Expériences de Bioinformatique sur Grille

Christophe Blanchet

Institut de Biologie et Chimie des Protéines

CNRS IBCP, LYON, FRANCE

christophe.blanchet@ibcp.fr

26 Mars 2010, «Grille et Biologie», IRD, Montpellier

Institute of Biology & Chemistry of Proteins

- Dir. Prof. G. Deléage, http://www.ibcp.fr, LYON, FRANCE
- ~180 people. Associated to CNRS and University of Lyon

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)

3 departments, 14 groups

- Topics such as cancer, extracellular matrix, tissue engineering, membranes, cell transport and signalling, bioinformatics and structural biology

Motivations in Bioinformatics

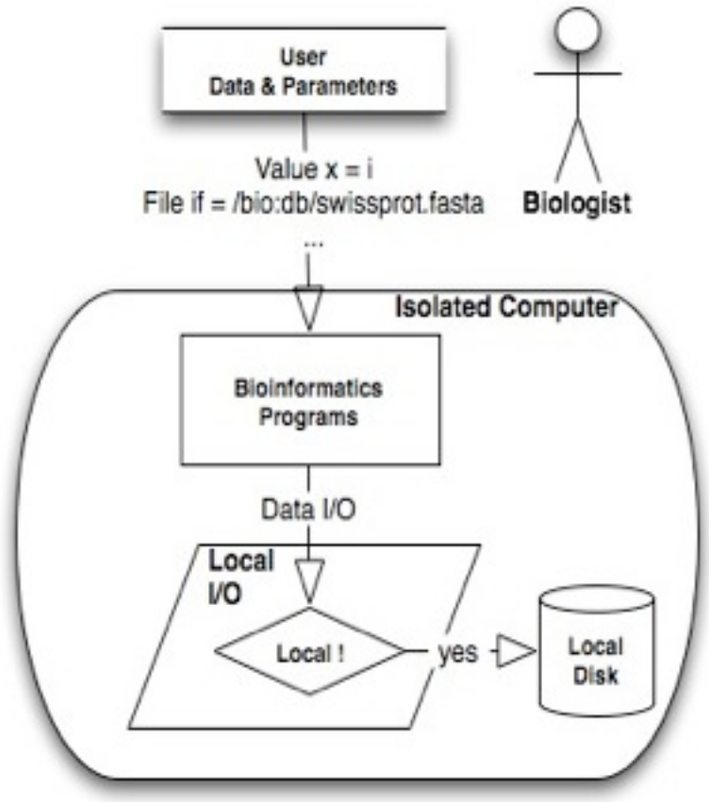
- Deluge of data
- Most common analyses have evolved to a larger scale,
  - from the study of a single gene/protein to a whole genome/proteome, from a single metabolic pathway to Systems Biology.
- Bioinformatics needs large-scale research infrastructures
  - to store very large biological data sets, 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), Pase* (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 alignement …



Transfering data among all these steps

OUTLINE

- GRIDs ...
- Bioinformatics experiments
  - Study the nucleosome
  - Large Scale Systems Biology on Grid
  - Proteins analysis and Structure determination on Clouds
- National Bioinformatics Grid: RENABI GRISBI

GRID...S

Source: CNRS IDRIS

Source: JÜLICH SUPERCOMPUTING CENTRE (JSC)

Source: CSCS

Source: CNRS IBCP

IBCP

Christophe Blanchet, CNRS IBCP

gbio-pbil.ibcp.fr

GRID...S

Source: CNRS IDRIS

Source: JÜLICH SUPERCOMPUTING CENTRE (JSC)

Source: CSCS

EGEE, DEISA, Grid5000,TeraGrid, SwissBioGrid, myGrid, Dgrid, …

=> RENABI GRISBI

=> ELIXIR, EGI, PRACE

ource: CNRS IBCP

Christophe Blanchet, CNRS IBCP

gbio-pbil.ibcp.fr

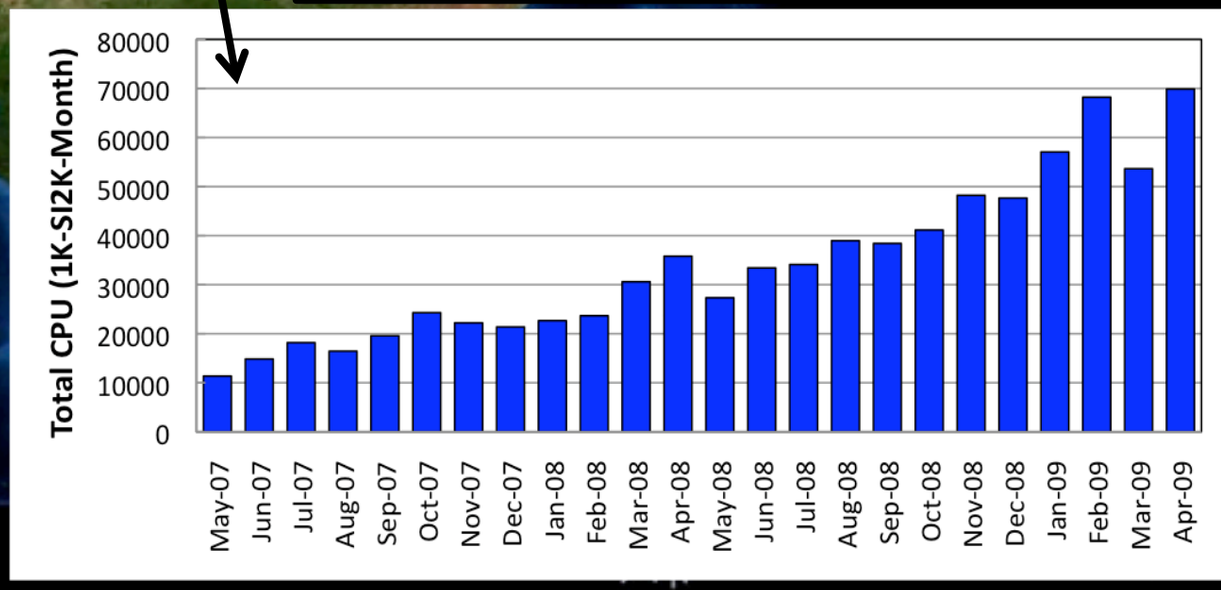
Enabling Grids for E-sciencE

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

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

Main Objectives
Expand/optimise 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

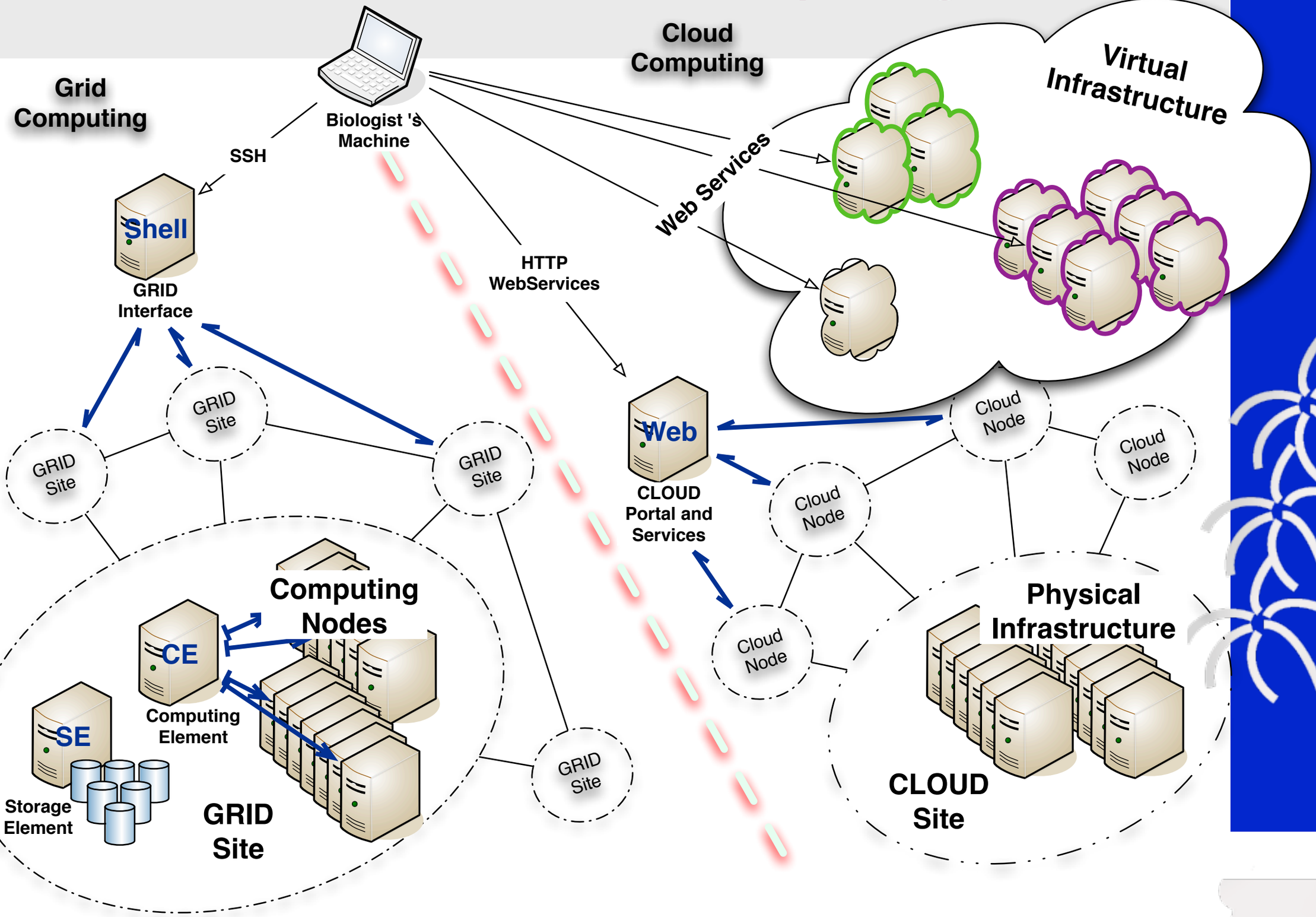
Consistent doubling every 12-18 months.
HEP largest users / contributors
AA/ES/other show strong increase



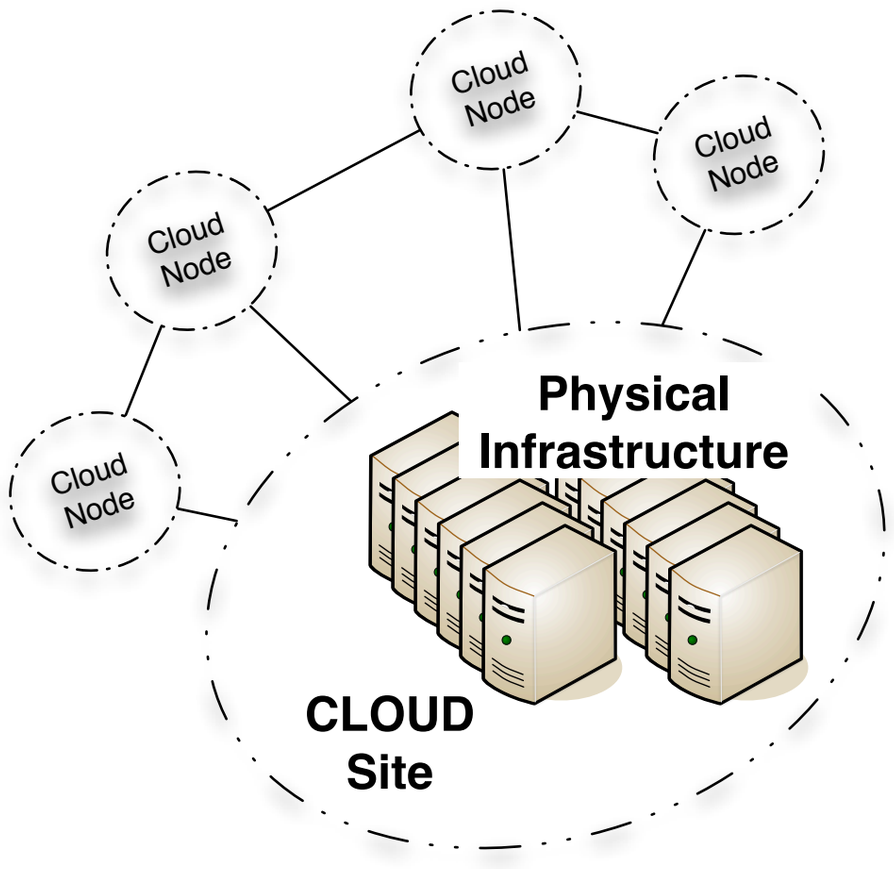
Current initiatives related to BIO

- 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



Cloud Infrastructures



- HIPerNET and Grid'5000
  - 9 sites, 5000 cores
  - HIPerNET 0.6

- IBCP's Eucalyptus cluster
  - 1 site, 40 cores
  - Eucalyptus 1.6.2

Eucalyptus

PAPER www.rsc.org/pccp | Physical Chemistry Chemical Physics

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

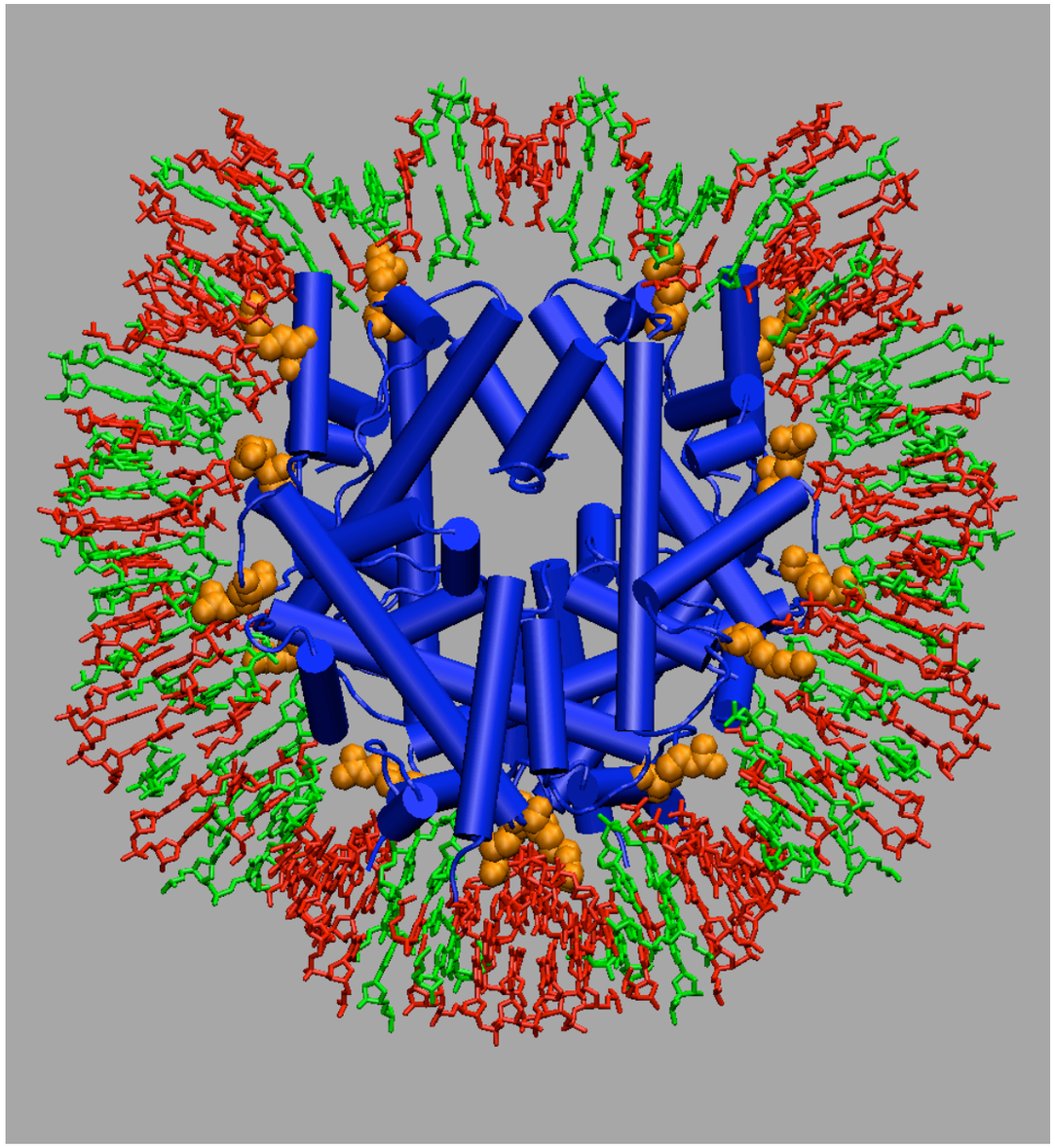
We use a physics-based approach termed ADAPT to analyse the sequence-specific interactions of three proteins which bind to DNA on the side of the minor groove. The analysis is able to estimate the binding energy for all potential sequences, overcoming the combinatorial problem *via* a divide-and-conquer 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

Contact: Richard Lavery (CNRS IBCP)

Nucleosome



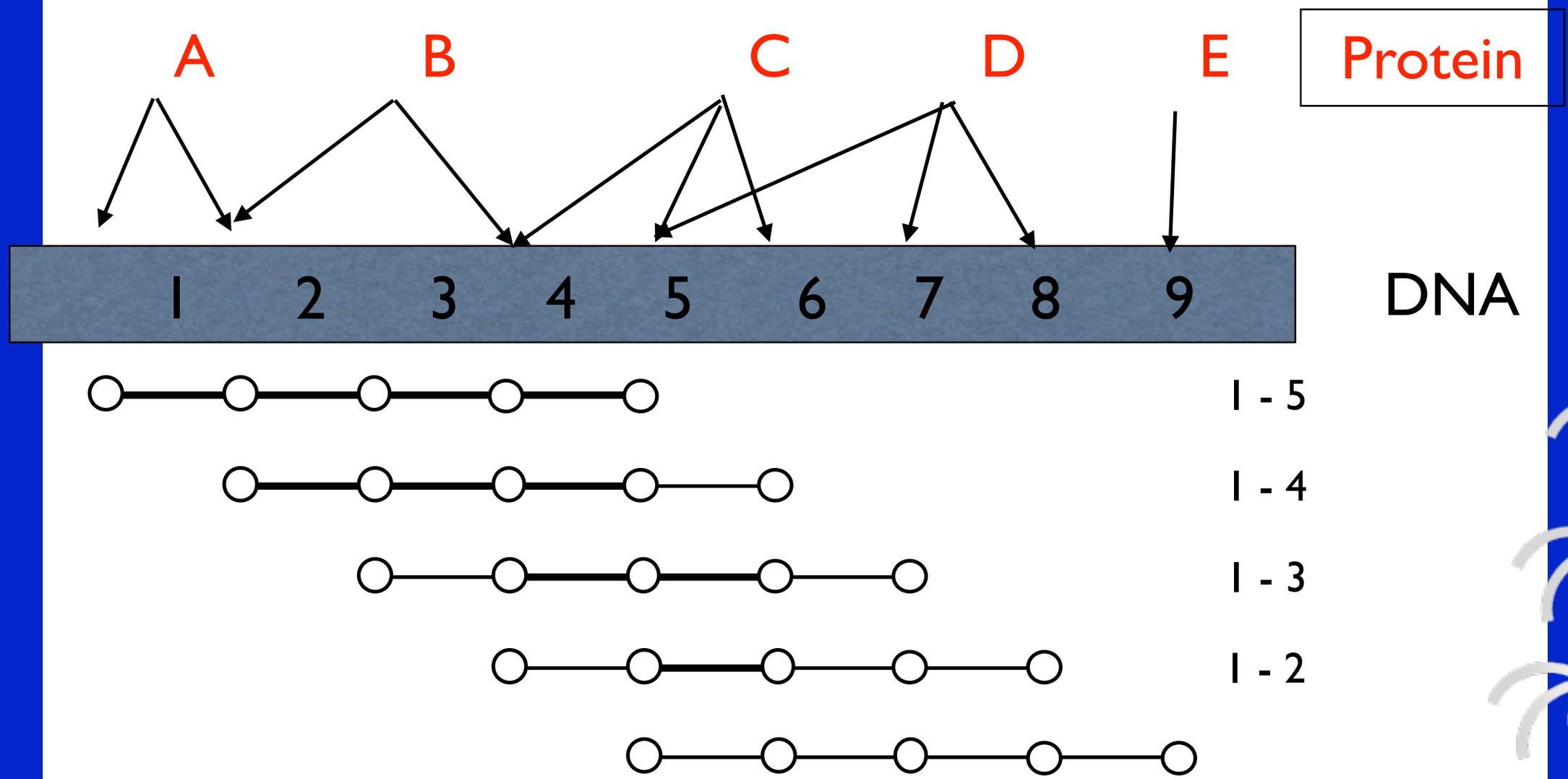
8 proteins, one DNA fragment
147 bp, 21,000 atomes

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

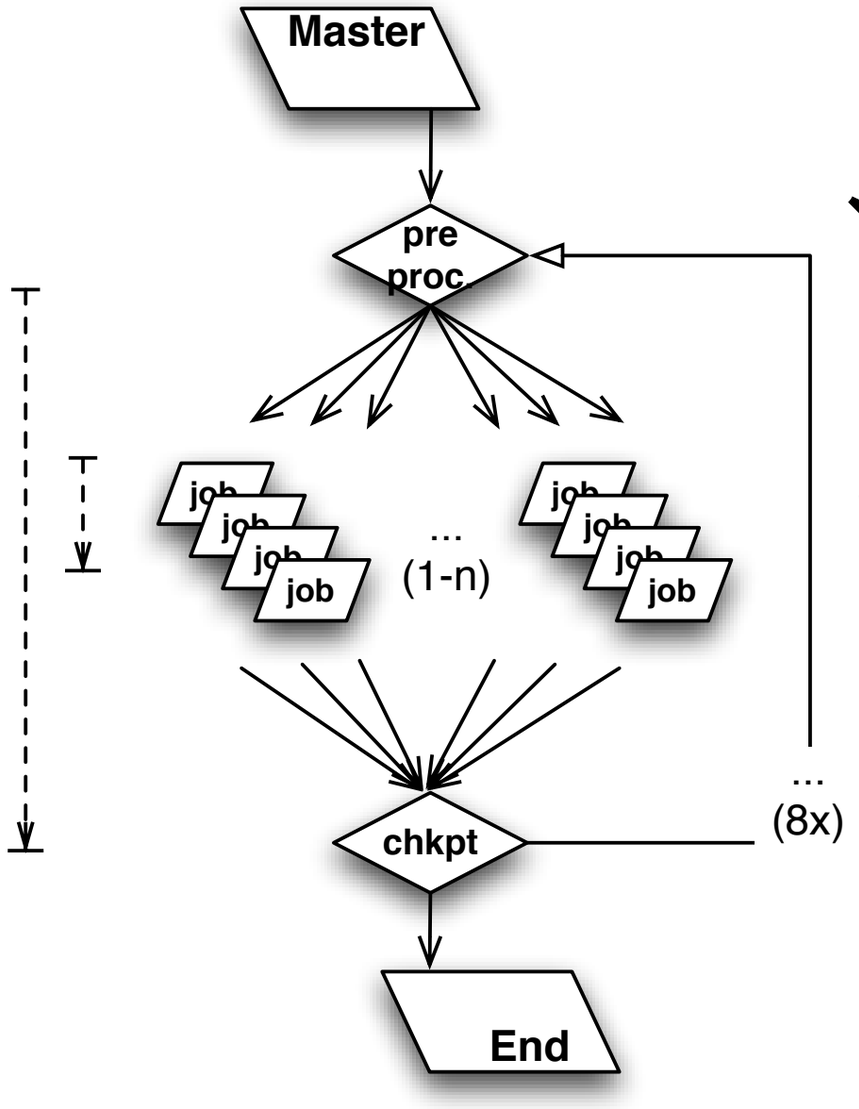
Contact: richard.lavery@ibcp.fr

Fragment of 5-plet



143 bp + 2 x 2 bp ends

Parallelization Model



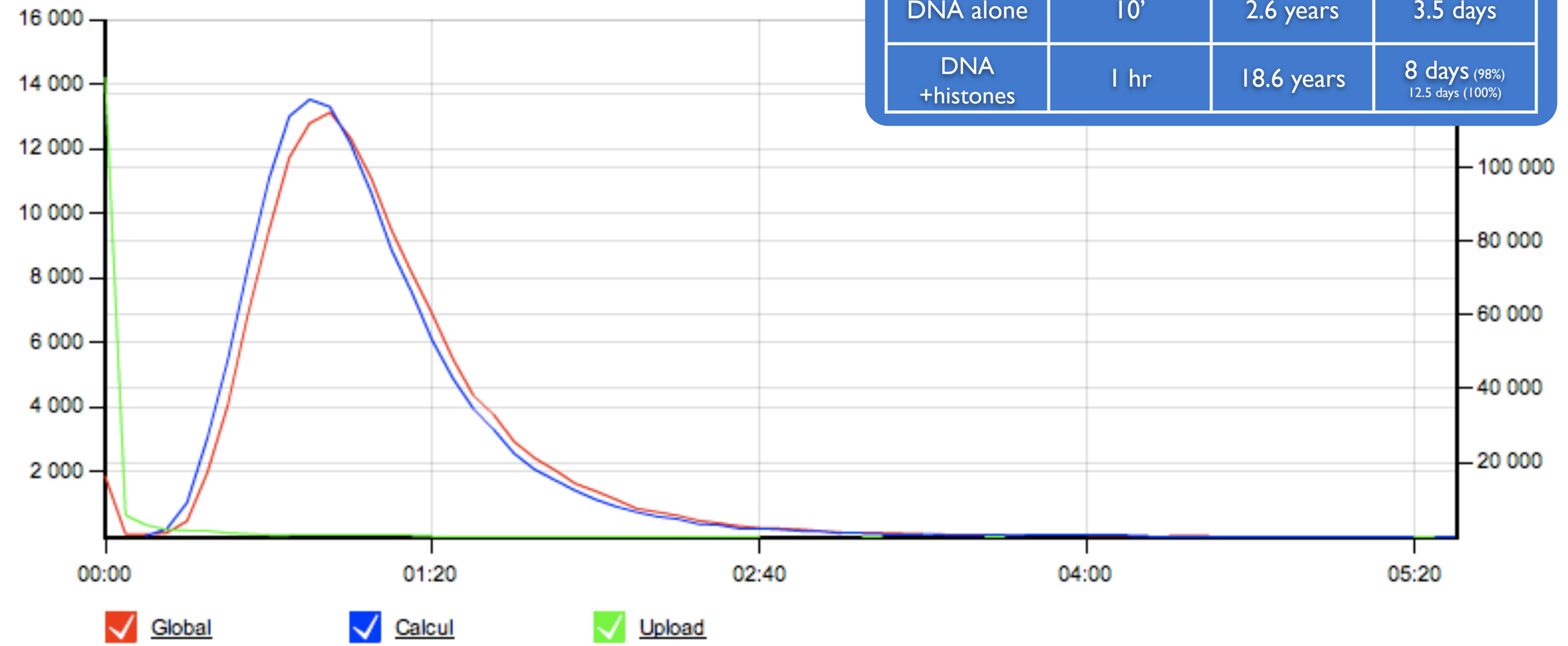
Nucleosome Application

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

DNA+histones: jobs duration

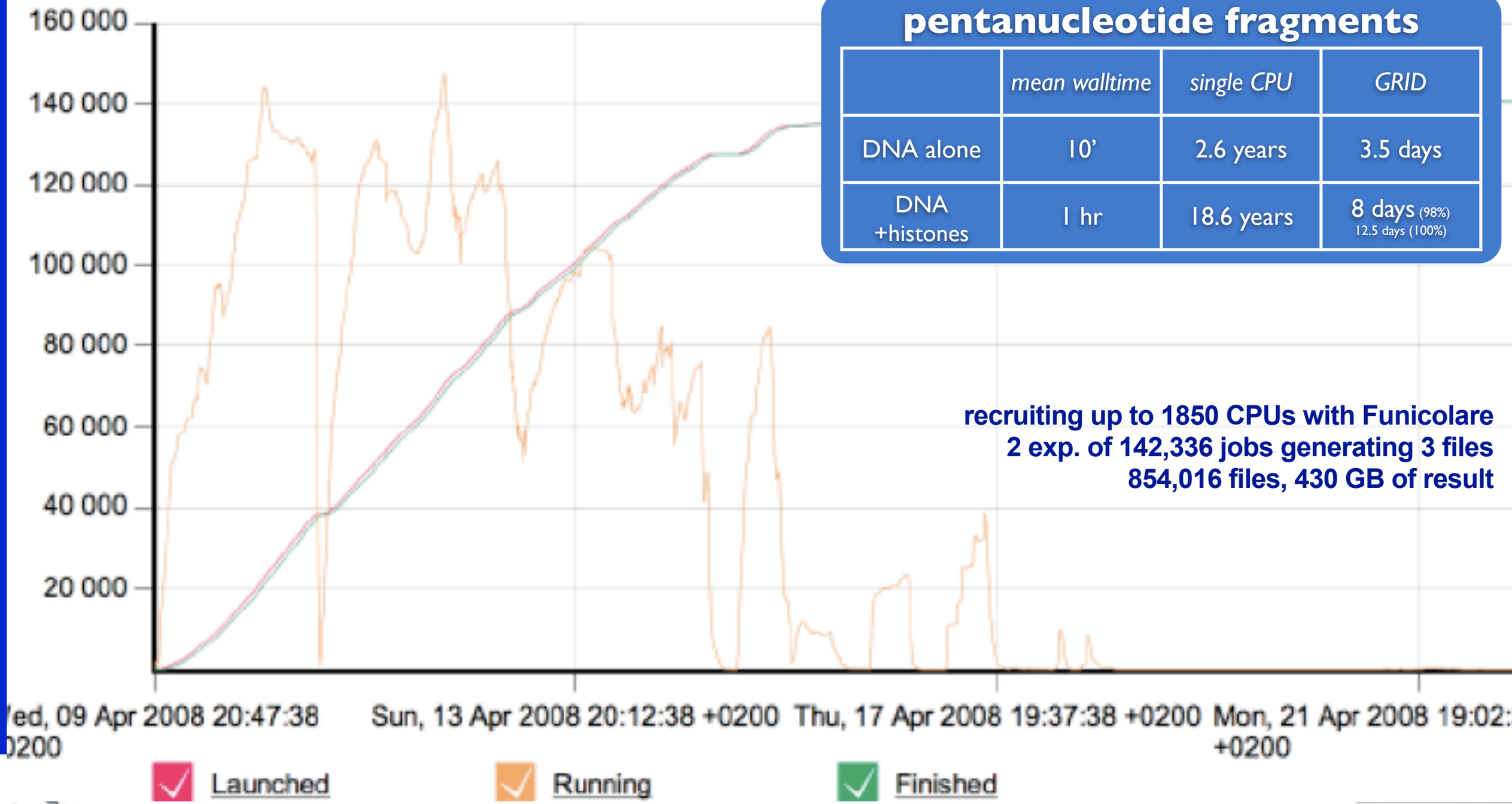
pentanucleotide fragments

	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%)



Christophe Blanchet, CNRS IBCP gbio-pbil.ibcp.fr

DNA+histones: jobs stats



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

Results

Grid added value

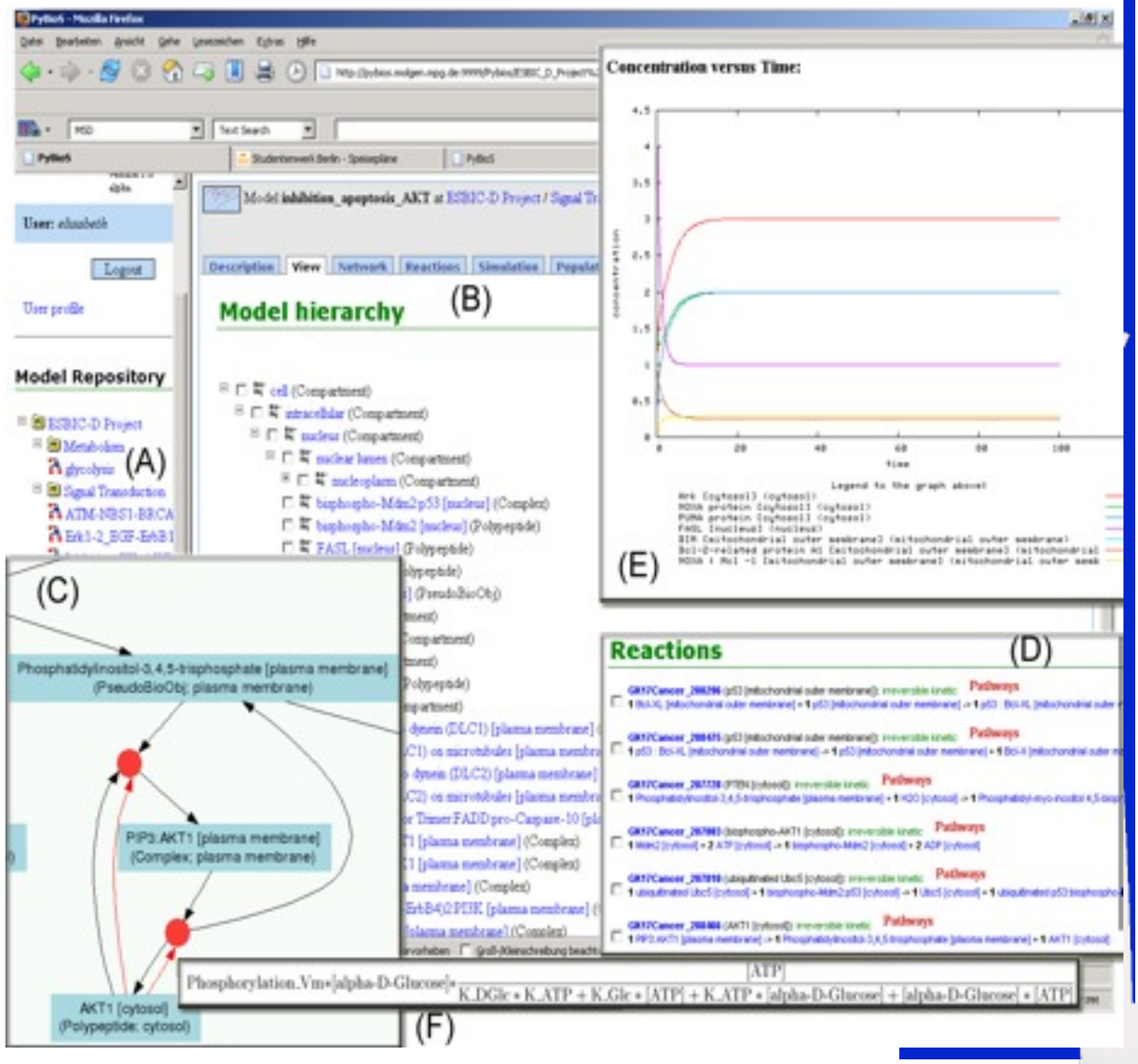
Uplet	*Single CPU*	*GRID*
N=4	3.5 years	4 days
N=5	21.2 years	11.5 days

- 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

Systems Biology on the Grid

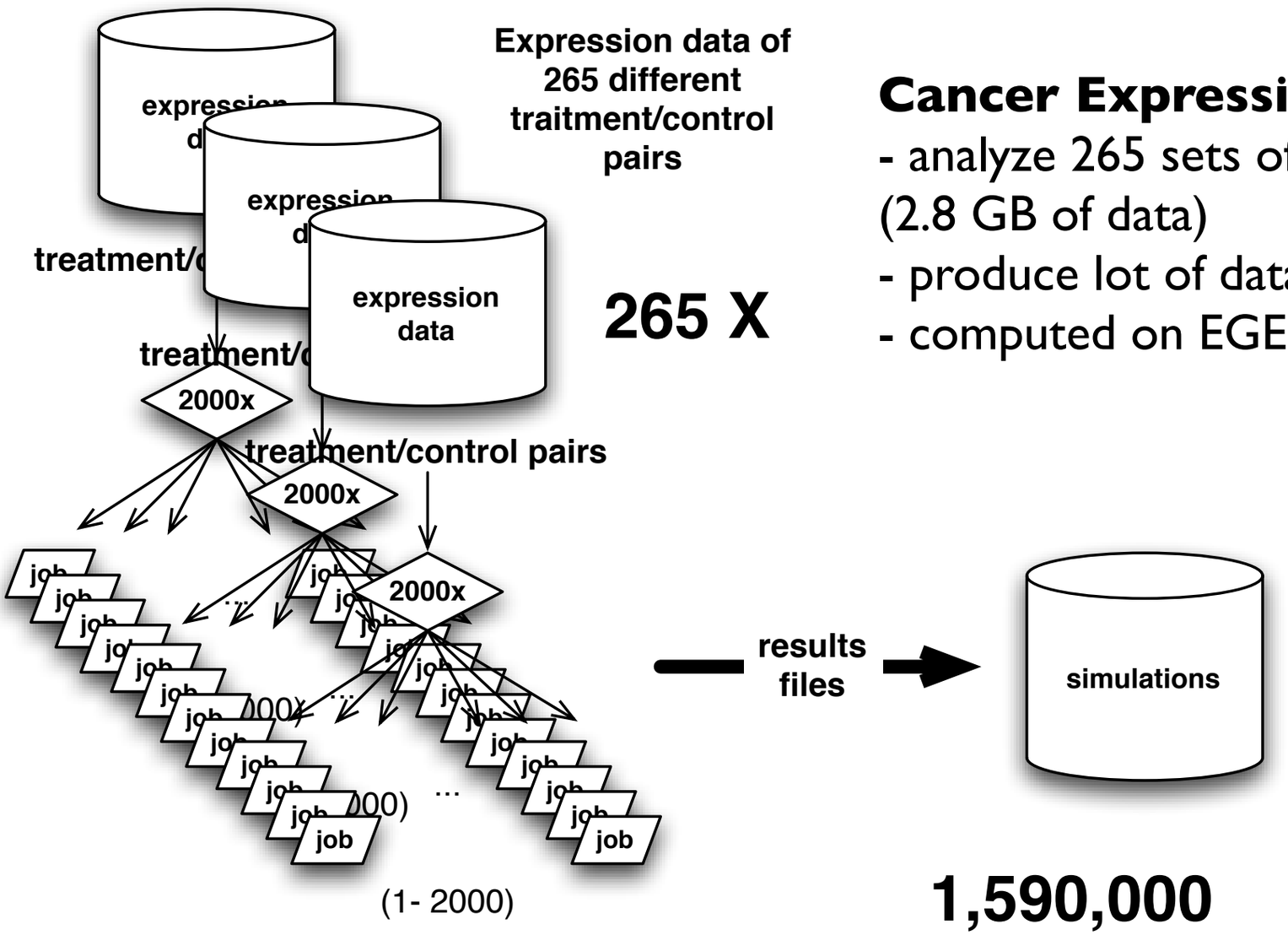
- **Scientific objectives**
  - **Systems Biology: modelling and simulation of biological systems**
  - **Analysis of system behaviour in the context of experimental high-throughput data**
  - **Monte-Carlo simulation on the Grid**
  - **Applications on**
    - **Human melanoma cell-lines**
    - **Type-2 diabetes**
    - **Cancer Expression Data**

PyBioS
(http://pybios.molgen.mpg.de/)



Contact: Ralf Herwig (MPI-MG)

Cancer data experiment

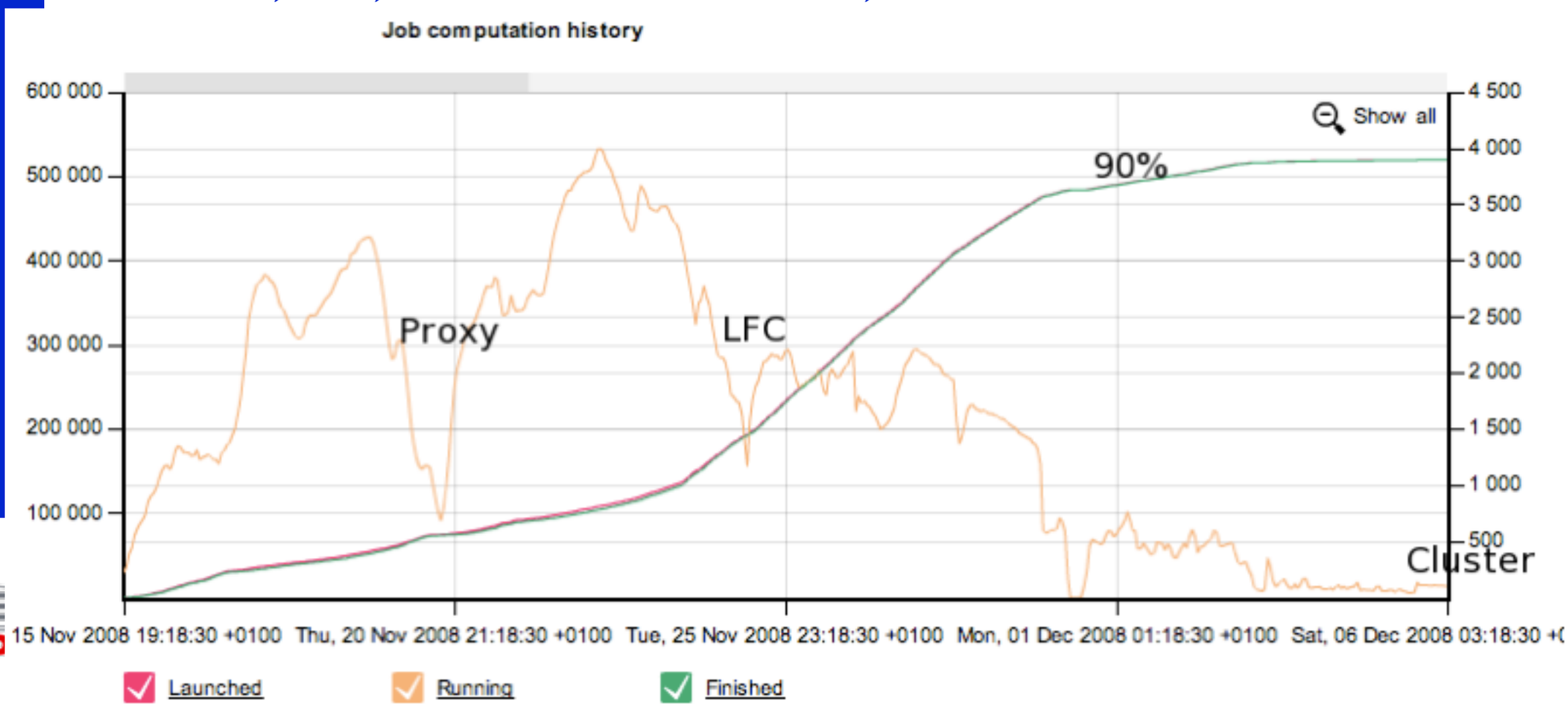


Cancer Expression data

- analyze 265 sets of expression data (2.8 GB of data)
- produce lot of data
- computed on EGEE grid

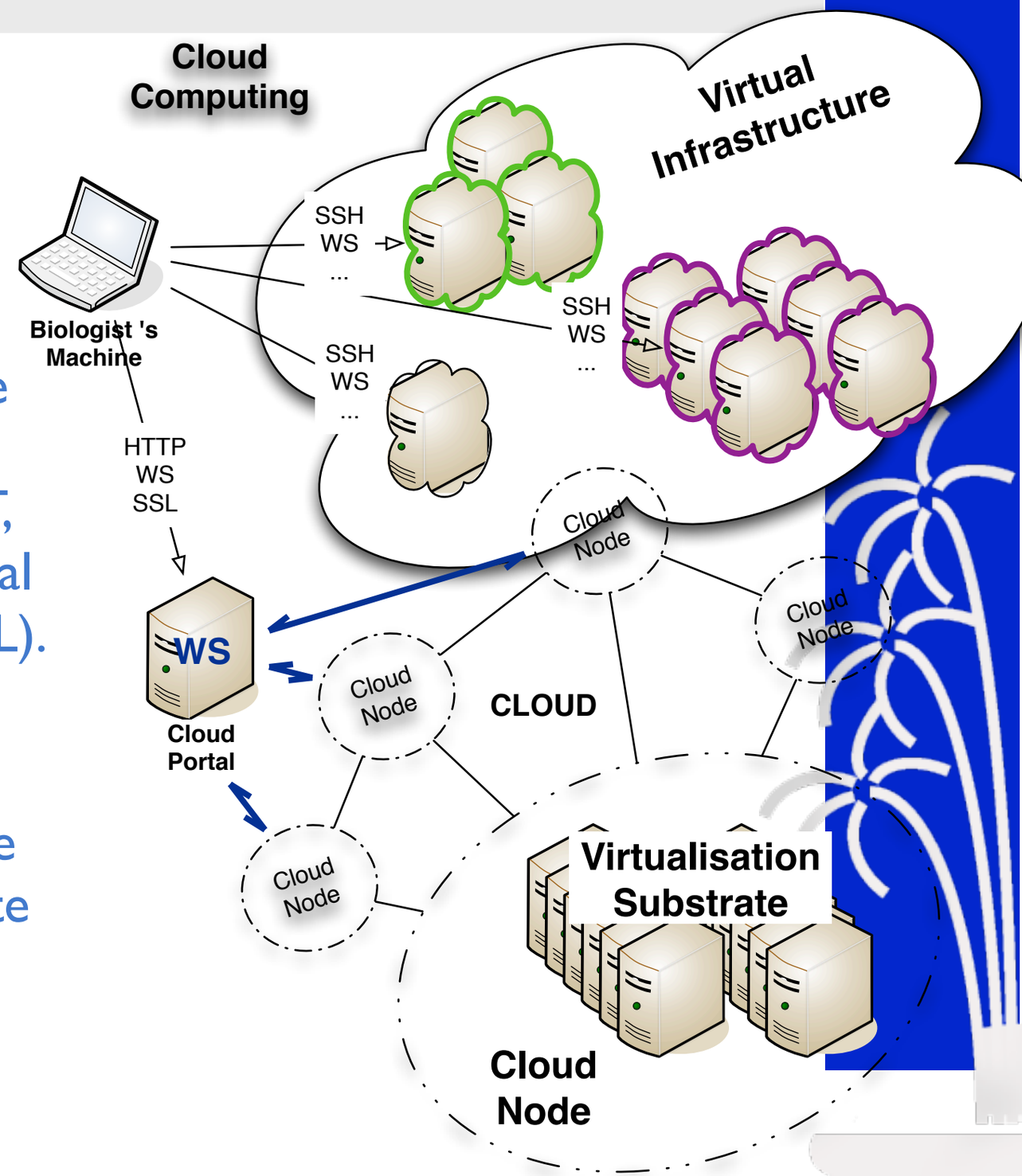
PyBioS on EGEE

- Total time : 102 years (910,000 hours)
  - Vs estimation of 530 Kh, 60 Y
  - ~ 97% after 17 days (2200x)
- Recruiting up to 4,000 CPUs with Funicolare
  - 3,792 WNs identified (! NAT)
- ~ 1,590,000 result files => 1,35 TB of data



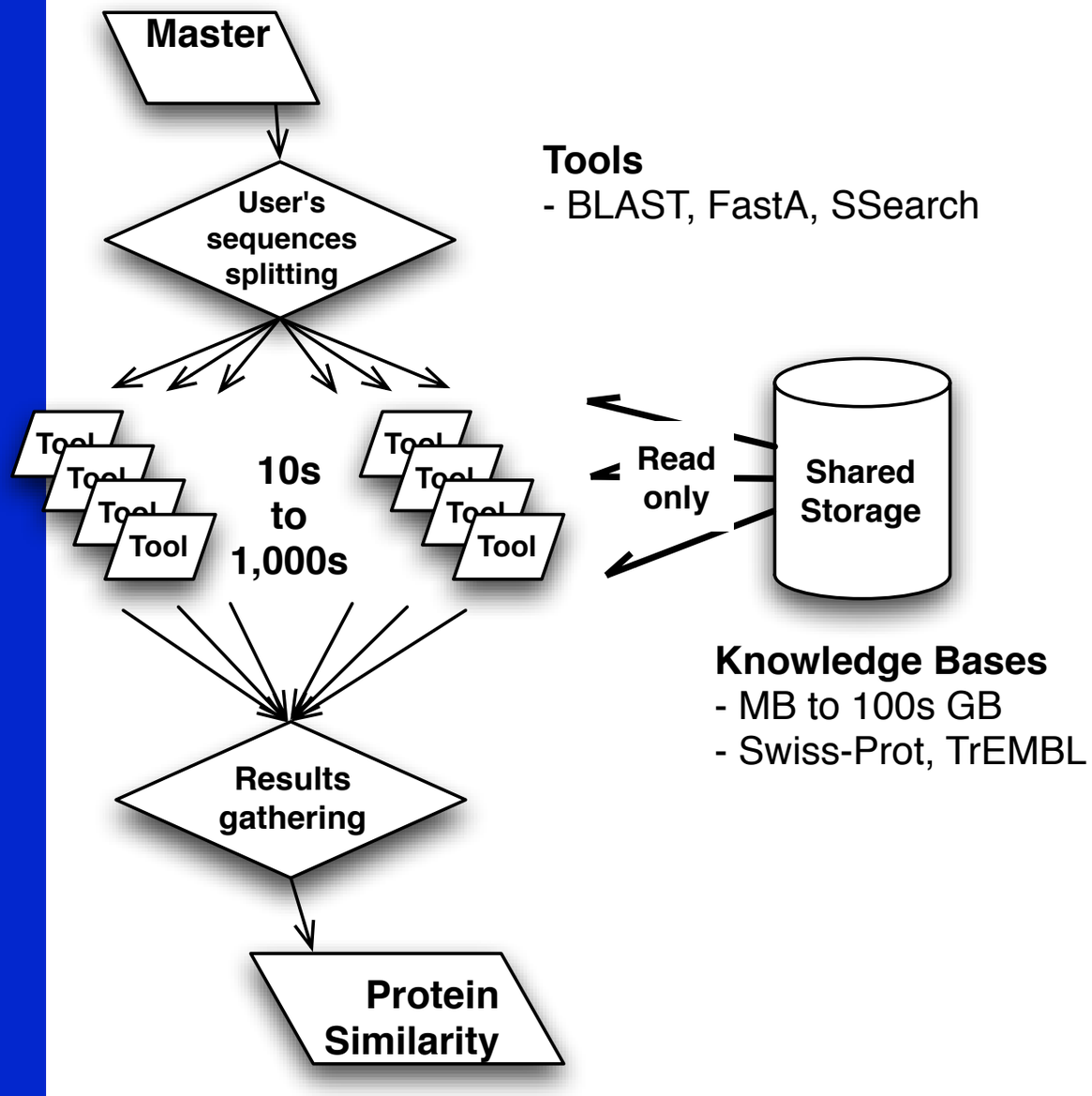
Bioinformatics applications on CLOUDs

- Cloud-enabling two applications
  - "**Proteins**" : analyzing large sets of proteins with bioinformatics tools (BLAST, SSearch, FastA) and biological data (Swiss-Prot and TrEMBL).
  - "**Solid-State NMR**": Automated assignment and structure calculation of large protein systems in solid-state NMR context.



Proteins

Application **Proteins**



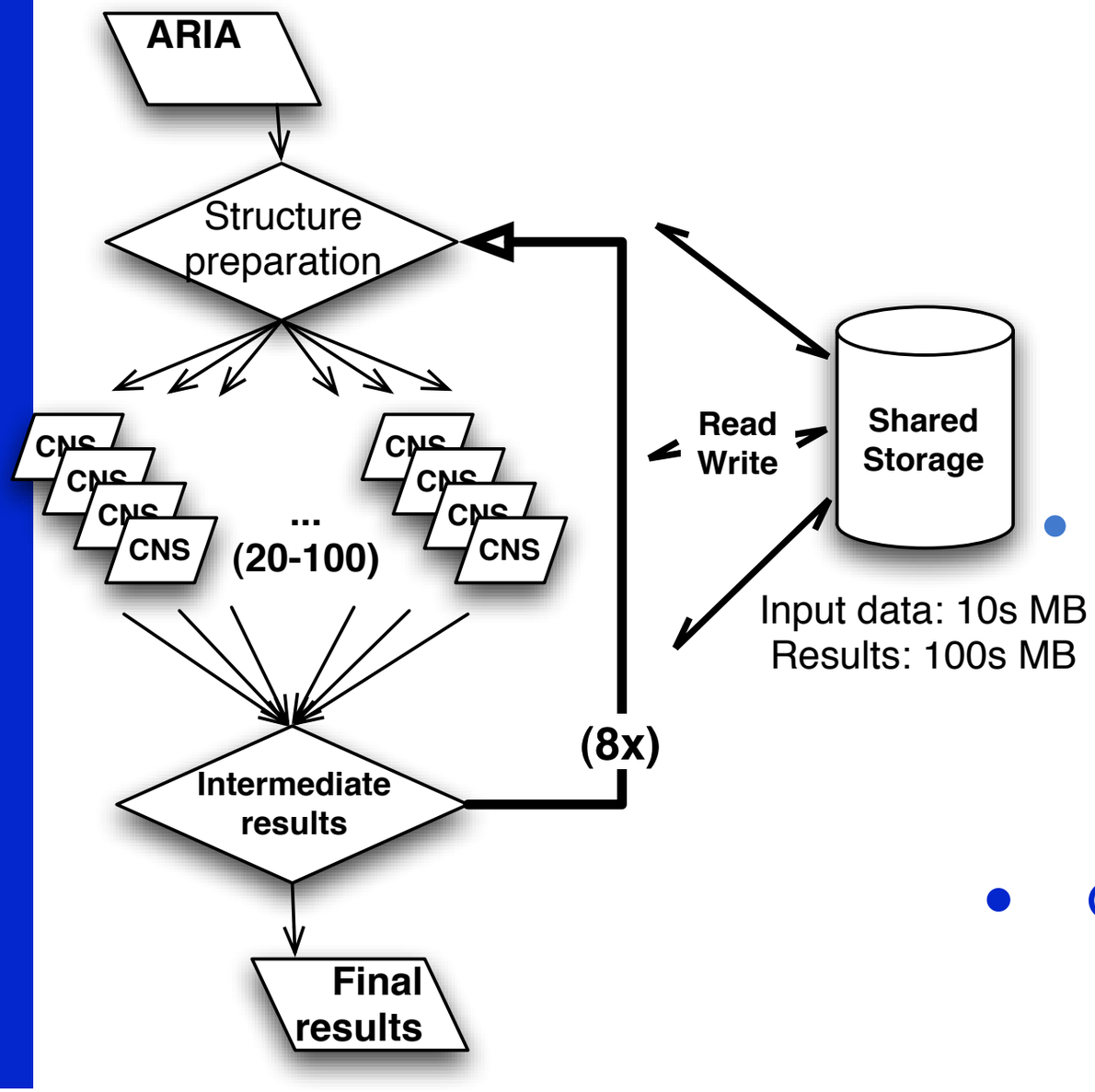
Tools
- BLAST, FastA, SSearch

Knowledge Bases
- MB to 100s GB
- Swiss-Prot, TrEMBL

- Protein sequence analysis at large scale : analyzing complete proteomes.
- Bioinformatics tools
  - use case with common software BLAST, SSearch and FastA,
- Biological data
  - Analyzing large sets of proteins obtained for example from Next Generation Sequencing
  - Using international databases Swiss-Prot and TrEMBL as knowledge bases

Solid-State NMR

Application **Solid-State NMR**

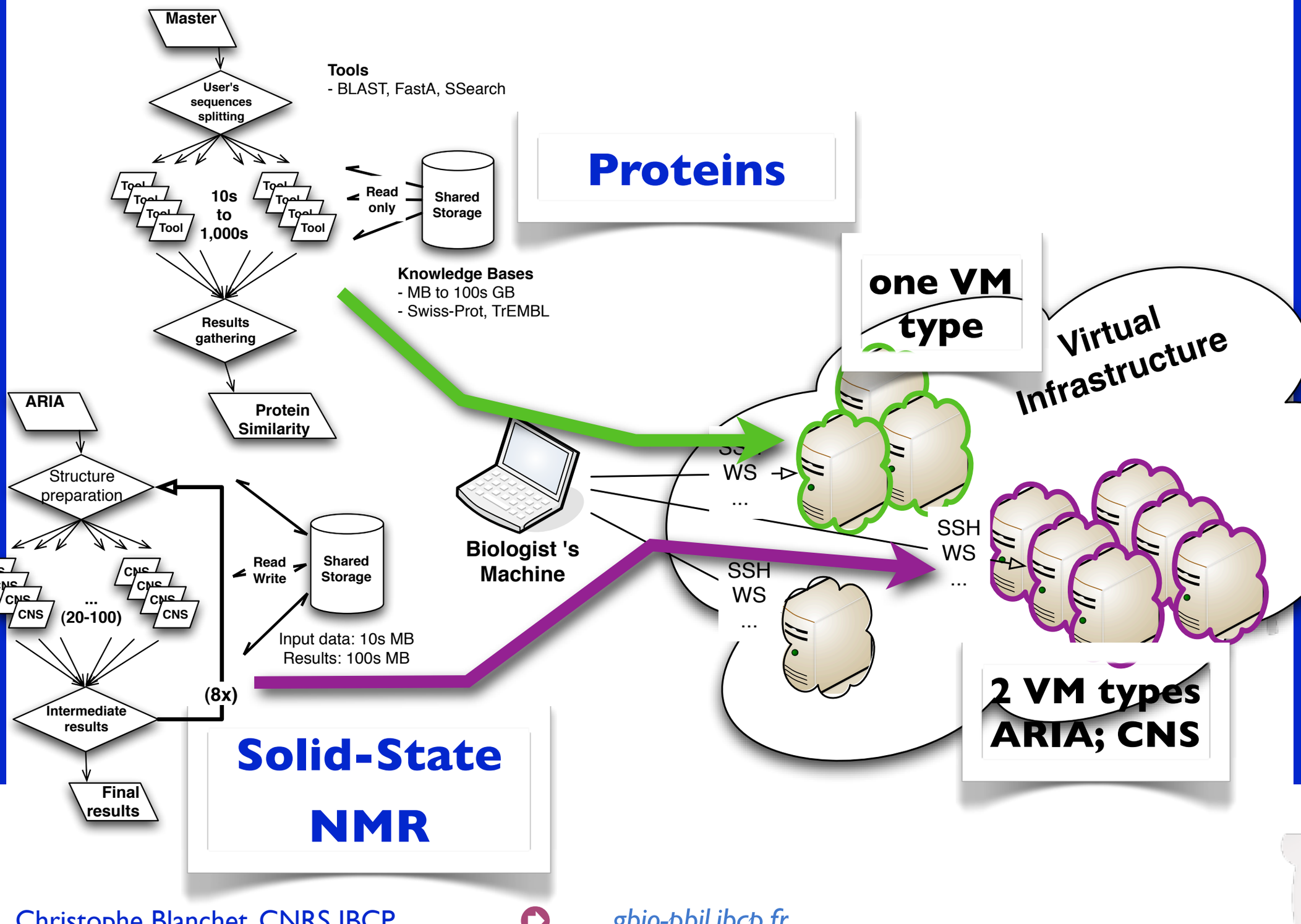


- ARIA (Ambiguous Restraints in Iterative Assignment), a software for automated NOE assignment and NMR structure calculation, speedup the NOE assignment process through the use of ambiguous distance restraints in an iterative structure calculation scheme.
- Rieping W., Habeck M., Bardiaux B., Bernard A., Malliavin T.E., Nilges M. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23, 381-382.
- Contact: Dr Anja Bockmann

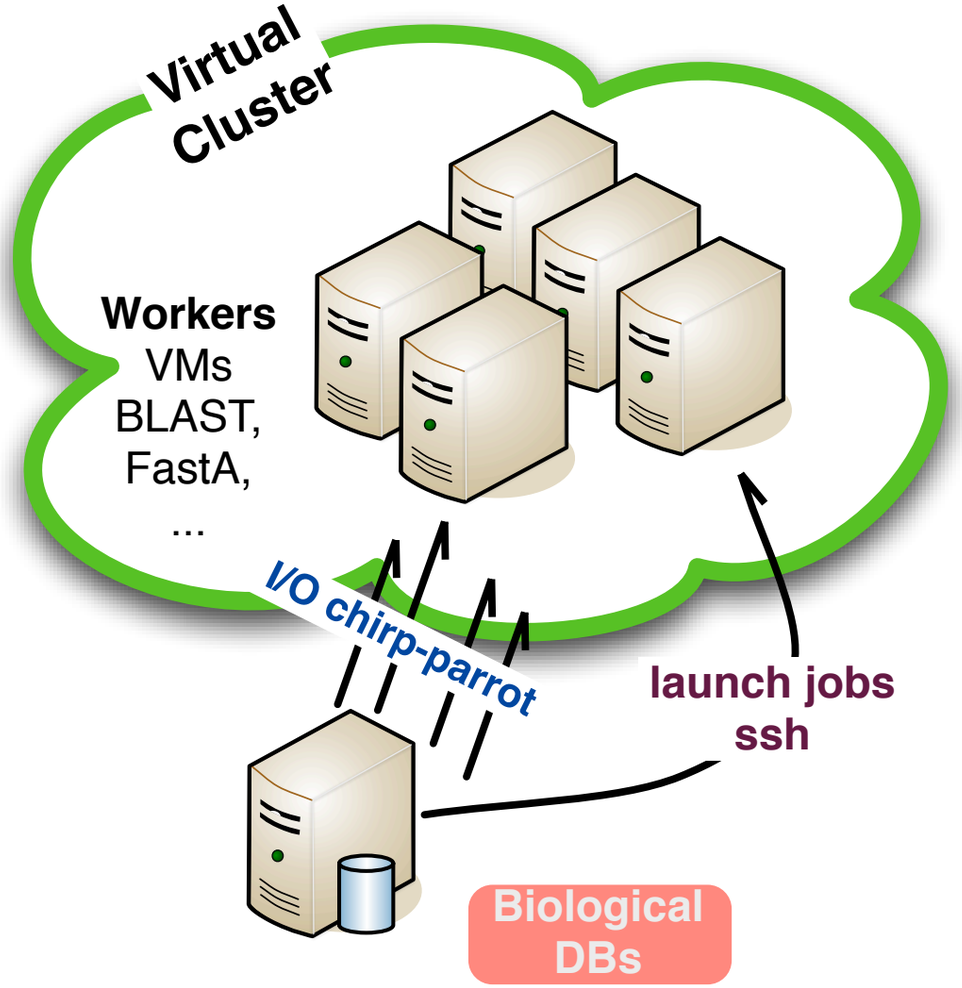
Bioinformatics Apps Requirements

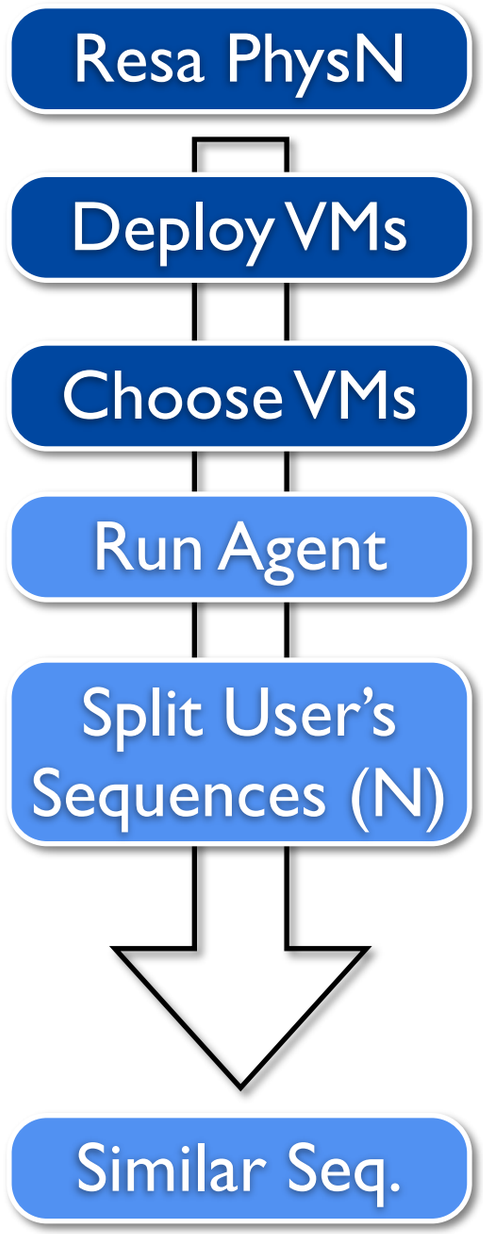
	Proteins	Solid-state NMR
Input data transfers	MB-GB	MB
Output data	MB-GB	GB
Databases (long-term storage)	yes + updates	no
Global FS	yes	yes
Software constraints	legacy, local IO	legacy, ARIA/CNSsolve
Middleware dependency	batch/parallel	batch
Message passing	no/MPI	no
Security	on-disk + data transfer encryption, access control on data	

Deploying on Cloud



Workflow


Virtual Cluster
Workers
VMs
BLAST,
FastA,
...
I/O chirp-parrot
launch jobs
ssh
Biological DBs
Master & Storage
Machine hipcal


Resa PhysN
Deploy VMs
Choose VMs
Run Agent
Split User's Sequences (N)
Run N BLAST (or Fasta, ...)
Similar Seq.

Solid-State NMR

Workflow

Resa PhysN

Deploy VMs

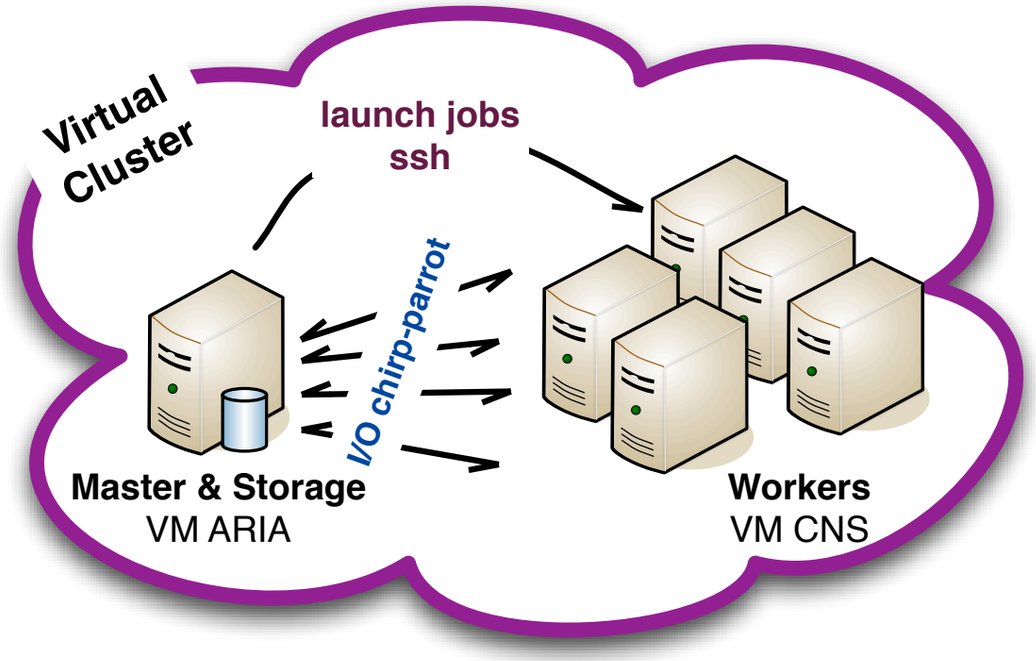
Choose VMs

Run Agent

Launch Chirp

Run Aria

Models



Define constraints

20-100 struct.

Calculate Struct. (CNS)

8 steps

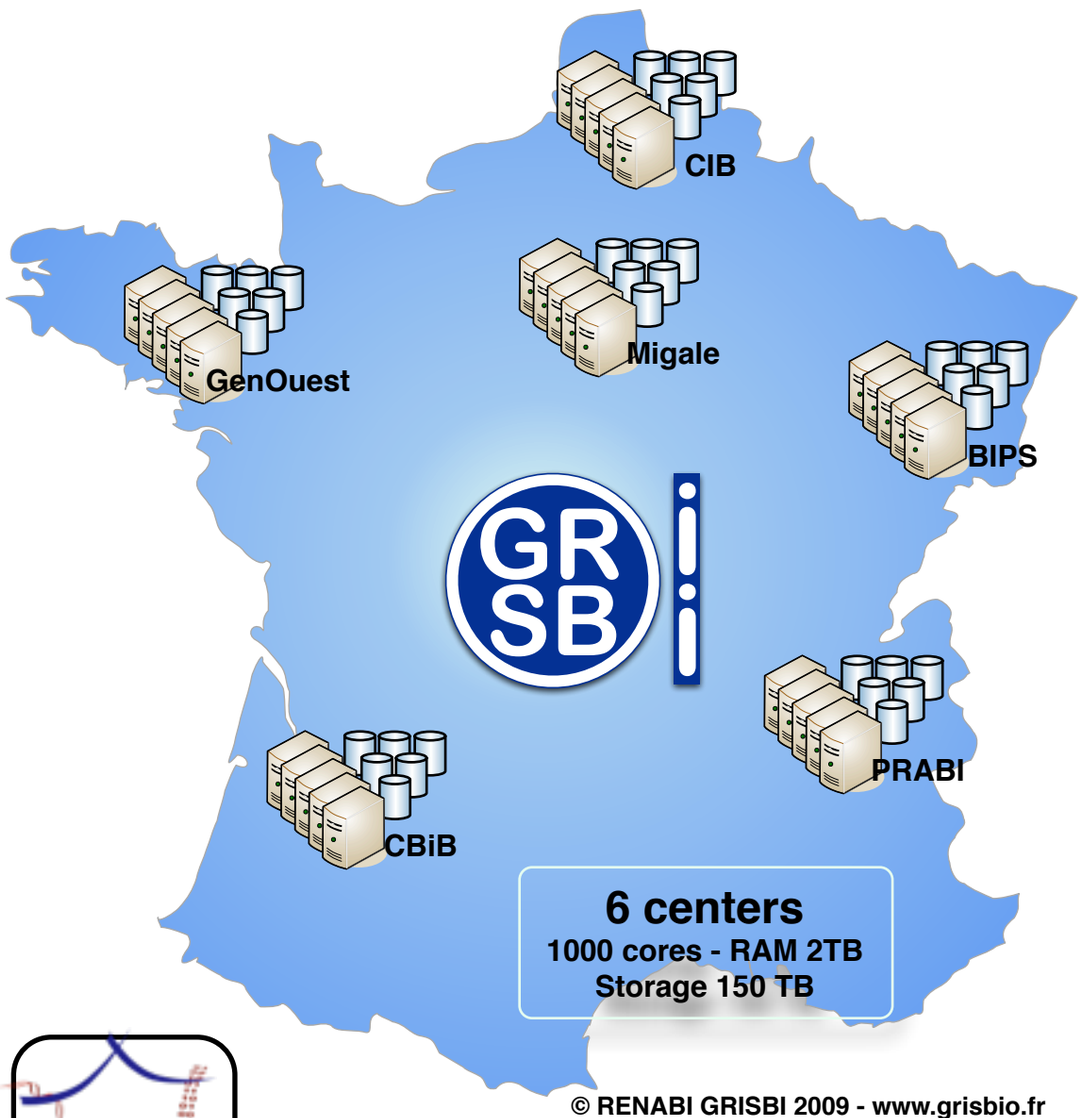
Write Structures

Select Structures

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

GRISBI - Grid, Support to Bioinformatics, www.grisbio.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.)

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)

Technical Committee

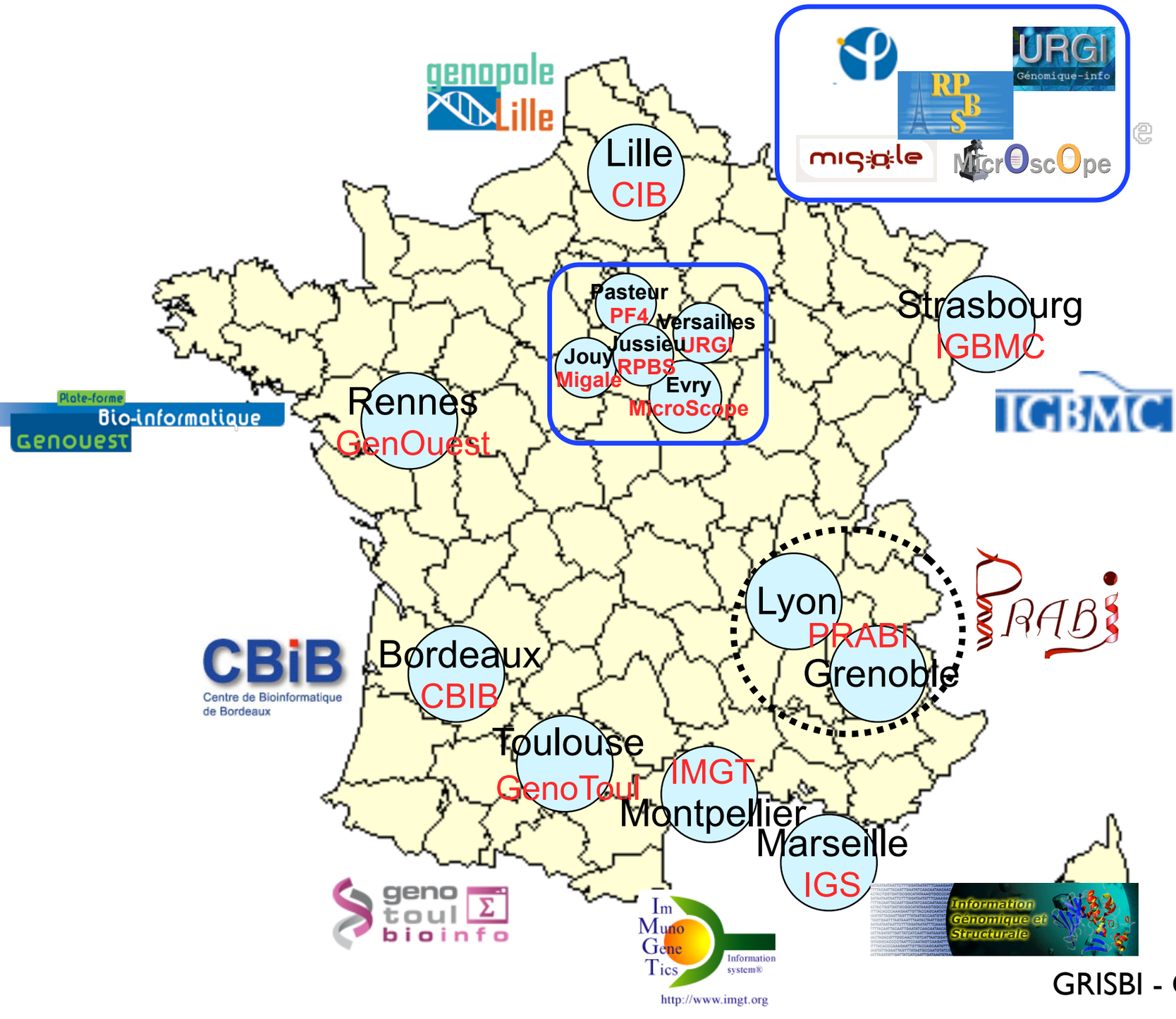
Christophe BLANCHET (Resp. Scient.)
Christophe CARON (Resp. Tech.)
Clément GAUTHEY (Engineer GRISBI)

Participants from the centers
Jean-Claude CHARR (CIB)
Stéphane DELMOTTE (PRABI)

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

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

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 | Emploi

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 :

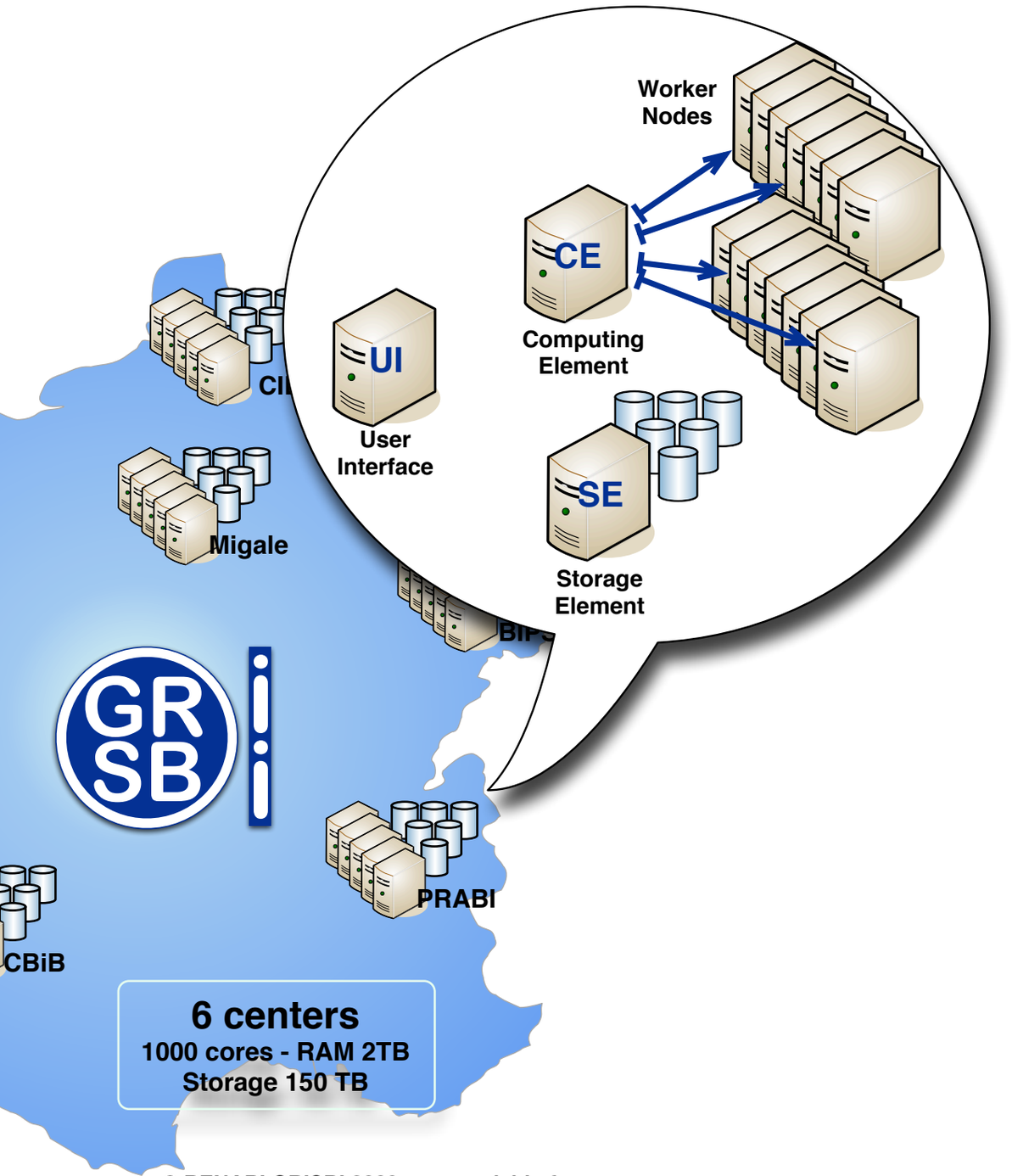
"Quelles actions à mener pour un passage à l'échelle ? : infrastructures, projets "

- cocktail de 18h00 à 19h30

25 novembre 2009 :

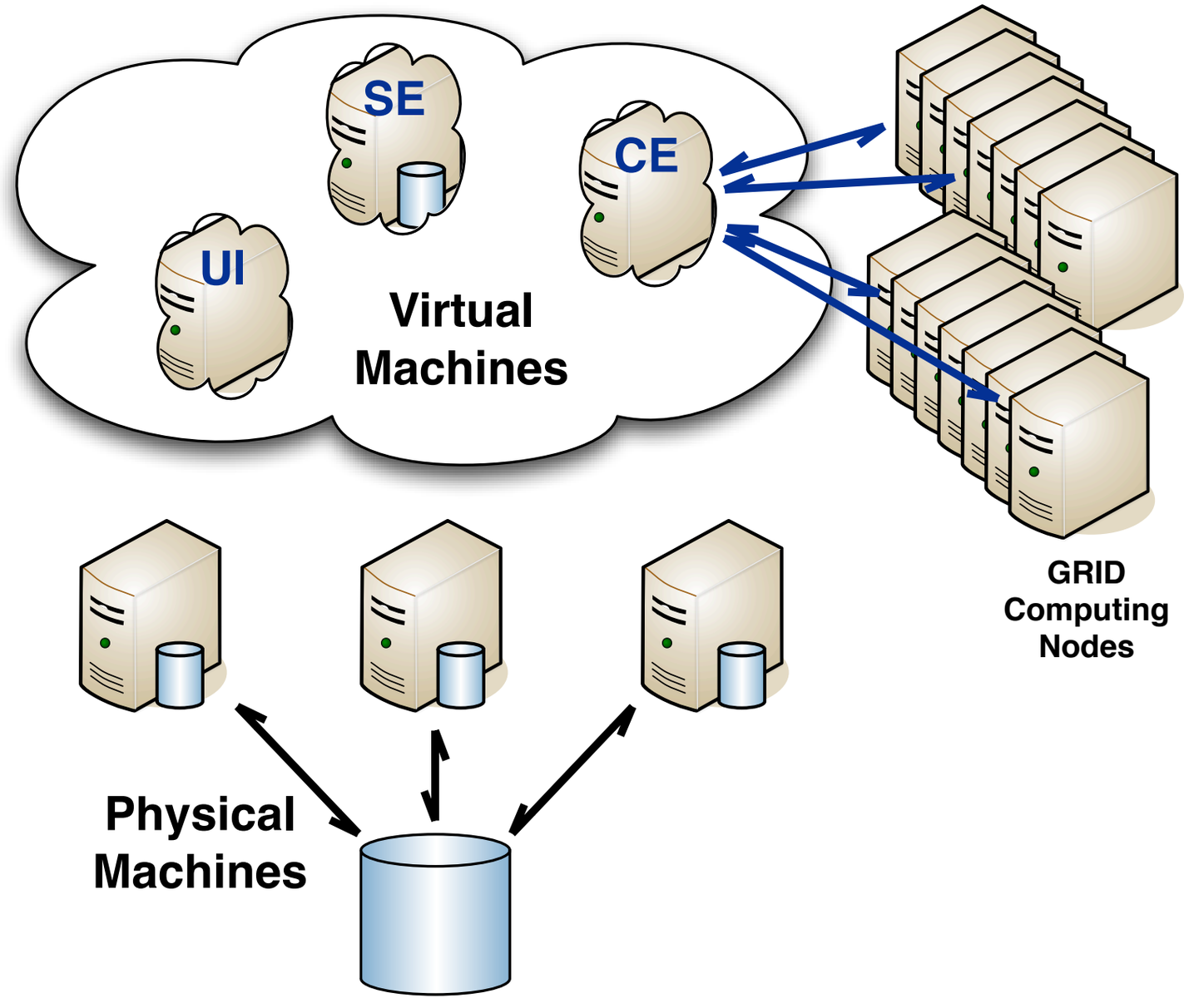
ormatics, www.grisbio.fr

Grid Infrastructure

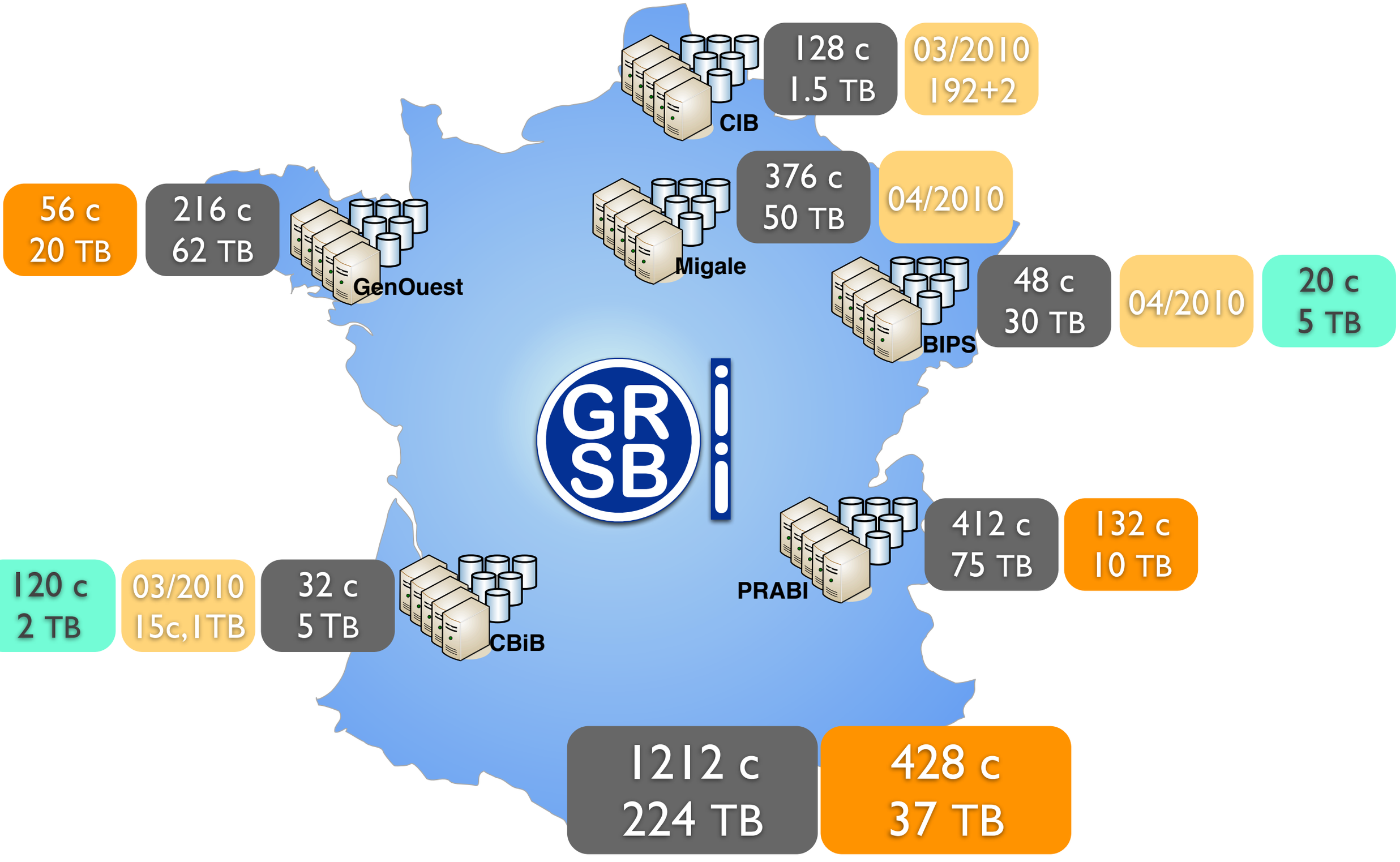


- Mutualize resources
  - resources of the 6 centers
  - monitoring using «ganglia»
- Manage Identities
  - Electronic certificates from the national CNRS PKI : **GRID2-FR**
  - Virtual organization (VO) **vo.renabi.fr** registered in the French ROC.
- Distribute storage
  - biomaj, Active Circle, Hadoop,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

Virtualisation/Cloud



GRID Infrastructure

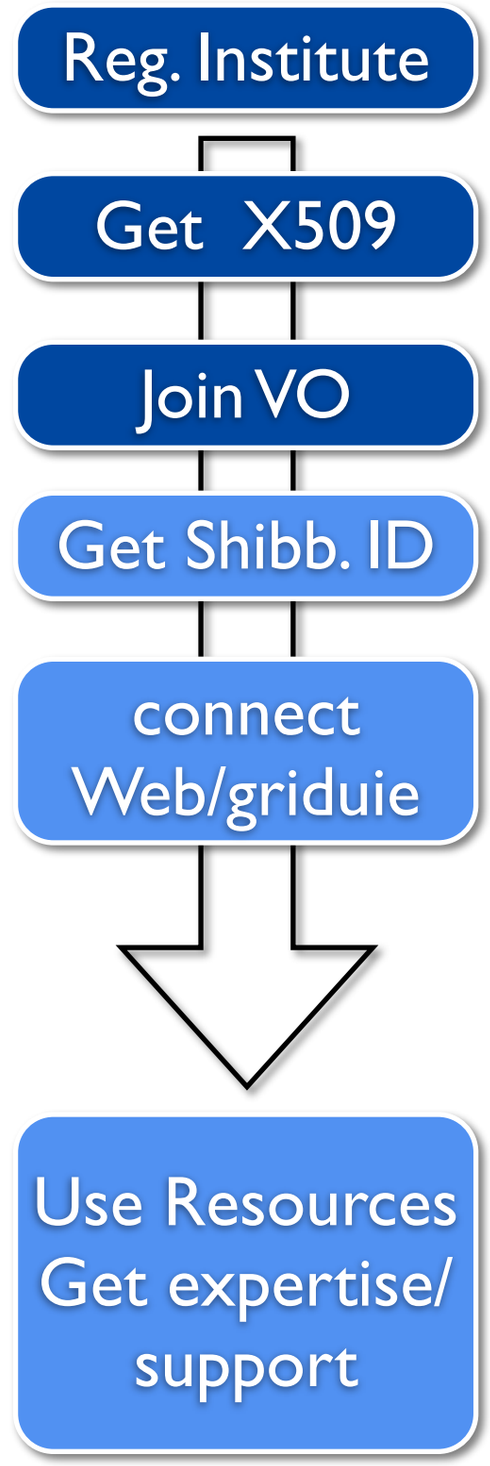


GRISBI - Grid, Support to Bioinformatics, www.grisbio.fr

Tight the Community

- Web: www.grisbio.fr
- E-mail
  - grisbi-tous@services.cnrs.fr
  - grisbi-tech@services.cnrs.fr
  - *bioinfo@sfbi.fr*
- Engineers in platforms
- Identity federation
  - certificat X509 / VOMS
  - federation shibboleth

- *and after ...*
- Support, call center



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

fr en

Plateforme nationale GRISBI

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 auquelle les grilles nationales peuvent répondre.

Inscriptions

Ouvertes aux agents CNRS, autres organismes, industriels

Modalités: *tba*

GRISBI – Calcul Scientifique sur Grille pour la Bioinformatique

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

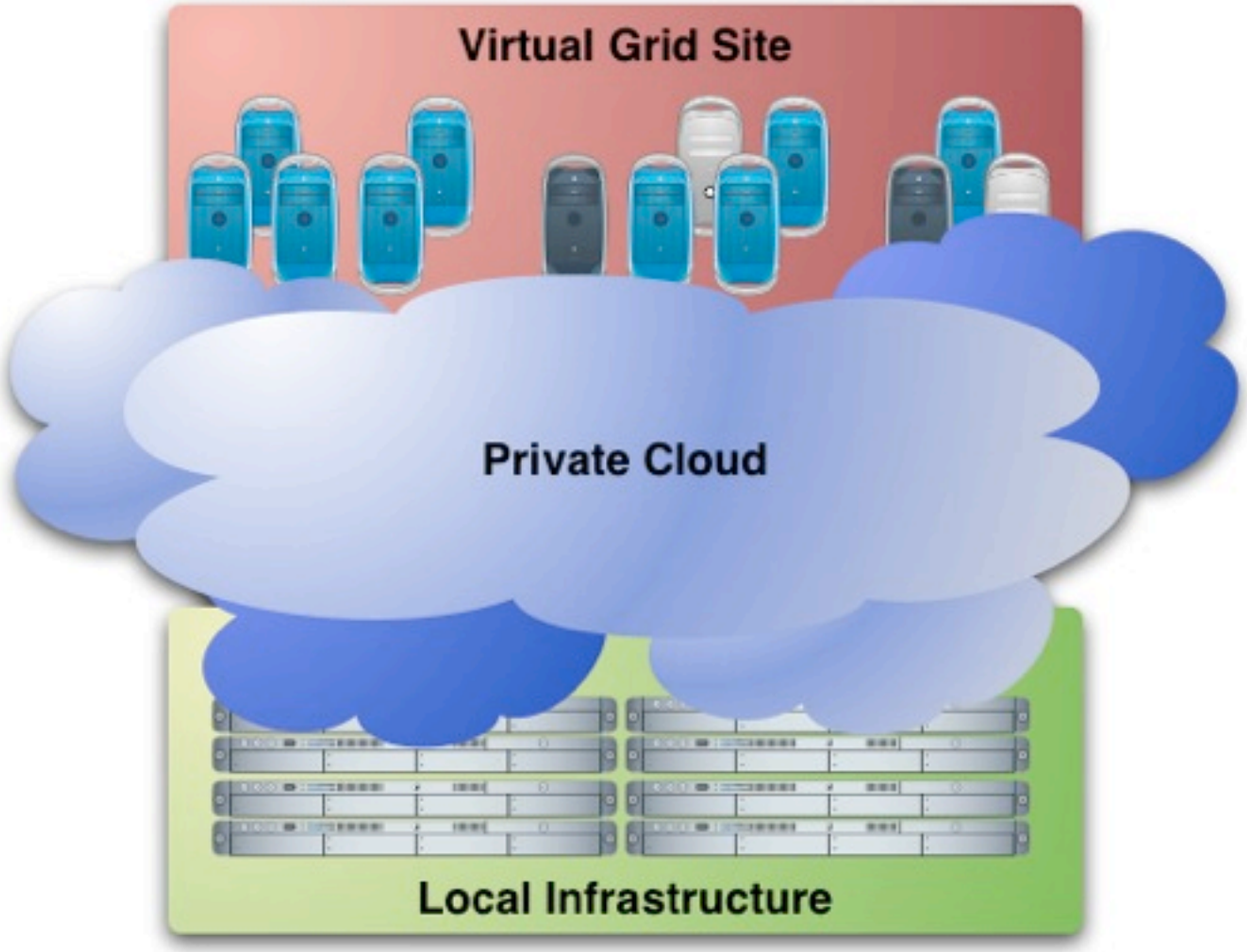
La Bioinformatique requiert maintenant des infrastructures de recherches permettant le stockage de grandes quantités de données complexes et hétérogène et le calcul massif pour leur analyse. Dans ce contexte nous pouvons tirer un avantage certain des nouveaux développements en

StratusLab Initiative

Enhancing Grid Infrastructures with Virtualization and Cloud Technologies

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 demosntrated the added value of these infrastructures
  - Study the nucleosome
  - 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

Agence Nationale de la Recherche, project HIPCAL (ANR-06-CIS6-005)

GIS IBISA through the project GRISBI PF 2008

The European Commission through the EU FP6 EMBRACE project, contract number LHSG-CT-2004-512092

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, R. Lavery, A. Loquet, A. Michon, F. Penin, K. Zakrzewska
GRISBI: and platforms members
Institut Pasteur: T.E. Malliavin, M. Nilges
MPI MG Berlin: R Herwig, C. Wierling
LIP: P. Vicat-Blanc and HIPCAL partners
EGEE: and partners