BioinfoGRID
Bioinformatics Grid Application for life science

BIOINFOGRID: Bioinformatics Grid Application for Life Science

Giorgio Maggi
INFN and Politecnico di Bari

giorgio.maggi@ba.infn.it

Project descriptions

- Bioinformatics Grid Application for Life Science (BIOINFOGRID) project.

BioinfoGRID
Bioinformatics Grid Application for life science

- The BIOINFOGRID projects proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure created by EGEE.
- In the BIOINFOGRID initiative we plan to evaluate genomics, transcriptomics, proteomics and molecular dynamics applications studies based on GRID technology.
- BIOINFOGRID will evaluate the Grid usability in wide variety of applications, and explore and exploit common solutions.

- The project start date: 1st January 2006
- kickoff meeting: last night

BIOINFOGRID: the Workpackages

WP	Description (Responsible Institution)
WP1	Genomics Applications in GRID (DKFZ)
WP2	Proteomics Applications in GRID (CNR-ITB - CILEA)
WP3	Transcriptomics Applications in GRID (UCAM-CLAB)
WP4	Database and Functional Genomics Applications (CNR-ITB - CILEA)
WP5	Molecular Dynamics Applications (CNRS)
WP6	Coordination of technical aspects and relation with RI Projects, user training, application support and resources integration. (INFN)
WP7	Dissemination and Outreach. (CNRITB, INFN, DKFZ, CNRS, UCAM-CLAB, CILEA)
WP8	Project Management Office (CNRITB)

The BIOINFOGRID resources connected to EGEE

BioinfoGRID

- Limited amount for the moment

		Computing Resources										Storage Resources		
Site ▼		GK#	Q#	RunJob	WaitJob	SlotLoad	MH#	Power	WN#	CPU#	CPULoad	Available	Total	%
INFN-T1		2	13	291	0	80%	8	-	-	-	-	373.1 TB	559.4 TB	33%
INFN-TORINO		1	9	38	0	79%	30	248K	24	48	0%	489.5 GB	2 TB	76%
INFN-TRIESTE		1	6	0	0	0%	3	4K	1	2	0%	10.7 GB	26.1 GB	59%
ITB-BARI		1	6	24	6084	100%	14	294K	12	48	20%	51 GB	58.7 GB	13%
SNS-PISA		1	6	0	0	0%	5	24K	3	6	33%	62.9 GB	67.7 GB	7%
SPACI-LECCE-IA64		1	6	0	0	0%	9	15K	7	14	0%	7.5 GB	29.4 GB	74%
SPACI-NAPOLI		1	6	2	3	67%	4	8K	3	3	0%	-	-	-
SPACI-NAPOLI-IA64		1	6	1	0	18%	9	29K	7	14	50%	49.5 GB	63.4 GB	22%
UNI-PERUGIA		1	5	1	0	4%	15	52K	13	26	4%	-	-	-
TOTAL		#39	41	262	1009	8087	2%	692	6M	532	1181	23%	400.8 TB	604.4 TB	33%

- Enabled VO:
  - Biomed, Dteam
  - Bio, Cms

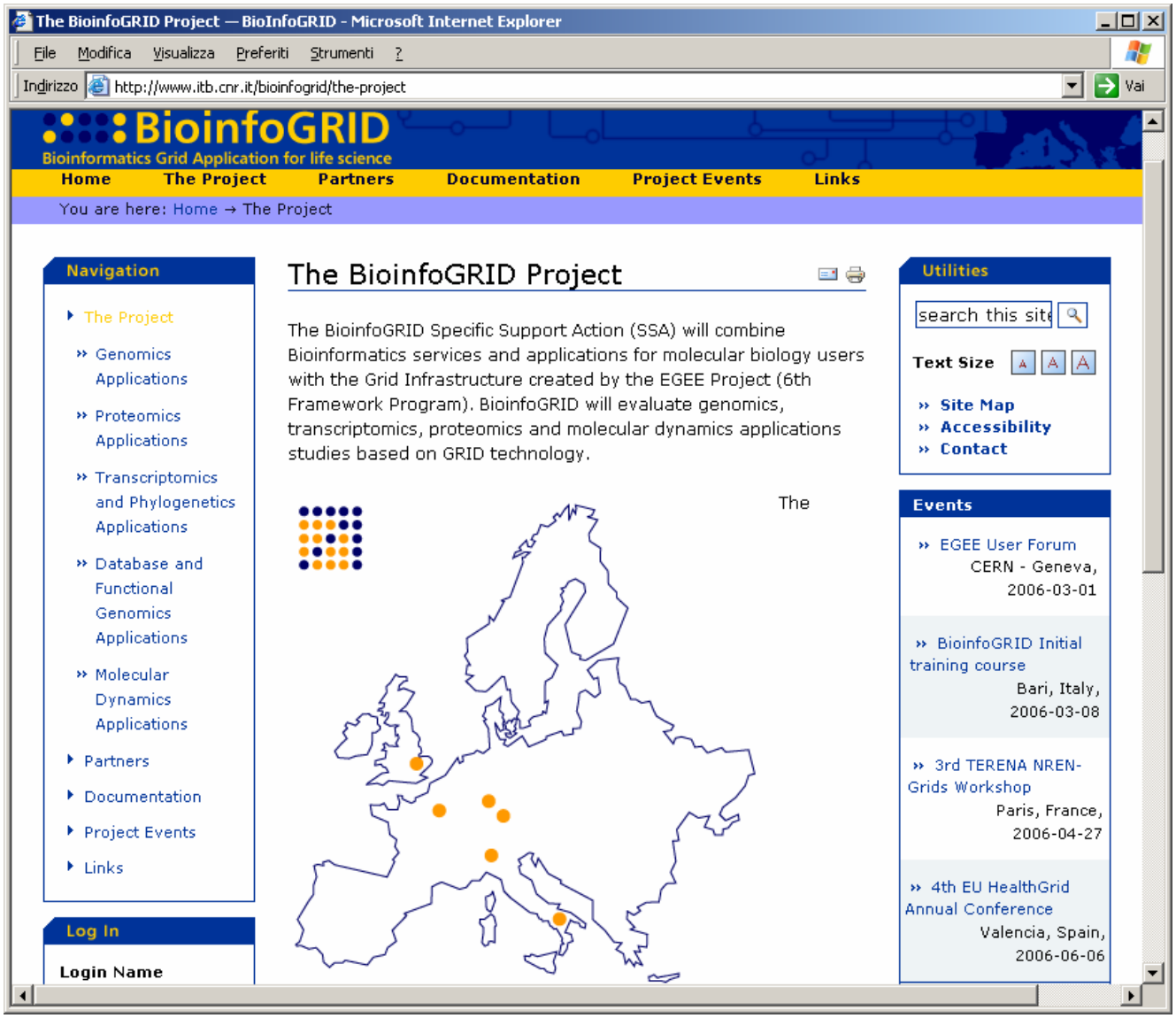
- More resources are expected to come in (in one year time)

Dissemination of knowledge: WP6 and WP7

Major objective of the BIOINFOGRID project are:

- raise the awareness, inside the bioinformatics community, about the potentialities offered by the Grid technology in solving Bioinformatics research problems.
- built a solid kernel of specialists with knowledge about the major aspects of bioinformatics applications on the GRID.
- evaluate and adopt common solutions to port the BIOINFOGRID applications to the Grid.

WP6 and WP7 activity: the web site



WP6 and WP7 activity

- **The BIOINFOGRID Initial training course**

  Bari 8-10 March 2006

- **Course objectives**
  - provide to bioinformatics users a general overview of the state of the art in the development of the Grid Middleware and infrastructures. In particular the state of LCG and gLite Middleware and of the EGEE infrastructure will be presented;
  - provide detailed technical information and precise instructions on how to use the GRID to enable new users to start using the Grid in the best possible way.
- **Course program (**http://www.itb.cnr.it/bioinfogrid/project-events/training-course-program**)**
  - The EGEE infrastructure,the GILDA t-infrastructure, the g-Lite middleware, Authorisation and Authentication, the Information System, the Workload Management System, the Data Management System, the Grid Monitoring
  - Practical
  - Access to Databases: AMGA, OGSA-DAI, G-DSE
  - Portals and workflows: GENIUS and TRIANA, Taverna, GRB
  - Examples of Bioinformatics applications and tools already deployed on the EGEE GRID infrastructure
  - Ethical issues

WP6 and WP7 activity

- The BIOINFOGRID Initial training course has been opened to other bioinformatics national projects
  (≈ 25 participants)
- Many thanks to EGEE for the really strong support.
- We expect that the trained BIOINFOGRID personnel will join the Biomed VO after the course.
- A second training course will be organized in about one year from now
- A project conference will be organized towards the end of the project
  - (I.e. one year and half from now)

Project applications (WP1 trough WP5)

The project will support studies on applications for:

- distributed laboratory management systems for microarray technology
- gene expression studies
- gene data mining
- analysis of cDNA data
- phylogenetics analysis
- distributed database access
- protein functional analysis
- molecular dynamics simulations in GRID

WP5

- Mandate: deploy Molecular Dynamics application in a grid environment
  - Evaluation of performances
- Goal: contribute to the WISDOM initiative dedicated to in silico drug discovery
  - Start from the results of WISDOM docking data challenge in 2005
  - Rerank the best hits using Molecular Dynamics
- Strategy: deployment of MD softwares on different grid infrastructures
  - Grid of PCs: EGEE-II
  - Grid of supercomputers: DEISA

WP4: Functional Analogous Finder

BioinfoGRID

Goal: compare gene products according to their described function instead of by the more conventional sequence comparison.

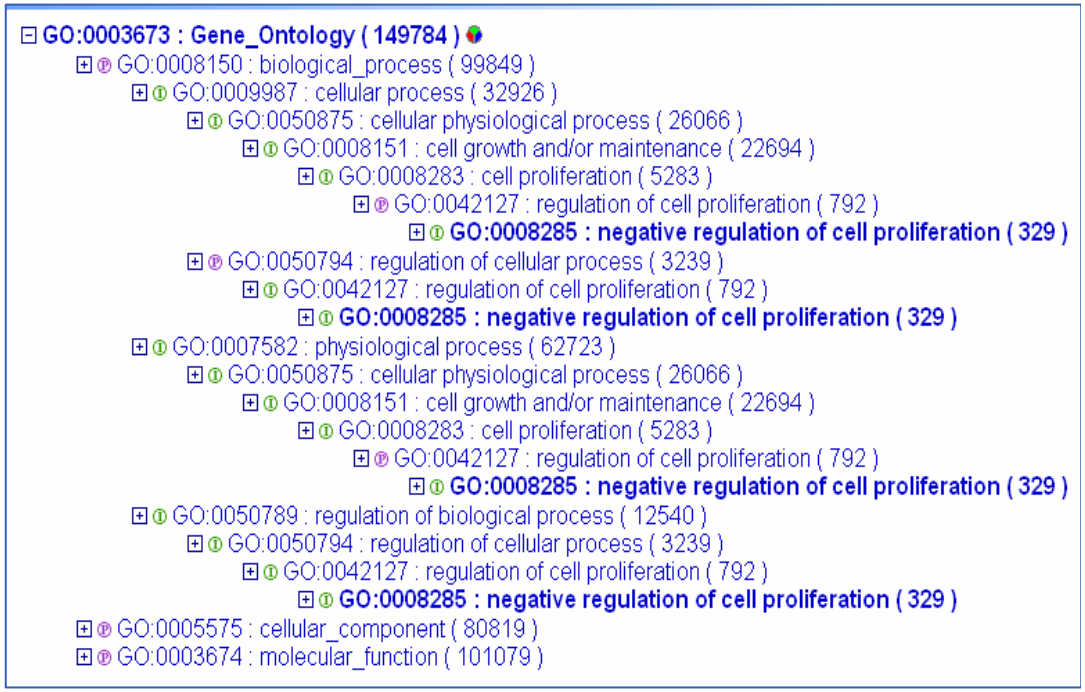
Data source: Gene Ontology (GO) and gene association

➔ 18800 GO-terms, ~ 1.3M gene products, 7.1M associations

Selection: only well described gene products are considered (>15 go terms) (≈1 million gene products)

Processing: one gene against all others ➔ 1 CPU hour

Output: text file with 100 best hits ➔ compiled as an additional packages of the GO DB

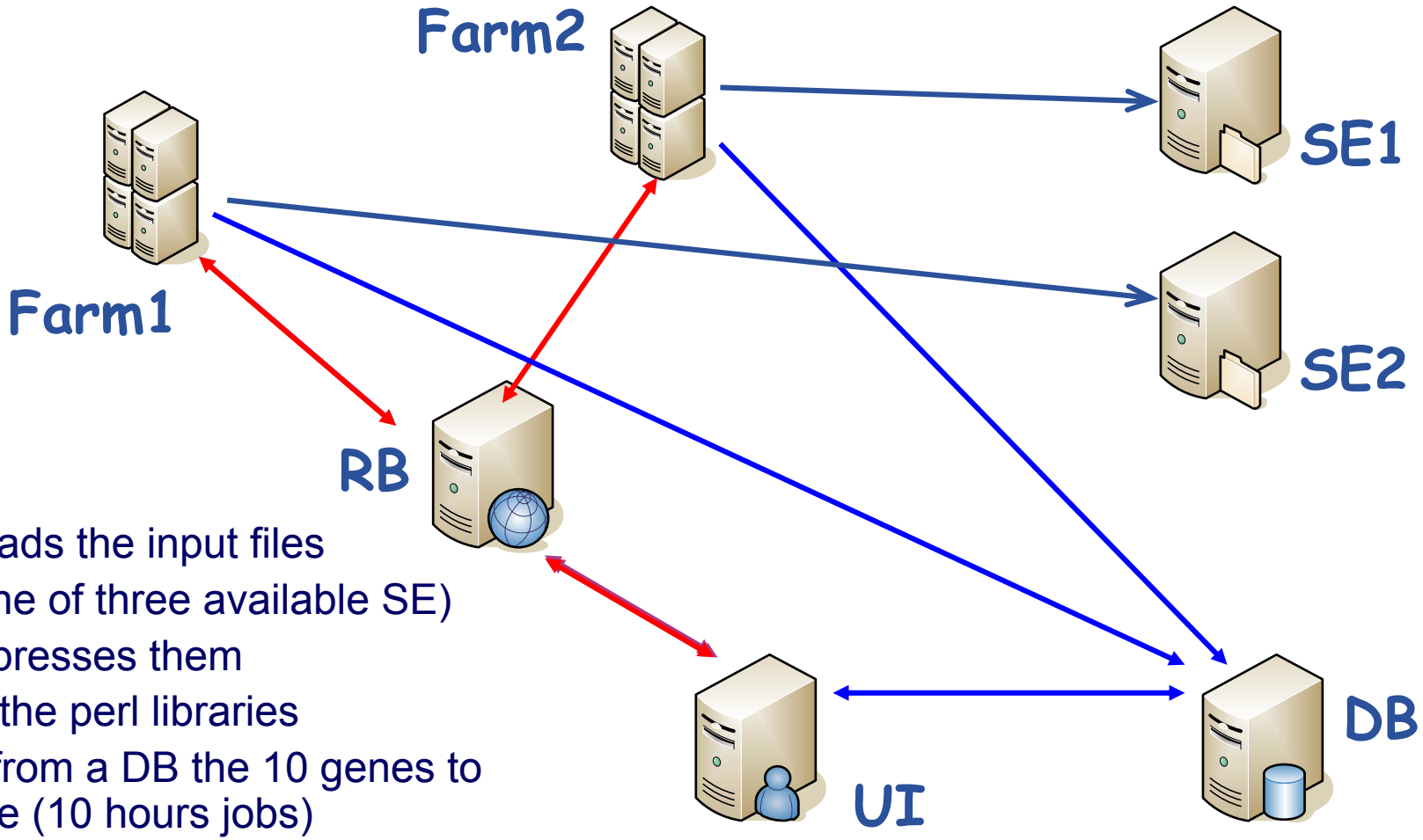


go terms associated to the BCL2_Human gene

Functional Analogous Finder: the grid implementation

- Build, once for all, a text file with all the gene products well described
  - The text file will be transferred at run time from the SE's where it is located to the WN
- The gene comparison is done by a perl script which uses statistical libraries
  - Need to install the perl libraries in every WN (do it at run time).
  - The libraries are "relocated" to avoid the need for "root" privileges
- The list of gene products to compare to all the others is stored into a central mySQL DB.
  - The DB keeps track of the completed gene products, the failed and the running ones.
  - The DB acts as "task queue" for automatic job submission

Functional Analogous Finder: when a job lands on a WN



- Downloads the input files (from one of three available SE)
- Decompresses them
- Installs the perl libraries
- Reads from a DB the 10 genes to compare (10 hours jobs)
  (chosen between the not completed genes or running ones from more then 48 hours)
- Start the perl script and the comparison

The job submitted is always the same
Don't have to worry about failed jobs

Functional Analogous Finder

- First test run (80.000 gene products) done over Christmas
- After optimization of the script and procedure
  - a new test run was started (100 k gene products)
  - It is running now using the VO “bio” on the Italian infrastructure
- We would like to extend the comparison to the full set of “well described” gene products (of the order of 1 million)
  - Using the biomed VO over the full EGEE infrastructure