Large Scale Deployment of Molecular Docking Application on Computational Grid infrastructures for Combating Malaria

Vinod Kasam[1, 2*], Jean Salzemann[1*], Nicolas Jacq[1], Astrid Mass[2], Vincent Breton[1]
[1]*Laboratoire de Physique Corpusculaire, Campus des Cézeaux - 24 av des Landais, 63177 Aubière cedex, France*
[2]*Fraunhofer Institut for Algorithms and Scientific Computing (SCAI), Schloss Birlingoven, 53754, Sankt Augustin, Germany.*
**Vinod kasam: kasam@clermont.in2p3.fr*
Jean Salzemann: salzemann@clermont.in2p3.fr

Abstract

Computational grids are solutions for several biological applications like virtual screening or molecular dynamics where large amounts of computing power and storage are required. The WISDOM project successfully deployed virtual screening at large scale on EGEE grid infrastructures in the summer 2005 and achieved 46 million dockings in 45 days, which is equivalent to 80 CPU years. WISDOM is one good example of a successful deployment of an embarrassingly parallel application. In this paper, we describe the improvements in our deployment. We screened ZINC database against four targets implicated in malaria. During more than 2 months and a half, we have achieved 140 million dockings, representing an average throughput of almost 80,000 dockings per hour. This was made possible by the availability of thousands of CPUs through different infrastructures worldwide. Through the acquired experience, the WISDOM production environment is evolving to enable an easy and fault-tolerant deployment of biological tools.

1. Introduction

WISDOM stands for World-wide In Silico Docking On Malaria. Malaria together with many other tropical and protozoan diseases is one of the most neglected diseases by the developed countries as well as by the pharmaceutical industries. Plasmodium is the protozoan genus causing malaria. Due to very high costs associated to the drug discovery process as well as due to late stage attrition rates, novel and cost effective strategies are absolutely needed for combating the neglected diseases, especially malaria [1].

In silico screening of chemical compounds against a particular target is termed as virtual screening. The costs associated to the virtual screening of chemical compounds are significantly low when compared to screening of compounds in experimental laboratory. Beside the costs, virtual screening is fast and reliable [2,3]. However, it is computationally intensive; docking a single compound within the active site of a given receptor requires about 1 minute CPU. With the development of combinatorial chemistry technology, millions of different chemical compounds are now available in electronic databases [4]. To screen all these compounds and store the results is a real data challenge. To address this problem, computational grid infrastructures are used. WISDOM-I [5] is the first large scale deployment of molecular docking application on EGEE grid infrastructure. It took place from August 2005 to September 2005 and achieved 41 million dockings which is equivalent to 80 years of CPU. The docking was performed on Plasmepsins, a aspartic protease involved in haemoglobin degradation. On the biological front, three scaffolds were identified, of them one is guanidino scaffold which is likely to be novel as it was not known as a plasmepsin inhibitor before [6]. With the success achieved by the WISDOM-I project both on the computation and biological sides, several scientific groups around the world proposed targets implicated in malaria which led to the second assault on malaria. WISDOM-II project deals with several targets which are both X-ray crystal models and homology models. Targets from different classes of proteins are also being tested.

Table 1. Structural features of all the targets used in WISDOM II. * Target preparation and Screening are under process.

Target	Activity	Structure	PDB id	Resolution	Ligand	Co-factor
GST [9]	Detoxification	Dimer	1Q4J	2.2	GTX	NO
P. falciparum DHFR Wild type	DNA synthesis	Polymer	1J3I	2.33	WR99210	NADPH
P. falciparum DHFR [11] Quadruple mutant	DNA synthesis	Polymer	1J3K	2.10	WR99210	NADPH
P. vivax DHFR [10] Wild type	DNA synthesis	Polymer	2BL9	1.90	Pyrimethamine	NADPH
P. falciparum DHFR Double mutant	DNA synthesis	Polymer	2BLC	2.25	Des-Chloropyrimethamine	NADPH
Plasmodium Tubulin [12]*	Cell division	Monomer	Homology model	-	-	GTP

2. Materials and methods

Virtual screening by molecular docking requires target structure, chemical compound database and docking software. Several targets are used in the current project and are described in the table 1: Dihydrofolate reductase (DHFR), Glutatione–S-trasferase (GST) and Tubulin. The chemical compound database used is ZINC and the docking software used is FlexX [7, 8].

3. Procedure

The goal of WISDOM II is two folds, the biological goal is to find the best hits against the targets implicated in malaria and the computational goal is to keep improving the relevance of computational grids in drug discovery applications. We are going to discuss in details the virtual screening experimental set up and the grid architecture and deployment.

3.1. Virtual screening experimental setup

The structural features of all the targets screened are given in table 1. The complete virtual screening experiment is segmented into five different phases. i: Target preparation, ii: Compound database, iii: Validation of the docking experiment, iv: Screening, v: Result analysis.

3.1.1. Target preparation. A standard protocol is used while preparing the target structures. The initial coordinates for all the target structures are obtained from Brookhaven protein database (http://www.pdb.org). Depending upon the inclusion of the significant residues, active site is defined as 8.0 Å for GST, 10.0 Å for PfDHFR and PvDHFR around the co-crystallized ligand. X-ray crystal models for tubulin are not available; hence a homology model had to be built. Tubulin will be discussed in detail in another paper. Structural details on the first three targets used for screening in WISDOM-II are described in table 1.

3.1.1. Compound database. The compound library used for WISDOM was obtained from the ZINC database [13]. The ZINC database is a collection of 4.3 million chemical compounds ready for virtual screening from different vendors. We have chosen to use the ZINC library because ZINC is an open source database and the structures have already been filtered according to the Lipinski rules. Moreover the data are available in different file formats (Sybyl mol2 format, sdf and smiles). A total of 4.3 million compounds were downloaded from the ZINC database and screened against 4 different targets (table 1).

3.1.3. Validation of the docking experiment. Re-docking against the co-crystallized compound is performed to check and tune the docking experiment requirements. Re-docking serves as a control for finally selecting the parameters for target structure.

3.2. Grid infrastructure and Deployment

3.2.1. Grid Infrastructures. The deployments were achieved on several grid infrastructures: Auvergrid [14], EELA [15], EGEE [16], EUChinaGrid [17] and EUMedGrid [18]. All these infrastructures are actually using the same middleware, gLite. While EGEE is the main infrastructure offering the largest resources, they are all interconnected with EGEE. In the case of Auvergrid, it is even more evident as all the resources

available through the Auvergrid Virtual Organization (VO) are also shared with several EGEE VOs.

3.2.2. WISDOM production Environment. WISDOM environment has been used two times in previous large-scale experiments, WISDOM-I in the summer 2005 and a second deployment against avian flu in the spring 2006 [19]. WISDOM environment keeps evolving in order to make it more user friendly and easier to use by non grid experts. The environment was initially using scripts that took care of submitting the jobs and following up their respective status on the grid, giving feedback to the end user. These scripts were reorganized in two different and independent tasks: the submission of the jobs, the follow-up of the jobs, and eventually their resubmission, as well as the collection of the job status and their publication on a web site. These two processes can be started and run simultaneously, the second one being fed from the information provided by the first one.

With the gained experience from the previous deployments [20], the environment was continuously maintained to improve the fault-tolerance of the system, in implementing, for instance, a persistent environment, that can be stopped and restarted at any time without risking of loosing important information. The user can also manage jobs finely, for instance force the cancellation and resubmission of a scheduled job. Along with this, we tried to minimize the cost of the environment in terms of disk space and CPU consumption for the User Interface: for example, we are generating dynamically the job files which also allow the user to modify on the fly the configuration of the resource brokers and the jobs requirements while saving disk space. This way, the user is sure that the next submissions will take these modifications into account. The figure 1 shows the overall architecture of the environment.

The jobs are submitted directly to the grid Workload Management System (WMS), and are executed on the grid Computing Elements (CEs) and Worker Nodes (WNs). As soon as it is running, a job transfers all the files stored on the Storage Elements (SEs) via the Data Management System (DMS) of the grid with the gridFTP protocol. Then the FlexX software can start to process the dockings. During the job lifetime the status is retrieved from the User Interface and some statistics are generated and collected to a remote server which hosts a relational database and outputs these statistics through a web site. Once the job is finished, the outputs are stored back on the grid Storage Elements (SEs) and the useful docking results are inserted directly into a relational database.

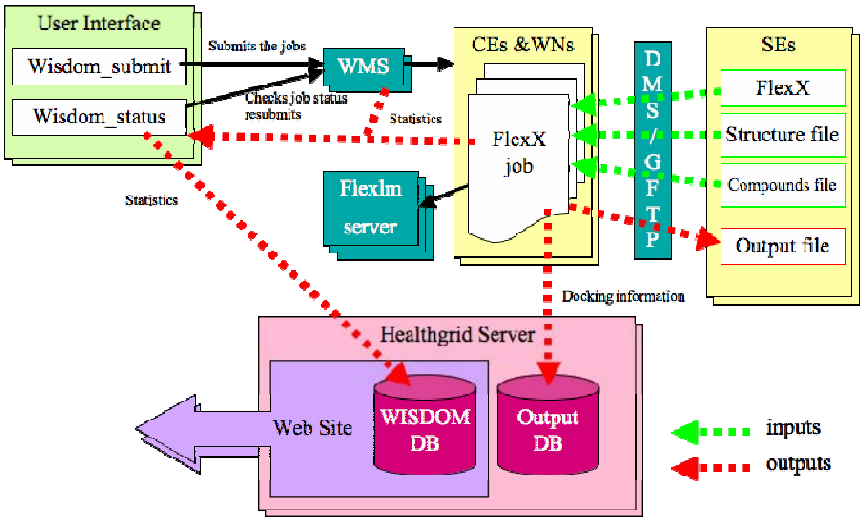

Figure 1. Schema of the production environment.

3.2.3. Data Challenge Deployment. The three groups of targets (GST, PvDHFR and PfDHFR) were docked against the whole ZINC database (4.3 millions of compounds) which was cut into 2,422 chunks of 1,800 compounds each. This splitting was chosen because we wanted to have an approximated processing time ranging from 20 to 30 hours for each job (one docking process takes from 40s to 1min depending on the CPU power and the data). The subsets had to be stored on the involved grid infrastructures. They were basically copied on a SE and were replicated on several locations whenever possible to improve fault-tolerance.

We define a WISDOM instance as being a target structure docked against the whole ZINC database, with a given parameter set. During the experiment a total number of 32 instances were deployed,

corresponding to an overall workload of 77,504 jobs, and up to 140,000,000 docking operations. On this total 32 instances, 29 were docked on EGEE, and 3 were run on Auvergrid, EELA and EuChinaGrid.

The job repartition was quite similar to the previous deployments, but here the United Kingdom and Ireland federation played an even bigger part. For instance, one of the British sites offered for quite a long period of time more than 1,000 free CPUs, which is half of the average used CPUs. The repartition of the jobs on the Auvergrid, EUChinaGrid and EELA infrastructures corresponds to 3% of the total 32 instances.

4. Results

4.1. Grid results analysis.

Table 2 shows the overall statistics of the deployment. The number of jobs was actually the number of awaited results, but far more jobs were actually submitted on the grid because of the resubmissions and failing jobs.

Anyhow the average docking throughput is coherent with the crunching factor, which basically represents the average number of CPUs used simultaneously all along the data challenge. If we consider 80,000 dockings per hour for 2,000 CPUs (the crunching factor) it means 40 dockings for one CPU per hour, which is coherent with the empiric observation of one docking process lasting approximately 1 minute on a 3.06 Intel Xeon processor. In the same logic, we can estimate that the instantaneous throughput peak would be obtained when the max number of CPUs were used (i.e.: 5,000), giving a throughput of approximately 200,000 dockings per hour.

In the table 2, the estimated grid success rate is the ratio between successful grid jobs on the total of submitted jobs. The success rate after output checking will consider just the jobs that succeeded in producing the good results; this score is thus lower. There are several explanations for these small values. Actually at the beginning of the data challenge, the observed grid success rate was about 80 to 90%, but it decreased constantly because of overload. The available disk space has also decreased dramatically on some WMS machines, the Resources Brokers, up to a point where some of the job data could not reach the CE. The Resource Brokers failed again to balance reasonably the jobs on the CEs, and some of them ended up with more than 500 jobs in queue. Actually, because of the automatic resubmission, this information should not be taken as an overall significant way to evaluate the efficiency of the grid, because the automatic resubmission guaranteed a successful job, and the aborted jobs are not staying a long time on the grid consuming useful resources. So one must keep in mind that the grid is a very dynamic system, and errors can occur at the last minute.

Table 2. Overall statistics of the deployment.

Number of Jobs	77,504
Total Number of completed dockings	156, 407,400
Estimated duration on 1 CPU	413 years
Duration of the experience	76 days
Average throughput	78,400 dockings/hour
Maximum number of loaded licences (concurrent running jobs)	5,000
Number of used computing elements	98
Average duration of a job	41 hours
Average crunching factor	1,986
Volume of output results	1,738 TB
Estimated distribution efficiency	39%
Estimated grid success rate	49%
Estimated success rate after output checking	37%

4.2. Preliminary docking results.

All the results are stored on the SEs of the Grid. Three different types of results are stored for each docking:

- Docking scores (free energy) of the ten best compound conformations after clustering.
- Interaction information between protein and compound for ten best conformations.
- 3D co-ordinates of the ten best conformations in MOL2 format.

As the result analysis of all the instances is under process, preliminary results of one instance, GST A chain score distribution is shown in figure 2. Approximately 100,000 compounds scored more than

its co-crystallized ligand (GTX) and these compounds will be given further focus.

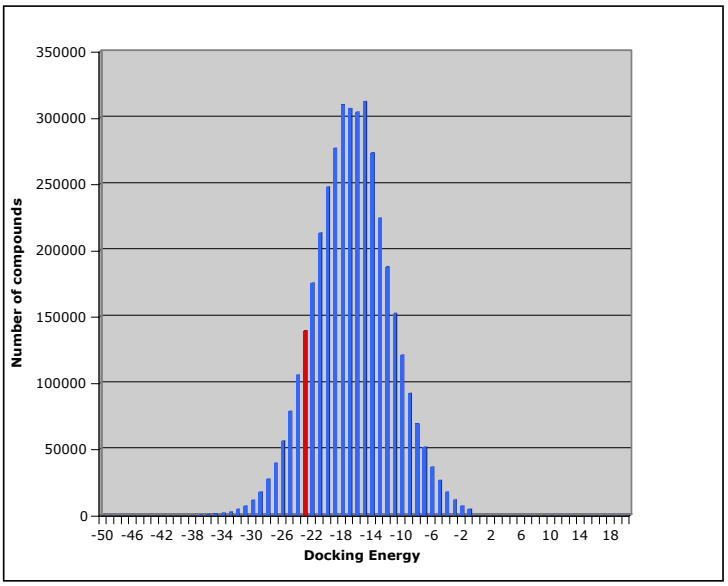

Figure 2. The repartition of docking energies of the ZINC database against GST A structure. (The red column represents a score of -24kj/Mol, which is the docking score of a co-crystallized ligand (GTX) of GST A chain).

4.3. Issues

As pointed out in the previous section and in the previous experiences of large scale deployment, the scheduling efficiency of the grid is still a major issue. The Resource Broker is still the main bottleneck, and even if used in high number (>15), is always a source of trouble. The sometimes unreliable and incomplete information provided by the Information System, which does not publish the available slots and VO limitations that would be mandatory to perform an efficient scheduling, is another point of failure. This deployment also shows that it is not possible to do a naive blacklisting of the failing resources, for the simple fact that virtually all the grid resources have produced aborted jobs, and this blacklisting should also take care of the site scheduled downtimes.

Another issue was that to be able to store and treat the data in a relational database. The machine hosting the database must have good performances to avoid its overloading by the number of queries coming from the grid. In this deployment we used a MySQL database and planned to put all the produced result in the same table, but finally we had to spit this database in several ones (one per target), because MySQL would not have been able to withstand the total number of records.

All these elements demonstrate clearly that even if the grid can show very good result in comparison to very simple architecture it is still missing robustness and reliability, and can indeed be improved performance-wise.

5. Conclusions

We have demonstrated the role and significance of computational grids in the life science applications such as structure based drug design. Large scale virtual screening on four different targets of malaria was performed in search for potential hits on several grid infrastructures: Auvergrid, EELA, EGEE, EUChinaGrid and EUMedGrid. One of our goals was to further demonstrate the impact of computational grids in life science applications like virtual screening where large amounts of computing power is required; we have achieved it by successfully screening the whole ZINC database for three malaria targets in 76 days instead of 413 years. We have reached during this ten-week period an average docking throughput of 78,400 dockings. MySQL databases are used for the analysis of the docking results, which will ease the final analysis of the virtual screening data. On the biological front, 1,738 Terabytes of valuable data has been produced. Analysis of the results (identification of the best hits) is under way: the best hits will be post processed by molecular dynamic simulations and tested in the experimental laboratories.

Acknowledgements. Information on WISDOM collaboration can be found at http://wisdom.healthgrid.org. The Enabling Grids for E-sciencE (EGEE) project is co-funded by the European Commission under contract INFSO-RI-031688. Auvergrid is a project funded by the Conseil Regional d'Auvergne. The BioinfoGRID project is co-funded by the European Commission under contract INFSO-RI-026808. The EMBRACE project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092. The authors acknowledge the contribution of Emmanuel Medernach (AuverGrid), Ignacio Blanquer (EELA), Gabri Aparico (EELA), Enrico Fattibene (EUChinaGrid), Kostas Koumantaros (EUMedGrid) and Domenico Vicinanza (EUMedGrid). We gratefully acknowledge the support of BioSolveIT who provided more than 6,000 free licenses for their commercial docking program FlexX.

6. References

[1] J. G. Breman, M. S. Alilio, A. Mills. Conquering the intolerable burden of malaria: what's new, what's needed: a summary. Am. J. Trop. Med. Hyg. 71S (2004) 1-15.

[2] K. H. Bleicher, H. J. Boehm, K. Mueller, A. I. Alanine. Hit And Lead Generation: Beyond High-Throughput Screening. Nat. Rev. Drug. Discov. , 2 (2003) 369-378.

[3] H. J. Boehm, G. Schneider. Virtual Screening For Bioactive Molecules. Chapter 1, High Throughput Screening and Virtual Screening Entry Points To Drug Discovery. Methods and Principles in Medicinal Chemistry, Volume 10, (2000).

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

[5] N. Jacq, J. Salzemann, Y. Legré, M. Reichstadt, F. Jacq, E. Medernach, M. Zimmermann, A. Maaß, M. Sridhar, K. Vinod-Kusam, J. Montagnat, H. Schwichtenberg, M. Hofmann, V. Breton, Grid enabled virtual screening against malaria, accepted for publication in Journal of Grid Computing (2007).

[6] V. Kasam, M. Zimmermann, A. Maaß, H. Schwichtenberg, A. Wolf, N. Jacq,V. Breton, M. Hofmann. Design of Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach on the EGEE Grid, submitted to Journal of Chemical Information and Modeling (2006).

[7] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, Predicting Receptor-Ligand interactions by an incremental construction algorithm, J. Mol. Biol. 261 (1996) 470-489.

[8] BioSolveIT Homepage: http://www.biosolveit.de

[9] M. Perbandt, C. Burmeister, R.D. Walter, C. Betzel, E. Liebau, Native and inhibited structure of a Mu class-related glutathione S-transferase from Plasmodium falciparum, J. Biol. Chem. 279 (2004) 1336-1342

[10] P. Kongsaeree, P. Khongsuk, U. Leartsakulpanich, P. Chitnumsub, B. Tarnchompoo, M.D. Walkinshaw, Y. Yuthavong. Crystal Structure of Dihydrofolate Reductase from Plasmodium Vivax: Pyrimethamine Displacement Linked with Mutation-Induced Resistance, Proc. Natl. Acad. Sci. USA. 2 (2005) 13046-13051.

[11] J. Yuvaniyama, P. Chitnumsub, S. Kamchonwongpaisan, J. Vanichtanankul, W. Sirawaraporn, P. Taylor, M.D. Walkinshaw, Y. Yuthavong. Insights into antifolate resistance from malarial DHFR-TS structures, Nat. Struct. Biol. 10 (2003) 357-365.

[12] Naomi S. Morrisette, Arpita Mitra, David Sept, L. David Sibley. Dinitroannilines bind Alpha-Tubulin to disrupt microtubules. Molecular Biology of Cell. 15 (2004) 1960-1968.

[13] A Free database for virtual screening. http://blaster.docking.org/zinc/ UCSF, University of California, San Francisco.

[14] Auvergrid, available at www.auvergrid.fr

[15] EELA, available at www.eu-eela.org

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

[17] EuChinaGrid, available at www.euchinagrid.org

[18] EuMedGrid, available at www.eumedgrid.org

[19] N. Jacq, V. Breton, H-Y. Chen,L-Y. Ho, M. Hofmann, H-C. Lee, Y. Legré, S. C. Lin, A. Maaß, E. Medernach, I. Merelli, L. Milanesi, G. Rastelli, M. Reichstadt, J. Salzemann, H. Schwichtenberg, M. Sridhar, V. Kasam, Y-T. Wu, M. Zimmermann. Virtual Screening on Large Scale Grids. Accepted for publication in Parallel Computing, (2007).

[20] H.-C. Lee, J. Salzemann, N. Jacq, H.-Y. Chen, L.-Y. Ho, I. Merelli, L. Milanesi, V. Breton, S. C. Lin, Y.-T. W. Grid-Enabled High-Throughput In Silico Screening Against Influenza A Neuraminidase, IEEE transactions on nanobioscience, (2006) 288-295.