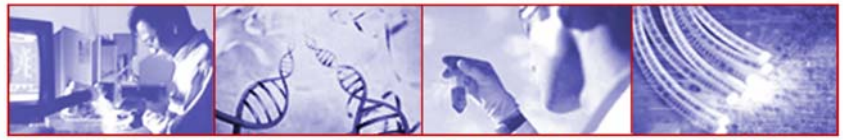
euchinagrid



Interconnection & Interoperability of Grids between Europe & China

Porting Biological Applications in Grid: An Experience within the EUChinaGrid Framework

G. La Rocca[(1)], *G. Minervini*[(2)], *P.L. Luisi*[(2)] *and F. Polticelli*[(2)]

[(1)] *INFN Catania, Italy*

[(2)] *Dept. of Biology, Univ. Roma Tre, Italy*

ISGC, 28.3.2007

FP6-2004-Infrastructures-6-SSA-026634

Information Society and Media

Outline

- The EUChinaGrid Project
  - Overview
  - Biological applications
    - Protein folding
    - “never born proteins”
- The software and its porting in Grid
  - Method
  - Input generation
  - “ab initio” prediction of protein structure
  - Integration in the GENIUS Grid portal

The EUChinaGRID Project (http://www.euchinagrid.org/)

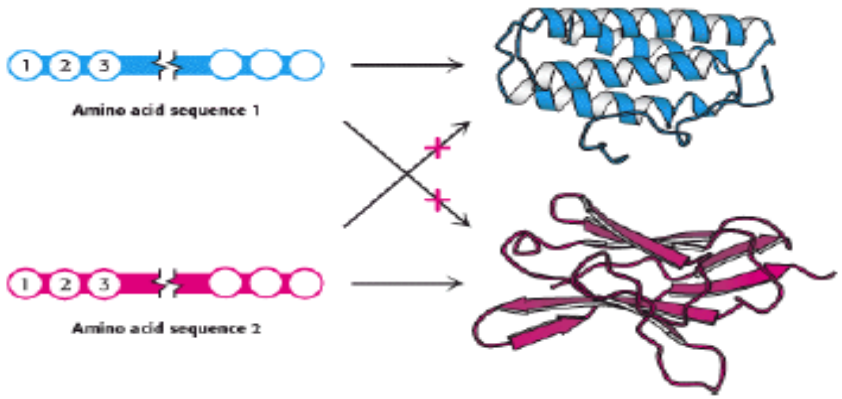
- ***Overview***
  - EUChinaGRID project is intended to provide specific support actions to foster the integration and interoperability of the Grid infrastructures in Europe (EGEE) and China (CNGrid).
  - The project promotes the migration of new applications on the Grid infrastructures by training new user communities and supporting the adoption of grid tools for scientific applications.

- ***WP4 - Applications***
  - The Workpackage is intended to validate the Intercontinental Infrastructure using scientific applications and make easier the porting of new applications relevant for scientific and industrial collaboration between Europe and China.
  - The activities within the WP4 are divided in three application fields:
    - ***A4.1:*** EGEE Applications (CMS and Atlas)
    - ***A4.2:*** Astroparticle Physics applications (the ARGO experiment)
    - ***A4.3:*** Biological applications

Infrastructures: CNGRID & EGEE

The Biological Applications

- **The protein folding "problem" and the structural genomics challenge**
  - The combination of the 20 natural amino acids in a specific sequence dictates the three-dimensional structure of the protein.
  - Protein function is linked to the specific three-dimensional arrangement of amino acids functional groups.
  - With the advancement of molecular biology techniques a huge amount of information on *protein sequences* has been made available but less information is available on structure and function of these proteins.
  - The "ab initio" prediction of protein structure is a key instrument to better understand the protein folding principles and successfully exploit the information provided by the "genomic revolution".



The protein sequences space

- The number of natural proteins, though apparently huge, represents just a tiny fraction of the theoretically possible protein sequences.
  - With 20 different co-monomers, a protein chain of just 60 amino acids can theoretically exist in 20^{60} chemically and structurally unique combinations.
- Estimates of the number of proteins present in nature vary from a minimum of 10^9 to a maximum of 10^{13}, thus the ratio between the number of existing proteins and those theoretically possible is very small.
  - A particularly suggestive example is that this ratio correspond to that between the volume of the hydrogen atom and that of the entire universe.

The "Never Born Proteins"

- **Rationale**
  - There exist a huge number of protein sequences that have never been exploited by biological systems, in other words enormous number of "never born proteins" (NBP).
  - The NBP pose a series of interesting questions for the biology and basic science in general:
    - Which are the criteria with which the existing proteins have been selected?
    - Natural proteins have peculiar properties in terms for example of thermal stability, solubility in water or amino acid composition?
    - Or else they represent just a subset of the possible protein sequences generated only by the contemporary action of contingency and physico-chemical forces?

The approach

- The problem is tackled by a "high throughput" approach made feasible by the use of the GRID infrastructure.
- A library of 10^7-10^9 random amino acid sequences of fixed length is generated (n=70).
- "ab initio" protein structure prediction software is used.
- Analysis of the structural characteristics of the resulting proteins in terms of:
  - Frequency of compact folds and characteristics of the corresponding amino acid sequences
  - Occurrence of novel yet unknown folds
  - Hydrophobicity/Hydrophilicity characteristics
  - Presence of putative catalytic sites
  - Experimental validation on "interesting" cases

Rosetta

- The Rosetta *ab initio* module (*developed by David Baker – University of Washington*) is a software application which allows the prediction of the three-dimensional structure of an amino acid sequences starting from a secondary structure of the sequence itself and a set of fragments extracted from the *Protein Data Bank (PDB).*

- The Protein Data Bank (http://www.wwpdb.org/) is a repository of proteins and nucleic acids that can be accessed for free by biologists and biochemists from around the world.

Rosetta: Method details

- **Module I - Input generation**
  - The query sequence is divided in fragments of 3 and 9 amino acids
  - The software extracts from the data base of protein structures the distribution of three-dimensional structures adopted by these fragments based on their specific sequence
  - For each query sequence is derived a fragments data base which contains all the possible local structures adopted by each fragment of the entire sequence.

- **Module II - *Ab initio* protein structure prediction**
  - The sets of fragments are assembled in a high number of different combinations by a Monte Carlo procedure.
  - The resulting structures are subjected to a energy minimization procedure using a semi-empirical force field.
  - The principal non-local interactions considered are hydrophobic interactions, electrostatic interactions, main chain hydrogen bonds and excluded volume.
  - The compatible structures both with local biases and non-local interactions are ranked according to their total energy resulting from the minimization procedure.

Rosetta: Module I

- The procedure for input generation is rather complex but computationally inexpensive (10 min of CPU time on a Pentium IV 3,2 GHz).
- Due to the many dependencies of module I (*Blast* and *psipred*), the input generation is carried out locally with a script that automatizes the procedure for a large dataset of sequences.
- Approximately 500 input datasets are currently being generated daily.

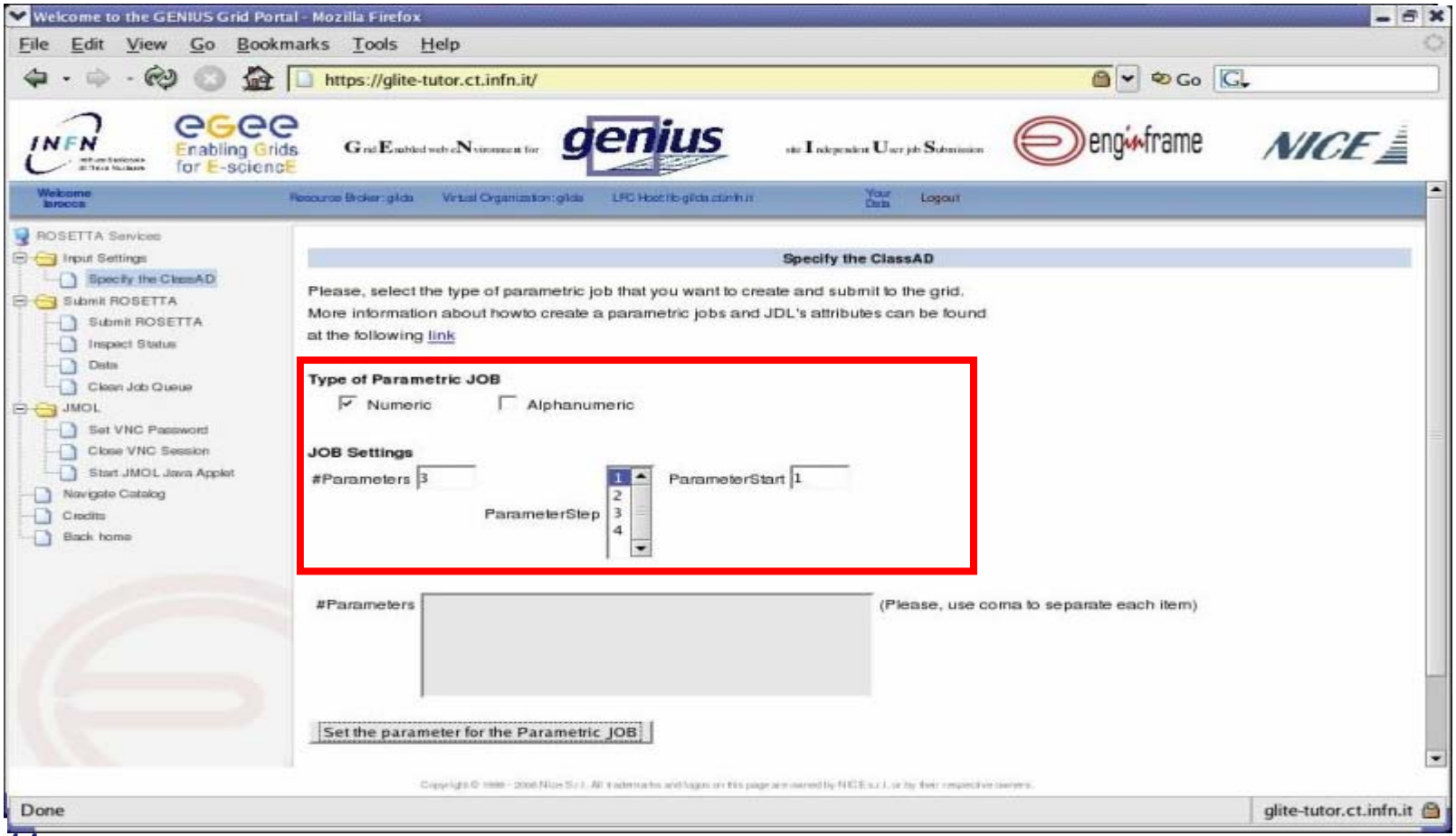
Rosetta: Module II

- Input
  - fragment files generated by module 1
  - secondary structure prediction using psipred
- In output the user obtains a number of structural models of the query sequence ranked by total energy
- A single run with just the lowest energy structure as output takes approx. 10-40 min of CPU time depending on the degree of refinement of the structure
- The Module II has been implemented in GRID through the use of the GENIUS Grid Portal (https://glite-tutor.ct.infn.it)
  - From this portal, exploiting the last feature of the gLite middleware, (www.glite.web.cern.ch/glite) it's possible submitting parametric jobs and run, in one shot, a large number of jobs (structure predictions).

The home – https://glite-tutor.ct.infn.it

Create the dynamic ClassAD /1

- After MyProxy initialization the user connects to the GENIUS portal to set up the parametric JDL, specifying the number of runs (equivalent to the number of amino acid sequences to be simulated) to be carried out.



Create the dynamic ClassAD /2

- Step 2. The user specifies the working directory and the name of the shell script.

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

ROSETTA Services
- Input Settings
  - Specify the ClassAD
- Submit ROSETTA
  - Submit ROSETTA
  - Inspect Status
  - Data
  - Clean Job Queue
- JMOL
  - Set VNC Password
  - Close VNC Session
  - Start JMOL Java Applet
- Navigate Catalog
- Credits
- Back home

JDL Attributes

With the next #services user can specify the attributes to customize his parametric job.
Please, use the **_PARAM_** item each time you want to indicate a parametric attribute.

Working Directory /home/larocca/rosetta-dir

Executable /home/larocca/ROSETTA-PARAM/rosetta.s Select...

Argument 2ptl_PARAM_tar.gz " _PARAM_ Argument Browse...

Executable, Standar Output and Standard Error

Enable Standard files

StdOutput file std_PARAM_out

StdError file std_PARAM_err

Create JDL Attributes for the Parametric JOB (1/4)

Done glite-tutor.ct.infn.it

Create the dynamic ClassAD /3

- Step 3. Input files (fragment libraries) are loaded as a single .tar.gz folder per amino acid sequence.

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

ROSETTA Services
- Input Settings
  - Specify the ClassAD
- Submit ROSETTA
  - Submit ROSETTA
  - Inspect Status
  - Data
  - Clean Job Queue
- JMOL
  - Set VNC Password
  - Close VNC Session
  - Start JMOL Java Applet
- Navigate Catalog
- Credits
- Back home

JDL Attributes

With the next 4 services user can specify the attributes to customize his parametric job.
Please, use the **_PARAM_** item each time you want to indicate a parametric attribute.

InputSandbox

☑ InputSandbox

Input File [ ] Browse... Not yet supported

Input File
/home/larocca/ROSETTA-PARAM/2ptl1.tar.gz
/home/larocca/ROSETTA-PARAM/2ptl2.tar.gz
Select... Clean

N.B.: Remember that the files to upload MUST contain the following items:

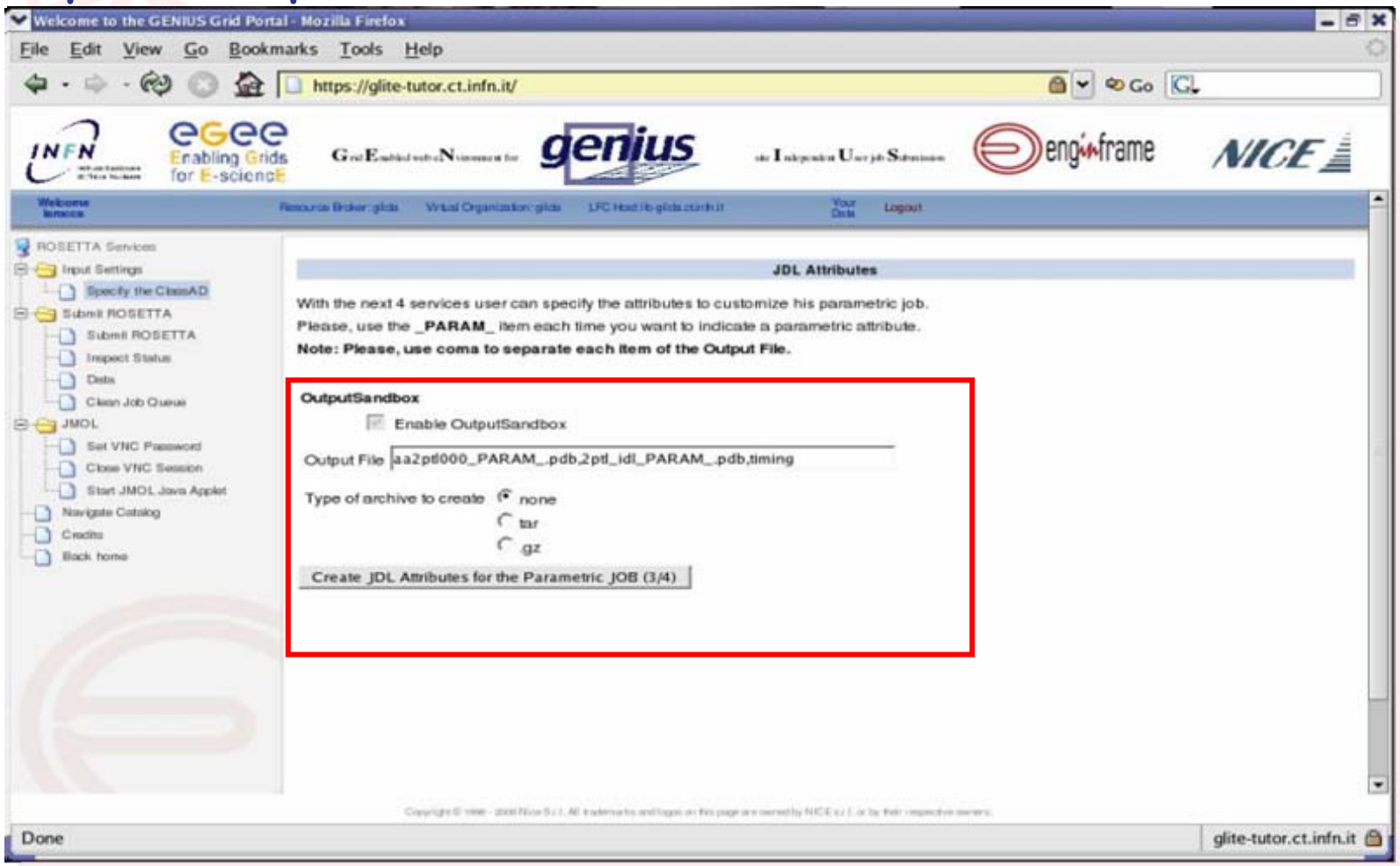
Type of archive to create ◉ none ○ tar ○ .gz

Create JDL Attributes for the Parametric JOB (2/4)

Done glite-tutor.ct.infn.it

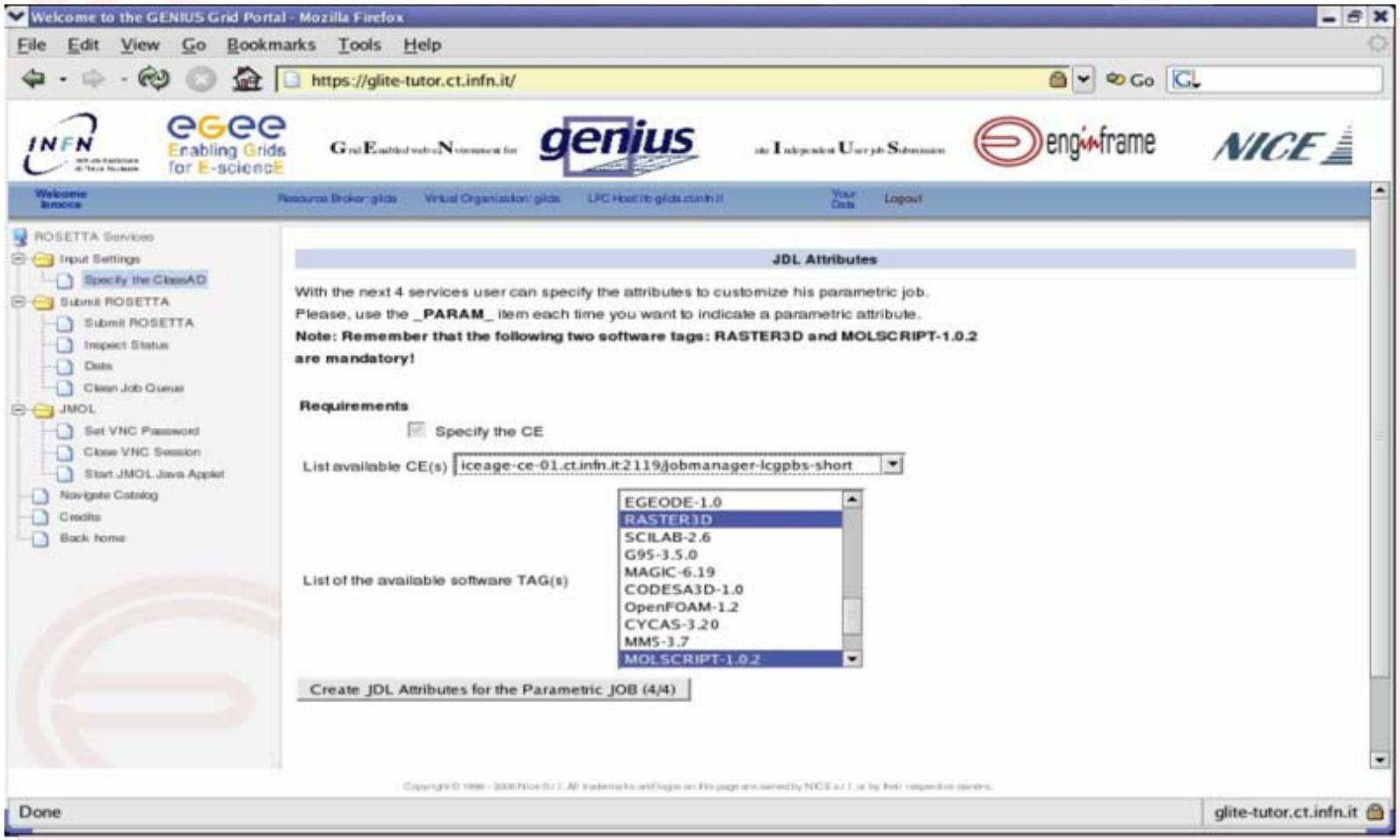
Create the dynamic ClassAD /4

- Step 4. Output files (initial and refined model coordinates) are specified in parametric form.



Create the dynamic ClassAD /5

- Step 5. The software requirements are specified in order to properly run ROSETTA.



Submit ROSETTA to the Grid /1

- Step 6. The parametric JDL file is generated and visualized to be inspected by the user.

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

ROSETTA Services
- Input Settings
  - Specify the ClassAD
- Submit ROSETTA
  - Submit ROSETTA
  - Inspect Status
  - Data
  - Clean Job Queue
- JMOL
  - Set VNC Password
  - Close VNC Session
  - Start JMOL Java Applet
- Navigate Catalog
- Credits
- Back home

Submit ROSETTA

Please, check the ClassAD before to submit it to the grid. If you need to change something use ~**Specify the ClassAD**~ service.

Note (*): You have to insert an unique production name.

```
{
JobType = "Parametric";
Parameters = 3;
ParameterStep = 1;
ParameterStart = 1;
Executable = "rosetta.sh";
Arguments = "2ptl_PARAM_.tar.gz '' _PARAM_";
StdOutput = "std_PARAM_.out";
StdError = "std_PARAM_.err";
InputSandbox = {"2ptl_PARAM_.tar.gz","rosetta.sh"};
OutputSandbox =
{"aa2ptl000_PARAM_.pdb","2ptl_idl_PARAM_.pdb","timing","std_PARAM_.out","std_PARAM_.err","2ptl_idl_P
Requirements = RegExp("iceage-ce-01.ct.infn.it:2119/jobmanager-lcgpbs-short",other.GlueCEUniqueID)
&& Member("MOLSCRIPT-1.0.2",other.GlueHostApplicationSoftwareRunTimeEnvironment) &&
(Member("RASTER3D",other.GlueHostApplicationSoftwareRunTimeEnvironment));
Rank = 3;
```

Production Name production (*)

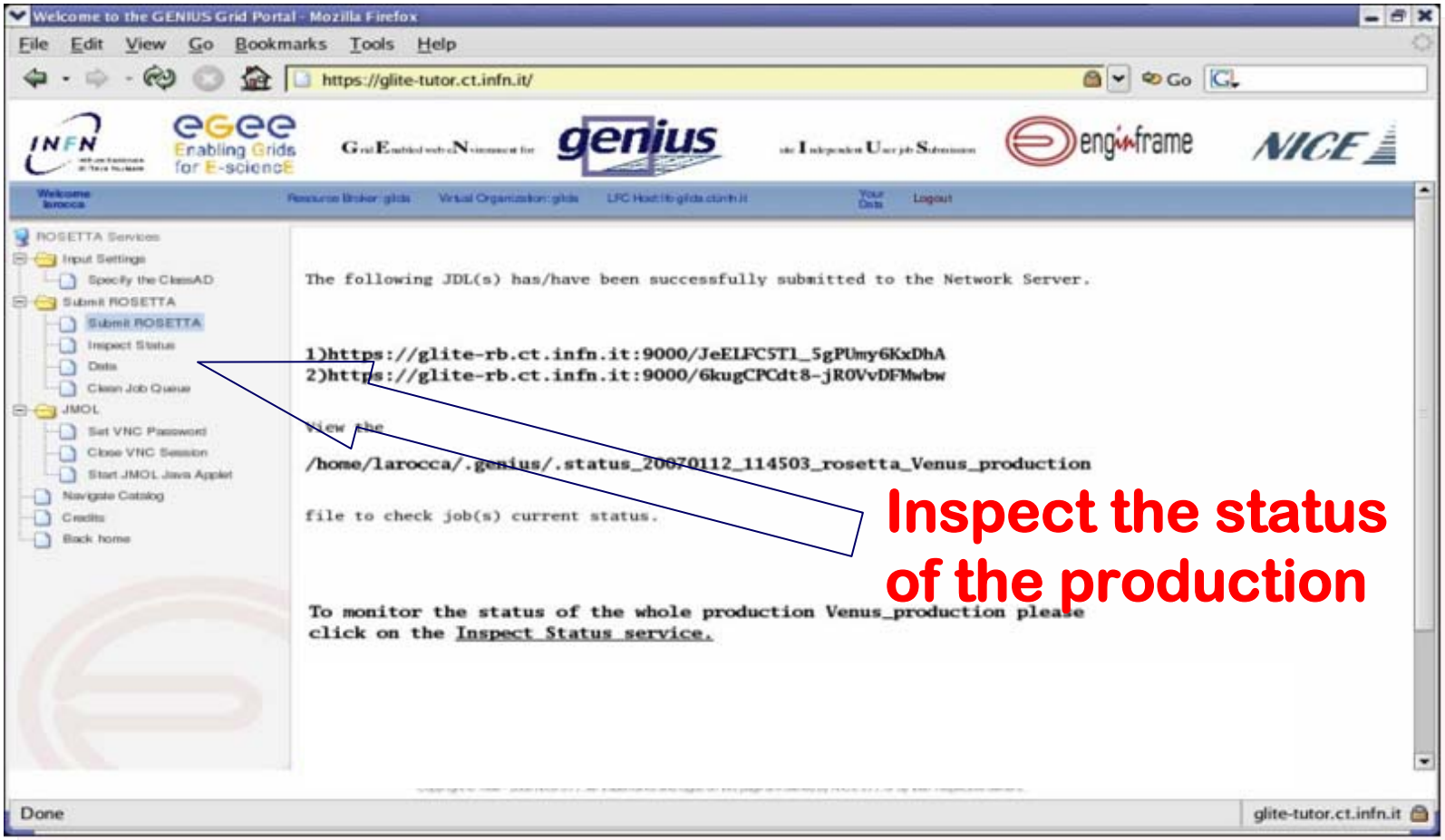
Submit ROSETTA to the GRID

Production Name

Done glite-tutor.ct.infn.it

Submit ROSETTA to the Grid /2

- Step 7. The parametric job is submitted and its status as well as the status of individual runs of the same job can be checked.



Inspect the status of the production

Inspect Status /1

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

ROSETTA Services
- Input Settings
  - Specify the ClassAD
- Submit ROSETTA
  - Submit ROSETTA
  - Inspect Status
  - Data
  - Clean Job Queue
- JMOL
  - Set VNC Password
  - Close VNC Session
  - Start JMOL Java Applet
- Navigate Catalog
- Credits
- Back home

ROSETTA Multi Queue

#	Production Name	Number of Events	Submission Time
3	production	2	Fri Jan 12 11:45:23 2007
2	MazingaZ	4	Fri Nov 17 12:58:05 2006
1	Venus	2	Thu Nov 16 11:56:27 2006

Done glite-tutor.ct.infn.it

Inspect Status /2

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

ROSETTA Services
- Input Settings
  - Specify the ClassAD
- Submit ROSETTA
  - Submit ROSETTA
  - Inspect Status
  - Data
  - Clean Job Queue
- JMOL
  - Set VNC Password
  - Close VNC Session
  - Start JMOL Java Applet
- Navigate Catalog
- Credits
- Back home

Production Name : **production**

Number of Events : **2**

Last Submission Time : **Thu Nov 16 11:56:27 2006**

Production Status: Executed

1)https://glite-rb.ct.infn.it:9000/JeELFC5Tl_5gPUny6KxDhA ==> iceage-ce-01.ct.infn.it:2119/jobmanager-lcgpbs-sh

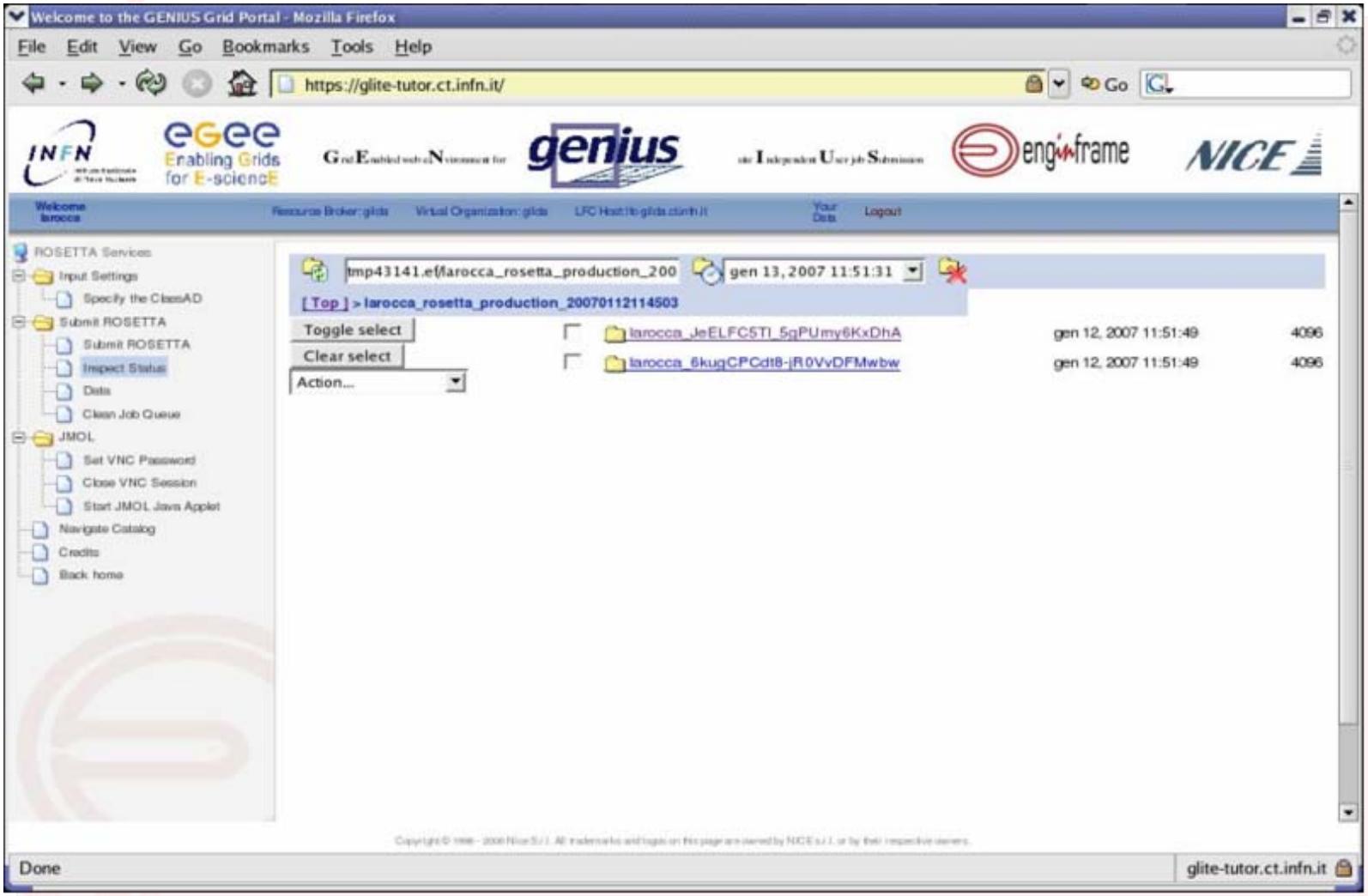
2)https://glite-rb.ct.infn.it:9000/6kugCPCdt8-jROVvDFNwbw ==> iceage-ce-01.ct.infn.it:2119/jobmanager-lcgpbs-sh

The Production has finished.

To inspect the outputs retrieved click on this link production

Done glite-tutor.ct.infn.it

Data Spooler



Navigate Catalog

Welcome to the GENIUS Grid Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://glite-tutor.ct.infn.it/

Getting Started Latest Headlines GILDA

INFN egee Enabling Grids for E-sciencE Grid Enabled web eNvironment for genius site Independent User job Submission engiwinframe NICE

Welcome larocca | Resource Broker: gilda | Virtual Organization: gilda | LFC Host: lfc-gilda.ct.infn.it | Your Data | Logout

- ROSETTA Services
  - Input Settings
    - Specify the ClassAD
  - Submit ROSETTA
    - Submit ROSETTA
    - Inspect Status
    - Data
    - Clean Job Queue
  - JMOL
    - Set VNC Password
    - Close VNC Session
    - Start JMOL Java Applet
  - Navigate Catalog
  - Credits
  - Back home

Navigate Catalog

Details of files from Catalog

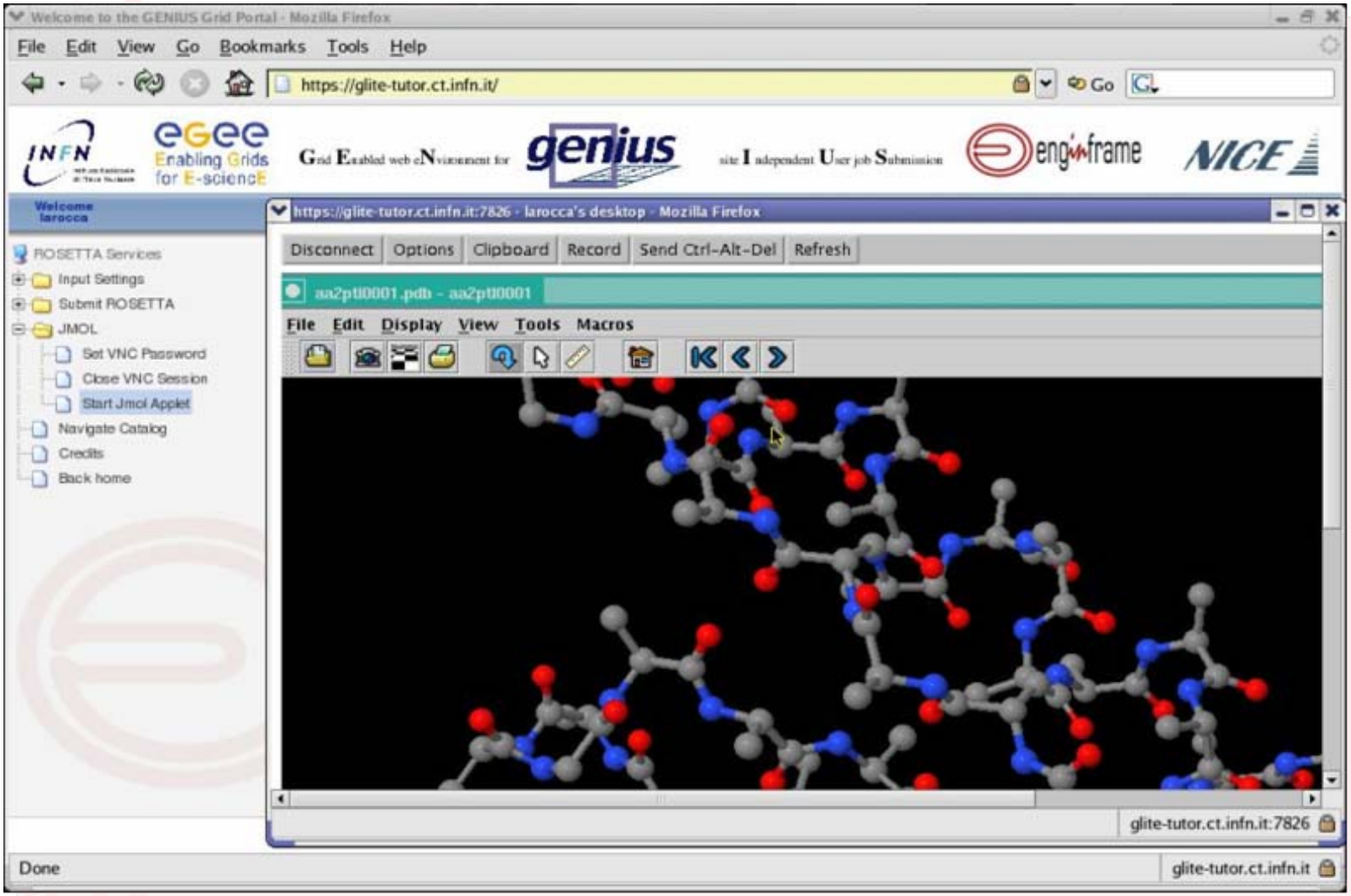
```
# file: /grid/gilda/ROSETTA/2ptl_idl1.pdb
# owner: /C=IT/O=GILDA/OU=Personal Certificate/L=MONACO/CN=MONACO01/Email=giuseppe.larocca@ct.infn.it
# group: gilda
user::rwx
group::rwx              #effective:rwx
other::r-x
------------------------------------

# file: /grid/gilda/ROSETTA/2ptl_idl2.pdb
# owner: /C=IT/O=GILDA/OU=Personal Certificate/L=MONACO/CN=MONACO01/Email=giuseppe.larocca@ct.infn.it
# group: gilda
user::rw-
group::rw-              #effective:rw-
other::r--
------------------------------------

# file: /grid/gilda/ROSETTA/aa2pt10001.pdb
# owner: /C=IT/O=GILDA/OU=Personal Certificate/L=MONACO/CN=MONACO01/Email=giuseppe.larocca@ct.infn.it
# group: gilda
user::rwx
group::rwx              #effective:rwx
other::r-x
------------------------------------
```


Done glite-tutor.ct.infn.it

JMOL Applet Java



Click here to inspect the typical output files produced by ROSETTA at the end of the prediction process

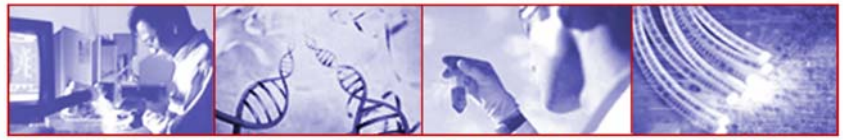
CONCLUSIONS

- We are currently accumulating data on NBP structures
- Collecting tools for analysis (structure and function analysis)
- Studying portability of other applications (e.g. function recognition software developed "in house") in GRID
- Envisioning application of ported tools for structural genomics initiatives on biomedically relevant targets
  - Example: prediction of the structure/function of the entire set of proteins of selected viral and microbial pathogens for target selection and *in silico* drug discovery

Contact us

- Giovanni Minervini (gminervini@uniroma3.it)
- Pier Luigi Luisi (luisi@mat.ethz.ch)
- Giuseppe La Rocca (giuseppe.larocca@ct.infn.it)
- Fabio Polticelli (polticel@uniroma3.it)

euchinagrid



Interconnection & Interoperability of Grids between Europe & China

Thank you for your attention !

FP6-2004-Infrastructures-6-SSA-026634

Information Society and Media