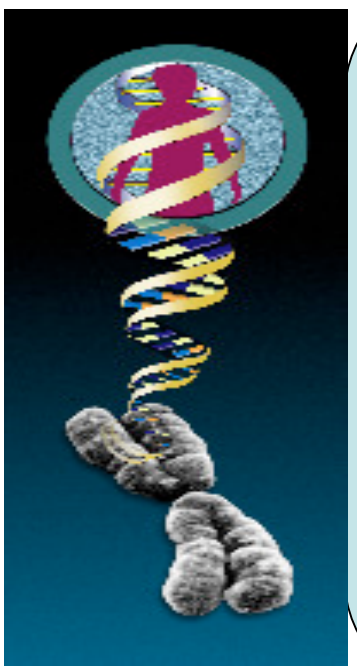
GriPPS
Bioinformatics on Grid
Protein pattern scanning taken as model

[Gri]-P-{PS} Grid Protein Pattern Scanning

> Institut de Biologie et Chimie des Protéines
> Centre de Calcul de l'IN2P3
> Laboratoire de l'Informatique du Parallélisme

Institut de Biologie et Chimie des Protéines — IN2P3 Centre de Calcul



The Grid Protein Pattern Scanning-GriPPS project has adapted bioinformatic algorithms of protein pattern scanning to the grid infrastructure. The behavior of the algorithms and the data have been studied on several experimental grids, as a model for the gridification of other common bioinformatic tools and databanks. The tested middleware are those from the projects DataGrid (EU FP5)/EGEE (EU FP6), e-Toile (Fr-RNTL) and GASP (Fr ACI GRID).

Protein pattern scanning: *PattInProt*
- Predict protein/gene function
- Cluster proteins into family

- Sequence annotation and biological crosslinks
- PattInProt is integrated into our software programs (e.g. MPSA) and web portals (e.g. NPS@ and GPS@)

GRID computing context
- Adapt PattInProt bioinformatic algorithm and data to the grid in order to foresee their behavior on a grid plateform
- Identify specific bioinformatic constraints on the grid
- Test on several middleware and model of grid: e-Toile, DataGrid/EGEE, DIET.

GRID benefit
- More complex analyses on larger data set, lower threshold
- Distributing sequence databanks
- Integrity and security of data and method software
- Recommendations on gridification of similar bioinformatic algorithms

PattInProt Algorithm

Usage

Pattern databank versus sequence databank

Two version

Identity:Exact matching

Similarity:Allowing biological mismatch to enhance sensitivity

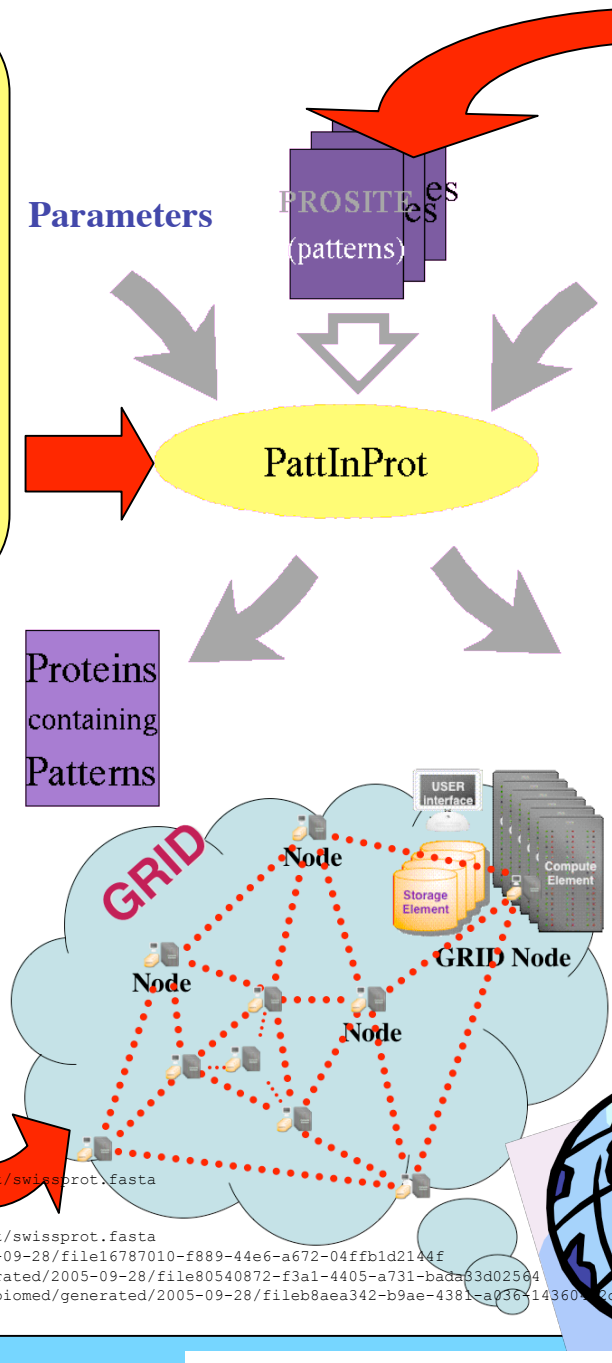
Optimization

Bit parallelism

Sequentialized recursive philosophy for gap expansion

Self indexed protein

Best starting pattern position



Biological Data

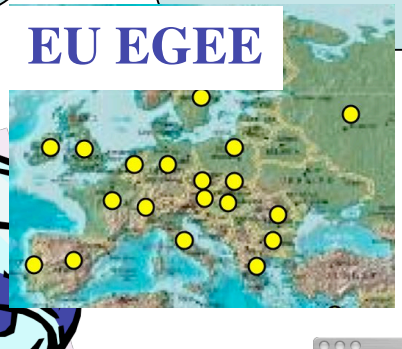
Protein: sequence of aminoacids
- Example: …**ELKRNQLTGIEAFEGASHIQELQLGENKIKEI**…
- Clustered into families according to their biological function.
- Part of a genome: rats, drosophilia, human …

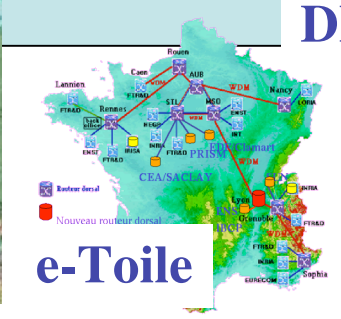
- Protein pattern:
  - Subset of regular expression.
    - Example : {ACH}-A-G-B-x(5,78)-Q-K(2)-[SBE]-N.
  - One or more pattern matched
    - Protein might belong to a given family
    - Hypothesis on the function of an unknown protein.

Gridification

Deploying Biological Resources into Grid context

- A Grid is:
  - Different computing resources, stated in remote places, linked with the grid middleware,
  - CPU, memory, disk storage, informations, etc.
  - Authentified access to the Grid
- GriPPS project have had access to several platforms (see below)





Gridified Tools and Data

- **Bioinformatic Tools**
  - PattInProt has been deployed on e-Toile, DIET and DataGRID/EGEE platforms
- **Biological Data**
  - **Sequence databanks:**
    - **Swiss-Prot, TrEMBL, …**
  - **Pattern databank**
    - **PROSITE**

```
> lcg-lg --vo biomed  lfn://genomics_gpsa/db/swissprot/swissprot.fasta
guid:1968c0ca-8131-4d6e-b539-b36cb6b8f8b0

> lcg-lr --vo biomed  lfn://genomics_gpsa/db/swissprot/swissprot.fasta
sfn://cclcgseli01.in2p3.fr/grid/biomed/generated/2005-09-28/file16787010-f889-44e6-a672-04ffb1d2144f
sfn://grid11.lal.in2p3.fr/var/storage/LCG/biomed/generated/2005-09-28/file80540872-f3a1-4405-a731-badad33d02564
sfn://marseillese01.mrs.grid.cnrs.fr/var/storage/LCG/biomed/generated/2005-09-28/fileb8aea342-b9ae-4381-a036-14360[illegible]2d1[illegible]
```

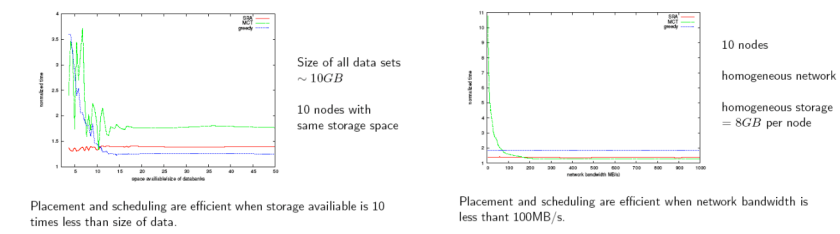
Results

A model for the gridification of bioinformatic resources on grid platform:
- Describing bioinformatic software and data with the XML language
- Using a common XML DTD through all the contexts: grid execution, portal, …
- Including biological semantic concepts into the XML description files

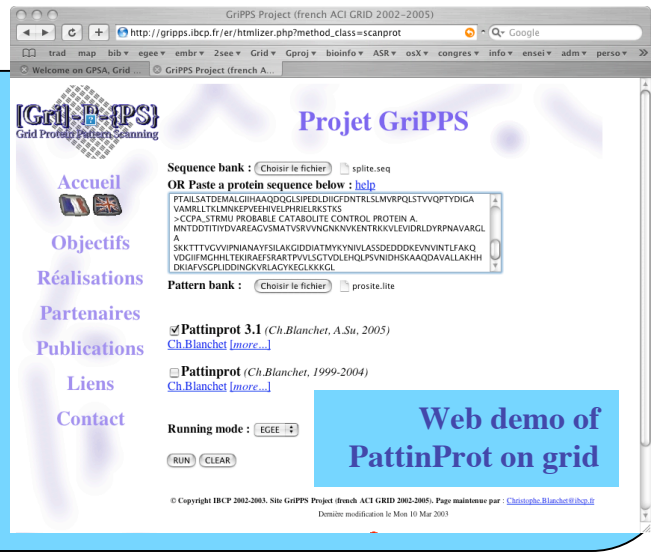
Gridifying tools and databanks on different grid platforms: e-Toile, DataGRID/EGEE, DIET

Grid Service Providing (GSP) for bioinformatic resources: protein sequence analysis

« Simultaneous scheduling of data replication and computation in Grids »



A. Vernois PhD thesis (LIP-IBCP)
Granted by ACI GRID 2002



Project granted by ACI GRID 2002

Action Concertée Incitative [ACI] Globalisation des Ressources Informatiques et des Données [GRID]

ministère délégué recherche et nouvelles technologies

http:// gripps.ibcp.fr

Contact: Christophe.Blanchet@ibcp.fr