GPS@: Bioinformatics grid portal for protein sequence analysis on EGEE grid

Blanchet, C., Combet, C., Lefort, V. and Deleage, G.
Pôle BioInformatique de Lyon – PBIL
Institut de Biologie et Chimie des Protéines
IBCP – CNRS UMR 5086
Lyon-Gerland, France
Christophe.Blanchet@ibcp.fr

Information Society and Media

www.eu-egee.org

TOC

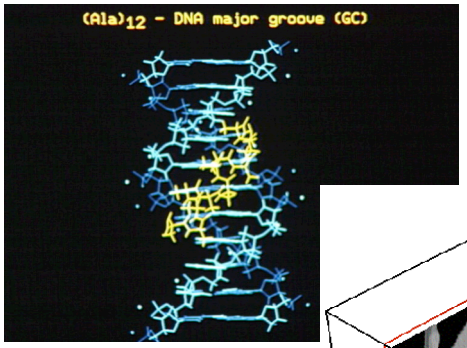
- **NPS@ Web portal**
  - Online since 1998
  - Production mode
- **Gridification of NPS@**
- **Bioinformatics description with XML-based Framework**
- **Legacy mode for application file access**
- **GPS@ Web portal for Bioinformatics on Grid**

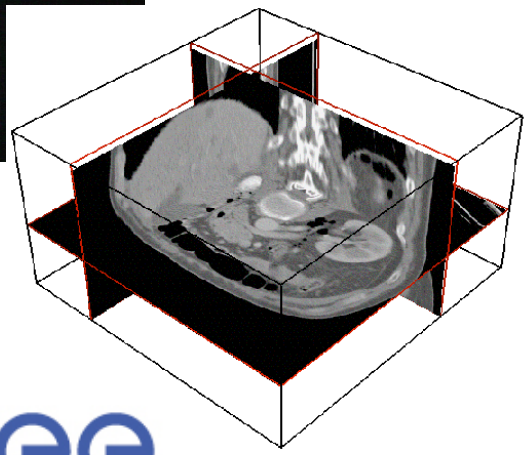
Institute of Biology and Chemistry of Proteins

- **French CNRS Institute, associated to Univ. Lyon1**
  - 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

EGEE Biomedical Activity





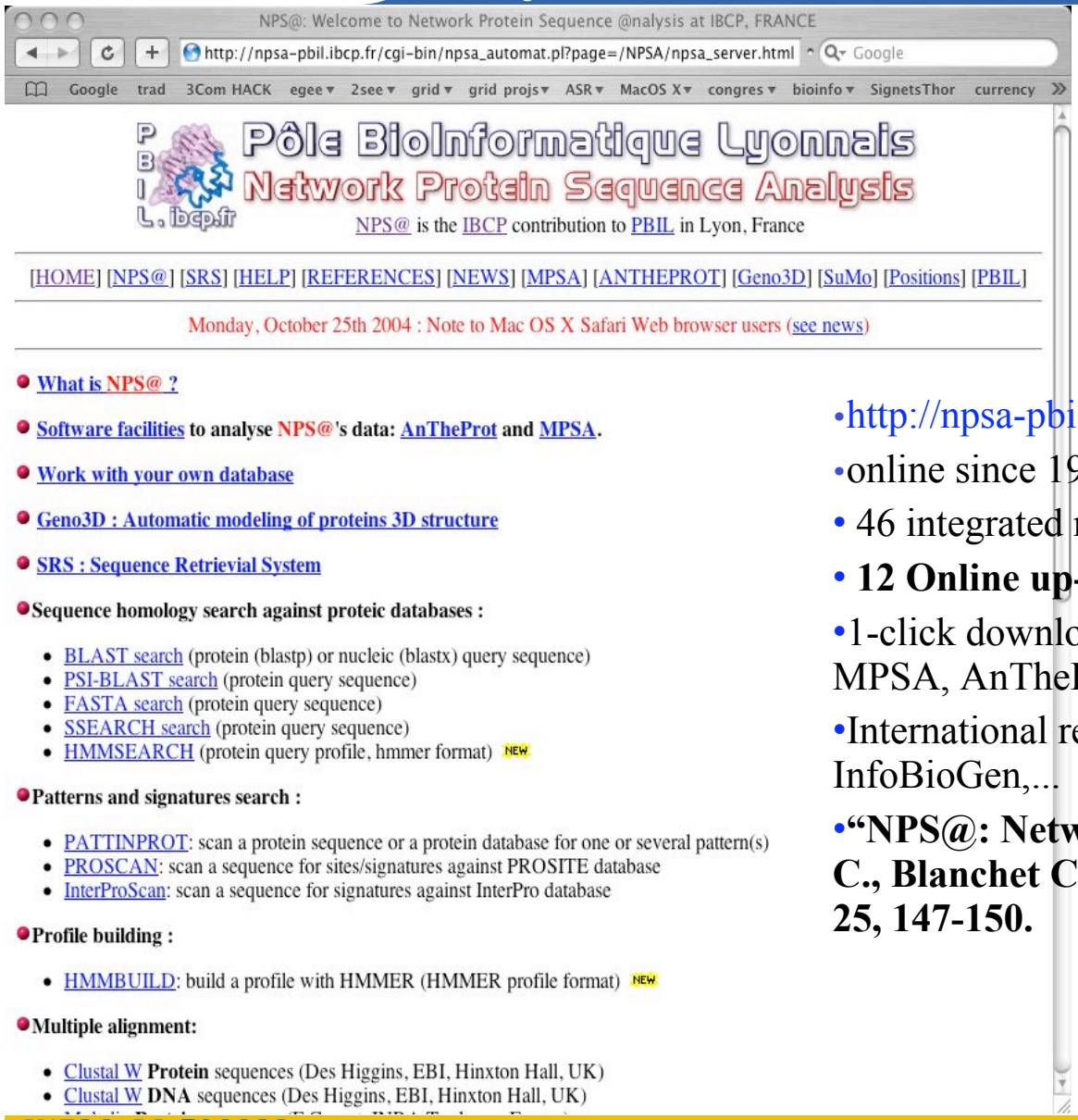
- **Chair:**
  - ♣ **Johan Montagnat**
  - ♣ **Christophe Blanchet (deputy)**
- **Biomedical activity area**
  - Bioinformatics
  - Medical imaging
  - Other health related areas
- **Three types of application**
  - **Pilots**: LCG-2 compliant applications at day 0
  - **Internal**: from project partners, to be deployed on
  - **External**: from other projects, to go through a selection procedure

Enabling Grids for E-sciencE

♣ **EGEE User Forum**, CERN, March 1-3th, 2006
http://egee-intranet.web.cern.ch/egee-intranet/User-Forum

NPS@: bioinformatic Web portal

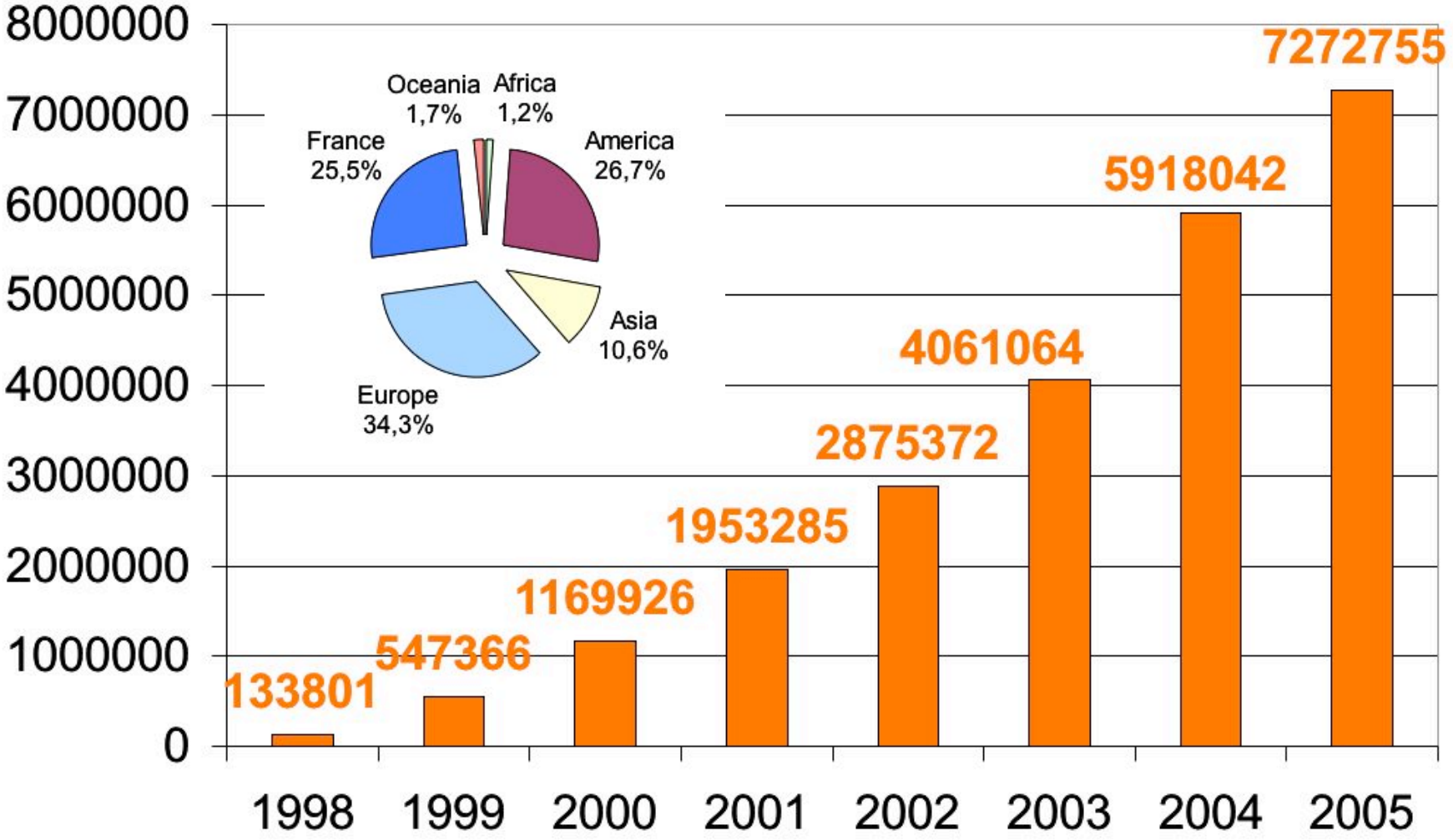
Enabling Grids for E-sciencE



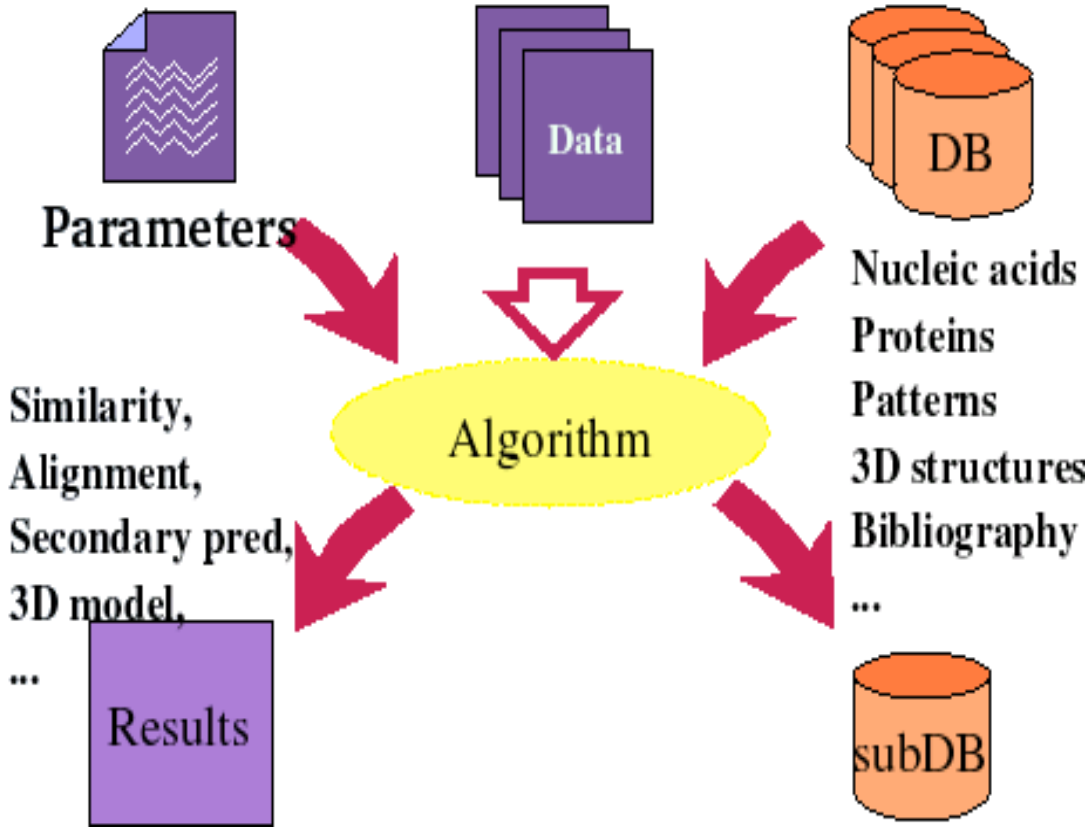
- http://npsa-pbil.ibcp.fr/
- online since 1998 ; NPS@ release 3
- 46 integrated methods for protein sequence analysis
- **12 Online up-to-date biological databanks**
- 1-click download of NPS@ results in biological softwares: MPSA, AnTheProt, Clustal X, RasMol, …
- International references: Expasy, University of California, InfoBioGen,...
- **"NPS@: Network Protein Sequence Analysis", Combet C., Blanchet C., Geourjon C. et Deléage G. *Tibs,* 2000, 25, 147-150.**

NPS@ hits

- More than 7 millions analyses since 1998
- More than 5000 analyses/day



Ex. of Bioinformatics Applications



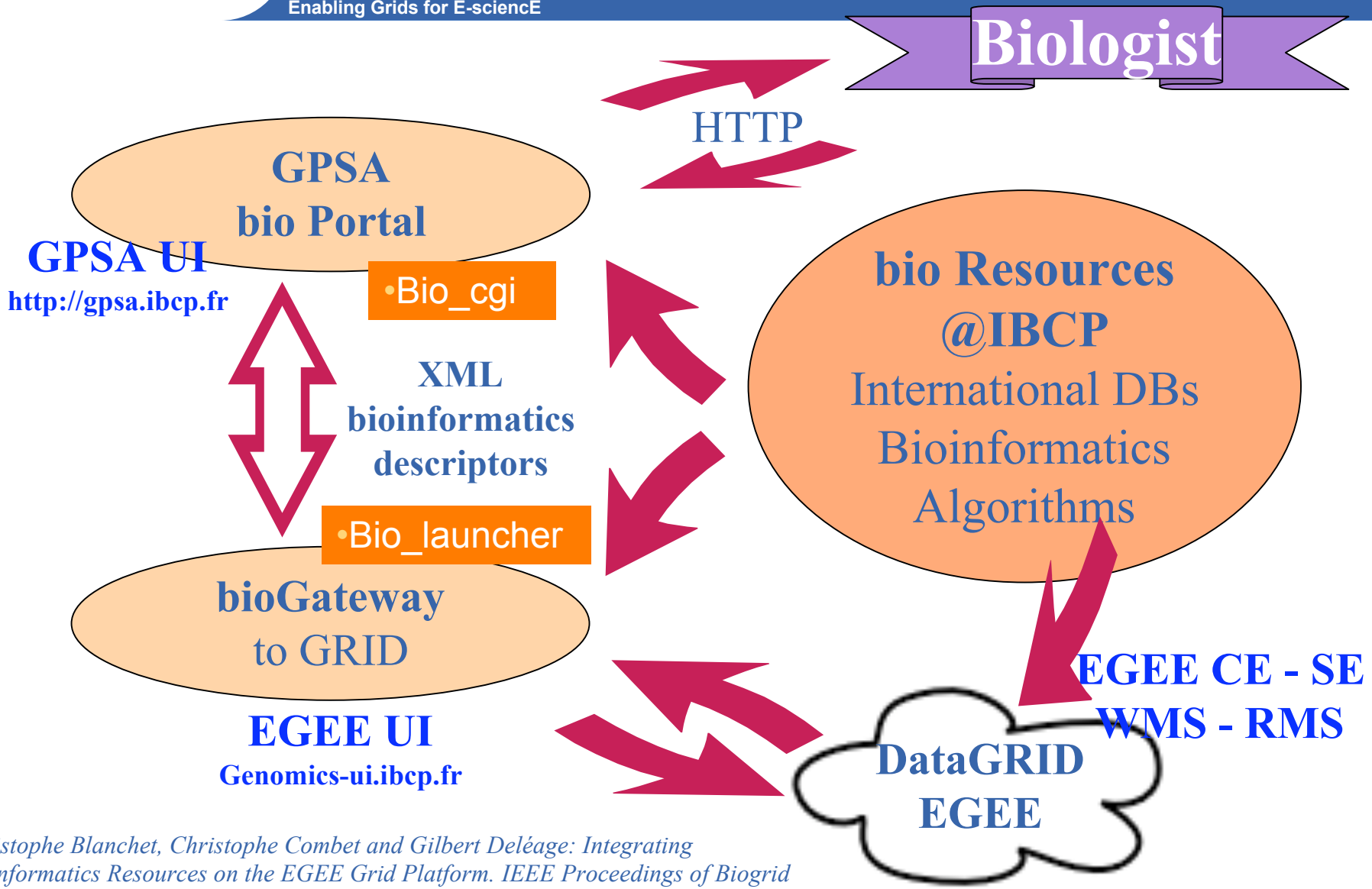
- **Different algorithms**
  - Sequence similarity,
  - Multiple alignment
  - Structural prediction
- **Numerous programs**
  - BioCatalog:
    - + 600 at end of 1990s
  - EMBOSS:
    - + 200 (world-famous)
- ***Data access***
  - *Text files*
  - *I/O standards with local file interface*
- *No modification of source codes to preserve generic model*

Expected benefits of gridification

Enabling Grids for E-sciencE

- **Biological data**
  - ♣ distribute international databases,
  - ♣ store more and large
- **Bioinformatics algorithms**
  - ♣ compute larger datasets
  - ♣ more complex workflows
- **NPS@ Web portal**
  - ♣ well-known Web interface
  - ♣ open to a wider user community.

XML integration into GPS@ portal



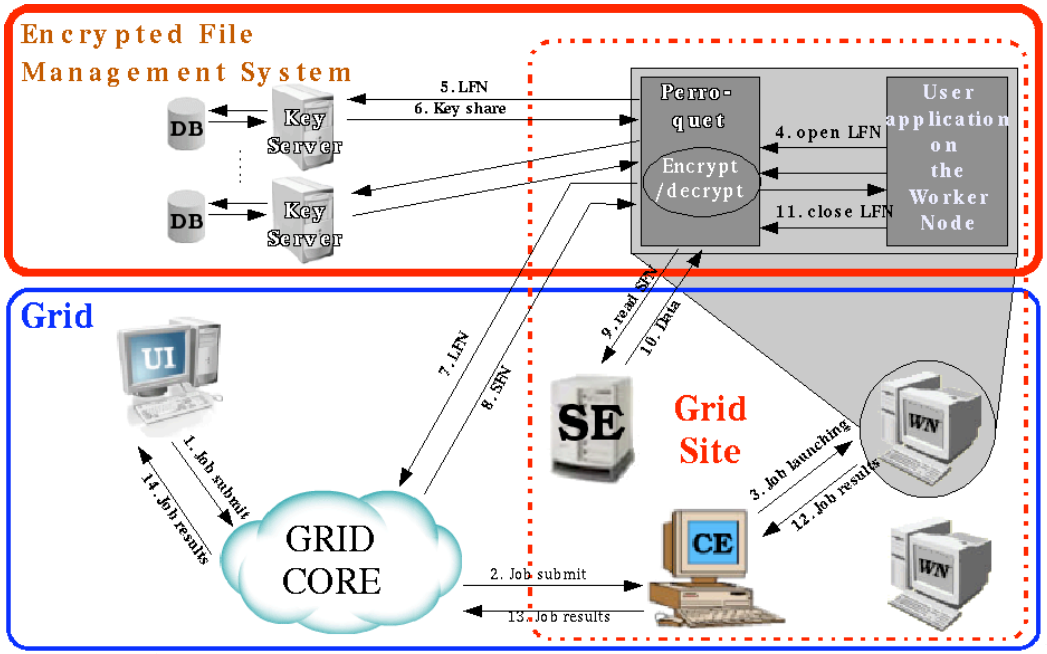
Christophe Blanchet, Christophe Combet and Gilbert Deléage: Integrating Bioinformatics Resources on the EGEE Grid Platform. IEEE Proceedings of Biogrid 2006, Singapore, May 16-19

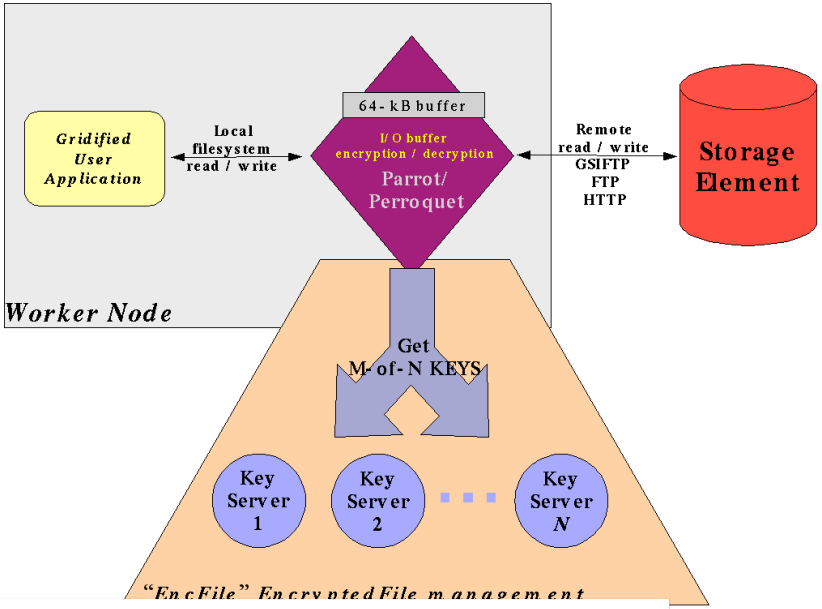
Bioinformatics description with XML-based Framework

```
<?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>
```

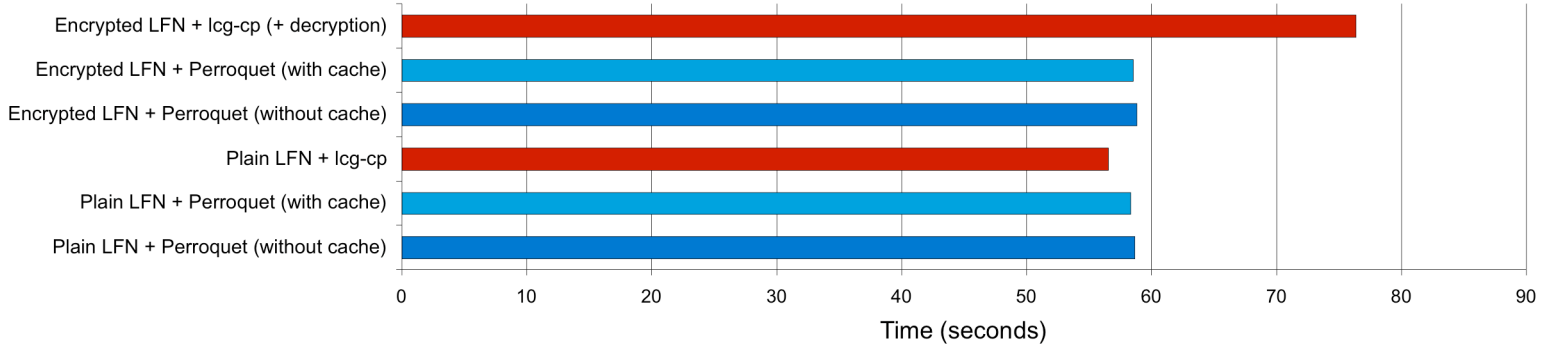
Legacy mode for file access

C. Blanchet, R. Mollon and G. Deleage: Building an Encrypted File System on the EGEE grid: Application to Protein Sequence Analysis. IEEE Proceedings of ARES 2006, Vienna, 20-22 April

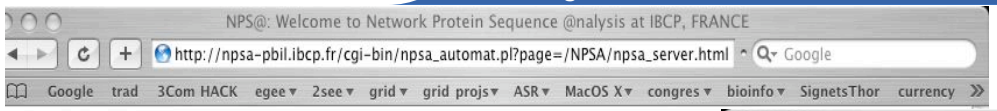




Time to download a 205-MB gridified file



GPS@ : Gridification of NPS@



Pôle Bioinformatique Lyo
Network Protein Sequence A

NPS@ is the IBCP contribution to PBIL in Lyon, Fran

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NE ANTHEPROT] [Geno

Monday, October 25th 2004 : Note to Web browser users

- What is NPS@ ?
- Software facilities to analyse NPS@'s data: AnTheProt and
- Work with your own database
- Geno3D : Automatic modeling of proteins 3D structure
- SRS : Sequence Retrievial System

Sequence homology search against proteic databases :

- BLAST search (protein (blastp) or nucleic (blastx) query sequence)
- PSI-BLAST search (protein query sequence)
- FASTA search (protein query sequence)
- SSEARCH search (protein query sequence)
- HMMSEARCH (protein query profile, hmmer format) NEW

Patterns and signatures search :

- PATTINPROT: scan a protein sequence or a protein database for one or several pattern(s)
- PROSCAN: scan a sequence for sites/signatures against PROSITE database
- InterProScan: scan a sequence for signatures against InterPro database

Profile building :

- HMMBUILD: build a profile with HMMER (HMMER profile format) NEW

Multiple alignment:

- Clustal W **Protein** sequences (Des Higgins, EBI, Hinxton Hall, UK)
- Clustal W **DNA** sequences (Des Higgins, EBI, Hinxton Hall, UK)

Pôle Bioinformatique Lyonnais
GRID Network Protein Sequence Analysis

GPS@ is the **grid port** of NPS@ from PBIL IBCP in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.ibcp.fr
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the NPS@ portal.

Take advantage of the EGEE Grid platform for your bioinformatics analysis on the NPS@ portal. Enabling Grids for E-sciencE

- What is NPS@ ?
- Software facilities to analyse NPS@'s data: AnTheProt and MPSA.
- Work with your own database
- Geno3D : Automatic modeling of proteins 3D structure
- SRS : Sequence Retrievial System

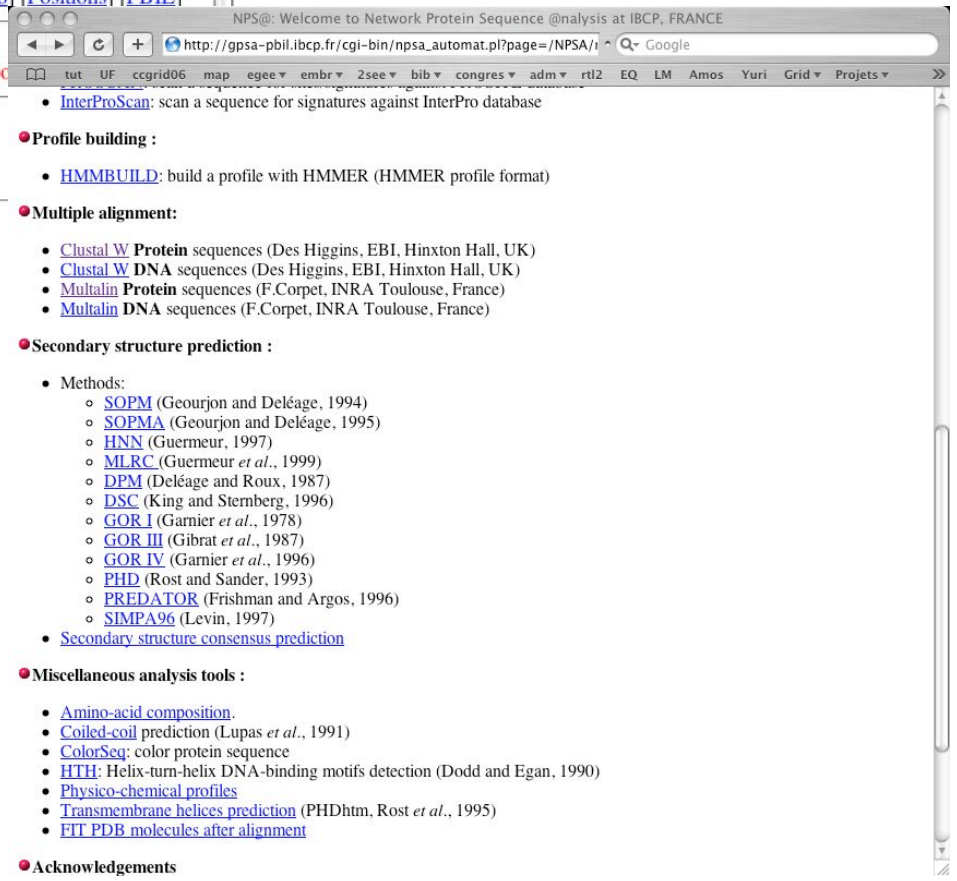
Sequence homology search against proteic databases :

- BLAST search (protein (blastp) or nucleic (blastx) query sequence)
- PSI-BLAST search (protein query sequence)
- FASTA search (protein query sequence)
- SSEARCH search (protein query sequence)
- HMMSEARCH (protein query profile, hmmer format)

Patterns and signatures search :

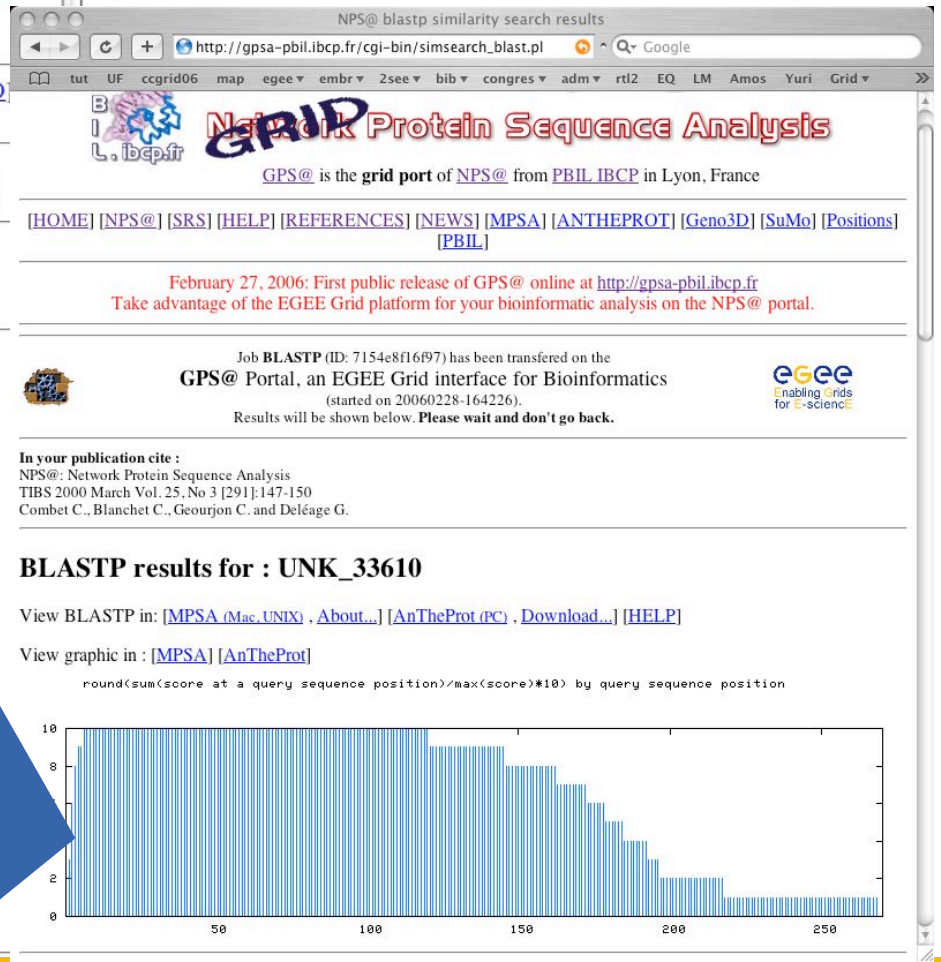
- PATTINPROT: scan a protein sequence or a protein database for one or several pattern(s)
- PROSCAN: scan a sequence for sites/signatures against PROSITE database
- InterProScan: scan a sequence for signatures against InterPro database

GPS@ portal on Grid



BLAST on grid through GPS@

Enabling Grids for E-sciencE



E.g. of gridified bio-resources

Resource	Grid Descriptor
Swiss-Prot	lfn://genomics_gpsa/db/swissprot/swissprot.fasta
And Blast	lfn://genomics_gpsa/db/swissprot/swissprot.fasta.phr
indexes	lfn://genomics_gpsa/db/swissprot/swissprot.fasta.pin
	lfn://genomics_gpsa/db/swissprot/swissprot.fasta.psq
TrEMBL	lfn://genomics_gpsa/db/trembl/trembl.fasta
PROSITE	lfn://genomics_gpsa/db/prosite/prosite.dat
	lfn://genomics_gpsa/db/prosite/prosite.doc
ClustalW	ESM tag "genomics_gpsa_clustalw"
SSearch	ESM tag "genomics_gpsa_ssearch"

- **Examples of biological databases and bioinformatics programs registered and deployed onto the EGEE grid.**
- **Database files have been registered as logical files into the replica manager system, with their own logical filename (LFN, lfn://),**
- **and programs with an tag of the experiment software manager (ESM tag).**

Conclusion and perspectives

- **GPS@ Web portal for Bioinformatics on Grid**
  - Access to grid resources of EGEE (computation and storage)
  - Well-known interface
- **Integration of legacy resources**
  - XML-based
  - Automatic deployment of legacy applications
- **Integration of EncFile tool**
  - Transparent and local file access to remote data
  - On-the-fly encryption/decryption
  - Good performances

- **Perspectives**
  - Short jobs: execution time < 5 minutes
  - EGEE TCG working group on “Short Deadline Job”