建構亞洲流行性疾病之高效能藥物篩選計算服務

High Throughput Virtual Screening Service against the Epidemic Diseases in Asia

Hsin-Yen Chen[1] His-Kai Wang[1] Eric Yen[1] Ying-Ta Wu[2] Simon C. Lin[3]

[1] Academia Sinica Grid Computing Center, Taiwan [2] Genomic Research Center, Academia Sinica, Taiwan [3] Institute of Physics, Academia Sinica, Taiwan

【**Abstract**】In the modern biomedical drug discovery, molecular docking simulation is a common method for predicting potential interacting complexes of small molecules on protein binding sites. However, searching all optimal conformations of a compound could be a time-consuming process. GAP Virtual Screening Service (GVSS), large scale in-silico protein-ligand virtual screening service, provides a production system to speed-up the searching process. We demonstrated how this high throughput in-silico massive molecular docking service benefits from state-of-the-art Grid technology to the activities for Avian Flu drug refinement and Dengue Fever drug discovery on EUAsiaGrid infrastructure since March 2009. Furthermore, these activities also facilitates more biomedical e-Science applications in Asia, such as other diseases and compounds profiling.

【**Keywords**】Drug Discovery、Molecular simulation、Grid Computing、High throughput computing、e-Science

1. Introduction

Since the first global Avian Flu Data Challenge was launched in 2005, Academia Sinica Grid Computing Centre (ASGC), within the EGEE collaboration, was devoted in developing and refining virtual screening for neglected and emerging diseases such as Avian Flu and Dengue Fever. Molecular docking simulation is a time consuming process to search exhaustively all correct conformations of a compound. However, the massive in silico processes benefit from the high throughput computing grid technology. Providing intensive computing power and effective data management, the production e-infrastructure (EUAsia VO) enables opportunities for in silico drug discovery on the important epidemic disease in Asia.

GAP and GVSS (GAP Virtual Screening Service) were developed with the docking engine of the AutoDock 3.0.5. GAP is a high-level application development environment for building up production-quality grid application services. It divides the grid application development work into three major stages: application gridification, complex job workflow design and user interface customization. GVSS is a Java-based graphical user interface and was designed for conducting large-scale molecular docking more easily on the gLite grid environment. The end users using GVSS are allowed to specify target and compound library, set up docking parameters, monitor docking jobs and computing resources, visualize and refine docking results, and finally download the final results.

We will foster the more biomed collaboration between the life science communities and IT people in Asia. There are other challenges to encourage more biomedical activities and integrate more dynamical resources to support the large scale virtual screening simulation in Asia. For example, scientists study the new target structure, therefore, he/she shall know how to model the target and prepare it using the AutoDock tools. One would also need a user-friendly GUI in order to join and access the collaboration, to submit the pipeline docking jobs, monitoring their

progress, visualize the docking and finally analyze the results. This is a good start that we initiate the first dengue fever data challenge on EUAsia Grid project. Most of the south Asia partners (UPM, AdMU, ASTI, ITB, IAMI, NECTEC, HAII) considered that they can seek for the more scientists to join this drug discovery at ISGC 2009.

2. High Throughput Virtual Screening Service

Molecular docking simulation is a useful method for predicting potential interacting complexes of small molecules in protein binding sites, that information are crucial to structure-based drug design (SBDD). Several docking programs, such as DOCK, GOLD, Autodock, Glide, LigandFit and FlexX [1-6] etc. have proved themselves useful in the pipeline of the in-silico drug discovery. The basic method behind molecular docking simulation is to generate all potential conformations of a docking molecule and evaluate among them for the most favorable orientation as the binding mode of the molecule by using a scoring function. To search exhaustively all correct conformations of a compound is a very time consuming process. Therefore, a successful docking simulation for large-scale high-throughput screening (HTS) will consume large computing resources [7]. For instance, it requires a few Tera-flops per job to exercise docking of thousands of compounds to one target protein. However, the existing tools are lack of simple way to provide concise procedures for regular users to arrange resources to conduct a massive molecular dockings.

Grid technology starts a new era of virtual screening due to its efficiency as well as its cost effectiveness [7,8]. The cost of traditional in vitro testing is usually extremely high when large-scale screening is conducted. Virtual screening provides scientists an effective tool to select the potential compounds for in vitro testing. As a result, virtual screening could indeed save enormous amount of money comparing to the traditional in vitro testing.

With the help of the high-speed computing and huge data managing capabilities of the Grid, possible drug components can be screened and studied very rapidly by the available computer modelling applications. This will free up medicinal chemists' time to better respond to instant, large-scale threats. Moreover, one can concentrate one's biological assays in the laboratory on the most promising components, the ones expected to have the greatest impact.

In this study, state-of-the-art Grid technology was deployed in order to execute a massive docking on Autodock [3]. Compared to the traditional high performance computing (HPC), the Grid computing environment has the capability of integrating the dynamic computing resources and the scalability in providing the computing services. In a production Grid environment, thousands of CPUs can be easily allocated and formed an enormous computing power for massive docking tasks. For example, more than 2000 computers under the Enabling Grids for E-sciencE (EGEE, funded by the European Commission) Biomed Virtual Organisation (VO, a.k.a Application Community) were employed in 4 weeks for Avian Flu Data Challenge (DC) in 2005 [9,10].

To make the large-scale molecular docking running on Grid environment, ASGC developed the GVSS (GAP-enabled virtual screening service) application package that incorporates the EGEE gLite middleware DIANE2/GANGA [12] and AMGA [13]. As illustrated by Figure 1, all computing jobs are managed by GAP/DIANE to distribute the Grid computing workers to the Grid. The computing results are managed by AMGA metadata catalogue to store on the storage elements. GVSS also uses Autodock as the docking engine. The GVSS was created by integrating several frameworks designed for Grid applications. They are described in the following subsections with further details.

3. Fostering Asia Biomed e-Science: In-silico docking against dengue fever

Dengue fever is now epidemic due to its geographic spread as well as increasing incidents. According to the World Health Organization (WHO), about 2.5 billion people, two fifths of the world's population, are at risk from dengue fever. It is estimated that approximately 50 million cases of dengue infection worldwide every year. A majority of cases (around 95 %) are among children of less than 15 years of age in many countries of South-East Asia.

Several flaviviruses are important human pathogens, including dengue virus, a disease against which neither a vaccine nor specific antiviral therapies currently exist. The dengue diseases are caused by the four antigenically distinct dengue virus serotypes, DENV 1-4. We will looked for the favorable drug leads to combat the related dengue, hepatitis C, West Nile, and Yellow fever viruses by the in silico simulation. The NS3 protease is an enzyme critical for virus replication, and its amino acid sequence and atomic structure are very similar among the different disease–causing flaviviruses. Furthermore, it is proven as drug target in Hepatitis C Virus (HCV) studies. Currently no drugs could effectively treat the disease. The discovery of both broad–spectrum and specific antiviral drugs is expected to significantly lower the mortality. The NS3 protease is an enzyme critical for dengue virus replication; therefore it was considered as the target for developing the therapeutic against dengue fever.

Inspired by the successful experiences on Avian Flu Data Challenges, ASGC coordinated the Dengue Fever Data Challenge via EUAsiaGrid VO in June 2009. The objective was to utilize the GAP-enabled screening service (i.e. GVSS) for structure–based computation to identify small molecule protease inhibitors. A total of 300,000 compounds from CDI (ChemDiv Inc., San Diego, USA) compound library, a commercially available compound library, were selected for virtual screening.

Ordinary users, usually not Grid experts, can select compounds and targets, submit simulation jobs, monitor the progress, and visualize the results by utilizing the GVSS.

This study aims to look for the favourable drug leads to combat the dengue viruses by the in silico simulation on the EUAsiaGrid production environment. We prepared the compound library, ZINC CDI that is a free database of commercially available compounds for virtual screening. It is composed of total 300,000 compounds that we delivered. For this phase, we proposed the well determined NS3 protein (PDB id is 2vbc) against the production ZINC CDI database on the GVSS.

Two hundred of CPU-cores from EUAsiaGrid VO were allocated. Eleven institutes from Asia and Europe joined this activity, they were: Genomics Research Center, Academia Sinica and Academia Sinica Grid Computing Centre (ASGC) from Taiwan; Ateneo de Manila University (AdMU) and Advanced Science and Technology Institute (AIST) from Philippines; Universiti Putra Malaysia (UPM) and MIMOS Berhad from Malaysia; Hydro and Agro Informatics Institute (HAII) and National Electronics and Computer Technology Center (NECTEC) from Thailand; Institute of Applied Mechanics and Informatics (IAMI), Vietnam; Institut Teknologi Bandung (ITB), Indonesia; and CESNET, Czech Republic. We estimated the total computing power is 4,167 CPU-day to support this. For the phase I, a total of 46 GB of the computing results from execution of the 300,000 jobs were generated. We analyzed the binging energy profile to seek for the more useful informatics database.

A large portion of computing jobs were submitted and executed on the ASGC computing resources. The rest of the computing jobs are also distributed to the other grid computing resources of EUAsia VO. The maximum number of the concurrent CPU resources is 100 CPUs and the average distribution efficiency is 69.44% (Table 1). The distribution efficiency reflects reasonable computing usage when sharing with other computing jobs. This further proves the availability and reliability of the EUAsia grid infrastructure.

Table 1: Statistical Summary of the DIANE Activity via GAP

Total number of completed dockings	300,000
Estimated duration on 1 CPU	4,167 days
Duration of the experiment	60 days
Cumulative number of Grid jobs	4,015
Maximum number of concurrent CPUs	100
Number of used Computing Elements	6
Crunching factor	69.44
Approximated distribution efficiency	69.44%

4. Summary

Contemporary drug discovery, strengthened by automatic high-throughput screening (HTS) techniques, can now screen millions of compounds in a time span of a week. However, the HTS approach is affected by the time required to develop stable screening assays for a disease target at hand. Molecular docking is a useful method to predict potential interacting complexes of small molecules in a protein binding site. That information can be applied as modeling HTS to funnel out a majority of less potential compounds from a huge chemical database or in-house collection so as to reduce time taken for useful information of active compounds.

Therefore, allocation of enough resources from a production Grid to accelerate the screening throughput in an effective way is a feature that present technology is capable of. ASGC develops GVSS to enable users to access Grid technology and resources seamlessly whilst providing flexible control over their docking jobs on the Grid. The GVSS GUI hides the complexity of deploying and utilizing massive molecular dockings. It allows users to submit and retrieve jobs easily and inspect docking results flexibly. The GVSS is freely available on the web (http://gap.grid.sinica.edu.tw).

We introduce GVSS which is a user-friendly graphical user interface desktop application for using this Grid-enabled virtual screening service. Through the GUI, the end-users can easily take the advantage of GRID computing resources for large-scale virtual screening. Furthermore, they can even upload their own target and ligands, and do the same docking process, visualization and analysis with this GUI, of course including the advanced refinement docking simulations. The end-users can finally have a real GRID-enabled desktop utility for the virtual screening service for their daily research. Moreover, we expected to foster the biomedical grid activities and promote the e-Science collaboration between partners in Asia and Europe.

FIGURE CAPTIONS

Figure 1: The GAP-based Virtual Screening Service (GVSS): Users prepare the large-scale virtual screening tasks in the GVSS Graphic User Interface, and then select the Grid computing resources to submit jobs. Those computing jobs are managed by GAP/DIANE to distribute the Grid computing workers to the Grid. The computing results are managed by AMGA metadata catalogue to store on the storage elements.

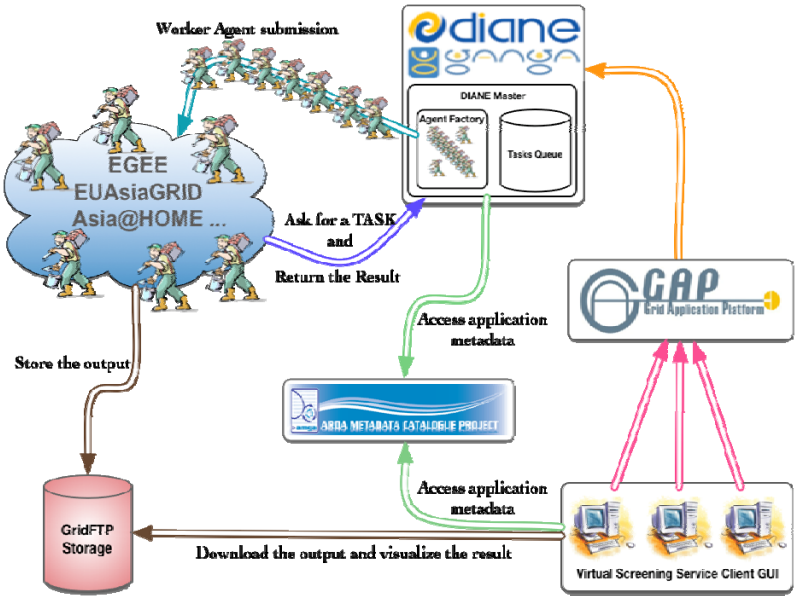
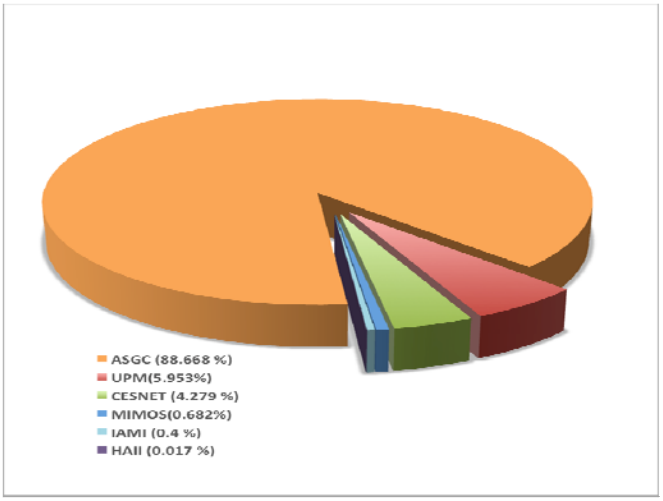


Figure 2: Distribution of the grid jobs via GAP in EUAsia infrastructure. The total computing power is 4,167 CPU-day to support this. Moreover, the storage of 46 GB of the computing results from execution of the 300,000 jobs were generated.



REFERENCES

1. Ewing, T.J.A., Makino, S., Skillman, A.G., et al.: DOCK4.0: search strategies for automated molecular docking of flexible molecule database. J. Comput. Aid. Mol. Des. **15**(5), 411-428 (2001), DOCK: UCSF DOCK, http://dock.compbio.ucsf.edu/Overview_of_DOCK/index.htm.
2. Meng, E.C., Gschwend, D.A., Blaney, J.M., and Kuntz, I.D.: Orientational sampling and rigid-body minimization in molecular docking. Proteins: Structure, Function, and Genetics, **17**, 266-278 (1993).
3. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J.: Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. J. Computational Chemistry, **19**, 1639-1662 (1998).
4. Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E., Francis, P., Shenkin, P.S.: Glide: a new approach for rapid, accurate docking and scoring.

1. Method and assessment of docking accuracy. J. Med. Chem., **47**, 1739-1749 (2004).

5. Venkatachalam, C.M., Jiang, X., Oldfield, T., Waldman, M.: LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model., **21**, 289-307 (2003).
6. Rarey, M., Wefing, S. and Lengauer, T.: Time-Efficient Docking of Flexible Ligands into Active Sites of Proteins. In: RAWLINGS, C. et al.(Editors): Proceedings of the Third International Conference on Intellegent Systems in Molecular Biology, 300-308, AAAI Press, Menlo Park, California (1995).
7. Chien, A., Foster, I., Goddette, D.: Grid technologies empowering drug discovery. Drug Discovery Today, **7**, Suppl. 176-180 (2002).
8. Buyya, R., Branson, K., Giddy, J., Abramson, D.: The Virtual Laboratory: a toolset to enable distributed molecular modeling for drug design on the World-Wide Grid, Concurrency Computat, Pract. Exper. **15**, 1-25 (2003).
9. Lee, H.-C., Salzemann, J., Jacq, N. Chen, H.Y., Ho, L.Y., Merelli, I., Milanesi, L., Breton, V., Lin, Simon C., Wu, Y.T.: Grid-enabled High-throughput in silico Screening against influenza A Neuraminidase, IEEE Transaction on Nanobioscience, **5**, 288-295 (2006).
10. Jacq, N., Breton, V., Chen, H.Y., Ho, L.Y., Hofmann, M., Kasam, V., Lee, H.C., Legré, Y., Lin, Simon C., Maaß, A., Medernach, E., Merelli, I., Milanesi, L., Rastelli, G., Reichstadt, M., Salzemann, J., Schwivhtenberg, H., Wu, Y.T., Zimmermann, M.: Virtual screening on large scale grids, Parallel Computing, **33**, 289-301 (2007).
11. Germain-Renaud1, C. , Loomis C., Mościcki J.T.: Scheduling for Responsive Grids. J. Grid Computingt. **6**, 15–27 (2008). Moscicki, J.T.: Distributed analysis environment for HEP and interdisciplinary applications Nuclear Ins. Methods Phys. Res. A, **502**, 426-429 (2003).
12. Moscicki, J.T., Brochu, F., Ebke, J., Egede, U., Elmsheuser, J., Harrison, K., Jones, R.W.L., Lee, H.-C., Liko, D., Maier, A., Muraru, A., Patrick, G.N.,

Pajchel, K., Reece, W., Samset, B.H., Slater, M.W., Soroko, A., Tan, C.L., van der Ster, D.C., Williams, M.: GANGA : A tool for computational-task management and easy access to Grid resources. Computer Physics Communications, **180**, 2303-2316 (2009).

13. Koblitz, B., Santos, N., Pose, V.: The AMGA metadata service. J. Grid Computing, **6**, 61-76 (2008).

【作者簡介】

陳信言 男

職　　稱： 中研院 資訊科技創新研究中心 研究助理

職　　務： 中研院 網格計算團體 專案經理

研究領域： Grid Computing、Computational Physics、High Performance Computing

聯絡電話： 886-2-27898306

聯絡郵箱： hychen@twgrid.org

王璽凱 男

職　　稱： 中研院 資訊科技創新研究中心 研究助理

職　　務： 中研院 網格計算團體 專案工程師

研究領域： Grid Computing、High Performance Computing

聯絡電話： 886-2-27898309

聯絡郵箱： hsikai.wang@twgrid.org

吳盈達 男

職　　稱： 中研院 基因體研究中心 副研究技師

研究領域： High Throughput Screening、Cheminformatics

聯絡電話： 886-2-27871237

聯絡郵箱： ywu@twgrid.org

嚴漢偉 男

職　　稱： 中研院 資訊科技創新研究中心 副研究技師

研究領域： Cloud & Grid Computing、Digital Library、Geospatial Informatics

聯絡電話： 886-2-27898375

聯絡郵箱： Eric.Yen@twgrid.org

林誠謙 男

職　　稱： 中研院 物理研究所 副研究員

職　　務： 中研院 網格計算團體 計畫主持人

研究領域： Grid & Cloud Computing、Computational Sciences、Information Physics

聯絡電話： 886-2-27896793

聯絡郵箱： Simon.Lin@twgrid.org