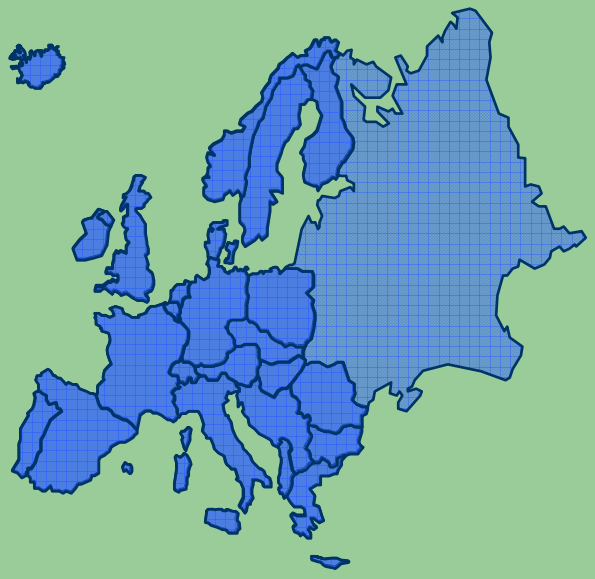
BIOINFOGRID

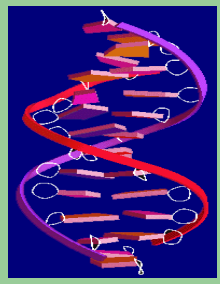
Bioinformatics Grid Application for Life Science .



COMMUNICATION NETWORK DEVELOPMENT

SPECIFIC SUPPORT ACTION

Andreas Gisel & Luciano Milanesi



Project goals

- In the BIOINFOGRID Specific Support Action (SSA) we propose combining Bioinformatics services and applications for molecular biology users with the Grid Infrastructure created by the EGEE project (6th Framework Program).

- In the BIOINFOGRID initiative we plan to evaluate genomics, transcriptomics, proteomics and molecular dynamics applications studies based on GRID technology.

Project goals

- The BIOINFOGRID SSA will establish a common ground for collaboration between the European Grid Infrastructure providers and the Bioinformatics research user community in various fields of Bioinformatics applications (Biology, Computational Chemistry, Medicine and Biotechnology).
- This will be achieved through specific studies for each reference application in the Bioinformatics domain in which experts of various disciplines can collaborate on the solution of highly complex problems.
- One of the key objectives is to “combine bioinformatics services and applications for molecular biology that have thousands of users, with the Grid Infrastructure created by the EGEE Project”.

Project Timescale

- 24 month full project duration
- Starting date 1st January 2006

Project Budget

- **1.054.000 Euro**

Project Partners

Partic. Role*	Participant name	Short name	Country
CO	Consiglio Nazionale delle Ricerche - Istituto Tecnologie Biomediche	CNR	Italy
CR	Istituto Nazionale di Fisica Nucleare	INFN	Italy
CR	Deutsches Krebsforschungzentrum	DKFZ	Germany
CR	Centre National de la Recherche Scientifique	CNRS	France
CR	The Chancellor, Master and Scholars of the University of Cambridge	UCAM	UK
CR	Consorzio Interuniversitario Lombardo Elaborazione Automatica	CILEA	Italy
CR	Steinbeis GmbH & Co	StC	Germany

Relationships to EGEE

- BIOINFOGRID aims to use
  - the EGEE infrastructure
  - and the gLite middleware
- BIOINFOGRID will join the Biomed VO and evaluate the opportunity of setting-up a specific BIOIFOGRID VO.
- BIOINFOGRID will bring resources to the EGEE infrastructure:
  - A farm of 28 CPU of ITB have been recently added to the INFN GRID/EGEE Infrastructure.
  - Of the order of 600 CPU will be available during next year in the BIOINFOGRID community (>10% of which will be accessible from the EGEE infrastructure)

Relationships to EGEE and other projects

- BIOINFOGRID partner are also involved in HEALTHGRID, EMBRACE, and EGEE.
- Potential collaboration with BIOSAPIENS, DILIGENT, ICEAGE, EUCHINAGRID, ETICS, EUMEDGRID, EELA, SEEGRID, BalticGrid, ISSSeG, eIRGSP, and BELIEF projects is envisaged.

Project applications
(First 5 WPs)

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

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.

Aims and Needs

Most of the Bioinformatics applications need, in a way or another, publicly available databases (DB).

We are looking for:

- the major public biological DBs deployed on the GRID (~600Gb, flatfile and RDBM),
- multiple copies of those DBs for an efficient and seamless access,
- a synchronized updating system since those DBs are regularly updated,
- a common middleware (API) for unified access.

Aims and Needs

Many Applications are data and/or processing intensive

We need:

- to parallelize the applications (divide them in many independent jobs),
- to have access to CPUs and storage (VO-biomed, VO-bioinfogrid),
- to have access to stable and/or temporary replications of certain DBs.

Aims and Needs

Bioinformatics applications run either as batch or on-demand.

- Simple, user friendly access to the GRID for “simple user”,
- the possibility to access CPUs outside VO-biomed / VO-bioinfogrid for “advanced users”,
- the possibility of launching clusters of jobs

are highly desirable.

Aims and Needs

Many applications involve several different tools

- An efficient workflow managing systems
- and an efficient data managing system

are highly desirable.

Applications ready to be deployed

Functional Analogous Finder

Goal: comparing gene product according to their described function instead of by the conventional sequence comparison.

Data source: Gene Ontology (GO) and gene association
- ➔ GODB: 18800 GO-terms, ~ 1.3M gene products, 7.1M associations

Processing: one gene against all others
- ➔ 1 CPU hour

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

Tested Solution: temporary distribution of GODB over several worker nodes and parallel processing of a sub list of input gene products

Applications ready to be deployed

Unfolding and Ion binding events in gramicidin

Goal: Study the effect of ions on stability and structure of the gramicidin in two different conformations.

Molecular Dynamics program

Entry: PDB structure file: 1000 atoms in (1), 2500 atoms in (2) ; starting parameters ➔ 1M

Process: 1) Run 10 MD trajectories each of 4 ns ➔ 266 CPU h.
2) Run 10 MD trajectories each of 2.5 ns ➔ 419 CPU h

Output: MD trajectory, describing time evolution of the system

Applications ready to be deployed

World-wide In Silico Docking On Malaria (WISDOM)

GOAL: virtual screening for in silico drug discovery for malaria

First step in 2005: docking data challenge

- Total of about 46 million ligands docked in 6 weeks
- 1TB of data produced
- Up 1000 computers in 15 countries used simultaneously corresponding to about 80 CPU years

BIOINFOGRID goal: extend WISDOM virtual screening pipeline to Molecular Dynamics

- Starting from EGEE data challenge results

Applications ready to be deployed

Analysis of the protein surface

Goal: Define a protein surface model by defining the exposed residues and to search for protein similarities by such models.

Entry: pdb files containing the atomic coordinates of proteins
pdb-database ➔ 28800 datasets

Process: 3 steps - volumetric description →15 min CPU/prot
- protein surface description → 10 min CPU/prot
- decompose in parametres → 15 min CPU/prot

Output: A database of protein surface parameters describing the surface characteristics in some protein functionality site.

Compare the protein surface to establish relationship between proteins → 25 min CPU/prot