

# Infrastructure de Grille pour la Bioinformatique - RENABI GRISBI -

Christophe Blanchet

**Institut de Biologie et Chimie des Protéines**

CNRS IBCP, LYON, FRANCE

[christophe.blanchet@ibcp.fr](mailto:christophe.blanchet@ibcp.fr)

27 Avril 2010, journée «Grille de Calcul», CPPM, Luminy



# Motivations in Bioinformatics

- Deluge of data
- Most common analyses have evolved to a larger scale,
  - from the study of a single gene/protein to a whole genome/proteome, from a single metabolic pathway to Systems Biology.
- Bioinformatics needs large-scale research infrastructures
  - to store very large biological data sets, complex and heterogeneous,
  - to make these data available for intensive scientific computing.
- Facing to these growing needs, we should take advantage of new developments in distributed and intensive computing
  - supercomputers or large clusters,
  - computing and data grids,
  - cloud computing and virtualization in datacenters
  - specific intensive-computing hardwares like GP-GPU, FPGA or Cell processor.

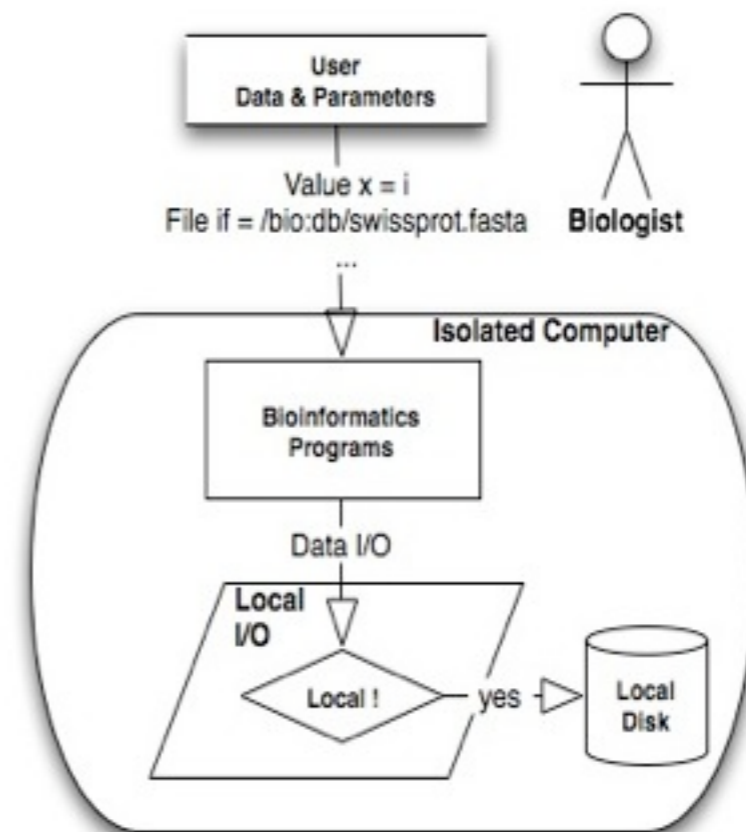
# Data and Tools

- InterPro, pFam, Genmark, Genezilla, Pred. Intron\*, Sys. Biology\*, Réseaux Méta\*, Ancêtres (hiador, MGR), Autodock, Docking@Grid \*, Base (stats), Pasa\* (Base), ASCQ\_me\*, R, MGA, Mauve, MathLab, Scilab, Show\*, R'mes\*, EMBOSS, Gromacs, ClustalW, Maft, MAST, MEME, Phred/Phrap, BLAST, FASTA, SSearch, MUSCLE, PhyML, Dialign, multalin, RepeatMasker, Amber, NAMD, JUMNA\*, ADAPT\*, MaxDo\*, Curves\*, Prophet\*, DALI, SUMO(\*), PattInProt\*, ...
- UniProt, Génolevures\*, Base, AcNuc (\*\*), GenBank, EMBL, PRODOM\*, Ensembl, Hogenom\*, Homolens\*, PDB, Génomes Complets, TransFac, Nr, SRS (\*\*), SUMO(\*), PROSITE, ABC, KEGG, ...

# Combining Data and Tools

1. Protein database
2. Semantic selection of sequences  
e.g.: species = human
3. Similarity selection of sequences  
e.g.: with BLAST
4. Aligning these subset of sequences  
e.g.: with ClustalW
5. Adding structural data  
e.g.: with insertion of protein secondary structure predictions : SOPMA, GOR4, PHD, ...
6. Building a 3D structural model from this multiple alignment ...

Transferring data among all these steps



# OUTLINE

- GRIDs ...
- Bioinformatics experiment
  - Study the nucleosome
- National Bioinformatics Grid: RENABI GRISBI

Source: CNRS IDRIS

# GRID...S



Source: JÜLICH SUPERCOMPUTING CENTRE (JSC)



Source: CSCS



Source: CNRS IBCP



Source: CNRS IDRIS

# GRID...S



Source: JÜLICH SUPERCOMPUTING CENTRE (JSC)



Source: CSCS

**EGEE, DEISA, Grid5000, TeraGrid,  
SwissBioGrid, myGrid, Dgrid, ...**

**=> RENABI GRISBI**

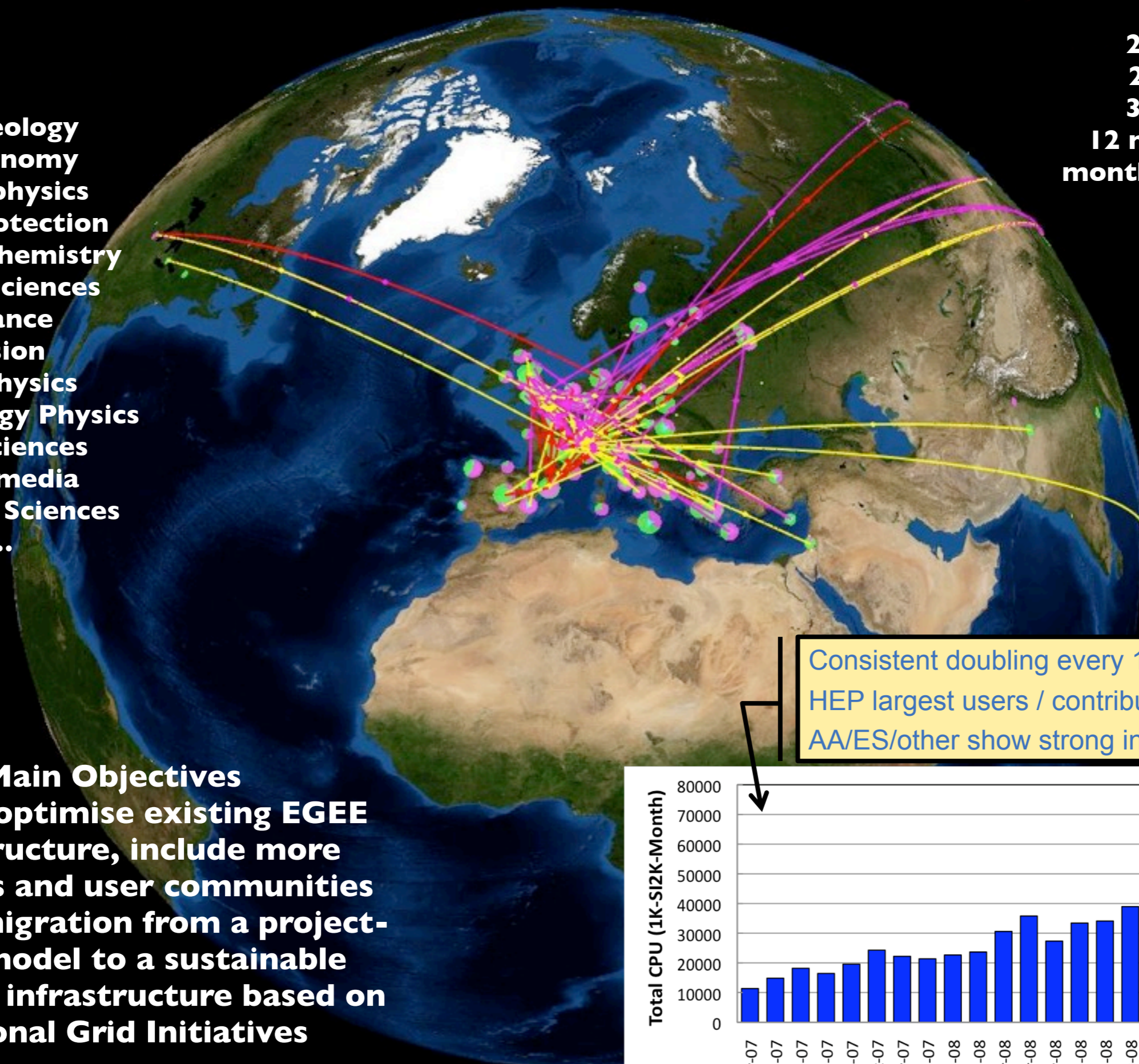
**=> ELIXIR, EGI, PRACE**



Source: CNRS IBCP

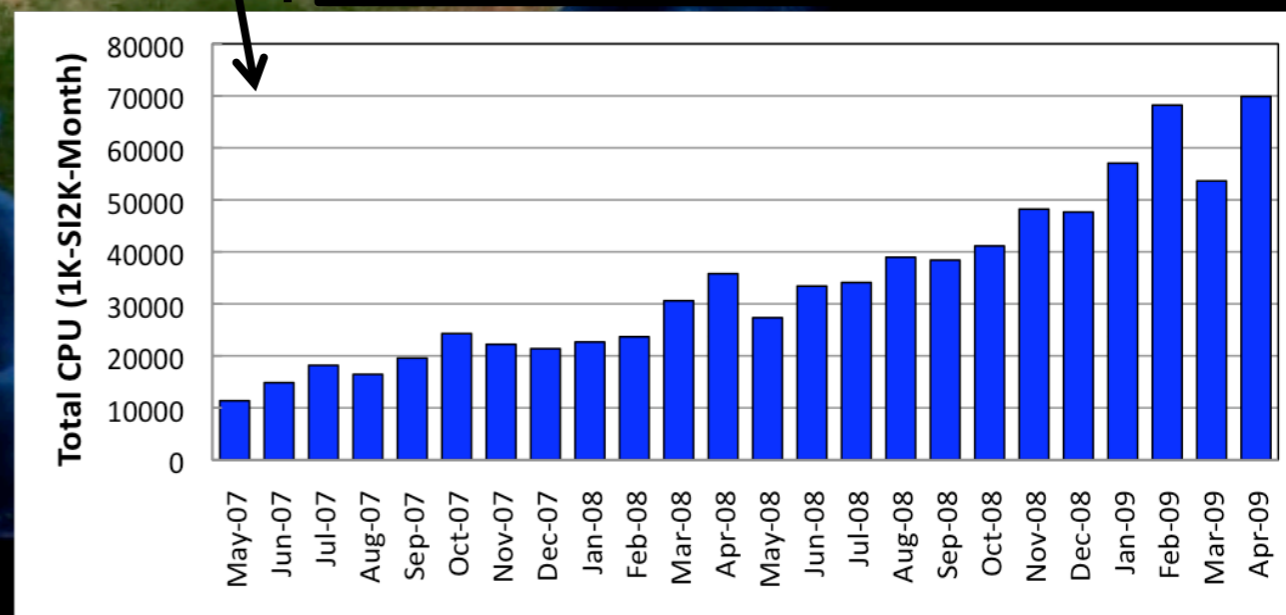
**Archeology**  
**Astronomy**  
**Astrophysics**  
**Civil Protection**  
**Comp. Chemistry**  
**Earth Sciences**  
**Finance**  
**Fusion**  
**Geophysics**  
**High Energy Physics**  
**Life Sciences**  
**Multimedia**  
**Material Sciences**  
 ...

**15,000 users**  
**140,000 CPUs**  
**(cores)**  
**260+ sites**  
**25Pb disk**  
**39Pb tape**  
**12 million jobs/**  
**month +45% in one**  
**year**



Consistent doubling every 12-18 months.  
 HEP largest users / contributors  
 AA/ES/other show strong increase

**Main Objectives**  
**Expand/optimize existing EGEE**  
**infrastructure, include more**  
**resources and user communities**  
**Prepare migration from a project-**  
**based model to a sustainable**  
**federated infrastructure based on**  
**National Grid Initiatives**





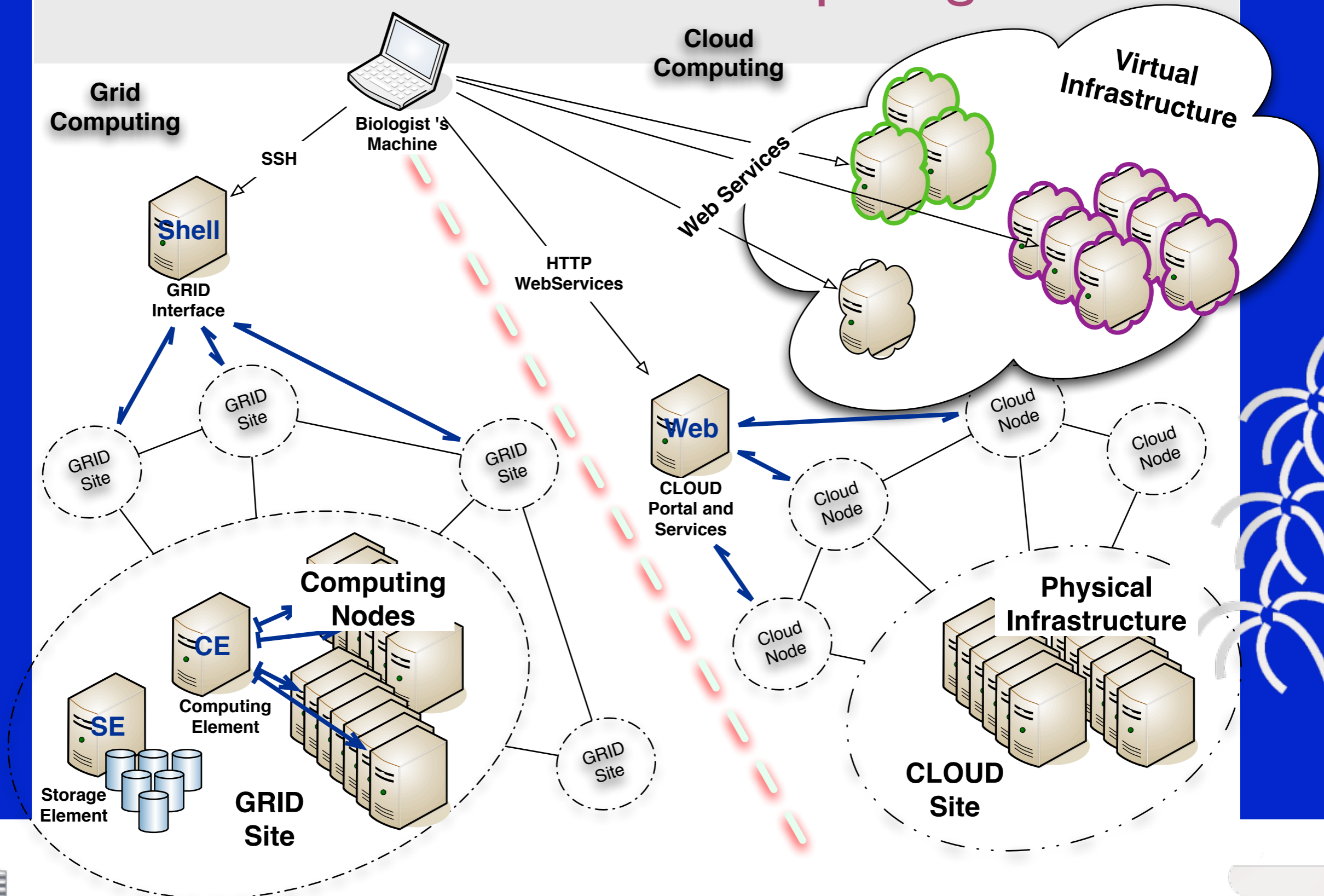
# Current initiatives related to BIO



- **EMBRACE, EU-FP6, 2005-2010**
  - Integrate the major databases and software tools in bioinformatics, using existing methods and emerging Grid service technologies.
  - <http://www.embracegrid.info>
- **ELIXIR, ESFRI 2008-2010**
  - Construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.
  - <http://www.elixir-europe.org>



# Grid et Cloud Computing



# Protein–DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies†

Krystyna Zakrzewska,\* Benjamin Bouvier, Alexis Michon, Christophe Blanchet and Richard Lavery

*Received 2nd June 2009, Accepted 24th September 2009*

*First published as an Advance Article on the web 7th October 2009*

DOI: 10.1039/b910888m

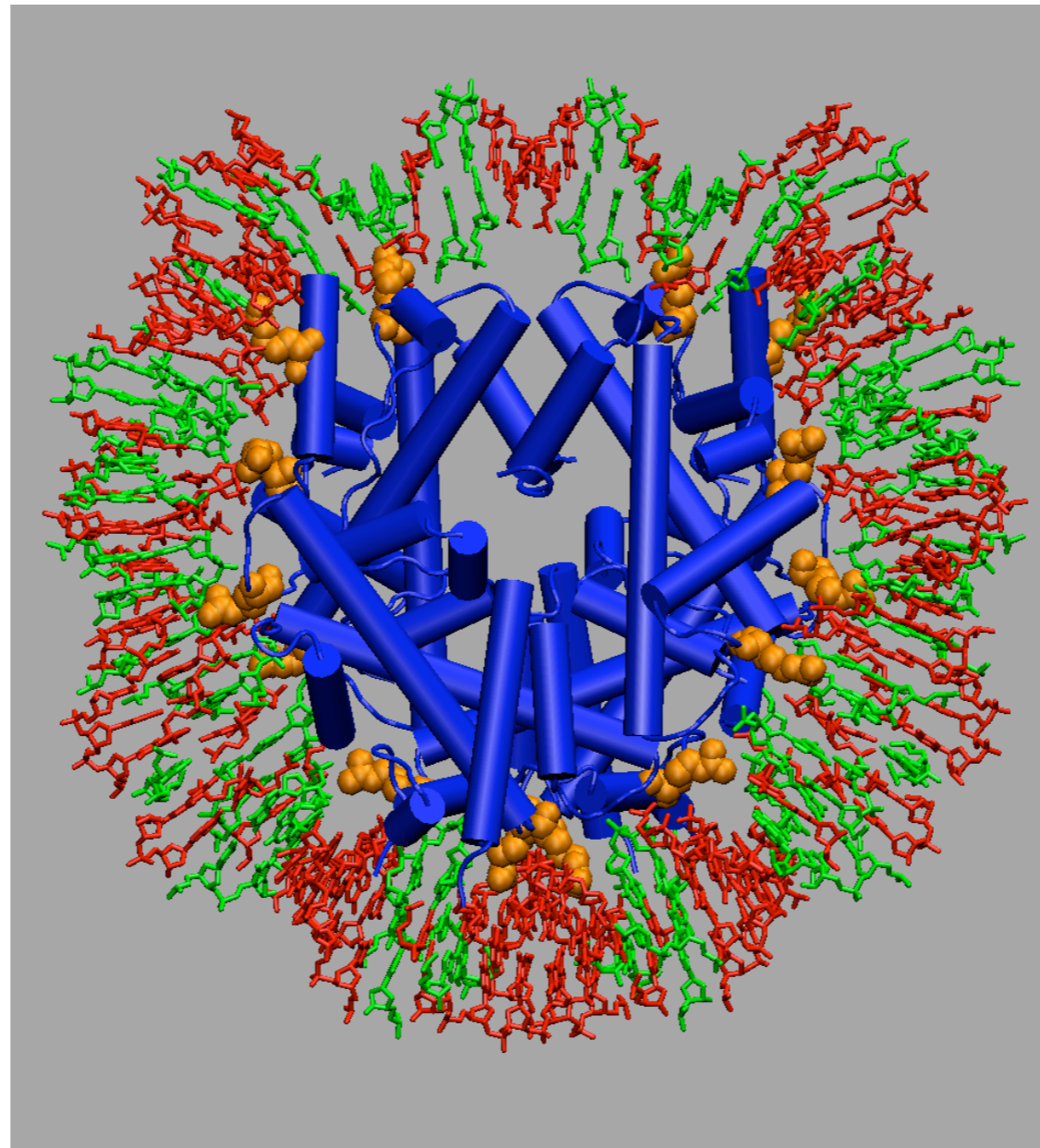
We use a physics-based approach termed ADAPT to analyse the sequence-specific interactions of three proteins which bind to DNA on the side of the minor groove. The analysis is able to estimate the binding energy for all potential sequences, overcoming the combinatorial problem *via* a divide-and-conquer approach which breaks the protein–DNA interface down into a series of overlapping oligomeric fragments. All possible base sequences are studied for each fragment. Energy minimisation with an all-atom representation and a conventional force field allows for conformational adaptation of the DNA and of the protein side chains for each new sequence. As a result, the analysis depends linearly on the length of the binding site and complexes as large as the nucleosome can be treated, although this requires access to grid computing facilities. The results on the three complexes studied are in good agreement with experiment. Although they all involve significant DNA deformation, it is found that this does not necessarily imply that the recognition will be dominated by the sequence-dependent mechanical properties of DNA.

Introduction

we can expect that these deformations (and, most notably



# Nucleosome



8 proteins, one DNA fragment  
147 bp, 21,000 atomes

# Study the Nucleosome : Grid added Value

- **Scientific Goal**

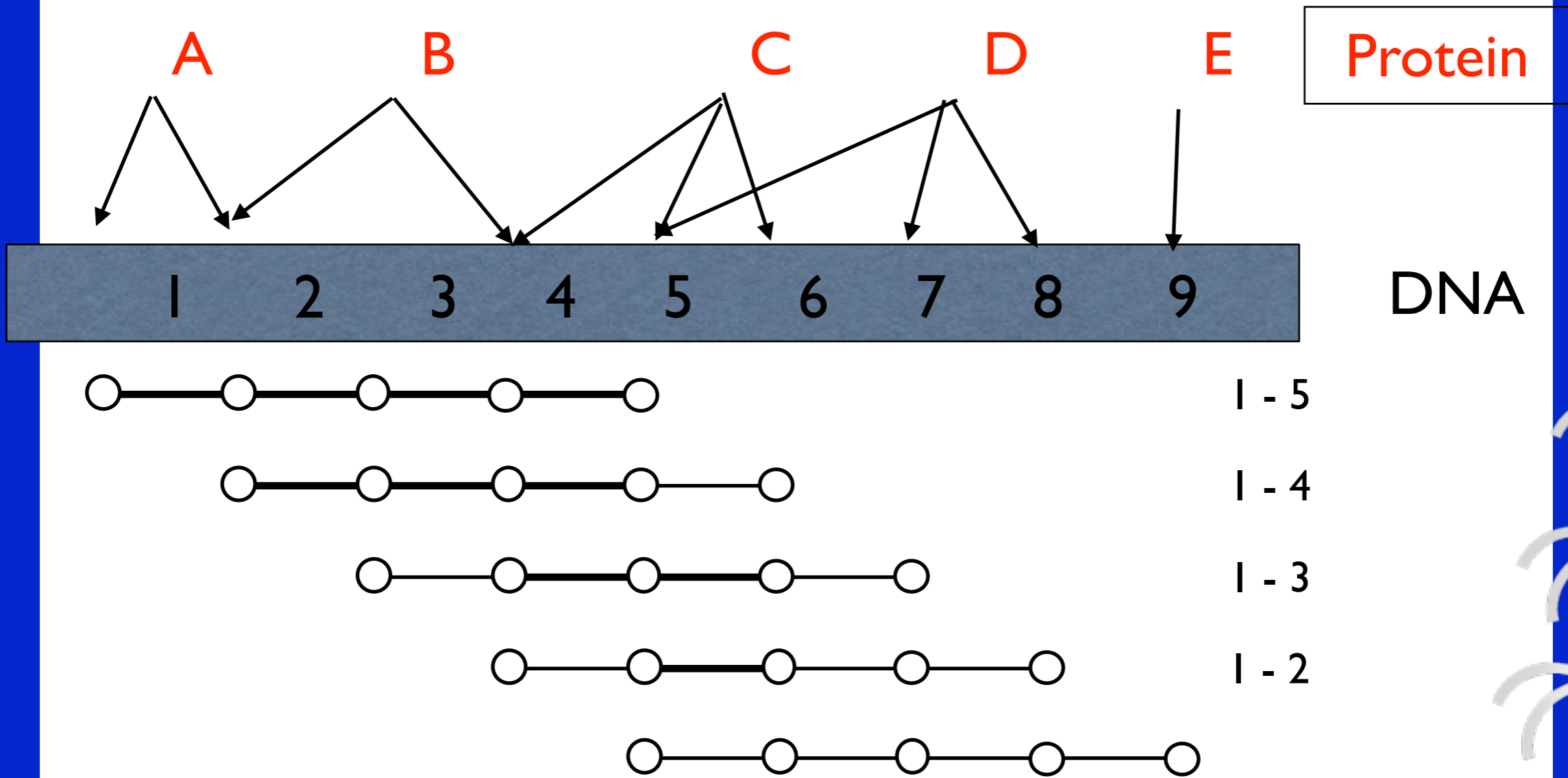
- Experimental data shows that proteins often find their target sites on DNA faster than simple diffusion would allow.
- nucleosome involves an eight protein complex binding to roughly 140 bp DNA fragments: 1086 potential sequences
- Determine the preferred nucleic base patterns for fixation of DNA on Histones

- **Methods: ADAPT**

- JUMNA program developed in our team
  - Fortran g77 libraries (g77 compiler)
- Reduction in combinatorial: overlapping fragments, "Nplets", involving N nucleotide pairs
  - moving one step along DNA: 143 positions
  - N=4, 44=256 sequences, 35,840 simulations
  - N=5, 45=1024 sequences, 146,432 simulations

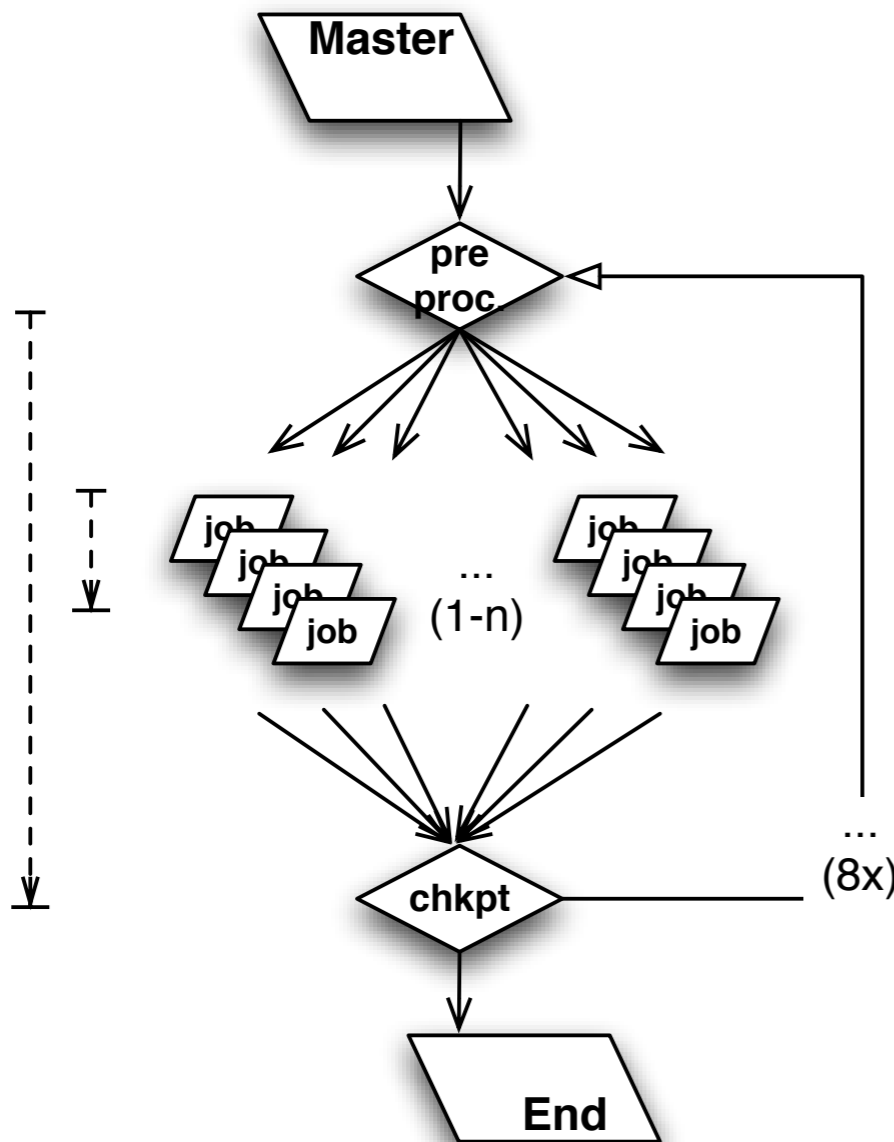
**Contact: [richard.lavery@ibcp.fr](mailto:richard.lavery@ibcp.fr)**

# Fragment of 5-plet



143 bp + 2 x 2 bp ends

# Parallelization Model



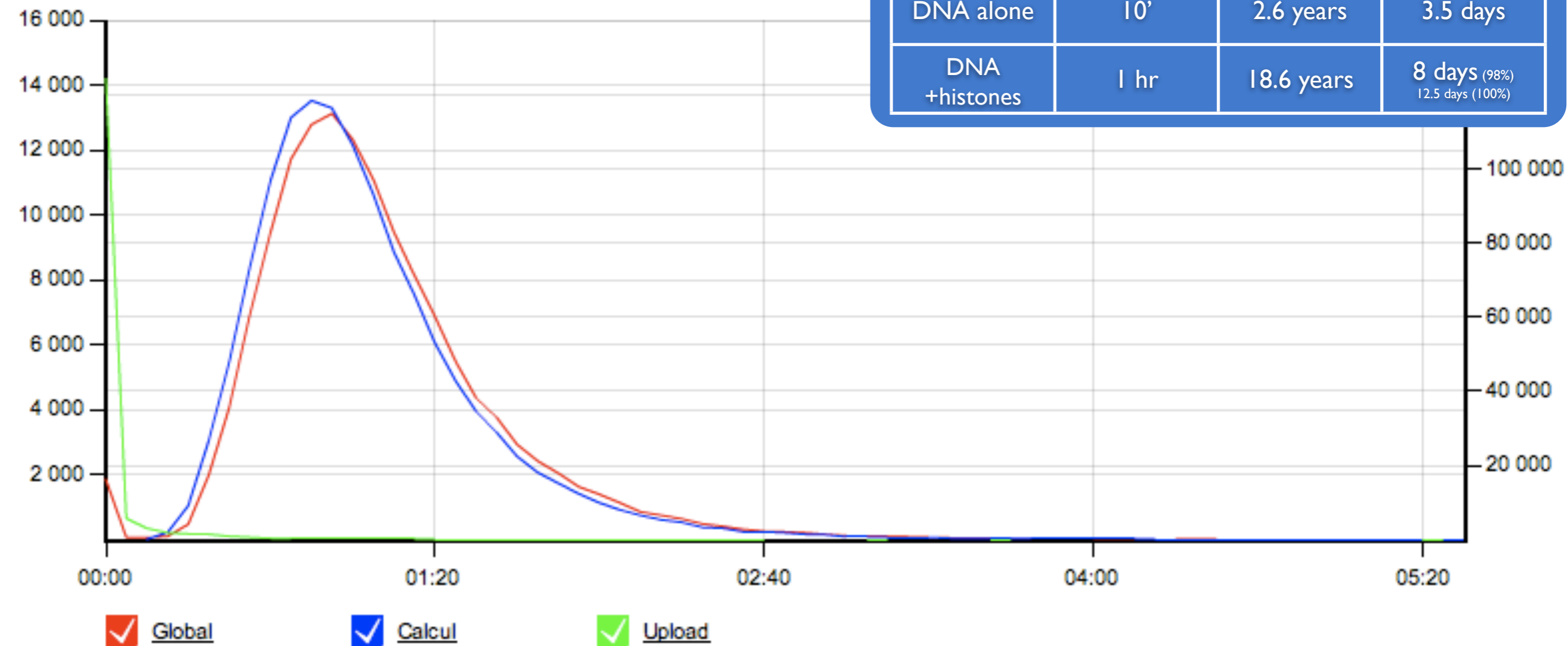
## Nucleosome Application

- independant tasks
- small input
- software 650 kB
- data 60 kB
- output
- small files
- but numerous, 3 files / job

# DNA+histones: jobs duration

## pentanucleotide fragments

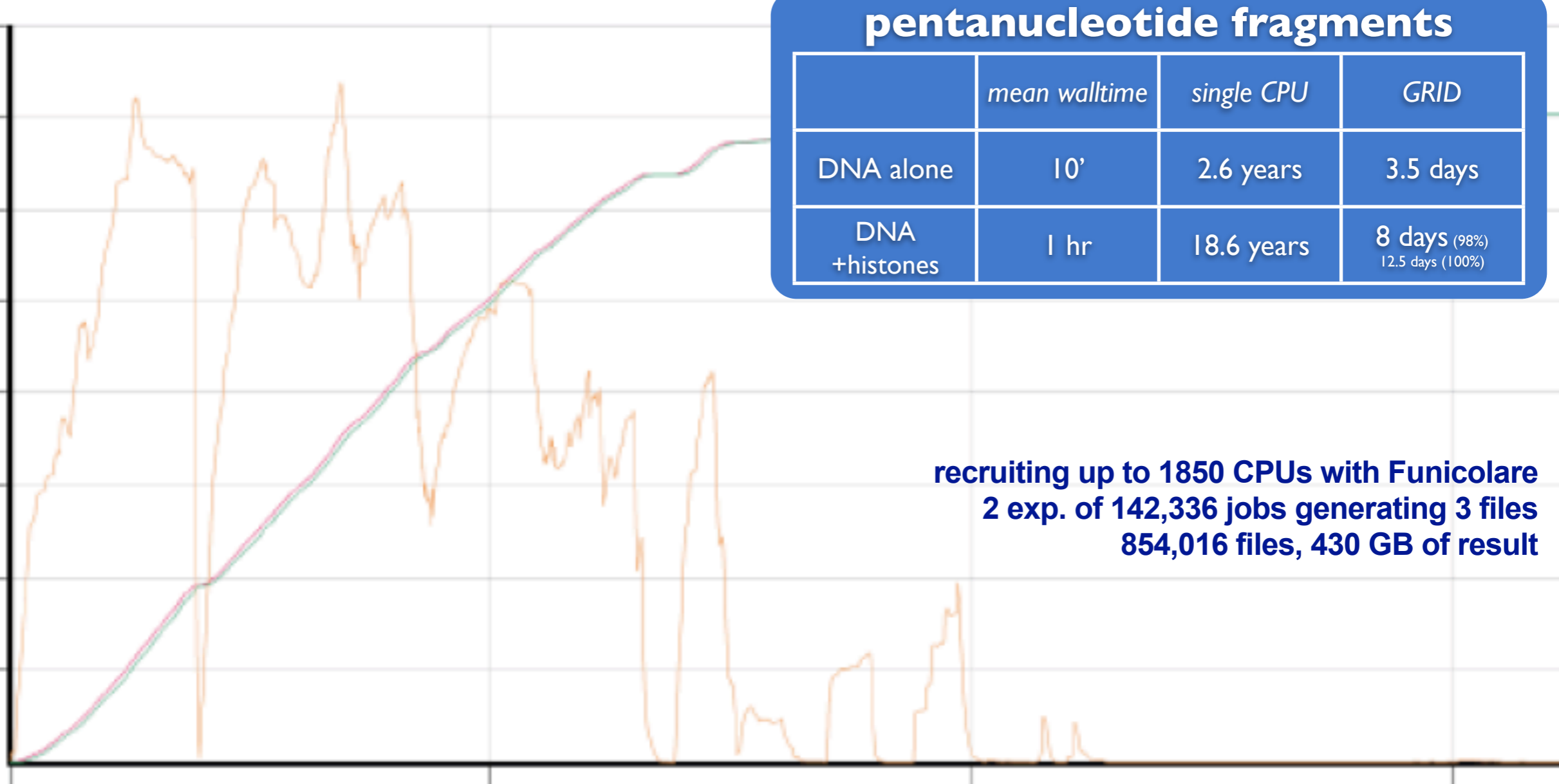
	mean walltime	single CPU	GRID
DNA alone	10'	2.6 years	3.5 days
DNA +histones	1 hr	18.6 years	8 days (98%) 12.5 days (100%)





# DNA+histones: jobs stats

160 000  
140 000  
120 000  
100 000  
80 000  
60 000  
40 000  
20 000



## pentanucleotide fragments

	mean walltime	single CPU	GRID
DNA alone	10'	2.6 years	3.5 days
DNA +histones	1 hr	18.6 years	8 days (98%) 12.5 days (100%)

recruiting up to 1850 CPUs with Funicolare  
2 exp. of 142,336 jobs generating 3 files  
854,016 files, 430 GB of result

Wed, 09 Apr 2008 20:47:38 +0200 Sun, 13 Apr 2008 20:12:38 +0200 Thu, 17 Apr 2008 19:37:38 +0200 Mon, 21 Apr 2008 19:02:38 +0200



Launched



Running



Finished



# Results

## Grid added value

<i>Uplet</i>	<i>Single CPU</i>	<i>GRID</i>
N=4	3.5 years	4 days
N=5	21.2 years	11.5 days

- The analysis depends linearly on the length of the binding site
- Complexes as large as the nucleosome can be treated
- Perspectives
  - Quantify the optimal positions of nucleosomes within the chromosomes of the human genome

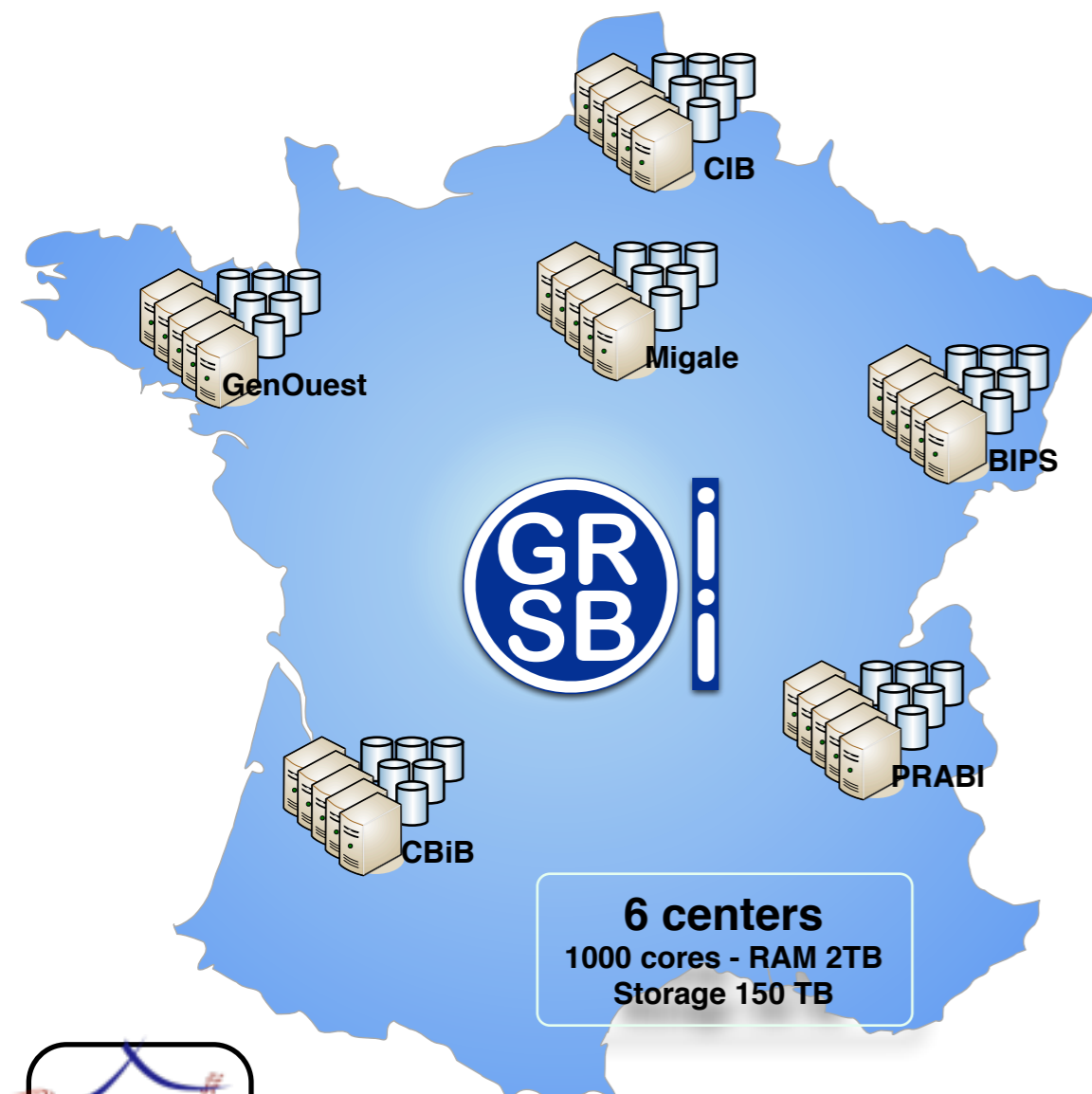


# GRISBI

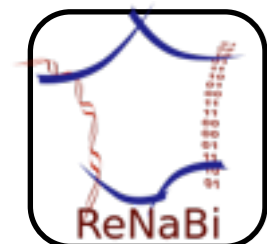
- Grid Support to Bioinformatics -

Make possible challenging bioinformatics applications dealing with large scale biological systems

- National Production infrastructure
  - RENABI, IBISA 2008-2010, Institut des Grilles 2009-2010
- 6 centers from RENABI
  - PRABI, MIGALE, GenOuest, CBIB Bordeaux, BIPS, CIB
  - 8 sites, with 7 CNRS institutes **IBCP Lyon, SBR Roscoff, CBiB Bordeaux, CIB Lille, IRISA Rennes, LBBE Lyon, MIGALE Jouy-en-Josas, BIPS Strasbourg**
  - 40 participants
- Computig resources
  - 1200 cores, 220 TB storage



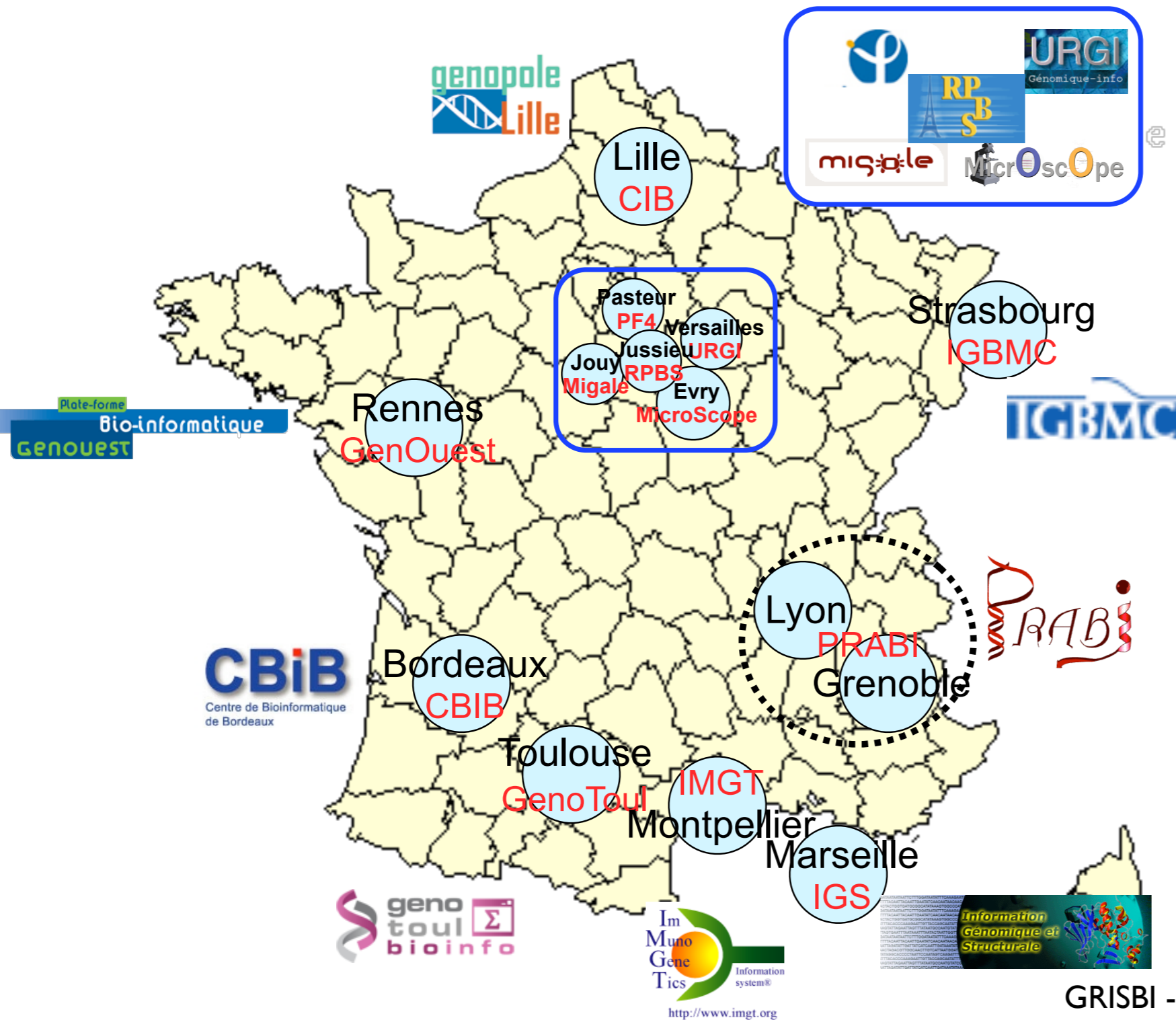
© RENABI GRISBI 2009 - [www.grisbio.fr](http://www.grisbio.fr)





# RENABI

National Network of Bioinformatics Centers



**Since 2004**  
**13 national centers**  
**(RIO/IBISA)**  
**Bioinformatics**  
**Services to the**  
**community**

**Coordinator:**  
**Dr Claudine Medigue**

**[www.renabi.fr](http://www.renabi.fr)**



# Organization

## Scientific Committee

Claudine MEDIGUE (RENABI & IBISA)  
Antoine De DARUVAR (CBiB & ELIXIR)  
Gilbert DELEAGE (CNRS INSB)  
Christian GAUTIER (PRABI)  
Jean-François GIBRAT (MIGALE)  
Bernard KLOAREG (GenOuest.)  
Richard LAVERY (PRABI.)  
Jacques NICOLAS (GenOUest)  
Olivier POCH (BIPS)  
El Ghazali TALBI (CIB)  
Christophe BLANCHET (Resp. Scient.)  
Christophe CARON (Resp. Tech.)

## Technical Committee

Christophe BLANCHET (Resp. Scient.)  
Christophe CARON (Resp. Tech.)  
Clément GAUTHEY (Engineer GRISBI)

## Participants from the centers

Jean-Claude CHARR (CIB)  
Stéphane DELMOTTE (PRABI)

## Steering Committee

Christophe BLANCHET (Resp. Scient.)  
Christophe CARON (Resp. Tech.)  
Olivier COLLIN (GenOuest)  
Tiphaine MARTIN (CBIB)  
Nouredine MELAB (CIB)  
Frédéric PLEWNIAK (BIPS)  
Franck SAMSON (MIGALE)  
Bruno SPATARO (PRABI)

Christelle ELOTO (PRABI)  
Daniel JACOB (CBIB)  
Tiphaine MARTIN (CBIB)  
Alexis MICHON (PRABI)  
Serge UGE (BIPS)  
Aurélien ROULT (GenOuest)



# GRISBI Collaborations

- Contact with French initiatives
  - Institut des Grilles CNRS UPS3107 (production grid)
  - GENCI (supercomputing coordination)
  - Grid5000 (computing Science grid)
- Participation to the biomedical survey of the Institut des Grilles - 420 answers - May 2008
- Participation to the Wh. Pap. of the Institut des Grilles - March 2009
- Participation to the report of the initiative “Pensez Pétaflops” - April 2009
- Collaborate with international GRID
  - EU FP7 EGEE (production grid), StratusLab (cloud computing)
  - Italy (LIBI), Sweden (BILS), Spain & Portugal (IBERGRID, INB), Switzerland (SWING), ...



Source: CNRS IDRIS



Source: ETH CSCS



Programme de Biograle 2009 - GenOuest Bioinformatics Platform

http://www.genouest.org/spip.php?article761

vizbi dg dc ip p w m a l t i ta euca bpel CCGH10 cfp

# GenOuest Bioinformatics Platform

Accueil du site La plate-forme Outils disponibles Banques Services proposés Sites hébergés Emplois

Accueil du site > Français > La plate-forme > Programme de Biograle 2009

## Programme de Biograle 2009

mardi 24 novembre 2009, par Anthony Bretaudeau

Les conférences se dérouleront dans la salle Métivier à l'IRISA.

**24 novembre 2009 :**

- ▶ A partir de 14h00 : Accueil des participants

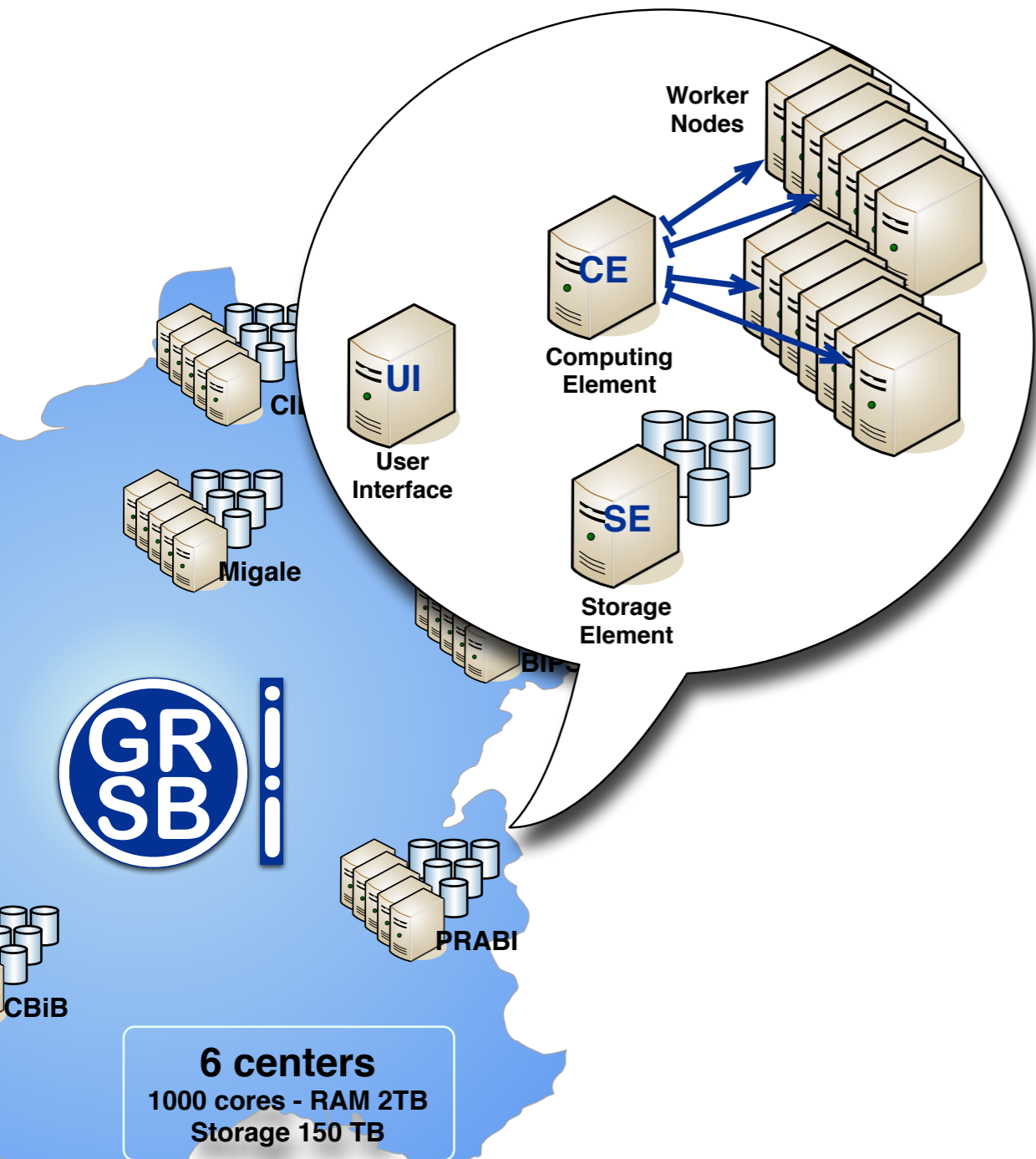
**conférences de 14h15 à 17h00**

- ▶ 14h15-14h45 : Brice Felden (Inserm) : Détection de gènes ARN et de leurs cibles ARN messagers dans le règne procaryote
- ▶ 14h45-15h30 : Krystyna Zakrzewska, Alexis Michon, Christophe Blanchet, Richard Lavery (IBCP Lyon) : Analyse des mécanismes de positionnement du nucléosome sur l'ADN : apports de la grille pour le calcul intensif en Bioinformatique
- ▶ 15h30-16h00 : Pause café - Discussions
- ▶ 16h00-16h30 : S. Penel, P. Calvat, Y. Cardenas (LBBE/PRABI et CC IN2P3) :  
Recherche de similarité de séquences à grande échelle. Calculs "BLAST" intensifs sur la plateforme TIDRA - Traitement de données et Informatique Distribuée en Rhône-Alpes.
- ▶ 16h30-17h00 : David Margery (INRIA Rennes - Bretagne Atlantique) : Le projet Grid 5000
- ▶ Table ronde de 17h00 à 18h00 :  
"Quelles actions à mener pour un passage à l'échelle ? : infrastructures, projets "
- ▶ cocktail de 18h00 à 19h30

**25 novembre 2009 :**



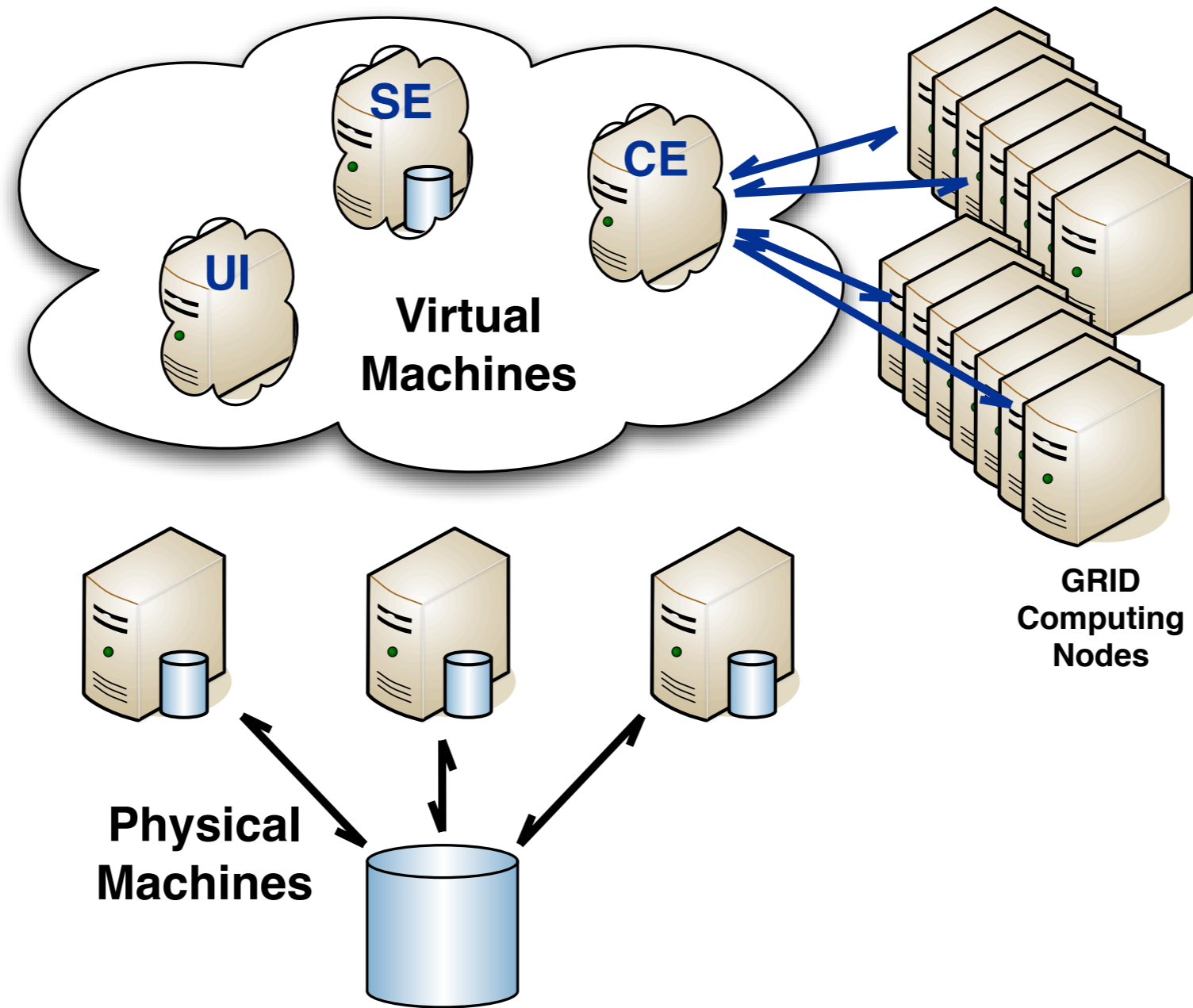
# Grid Infrastructure



- **Mutualize resources**
  - resources of the 6 centers
  - monitoring using «ganglia»
- **Manage Identities**
  - Electronic certificates from the national CNRS PKI : **GRID2-FR**
  - Virtual organization (VO) **vo.renabi.fr** registered in the French ROC.
- **Distribute storage**
  - biomaj, Active Circle, Hadoop, GPFS, XtremFS, SRM, LFC, Isilon FS...
- **Schedule Jobs**
  - gLite, Gridway, Globus, Diet, OpenNebula, ...

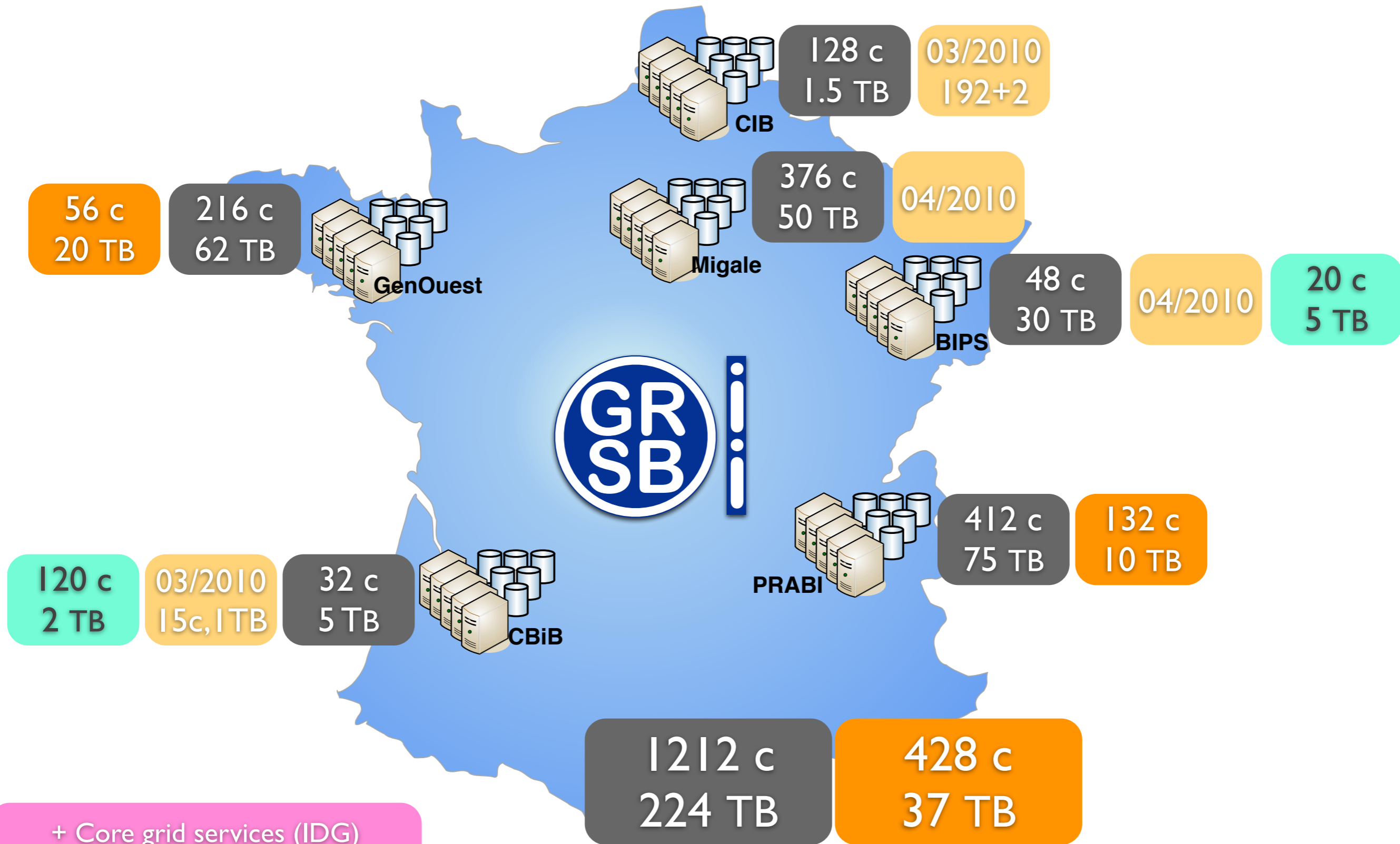


# Virtualisation/Cloud





# GRID Infrastructure

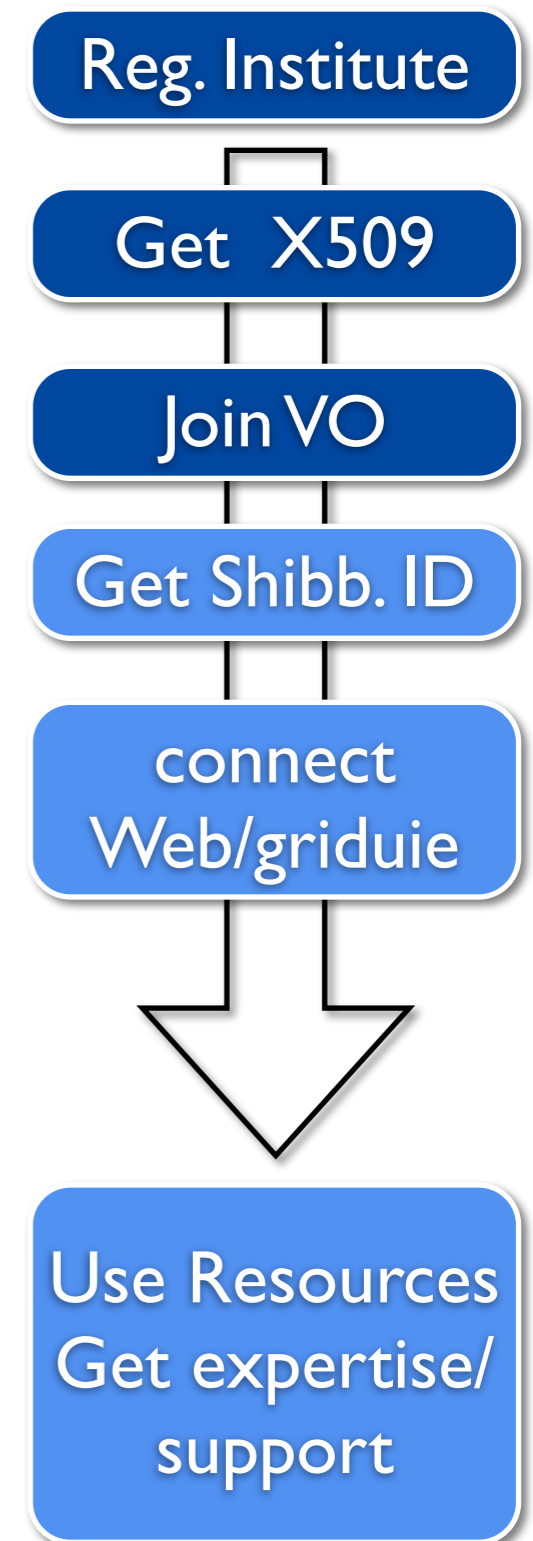


+ Core grid services (IDG)  
WMS (GRIF), LFC&VOMS (CC)



# Tight the Community

- Web: [www.grisbio.fr](http://www.grisbio.fr)
- E-mail
  - [grisbi-tous@services.cnrs.fr](mailto:grisbi-tous@services.cnrs.fr)
  - [grisbi-tech@services.cnrs.fr](mailto:grisbi-tech@services.cnrs.fr)
  - [bioinfo@sfbi.fr](mailto:bioinfo@sfbi.fr)
- Engineers in platforms
- Identity federation
  - certificat X509 / VOMS
  - federation shibboleth
- *and after ...*
- *support, call center, training, ...*





# Ecole GRISBI 2010

Ecole GRISBI 2010

http://www.grisbio.fr/fr/evenements/ecole2010/ Google

dg dc ip p w m a l t i ta euca bpel CCGH10 cfp grisbi embr egee3 vizbi djgo adm projets GBIO ASR

 **Plateforme nationale GRISBI** [fr](#) [en](#)  
*Grille, Support pour la Bioinformatique*

**Rejoignez la communauté GRISBI !**

Accueil Comités **Évènements** Documentation Adhésion  
Ecole GRISBI 2010 BIOGRALE 2009

**Ecole Thématique CNRS**  
Station Biologique de Roscoff  
27 sept. - 01 oct. 2010

Les objectifs sont de diffuser les aspects scientifiques et techniques liés au passage de la Bioinformatique à grande échelle, mutation que notre discipline est en train de réaliser, au sein des laboratoires de recherche en Biologie/Bioinformatique. Les retombées attendues sont de diffuser les compétences liées à l'utilisation de grilles pour la Bioinformatique auprès des plateformes de Bioinformatiques nationales, et auprès des laboratoires de recherche qui se trouve actuellement confrontés à des applications biologiques requérant de grandes ressources informatiques, difficulté auquel les grilles nationales peuvent répondre.

**Inscriptions**  
Ouvertes aux agents CNRS, autres organismes, industriels  
Modalités: *th*



**GRISBI – Calcul Scientifique sur Grille pour la Bioinformatique**

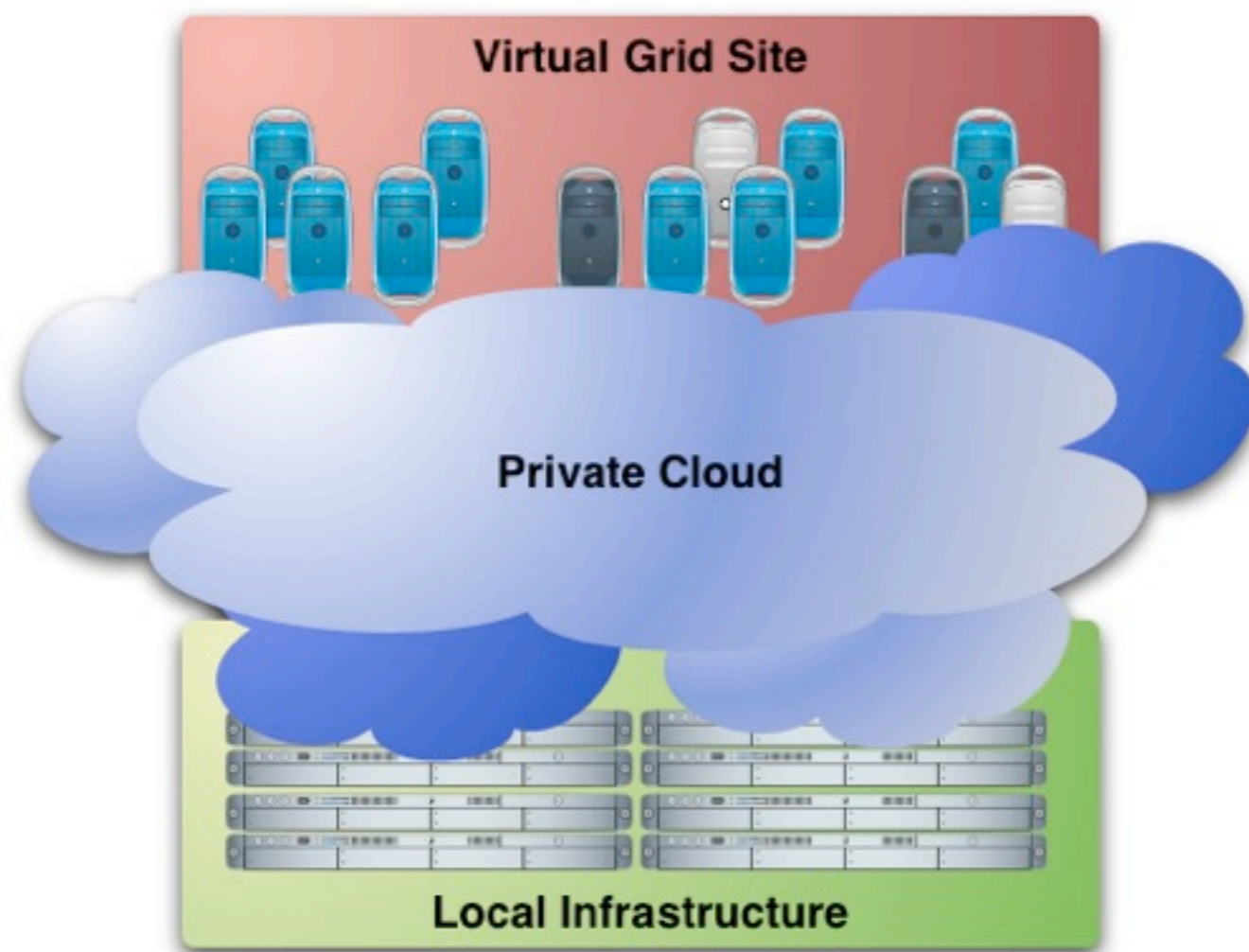
**Préparer l'Infrastructure de Recherche pour la Bioinformatique de demain !**

La Bioinformatique requiert maintenant des infrastructures de recherches permettant le stockage de grandes quantités de données complexes et hétérogène et le calcul massif pour leur analyse. Dans ce contexte nous pouvons tirer un avantage certain des nouveaux développements en

[www.grisbio.fr](http://www.grisbio.fr)

## Key Concept 1: Cloud Computing for Resource Provisioning in Grid Sites

- Integration of private cloud technologies and services into existing grid sites to **transform the local infrastructure into a private cloud** to deploy a virtualized grid site
- **Benefits:** Enhanced flexibility, elasticity, energy efficiency, utility of grid resources. Users would benefit indirectly via the improved stability, reliability, and robustness of the infrastructure.



Perspective of future  
Collaboration  
RENABI GRISBI  
with StratusLab

# CONCLUSIONS

- Several kind of distributed computing infrastructures are available to Bioinformatics usage:
  - Grid, Cloud
  - but also with high level interfaces: Web Services
- Several Bioinformatics experiments have demonstrated the added value of these infrastructures
  - Study the nucleosome
  - and also Large Scale Systems Biology on Grid, Proteins analysis and Structure determination on Clouds
- Perspectives for the French Bioinformatics community with the RENABI GRISBI infrastructure
  - IBISA, ELIXIR, Institut des Grilles

# Acknowledgment

CNRS - Centre National de la  
Recherche Scientifique

University of Lyon 1

GIS IBISA through the project GRISBI  
PF 2008

The European Commission through the  
EU FP7 EGEE III project, contract  
number INFSO-RI-222667.

**CNRS IBCP:** A. Bockmann, E. Bettler, C. Combet, G. Deléage,  
C. Eloto, C. Gauthey, A. Joseph, R. Lavery, A. Loquet, A. Michon, F.  
Penin, K. Zakrzewska

**GRISBI:** and platforms members

**EGEE:** and partners

