

Proposal for a Data Challenge

FIRST DATA CHALLENGE ON *IN SILICO* DRUG DISCOVERY

Document identifier:	Proposal for a first data challenge on <i>in silico</i> drug discovery v1.4.doc
Date:	18/05/2005
Activity:	NA4
Authors:	N. Jacq, M. Reichstadt, V. Breton, M. Zimmermann
Document status:	Draft
Document link:	

Abstract: This document is a Proposal for a first data challenge on *in silico* drug discovery. The general aspects of the application are described, the requirements are defined and a planning is proposed. This data challenge is a scalability step through towards a full *in silico* drug discovery platform. The use case is the malaria, a neglected disease.

Document Log

Issue	Date	Comment	Author
1	04/05/2005	First version	M. Reichtadt, N. Jacq
1.1	06/05/2005	Modifications	V. Breton
1.2	11/05/2005	Comments	J. Montagnat
1.3	15/05/2005	Comments	M. Zimmermann
1.4	18/05/2005	Data challenge resources update	N. Jacq

Terminology

DD	Drug Discovery
DHF	Dengue Hemorrhagic Fever
EGEE	Enabling Grids for E-science in Europe
GGUS	Global grid user support
HTS	High-throughput Screening
HTD	High-throughput Docking
<i>In silico</i>	On computers
Lead	Best potential drug found by virtual screening process
PDB	Protein Data Base
SIMDAT	Data Grids for Process and Product Development using Numerical Simulation and Knowledge Discovery
VO	Virtual Organisation
VOMS	Virtual Organisation Membership Service

CONTENT

1. PROJECT PRESENTATION.....	4
1.1. ACTIVITY CONTEXT	4
1.2. DATA CHALLENGE OBJECTIVES	5
2. DATA CHALLENGE DESCRIPTION	7
2.1. SUMMARY	7
2.2. DATA CHALLENGE RESOURCES	7
2.3. DATA CHALLENGE WORKFLOW	8
2.4. DATA CHALLENGE OUTPUT	8
3. WORK PLAN.....	9
3.1. DATA CHALLENGE STRUCTURE	9
3.2. TEAMS INVOLVED IN THE PROJECT	9
3.3. DOCUMENTATION	10
4. IMPLEMENTATION PLANNING	11
5. FOLLOW-UP	12
5.1. BIOLOGY	12
5.2. BIOMEDICAL INFORMATICS	12
5.3. GRID DEPLOYMENT	12

1. Project Presentation

1.1. Activity context

In silico Drug Discovery in a grid environment

The challenges in Drug Discovery (DD) today are in understanding diseases, elucidating metabolism pathways and exploiting information at molecular and cellular level, which should result in reducing DD projects life cycle and increasing the probability of success after lead optimization.

Information Technology can transform the DD process through comprehensive, reliable data and information seamlessly integrated for easy navigation, the simulation of biomolecular processes, data mining and e-collaboration for dispersed teams. Grids are creating a fertile ground for the development of such new services. The grid technology provides the collaborative IT environment to enable a quicker drug development process from molecular biology research to goal-oriented field work. A pharmaceutical Grid should be a shared *in silico* resource to guarantee and preserve knowledge in the areas of discovery, development, manufacturing, marketing and sales of new drug therapies, and should cover the following aspects:

- a resource that provides CPU power to perform computational intensive tasks in a transparent way by means of an automated job submission and distribution facility
- a resource that provides transparent and secure access to the storage and archiving of large amounts of data in an automated and self-organized mode
- a resource that connects, analyses and structures data and information in a transparent mode according to pre-defined rules (science or business process based)

Drug discovery efforts are traditionally associated with pharmaceutical companies. The path from initial discovery of a biological target to the transition into development of a single lead compound into a drug takes up to 12 years in a highly sequential process. Pharmaceutical GRIDS hold the power to change the DD process in significant ways: They can speed up the DD process by enabling large-scale computational approaches previously considered infeasible, such as high-throughput docking (HTD). More importantly, due to their innate characteristics, public GRIDS foster collaboration between academic labs.

The motivating perspective is to enhance the ability of both pharmaceutical industry and academic research institutions to share diverse, complex and distributed information on a given disease for collaborative exploration and mutual benefit. The goal is to lower the barrier to such substantive interactions in order to produce cheaper drugs and insecticides to address diseases affecting third world development and to increase the return on investment for new drugs in the developed countries.

Wide *in silico* docking on malaria in a grid environment

We want to propose new inhibitors for the targets implicated by malaria by using a docking approach on the grid. This application will be jointly developed by the LPC Clermont-Ferrand and SCAI Fraunhofer, in collaboration with the SIMDAT and EGEE projects, the INSTRUIRE regional grid in Auvergne and the *CampusGRID* Bonn Aachen regional GRID. The challenge of WISDOM is to demonstrate the relevance and the impact of the grid approach to address Drug Discovery for neglected diseases.

It starts by addressing massive *in silico* docking which is a computational intensive task. Docking is the first step towards *in silico* drug design. Basically, 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 (the "target") 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 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 available open source by chemistry companies which can produce them.

With the help of the grid, large scale *in silico* experimentation is possible. Large resources are needed in order to test a family of targets, a large enough 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.

Dealing with malaria

A neglected disease has been selected as target to prove the feasibility of the concept: Malaria. The number of cases and deaths from malaria increases in many parts of the world. There are 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. For both vector control and chemotherapy, knowing the gene sequences of *Anopheles* and *Plasmodium* species should lead to discovery of targets against which new insecticides or anti-malarial drugs can be produced. However, such discoveries are likely to be patented and only developed at prices unaffordable to governments or villagers in tropical countries.

EGEE

The EGEE (Enabling Grid for E-science) project 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 project is also focussed on attracting a wide range of new users to the Grid.

The EGEE infrastructure has key features that make it particularly relevant for *in silico* drug discovery:

- A consistent, robust and secure Grid infrastructure.
- A continuously improving middleware delivering reliable services to users. This middleware is expected to provide more and more secure data management services.
- A willingness to attract new users from industry as well as science and to ensure they receive the highest standard of training and support.

These features answer the needs of our project of production service for a wide *in silico* drug discovery.

1.2. Data challenge objectives

In silico drug discovery is one of the most promising approaches to speed up and reduce the cost to develop new drugs. Many initiatives around the world (grid.org, World community grid, Decryphon) use pervasive grids for virtual screening and large scale genomics comparative analysis relevant to drug discovery. While pervasive grids only provide distributed computing, EGEE offers additional facilities in terms of data management and security. The main goal of this data challenge is to demonstrate to the research communities active in the area of drug discovery (molecular chemistry labs, pharmaceutical laboratories) the relevance of grid infrastructures in comparison to cluster computing and pervasive grids.

The objectives can be articulated as follows:

Biological goal

The malaria has been identified as the target disease to demonstrate the feasibility of the concept because it is a neglected disease killing millions worldwide. Very few drugs are now efficient and there is a major need for new drugs and vaccines. The targets for malaria are selected from the PDB. Structures are resolved by X-ray crystallography. One promising family of proteins studied exclusively in malaria is plasmepsin, an aspartic protease. Many parasites which cause serious human or livestock diseases produce aspartic proteinases which perform critical functions for parasite viability. This makes them excellent targets for the design of novel anti-parasitic drugs. The aspartic proteinases are particularly suitable for this since selective inhibition of several members of this class of enzyme has already been achieved, eg. HIV proteinase, *Candida* proteinase. Selectivity is clearly essential to drug development to allow potent inhibition of the target enzyme without effect on the human aspartic proteinases pepsin, gastricsin, renin, cathepsin D and cathepsin E or the recently discovered Napsins and BACE enzymes.

The malaria parasite *Plasmodium falciparum* may produce up to nine aspartic proteinases (the plasmepsins) and a closely related protein -HAP. Three of these enzymes (plasmepsins I, II and IV) are located in the parasite food vacuole and are involved in the digestion of host red cell haemoglobin by the parasite. HAP is also found in this organelle but its function is presently unknown. Inhibition of the plasmepsins can lead to parasite death.

Source :

- Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets, Graham H. Coombs, Daniel E. Goldberg, Michael Klemba, Colin Berry, John Kay and Jeremy C. Mottram, *TRENDS in Parasitology* Vol.17 No.11 November 2001
- <http://sites.huji.ac.il/malaria/maps/hemoglobinpolpath.html>

Biomedical informatics goal

The number of life science applications that have been successfully used on grids is very limited. Docking is – along with BLAST homology searches and some folding algorithms – one of the most prominent applications that have successfully been executed on grid testbeds. Moreover, docking is the only application for distributed computing that has prompted the uptake of grid technology in the pharmaceutical industry. Therefore, with proposing a large scale docking experiment here we aim at one of the most serious applications for the grid data challenge on the biomedical informatics side. Docking is the focus of this data challenge but it is only the first step to design a full drug discovery pipeline from target identification to lead optimization. Several data challenges are foreseen in the future when the next steps for *in silico* drug discovery (molecular dynamics, lead optimization, toxicity,...) are deployed on grid infrastructures.

Grid goal

The purpose of our Data Challenge is to deploy a CPU consuming application generating large data flows – millions of files summing up to a few TB – to test the grid infrastructure and services. It is also a mean to run a scientific challenge that could not be solved without the grid. Indeed, the grid added value lies not only in the computing resources made available, but also already in the permanent storage of the data. In a very close future, we expect improved data management middleware services to allow automatic update of compound database and the design of a grid knowledge space where biologists can analyze output data. During the data challenge, we expect that part of the infrastructure resources are dedicated to the application.

2. Data challenge description

2.1. Summary

The first data challenge aims at docking a subset of the ZINC database (3.3 Million compounds) against 5 different structures of plasmepsin, a protein target from the Malaria parasite Plasmodium. This molecule is potentially very interesting because it acts on human haemoglobin. Inhibiting its action would prevent the parasite from feeding itself with the human blood. Two docking algorithms will be used to find the best hits. Autodock (<http://w3.to/autodock>) is an open source algorithm developed by the Scripps Research Institute. SCAI Fraunhofer is known world-wide for developing one of the best docking algorithms, FlexX (<http://www.biosolveit.de/>). SCAI will give access to FlexX within the framework of this collaboration for a limited time. A server solution is studied to solve the license issue. ZINC is an open source databases of compounds (<http://zinc.docking.org>). The information on the targets can be found in the protein data bank (PDB, <http://www.rcsb.org/pdb>).

2.2. Data challenge resources

Tests performed in December 2004 and in April 2005 have shown that three scenarios are possible. They are described below depending on the data challenge duration and the available resources.

We propose to perform the scenario 1 during the summer 2005. This data challenge will be deployed on the biomedical VO.

Data challenge description	Scenario 1	Scenario 2	Scenario 3
Duration	3 weeks	3 weeks	1 month
CPU time	80 years CPU	160 years CPU	320 years CPU
Grid performance	70%	60%	50%
Number of CPU	2 000	4600	8 300
Number of grid jobs (20h)	30 000	60 000	140 000
Storage	6 TB	12 TB	24 TB
Docking workflow description			
Number of software / targets / compounds / parameter settings	2 / 5 / 5.10 ⁵ / 4	2 / 5 / 5.10 ⁵ / 8	2 / 5 / 10 ⁶ / 8
Objective	Selection of the best hits with short analysis	Selection of the best hits with real analysis	Real docking experimentation at a large scale

- The grid performance coefficient takes into account grid inefficiencies due to job submission failure, aborted jobs, and loss of resources due to concurrent jobs by other users, etc.
- About 1/3 of the jobs are using FlexX as docking algorithm and produce each one about 500 MB of compressed data. The other 2/3 of the jobs are using Autodock and produce each one about 50 MB of compressed data. These data will be further processed to select the most promising compounds to reduce job output by a factor 100 and select the 1% most promising hits.

2.3. Data challenge workflow

The aim is to launch an automatic, optimized and fault tolerant workflow using the grid resources and services. This is achieved through different steps:

Platform deployment

The docking platform is composed of the input files (the targets 3D structure), the database (the compounds 3D structure) and the software (a docking workflow with different parameters and scoring settings). First this platform must be deployed on the different grid elements. The targets structures and the compressed compounds database will be copied and registered on each storage element to be closed to the computing elements. The docking software will be stored on each computing element, and the tags will be published in the Information System.

Infrastructure testing

Each needed grid element (CE, SE, RB) is automatically tested before the data challenge to avoid aborted jobs and optimize the use of the infrastructure. Elements with errors will be deleted from the resources list for the data challenge.

Data challenge submission

The application development will allow executing only one command line with different parameters to launch the data challenge. Its deployment is achieved by successful steps: computation of the number of compounds for each job, JDL creation with requirements and ranking, job submission on RBs, status report, fault tolerance with new submission, output retrieval, output checking and statistics calculation. The job on a CE downloads the targets and the compressed database, extracts only the necessary compounds, processes the docking with published software and stores the registered results in a storage element. Necessary scratch space for the database of compounds on the WN (to avoid the simultaneous transfer of about 500 to 1200 small files) and the output files (before the compression) is about 5 GB.

2.4. Data challenge output

The result will be about 6 TB to 24 TB of compressed files depending of the scenario. A HPSS storage element will be used to permanently store the data challenge output files. We plan to make the results accessible to everyone. Different statistics on job execution time, success rate, transfer time, grid time as well as grid acceleration will be produced to check the achieved performance.

3. Work plan

3.1. Data challenge structure

The project is structured in 4 work packages : Molecular Modelling (MM), Cheminformatics (CI), Grid Resources Management (GRM) and Grid Application Deployment and Development (GADD). These 4 work packages work together on the following tasks:

- Target preparation - MM
- Compounds database preparation - CI
- Docking workflow preparation – MM, CI
- Platform deployment on the grid – GADD
 - Target deployment
 - Compounds database deployment
 - Docking workflow deployment
- Data challenge preparation
 - Grid resources integration – GRM
 - Grid workflow preparation – GADD
 - Docking workflow integration – GADD
- Data challenge management
 - Grid resources monitoring – GRM
 - Grid workflow monitoring – GADD

3.2. Teams involved in the project

This Data Challenge will be managed by SCAI Fraunhofer and LPC Clermont-Ferrand teams which will be in charge of the working groups described above. The people in charge of the different work packages are the following

- Work package manager
 - Molecular Modelling : Astrid Maass
 - Cheminformatics : Marc Zimmermann
 - Grid Resources Management : Yannick Legré
 - Grid Application Deployment and Development : Nicolas Jacq
- Developers
 - Molecular Modelling : Astrid Maass, Vinod Kumar Kasam, Nicolas Jacq
 - Cheminformatics : Marc Zimmermann, Vinod Kumar Kasam, Nicolas Jacq
 - Grid Resources Management : Yannick Legré, Nicolas Jacq, *Horst Schwichtenberg*
 - Grid Application Deployment and Development : Nicolas Jacq, Matthieu Reichstadt, Jiri Kraus, Mahendrakar Sridhar
- Users :

- Data Challenge : Nicolas Jacq
- Molecular Modelling : Astrid Maass
- Cheminformatics : Marc Zimmermann

3.3. Documentation

The documentation is useful for the collaborative work in a team and between partners, and for the maintenance and the perennality of the application. In particular, the documentation will considerably decrease the time for the preparation of a next data challenge. Our documentation will be composed of :

- The Proposal for a data challenge
- The Functional and technical specifications for :
 - Target preparation
 - Compounds database preparation
 - Docking workflow preparation
 - Platform deployment on the grid
 - Data challenge preparation
- The Tests report for each development
- The Developer manual for each development
- The User manual for each development

4. Implementation planning

MM : Molecular Modelling ; CI : ChemInformatics ; Grid Resources Management: GRM ; Grid Application Deployment and Development : GADD.

<i>Tasks</i>	<i>Planning (deadline)</i>	<i>MM manager</i>	<i>CI manager</i>	<i>GRM manager</i>	<i>GADD manager</i>	<i>MM Developers</i>	<i>CI Developers</i>	<i>GRM Developers</i>	<i>GADD Developers</i>	<i>MM, CI Users</i>	<i>GADD User</i>
Proposal for a data challenge	18/05/2005	X	X	X	X					X	X
Functional and technical specifications	27/05/2005	X	X	X	X	X	X	X	X		
Development & unit test – MM	17/06/2005					X					
Development & unit test – CI	17/06/2005						X				
Development – GRM	17/06/2005							X			
MM, CI users test	24/06/2005	X	X							X	
Development & unit test - GADD	01/07/2005								X		
GADD user test	13/07/2005			X	X						X
Data challenge	18/07/2005			X	X						X

5. Follow-up

5.1. Biology

The biological aim is to propose new inhibitors for a target implicated by the malaria, the plasmepsin. The target-compound complexes calculated during the data challenge will be published and freely available on the grid for the biological community interested by the malaria and more specifically by the plasmepsin target. The data challenge results will be stored on permanent grid storage and a grid portal will give an easy access to the data (targets, compounds and results). Information will be addressed to different actors of the research on malaria or to pharmaceutical industries like The Roll Back Malaria Partnership (<http://www.rbm.who.int/>), The National Center for Scientific Calculations of Venezuela (<http://www.cecalc.ula.ve>), Novartis (<http://www.novartis.com/>) or GlaxoSmithKline (<http://www.gsk.com>).

5.2. Biomedical informatics

Scoring function of the docking software will rank the 3,3 million results. But this list will not be sufficient to sort the target-compound complex. Post-treatment will be processed on the results.

- The 1% most promising hits will be reprocessed on Fraunhofer cluster using the FlexX algorithm to compare Autodock and FlexX scores.
- Post-filtering by FlexPharm
- Clustering of similar conformations
- Checking pharmacophoric points of each conformation
- Doing statistics on the score distribution - picking high ranking, low ranking and seeds
- Re-ranking for interesting compounds
- Sorting and assembly of data by vHTS Explorer

5.3. Grid deployment

After this data challenge, three actions are expected to enrich the *in silico* drug discovery process:

- The compounds database ZINC is regularly updated on a web site because the number of compounds is increasing each month. That is why the grid-enabled Drug Discovery platform could benefit of an automatic update and replication service on the grid. This service is under development within the Rugby bioinformatics grid project (<http://rugbi.in2p3.fr>) and could be interfaced with LCG or gLite API.
- Another challenge will be the management of the huge docking output on the grid. We want to progressively build a knowledge space around these results to extract and process the most interesting information and enrich the data with the results found later by other *in silico* drug discovery processes.
- Our data challenge is a scalability step towards a full *in silico* Drug Discovery platform. Once the gLite data management facilities will be available as well as workflow managers like Taverna (<http://taverna.sourceforge.net>), we intend to further extend the drug discovery workflow to include more precise molecular dynamics computations using quantum chemistry software like NAMD (<http://www.ks.uiuc.edu/Research/namd/>) for the hits selected at the docking step. NAMD is a parallelized and free for academic software that

selects more specifically the drug candidates. This software is even more CPU consuming than docking software, so it has to be executed only on the best hits.