



Enabling Grids for E-science

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

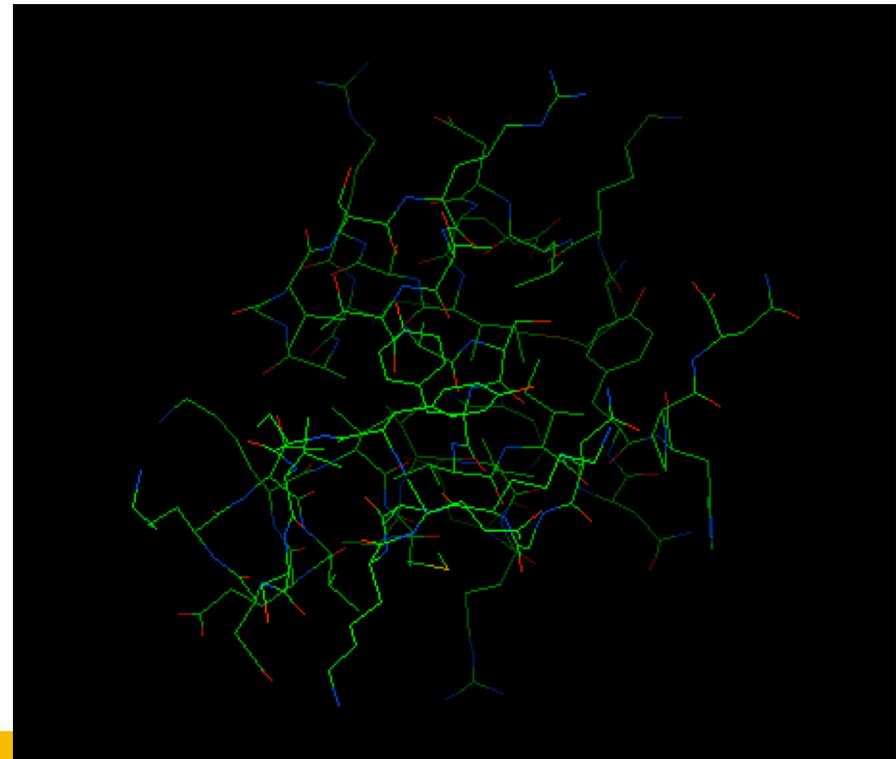
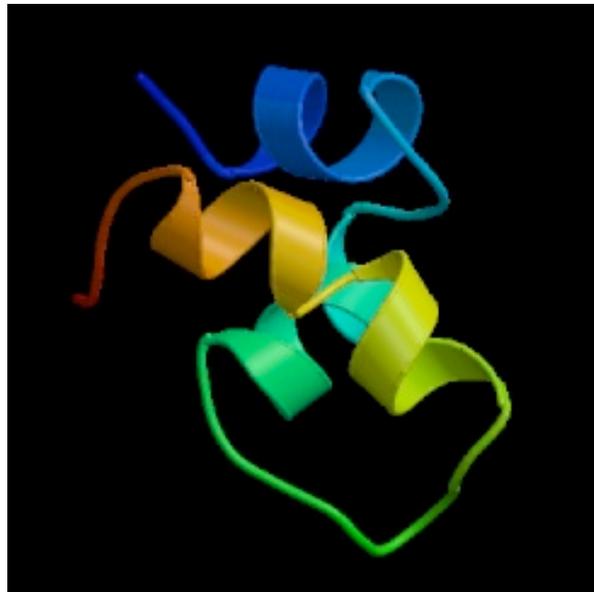


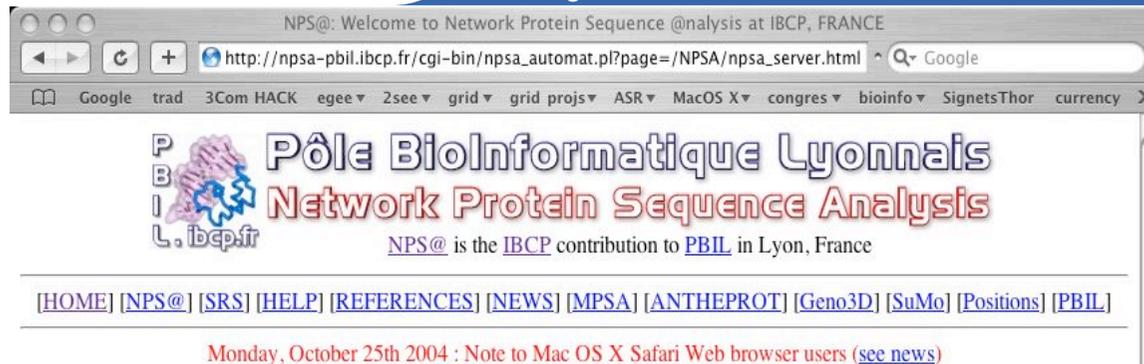
- **French CNRS Institute, associated to Univ. Lyon 1**
 - Life Science
 - About 160 people
 - <http://www.ibcp.fr>
 - Located in Lyon, France
- 
- **Study of proteins in their biological context**
 - Approaches used include : integrative cellular (cell culture, various types of microscopies) and molecular 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

- **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
 MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKVMVAVVREH
 NYHPNAVAAGLRLQHSHHHH



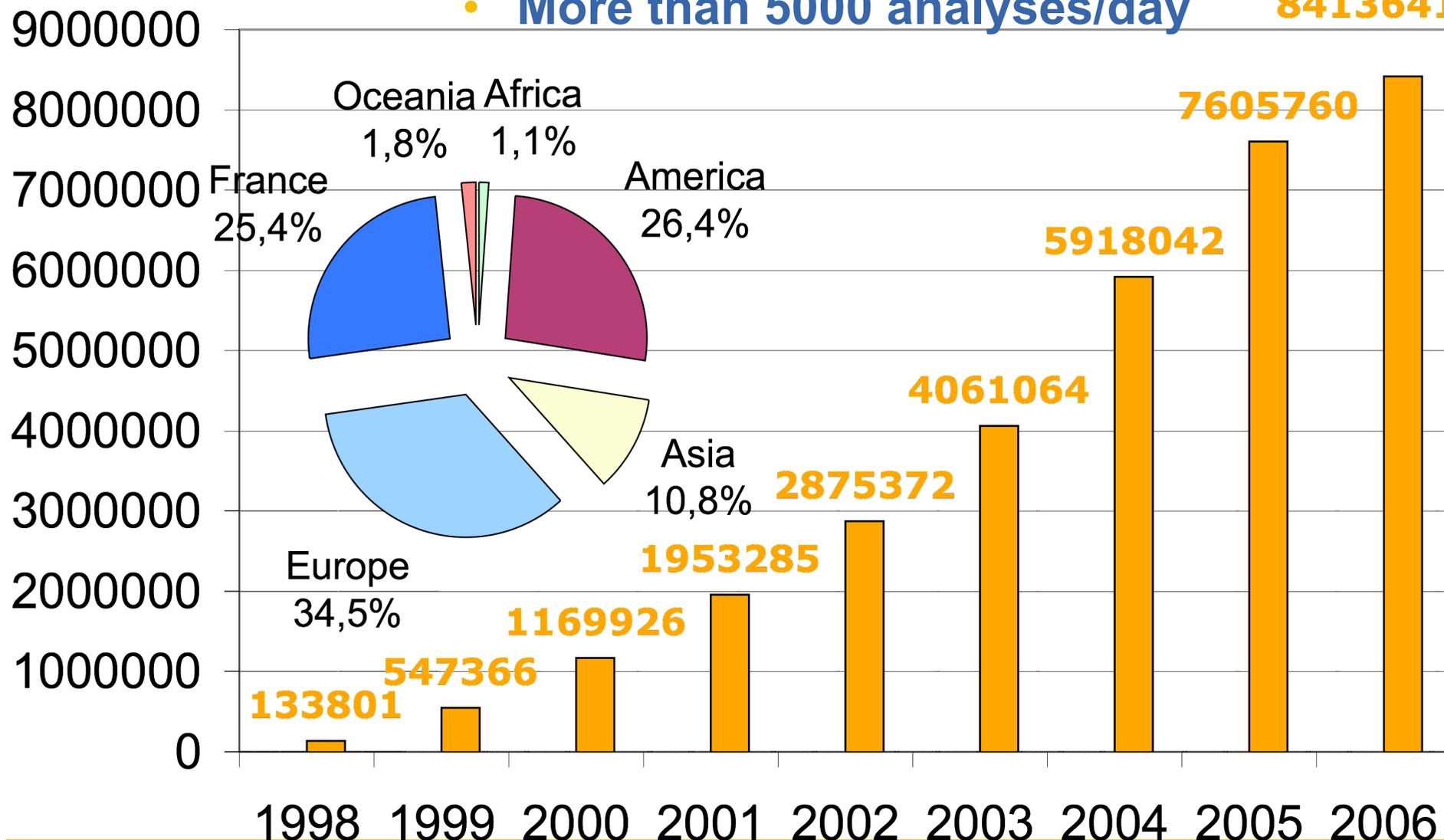


- [What is NPS@ ?](#)
- [Software facilities](#) to analyse NPS@'s data: [AnTheProt](#) and [MPSA](#).
- [Work with your own database](#)
- [Geno3D : Automatic modeling of proteins 3D structure](#)
- [SRS : Sequence Retrieval System](#)
- **Sequence homology search against proteic databases :**
 - [BLAST 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 p
 - [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)

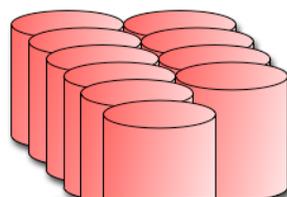
- **Network Protein Sequence Analysis (NPS@ release 3)**
<http://npsa-pbil.ibcp.fr>
- **Online since 1998**
- **46 integrated methods for protein sequence analysis**
- **12 Online up-to-date biological databanks**
- **International pointers: Expasy (Ch) , University of California, ...**
- **Ref.: “ NPS@: Network Protein Sequence Analysis”, Combet C., Blanchet C., Geourjon C. et Deléage G. Tibs, 2000, 25, 147-150.**

- More than 8 millions analyses since 1998

- More than 5000 analyses/day **8413641**



- **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)**



Databases



Software

- **Numerous**
 - BioCatalog: + 600 at end of 1990s
 - EMBOSS toolkit: + 200
- **Heterogeneous**
 - Bioinformatics algorithm: Sequence similarity, Multiple alignment, Structural prediction, ...
 - Execution: sequential, MPI, openMP, ...

SIZE OF SOME BIOLOGICAL DATABASES

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

Data I/O

- *Text files*
- *Specific format*
- *Local I/O only*

Describe to gridify!

Description of an application

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

XML-based:

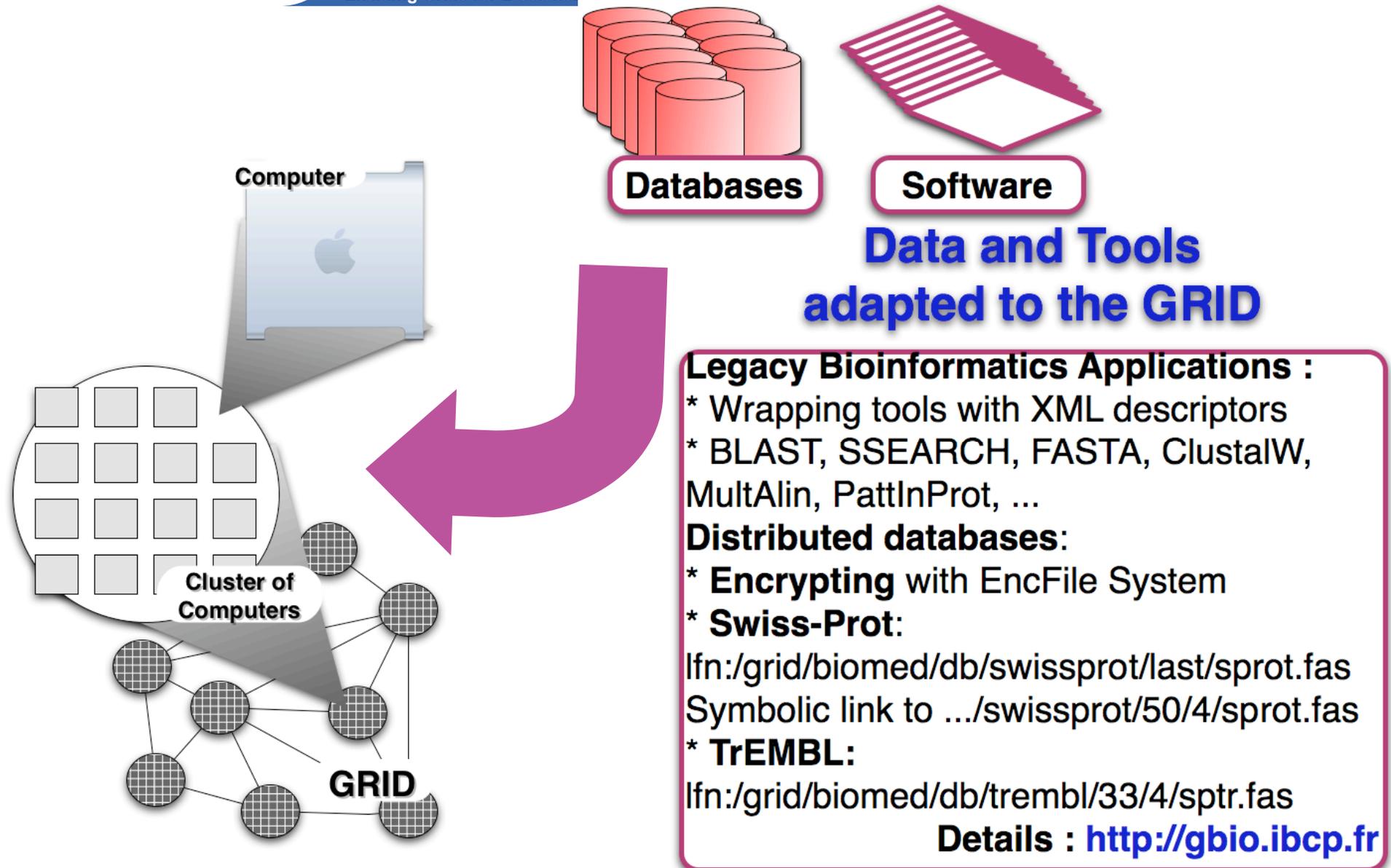
- Bio_methods.dtd
- Blast.xml, pattinprot.xml, clustalw.xml,
- ...

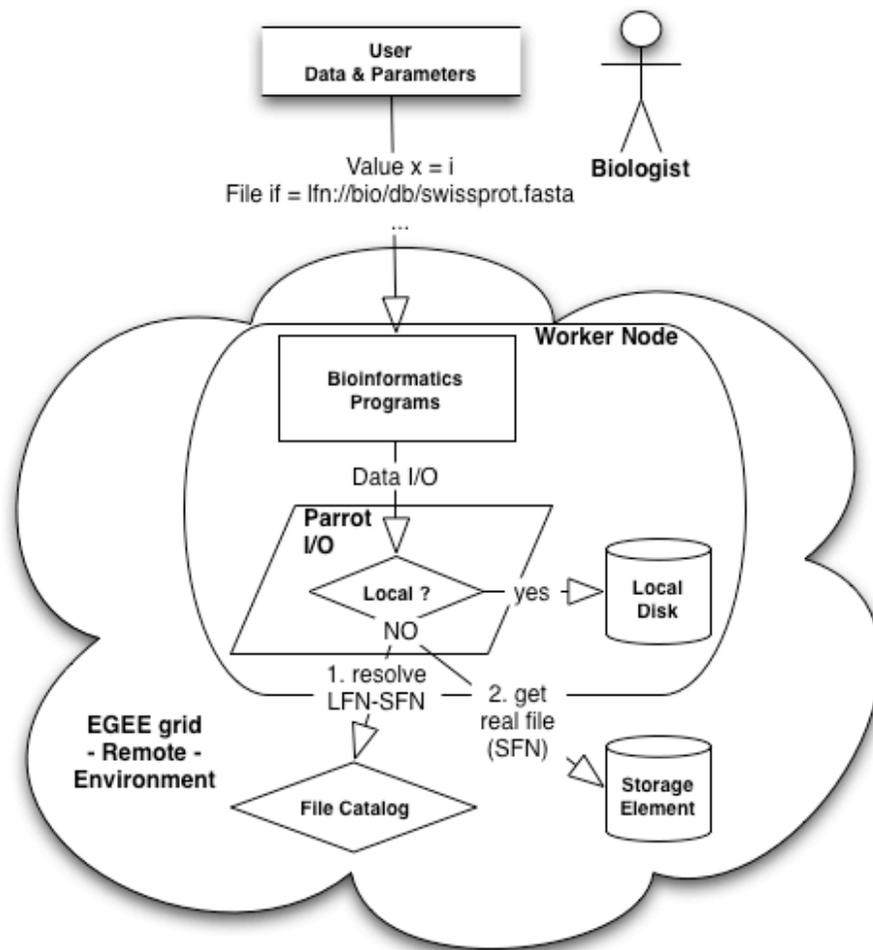
Use application descriptors

- 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>
```





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, ...

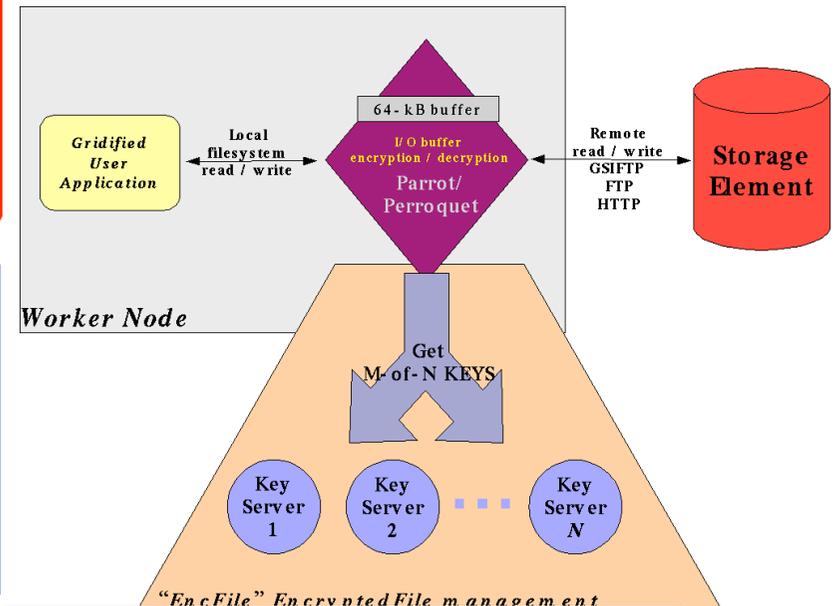
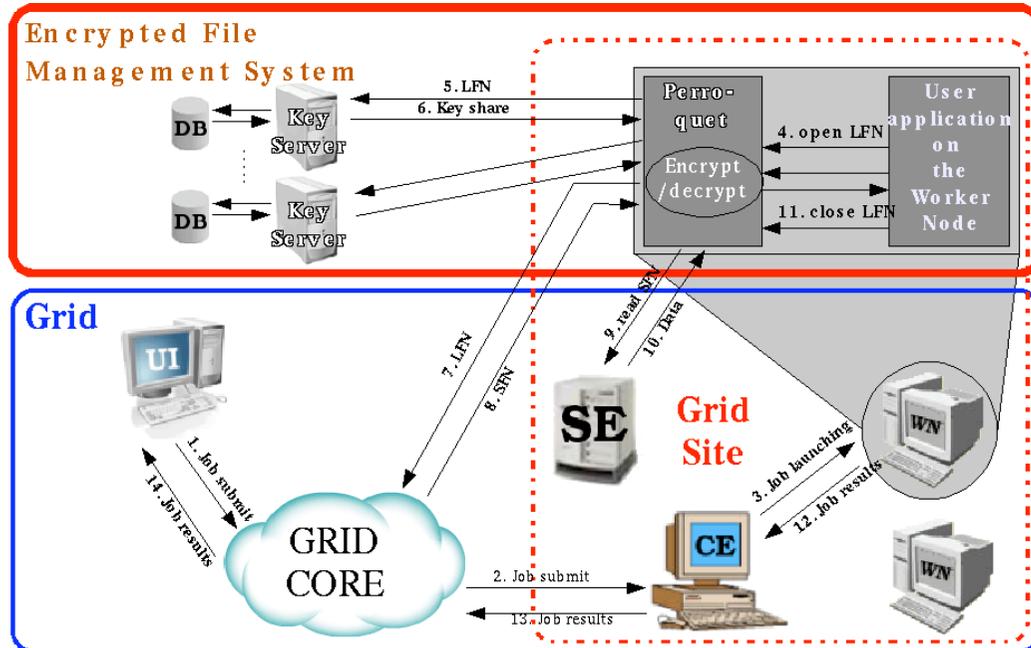
Conclusion

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

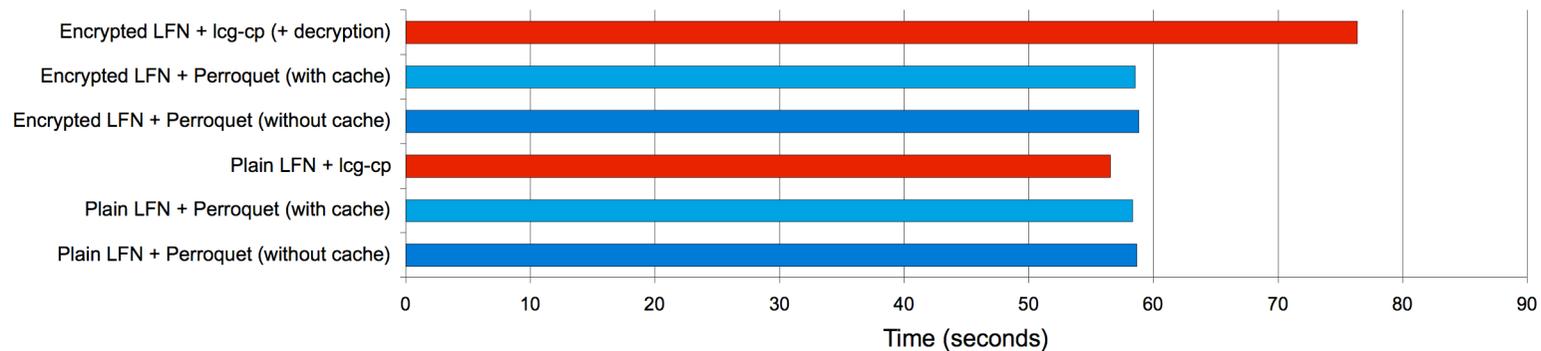
Christophe Blanchet, Rémi Mollon, Douglas L. Thain and Gilbert Deléage
 Grid Deployment of Legacy Bioinformatics Applications with Transparent Data Access
 To be presented at Conference GRID'2006, Barcelona, Sept. 28-29, 2006

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



NPS@ blastp similarity search results

http://gpsa-pbil.ibcp.fr/cgi-bin/simsearch_blast.pl

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.

Job BLASTP (ID: 7154e8f16f97) has been transferred on the GPS@ Portal, an EGEE Grid interface for Bioinformatics (started on 20060228-164226).
Results will be shown below. Please wait and don't go back.

EGEE
Enabling Grids for E-science

In your publication cite :
NPS@: Network Protein Sequence Analysis
TIBS 2000 March Vol. 25, No 3 [291]:147-150
Combet C., Blanchet C., Geourjon C. and Deléage G.

Computation Virtualization:

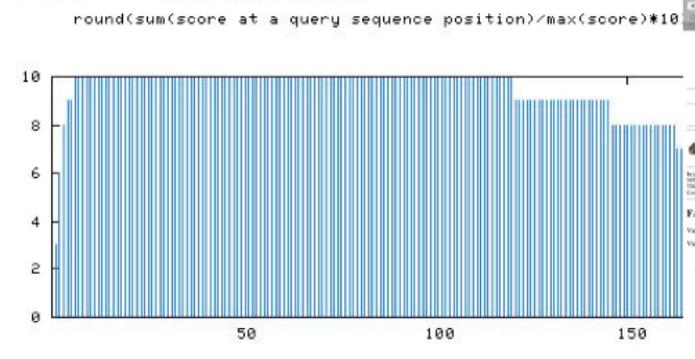
Web portal GPS@

BLAST on GRID
But also ...

BLASTP results for : UNK_33610

View BLASTP in: [MPSA (Mac, UNIX)] , [About...] [AnTheProt]

View graphic in : [MPSA] [AnTheProt]



FASTA results for : UNK_84170

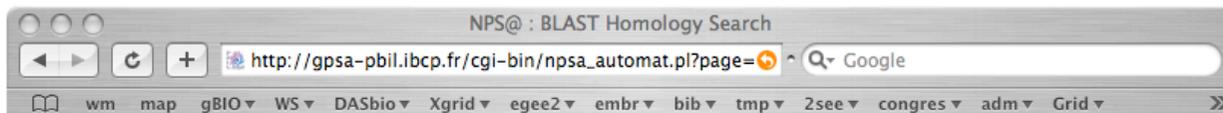
SEARCH results for : UNK_66550

CLUSTALW

Pair a protein sequence database in ProteinFasta format below :

Output: fasta ->

SSearch, Fasta, ClustalW
And others bioinformatics tools we have ported on GRID ...



Pôle BioInformatique Lyonnais

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.

Work supported in part by projects: French ACI Grid GriPPS, EU-FP6 EGEE and EU-FP6 EMBRACE.



BLAST search on protein sequence databank

[\[Abstract\]](#) [\[NPS@ help\]](#) [\[Original server\]](#)

Program:

Database:

Sequence name (optional):

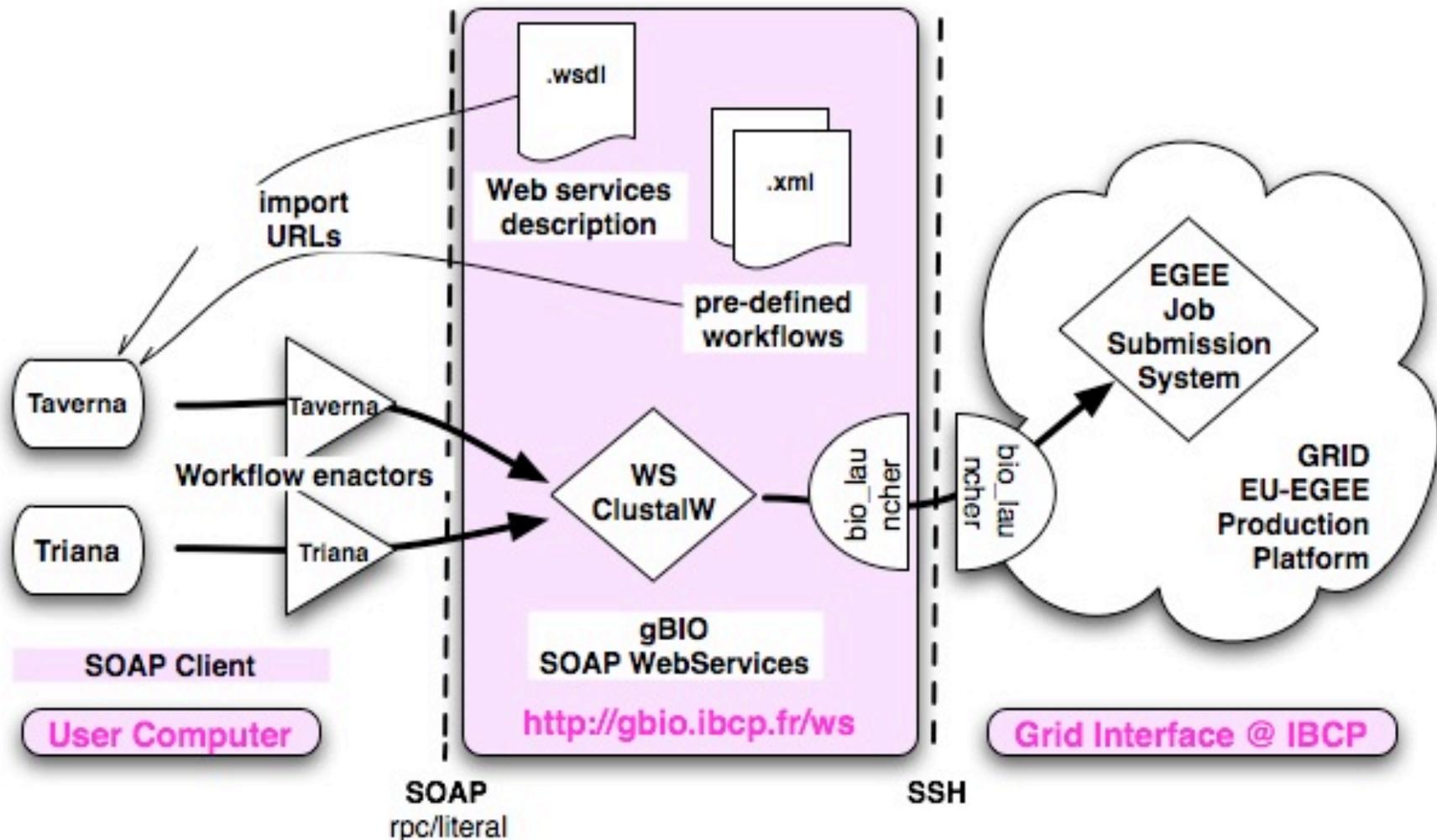
Paste a protein/nucleic sequence below : [help](#)

```
MKKITTYDLAELSGVSAVSAILNGNWKKRRISAKLAEKVTRIAEEQGYAINRQASMLR
SKKSHVIGMIIPKYDNRYFGSIAERFEEMARERGLLPITCTRRRPELEIEAVKAMLSWQ
VDWVVATGATNPKISALCQAGVPTVNLDPGLSPLSDNYGGAKALTHKILANSAR
RRGELAPLTFIGRRRATITPASVYAASTMRIASWGLACRRRIFWLPAIRKATLRTACRS
LAARRRCCRCGYLLTRRYPWKGLCAGCRRWV
```

Use the GRID resources from 

User : public@193.55.43.12. Last modification time : Fri Jan 20 10:11:15 2006. Current time : Fri Sep 22 14:29:02 2006
This service is supported by [Ministere de la recherche](#) (ACC-SV13), [CNRS](#) (IMABIO, COMI, GENOME) and [Région Rhône-Alpes](#) (Programme EMERGENCE) . [Comments](#).

Grid Protein Sequence Analysis (GPS@)



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





Advanced model explorer

Workflow Metadata for 'gBIOclustalwGrid'

Load Save New subworkflow Offline Reset

Workflow object	Retrie	Delay	Backof	Thread	Critica
Workflow model					
Workflow inputs					
sequences					
Workflow outputs					
alignment					
Processors					
gBIOclustalwGrid	0	0	1	1	
sequence-bank 'text/plain'					
attachmentList l("")					
result 'text/plain'					
Data links					
sequences-gBIOclustalwGrid:sequence-bank					
gBIOclustalwGrid:result-alignment					
Control links					

Available services

Search Watch loads

- Available Processors
 - Local Services
 - Biomart service @ http://www.biomart.org/
 - WSDL @ http://gbio.ibcp.fr/ws/gBIO.wsdl
 - porttype: gBioWSPortType [RPC]
 - gBIOclustalw
 - gBIOclustalwGrid

Enactor invocation

Save as XML Save to disk Save to disk as website Excel

Status Results Process report

alignment

```

CLUSTAL W (1.83) multiple sequence alignment

CCPA_STRM5      MNTDDTITIYDVAREAGVSMATVSRVVGNGK---NVKENTRKKVLEVIDRLD
DEGA_BACSU      ----MKTTIYDVAKAAGVSIITTVSRVINNTG---RISDKTRQKVMNVMNEMA
FRUR_ECOLI      -----MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKVMVAVVREHN
RBTR_KLEAE      ---MKKTIYDLAELSGVSAVSAVAILNGNWKRRRISAKLAEKVTRIAEEQQ
                :   . : . : . : * * * * : * : . : . : . : * * * * : .

CCPA_STRM5      GLASKKTTTVGVVIPNIANAYFSILAKGIDDIATMYKYNIVLASSDEDDDDKE
DEGA_BACSU      ALTCRRTNMIALVAPDISNPFYGELAKSIEERADELGFQMLICSTDYDPKKE
FRUR_ECOLI      FRURAGRTRSIGLVPDLENTSYTRIANYLERQARQRYQLLIACSEDQPDNE
RBTR_KLEAE      MLRSKSHVIGMIIPKYDNRYPGSAERFEEMARERGLLPIITCTRRRPELE
                * . : . : . : * . : . : * : . : . : * : . : . : *

CCPA_STRM5      AKQVDGIIIFMG---HHLTEKIRAEF SRARTPVVLSGTVDLEHQLPVSNIDHS
DEGA_BACSU      QKKVDGIIIFATGIESHDSMSALEE IASEQIPIAMISQDKPLLPMDIIVDDV
FRUR_ECOLI      QRQVDAIIVSTS---LPPEHPFYQRWANDPFPIVALDRALDREHFTSVVGADQ
RBTR_KLEAE      SWQVDWVATG---ATNPDKISALCQQAGVPTVNLDLPG---SLSPSVISDN
                : * * : . : . : . : * . : . : * : . : . : * : . : . : *
    
```

Workflow diagram

Save as Configure diagram

```

graph TD
    subgraph Workflow_Inputs
        sequences[sequences]
    end
    sequences --> gBIOclustalwGrid[gBIOclustalwGrid]
    gBIOclustalwGrid --> subgraph Workflow_Outputs
        alignment[alignment]
    end
    
```

Workflow diagram

Save as Configure diagram

Rendering done.

The screenshot displays the Triana workflow editor interface. The main workspace shows a workflow with three nodes: 'Protein Sequences' (purple), 'gBIOclustalwGrid' (red), and 'Multiple Alignment' (purple), connected by a black line. A 'Protein Sequences' dialog box is open, showing a list of protein sequences. A 'StringViewer' dialog box is also open, displaying the output of the 'gBIOclustalwGrid' node, which is a CLUSTAL W multiple sequence alignment. The alignment shows several protein sequences aligned together, with gaps represented by dashes. The sequences include CCBA_BACSU, CCBA_BACME, CCBA_STRMU, DEGA_BACSU, FRUR_ECOLI, RBTR_KLEAE, and CCBA_BACME.



- **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



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)

