



Enabling Grids for E-science

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

www.eu-egee.org



Information Society
and Media



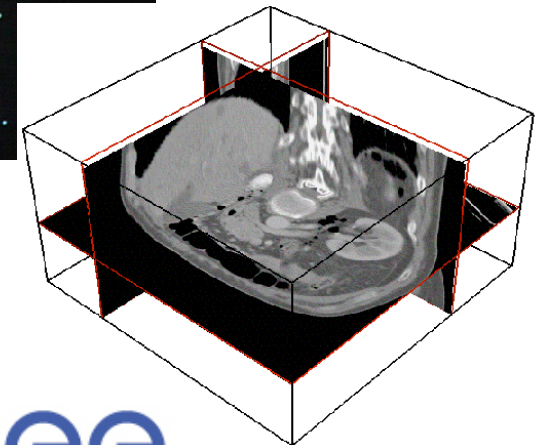
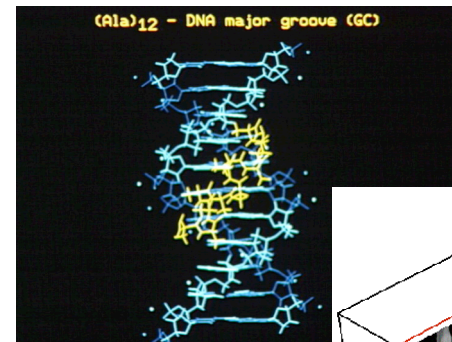
- **NPS@ Web portal**
 - Online since 1998
 - Production mode
- **Gridification of NPS@**
- **Bioinformatics description with XML-based Framework**
- **Legacy mode for application file access**
- **GPS@ Web portal for Bioinformatics on Grid**

- **French CNRS Institute, associated to Univ. Lyon1**
 - Life Science
 - About 160 people
 - <http://www.ibcp.fr>
 - Located in Lyon, France



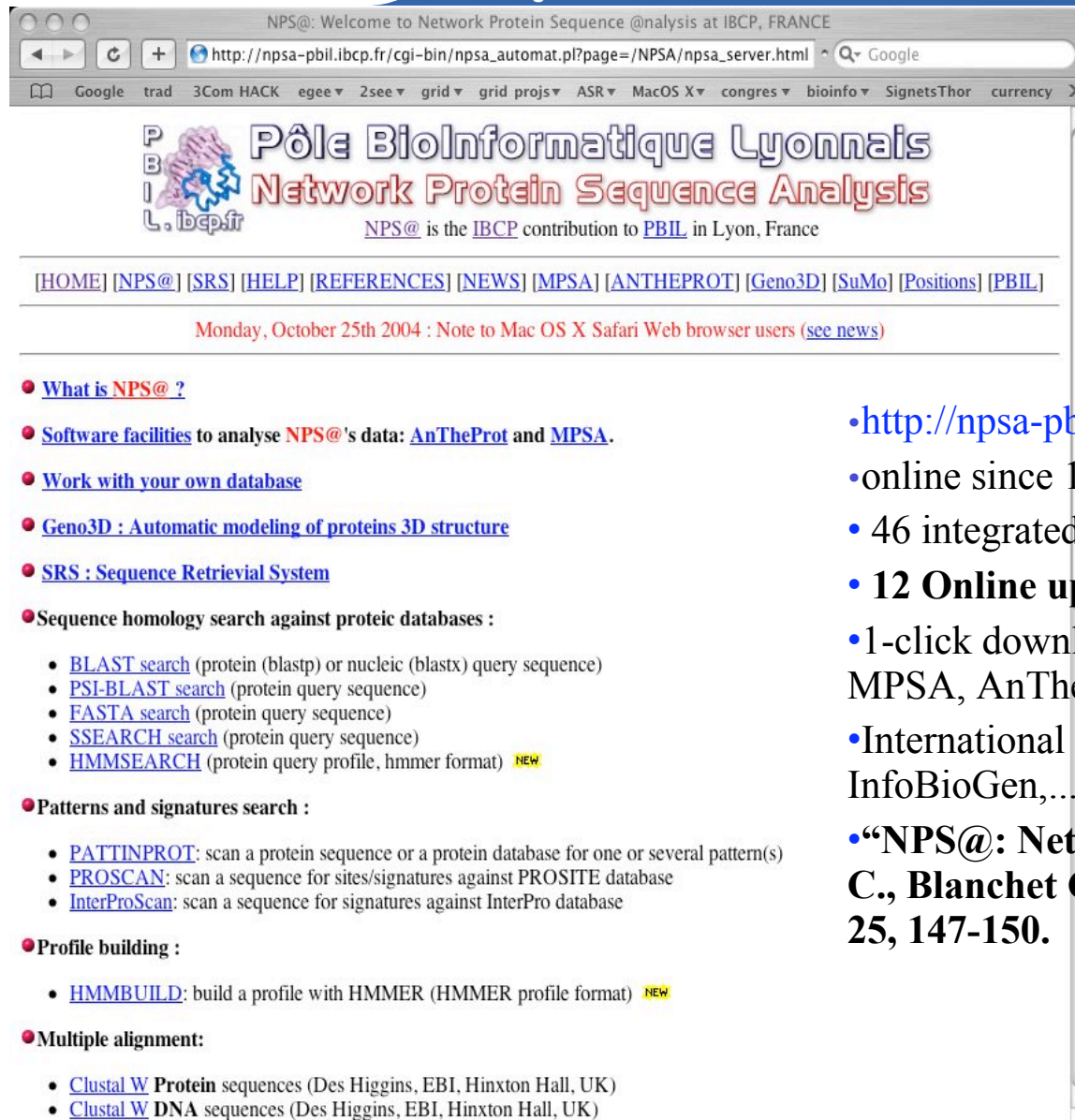
- **Study of proteins in their biological context**
 - ♣ Approaches used include integrative cellular (cell culture, various types of microscopies) and molecular techniques, both experimental (including biocrystallography and nuclear magnetic resonance) and theoretical (structural bioinformatics).
-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

- **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 for E-science
 - **External:** from other projects, to go through a selection procedure

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



NPS@: Welcome to Network Protein Sequence @analysis at IBCP, FRANCE

http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html

Google trad 3Com HACK egee 2see grid grid projs ASR MacOS X congres bioinfo SignetsThor currency

P
B
I
L.ibcp.fr

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis

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

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

Monday, October 25th 2004 : Note to Mac OS X Safari Web browser users ([see news](#))

- [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 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)

• <http://npsa-pbil.ibcp.fr/>

• online since 1998 ; NPS@ release 3

• 46 integrated methods for protein sequence analysis

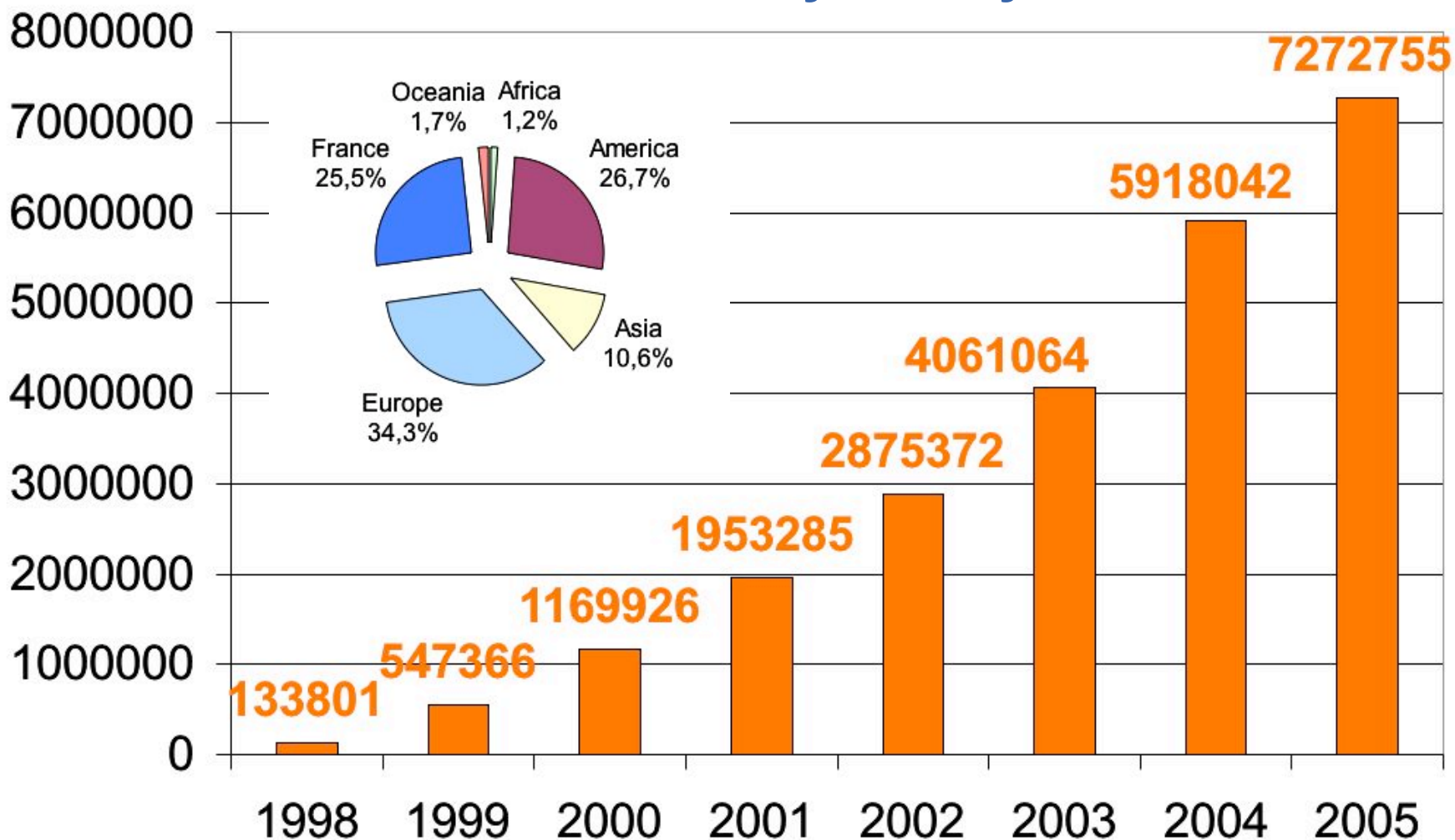
• **12 Online up-to-date biological databanks**

• 1-click download of NPS@ results in biological softwares: MPSA, AnTheProt, Clustal X, RasMol, ...

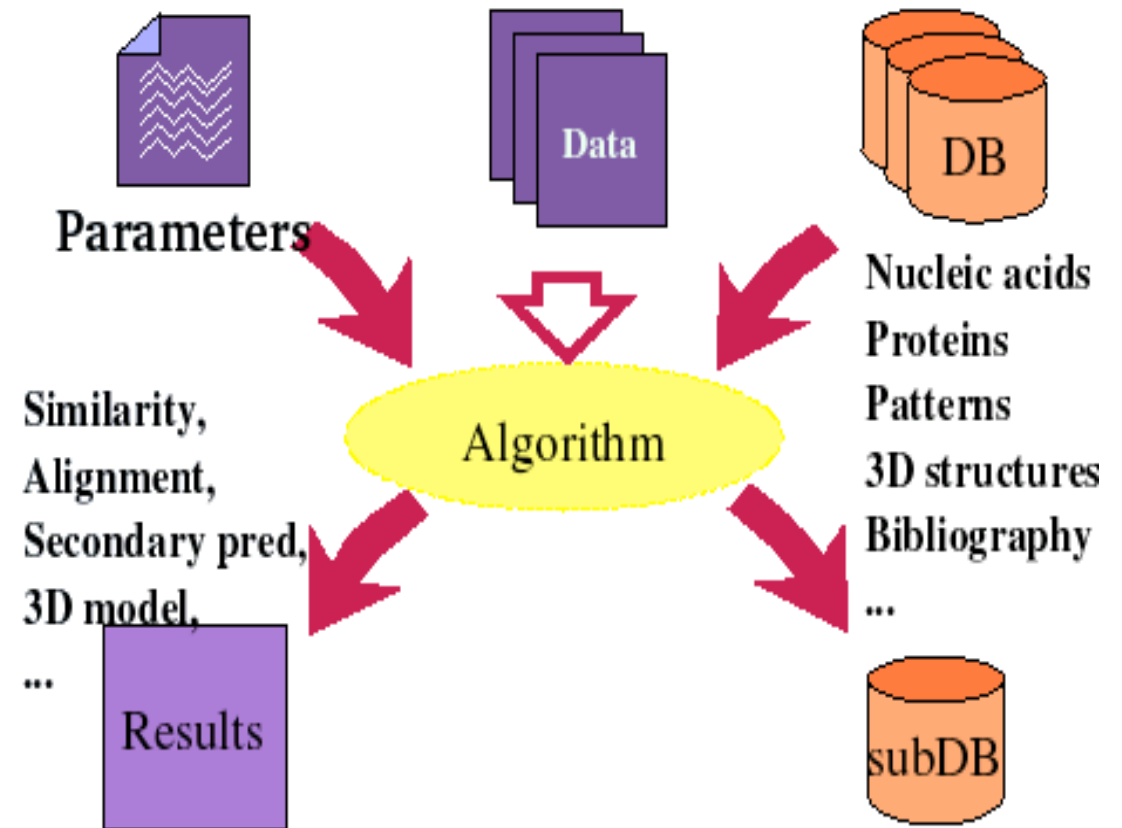
• International references: Expasy, University of California, InfoBioGen,...

• **“NPS@: Network Protein Sequence Analysis”, Combet C., Blanchet C., Geourjon C. et Deléage G. *Tibs*, 2000, 25, 147-150.**

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



- **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*





NPS@: Welcome to Network Protein Sequence @nalysis at IBCP, FRANCE

http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis
NPS@ is the IBCP contribution to PBIL in Lyon, France

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

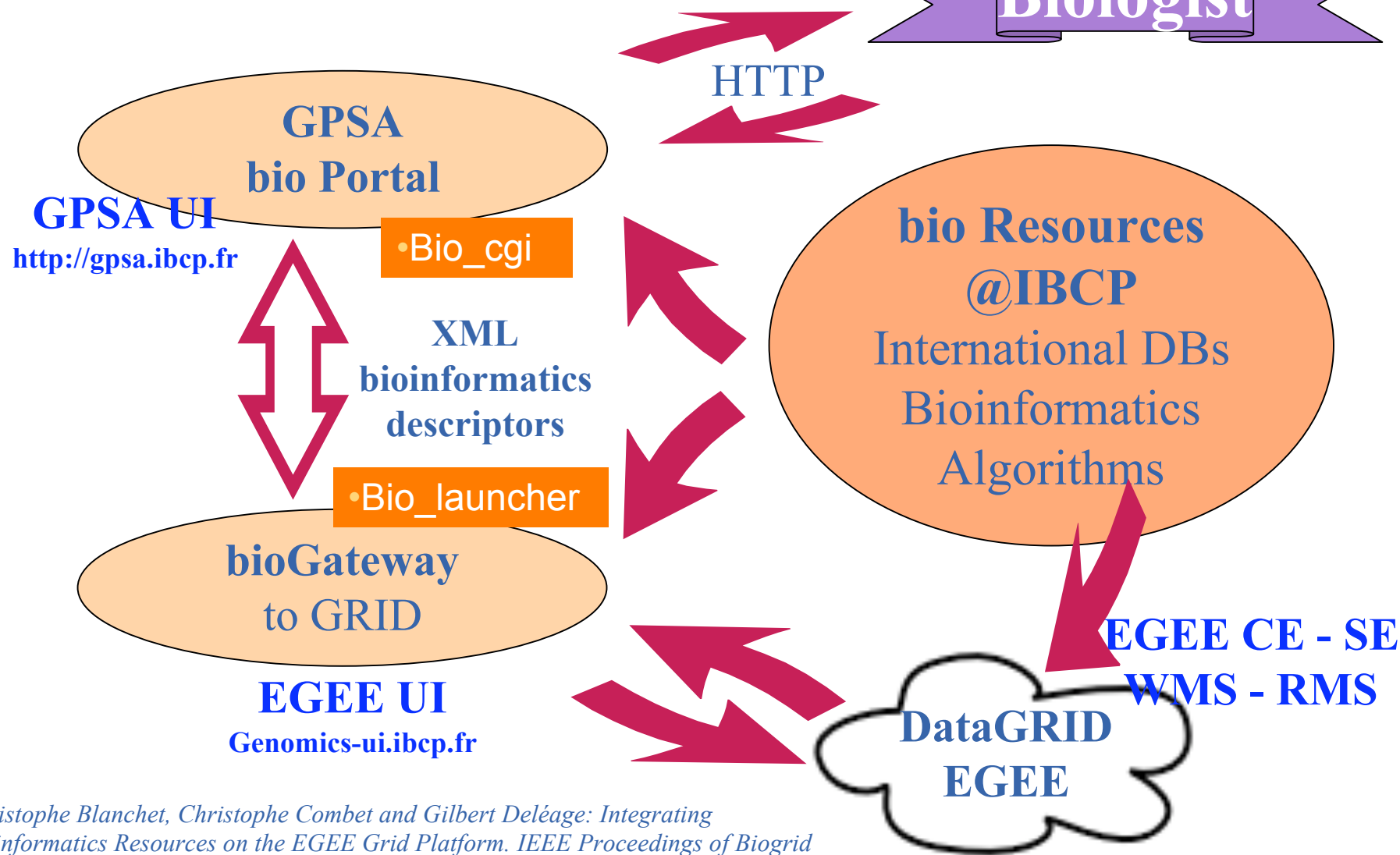
Monday, October 25th 2004 : Note to Mac OS X Safari Web browser users ([see news](#))

- What is NPS@ ?
- Software facilities to analyse NPS@'s data: [AnThePr](#) | [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 more patterns
 - [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)

- **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.



Biologist



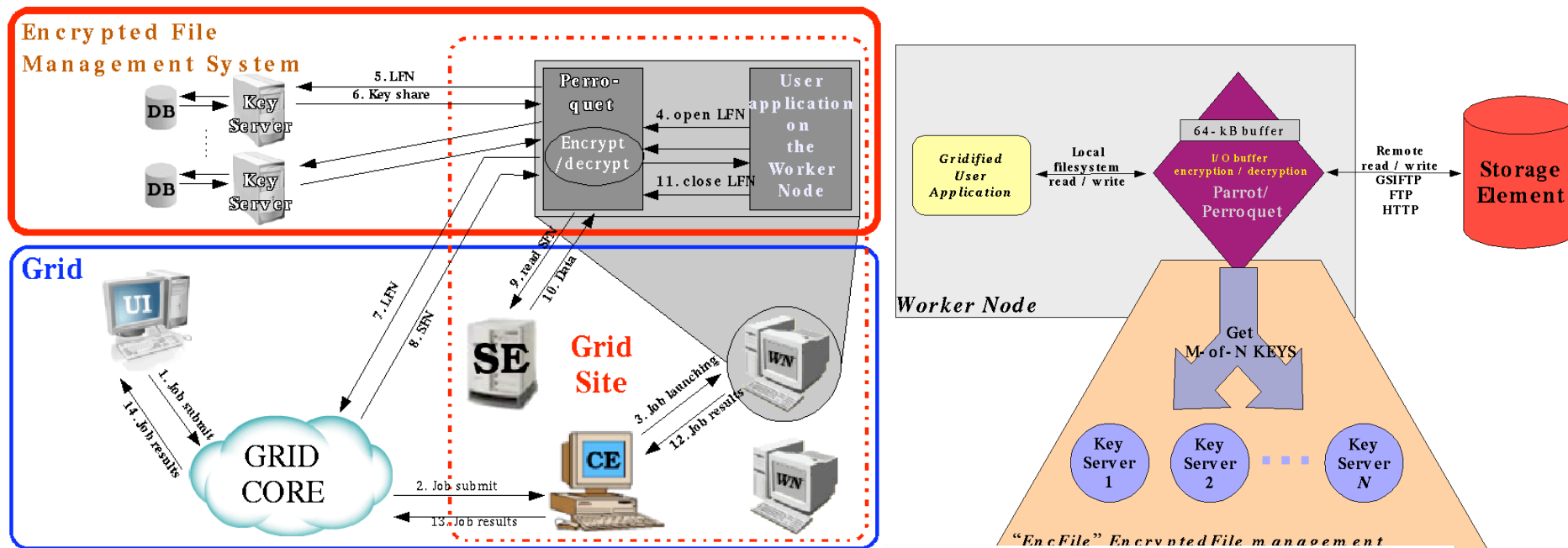
Christophe Blanchet, Christophe Combet and Gilbert Deléage: Integrating Bioinformatics Resources on the EGEE Grid Platform. IEEE Proceedings of Biogrid 2006, Singapore, May 16-19

```

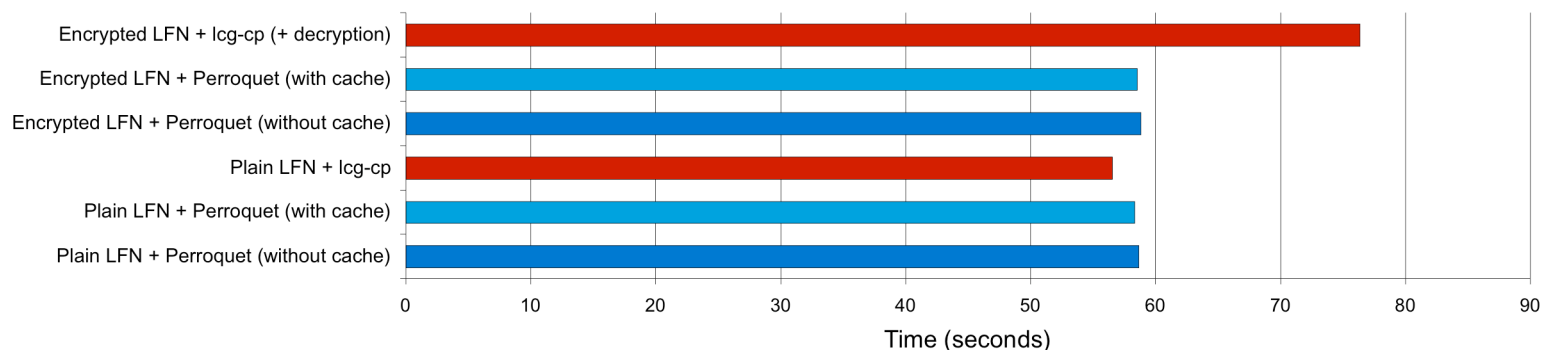
<?xml version="1.0"?>
<!DOCTYPE bio_method SYSTEM "/opt/bio/etc/bio_method.dtd">
<bio_method version="2.0" mode="egEE" >
  <method name="PATTINPROT" class="scanprot" type="sequential"
  root="/var/www/gpsa.ibcp.fr/pbil/servers/gpsa/w3-gpsa/" >
    <bio_binary path="gbio_lfn://PATTINPROT/newpattinprot" 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>

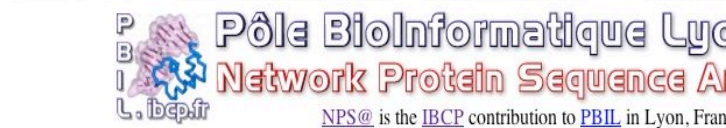
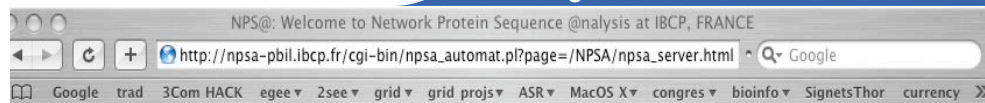
```

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





[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [ANTHEPROT] [Geno3D]

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


- ▶ **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 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)



[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.

- **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)
- **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

NPS@: Welcome to Network Protein Sequence @analysis at IBCP, FRANCE

http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/1

tut UF ccgrid06 map egee embr 2see bib congres adm rtl2 EQ LM Amos Yuri Grid Projets



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.

Take advantage of the EGEE Grid platform
 for your bioinformatics analysis on the NPS@ portal.



• [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)

• **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

NPS@: Welcome to Network Protein Sequence @analysis at IBCP, FRANCE

http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/1

tut UF ccgrid06 map egee embr 2see bib congres adm rtl2 EQ LM Amos Yuri Grid Projets

- [InterProScan](#): scan a sequence for signatures against InterPro database
- **Profile building :**
 - [HMMBUILD](#): build a profile with HMMER (HMMER profile format)
- **Multiple alignment:**
 - [Clustal W Protein](#) sequences (Des Higgins, EBI, Hinxton Hall, UK)
 - [Clustal W DNA](#) sequences (Des Higgins, EBI, Hinxton Hall, UK)
 - [Multalin Protein](#) sequences (F.Corpet, INRA Toulouse, France)
 - [Multalin DNA](#) sequences (F.Corpet, INRA Toulouse, France)
- **Secondary structure prediction :**
 - Methods:
 - [SOPM](#) (Geourjon and Deléage, 1994)
 - [SOPMA](#) (Geourjon and Deléage, 1995)
 - [HNN](#) (Guermeur, 1997)
 - [MLRC](#) (Guermeur *et al.*, 1999)
 - [DPM](#) (Deléage and Roux, 1987)
 - [DSC](#) (King and Sternberg, 1996)
 - [GOR I](#) (Garnier *et al.*, 1978)
 - [GOR III](#) (Gibrat *et al.*, 1987)
 - [GOR IV](#) (Garnier *et al.*, 1996)
 - [PHD](#) (Rost and Sander, 1993)
 - [PREDATOR](#) (Frishman and Argos, 1996)
 - [SIMPAA96](#) (Levin, 1997)
 - [Secondary structure consensus prediction](#)
- **Miscellaneous analysis tools :**
 - [Amino-acid composition](#).
 - [Coiled-coil](#) prediction (Lupas *et al.*, 1991)
 - [ColorSeq](#): color protein sequence
 - [HTH](#): Helix-turn-helix DNA-binding motifs detection (Dodd and Egan, 1990)
 - [Physico-chemical profiles](#)
 - [Transmembrane helices prediction](#) (PHDhtm, Rost *et al.*, 1995)
 - [FIT PDB molecules after alignment](#)
- **Acknowledgements**

Enabling Grids for E-science

NPS@ : BLAST Homology Search

http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat

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] [PBL]

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.

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] [PBL]

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.

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.

BLAST search on protein sequence databank

[Abstract] [NPS@ help] [Original server]

Program: blastp : protein sequence versus protein sequence databank

Database : UNIPROT-SWISSPROT

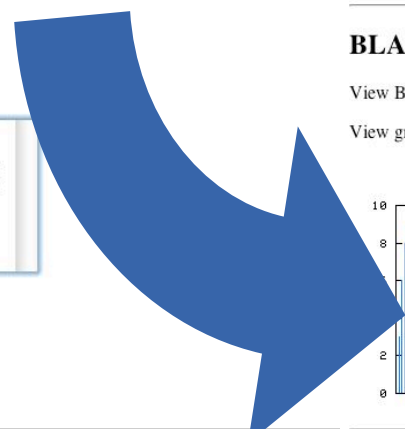
Sequence name (optional) :

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

```
MKKITIYDLAELSGVSASAVSAILNGNWKRRISAKLAEKVTRIAEEQGYAINRQASMLR
SKKSHVIGMIIPKYDNRVFGSIAERFEEMARERGLPIITCTRRRPELEIEAVKAMLSWQ
VDWVAVATGATNPDKISALCQQAGVPTVNLDPGLSPLSPVISDNYGGAKALTHKILANSA
RRRGELAPLTFIGRRATITPASVYAASTMRIASWGLACRRRIFWLPAIRKATLRTACRSQ
LAARRRCRGRYLLTRRYPWKGLCAGCRWW
```

Use the GRID resources from

SUBMIT CLEAR

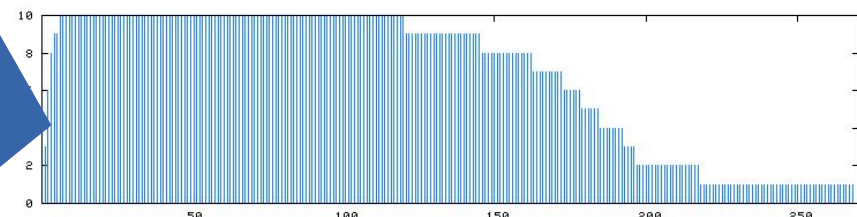


BLASTP results for : UNK_33610

View BLASTP in: [MPSA (Mac UNIX)] [About...] [AnTheProt (PC)] [Download...] [HELP]

View graphic in: [MPSA] [AnTheProt]

round(sum(score at a query sequence position)/max(score)*10) by query sequence position



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

- **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”