

---

# Second Data Challenge on *In Silico* Drug Discovery

## *IN SILICO* DOCKING ON GRID INFRASTRUCTURES TO ACCELERATE STRUCTURE-BASED DESIGN AGAINST INFLUENZA A NEURAMINIDASES

---

Document identifier:	<b>Data Challenge on Drug Discovery against H5N1 v6.0.doc</b>
Date:	<b>30/03/2006</b>
Activity:	<b>NA4</b>
Authors:	<b>N. Jacq, J. Salzemann, V. Breton, Y.T. Wu, H.C. Lee, Y.C. Chen, L. Milanesi</b>
Document status:	<b>Draft</b>
Document link:	<a href="https://edms.cern.ch/document/711157/4">https://edms.cern.ch/document/711157/4</a>

---

**Abstract:** This document is a proposal for a second data challenge on *in silico* drug discovery. After the first biomedical data challenge on drug discovery for WISDOM application during the summer 2005, this application will focus on Influenza H5N1, the avian flu. The use case aims to deploy an *in silico* docking at a large scale on grid infrastructure to accelerate structure-based design against Influenza A neuraminidases.

The general aspects of the application are described, the requirements are defined, a detailed planning of work is proposed. This data challenge is an improvement step through towards a virtual screening service at a large scale. Timing is very important to avoid a clash with the LHC SC4 so it will use the shortest available technical path to achieve production deployment. The data challenge will take place in April 2006.



**PUBLIC**

### Document Log

Issue	Date	Comment	Author
1	10/03/2006	First version	N. Jacq, J. Salzemann, V. Breton
2	14/03/2006	Adding the plan of using DIANE in the data challenge	H. C. Lee
4	17/03/06	Improved description of biological goals	Y.T. Wu
4.2	17/03/06	Contribution of BIOINFOGRID	L. Milanesi
4.3	22/03/06	Update of submission plan of DIANE framework – Resources contribution of TWGrid	H.Y. Chen, H.C. Lee
5	23/03/06	Version 5 of the draft	N. Jacq
6	30/03/06	Version 6 of the draft	V. Breton

### Terminology

<b>DD</b>	Drug Discovery
<b>EGEE</b>	Enabling Grids for E-science in Europe
<b>GGUS</b>	Global grid user support
<b>HTS</b>	High-throughput Screening
<b>HTD</b>	High-throughput Docking
<b><i>In silico</i></b>	On computers
<b>Lead</b>	Best potential drug found by virtual screening process
<b>PDB</b>	Protein Data Base
<b>VO</b>	Virtual Organisation
<b>WISDOM</b>	Wide In Silico Docking On Malaria
<b>DIANE</b>	Distributed Analysis Environment

## **CONTENT**

<b>1. PROJECT PRESENTATION.....</b>	<b>4</b>
1.1. ACTIVITY CONTEXT .....	4
1.2. DATA CHALLENGE OBJECTIVES .....	4
1.3. LIST OF PARTICIPANTS .....	4
1.4. PLANNING .....	6
<b>2. DATA CHALLENGE DESCRIPTION .....</b>	<b>7</b>
2.1 SUMMARY .....	7
2.2 DATA CHALLENGE RESOURCES .....	7
2.3 DATA CHALLENGE PREPARATION .....	8
2.4 DATA CHALLENGE WORKFLOW .....	10
2.5 DATA CHALLENGE OUTPUT .....	12
<b>3. WORK PLAN.....</b>	<b>14</b>
3.1 DATA CHALLENGE DEPLOYMENT .....	14
3.2 DATA CHALLENGE STRUCTURE .....	14
3.3 DATA CHALLENGE PARTICIPANTS .....	14
3.4 DOCUMENTATION .....	15
<b>4. IMPLEMENTATION PLANNING .....</b>	<b>16</b>
<b>5. FOLLOW-UP.....</b>	<b>17</b>
5.1 BIOLOGY .....	17
5.2 BIOMEDICAL INFORMATICS .....	17
5.3 GRID DEPLOYMENT .....	17

## 1. Project Presentation

### 1.1. Activity context

The H5N1 virus transmission to human has been observed since 1997, but there has been experience of the subtype N1 at least since 1918. However, scientists showed that the N1 and N2 subtypes could evolve into variants under drug stress. Therefore, the data challenge is going to study the impact of point mutation on drug resistance. The goal is to screen a large set of compounds against the same target, the NA, with various structures predicted from homology methods.

This application will be jointly deployed by the Corpuscular Physics Laboratory of Clermont-Ferrand, CNRS/IN2P3, France; Genomics Research Center, Academia Sinica, Taiwan; Grid Computing Centre, Academia Sinica, Taiwan; Institute for Biomedical Technologies, CNR, Italy, in collaboration with the EGEE project, the AuverGrid regional grid in Auvergne, and the TWGrid. This work will take place within collaboration between these three projects, the EMBRACE network of excellence and the BioInfoGrid project. The challenge of the wide *in silico* docking on avian flu in a grid environment is to accelerate structure-based design against Influenza A Neuraminidases.

### 1.2. Data challenge objectives

The objective is to deploy a first data challenge on avian flu, a *in silico* docking at a large scale on grid infrastructure to accelerate structure-based design against Influenza A Neuraminidases.

The biological goal is to find potential compounds that can inhibit the activities of Influenza A neuraminidase N1 subtype variants. Viral neuraminidase is an enzyme that helps release of new virions by cleaving human host receptors, which action is essential for virus proliferation and infectivity. The development of drug resistance variants is one of the potential concerns of influenza pandemics. Therefore, the idea is to compile the results from *in silico* screening to know the kinds of compounds and chemical groups (fragments) to be equipped for blocking the active neuraminidases if mutations are to occur at some specific sites.

The biomedical goal is to accelerate the discovery of novel potent inhibitors thru minimizing non-productive trial-and-error approaches and improving the efficiency of high throughput screening thanks to the grid infrastructure and the submission environment developed for the WISDOM experience. Grid performance metrics will be compared between the first and the second data challenge.

The grid goal is to reproduce a grid-enabled *in silico* process with a shorter time of preparation. The quality of usage and quality of process needs to be improved. In addition to re-exercising the same framework used for the first data challenge, a light-weight framework adopted by ASGC for building high-throughput docking service, DIANE, will also take part in the second data challenge. The purpose is to investigate how far the DIANE's special scheduling and failure recovery mechanisms could achieve for the concerned qualities under the pressure of handling jobs in large scale. Another objective is to deploy the application on 3 different grid infrastructures with the same grid technology: AuverGrid, TWGrid, EGEE. In the same time, it is the occasion to test new submission strategy and new developments in the perspective of the second data challenge against neglected diseases in preparation for fall 2006.

### 1.3. List of participants

Partners involved in the project are:

The Corpuscular Physics Laboratory of Clermont-Ferrand, CNRS/IN2P3, France

The Genomics Research Center, Academia Sinica, Taiwan

The Grid Computing Centre, Academia Sinica, Taiwan (ASGC)

**SECOND DATA CHALLENGE ON *IN SILICO*  
DRUG DISCOVERY**

Doc. Identifier:  
**Data Challenge on Drug  
Discovery against H5N1  
v6.0.doc**

Date: 16/03/2006

The Institute for Biomedical Technologies, CNR, Italy  
The Biomedical Task Force, EGEE  
EMBRACE WP3  
BioinfoGrid

Organisation	Actors	Role	Contact email
LPC	Jean Salzemann Nicolas Jacq Nathanaél Verhaegue Matthieu Reichstadt Emmanuel Medernach Arnaud Fessy Yannick Legré Vincent Breton	Improving the grid workflow environment Supervising the data challenge deployment on the EGEE and Auvergrid infrastructures	<a href="mailto:breton@clermont.in2p3.fr">breton@clermont.in2p3.fr</a> <a href="mailto:jacq@clermont.in2p3.fr">jacq@clermont.in2p3.fr</a> <a href="mailto:salzemann@clermont.in2p3.fr">salzemann@clermont.in2p3.fr</a>
GRC	Ying-Ta Wu	Preparing the targets and the docking software Participating to the data challenge deployment on the EGEE and TWGrid infrastructures Analysing docking outputs	<a href="mailto:ywu@gate.sinica.edu.tw">ywu@gate.sinica.edu.tw</a>
ASGC	Hung-Chun Lee Li-Yung Ho Hsin-Yen Chen Simon C. Lin Eric Yen	Participating to the data challenge deployment on the EGEE infrastructure and TWGrid Analysing docking outputs Preparing the DIANE's runtime environment on the UI of TWGrid	<a href="mailto:Hung-Chun.Lee@cern.ch">Hung-Chun.Lee@cern.ch</a> <a href="mailto:liyungho@gate.sinica.edu.tw">liyungho@gate.sinica.edu.tw</a> <a href="mailto:hychen@mail.twgrid.org">hychen@mail.twgrid.org</a> <a href="mailto:Simon.Lin@twgrid.org">Simon.Lin@twgrid.org</a> <a href="mailto:Eric.Yen@twgrid.org">Eric.Yen@twgrid.org</a>
ITB	Luciano Milanesi Ermanna Rovida Pasqualina D'Ursi Ivan Merelli	Participating to the preparation of the targets and the docking software Participating to the data challenge deployment on the EGEE and TWGrid infrastructures	<a href="mailto:luciano.milanesi@itb.cnr.it">luciano.milanesi@itb.cnr.it</a> <a href="mailto:ermanna.rovida@itb.cnr.it">ermanna.rovida@itb.cnr.it</a> <a href="mailto:parqualina.dursi@itb.cnr.it">parqualina.dursi@itb.cnr.it</a> <a href="mailto:ivan.merelli@itb.cnr.it">ivan.merelli@itb.cnr.it</a>
BTF	Christophe Blanchet Johan Montagnat	the EGEE biomed task force will help with the deployment of the data	<a href="mailto:c.blanchet@ibcp.fr">c.blanchet@ibcp.fr</a> <a href="mailto:johan.montagnat@unice.fr">johan.montagnat@unice.fr</a>

**PUBLIC**

**SECOND DATA CHALLENGE ON *IN SILICO*  
DRUG DISCOVERY**

*Doc. Identifier:*  
**Data Challenge on Drug  
Discovery against H5N1  
v6.0.doc**

*Date:* 16/03/2006

---

	challenge.	
--	------------	--

**1.4. Planning**

The data challenge will take place in April 2006 in order to avoid a clash with LHC service challenge 4 before migration of the infrastructure to gLite3.0.

## **2. Data challenge description**

### **2.1 Summary**

The second biomedical data challenge aims at docking

- 300,000 compounds (200,000 compounds from the ZINC database and 100,000 compounds from a chemical combinatorial library). Filters to select subsets are the presence of a benzyl, six or five member-ring as the compound core (scaffold), the presence of at least one acid group, and drug-like consideration.
- against 16 different target structures of Influenza A neurominidases. 5 structures from the same target protein (from PDB) are prepared by considering each with one amino acid mutated only. 4 possible points are expected from literatures and the analysis within the range of binding pocket. These preparations are compared with crystal structures of other subtypes, except N1, available in PDB.
- with Autodock, an open source algorithm developed by the Scripps Research Institute. Autodock carries out quick conformation search of 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. Using AutoDock to evaluate one compound structure for 10 poses within the target enzyme would take 150 Kilobyte storage and 15 minutes on an average PC.

The jobs need to be done under controlled conditions, with a small test set of target and compounds. Timing is very important to avoid a clash with the LHC Service Challenge 4 so it will use the shortest available technical path to achieve production deployment.

### **2.2 Data challenge resources**

The table presents the resources needed for the data challenge scenario. They need to be modified in function of the number of parameter settings and the number of compounds in the test set. The CPU needed is 137 years CPU. The grid performance for CPU time observed during the first data challenge is 80%. The number of CPU needed during 4 weeks is so increased up to 2150. The grid performance for the number of jobs was lower during the first data challenge: 63%. So the total number of jobs will be approximatively 82 200. The storage space needed is 1,4 TB (with 1 back-up).

<b>Data challenge description</b>	<b>Scenario</b>
Duration	4 weeks
CPU time	137 years CPU
Grid performance for CPU	80%
Number of CPU	2 150
Grid performance for the number of jobs	63%
Number of grid jobs (20h)	82 200
Storage (1 back-up)	1,4 TB
<b>Docking workflow description</b>	
Number of software / targets / compounds / parameter settings	1 / 16 / 3.10 <sup>5</sup> / 1

## SECOND DATA CHALLENGE ON *IN SILICO* DRUG DISCOVERY

Doc. Identifier:  
Data Challenge on Drug  
Discovery against H5N1  
v6.0.doc

Date: 16/03/2006

---

Number of compounds by job	80
----------------------------	----

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.

The next table presents the available resources for the data challenge.

Resources	Available CPU	Dedicated CPU	Available storage
AuverGrid	500-600		700 GB
TWGrid	120	100	1 TB
Biomed resources in EGEE	6000		

### 2.3 Data challenge preparation

Before starting the data challenge, several preparation steps are necessary:

#### - Preparing docking material

- Target structures in pdbqs format
  - 16 structures
  - The name can be T[YY]IAN
- Compound database in pdbq format
  - The database will be divided in subsets of 80 compounds to build job of 20h. We will obtain approximatively 3750 subsets.
  - Name of subsets has this form: D[XXXX]IAN
  - Subsets can be compressed in tgz
- Software binaries are not the usual Autodock binaries
- Parameter settings: macro files. Only 1 parameter settings
- There are no intermediary file
- Needs to keep STDOUT/STDERR of the process
- Testing docking workflow in a local machine by LPC and obtaining awaiting outputs
  - Name of compressed outputs directory has this form: T[YY]D[XXXX]IAN.tgz
  - Evaluate process time and output size to define DC workload (number and size of the instance)
  - Evaluate max scratch space necessary in the WN

#### - Integrating grid resources

- Contact ROC managers and grid resources to inform, ask dedicated resources, particularly in AuverGrid and TWGrid, and ask migration planning on gLite (normally from 1st may)
  - Prepare an email with a clear message about issue of this DC to motivate computing centers.

- Determining available resources in the 3 grids : number, stability and number of resources dedicated for biomed VO and this data challenge (RBs, CEs, SEs)
- Contact User Support, EIS team and JRA2 team for information
- Checking RLS/RMC availability. If it is not available, we will just use GridFTP transfers without registration.

- **Improving WISDOM grid workflow**

- Change organization of the environment
  - adapt environment to the docking material: repositories, user instruction
  - Environment can be articulated hierarchically. For instance:
    - Instance: T[YY]IAN
    - Sequence: resubmission number of an instance in case of failures
    - Workload: T[YY]D[XXXX]IAN
    - JobID: Unique by workload and sequence.
- Change workload creation
  - usage of subsets instead of all database
  - change variables in the grid parameters file
- change submission strategy
  - slow but secure submission of jobs : each 2 minutes on the selected RBs with a round-robin, with Atlas ranking, fuzzyrank attribute and limited resources requirements.
- change jobs monitoring
  - no automatic resubmission
  - robust save of workload and jobID
- change output management
  - storage only on the grid of docking output and grid output
  - what presentation of the docking output for GRC ? 1 unique file by target (43 GB) ?
  - grid output contains all relevant information about job success and job statistics
  - no output sandbox to retrieve
  - Control of the grid output
    - Ok => statistics to be saved
    - KO => preparation for resubmission of failed jobs
    - Case where lcg-utils for data transfer failed but gridFTP succeeded
- Monitoring tool
  - study GridIce or keep WISDOM environment
- Statistics tool
  - study GridIce or keep WISDOM database
  - web site migration

- **Preparing DIANE runtime environment**
  - o Install DIANE on a separate UI of TWGrid
  - o Synchronize DIANE's workflow of AutoDock with the data challenge workflow, especially the policy of job taking and result archiving, job execution statistics
  
- **Installing on the grid the docking material**
  - o Install software in software repository of each CE
  - o Install each compressed subset of the compound database in 3 SEs
  - o The target structures and parameter settings are sent with each job (provided by the user as a docking service)
  - o Docking and grid outputs repositories if gridFTP transfer is used must be prepared
  
- **Testing grid job with docking material on each node by WISDOM and DIANE environment**
  - o Test of docking job in each CE
  - o Test of the RBs, CEs, SEs, grid services in the same time
  - o Clean eventual outputs
  
- **Integrating data challenge users**
  - o AuverGrid, TWGrid, BIOINFOGRID, Biomed Task Force users
  - o Each user must have a deputy to replace him in case of problem
  - o Determining User Interfaces (LPC, Taiwan, BIOINFOGRID, Lyon etc.)
  - o Defining instances submission plan by user depending its availability
  
- **Installing environment for grid workflow on all UIs**
  - o Ensure to have several GB of storage space in all User Interface
  - o Instruction for user
  - o Testing the environment before the real data challenge with a large number of jobs

## **2.4 Data challenge workflow**

The aim is to launch an automatic workflow using the grid resources and services. Quality of the process (production of output and grid success rate) and quality of usage (simplified process for users). The application development will allow executing only one command line with different parameters to launch the data challenge.

### **Instance submission plan**

This plan is still a hypothesis. It depends of the number of available resources, RBs, User Interfaces.

We can deploy 16 instances of 3 750 jobs. Number of User Interface to submit jobs can be 5: 1 or 2 in LPC, 1 in Academia Sinica, 1 in BIOINFOGRID, 1 or 2 by Biomedical Task Force (Lyon, Nice, Valencia...). So there will be only 3 instances by UI without the resubmission instances (that will be submitted manually from the same UI or by another UI). Duration of the instance submission phase depends

of the number of RBs. For 10 RBs, the phase will be 12,5 hours. Checking jobs can be done in parallel during submission by status command line, lcg-infosites command line, or via GridIce or Imperial College interface.

### **DIANE submission plan**

To maintain a sustained docking throughput provided by the DIANE framework, an instance of DIANE master will be run as long as possible on the User Interface; while a certain number of DIANE workers (100-150) are submitted separately from the same UI to the Grid. The number of submitted worker is equivalent to the number of Grid jobs. Instead of processing a constant number of docking tasks in single Grid job, the number of docking tasks a single DIANE worker will take is dependent on the overall performance given by the machine that the worker runs on. Concerning that the master process might be terminated due to unexpected reason, the DIANE master should keep track of the finished dockings during the runtime so that new master instance only needs to schedule the unfinished dockings. In the beginning of the data challenge, only one DIANE master process will be run on the UI. A reasonable number of concurrent DIANE master processes will be given to ramp up the data challenge throughput.

### **Its deployment is achieved by successful steps:**

- job script creation with docking requirements
- JDL creation with requirements and ranking
- job submission on RBs

NB: The workload creation will be done separately before each submission. The grid parameter file will be able to be modified during the process in function of the grid status.

- Robust jobID save and process status save (in case of UI failure)
- status report
- job report checking, save workload for resubmission if failure (no done (success))
- grid output checking (after file download from the grid), save workload for resubmission if failure
- failures identification during process, even if job succeeded (transfer for instance)
- Statistics calculation done with log files of jobs, time of each job, failures nature. Information stored is the same than the WISDOM data challenge. A success rate can be done by Instance, Sequence and workload
- Checking individual docking result in each job by employing DIANE's active feedback feature. The failed dockings are collected for further analysis and error tracking. A successful docking is recognized by matching the string: "autodock3: Successful Completion" inside the output dlg.

NB: Failures need to be classified by category (WMS, DMS, site, docking and user)

### **The grid job on a Worker Node is achieved by successful steps:**

- download the targets and the compressed subset database
- decompress subset
- run the docking with published software
- Checking the docking output
- Store the registered output in a storage element (dlg, STDOUT, STDERR, intermediary files ? in a compressed repository)

- checking the storage, delete stored output if failure
- store the grid output in a storage element with process information
- checking the storage, do again if necessary

NB: CPU time, data transferred size, data transferred time, data transferred failures will be saved in the grid output

**The data challenge user will be responsible of:**

- maintenance of his User Interface, save the data of the User Interface if possible
- submission of the instance at the right time

NB: It is very IMPORTANT to respect the planning of the submissions. All delay MUST be reported to the supervisor.

NB: It is very important to check regularly the beginning of the instance during 1 hour to prevent the RBs failures ; and the beginning of the status report to be informed by the grid of the grid failures

- follow-up of its instance by WISDOM environment (or another environment)
- inform regularly the supervisor about the status of the instance (ok, ko)
- send him failure information to solve the problem as soon as possible
- Change the grid parameter file during the submission process in case of failure

**The data challenge supervisor will be responsible of:**

- same tasks than a user
- is the contact point of each actor (also for user support and site administrator)
- centralise all information during the data challenge

## **2.5 Data challenge output**

The result will be about 1,4 TB (with a back-up copy), stored on 2 disk SE located in TWGrid and AuverGrid. A third copy could be done on a HPSS storage element to permanently store the data challenge output files (CCIN2P3).

There will be 3750 output files of 12 MB in dlg format by instance, so 60 000 output files. GRC of Academia Sinica wants to apply a script to compute the enrichment factor of compound library from the docking outputs, to list the interacting residues, to re-evaluate those docked poses of selected topmost compounds.

The mean to analyse the results are still under discussion

- Script applied by docking run ?
- Script applied for each instance (16 runs on a output file of 43 GB)
- Building a relational database to store few parameters at the aim to classify outputs ?

A relational database based on the information extracted with the script and the location of the output (in what file it is stored) could be build. BIOINFOGRID can participate in the database building.

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. Grid process information will be stored in a Mysql database, and published on a web-site (in a similar way than for WISDOM data challenge).

**SECOND DATA CHALLENGE ON *IN SILICO*  
DRUG DISCOVERY**

*Doc. Identifier:*  
**Data Challenge on Drug  
Discovery against H5N1  
v6.0.doc**

*Date:* 16/03/2006

---

### **3. Work plan**

#### **3.1 Data challenge deployment**

This data challenge will be deployed on 3 grids infrastructures using LCG-2 middleware: AuverGrid, TWGrid and the biomedical VO resources. The AuverGrid users, the TWGrid users, the BIOINFOGRID users and the Biomedical Task Force will support the jobs submission, the jobs follow-up and the data collect during the data challenge. To avoid a clash with the LHC Service Challenge 4, the deployment will take place in April 2006.

#### **3.2 Data challenge structure**

The project is structured in 3 work packages: Docking Preparation (DP), Grid Resources Management (GRM) and Grid Application Deployment and Development (GADD). The 3 work packages work together on the following tasks:

- Docking workflow preparation – DP
  - Target structures preparation
  - Compounds database preparation
  - Parameter settings preparation
  - Preparation of the docking output analysis
- Platform deployment on the grid – GADD
  - Compounds database deployment
  - Docking software 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

#### **3.3 Data challenge participants**

- DP
  - Contact point: Y.T. Wu
  - E. Rovida
  - P. D'Ursi
  - N. Jacq
- GRM
  - Contact point: J. Salzemann
  - TWGrid : H.C. Lee
  - AuverGrid : E. Medernach
  - EGEE : Y. Legré

## SECOND DATA CHALLENGE ON *IN SILICO* DRUG DISCOVERY

*Doc. Identifier:*  
**Data Challenge on Drug  
Discovery against H5N1  
v6.0.doc**

*Date:* 16/03/2006

---

- GADD
  - Contact point: H.C. Lee, J. Salzemann
  - M. Reichstadt
  - N. Jacq
  
- Users (and deputy)
  - J. Salzemann (N. Jacq)
  - M. Reichstadt (A. Fessy or N. Verhaegue)
  - Li-Yung Ho (Hurng-Chun Lee)
  - I. Merelli, C. Arlandini (L. Milanesi)
  - BTF user (BTF deputy)

### 3.4 Documentation

Our documentation will be composed of :

- The Proposal for the data challenge
- The user manual for the data challenge deployment by the grid application

**SECOND DATA CHALLENGE ON *IN SILICO*  
DRUG DISCOVERY**

*Doc. Identifier:*  
**Data Challenge on Drug  
Discovery against H5N1  
v2.doc**

*Date:* 16/03/2006

---

**4. Implementation planning**

DP: Docking Preparation; Grid Resources Management: GRM; Grid Application Deployment and Development: GADD.

<i>Tasks</i>	<i>Planning (deadline)</i>	<i>DP</i>	<i>GRM</i>	<i>GADD</i>
Proposal for the DC	13/03/2006	X	X	X
Docking material Grid app. development	24/03/2006	X		
Grid app. test Infrastructures available	03/04/2006		X	X
End of the DC	30/04/2006			X

The detailed planning is available in the excel file joined to the proposal.

## **5. Follow-up**

### **5.1 Biology**

The potential compounds are identified and selected by ranking binding free energies of resulted docked models obtained from this data challenge. 1000 of topmost ranked (lowest binding free energy) docked complexes (compound against target) will be refined with interaction potential and re-ranked. At least 50 compounds will be assayed experimentally at identified laboratories. Y.T. Wu (GRC, Academia Sinica) can participate in one of these laboratories for the assay.

### **5.2 Biomedical informatics**

One possibility is a service website to be included in pandemics watch.

### **5.3 Grid deployment**

After this data challenge, four actions are expected:

- Next data challenge against several targets of neglected diseases in fall 2006
- Docking service at a large scale on grid infrastructure using Taverna tool.
- Molecular Dynamics simulations deployment on grid environment
- BioinfoGRID applications on grid infrastructure in fall 2006, 2007