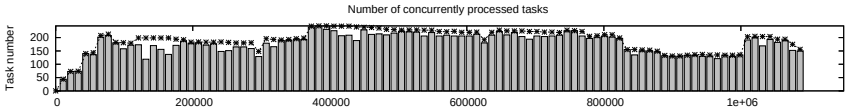

Fig. 3. The resource utilization of a DIANE job. The curve with crosses on it illustrates the number of CPUs available for doing the dockings; while the bars indicate the concurrent executing dockings (i.e. the utilized CPUs).