GPSA

Grid Protein Sequence Analysis

Christophe Blanchet

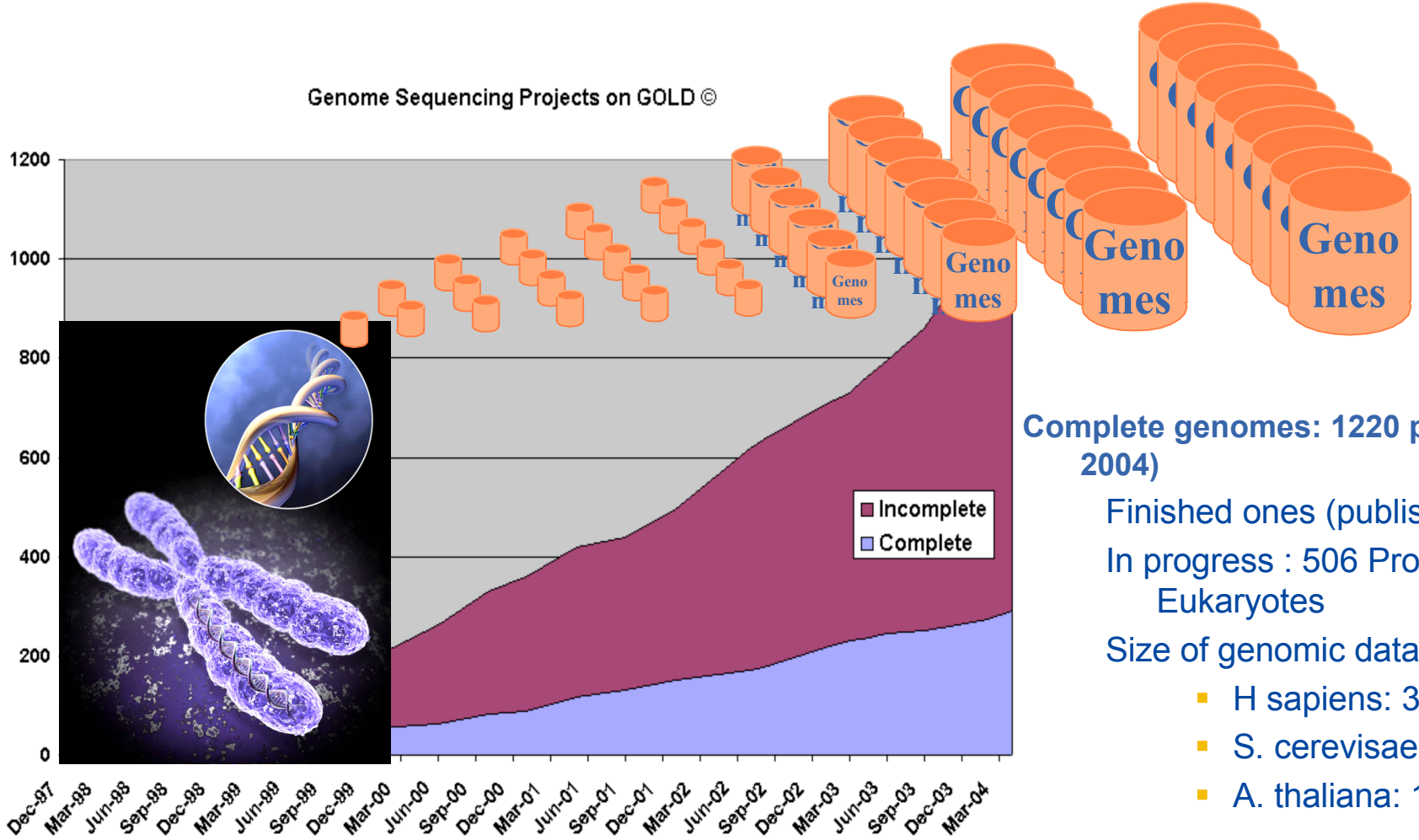
CNRS IBCP

EGEE conference 2, Den Haag, 24 nov 2004

Contents

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

Complete genome projects



Complete genomes: 1220 projects (nov. 2004)

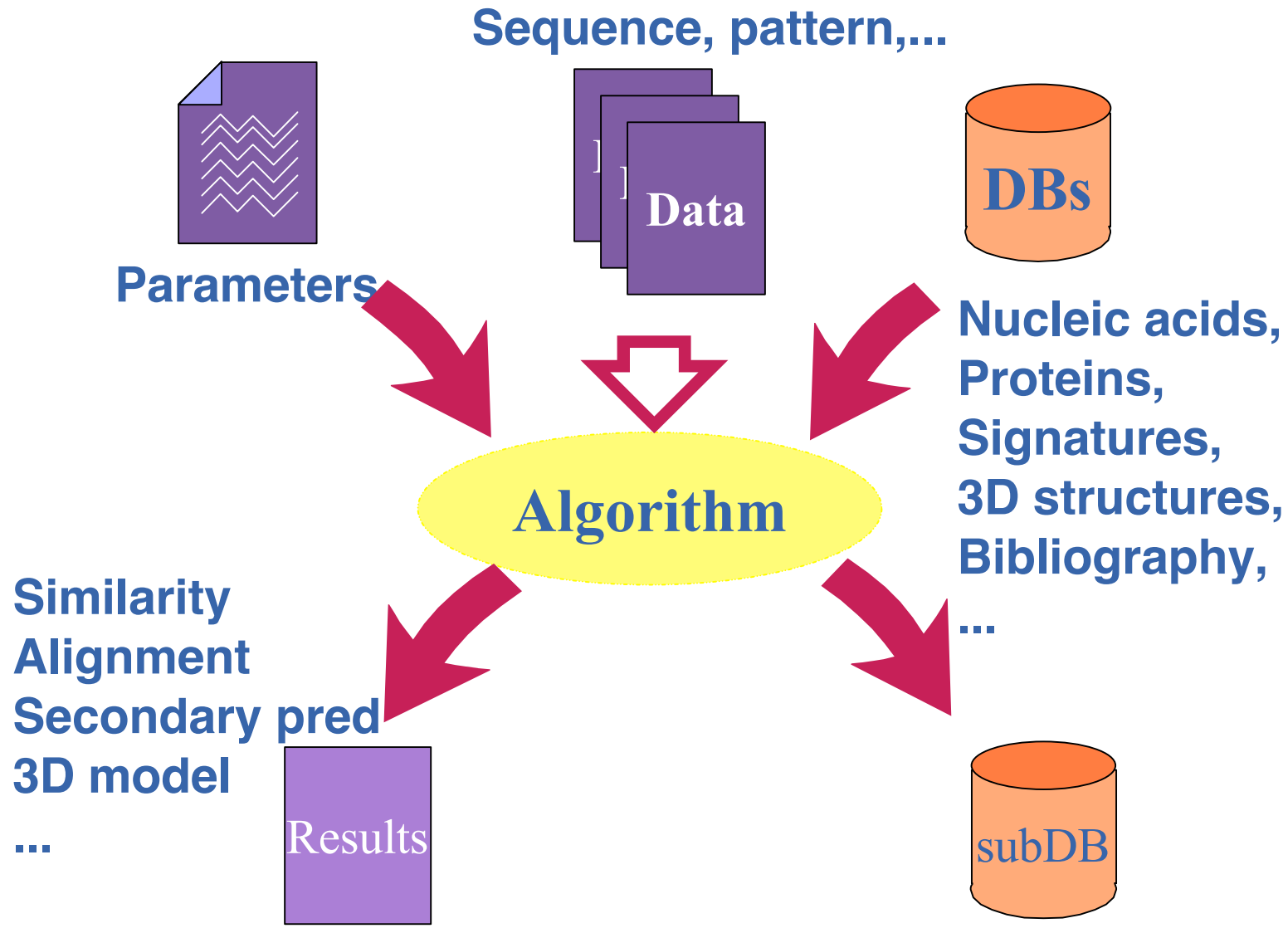
- Finished ones (published): 207
- In progress : 506 Prokaryotes, 418 Eukaryotes
- Size of genomic data
  - H sapiens: 3.000 Mb (24)
  - S. cerevisae: 120 Mb,
  - A. thaliana: 116 Mb

- **Large scale Bioinformatics**
  - Ratio of 1000 between raw and annoted data
  - Updating data and analyses (in automatic way)

World-used methods for protein sequence analysis

- **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 algorithm model



TOC

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

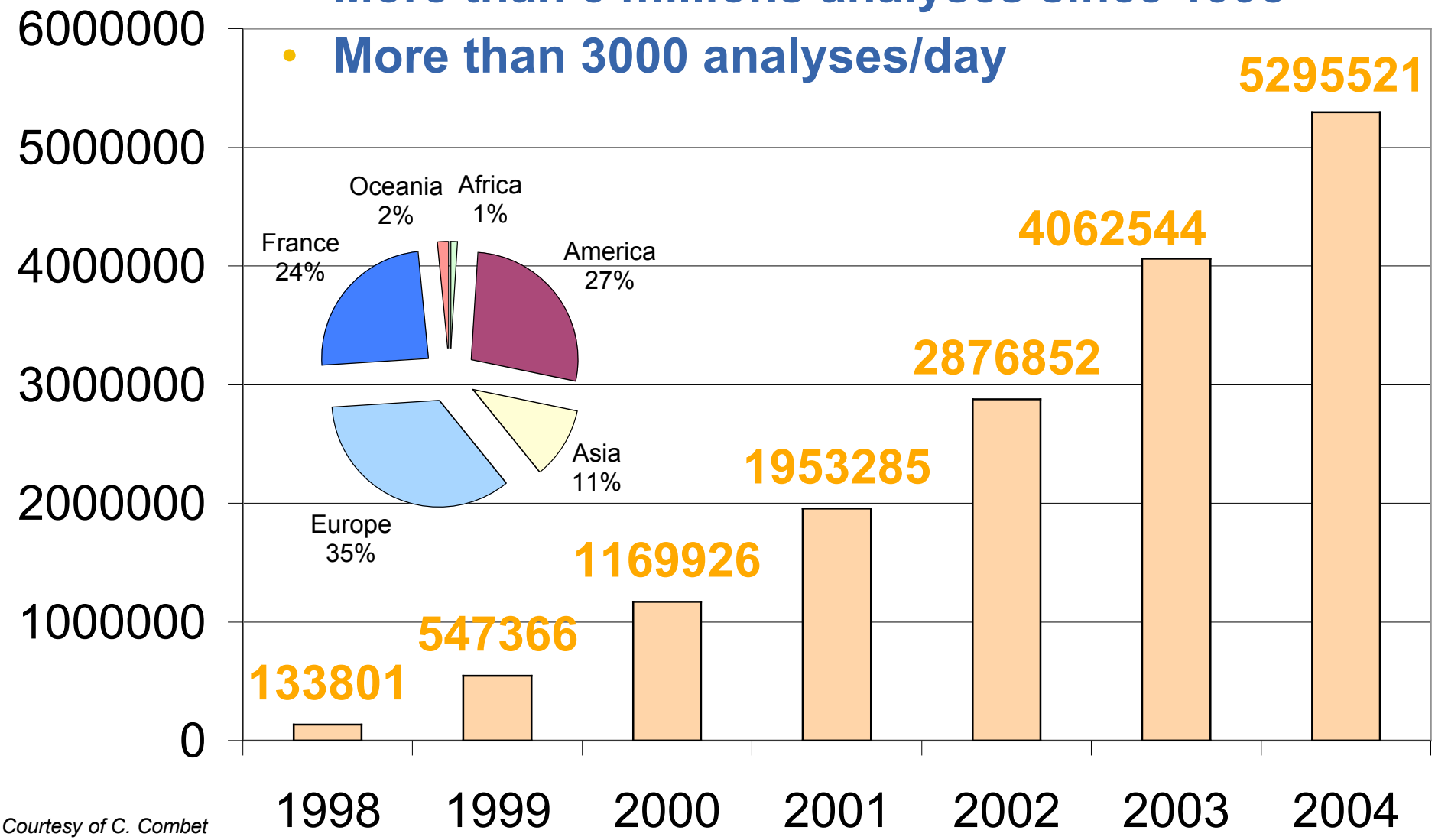
Genomic web portal : NPS@

- 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.**

NPSA hits

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



Courtesy of C. Combet

TOC

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

GPSA-Grid Protein Sequence Analysis

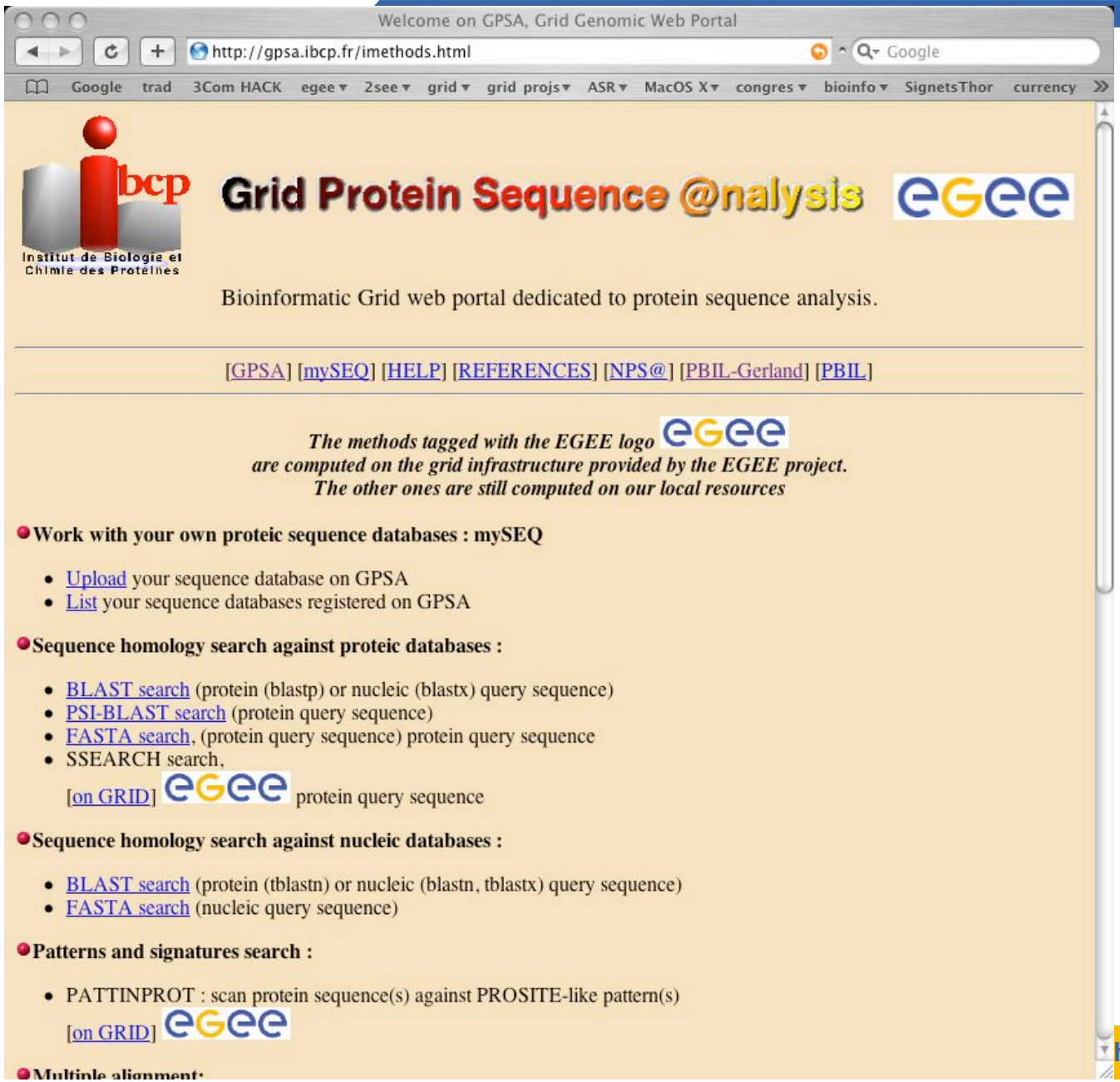
- **EGEE – Enabling Grids for E-Science in Europe.**
- **NA4 activity - Biomedical Applications**

- **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/

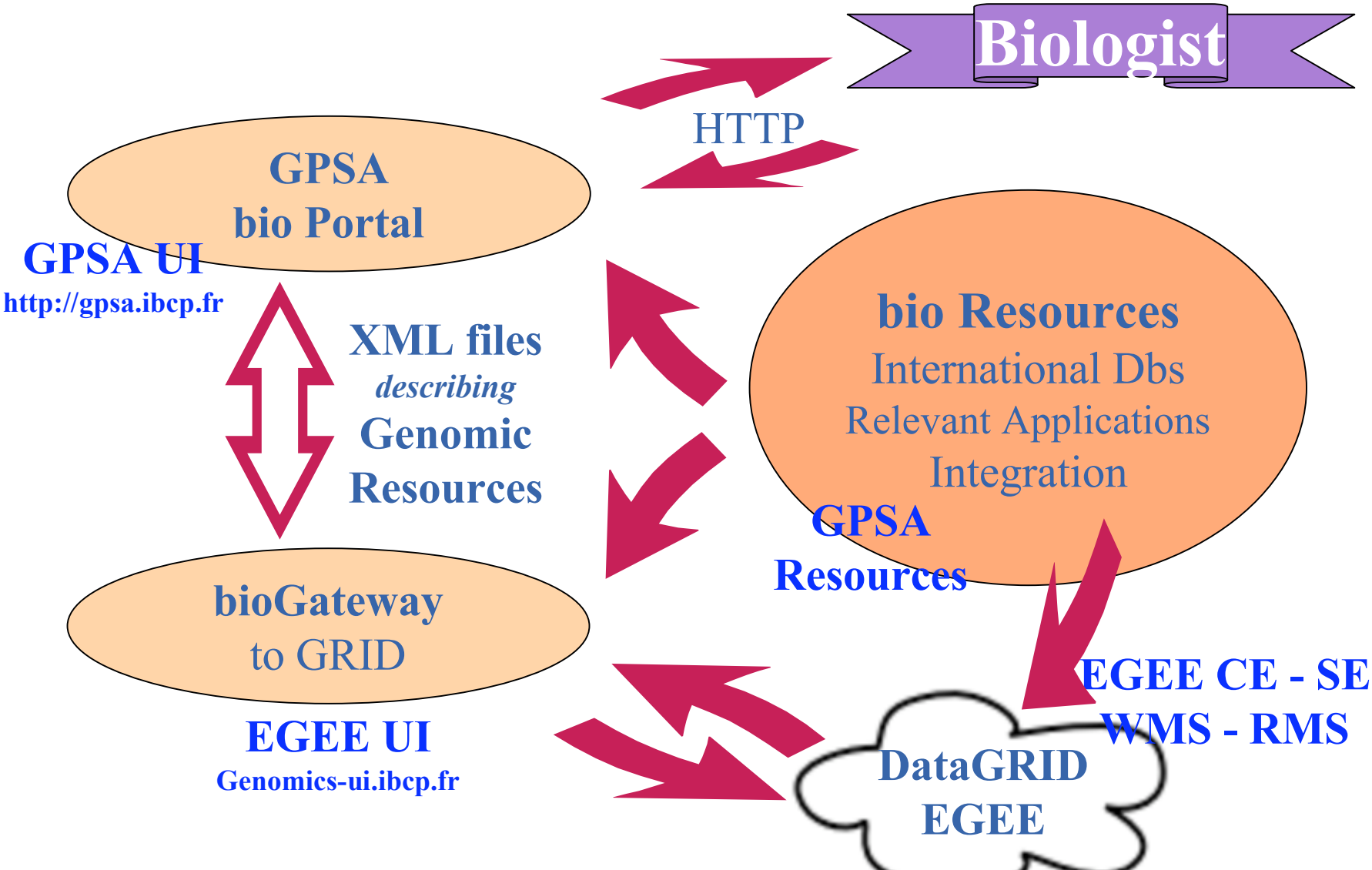
GPSA: grid benefits

- **NPS@ - network Protein Sequence Analysis**
  - Production web portal hosting proteins databases and algorithmes for sequences analysis.
  - Online since 1998. Hosted by a cluster of 14 CPUs
  - Currently, strong restrictions in terms of databanks and algorithmes 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

GPSA: welcome



GPSA : Portal Architecture



Bioinformatic Resources on EGEE (Nov 2004)

- **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

Bioinformatic resources on EGEE Perspective (Feb 2005)

- **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@

Demo Time

CNRS IBCP

Institute of Biology and Chemistry of Proteins

Lyon, France