WISDOM: A Grid-Enabled Drug Discovery Initiative Against Malaria

Vincent Breton, D. Kim, G. Rastelli

14

WISDOM: A Grid-Enabled Drug Discovery Initiative Against Malaria

Vincent Breton, Doman Kim, and Giulio Rastelli

CONTENTS

14.1 Introduction 354
14.2 Grid-Enabled Drug Discovery 354
14.2.1 *In Silico* Drug Discovery: Requirements and Grid Added Value 354
14.2.1.1 Requirements 356
14.2.1.2 Grid Added Value 357
14.2.2 Grid-Enabled Virtual Screening 359
14.2.2.1 The Virtual Screening Pipeline 359
14.3 Virtual Screening on the Grid: The WISDOM Initiative 361
14.3.1 Historical Perspective 361
14.3.2 First Data Challenge on Malaria 362
14.3.2.1 Preparation 362
14.3.2.2 Deployment 365
14.3.2.3 Results 366
14.3.3 Molecular Dynamics on the Grid 366
14.3.3.1 Introduction 366
14.3.3.2 Deployment 367
14.3.3.3 Molecular Dynamics Refinement and Rescoring Procedure 368
14.3.3.4 Results 369
14.3.4 *In Vitro* Tests 370

14.4 Second Data Challenge on Malaria 370
14.4.1 Introduction 370
14.4.2 Evolution of the Production Environment 371
14.4.3 Data Challenge Deployment 374
14.4.4 Postdocking Analysis and MD Refinement 376
14.5 Conclusion and Perspectives 377
Acknowledgments 379
References 379

14.1 Introduction

The goal of this chapter is to present the WISDOM initiative, which is one of the main accomplishments in the use of grids for biomedical sciences achieved on grid infrastructures in Europe. Researchers in life sciences are among the most active scientific communities on the EGEE infrastructure. As a consequence, the biomedical virtual organization stands fourth in terms of resources consumed in 2007, with an average of 7000 jobs submitted every day to the grid and more than 4 million hours of CPU consumed in the last 12 months. Only three experiments on the CERN Large Hadron Collider have used more resources. Compared to particle physics, the use of resources is much less centralized as about 40 different scientific applications are now currently deployed on EGEE. Each of them requires an amount of CPU which ranges from a few to a few hundred CPU years. Thanks to the 20,000 processors available to the users of the biomedical virtual organization, crunching factors in the hundreds are witnessed routinely. Such performances were already achieved on supercomputers but at the cost of reservation and long delays in the access to resources. On the contrary, grid infrastructures are constantly open to the user communities.

Such changes in the scale of the computing resources made continuously available to the researchers in biomedical sciences open opportunities for exploring new fields or changing the approach to existing challenges. In this chapter, we would like to show the potential impact of grids in the field of drug discovery through the example of the WISDOM initiative.

14.2 Grid-Enabled Drug Discovery

14.2.1 *In Silico* Drug Discovery: Requirements and Grid Added Value

The pharmaceutical R&D enterprise presents unique challenges for information technologists and computer scientists. The diversity and complexity of

the information required to arrive at well-founded decisions based on both scientific and business criteria is remarkable and well recognized in the industry.

Drug discovery is the process by which drugs are discovered and/or designed. Drug candidates are inputs to the drug development process. Recent progress in genomics, transcriptomics, proteomics, high throughput screening, combinatorial chemistry, molecular biology, and pharmacogenomics has radically changed the traditional physiology-based approach to drug discovery where the organism is seen as a black box. *In silico* drug discovery contributes to increasing biological system knowledge. The efficiency gains of such an integrated knowledge system could correspond to save 35% costs, or about US$300 million, and 15% time, or two years of development time per drug.

In silico drug discovery is one of the most promising strategies to speed-up the drug discovery process. It is important to know and control the *in silico* process, which is described below. Figure 14.1 shows the different phases of a drug discovery process with their approximate duration, their success rate and the corresponding *in silico* contributions.

A target is a cellular or genetic molecule which is believed to be associated with a desired change in the behavior of diseased cells and on which drugs usually act. The target identification and validation aims to isolate and select it. *In silico* drug discovery contributes to the target discovery by gene expression analysis, target function prediction and target three-dimensional (3D) structure prediction for postprocessing.

To identify a lead compound, a substance affecting the target selected in a drug-like way, two different *in silico* pipelines can be used which speed up the process and reduce costs avoiding useless *in vitro* tests: *de novo* design and virtual screening. *De novo* design builds iteratively a compound from the structure of a protein active site. Virtual screening selects *in silico* the best compound from a molecule database.

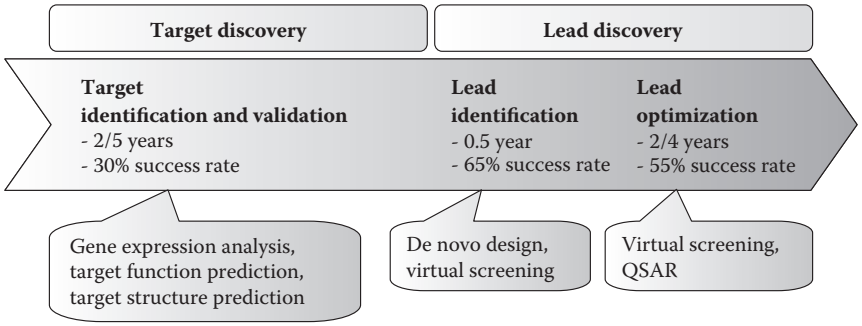

FIGURE 14.1 Representation of the different phases of the drug discovery process with their duration, their success rate and the corresponding *in silico* contributions.

Lead optimization addresses the development from the most promising lead compounds to a safe and effective drug. Instead of expensive and longer *in vitro* and *in vivo* tests, evaluation of basic chemical properties can be achieved by virtual screening and quantitative structure activity relationship (QSAR). QSAR is a quantitative correlation process of chemical structure with well-defined methods, such as optimization for pharmaceutical properties [absorption, distribution, metabolism, excretion and toxicity (ADMET)] or efficacy against the target organism.

In silico drug discovery contributes to increasing biological system knowledge, to managing data in a collaboration space, to speeding up analysis and consequently increasing the low success rate of the traditional "wet" approach. The efficiency gains of such an integrated knowledge system could correspond to save 35% costs, or about US$300 million, and 15% time, or two years of development time per drug.

Nevertheless, in spite of increasing levels of investment in *in silico* techniques, there is a steady decline in the number of new molecules that enter clinical development and reach the market. Many factors have changed over the past ten years, particularly the domination of the target-based drug discovery paradigm, favoring screening and rational drug discovery programs. A new approach aims to integrate rational drug discovery with a strong physiology and disease focus.

14.2.1.1 Requirements

Reducing the research time and cost in the discovery stage and enhancing information about the leads are key priorities for pharmaceutical companies worldwide [1]. To achieve this goal, *in silico* drug discovery must meet the following requirements: **Q1**

- The *in silico* drug discovery process includes the management of a large variety and quantity of scientific data; for example, images, sequences, models, and databases. Data integration is thus a challenge to increase knowledge discovery but also to ease the complex workflow. This implies data format standardization, dataflow definition in a distributed system, infrastructure and software providers for data storage, services for data and metadata registration, data manipulation, and database updates.
- The *in silico* drug discovery process also includes the management of a large variety and quantity of software. Software integration is another challenge to build efficient and complex workflows and to ease data management and data mining. Software can be provided in a distributed environment such as a Web server on the Internet. Different experts are absolutely necessary to maintain and update software and workflows to propose new methods or pipelines, to use remote services, exploit outputs, and finally to propose compounds for assay. A software workflow

will assist the scientist and the decision-maker in organizing their work in a flexible manner, and in delivering the information and knowledge to the organization.

- Deploying intensive computing is a challenge for *in silico* drug discovery. For instance, computing 1 million docking probabilities or modeling 1000 compounds on one target protein requires in the order of a few TFlops in one day. Very large computing resources are also needed to describe accurately protein structure models by computational methods based on all-atom physics-based force fields including implicit solution. Computing power is also required for bioinformatics resource centers where server access is saturated by the large number of short tasks requested by users.
- Joining new information technologies with life sciences to enable *in silico* drug discovery requires strong remote collaboration between different public and private experts when addressing neglected and emerging infectious diseases. It also involves strong sharing of resources: data and knowledge, software and workflow, and infrastructures such as computing, storage, and networks. The collaboration space needs experts to maintain the resources. Having tools and data accessible to everyone in collaboration requires intuitive interfaces that need to be maintained. These interfaces reduce the development time of new methods. They also help the integration of data and software from *in silico* drug discovery but also from experimental processes.
- Security is a key challenge for pharmaceutical industries but also for academic institutes in most cases. Effective protection of intellectual properties and sensitive information requires, for instance, authentication of users from different institutions, mechanisms for management of user accounts, and privileges and support for resource owners to implement and enforce access control policies.

In summary, the main requirements to develop *in silico* drug discovery are data and software integration, intensive computing deployment, remote collaboration and resources sharing, and, of course, security. Thus, there is a need for a powerful and secured environment sharing and integrating remote resources such as tools, data, computing, and storage.

14.2.1.2 Grid Added Value

The grid added value in the development of *in silico* drug discovery for neglected and emerging infectious diseases has multiple dimensions:

- Grids offer unprecedented opportunities for resource sharing and collaboration.

- Grids open exciting perspectives for handling information flows.
- Grids provide the resources to speed up the execution of time-consuming software.

Grids offer unprecedented opportunities for resource sharing and collaboration:

- The sharing of resources in a cross-organizational collaboration space between the pharmaceutical industry and academic research institutions, and between developed and least developed countries
- The creation of as a virtual laboratory for the different actors, increasing cooperation and communication between partners
- The mobilizing of resources routinely or in an emergency
- The sharing of diverse, complex, large, and distributed information for collaborative exploration and mutual benefit
- The use of new information technology such as large databases or time-consuming software
- The optimal exploitation of resources by taking advantage of spare computing cycles or by maximizing the use of high-performance computing platforms usage
- The reduction of hardware costs

Grids open also exciting perspectives for handling information flows:

- The deployment of services for healthcare and research centers in endemic regions
- The deployment of infrastructures to collect data and improve disease surveillance and monitoring
- The building of knowledge space with genomics and medical information (epidemiology, status of clinical tests, drug resistances, etc.)
- Access to relevant data, periodically updated databases and publications
- The federation of regional or international databases for disease study and monitoring of vector control, clinical trials, and drug delivery
- The provision of transparent and secure access to storage and the archiving of large amounts of data in an automated and self-organized fashion
- Connection, analysis and structuring of data and information in a transparent mode according to predefined rules (science- or business-process-based)

Finally, grids provide the resources to speed up the execution of time-consuming software:

- Access to large computing resources for *in silico* drug discovery, data analysis and mathematical modeling
- The application of high-performance computing to new areas
- The production of additional or more accurate analyses
- The facilitation of the exchange of tools and workflows between scientists
- The performance of computing intense tasks in a transparent way by means of an automated job submission and distribution facility
- Access to services and resources 24 hours a day
- The running of the same job on many platforms across different sites
- Access to computing resources by a single efficient path

Grids are unique tools for collecting and sharing information, networking experts, mobilizing resources routinely or in an emergency. A grid is thus an appropriate environment to develop *in silico* drug discovery.

14.2.2 Grid-Enabled Virtual Screening

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. Docking is only the first step of virtual screening since the docking output data has to be processed further [4].

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 modeling of 1000 compounds on one target protein. Access to very large computing resources is therefore needed for successful high throughput virtual screening [5].

14.2.2.1 The Virtual Screening Pipeline

Screening of chemical compounds against a target is an important step in the drug discovery process. Virtual screening is the process that screens the chemical compounds *in silico* against a target. The prerequisite to set up a virtual screening experiment is knowledge on the target, against which the screening has to be performed, and on the chemical compound libraries. Most of the information related to the targets is available in the literature, whether it is digital or paper-based.

Docking is the method of first choice for rapid *in silico* screening of large ligand databases for drug research, since it is based on a rational physical model. Basically, protein-compound docking is about computing 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; for example, the biological cycles of a given pathogen (parasite, virus, bacteria, etc.). The goal is to identify which molecules could dock on the protein 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 are available either open-source or licensed.

However, there is very often a compromise between speed and accuracy of results (in terms of the actual binding mode as well as the calculated affinity values) concerning the best scoring docking solutions. Docking methods usually generate a number of possible orientations of ligands in the binding site of the receptor, and the "correct" one (e.g. the orientation observed in the crystal structure) may not necessarily be ranked among the first docking solutions. This is due to deficiencies in orientation sampling and to the approximate nature of the scoring functions. Thus, it seems reasonable to subject screening results to postdocking refinement; for example, using molecular dynamics or similar methods that can describe biomolecular structure and energy in more details.

An example of workflow to deploy virtual screening that additionally includes post-docking refinement goes through the following steps:

Step 1 Selection of the target, the chemical compound database and the docking software.

Step 2 *Preparation*. If the selected target is a X-ray crystal structure with a bound ligand, then it requires preparing the binding site of the protein by taking 6-8 Angstroms from the cocrystallized ligand, taking care the significant amino acids for the activity are included in the binding site. Information on the significant amino acids can be obtained either from the literature or from the Brookhaven protein database. Regarding the preparation of the compounds to dock, open source database providing ready to dock molecules are made available on internet by the companies selling the compounds, but both target and compound have to be prepared according to the needs of the software. **Q2**

Step 3 *Docking*. Access to data analysis and visualization software is required at this point.

Step 4 *Postdocking analysis*. Results are analyzed based on the docking energy score and binding mode of the compound inside the binding site.

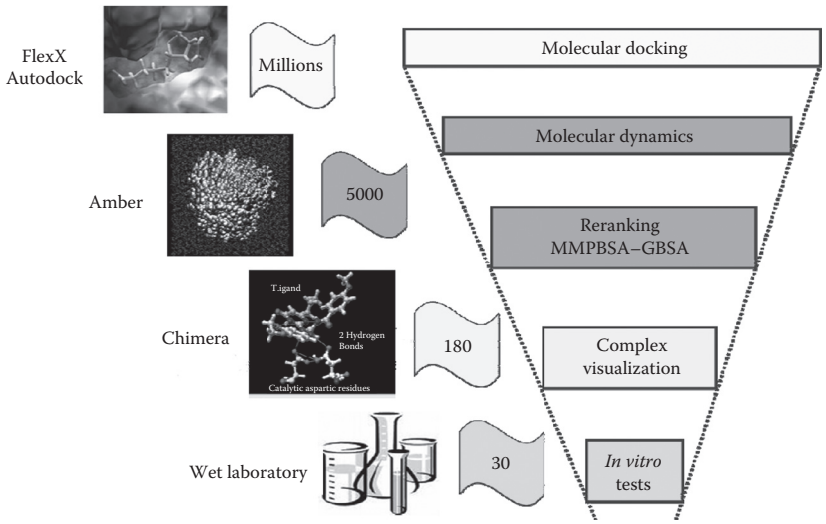

FIGURE 14.2 Example of a virtual screening workflow (credit: A. Da Costa).

Step 5 Refinement of docking orientations using molecular dynamics and reranking of the best compounds with more accurate scoring functions.

Step 6 *Postmolecular dynamics*. analysis. Access to data analysis and visualization software is required at this point.

Step 7 Selection of a few hundreds of compounds for *in vitro* testing.

Figure 14.2 illustrates this workflow with some popular software packages.

14.3 Virtual Screening on the Grid: The WISDOM Initiative

14.3.1 Historical Perspective

The WISDOM initiative was born from discussions that took place in July 2003 at the PharmaGrid conference in Welwyn, UK. Organized yearly since 2001 by the PRISM forum [6], the PharmaGrid cycle of conferences was aimed at steering exchanges between the leaders of the IT departments of the pharmaceutical laboratories and the developers of the grid technology. During the conference in Welwyn, the idea was discussed to identify a few topics where private and public partners could work together on the development and deployment of pilot projects on the grid.

One of the topics was *in silico* drug discovery for neglected diseases. Proposed by Martin Hofmann (SCAI, Fraunhofer Institute) and Vincent Breton, the idea raised interest from Manuel Peitsch who was at that time Global Head of Informatics and Knowledge Management at Novartis. After the conference, Novartis started a project focused on finding *in silico* new hits against dengue in partnership with University of Basel. As it was turning out difficult for other partners to contribute to this dengue project, the idea to launch a second project on an open-grid infrastructure emerged. In view of previous experience, we decided to start the project without involving a pharmaceutical laboratory.

Discussions started in the spring of 2004 with Nicolas Jacq, who was a PhD student at that time at LPC Clermont-Ferrand. The first decision was to choose a disease on which to focus our efforts. We did not have any *a priori* so we decided to ask one of our friends, the African pastor Joany Bazemo from Burkina-Faso, what was the worst plague in Africa. When asked, he did not hesitate in replying that malaria was the worst plague because it was killing children in hundreds of thousands. Once the disease was chosen, the next decision was to choose a biological target. This was entrusted to Vinod Kasam, a masters student at SCAI Fraunhofer, who studied the literature and identified the family of plasmepsins as a new potentially promising target. Plasmepsins are involved in the degradation of hemoglobin by the malaria vector, *Plasmodium falciparum*. Structures for plasmepsin II and IV were available in the Protein Data Base so all the ingredients were there for launching our first large virtual screening deployment on the grid. Large-scale deployments are called data challenges in the EGEE jargon.

14.3.2 First Data Challenge on Malaria

14.3.2.1 Preparation

Having identified the disease and the targets, there was still a lot of work to do in order to prepare the data challenge. On the biochemical side, three ingredients were required:

- 3D structures of targets
- A docking software
- A database of drug-like molecules to dock against the targets

We had already identified the 3D structures of interest in the PDB [7] database.

The docking software initially selected was Autodock [8], an open-source algorithm developed by the Scripps Research Institute. SCAI Fraunhofer is known worldwide for developing one of the best docking algorithms, FlexX [9] but FlexX could not be deployed on multiple grid sites because it

is licensed software. We had planned, however, that the best compounds selected with Autodock would be reprocessed on Fraunhofer cluster using the FlexX algorithm to compare Autodock and FlexX scores.

Libraries of compound 3D structures are made available open source by chemistry companies which can produce them. ZINC is one of these open source databases of compounds [10] which is constantly growing and was already providing the 3D structures of more than 3.4 million compounds in 2005.

Tests performed in December 2004 showed that docking the whole of ZINC against one target using Autodock represented about 35,000 jobs of approximately 20 hours, corresponding to about 80 years CPU. It soon appeared relevant from a biochemical perspective to perform docking not only on one but rather on three plasmepsin II structures (1lee, 1lf2, 1lf3) and one plasmepsin IV structure (1ls5) obtained from the Brookhaven Protein Data Base. In view of this increase of the number of targets, it was decided to focus on a subset of ZINC, the ChemBridge database collecting "only" 500,000 compounds. Finally, thanks to our collaborators from SCAI Fraunhofer, the BioSolveIT company distributing the FlexX software made graciously available up to 3000 free licenses of the software for deployment on the grid. This was extremely important as we had only a limited experience with Autodock.

Preparation of the data challenge involved on the biochemical side a number of steps in relation to the preparation of the targets and the validation of the docking procedure. On the grid side, this was the first time a biomedical application was going to require about one century of CPU time. At that time, only the LHC experiments had been deploying data challenges and the know-how for such large-scale deployments was shared by only a few experts. Preparation of the deployment included the development of an environment for job submission and output data collection. This environment had to be able to handle the submission of about 70,000 15-hour long jobs and the collection of the output data. A major issue was to handle job resubmission whenever a job failed for any reason, as the grid success rate was typically of the order of 80–85%. Large-scale tests were made on the French regional grid AuverGrid to validate the environment and to identify potential issues and bottlenecks. Other issues were raised by the data challenge, like the usage of licensed software on the grid or the need for a high throughput job submission scheme.

Based on the experience acquired during the testing phase on AuverGrid, the WISDOM production system (see Figure 14.3) [11] was developed in Perl, except the multithreaded job submission tool in Java. Two packages, wisdom_install and wisdom_test, were developed for installing the application components on the resources and for testing these components, together with the resources and grid services.

For the user submitting jobs, the entry point was a simple command line tool. Its users during the data challenge were members of the

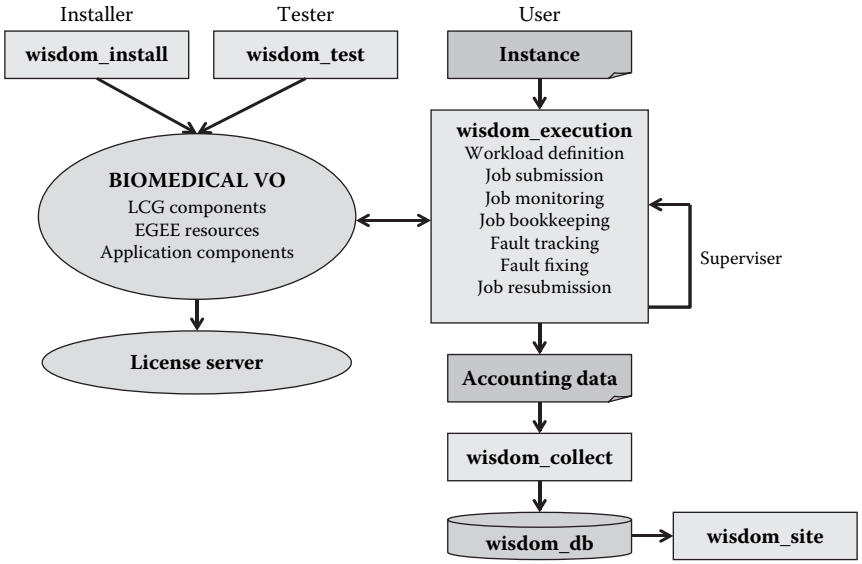

FIGURE 14.3 Design of the WISDOM production system.

Biomedical Task Force, which gathers a team of engineers with recognized expertise in application development and deployment. This software environment was developed to allow the submission and monitoring of job sets which were called *instances*. The different jobs of a given instance have the same target input and docking software. They only differ by the molecules of the compound library which are docked. Tasks needed to submit an instance were automatically executed by the WISDOM execution system. The user, authenticated by a proxy certificate, had to start his or her instance execution following a precise submission schedule to avoid too much competition between the computation participants, which would lead to a grid overload. Once the computation had started, the WISDOM environment took care of monitoring jobs and registering results. The user only had to check regularly if the process ran correctly, up to the end of all the jobs belonging to the instance. The overall process progression could be monitored through an output file for follow-up messages and an error file in case of any problems.

For each instance, a configuration file contained the instance information (software, target, database, parameter settings) and the grid parameters (number of jobs for the instance, resource brokers, computing elements, storage elements). A shell script and a job description language file were created for each job and used by the submission tool.

On the worker node, after the environment was configured, the shell script downloaded the database file from a storage element chosen by the

information system using the LCG API. Binaries were then called with the target and parameter settings transferred with the job. The compressed result was stored on a storage element and registered in the grid file catalog. A backup copy was also generated on another storage element. For the sake of simplification, the most relevant metadata relative to the output (software, parameter settings, compounds database, etc.) were stored in the name of the output itself. Output, errors and accounting messages were transferred on the user interface. Q3

Execution of FlexX jobs required handling floating licenses. Each job using FlexX software was contacting the Flexlm server at the beginning of the job and asked for a license, namely an ASCII file with specific keys generated for this server. Then the job was able to run without connection to the license server. Accessing floating licenses on the grid behind firewalls required known IP and the opening of two specific ports for institutes hosting worker nodes.

14.3.2.2 Deployment

The deployment took place in July and August 2005. During this period, ten users launched jobs from five user interfaces, monitoring the process with the help of the WISDOM environment and interacting with the user support of the EGEE project and the nodes administrators. For a total of 80 CPU years, 72,751 jobs were launched, producing 1 TB of data (500 GB, doubled for the back-up). Detailed analysis of the deployment can be found in [11] from which Table 14.1 is extracted.

TABLE 14.1

Main Metrics from WISDOM First Data Challenge

Metrics	Total	FlexX Phase	Autodock Phase
Cumulated number of docked compounds (in millions)	4127	3141	987
Effective duration	37 days	22 days	15 days
Number of docked compounds/hour	46475	59488	27417
Crunching factor	662	411	1031
Number of jobs submitted	72751	41520	31231
Number of grid computing elements used	58	56	57
Number of resource brokers used	12	12	11
Maximum number of jobs running in parallel on the grid	1643	1008	1643
Volume of output data	946 GB	506 GB	440 GB
Total CPU time	80 years	29.5 years	50.5 years
Effective CPU time used by successful jobs	67.2 years	24.8 years	42.4 years

The data challenge was very useful to identify the limitations and bottlenecks of the EGEE infrastructure. The WISDOM production system developed to submit the jobs on the grid accounted for a small fraction of the failures as well as the grid management system. On the other hand, the resource brokers were observed to significantly limit 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 distributed the jobs on the grid. As a consequence, 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.

14.3.2.3 Results

Postprocessing 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 1000 compounds based on scoring were selected thanks to postprocessing 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 postprocessing. 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.

Several scaffolds were identified in the 100 compounds selected for postprocessing. The urea, thiourea, and guanidino scaffolds were the most frequently observed in the top 1000 compounds. Some of the compounds identified were similar to already known plasmepsin inhibitors, like the urea analogs which were already established as micromolar inhibitors for plasmepsins (Walter Reed compounds) [12]. These results were already an indication that the overall approach was sensible and that large-scale docking on computational grids had the potential to identify new inhibitors, such as the guanidino analogs. To confirm these results, a new step had to be implemented: the refinement of the compound selection using molecular dynamics computations.

14.3.3 Molecular Dynamics on the Grid

14.3.3.1 Introduction

While docking methods have been significantly improved in the last few years by including more thorough compound orientation searches, additional energy contributions, and/or refined parameters in the force field, it is generally agreed that docking results need to be postprocessed with

more accurate modeling tools before biological tests are undertaken. Molecular dynamics (MD) [13] has great potential at this stage: first, 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; second, it partially solves conformation and orientation search deficiencies which might arise from docking; third, it allows the reranking of molecules based on more accurate scoring functions.

Just at the time when the results of the first WISDOM data challenge were being analyzed at SCAI Fraunhofer by Vinod Kasam, the BioinfoGRID project was launched in January 2006. LPC Clermont-Ferrand contribution to the project focused on the deployment of MD computations on the grid.

14.3.3.2 Deployment

Work started on the identification of the relevant molecular dynamics software and on the choice of the grid infrastructure on which to deploy the computations. Contacts were established with two groups at SCAI Fraunhofer and University of Modena which expressed interest to rerank the best hits coming out of the first WISDOM data challenge. Both groups were using the Amber software for molecular modeling which raised again the problem of deploying a licensed software on a grid. Contacts were established with the institution distributing Amber regarding the license policy on the grid.

The outcome of the negotiation was that we were allowed to deploy Amber on the grid under the following conditions:

- Each cluster deploying Amber had to have at least one license.
- Grid users allowed to use Amber had to come from one of the laboratories owning an Amber license.
- Grid users allowed to use Amber under the conditions described above could deploy their computations on all the grid clusters.

Regarding the choice of infrastructure, contacts were established with the EGEE biomedical virtual organization and several groups involved in DEISA. A clear preference for deployment on EGEE was expressed by the groups collaborating with us at University of Modena and SCAI Fraunhofer. Both had developed MD procedures which were well fitted for running on clusters. As a consequence, we focused our efforts on deploying these procedures on EGEE.

From September 2006, we started to investigate in details the deployment of the MD procedure designed by Giulio Rastelli (University of Modena and Reggio Emilia) on EGEE and based on the Amber software package. Amber [14] is a suite of different tools that carry out molecular dynamics simulations. The simulations in Amber can be divided into

three phases and different programs of the Amber distribution are responsible for performing these steps:

- Preparatory phase
- Simulatory phase
- Analysis phase

Encoding these steps in separate programs has some important advantages. First, it allows individual pieces to be upgraded or replaced with minimal impact on other parts of the program suite. Second, it allows different programs to be written with different coding practices: LEAP is written in C using X-window libraries, ptraj and antechamber are text-based C codes, mm-pbsa is implemented in Perl, and the main simulation programs are coded in Fortran 90. Third, this separation often eases porting to new computing platforms: only the principal simulation codes (*sander* and *pmemd*) need to be coded for parallel operations or need to know about optimized (perhaps vendor-supplied) libraries. The preparation and analysis programs are carried out on local machines on a user's desktop, whereas time-consuming simulation tasks are sent to a batch system on a remote machine; having stable and well-defined file formats for these interfaces facilitates this mode of operation.

14.3.3.3 Molecular Dynamics Refinement and Rescoring Procedure

An automated multistep procedure for the refinement and rescoring of docking screening results was designed and validated at University of Modena in Giulio Rastelli team [15]. The workflow, based on the Amber package, is able to automatically and efficiently refine docking poses, which sometimes may not be accurate, and rank the compounds based on more accurate scoring functions. The procedure requires as input a pdb file containing the structure of the protein and a mol2 file containing the coordinates of the docked ligands. The coordinates of the docked ligand and target structure are merged to create the complex.

The topology files are created using antechamber. The ligand atoms are described with GAFF (general Amber force-field) atom types and AM1-BCC charges. In order to avoid the time-consuming procedure of charge calculation, atomic charges of the ligand are read from the original mol2 file and not computed during the procedure; this choice adds the advantage that charge calculation of ligands can be performed only once, and obtained mol2 files used for many other target proteins. The ligand charges are calculated using antechamber by means of a separate script. Interestingly, the same set of AM1-BCC charges can also be exploited for automated ligand docking and MD. Missing gaff force field parameters for the ligand are automatically assigned by parmcheck. The

ligand, receptor, and complex topologies are written using leap utility (Amber 9). Minimization, MD, and final reminimization of the complexes are performed using *sander* with a distance-dependent dielectric constant $\varepsilon = 4r$. For each of these steps, the procedure enables the user to set ad hoc refinement options depending on the application. After refinement of the complex, a pdb file is generated as output and the final coordinates of the ligand, receptor, and complex are updated and used to compute the binding free energy evaluation using Amber MM-PBSA and MM-GBSA. The free energy results (ΔG_{MM}, ΔG_{solv}, and $\Delta G'_{bind}$) are written to an output file and compounds are ranked on the basis of their binding free energy [15].

This workflow was implemented in the EGEE grid using the WISDOM production environment designed for the large-scale docking experiments to deploy the MD procedure on the grid. The wide CPU availability of EGEE grid allowed the submission of a large subset of best docking complexes to the MD refinement and rescoring procedure despite its high computational demand.

The first successful deployment of the MD procedure was achieved in December 2006 on fake data. At preparation stage, all the required input files and Amber executables were stored on the storage element (SE) of the grid. At execution stage, jobs lasting approximately 20 hours were submitted corresponding to 50 compound subsets. Each subset with 50 compounds was submitted on one worker node along with the Amber executables, target structure, and the main script controlling the MD refinement and rescoring operations. After the jobs were finished the result files (structures and energy scores) were copied and stored on the storage element.

Following the successful test, the 5000 best compounds coming out of the first data challenge on plasmepsin II were refined and reranked. Computing resources from the EGEE Biomed virtual organization were used exclusively; 100 CPUs were used in parallel, all of them belonged to our local cluster at Clermont Ferrand, due to licensing issues. One single simulation was consuming ~20 CPU minutes on an Intel Xeon 3.05 machine. The estimated CPU time if the simulations were to be performed on one machine was therefore expected to be 124 days. By using EGEE infrastructure the simulation time was significantly brought down to 7 days.

14.3.3.4 Results

After rescoring the 5000 best docking results by Molecular Dynamics with Amber and MM-PBSA and MM-GBSA, the next step was to select the best compounds in the perspective of *in vitro* tests. The starting points were the two ranked lists of compounds, one according to MM-PBSA and the other according to MM-GBSA free energies. One hundred complexes of each list were analyzed manually. Each complex was visualized in 3D with UCSF Chimera software in order to determine the molecular interactions

between protein and ligand in the complex. The major criteria for selection were the ligand-making interactions to the two catalytic residues of plasmesin II, Asp 34 and Asp214. Second, the interaction with other key amino acids was checked: Gly36, Val78, Gly216, Ser79. The complexes with no interaction with at least one of the two amino acids of the catalytic dyad were rejected. The complexes which were kept had at least one main interaction to amino acids of catalytic dyad. In total, 30 out of 200 compounds were selected for *in vitro* tests.

14.3.4 *In Vitro* Tests

The 30 compounds selected for testing showed submicromolar or nanomolar IC_{50} values against recombinant *P. falciparum* plasmepsin II, using the inhibition assay based on FRET substrate degradation, which is well documented in the literature [16,17]. In the parasite, plasmepsin II is translated as an inactive zymogen containing a 124 amino acid-long N-terminal prosequence that has a membrane-spanning domain. Within the food vacuole the prosequence is removed by a calpain-like maturase, and active plasmepsin II is released [18]. Plasmepsin II was expressed well from a pET3d construct that contained Glu124 after initiator Met, and inclusion body of the protein was refolded and purified to near homogeneity as judged by SDS-PAGE. Based on molecular mass standards, recombinant plasmepsin II migrates at about 37 kDa.

Pepstatin A, a general inhibitor of aspartic proteases of microbial origin [19], was also reported to inhibited hemoglobin degradation by extracts of digestive vacuoles of *P. falciparum* [20]. In the *in vitro* inhibition test, the recombinant plasmepsin II activity without inhibitor was used as negative control and pepstatin A was used as a positive control. Pepstatin A showed inhibition (IC_{50} 80 nM) against the recombinant plasmepsin II activity. Among 30 compounds tested, six compounds had IC_{50} values below 80 nM. These results are extremely encouraging and suggest that the overall approach used to select the candidates is sensible for discovery of new plasmepsin inhibitors. The six compounds are currently being evaluated for their antimalarial activities *in vitro*.

14.4 Second Data Challenge on Malaria

14.4.1 Introduction

With the success achieved by the first data challenge 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. The WISDOM-II project dealt with several targets, which were both X-ray crystal structures and homology models. Targets from different classes of proteins were also being tested; reductases such as malarial dihydrofolate reductase (DHFR) and transferases such as GST as can be seen from Table 14.2. The same procedure described previously for the first data challenge was applied for target preparation. We had again the privilege to be able to use the FlexX software thanks to the generous support of BioSolveIT. Q3

DHFR and dihydropteroate synthase (DHPS) are two enzymes that belong to the folate biosynthetic pathway. The antifolates are the most exploited class of antimalarials, to which belong well-known molecules like pyrimethamine and cycloguanil. To date, the most widely used antifolate is a combination of pyrimethamine, a DHFR inhibitor, and sulfadoxin, a DHPS inhibitor. Nevertheless, their synergic action that results in enhanced activity is seriously compromised by drug resistance and hypersensibilization. For example, drug resistance is due to point mutations of various amino acids in the DHFR and DHPS (*P. falciparum* and *P. vivax*) active sites, and severely decreases drug efficacy. While most antimalarial research has been conducted on *P. falciparum* DHFR, there is growing interest on *P. vivax* DHFR, a less-studied target that is getting increasingly important because mixed falciparum and vivax infections are increasing, and parasites have developed resistance to both. Therefore, there remains a pressing need of new molecules apt to selectively bind these targets. To date, six different malarial DHFR crystal structures are available in the Protein Data Bank. These are the structures of wild-type *P. falciparum* DHFR (PDB code 1J3I), and of its C59R + S108N (1J3J) and N51I + C59R + S108N + I164L (1J3K) highly resistant mutants. In addition, four crystal structures of *P. vivax* DHFR are available, the wild-type structure (2BL9, 2BLB) and the S58R + S117N resistant mutant structures (2BLA, 2BLC). Hence, we thought it was a great opportunity to make use of these available resources to develop technology-based *in silico* screening on these targets. We chose *P. falciparum* DHFR (wt and the highly resistant quadruple mutant) and *P. vivax* DHFR (wt and the double mutant) structures (Table 14.2).

14.4.2 Evolution of the Production Environment

Following the experience acquired during the first data challenge on malaria and the first data challenge on avian flu [21], the WISDOM environment was 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

TABLE 14.2

Structural Features of all the Targets Used in the Second WISDOM Data Challenge

Target	Activity	Structure	PDB id	Resolution (Å)	Ligand	Cofactor
GST	Detoxification	Dimer	1Q4J	2.2	GTX	NO
P. falciparum DHFR (wild-type)	DNA synthesis	Bifunctional (with TS)	1J3I	2.33	WR99210	NADPH
P. falciparum DHFR (quadruple mutant)	DNA synthesis	Bifunctional (with TS)	1J3K	2.10	WR99210	NADPH
P. vivax DHFR (wild-type)	DNA synthesis	Bifunctional (with TS)	2BL9	1.90	Pyrimethamine	NADPH
P. falciparum DHFR (double mutant)	DNA synthesis	Bifunctional (with TS)	2BLC	2.25	Des-chloropyrimethamine	NADPH
Tubulin (*Plasmodium*)	Cell division	Monomer	Homology model	—	—	GTP

These two processes can be started and run simultaneously, the second one being fed from the information provided by the first one.

As seen during the deployment against avian flu [21], we removed the automatic resubmission of jobs in case of a job failure, because we thought that the overload generated by this automatic resubmission was a major cause for the poor reliability we observed during the first WISDOM deployment. Removing the automatic resubmission of jobs helped indeed a lot to improve overall reliability, but this induced a large amount of work for the users, who had to handle manual resubmissions of the instances after all the jobs had finished. The environment was further modified and enhanced to become more a "launch and forget" system, to carry all the tedious tasks and relieve the users whenever possible, such as automatic resubmission of jobs, automatic storage of job results in a relational database, automatic and real-time update of the experiment statistics and status, viewable through a Web site, and so on. As a consequence, the WISDOM environment was easily customized, tested and successfully used by the EELA [22] and EUChinaGRID [23] virtual organization.

The main objective was also 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 losing important information. This also proved to be also very useful as it enabled the whole maintenance of the scripts and code and improved the interactivity with the user, as the user could also manage jobs in fine detail; 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. While a 500 MB quota account was just enough to handle a single instance during the deployment on avian flu [21], it was now enough to handle no less than 20 instances simultaneously. Most of the job files were now generated dynamically which allowed also the user to modify on the fly the configuration of the resource brokers and the job requirements. This way, the user was sure that the next submissions would take these modifications into account. Figure 14.4 describes the overall architecture of the environment and the deployment process:

- The user is interacting with the system through the two main scripts (wisdom_submit and wisdom_status) deployed on the user interface. These scripts will take care automatically of job files generation, submission, status follow-up, and eventually resubmission.
- The jobs are submitted directly to the grid workload management system, and are executed on the grid computing elements. As soon as it is running, a job transfers all the files stored on the storage elements via the data management system of the grid, and the FlexX software, which asks a floating license for the FlexLm license server, can start to process the dockings.

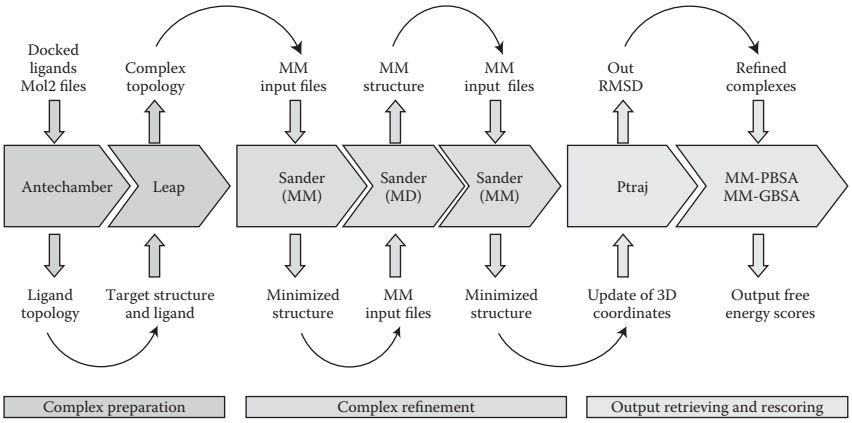

FIGURE 14.4 Schematic representation of the automated multistep refinement and rescoring procedure of the ligand-target complexes based on Amber. (From Luker, K.E. et al., *Molecular and Biochemical Parasitology*, 79, 1996, 71–78.)

- 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 via the data management system and the useful docking results are inserted directly from the grid to a relational database where they can later be more easily queried and analyzed.

14.4.3 Data Challenge Deployment

The deployment was performed on several grid infrastructures (Auvergrid [24], EELA, EGEE, EUChinaGrid) and involved at least one manager to oversee the process on each of them. The three groups of targets (GST, *P. vivax* and *P. falciparum* DHFR) were docked against the whole ZINC database (4.3 million ligands). The database was actually cut into 2422 chunks of 1800 ligands 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 40 seconds to 1 minute depending on the CPU power). The subsets had to be stored on the grid infrastructures. They were basically copied on a storage element and registered
Q4 on the grid file catalog (LFC) and were also replicated on several locations whenever possible to improve fault-tolerance. We defined a WISDOM instance as being a target structure docked against the whole ZINC database, with a given parameter set. Table 14.3 shows the instances deployed on the different infrastructures.

TABLE 14.3

Instances Deployed on the Different Infrastructures during the WISDOM-II Data Challenge

Target Structures	Number of Instances Deployed
GST (A chain)	4 on EGEE
GST (B chain)	4 on EGEE
2BL9 (*P. vivax* wild-type DHFR)	3 on EGEE, 1 on EELA
2BLC (*P. vivax* double mutant DHFR)	3 on EGEE, 1 on AuverGrid
Dm_vivax (*P. vivax* DHFR 2BLC minimized)	4 on EGEE
Wt_vivax (*P. vivax* DHFR 2BL9 minimized)	4 on EGEE
1J3K (*P. falciparum* quadruple mutant DHFR)	4 on EGEE
1J3I (*P. falciparum* wild-type DHFR)	3 on EGEE, 1 on EuChinaGRID

A total number of 32 instances were deployed, corresponding to an overall workload of 77,504 jobs, and up to 140 million docking operations. As shown in Figure 14.5, the environment included a FlexLm server that was providing the floating licenses for the FlexX commercial software. During the first WISDOM deployment in 2005, the license server was identified as a potential bottleneck and point of failure because we had just one server available at this time. For WISDOM-II, we started the deployment with only one server as well. Very soon, up to three servers were made available at the SCAI Fraunhofer institute (http://www.scai.fraunhofer.de), with 3000 licenses available on each server. The FlexX software binaries were

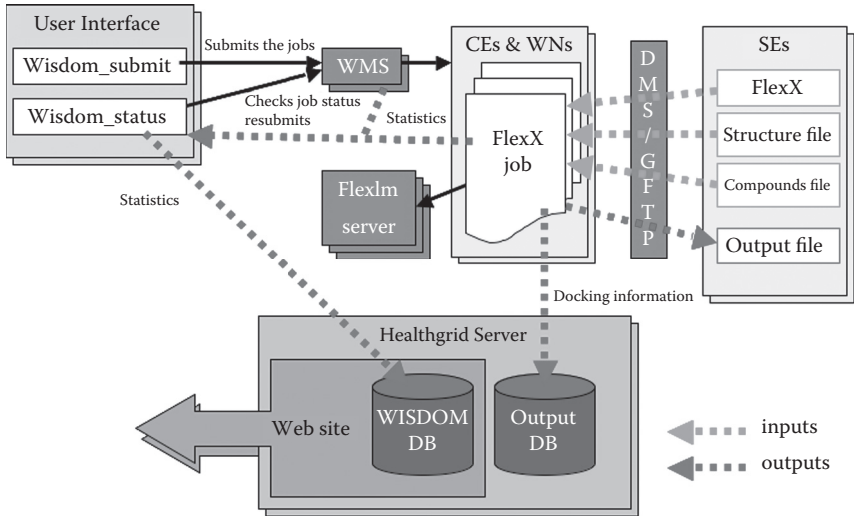

FIGURE 14.5 Evolution of the production environment.

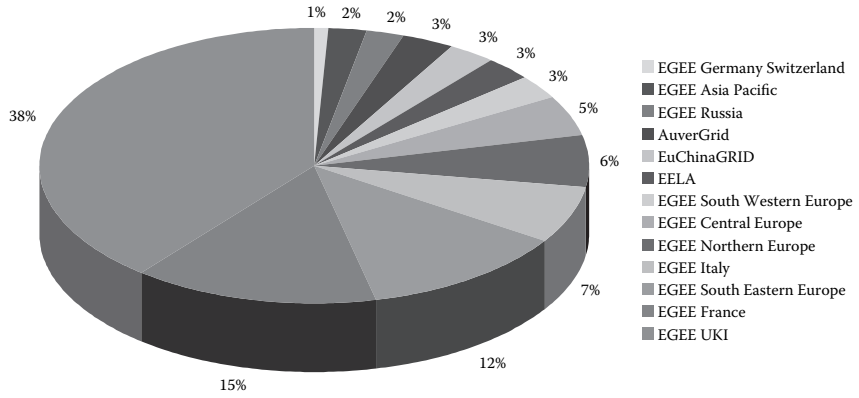

FIGURE 14.6 Distribution of jobs on the different grid federations.

stored like all the inputs on the grid storage elements and were installed on the fly on each worker node at the beginning of the job.

As the average duration of a job was around 30 hours, we submitted one instance per day, with a delay of 30 seconds between each submission. As one instance was submitted in about 20 hours, the submission process was quite continuous during the first month of deployment. The jobs were submitted to 15 resource brokers in a round-robin order. At the end of a job, the results were stored on the grid storage elements, and directly into a relational database.

Figure 14.6 shows the repartition of the jobs on the different grid federations. It also shows the AuverGrid, EuChinaGRID, and EELA infrastructures contribution. Each of these three infrastructures ran one single instance which corresponds to 3% of the total 32 instances.

14.4.4 Postdocking Analysis and MD Refinement

Analysis of the docking results was carried out on the four targets docked. The University of Modena took charge and analyzed results obtained on DHFR targets from *P. vivax* and *P. falciparum*. CSIR, in collaboration with LPC Clermont-Ferrand, analyzed results on GST. Two criteria were employed for result analysis: statistical analysis based on scoring values and analysis of compounds making interactions to key residues of the receptor. Based on these two criteria, the best compounds coming out of the docking step were selected and refined using the MD procedure described previously.

The results refined with MD at the University of Modena for wild-type *P. falciparum* DHFR showed that the identification of novel families of antimalarials whose structures are not related to classical antifolates is indeed possible. For wild-type *P. falciparum* DHFR, the best 15,000 compounds resulting from the docking screening (FlexX) have been refined with

molecular dynamics procedure, and the molecules have been rescored using MM-PBSA and MM-GBSA. Hydrogen bond analyses of the 15,000 ligands interacting at the active site revealed a significant enrichment of hydrogen bonds with DHFR key residues Asp54, Ile14, and Ile164 (as well as their combination) in the first ranking positions. This finding may validate the approach, because interaction with these residues is known to be required for biological activity.

In perspective, MD refinement will also be performed on the results obtained with the quadruple mutant DHFR structure and the results generated with *P. vivax* DHFRs. Once the analyses are complete, a final selection of potential inhibitors will be performed and the best candidates will be tested *in vitro* at Mahidol University (Thailand) in Worachart Sirawaraporn's laboratory. MD refinement of the results obtained on GST has been performed in collaboration with the University of Modena and the results of this rescoring step are presently under analysis.

14.5 Conclusion and Perspectives

The WISDOM initiative has demonstrated how the grid can significantly change the approach to screening which is a mandatory step on the road to drug discovery. In the last couple of years, a number of targets have been screened *in silico* using grid infrastructure resources. The molecules that have been selected through this process have shown a significant inhibition activity *in vitro* and some of them are under a patenting process. The grid added value has been clearly demonstrated by the remarkable increase of the virtual screening throughput up to 100,000 docked compounds per hour using 5000 computers on the EGEE grid.

Following this success, the WISDOM initiative, started from discussions between Martin Hofmann and Vincent Breton at the PharmaGrid conference in Welwin City in July 2003, has turned into a collaboration which has seven partners today:

- CNRS in Clermont-Ferrand (http://clrpcsv.in2p3.fr) with expertise in grid technology and Web services for life sciences
- SCAI Fraunhofer institute in Bonn, with expertise in bio- and chemoinformatics, in text-mining technology and author of the FlexX docking software
- University of Modena, with expertise in molecular dynamics for virtual screening and drug design on malaria
- CNR-ITB Institute for Biomedical Technologies in Milano, with expertise in high performance and GRID computing applications for bioinformatics in life science

- HealthGrid (http://healthgrid.org) with expertise in grid technology for life sciences and in infrastructure support
- Chonnam National University with expertise in biochemistry and *in vitro* validation of virtual screening
- KISTI in Korea with expertise in grid technology

Beside this core group, a number of biological laboratories have expressed interest to propose targets or to perform *in vitro* tests of the best molecules selected *in silico*:

- Mahidol University in Thailand
- University of Los Andes in Venezuela
- CSIR and University of Pretoria in South Africa

These very satisfactory results obtained both on the biochemical side and on the grid side of the project should not hide that very significant progress can be achieved on both sides.

The virtual screening pipeline described in Figure 14.2 can be refined at several levels.

- In terms of target preparation, it is well known that the PDB is full of bugs, especially for the structures which have been downloaded many years ago.
- In terms of ligand selection, our approach has been very naïve as we have been docking all the compounds available in ZINC and Chembridge. Software packages exist to perform a first selection of the ligands on physical and chemical criteria to reduce the number of dockings.

The availability of three-dimensional macromolecular coordinates is a prerequisite for many types of studies. PDB contains many anomalies ranging from proteins with small deviations from normal geometry, to structures that fit their submitted experimental data very poorly.

Led by CMBI, a massive rerefinement project was launched to obtain a better match between the experimental data and the atomic parameters (coordinates, B-factors) in the structure models while not compromising the geometric quality of the structure models [25]. The rerefinement of such a vast number of structure models requires enormous amounts of computer power. The rerefinement of 18,738 PDB files with 18,738 independent jobs requiring from 1 to 24 hours of CPU each is a very good example of a parallel project ideally suitable for deployment on a computing grid.

The rerefinements of the structure models were done on a hybrid computer environment consisting of the Biomed and EMBRACE [26] virtual organizations on EGEE grid infrastructure as well as several clusters of

EMBRACE partners in Europe. Each grid job consisted of 20 proteins that would run for 20 hours approximately on queues of 72 hours and were managed using the WISDOM production environment. More than 90% of the total of 17 years of CPU time was finished in two months. The 18,738 rerefinements were completed in four months. The rerefined structure models are available from http://www.cmbi.ru.nl/pdb_redo/.

Selection of ligands can significantly reduce the number of compounds docked. It can be based on toxicity or on the properties of the target binding site. We are presently investigating existing open-source software packages to filter the ligands before their docking on the grid.

Acknowledgments

The authors would like to acknowledge the contribution from numerous collaborators to the research activities described in this chapter: we would like to particularly thank Vincent Bloch, Ana Da Costa, Gianluca Degliesposti, Matteo Diarena, Géraldine Fettahi, Nicolas Jacq, Yannick Legrè, Simon Nowak, and Jean Salzemann.

The work described in this chapter was partly supported by grants from the European Commission (BioinfoGRID, EGEE, Embrace), the French Miinistry of Research (AGIR, GWENDIA) and the regional authorities (Conseil Régional d'Auvergne, Conseil Général du Puy-de-Dôme, Conseil Général de l'Allier).

The Enabling Grids for E-sciencE (EGEE) project is cofunded by the European Commission under contract INFSO-RI-031688. The BioinfoGRID project is cofunded by the European Commission under contract INFSO-RI-026808. The EMBRACE project is cofunded by the European Commission under the thematic area "Life sciences, genomics and biotechnology for health," contract number LHSG-CT-2004-512092. The SHARE project is cofunded by the European Commission under contract number FP6-2005-IST-027694.

AuverGrid is a project funded by the Conseil Regional d'Auvergne. The AGIR and GWENDIA projects are supported by the French Ministry of Research.

References

1. The Innovative Medicines Initiative (IMI) Strategic Research Agenda. Creating Biomedical R&D Leadership for Europe to Benefit Patients and Society, February 2008. Available at http://www.imi.europa.eu/docs/imi-gb-006v2-15022008-research-agenda_en.pdf.

2. Lyne, P.D. Structure-based virtual screening: An overview. *Drug Discovery Today* 7, 2002, 1047–1055.
3. Congreve, M. et al. Structural biology and drug discovery. *Drug Discovery Today* 10, 2005, 895–907. **Q5**
4. Ghosh, S. et al. Structure-based virtual screening of chemical libraries for drug discovery. *Current Opinion in Chemical Biology* 10, 2006, 194–202. **Q5**
5. Chien, A. et al. Grid technologies empowering drug discovery. *Drug Discovery Today* 7, 2002, 176–180. **Q5**
6. PRISM forum. Available at: http://www.prismforum.org.
7. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research* 28, 2000, 235–242.
8. 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. *Journal of Computational Chemistry* 19, 1998, 1639–1662.
9. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. Predicting receptor-ligand interactions by an incremental construction algorithm. *Journal of Molecular Biology* 261, 1996, 470–489.
10. Irwin, J.J. and Shoichet, B.K. ZINC—a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 45 (1), 2005, 177–182.
11. Jacq, N., Salzemann, J., Legré, Y., Reichstadt, M., Jacq, F., Medernach, E., Zimmermann, M., et al. Grid enabled virtual screening against malaria. *Journal of Grid Computing* (in press). **Q6**
12. Silva, A.M. et al. Structure and inhibition of plasmepsin II, A haemoglobin degrading enzyme from Plasmodium falciparum. *Proceedings of the National Academy of Sciences USA* 93, 1996, 10034–10039. **Q5**
13. Lamb, M.L. and Jorgensen, W.L. Computational approaches to molecular recognition. *Current Opinion in Chemical Biology* 1, 1997, 449.
14. Case, D.A. et al. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 26, 2005, 1668–1688.
15. Ferrari, A.M., Degiesposti, G., Sgobba, M., and Rastelli, G. Validation of an automated procedure for the prediction of relative free energies of binding on a set of aldose reductase inhibitors. *Bioorganic and Medicinal Chemistry* 15, 2007, 7865–7877.
16. Luker, K.E., Francis, S.E., Gluzman, I.Y., and Goldberg, D.E. Kinetic analysis of plasmepsin I and II aspartic protease of the *Plasmodium falciparum* digestive vacuole, *Molecular and Biochemical Parasitology* 79, 1996, 71–78.
17. Matayoshi, E.D., Wang, G.T., Krafft, G.A., and Erickson, J. Novel fluorogenic substrates for assaying retroviral proteases by resonance energy transfer. *Science* 247, 1990, 954–958.
18. Banerjee, R., Francis, S.E., and Goldberg, D.E. Food vacuole plasmepsins are processed at a conserved site by an acidic convertase activity in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* 129, 2003, 157–165.
19. Morishima, H., Takita, T., Aoyagi, T., Takeuchi, T., and Umezawa, H. The structure of pepstatin. *Journal of Antibiotics (Tokyo)* 23, 1970, 259–262.
20. Gluzman, I.Y., Francis, S.E., Oksman, A., Smith, C.E., Duffin, K.L., and Goldberg, D.E. Order and specificity of the *Plasmodium falciparum* hemoglobin degradation pathway. *Journal of Clinical Investigation* 93, 1994, 1602–1608.

21. Lee, H.-C., Salzemann, J., Jacq, N., Chen, H.-Y., Ho, L.-Y., Merelli, I., Milanesi, L., Breton, V., Lin, S.C. and Wu, Y-T. Grid enabled high throughput *in silico* screening against influenza A neuraminidase. *IEEE Transactions on Nanobioscience*, 5 (4), 2006, 288–295.
22. EELA. Available at: http://www.eu-eela.org.
23. EUChinaGRID. Available at: http://www.euchinagrid.org.
24. AuverGrid. Available at: http://www.auvergrid.org.
25. Joosten, R.P., Salzemann, J., Blanchet, C., Bloch, V., Da Costa, A.L., Diarena, M., Fabbretti, R., et al. Re-refinement of all X-ray structures in the PDB. *Proteins* (submitted). **Q6**
26. EMBRACE. Available at: http://www.embracegrid.info.
27. Kasam, V., Maaß, A., Schwichtenberg, H., Zimmermann, M., Wolf, A., Jacq, N., Breton, V., and Hofmann, M. Design of plasmepsine inhibitors: A virtual high throughput screening approach on the EGEE Grid. *Journal of Chemical Information Modeling* 47 (5), 2007, 1818–1828. **Q7**