GPS@, Web interface for Protein Sequence Analysis on Grid

Blanchet C., Mollon R., Combet C. and Deleage G.

Pôle Bioinformatique de Lyon – PBIL

Institut de Biologie et de Chimie des Protéines

IBCP CNRS UMR 5086

Lyon – Gerland, France

Christophe.Blanchet@ibcp.fr

www.eu-egee.org

EGEE and gLite are registered trademarks

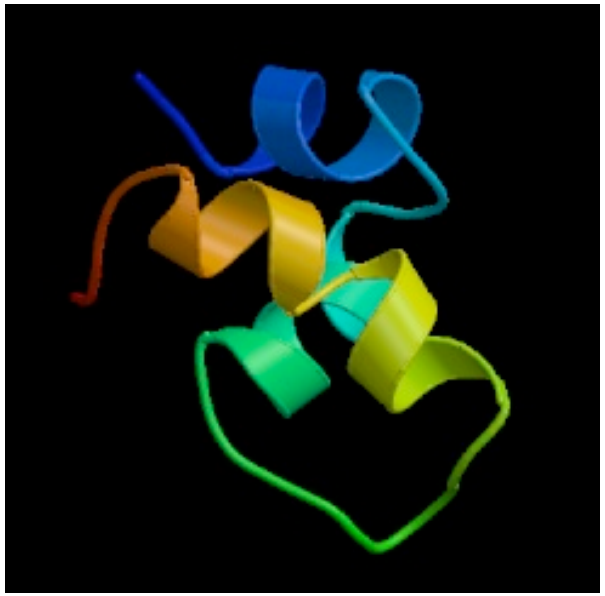
Institute of Biology and Chemistry of Proteins

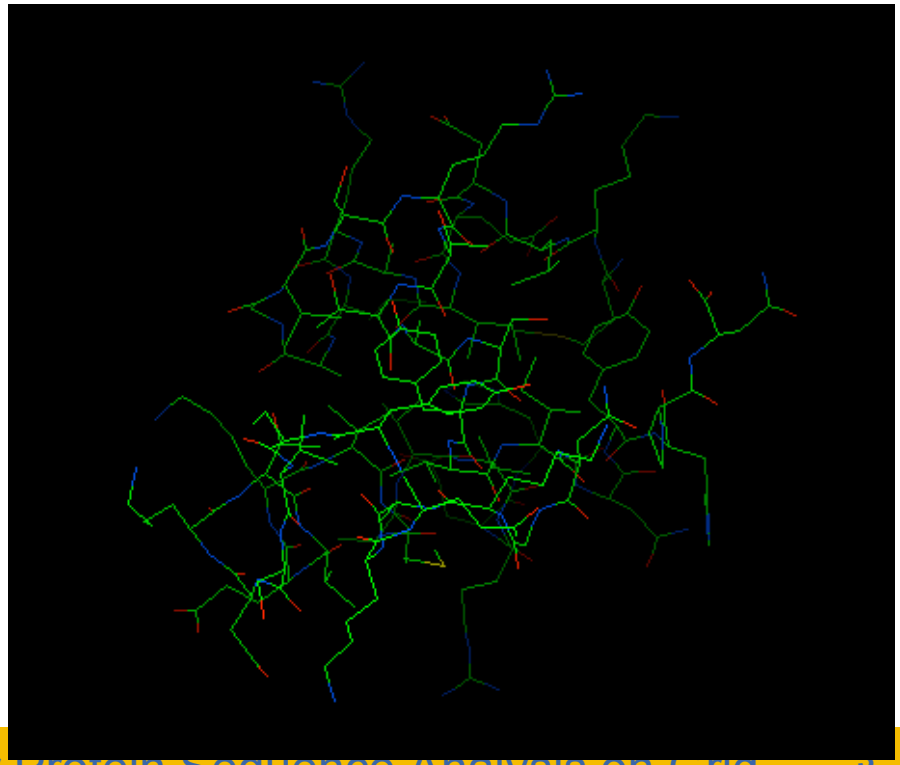
- **French CNRS Institute, associated to Univ. Lyon 1**
  - Life Science
  - About 160 people
  - http://www.ibcp.fr
  - Located in Lyon, France

- **Study of proteins in their biological context**
  - Approaches used include : integrative cellular (cell culture, various types of microscopies) and molecular techniques, both experimental (including biocrystallography, and nuclear magnetic resonance) and theoretical (structural bioinformatics)
- **Three main departments, bringing together 13 groups**
  - Topics such as cancer, extracellular matrix, tissue engineering, membranes, cell transport and signalling, bioinformatics and structural biology

Protein ?

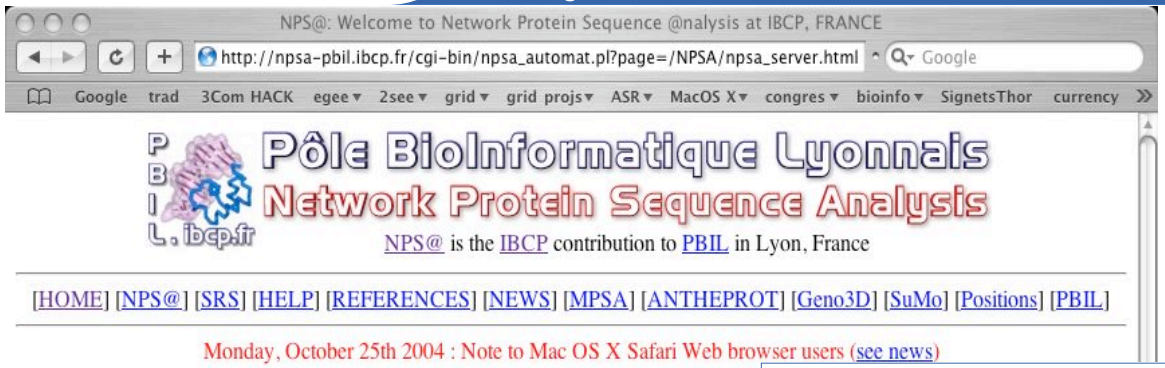
- **FRUCTOSE REPRESSOR DNA-BINDING DOMAIN, NMR, MINIMIZED STRUCTURE**
  - Penin, F., Geourjon, C., Montserret, R., Bockmann, A., Lesage, A., Yang, Y., Bonod-Bidaud, C., Cortay, J.C., Negre, D., Cozzone, A.J., Deleage, G
    - >1UXC:_|PDBID|CHAIN|SEQUENCE
      MKLDEIARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREH
      NYHPNAVAAGLRLQHHHHHH





NPS@: Bioinformatics Web Portal

Enabling Grids for E-sciencE

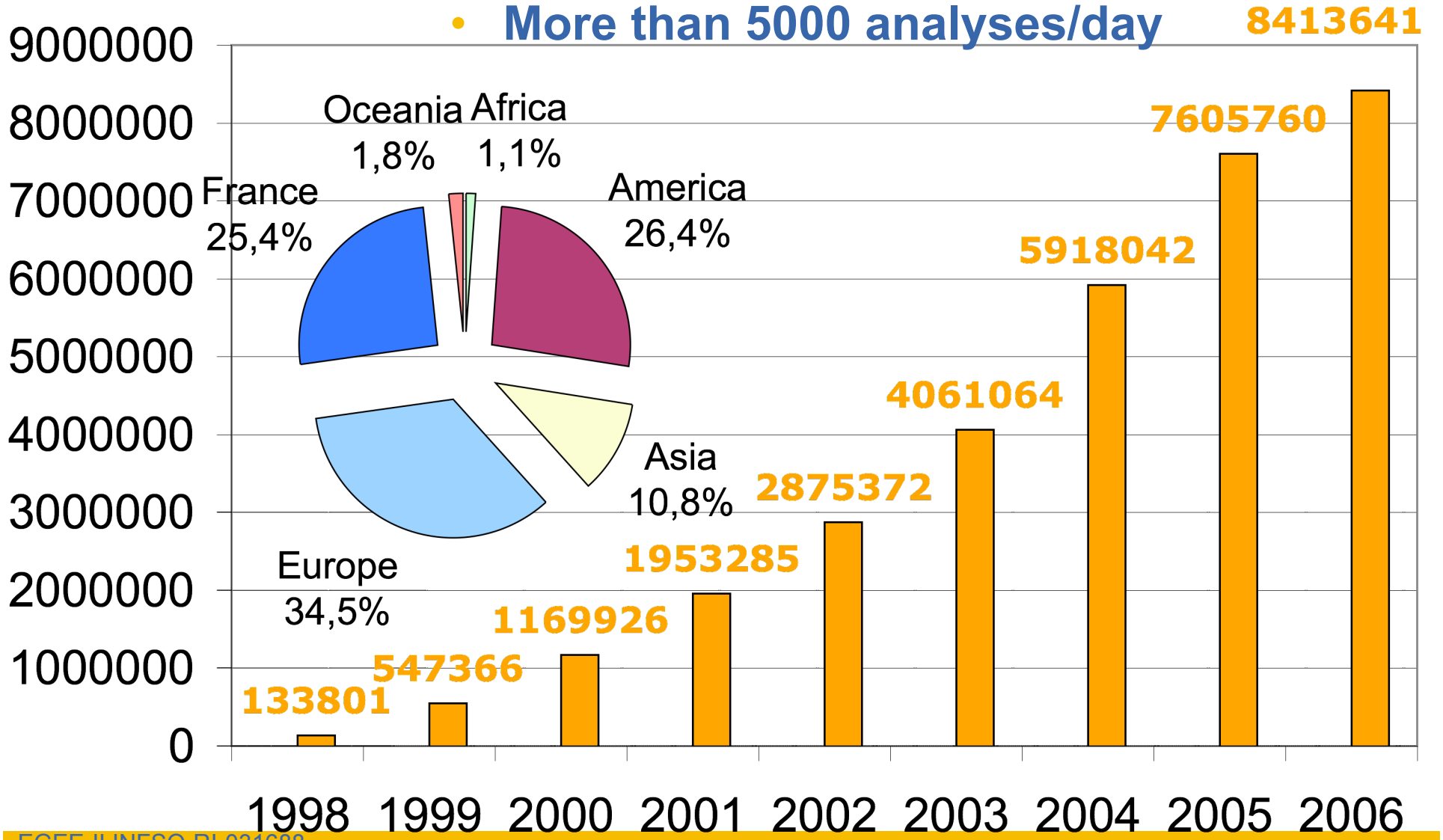


- **Network Protein Sequence Analysis (NPS@ release 3)**
  http://npsa-pbil.ibcp.fr
- **Online since 1998**
- **46 integrated methods for protein sequence analysis**
- **12 Online up-to-date biological databanks**

- **International pointers: Expasy (Ch) , University of California, ...**
- **Ref.: " NPS@: Network Protein Sequence Analysis", Combet C., Blanchet C., Geourjon C. et Deléage G. Tibs, 2000, 25, 147-150.**

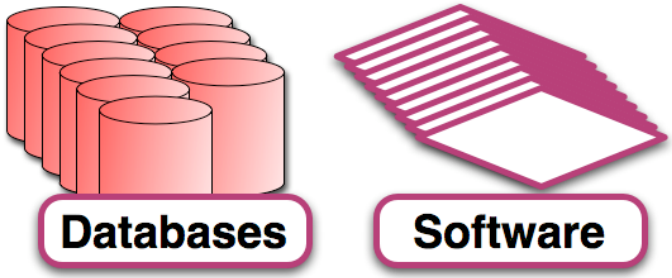
NPS@ hits

- More than 8 millions analyses since 1998
- More than 5000 analyses/day



Biological Data and Tools

- **Numerous**
  - + 800 (Galperin *et al.*, 2006)
- **Heterogeneous**
  - Data & metadata!
    - Swiss-Prot: 12 % of data, 88% MD
    - TrEMBL: 19% data, 81% MD
  - Size: kB to GB
  - Authors & initial location
  - Storage: file, object, image
  - Format: EMBL, GenBank, Pearson-Fasta, PDB, pubmed, …
- **Updatable !!**
- **In some case sensitive (Patient, Industrial, Scientific)**



- **Numerous**
  - BioCatalog: + 600 at end of 1990s
  - EMBOSS toolkit: + 200
- **Heterogeneous**
  - Bioinformatics algorithm: Sequence similarity, Multiple alignment, Structural prediction, …
  - Execution: sequential, MPI, openMP, …
- ***Data I/O***
  - *Text files*
  - *Specific format*
  - *Local I/O only*

SIZE OF SOME BIOLOGICAL DATABASES

Name	Nature	Rel.	Entries	Size (MB)
GenBank	Gene Sequence	153	56,620,500	224,000
EMBL	Gene Sequence	86	69,783,593	~100,000
Swiss-Prot	Protein Sequence	49.5	216,380	824
TrEMBL	Protein Sequence	32.5	2,807,081	6,347
PROSITE	Protein Signature	19.25	1,411	14
pFAM-A	Protein Signature	19.0	8,183	2,104
PDB	Protein Structure	04/2006	36,121	23,316

Grid-enabling tools with XML wrapping

Describe to gridify!

Description of an application

- Arguments, output, options,
- Introducing attribute about biological semantic

XML-based:

- Bio_methods.dtd
- Blast.xml, pattinprot.xml, clustalw.xml, …

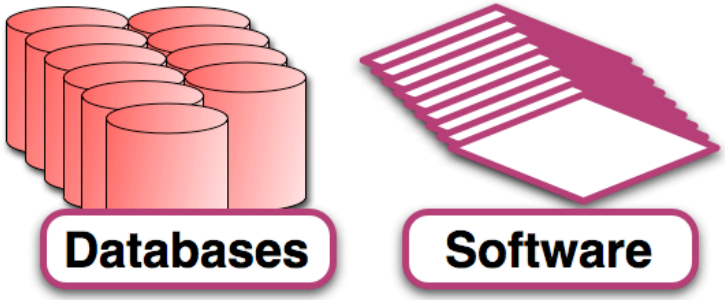
Use application descriptors

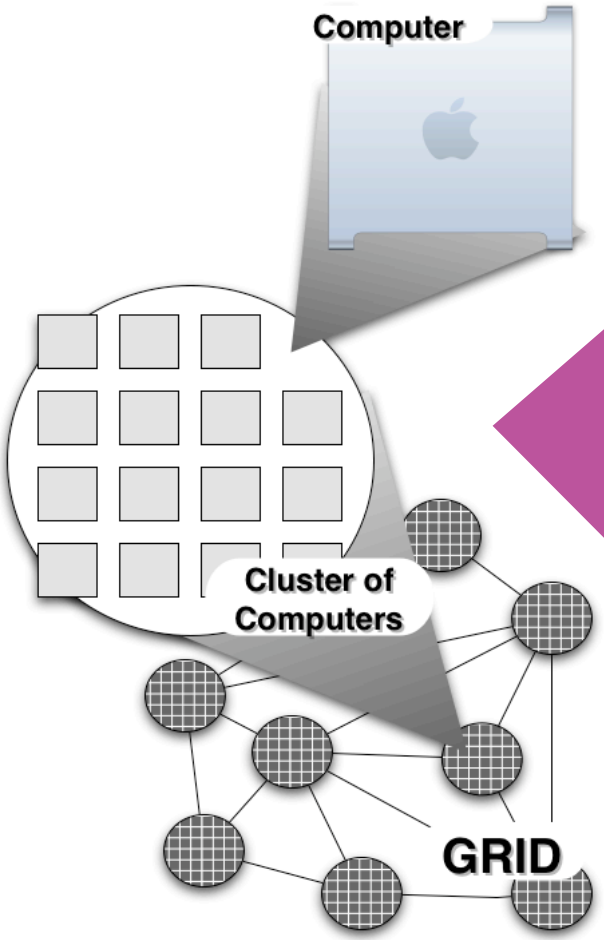
- To build user interface: *bio_cgi* for html
- To manage jobs: *bio_launcher* to launch job on cluster or grid

Example of pattinprot.xml

```
<?xml version="1.0"?>

<!DOCTYPE bio_method SYSTEM  "/opt/bio/etc/bio_method.dtd">

<bio_method version="2.0" mode="egee" >

<method name="PATTINPROT" class="scanprot" type="sequential"
root="/var/www/gpsa.ibcp.fr/pbil/servers/gpsa/w3-gpsa/" >

<bio_binary path="gbio_lfn://PATTINPROT/newpattinprot"
arch="i686" version="1" />

<bio_parameter usage="cliIO" >

<parameter class="sequence_bank" type="file" option="-p"
value="gbio_lfn://WORK_SPACE/PATTINPROT_0.inputdata"
visibility="external" IO="in" />

<parameter class="pattern_bank" type="file" option="-m"
value="gbio_lfn://WORK_SPACE/PATTINPROT_1.inputdata"
visibility="external" IO="in" />

<parameter class="result" type="file" option="-r"
link="biodata" value="gbio_lfn://WORK_SPACE/pattinprot.out"
visibility="external" IO="out" />

</bio_parameter>

</method>

</bio_method>
```

Grid-enabled Biological Resources





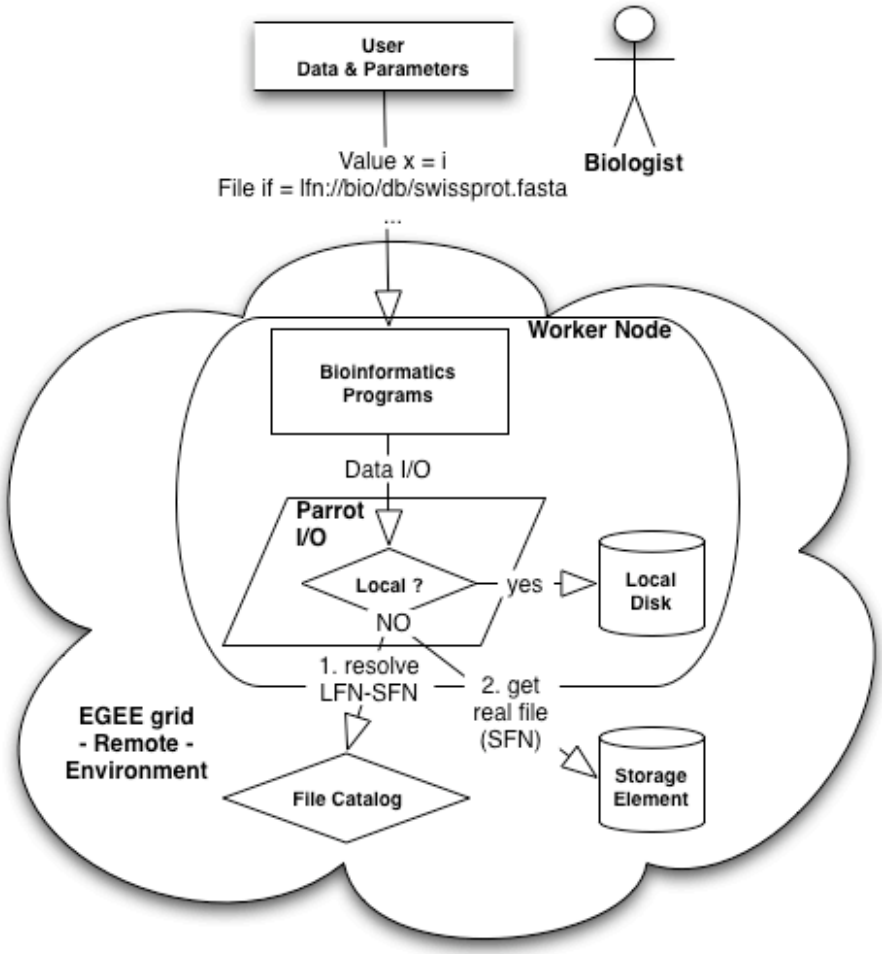
Data and Tools adapted to the GRID

Legacy Bioinformatics Applications :
* Wrapping tools with XML descriptors
* BLAST, SSEARCH, FASTA, ClustalW, MultAlin, PattInProt, ...

Distributed databases:
* **Encrypting** with EncFile System
* **Swiss-Prot**:
lfn:/grid/biomed/db/swissprot/last/sprot.fas
Symbolic link to .../swissprot/50/4/sprot.fas
* **TrEMBL:**
lfn:/grid/biomed/db/trembl/33/4/sptr.fas

Details : http://gbio.ibcp.fr

Data Virtualization



Perroquet

- ***IBCP's extension to Parrot tool***
- Adding EGEE file namespace
  - URL recognition
- Adding EGEE name resolution
  - Querying File Catalog (RLS, LFC)

Based on *Parrot* Tool

- Collaboration with D. L. Thain (Univ. ND, USA).
  Paper accepted at GRID'2006
- Custom I/O: chirp, ftp, gsi-ftp, …

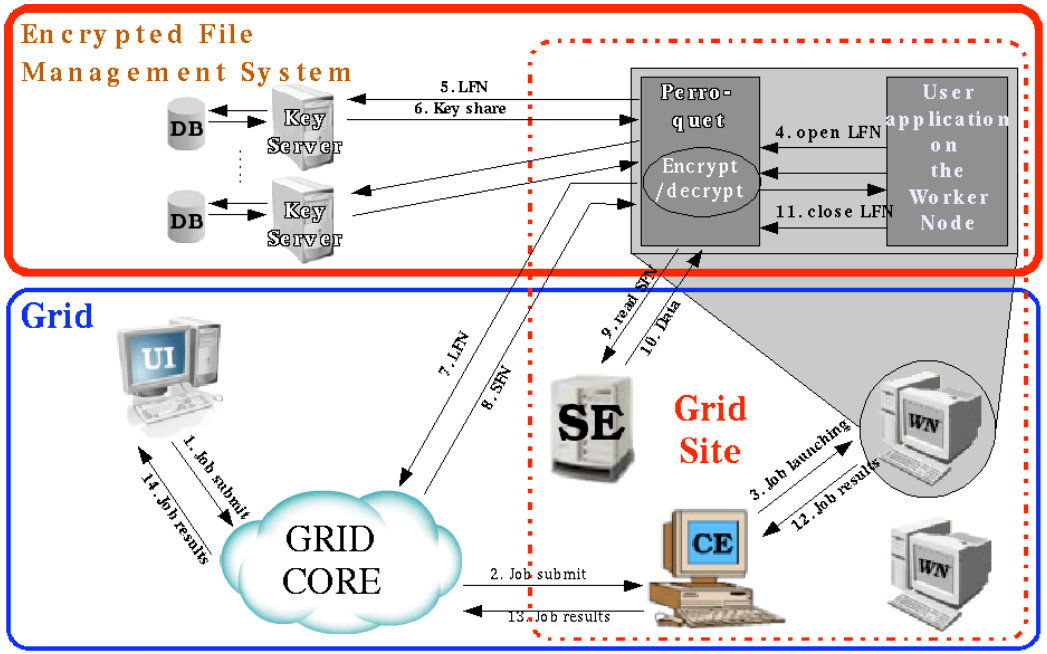
Conclusion

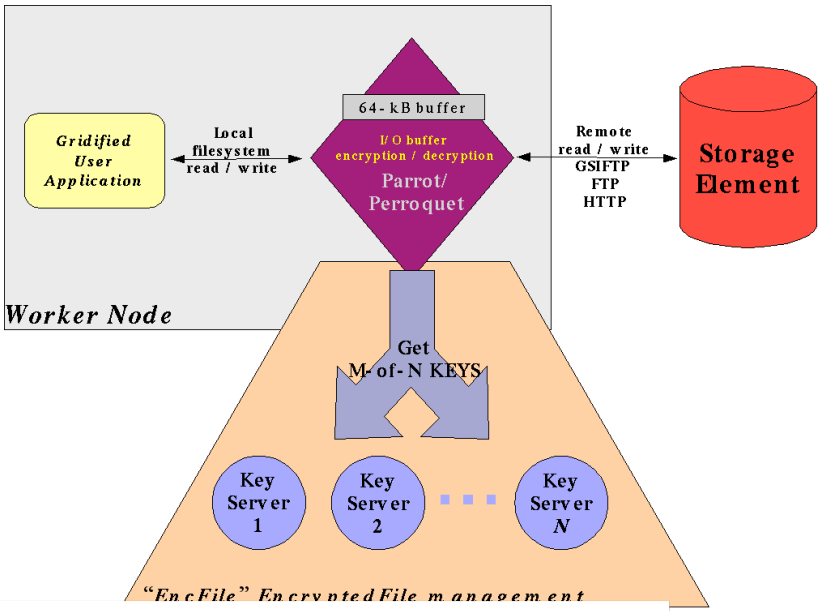
- Grid-enabling legacy applications on EGEE Grid
- Good performances

Christophe Blanchet, Rémi Mollon, Douglas L. Thain and Gilbert Deléage
Grid Deployment of Legacy Bioinformatics Applications with Transparent Data Access
To be presented at Conference GRID'2006, Barcelona, Sept. 28-29, 2006

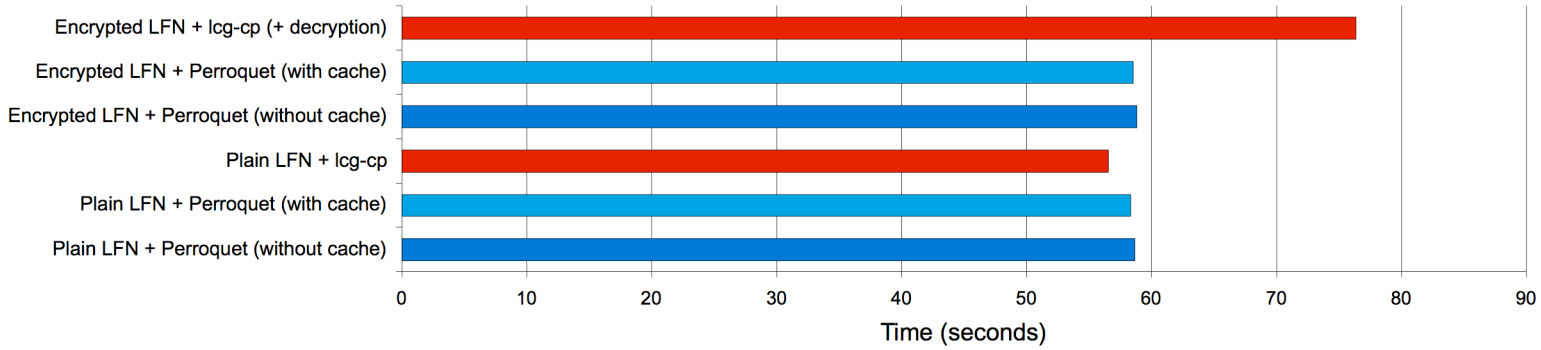
Secure Data Virtualization

C. Blanchet, R. Mollon and G. Deleage.
Building an Encrypted File System on the EGEE grid: Application to Protein Sequence Analysis.
IEEE Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06)



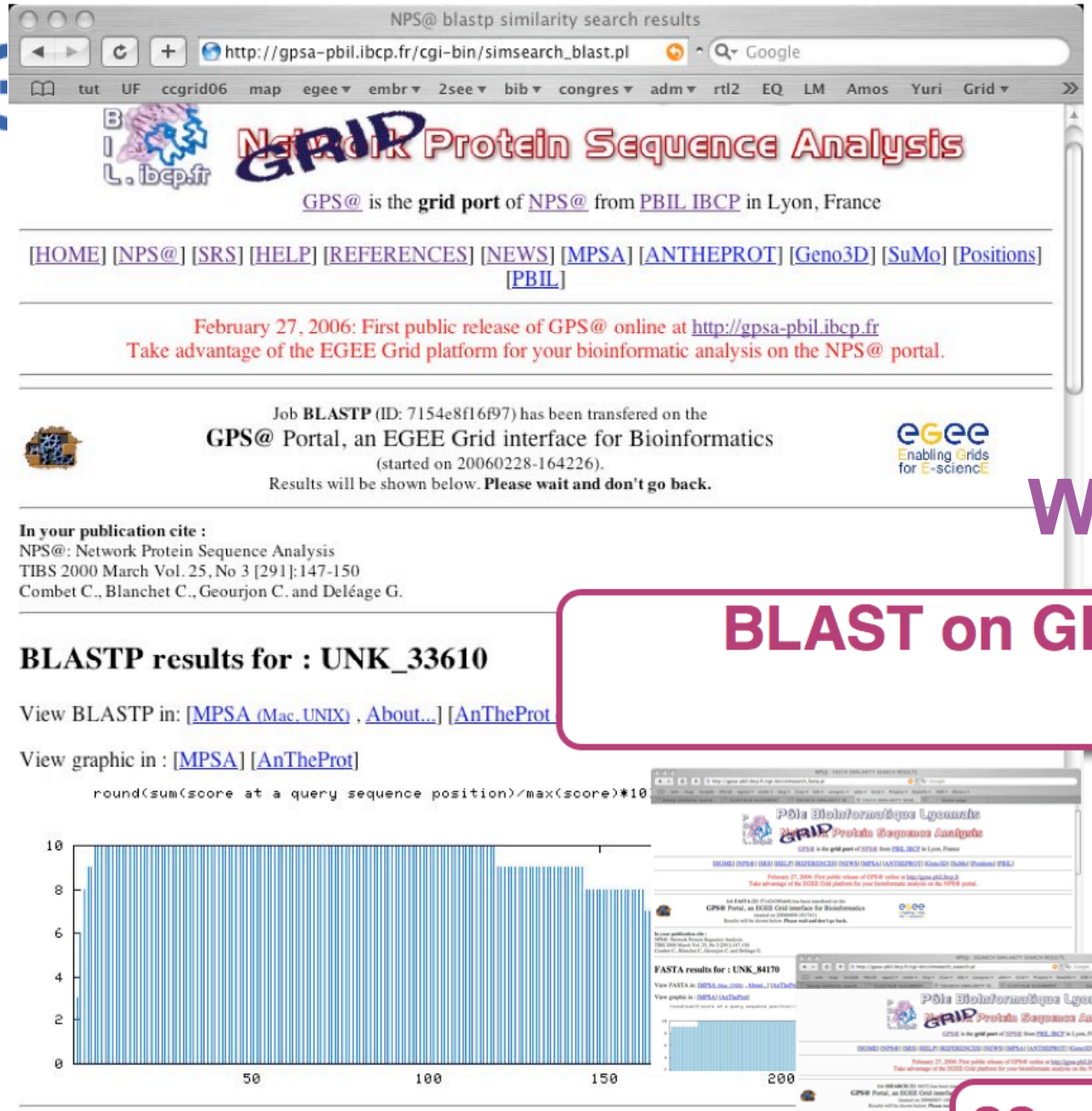


Time to download a 205-MB gridified file



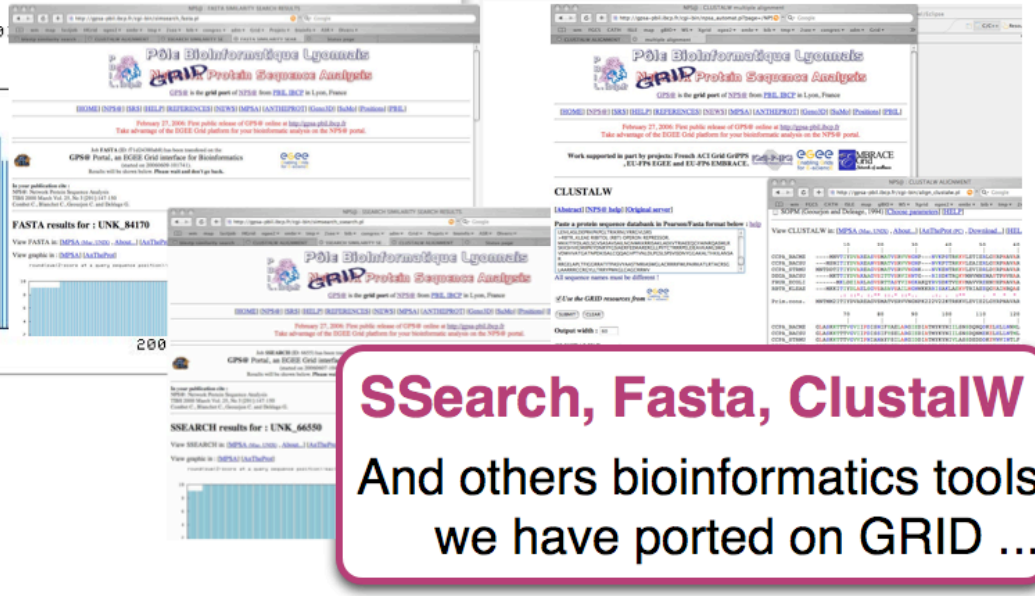
Computation Virtualization:

Web portal GPS@



BLAST on GRID

But also ...

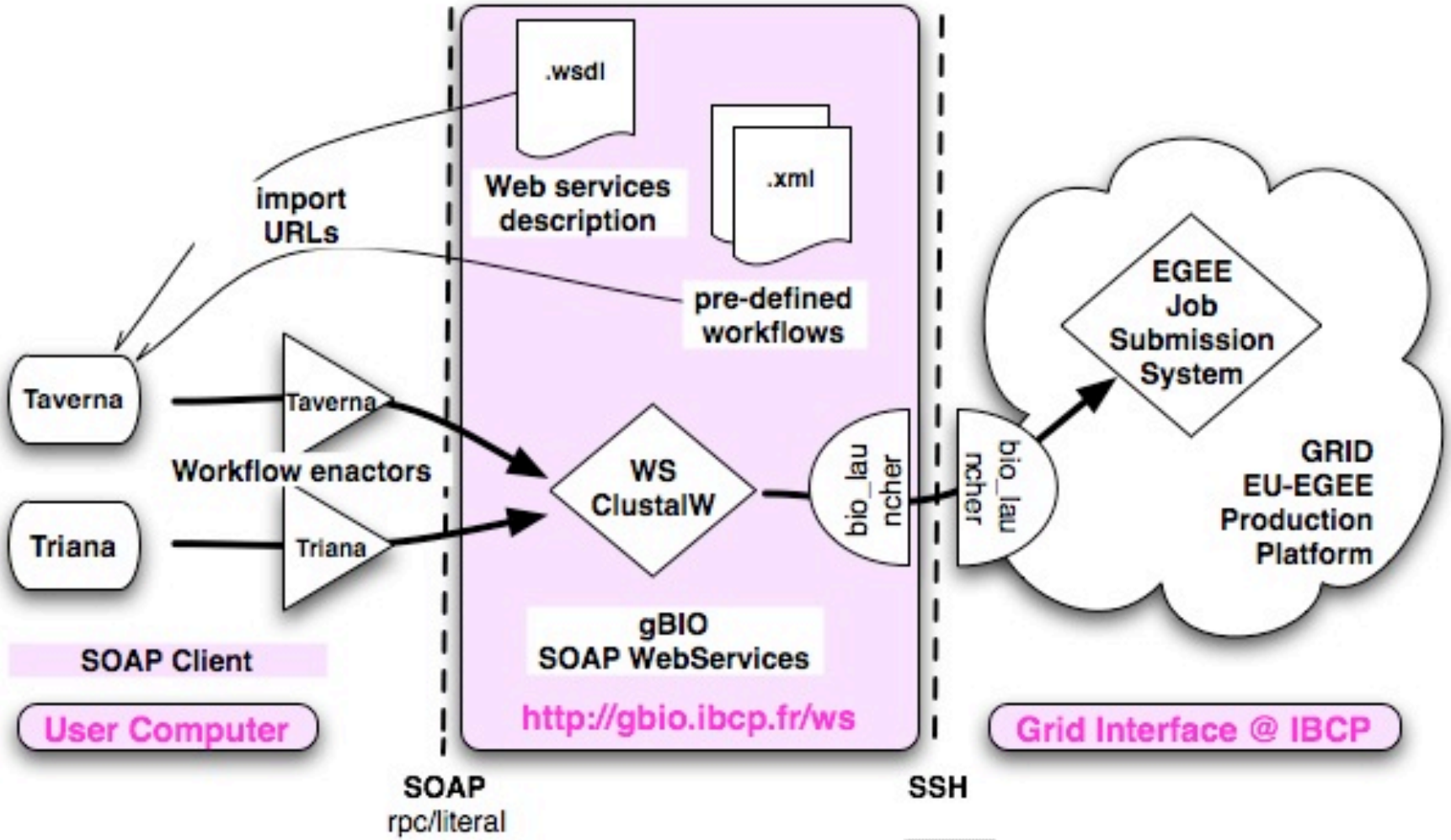


SSearch, Fasta, ClustalW

And others bioinformatics tools we have ported on GRID ...

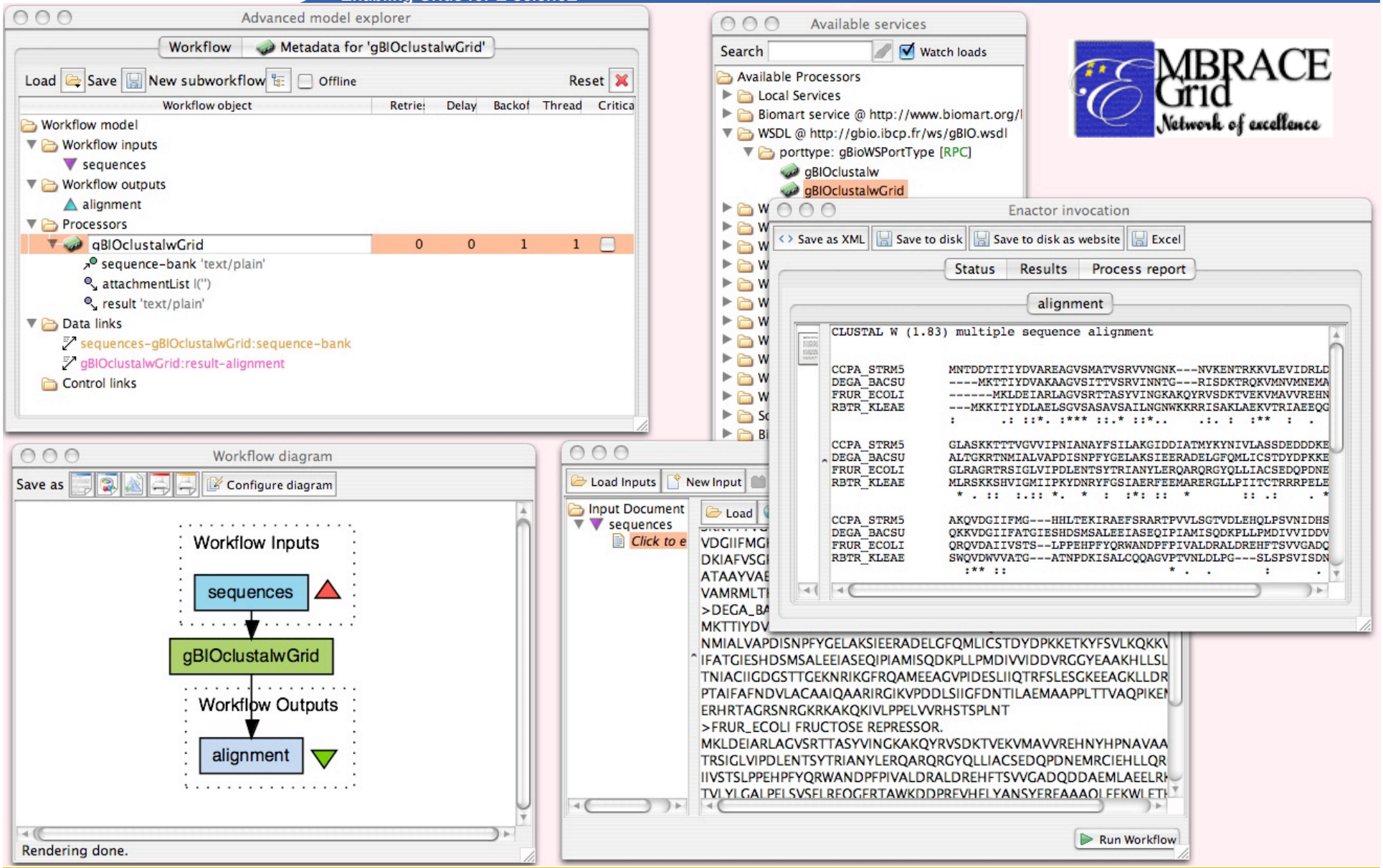
Grid Protein Sequence Analysis (GPS@)

Comp. Virtualization: Web Services

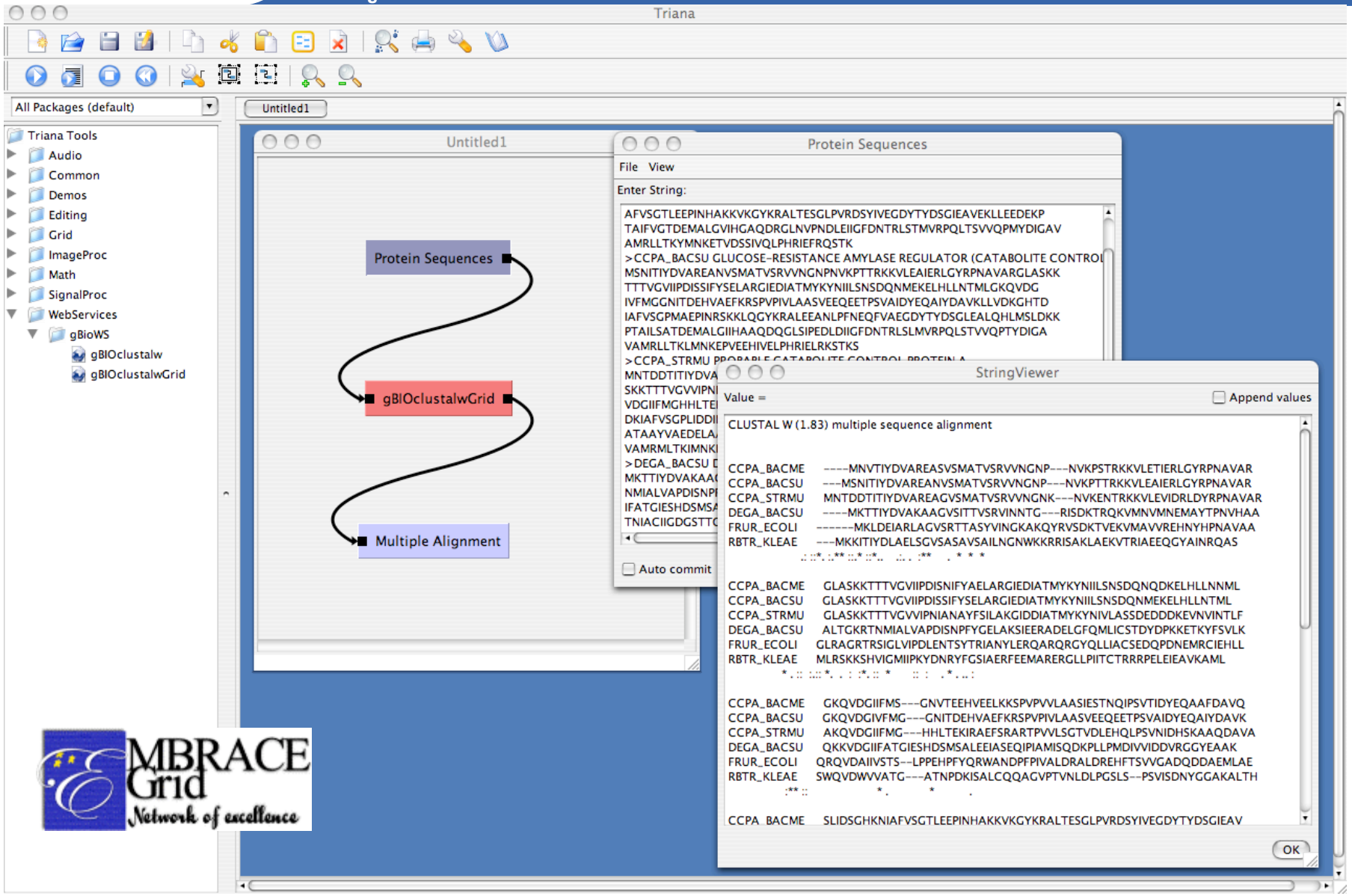


Joined work with EU-FP6 EMBRACE (LHSG-CT-2004-512092)

Bioinformatics Workflow: Taverna



Bioinformatics Workflow : Triana



Conclusion

- **Grid-enabling Biological data and tools**
- **Deploying biological databases**
- **Workload management**
  - Wrapping tools with generic XML-based framework
  - Toolkit to remotely run jobs: *bio_launcher*
- **Data management**
  - Grid-enabling legacy application with remote data access
  - Transparent encryption system of data (mgmt and access)
- **User interface**
  - **Web Portal**: **GPS@** at gpsa-pbil.ibcp.fr
  - **Web Services**: details at gbio.ibcp.fr/ws
- **Pending issues**
  - Short Jobs (<5 min): SDJ workgroup has decreased grid middleware overhead to 2 min: only one site (LAL) enabled
    => deploying this SDJ recommendation on other biomed sites
  - Data management: still some security issues in gLite data management system, DLI interface cannot be activated on biomed LFC
    => RB should use user credential to access DLI interface on LFC

Acknowledgement

Science collaborators

- D.G. Thain (Univ. ND, US)
- Y. Denneulin (IMAG, Fr)
- Members of the EU-FP6 EGEE project

Team collaborators

- C. Blanchet
- R. Mollon (EGEE fellow)
- V. Daric (EMBRACE fellow)
- C. Combet
- G. Deléage (Team Leader)

Work supported in part by projects:
French ACI GriPPS (FR-GRID-PPL02-05),
EU-FP6 EGEE-II (INFSO-RI-031688)
EU-FP6 EMBRACE (LHSG-CT-2004-512092)