Demonstration of In Silico docking at a large scale on grid infrastructure

N. Jacq, J. Salzemann, Yannick Legre, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, F. Jacq, M. Zimmermann, A. Maas, M. Sridhar, K. Vinod-Kusam, H. Schwichtenberg, et al.

Demonstration of *In Silico* Docking at a Large Scale on Grid Infrastructure

Nicolas JACQ[a+b], Jean SALZEMANN[a], Yannick LEGRE[a], Matthieu REICHSTADT[a], Florence JACQ[a], Marc ZIMMERMANN[c], Astrid MAAß[d], Mahendrakar SRIDHAR[c], Kasam VINOD-KUSAM[c], Horst SCHWICHTENBERG[d], Martin HOFMANN[c], Vincent BRETON[a]

[a]*Laboratoire de Physique Corpusculaire, Université Blaise Pascal/IN2P3-CNRS UMR 6533, France*
[b]*Communication & Systèmes, CS-SI, France*
[c]*Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department of Bioinformatics, GERMANY*
[d]*Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department of Simulation Engineering, GERMANY*

Abstract. WISDOM stands for World-wide *In Silico* Docking On Malaria. First step toward enabling the in silico drug discovery pipeline on a grid infrastructure, this CPU consuming application generating large data flows was deployed successfully on EGEE, the largest grid infrastructure in the world, during the summer 2005. 46 million docking scores were computed in 6 weeks. The proposed demonstration presents the submission of *in silico* docking jobs at a large scale on the grid. The demonstration will use the new middleware stack gLite developed within the EGEE project.

Keywords: Large scale deployment; grid infrastructure; in silico docking; drug discovery; malaria

Malaria is a dreadful disease affecting 300 million people and killing 1.5 million people every year [1]. Drug resistance has emerged for all classes of antimalarials except artemisinins. This example illustrates the real need for new drugs against neglected diseases. The World Health Organization declared its will to help the development of new drugs against neglected diseases starting from hits proposed by the fundamental researchers.
Advance in combinatorial chemistry has paved the way for synthesizing large numbers of diverse chemical compounds. Thus there are millions of chemical compounds available in the laboratories and also in 2D, 3D electronic databases, but it is nearly impossible and very expensive to screen such a high number of compounds in the experimental laboratories by high throughput screening (HTS). Besides the high costs, the hit rate in HTS is quite low, about 10 to 100 per 100,000 compounds when screened on targets such as enzymes [2].

An alternative is high throughput virtual screening by molecular docking, a technique which can screen millions of compounds rapidly, reliably and cost effectively. Screening millions of chemical compounds *in silico* is a complex process. Screening each compound, depending on structural complexity, can take from a few minutes to hours on a standard PC, which means screening all compounds in a single database can take years. Computation time can be reduced very significantly with a large grid gathering thousands of computers [3,4].

WISDOM (World-wide *In Silico* Docking On Malaria) is an European initiative to enable the *in silico* drug discovery pipeline on a grid infrastructure. Initiated and implemented by Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) in Germany and the Corpuscular Physics Laboratory (CNRS/IN2P3) of Clermont-Ferrand in France, WISDOM has deployed a large scale docking experiment on the EGEE [5] infrastructure. Three goals motivated this first experiment. The biological goal was to propose new inhibitors for a family of proteins produced by *Plasmodium falciparum*. The biomedical informatics goal was the deployment of *in silico* virtual docking on a grid infrastructure. The grid goal is the deployment of a CPU consuming application generating large data flows to test the grid operation and services. Relevant information can be found on http://wisdom.eu-egee.fr and http://public.eu-egee.org/files/battles-malaria-grid-wisdom.pdf.
The first large scale docking experiment ran on the EGEE grid production service from 11 July 2005 until 19 August 2005. It saw over 46 million docked ligands, the equivalent of 80 years on a single PC, in about 6 weeks. Usually in silico docking is carried out on classical computer clusters resulting in around 100,000 docked ligands. This type of scientific challenge would not be possible without the grid infrastructure - 1700 computers were simultaneously used in 15 countries around the world. WISDOM demonstrated how grid computing can help drug discovery research by speeding up the whole process and reduce the cost to develop new drugs to treat diseases such as malaria. The sheer amount of data generated indicates the potential benefits of grid computing for drug discovery and indeed, other life science applications. Commercial software with a server license was successfully deployed on more than 1000 machines in the same time.

First docking results show that 10% of the compounds of the database studied may be hits. Top scoring compounds possess basic chemical groups like thiourea, guanidino, amino-acrolein core structure. Identified compounds are non peptidic and low molecular weight compounds. Future plans for the WISDOM initiative is first to process the hits again with molecular dynamics simulations in a grid environment. A second data challenge planned for the fall of 2006 is also under preparation to improve the quality of service and the quality of usage of the data challenge process on gLite. *In silico* docking at a large scale is a first step toward enabling the *in silico* drug discovery pipeline on a grid infrastructure against neglected diseases.
Key issues in the pharmaceutical process include cost and time reduction in a drug discovery development, security and data protection, fault tolerant and robust services and

infrastructure, and transparent and easy use of the interfaces. With the help of the grid, a such large scale *in silico* experimentation is possible.

The aim of the demonstration is to show the submission of *in silico* docking jobs at a large scale on the grid using the environment Taverna [6]. The user will prepare the jobs with a protein target, a compounds database, a docking software and a set of parameters, submit them on the grid, check their status and receive their output. Demonstration conditions will be similar to a large scale submission of jobs to reduce the necessary time for the *in silico* docking process. The demonstration will use the new middleware stack gLite [7], Lightweight Middleware for Grid Computing, developed within the EGEE project.

References

[1] J. Weisner, R. Ortmann, H. Jomaa, M. Schlitzer, Angew. New Antimalarial drugs, Chem. Int. 42 (2003) 5274-529.

[2] R.W. Spencer, Highthroughput virtual screening of historic collections on the file size, biological targets, and file diversity, Biotechnol. Bioeng 61 (1998) 61-67.

[3] A. Chien, I. Foster, D.Goddette, Grid technologies empowering drug discovery, Drug Discovery Today, 7 Suppl 20 (2002) 176-180.

[4] R. Buyya, K. Branson, J. Giddy, D. Abramson, The Virtual Laboratory. A Toolset to Enable Distributed Molecular Modeling for Drug Design on the WorldWide Grid, Concurrency Computat.: Pract. Exper. 15 (2003) 1–25.

[5] F. Gagliardi, B. Jones, F. Grey, M.E. Bégin, M. Heikkurinen, Building an infrastructure for scientific Grid computing: status and goals of the EGEE project, Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 363 (2005) 1729-1742 and EGEE Homepage: http://public.eu-egee.org

[6] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, P. Li, Taverna: A tool for the composition and enactment of bioinformatics workflows Bioinformatics Journal 20(17) pp 3045-3054, 2004, doi:10.1093/bioinformatics/bth361.

[7] gLite Homepage: http://glite.web.cern.ch/glite/