



Enabling Grids for E-science

GPSA

Grid Protein Sequence Analysis

Christophe Blanchet

CNRS IBCP

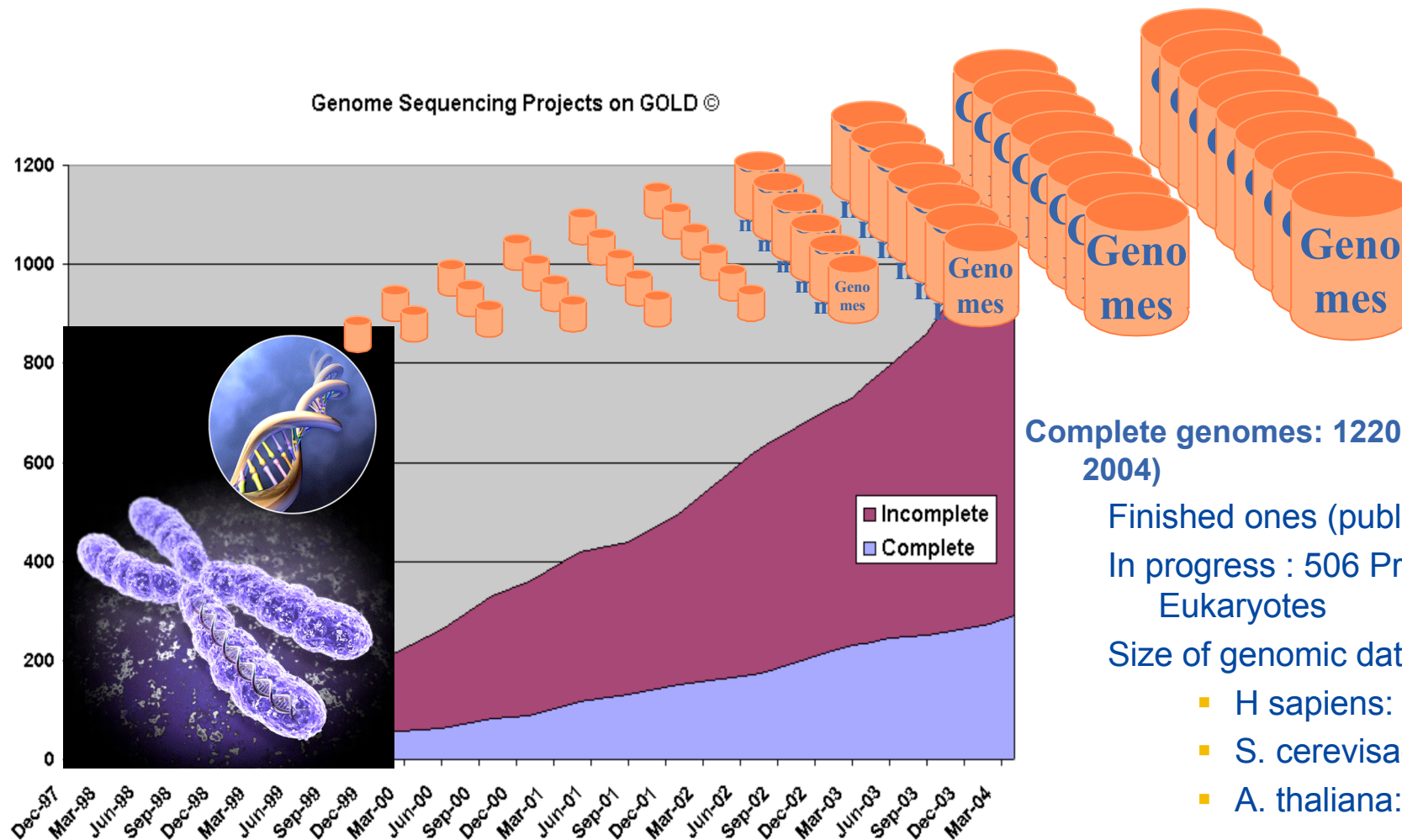
EGEE conference 2, Den Haag, 24 nov 2004

www.eu-egee.org



INFSO-RI-508833

- • **Bioinformatic context**
 - **NPS@: genomic web portal**
 - **GPS@: genomic grid portal**
 - **GPS@ demo**



Complete genomes: 1220 projects (nov. 2004)

Finished ones (published): 207

In progress : 506 Prokaryotes, 418 Eukaryotes

Size of genomic data

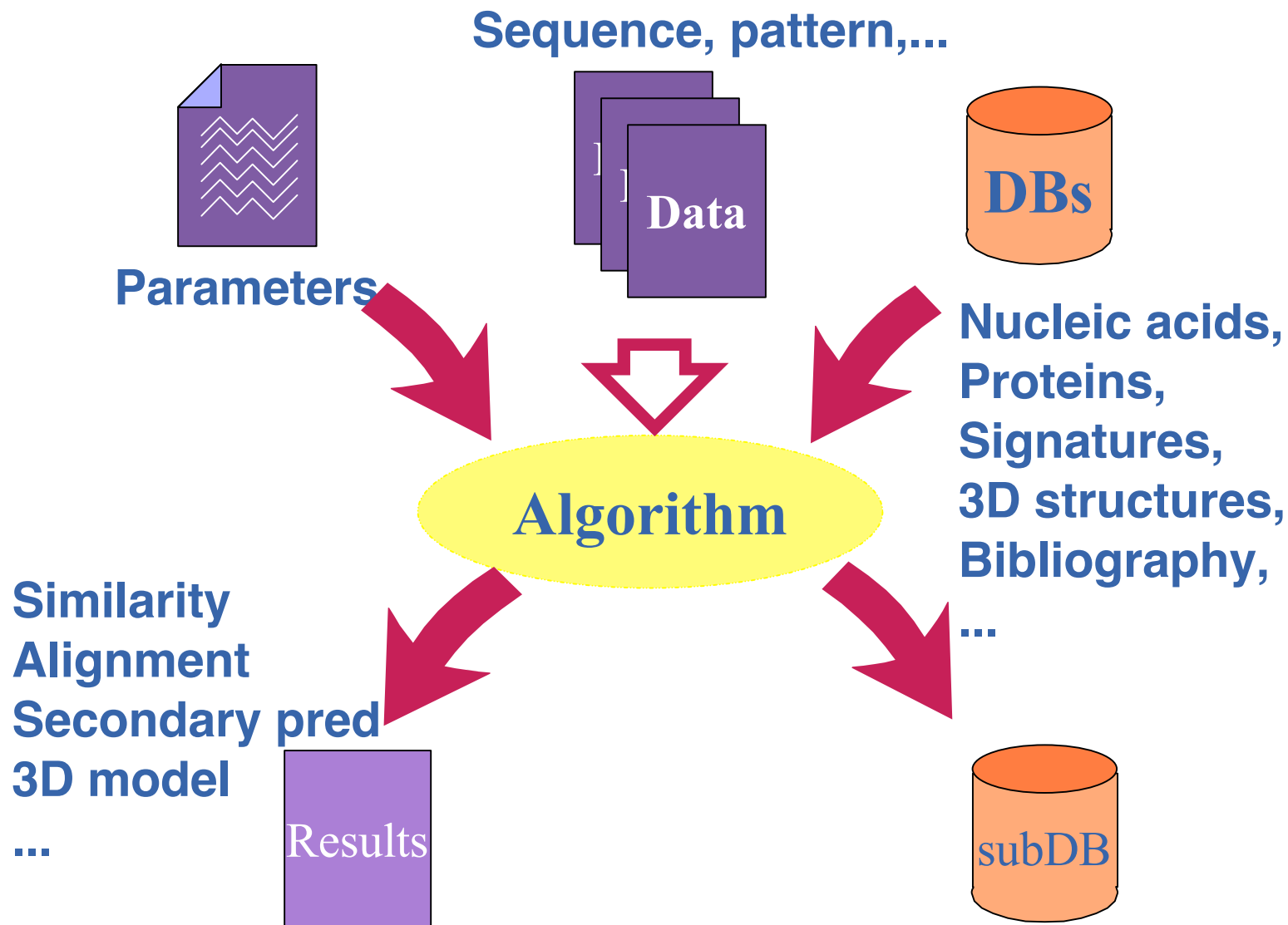
- H sapiens: 3.000 Mb (24)
- S. cerevisiae: 120 Mb,
- A. thaliana: 116 Mb

- Large scale Bioinformatics

Ratio of 1000 between raw and annotated data

Updating data and analyses (in automatic way)

- **Pairwise alignment: similarity**
 - BLAST, FASTA, SSEARCH
- **Pattern scanning:**
 - ScanProsite, **PattInProt**
- **Multiple alignment**
 - Clustal W, MultAlin
- **Secondary structure prediction**
 - GOR, Predator, PHD, SIMPA, **SOPMA, HNN,...**
- **Misc.**
 - Physico-chemical profiles transmembran section,



- **Bioinformatic context**
- • **NPS@: genomic web portal**
- **GPS@: genomic grid portal**
- **GPS@ demo**

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

• 34 integrated methods for protein sequence analysis

• **Hypertext cross-links to 11 international biological databanks**

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

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

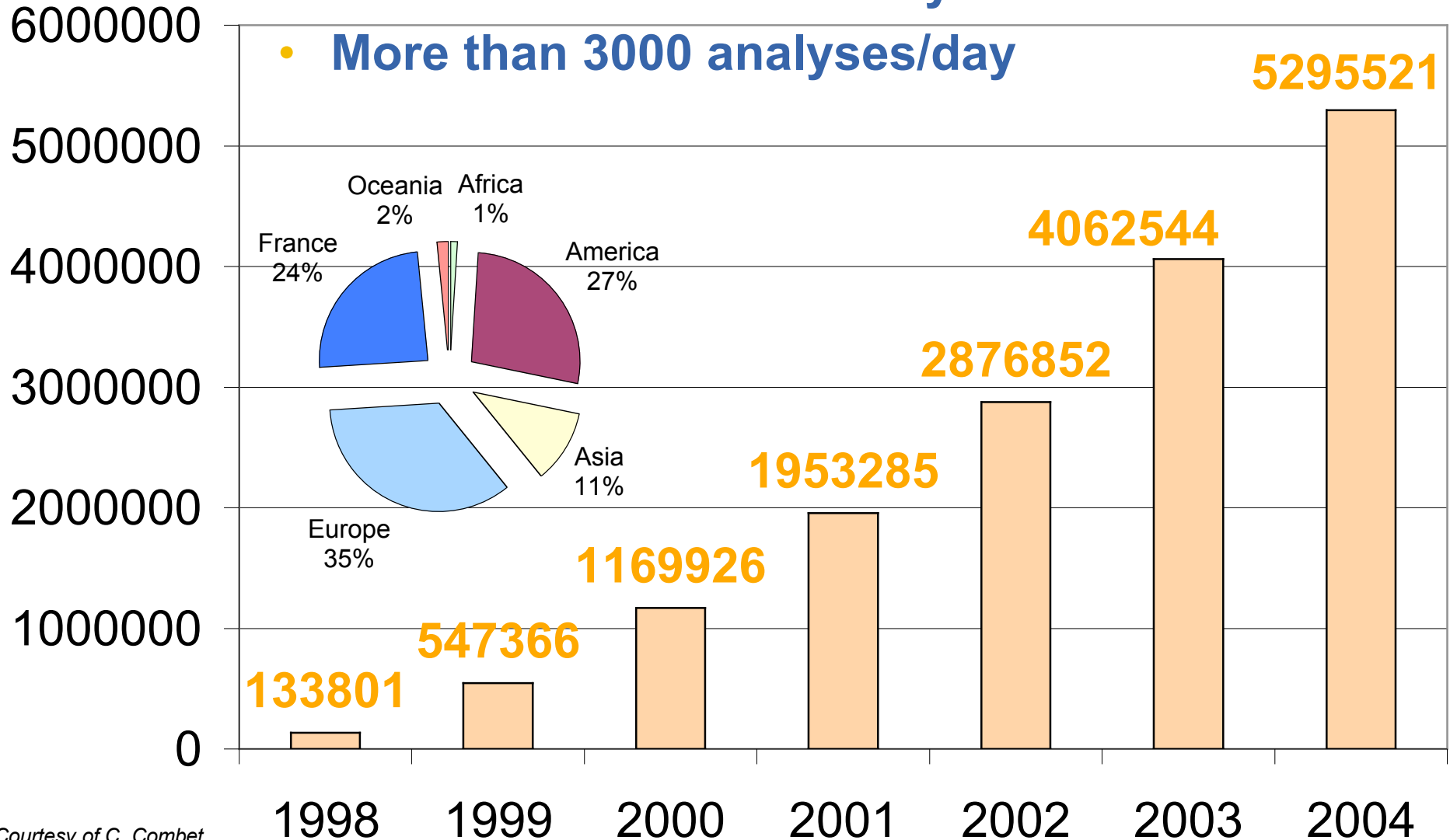
• International references: Expasy, University of California, InfoBioGen,...

• **« NPS@: Network Protein Sequence Analysis »**

Combet C., Blanchet C., Geourjon C. et Deléage G.

Tibs, 2000, 25, 147-150.

- More than 5 millions analyses since 1998
- More than 3000 analyses/day



Courtesy of C. Combet

- **Bioinformatic context**
- **NPS@: genomic web portal**
- • **GPS@: genomic grid portal**
- **GPS@ demo**


- **EGEE – Enabling Grids for E-Science in Europe.**
- **NA4 activity - Biomedical Applications**
- **GPSA**
 - One of the 3 biomed pilot applications (available at day 0)
 - GPS@ will be an integrated grid portal devoted to molecular bioinformatics.
 - GPSA is a porting experiment of the NPSA (Network Protein Sequence Analysis) services onto the EGEE grid. The current version is under development and deployed on LCG2.
 - <http://gpsa.ibcp.fr/>


- **NPS@ - network Protein Sequence Analysis**
 - Production web portal hosting proteins databases and algorithms for sequences analysis.
 - Online since 1998. Hosted by a cluster of 14 CPUs
 - Currently, strong restrictions in terms of databanks and algorithms due to **limited resources**.
 - Therefore the number of users connecting to the portal and the size of the data sets are restricted by the server but they will have to process.
- **GPSA: bringing the EGEE grid services to the NPSA genomic portal**
 - The same user community will be eager to **transparently** use the grid version of the same service once available.
 - Providing biologists with more and larger databanks
 - Providing biologists with more bioinformatic methods and connections between them (ClustalW and PSSP)
 - Allowing analysis of larger datasets

Welcome on GPSA, Grid Genomic Web Portal

http://gpsa.ibcp.fr/imethods.html

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






Grid Protein Sequence @analysis 

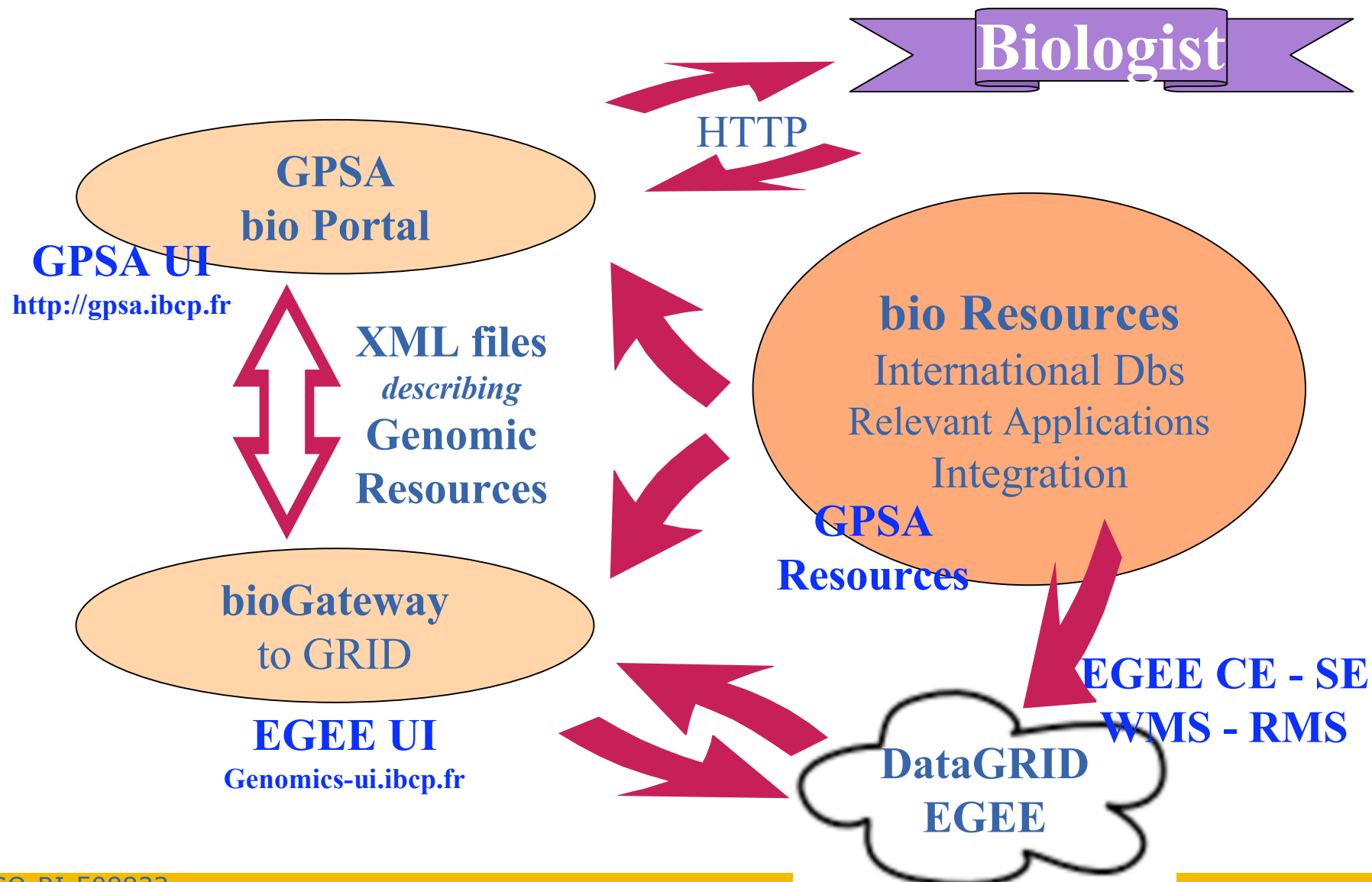
Institut de Biologie et Chimie des Protéines

Bioinformatic Grid web portal dedicated to protein sequence analysis.

[\[GPSA\]](#) [\[mySEQ\]](#) [\[HELP\]](#) [\[REFERENCES\]](#) [\[NPS@\]](#) [\[PBIL-Gerland\]](#) [\[PBIL\]](#)

The methods tagged with the EGEE logo  are computed on the grid infrastructure provided by the EGEE project. The other ones are still computed on our local resources

- **Work with your own proteic sequence databases : mySEQ**
 - [Upload](#) your sequence database on GPSA
 - [List](#) your sequence databases registered on GPSA
- **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) protein query sequence
 - SSEARCH search,
 - [\[on GRID\]](#)  protein query sequence
- **Sequence homology search against nucleic databases :**
 - [BLAST search](#) (protein (tblastn) or nucleic (blastn, tblastx) query sequence)
 - [FASTA search](#) (nucleic query sequence)
- **Patterns and signatures search :**
 - PATTINPROT : scan protein sequence(s) against PROSITE-like pattern(s)
 - [\[on GRID\]](#) 
- **Multiple alignment:**



- **Protein analysis application**

- Similarity: Ssearch
- Multiple Alignment: ClustalW
- Functionnal site and signatures: PattInProt
- Prot Secondary Pred: Predator, Gor4, Simpa96

- **Data**

- Sequence Bank: SWISSPROT
- Pattern and profile bank: PROSITE

- **Deploying more protein databanks (using RMS)**
- **Integrating gridified databanks**
 - in progress into GPSA
 - to be done into NPS@
- **Deploying bioinformatic algorithms (using ESM)**
 - Tested with ClustalW and SSEARCH algorithms (in LAL site)
- **Integrating more bioinformatic algorithms**
 - Currently 4 in GPSA (chosen as models for the others)
 - Currently 1 into NPS@



CNRS IBCP

Institute of Biology and
Chemistry of Proteins
Lyon, France

