Virtual Screening on Large Scale Grids

N. Jacq, Vincent Breton, H.-Y. Chen, L.Y. Ho, M. Hofmann, V. Kasam, H.-C. Lee, Yannick Legre, S. C. Lin, A. Maaß, et al.

Virtual Screening on Large Scale Grids

Nicolas Jacq[1], Vincent Breton[1], Hsin-Yen Chen[2], Li-Yung Ho[2], Martin Hofmann[3], Vinod Kasam[1+3], Hurng-Chun Lee[2], Yannick Legré[1], Simon C. Lin[2], Astrid Maaß[3], Emmanuel Medernach[1], Ivan Merelli[4], Luciano Milanesi[4], Giulio Rastelli[5], Matthieu Reichstadt[1], Jean Salzemann[1], Horst Schwichtenberg[4], Ying-Ta Wu[2], Marc Zimmermann[3]

[1] Laboratoire de Physique Corpusculaire, IN2P3 / UMR CNRS 6533,
24 avenue des Landais, 63177 AUBIERE, France
jacq@clermont.in2p3.fr
[2] Academia Sinica,
No. 128, Sec. 2, Academic Rd., NanKang, Taipei 115, Taiwan
ywu@gate.sinica.edu.tw
[3] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI),
Schloss Birlinghoven, 53754, Sankt Augustin, Germany
martin.hofmann@scai.fraunhofer.de
[4] CNR-Institute for Biomedical Technologies,
Via Fratelli Cervi 93, 20090 Segrate (Milan), Italy
luciano.milanesi@itb.cnr.it
[5] Dipartimento di Scienze Farmaceutiche, Università di Modena e Reggio Emilia,
Via Campi 183, 41100 Modena, Italy
rastelli.giulio@unimore.it

Abstract: Large scale grids for in silico drug discovery open opportunities of particular interest to neglected and emerging diseases. In 2005 and 2006, we have been able to deploy large scale virtual docking within the framework of the WISDOM initiative against malaria and avian influenza requiring about 100 years of CPU on the EGEE, Auvergrid and TWGrid infrastructures. These achievements demonstrated the relevance of large scale grids for the virtual screening by molecular docking. This also allowed evaluating the performances of the grid infrastructures and to identify specific issues raised by large scale deployment.

Keywords: large scale grids, virtual screening, malaria, avian influenza

1 Introduction

In silico drug discovery is one of the most promising strategies to speed-up the drug development process [1]. Virtual screening is about selecting in silico the best candidate drugs acting on a given target protein [2]. Screening can be done in vitro but it is very expensive as there are now millions of chemicals that can be synthesized [3]. A reliable way of in silico screening could reduce the number of molecules

required for in vitro and then in vivo testing from a few millions to a few hundreds.

In silico drug discovery should foster collaboration between public and private laboratories. It should also have an important societal impact by lowering the barrier to develop new drugs for rare and neglected diseases [4]. New drugs are needed for neglected diseases like malaria where parasites keep developing resistance to existing drugs or sleeping sickness for which no new drug has been developed for years. New drugs against tuberculosis are also needed as the treatment now takes several months and is therefore hard to manage in developing countries.

However, in silico virtual screening requires intensive computing, in the order of a few TFlops per day to compute 1 million docking probabilities or for the molecular modelling of 1,000 compounds on one target protein. Access to very large computing resources is therefore needed for successful high throughput virtual screening [5]. Grids now provide such resources. A grid infrastructure such as EGEE [6] today provides access to more than 30,000 computers and is particularly suited to compute docking probabilities for millions of compounds. Docking is only the first step of virtual screening since the docking output data has to be processed further [7].

After introducing the case for large scale in silico docking on grids, this paper will present the context and objectives of the two initiatives against malaria and avian influenza in chapters 2 and 3. In chapter 4, the grid infrastructures and the environments used for large scale deployment are briefly described. Results and performance of grid environment are discussed in chapter 5. In chapter 6, we will also suggest some perspectives for the coming years.

2 Large Scale In Silico Docking against Malaria

2.1 Introduction

The number of cases and deaths from malaria increases in many parts of the world. There are about 300 to 500 million new infections, 1 to 3 million new deaths and a 1 to 4% loss of gross domestic product (at least $12 billion) annually in Africa caused by malaria. The main causes for the comeback of malaria are that the most widely used drug against malaria, chloroquine, has been rendered useless by drug resistance in much of the world [8,9] and that anopheles mosquitoes, the disease carrier, have become resistant to some of the insecticides used to control the mosquito population.

Genomics research has opened up new ways of finding novel drugs to cure malaria, vaccines to prevent malaria, insecticides to kill infectious mosquitoes and strategies to prevent development of infectious sporozoites in the mosquito [10]. These studies require more and more in silico biology; from the first steps of gene annotation via target identification to the modeling of pathways and the identification of proteins mediating the pathogenic potential of the parasite. Grid computing supports all of these steps and, moreover, can also contribute significantly to the monitoring of ground studies to control malaria and to the clinical tests in plagued areas.

A particularly computing intensive step in the drug discovery process is virtual

screening which is about selecting in silico the best candidate drugs acting on a given target protein. Screening can be done in vitro using real chemical compounds, but this is a very expensive not necessarily error-free undertaking. If it could be done in silico in a reliable way, one could reduce the number of molecules requiring in vitro and then in vivo testing from a few millions to a few hundreds [11]. Advance in combinatorial chemistry has paved the way for synthesizing millions of different chemical compounds. Thus there are millions of chemical compounds available in pharmaceutical laboratories and also in a very limited number of publicly accessible databases.

The WISDOM experiment ran on the EGEE grid production service during summer 2005 to analyze one million drug candidates against plasmepsin, the aspartic protease of plasmodium responsible for the initial cleavage of human haemoglobin.

2.2 Objectives

Grid Objective

A large number of applications are already running on grid infrastructures. Even if many have passed the proof of concept level [12], only few are ready for large-scale production with experimental data. Large Hadron Collider experiments at CERN, like the ATLAS collaboration [13], have been the first to test a large data production system on grid infrastructures [14]. In a similar way, WISDOM [15] aimed at deploying a scalable, CPU consuming application generating large data flows to test the grid infrastructure, operation and services in very stressing conditions.

Docking is – along with BLAST [16] homology searches and some folding algorithms – one of the most prominent applications that have successfully been demonstrated on grid testbeds [17,18]. It is typically an embarrassingly parallel application, with repetitive and independent calculations. Large resources are needed in order to test a family of targets, a significant amount of possible drug candidates and different virtual screening tools with different parameter/scoring settings. This is both a computational and data challenge problem to distribute millions of docking comparisons with millions of small compound files.

Moreover, docking is the only application for distributed computing that has prompted the uptake of grid technology in the pharmaceutical industry [19]. The WISDOM scientific results are also a means of making a demonstration of the EGEE grid computing infrastructure for the end users community, of illustrating the usefulness of a scientifically targeted Virtual Organization, and of fostering an uptake of grid technologies in this scientific area [20].

Biological Objective

Malaria is a dreadful disease caused by a protozoan parasite, plasmodium. A new strategy to fight malaria investigated within WISDOM aims at the haemoglobin metabolism, which is one of the key metabolic processes for the survival of the parasite. Plasmepsin, the aspartic protease of plasmodium, is responsible for the initial cleavage of human haemoglobin and later followed by other proteases [21]. There are

ten different plasmepsins coded by ten different genes in *Plasmodium falciparum* [8]. High levels of sequence homology are observed between different plasmepsins (65-70%). Simultaneously they share only 35% sequence homology with its nearest human aspartic protease, Cathepsin D4 [9]. This and the presence of accurate X crystallographic data make plasmepsin an ideal target for rational drug design against malaria.

Docking is a first step for in silico virtual screening. Basically, protein-compound docking is about estimating the binding energy of a protein target to a library of potential drugs using a scoring algorithm. The target is typically a protein, which plays a pivotal role in a pathological process, e.g. the biological cycles of a given pathogen (parasite, virus, bacteria…). The goal is to identify which molecules could dock on the protein's active sites in order to inhibit its action and therefore interfere with the molecular processes essential for the pathogen. Libraries of compound 3D structures are made openly available by chemistry companies, which can produce them. Many docking software products are available either open-source or through a proprietary license.

3 In silico Docking against Avian Influenza

3.1 Introduction

The first large scale docking experiment focused on virtual screening for neglected diseases but new perspectives appear also for using grids to address emerging diseases. While the grid added value for neglected diseases is related to their cost effectiveness as compared to in vitro testing, grids are also extremely relevant when time becomes a critical factor. A collaboration between Asian and European laboratories has analyzed 300,000 possible drug compounds against the avian influenza virus H5N1 using the EGEE grid infrastructure in April and May 2006 [22]. The goal was to find potential compounds that can inhibit the activities of an enzyme on the surface of the influenza virus, the so-called neuraminidase, subtype N1. Using the grid to identify the most promising leads for biological tests could speed up the development process for drugs against the influenza virus.

3.2 Objectives

Grid Objective

Beside the biological goal of reducing time and cost of the initial investment on 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 environment based on what has been learnt in the previous challenge against malaria; the other is to test another environment which enables users to control the massive

molecular dockings on the grid. Therefore, two Grid tools were used in parallel in this second large scale deployment. An enhanced version of WISDOM high-throughput workflow was designed to achieve the first goal and a lightweight framework called DIANE [23,24] was introduced to carry a significant fraction of the deployment for implementing and testing the new scenario.

Biological Objective
The potential for reemergence 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. Indeed, an increase of transmission incidents suggests a risk of human-to-human transmission [25]. In addition, the report of the development of drug resistant variants [26] is of potential concern. Two of the present drugs (oseltamivir and zanamivir) were discovered through structure-based drug design targeting influenza neuraminidase (NA), a viral enzyme that cleaves terminal sialic acid residues 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 modelled NA variant structures and structure-based design.

To achieve this goal, molecular docking engines, such as AutoDock [27], carry out a quick conformational search for 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 engines 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.

4 The Grid Tools

4.1 The Grid Infrastructures

Compared to WISDOM which used only the EGEE infrastructure, the large scale deployment against avian influenza used three infrastructures which are sharing the same middleware (LCG2) and also common services: Auvergrid, EGEE and TWGrid.
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 science, environment and chemistry.
TWGrid is responsible for operating a Grid Operation Centre in Asia-Pacific region. Apart from supporting the worldwide 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 EGEE project [6] (Enabling Grid for E-sciencE) brings together experts from over 27 countries with the common aim of building on recent advances in grid technology and developing a service grid infrastructure which is available to scientists 24 hours a day. The project aims to provide researchers in academia and industry with access to major computing resources, independent of their geographic location. The EGEE infrastructure is now a production grid with a large number of applications installed and used on the available resources. The infrastructure involves more than 180 sites spread out across Europe, America and Asia.

The two applications described in this paper were deployed within the framework of the biomedical Virtual Organization. Resource nodes available for biomedical applications scale up to 3,000 CPUs and 21 TB disk space. The resources are spread over 15 countries.

4.2 The WISDOM Production Environment

A large-scale deployment requires the development of an environment for job submission and output data collection. The EGEE Workload Management System [6] is responsible for the management and monitoring of jobs. The Job Description Language is used to describe the jobs and their requirements like targeted workstations. A set of services running on the Resource Broker machine matches job requirements expressed in the Job Description Language files to the available resources (as gathered from the Information System), distributes the jobs over the matching Computing Elements, tracks the job status, and allows users to retrieve their job output. Computing Elements are frontals to computing centre clusters. They plan jobs execution over the available Worker Nodes through batch schedulers.

A number of issues need to be addressed to achieve significant acceleration from the grid deployment:

- Grid performances are impacted by the amount of data moved around at job submission. Consequently, the files providing the 3D structure of targets and compounds should preferably be stored on grid storage elements in preparation for the data challenge.
- The rate at which jobs are submitted to the grid resource brokers must be carefully monitored in order to avoid their overload. The job submission scheme must take into account this present limitation of the EGEE brokering system.
- Grid submission process introduces significant delays for instance at the level of resource brokering. The jobs submitted to the grid computing nodes must be sufficiently long in order to reduce the impact of this middleware overhead.

The WISDOM production environment [28] was designed to achieve production of a large amount of data in a limited time using EGEE, Auvergrid and TWGrid middleware services. Three packages were developed in Perl and Java. Their entry points are a simple command line tool. The first package installs the application components (software, compounds database…) on the grid computing nodes. The second package tests these components. The third package monitors the submission

and the execution of the WISDOM jobs. Figure 1 illustrates the packages and the steps of the jobs execution.

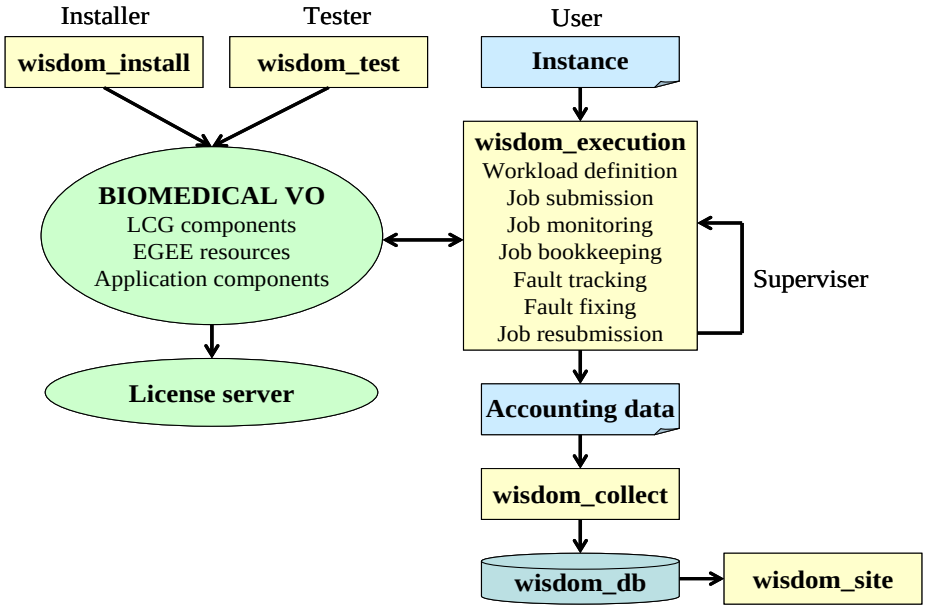

Fig. 1. Design of the WISDOM production system

The environment was improved to address limitations and bottlenecks identified during the first data challenge against Malaria deployed in the summer of 2005 on the EGEE infrastructure.

- The number of resource broker machines and the rate at which the jobs were submitted to these grid services were reduced to avoid their overload.
- Another improvement concerned the resubmission process after a job failure, which was redesigned to avoid "sinkhole" effect on a failing grid-computing node. Automatic resubmission was replaced by the manual intervention of the WISDOM production user.
- Mis-configuration of the grid sites was an important cause of failure (see table 1). In the first data challenge, the compound database was installed on each Storage Element of the grid, and the docking binary was installed on each Software Repository of the grid Computing Elements. The Software Repository is dedicated to a Virtual Organization for installing binaries on each grid site. But many jobs failed to access database and software because of site mis-configuration appearing during the deployment. In the data challenge against avian influenza, all the input data were stored on only three Storage Elements. Advantages are the simpler management of only three grid sites and the use of basic transfer commands instead of more complex software repository commands.

4.3 The DIANE Framework

DIANE (Distributed Analysis Environment, http://cern.ch/diane/) is a lightweight distributed framework for parallel scientific applications in master-worker model [23,24]. It assumes that a job may be split into a number of independent tasks, which is a typical case in many scientific applications. It has been successfully applied in a number of applications ranging from image rendering to data analysis in high-energy physics.

As opposed to standard message passing libraries such as MPI [29], the DIANE framework takes care of all synchronization, communication and workflow management details on behalf of the application. The execution of a job is fully controlled by the framework, which decides when and where the tasks are executed. Thus the application is very simple to program and contains only the essential code directly related to the application itself without bothering about networking details. Aiming to efficiently bridge underlying distributed computing environments and application centric user interface, DIANE itself is a thin software layer which can easily work on top of more fundamental middleware such as LSF, PBS or the grid Resource Brokers. It may also work in a standalone mode and does not require any complex underlying software.

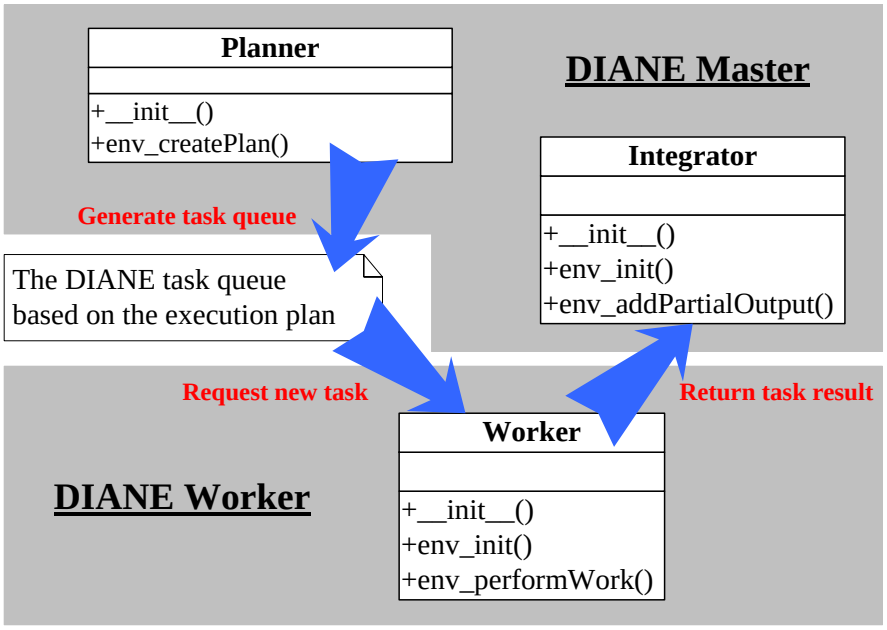

Fig. 2. Template of DIANE application plug-ins

As a framework, DIANE provides an adapter for applications. Figure 2 shows the template of DIANE application plug-ins. A complete DIANE application plug-in should implement three major Python objects: the Planner and the Integrator objects implement the job splitting and result merging, respectively; while the logic of the

Worker object concentrates on the execution of the individual task. When a DIANE job is started by a user, both the Planner and the Integrator objects are invoked by a master agent usually executed on the user's desktop, and typically the worker agents are submitted to run on distributed CPUs such as the grid Worker Nodes.

Once the worker agent is launched, it first registers itself with the master agent. In the second step, a channel is established for pulling the tasks from the queue held by the master agent. When the worker agent has finished the individual task, the result is returned and merged on the master. The pulling-executing-returning cycle will iterate until all the tasks are accomplished. The same channel is also used to profile the worker agent's health and to support user interaction with the task. The whole DIANE framework is written in Python and the communication between the master agent and the worker agents is based on the CORBA protocol.

Further details on DIANE can be found in [23] and [24].

5 Results

5.1 WISDOM Results

The WISDOM experiment ran on the EGEE grid production service from 11 July 2005 until 19 August 2005. It saw over 42 million docked compounds, the equivalent of 80 years on a single PC, in about 6 weeks. FlexX [30], a commercial docking software with a license server, was successfully deployed during three weeks on more than 1,000 machines at the same time. Then Autodock, a free for academic docking software, was deployed during three weeks and used up to 1,700 computers 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 analysis of the performance and the cause of failures are detailed in [28] and summarized below.

Grid Performances

The large scale deployment was a very useful experience to identify the limitations and bottlenecks of the EGEE infrastructure.

The success rate formula used for measuring EGEE Quality of Service stands as follows:

Successful jobs / (submitted jobs – cancelled jobs)

where successful jobs are jobs which have been executed successfully, submitted jobs are the jobs launched by the user from the User Interface, cancelled jobs are jobs cancelled by the user. The proposed definition of the success rate is not completely relevant from a user point of view as a successful job as seen from the grid can be unsuccessful from a user perspective if it did not produce the expected output data. As a consequence, our definition of the success rate during these experiments took into

account the quality of the data produced. Successful jobs were defined as jobs executed successfully on the grid which produced the expected data. The content of the output data file was checked through an automatic procedure. All sources of failures were registered during the 6 weeks; the failure rate was defined as the ratio of failures of a given origin to the total number of jobs submitted minus the number of jobs cancelled. We also introduced the concept of grid success rate defined as the success rate after checking output data and subtracting WISDOM and server license failures. All the unsuccessful tasks are resubmitted and recomputed by the WISDOM production environment.

The success rate after checking output data was 46% while the grid success rate was of the order of 63%. This means that a large fraction of the jobs had to be resubmitted. This generated a significantly larger workload for the users. The different sources of failures are identified on table 1 with their corresponding rates.

Table 1. Origin of failures during the WISDOM deployment with their corresponding rates

	Rate	Reasons
Success rate after checking output data	46 %	
Workload Management failure	10 %	Overload, disk failure
		Misconfiguration, disk space problem
		Air-conditioning, power cut
Data Management failure	4 %	Network / connection
		Power cut
		Other unknown causes
Sites failure	9 %	Misconfiguration, tar command, disk space
		Information system update
		Job number limitation in the waiting queue
		Air-conditioning, power cut
Unclassified	4 %	Lost jobs
		Other unknown causes
Server license failure	23 %	Server failure
		Power cut
		Server stop
WISDOM failure	4 %	Job distribution
		Human error
		Script failure

Two docking tools were used on the grid during WISDOM deployment, namely Autodock and FlexX. In the phase where proprietary licensed software FlexX was deployed on the grid, the dominant origin of failure was the license server in charge of distributing tokens to all the jobs running on the grid. Each job using FlexX software was contacting the Flexlm server at the beginning of the job, asked for a license and then was able to run without connection to the license server. When the server is down (for example by a power cut), the jobs can not connect to the server and are considered to have failed. The development of a grid service to manage proprietary licensed software is under way to address this single point of failure ignored by the information system.

The second most important sources of failure were workload management and site failures, including overload, disk failure, node mis-configuration, disk space problem,

air-conditioning and power cut. To improve the job submission process, an automatic resubmission of jobs was included in the WISDOM execution environment. However, the consequence of automatic resubmission was the creation of several "sinkhole" effects where all the jobs were attracted to a single node. These sinkhole effects were observed when the status of a Computing Element was not correctly described in the information system. If a Computing Element already loaded is still viewed as completely free by the Information System, it keeps receiving jobs from the Resource Broker. If the Computing Element gets down, all jobs are aborted. Even if the Computing Element can support the excessive number of jobs, the processing time is going to be very long.

The WISDOM production system developed to submit the jobs on the grid accounted for a small fraction of the failures, as did also the grid data management system. About 1 terabyte of data was produced by the 72,000 jobs submitted. Collection and registration of these output data turned out to be a heavy task. The grid data management services allowed replicating all the output files for backup. However, they did not allow storing all the results directly in a database to ease the final analysis by the end-user. WISDOM experience was the opportunity to learn from biochemists the relevant parameters to extract from the docking output file so that the next implementation of the production environment will offer this service to the end-users.

Finally, unclassified failures accounted for 4% of inefficiency. This illustrates the work which is still needed to improve grid monitoring.

In terms of grid throughput, the resource brokers significantly limited the rate at which the jobs could be submitted. Another significant source of inefficiency came from the difficulty for the grid information system to provide all the relevant information to the resource brokers when they distribute the jobs on the grid. Therefore, job scheduling was a time-consuming task for the WISDOM users during all the data challenge due to the encountered limitations of the information system, the computing elements and the resource brokers. The achieved throughput averaged about 12 docked compounds per second during the whole computing challenge.

Several of the issues identified during WISDOM deployment were improved for the second large scale docking targeted on avian influenza which is described in the next chapter.

WISDOM Biological Results

Post-processing of the huge amount of data generated was a very demanding task as millions of docking scores had to be compared. At the end of the large scale docking deployment, the best 1,000 compounds based on scoring were selected thanks to post-processing ranking jobs deployed on the grid. They were inspected individually. Several strategies were employed to reduce the number of false positives. A further 100 compounds were selected for post-processing. These compounds had been selected based on the docking score, the binding mode of the compound inside the binding pocket and the interactions of the compounds to key residues of the protein.

There are several scaffolds in the 100 compounds selected for post processing. The scaffolds urea, thiourea, and guanidino analogues are most repeatedly identified in the top 1,000 compounds. Some of the compounds identified were similar to already known plasmepsin inhibitors, like the urea analogues which were already

established as micro molar inhibitors for plasmepsins (Walter Reed compounds) [31]. This indicates that the overall approach is sensible and that large scale docking on computational grids has potential to identify new inhibitors. The guanidino analogues can become a novel class of plasmepsin inhibitors.

5.2 Results from the Large Scale Deployment against Avian Influenza

Table 2 summarizes the achieved deployments using WISDOM and DIANE environments. The crunching factor represents the gain of time obtained thank to the grid deployment. It is simply obtained by dividing the total CPU time by the execution duration. The approximated distribution efficiency is defined as the ratio between the crunching factor and the maximum number of concurrently running CPUs.

Table 2. Statistical summary of the WISDOM and DIANE activities

	WISDOM	DIANE
Total number of completed dockings	2 * 106	308,585
Estimated duration on 1 CPU	88.3 years	16.7 years
Duration of the experience	6 weeks	4 weeks
Cumulative number of grid jobs	54,000	2,585
Maximum number of concurrent CPUs	2,000	240
Number of used Computing Elements	60	36
Crunching factor	912	203
Approximated distribution efficiency	46%	84%

The WISDOM production environment was used to distribute 54,000 jobs on 60 grid Computing Elements to dock about 2 million pairs of target and chemical compound for a total amount of about 88 CPU years. Because the grid resources were used by other Virtual Organizations during the deployment, a maximum of 2,000 CPUs were concurrently running at the same time. For the DIANE part, we were able to complete 308,585 docking runs (i.e. 1/8 of the whole deployment) in 30 days using the computing resources of 36 grid nodes. A total number of 2,585 DIANE worker agents have been running as grid jobs during that period and the DIANE master concurrently maintained 240 of them. The distribution of those grid jobs in terms of the regions of the world is shown on Figure 3. About 600 gigabytes of data has been produced on the grid during the deployment.

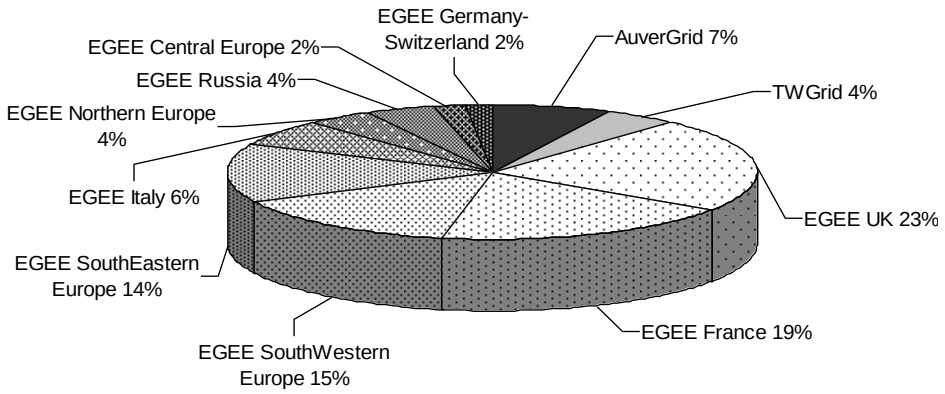

Fig. 3. Distribution of the grid jobs in different region.

Grid Performances

Since a grid is a dynamic system in which the status of resources may change without central control, transient problems occur which cause job failures. In the WISDOM activity, the success rate after checking output data is 70% and the grid success rate is 80%; the observed failures were mainly due to errors at job scheduling time because of misconfiguration of grid Computing Elements and frequent last-minute errors in the transfer of results to the grid Storage Elements. Compared to the previous large scale deployment, improvement is significant as the observed success rates were respectively 46 and 63%. The last-minute error in output data transfer is particularly expensive since the results are no longer available on the grid Worker Node although they might have been successfully produced.

In DIANE, similar job failure rates were also observed; nevertheless, the failure recovery mechanism in DIANE automated the re-submission and guaranteed a finally fully complete job. On the other hand, the feature of interactively returning part of the computing efforts during the runtime (e.g. the output of each docking) also introduces a more economical way of using the grid resources.

For the instances submitted using WISDOM production environment, the overall crunching factor was about 912. The corresponding distribution efficiency was estimated to 46%. This is due to the known issue of long job waiting time in the current EGEE production system.

The task pull model adopted by DIANE allows isolating the scheduling overhead of the grid jobs and is therefore expected to achieve a better distribution efficiency. During the deployment, DIANE was able to push the efficiency to higher than 80% within the scope of handling intermediate scale of distributed docking. A fare comparison with WISDOM can only be made if the improvement of the DIANE framework is tested in a large scale as the exercise of WISDOM.

Biological Results

Two sets of re-ranked data for each target were made (QM-MM method and selection by the study of the interactions between the compound and the target active site). The

top 15% is about 45,000 compounds for each target. This set will be publicly available for the scientific community working on the avian influenza neuraminidase. The top 5% is about 2,250 compounds for each target. This set will be refined by different methods (molecular modeling, molecular dynamics…). The analysis will indicate which residue mutation is critical, which chemical fragments are preferred in the mutation sub sites and other information for lead optimization to chemists. Finally, at least 25 compounds will be assayed experimentally at the Genomic Research Center from Academia Sinica.

6 Perspectives

The results obtained by the first two high throughput virtual docking deployments have opened important perspectives. On the grid side, developments are under way to further improve the WISDOM and DIANE environments to improve the quality of service offered to the end-users. From an application point of view, beyond the necessity to analyze the results obtained on malaria and avian influenza, it is particularly relevant to further enable the deployment of virtual screening on large scale grids.

6.1 Perspective on the Grid Side: WISDOM-II

The impact of the first WISDOM deployment has significantly raised the interest of the research community on neglected diseases so that several laboratories all around the world have expressed interest to propose targets for a second large scale deployment called WISDOM-II. Contacts have been established with research groups from Italy, United Kingdom, Venezuela, South Africa and Thailand and discussions are under way to finalize the list of targets to be docked.

Similarly, several grid projects have expressed interest to contribute to the WISDOM initiative by providing computing resources (Auvergrid, EELA (http://www.eu-eela.org/), South East Asia Grid, Swiss Biogrid (http://www.swissbiogrid.com/)) or by contributing to the development of the virtual screening pipeline (Embrace (http://www.embracegrid.info/), BioinfoGRID (http://www.bioinfogrid.eu/)).

6.2 From Virtual Docking to Virtual Screening

While docking methods have been significantly improved in the last years by including more through compound orientation search, additional energy contributions and/or refined parameters in the force field, it is generally agreed that docking results need to be post-processed with more accurate modeling tools before biological tests are undertaken. Molecular dynamics (MD) [32] has great potential at this stage: firstly, it enables a flexible treatment of the compound/target complexes at room temperature for a given simulation time, and therefore is able to refine compound orientations by finding more stable complexes; secondly, it partially solves

conformation and orientation search deficiencies which might arise from docking; thirdly, it allows the re-ranking of molecules based on more accurate scoring functions.

Figure 4 illustrates how the best hits coming out the docking step need to be further processed and analyzed with MD.

However, for the same number of compounds, MD analysis requires much heavier computing than docking. Consequently, MD can only be applied to a restricted number of compounds, usually the best hits coming out of the docking step. MD and subsequent free-energy analysis most often changes significantly the scoring of the best compounds and it is therefore very important to apply it to as many compounds as possible. As a consequence, grids appear very promising to improve the virtual screening process by increasing the number of compounds that will be processed using MD.

For instance, running MD analysis with Amber 8 [33] software on one protein target and 10,000 compounds requires about 50 CPU years for a model with a few thousands atoms and a few tens of thousands of steps. This process produces only tens gigabytes of data. One run with only one compound takes 44 hours (computing time heavily depends on choice of conditions, like explicit water simulations, or generalized born simulations).

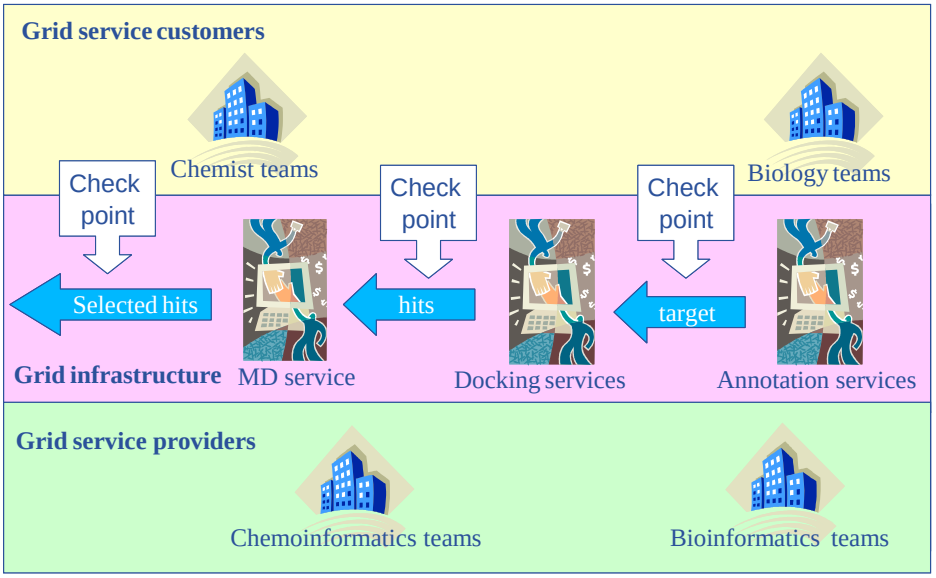

Fig. 4. First steps of the *in silico* drug discovery pipeline on the grid

Deployment of MD computing is relevant on grids of clusters such as EGEE and grids of supercomputers such as DEISA (http://www.deisa.org/), which is an European Supercomputing Service built on top of existing national services and based on the deployment and operation of a persistent, production quality, distributed supercomputing environment with continental scope. MD computations of large molecules are significantly faster on a supercomputer.

Within the framework of the BioinfoGRID European project which aims to expand the grid awareness inside the bioinformatics community, focus will be put on the reranking of the best scoring compounds coming out of WISDOM. The goal will be to deploy on at least one of the European grid infrastructures MD software to rerank the best compounds before in vitro testing.

7 Conclusion

Large scale grids offer unprecedented perspectives for in silico drug discovery. This paper has presented pioneering activities in the field of grid enabled virtual screening against neglected and emerging diseases in Europe. This activity started with a large scale docking deployment against malaria on the EGEE infrastructure in 2005, which was followed by another large scale deployment focused on avian influenza in the spring of 2006. A second large scale deployment on neglected diseases is foreseen to take place in the fall: it will involve several European projects, which will bring additional resources and complementary contributions in order to enable a complete virtual screening process on the grid.

These deployments were achieved using two different systems for submission and monitoring of virtual docking jobs. The WISDOM system was designed to achieve production of a large amount of data in a limited time while the DIANE framework was designed as a lightweight distributed framework for parallel scientific applications in master-worker model. Both systems were able to provide high throughput virtual docking. Development of a new system merging functionalities from both WISDOM and DIANE frameworks is under way in the perspective of WISDOM-II, second large scale deployment on neglected diseases.

Acknowledgment. This work was supported in part by Auvergrid, EGEE and TWGrid projects. This work took place in collaboration with the Embrace network of excellence and the BioinfoGRID project. 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 TWGrid is funded by the National Science Council of Taiwan. The authors express particular thanks to the site managers in EGEE, TWGrid, AuverGrid for operational supports, the LCG ARDA group for the technical support of DIANE, and the Biomedical Task Force for its participation to the WISDOM deployment. The following institutes contributed computing resources to the data challenge: ASGC (Taiwan); NGO (Singapore); IPP-BAS, IMBM-BAS and IPP-ISTF (Bulgaria); CYFRONET (Poland); ICI (Romania); CEA-DAPNIA, CGG, IN2P3-CC, IN2P3-LAL, IN2P3-LAPP and IN2P3-LPC (France); SCAI (Germany); INFN (Italy); NIKHEF, SARA and Virtual Laboratory for e-Science (Netherlands); IMPB RAS (Russia); UCY (Cyprus); AUTH FORTH-ICS and HELLASGRID (Greece); RBI (Croatia); TAU (Israel); CESGA, CIEMAT,

CNB-UAM, IFCA, INTA, PIC and UPV-GryCAP (Spain); BHAM, University of Bristol, IC, Lancaster University, MANHEP, University of Oxford, RAL and University of Glasgow (United Kingdom).

References

1. BCG Estimate: A Revolution in R&D, The Impact of Genomics. (2001)
2. Lyne, P.D.: Structure-based virtual screening: an overview. Drug Discov. Today 7 (2002) 1047–1055
3. Congreve, M., et al.: Structural biology and drug discovery. Drug Discov Today 10 (2005) 895-907
4. Nwaka, S., Ridley, RG.: Virtual drug discovery and development for neglected diseases through public-private partnerships. Nat Rev Drug Discov 2 (2003) 919-28
5. Chien, A., et al.: Grid technologies empowering drug discovery. Drug Discovery Today 7 (2002) 176-180
6. Gagliardi, F., et al.: 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 http://www.eu-egee.org/
7. Ghosh, S., et al.: Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol. 10 (2006) 194-202
8. Coombs, G. H., et al.: Aspartic proteases of plasmodium falciparum and other protozoa as drug targets. Trends parasitol. 17 (2001) 532-537
9. Weisner, J., et al.: Angew. New Antimalarial drugs, Chem. Int. 42 (2003) 5274-529
10. Curtis, C.F., Hoffman, S.L.: Science 290 (2000) 1508-1509
11. Spencer, R.W.: Highthroughput virtual screening of historic collections on the file size, biological targets, and file diversity. Biotechnol. Bioeng 61 (1998) 61-67
12. Jacq, N., et al.: Grid as a bioinformatics tool, Parallel Computing. 30 (2004) 1093-1107
13. Campana, S., et al.: Analysis of the ATLAS Rome Production Experience on the LHC Computing Grid. IEEE International Conference on e-Science and Grid Computing (2005)
14. Bird, I., et al.: Operating the LCG and EGEE production Grids for HEP. Proceedings of the CHEP'04 Conference (2004)
15. Breton, V., et al.: Grid added value to address malaria. Proceedings of the 6-th IEEE International Symposium on Cluster Computing and the Grid 40 (2006) and http://wisdom.healthgrid.org/.
16. Altschul, S.F., et al.: Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403-410
17. Buyya, R., et al., The Virtual Laboratory: A Toolset to Enable Distributed Molecular Modelling for Drug Design on the World-Wide Grid, J. Concurrency and Computation: Practice and Experience, (2002).
18. Richards, W.G., Virtual screening using grid computing: the screensaver project, Nature Reviews Drug Discovery 1 551-555 (2002).
19. Ziegler, R.: Pharma GRIDs: Key to Pharmaceutical Innovation ?, Proceedings of the HealthGrid conference 2004 (2004)
20. Jacq, N., et al.: Demonstration of In Silico Docking at a Large Scale on Grid Infrastructure. Studies in Health Technology and Informatics 120 (2006) 155-157
21. Francis, S. E., et al.: Hemoglobin metabolism in the malaria parasite plasmodium falciparum. Annu.Rev. Microbiol. 51 (1997) 97-123
22. Lee, H.-C., et al.: Grid-enabled High-throughput in silico Screening against influenza A Neuraminidase. to be published in IEEE Transaction on Nanobioscience (2006)
23. Moscicki, J.T.: DIANE - Distributed Analysis Environment for GRID-enabled Simulation and Analysis of Physics Data. NSS IEEE 2004 (2004)

24. Moscicki, J.T., et al.: Biomedical Applications on the GRID: Efficient Management of Parallel Jobs. NSS IEEE 2003 (2003)
25. Li, K.S., et al.: Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430 (2004) 209-213
26. de Jong, M. D., et al.: Oseltamivir Resistance during Treatment of Influenza A (H5N1) Infection. N. Engl. J. Med. 353 (2005) 2667-72
27. Morris, G. M., et al.: Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. J. Computational Chemistry 19 (1998) 1639-1662
28. Jacq, N., et al.: Grid-enabled Virtual Screening against malaria, accepted for publication in Journal of Grid Computing, (2007).
29. Gropp, W., Lusk, E.: Dynamic process management in an MPI setting. Proceedings of the 7th IEEE Symposium on Parallel and Distributed Processing (1995)
30. Rarey, M., et al.: A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261: (1996) 470-489
31. Silva, A.M., et al.: Structure and inhibition of plasmepsin II, A haemoglobin degrading enzyme from Plasmodium falciparum. Proc. Natl. Acad. Sci. USA 93 (1996) 10034-10039
32. Lamb, M. L., Jorgensen, W. L.: Computational approaches to molecular recognition. Curr. Opin. Chem. Biol. 1 449 (1997)
33. Case, D.A., et al.: The Amber biomolecular simulation programs. J. Computat. Chem. 26 (2005) 1668-1688