Décrypthon Grid
Grid Resources Dedicated to Neuromuscular Disorders

N. Bard, E. Caron, F. Desprez
GRAAL/Avalon Research Team
LIP ENS Lyon / INRIA, France

A. Friedrich, L. Moulinier, N.H. Nguyen, O. Poch
IGBMC, France

M. Heymann
Luxembourg Clinical Proteomics
CRP-Santé, Luxembourg

R. Bolze
USC Information Science Institute
USA

T. Toursel
AFM
France

DECRYPTHON

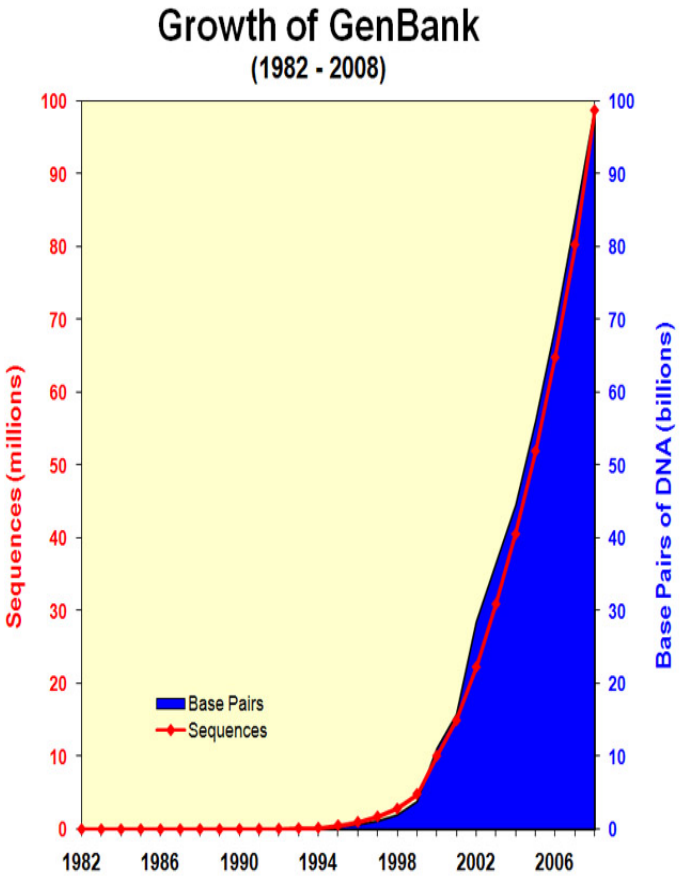
AFM IBM CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Introduction

Thanks to new high throughput technologies, we now have to cope with vast amounts of **data**.

The multiplication of the algorithms in the cascades of processing also generate an increasing need of **CPU** resource.

Transparency and **simplicity** represent the holy grail for distributed platforms (maybe even before performance) !



The size of Biological Datasets is increasing exponentially.

A long collaboration between private and public partners

The Décrypthon project started in 2001 when we invited the Internet users to help us study proteomes by giving unused computer cycles. In 2004 the Decrypthon program was created, whose aim was to build a dedicated university grid, and a desktop voluntary grid

DECRYPTHON

http://www.decrypthon.fr/

AFM, a non government organization.

- Coordination of calls for projects toward the scientific
- community.
- Funding of research projects.
- Scientific strategy of the AFM : curing neuromuscular
- disorders and rare diseases that have most of
- the time a genetic origin.

AFM

IBM, a well known private partner.

- Expert in Grid Computing and Life sciences.
- Gift of supercomputers in 6 universities.
- (Shared University Research program)

IBM

CNRS, the French public research organism.

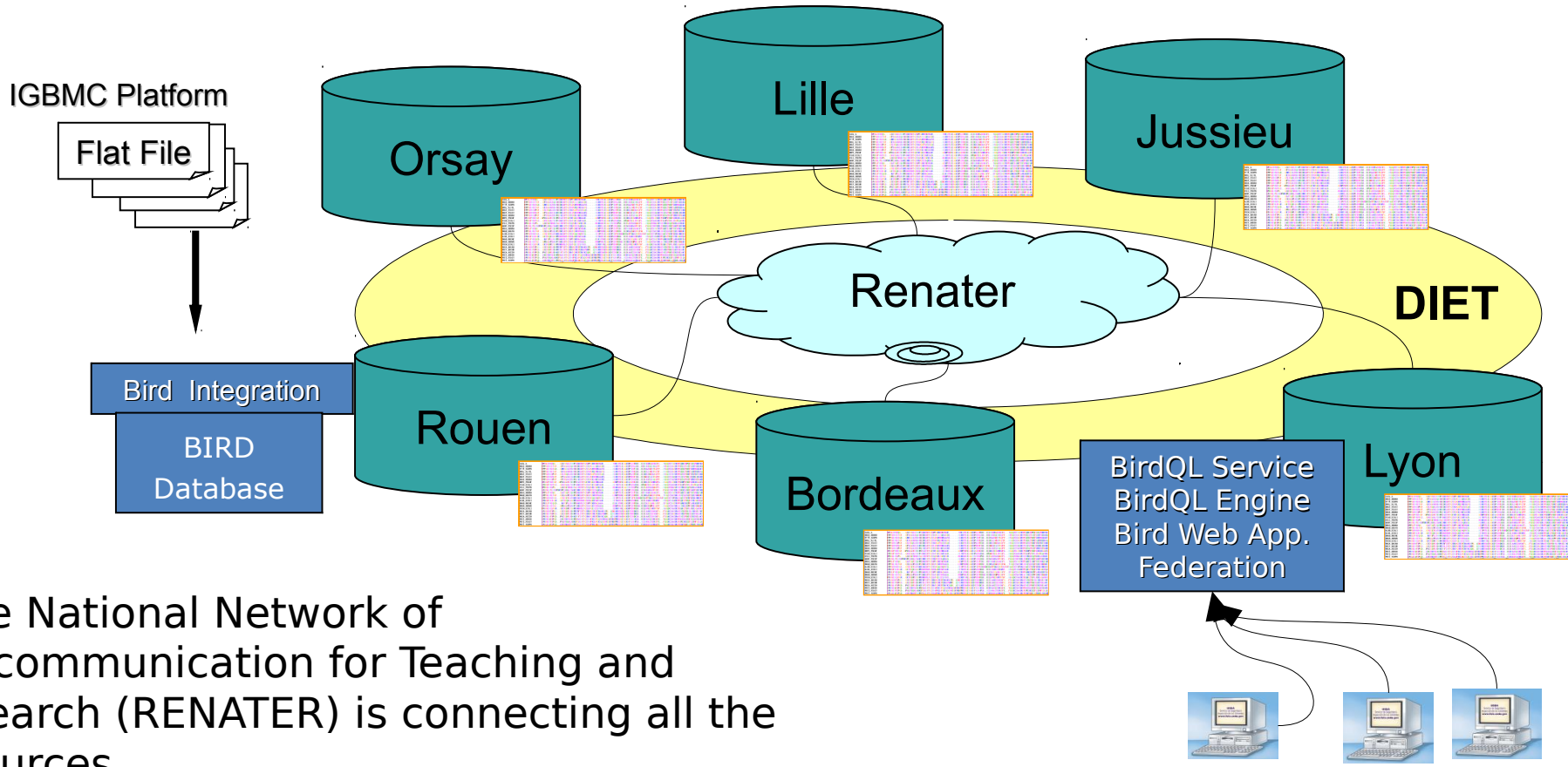
- Scientific steering of the program.
- Scientific and technological expertise for porting applications on the grid.

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Strongly interactive partners

- **ENS Lyon – AVALON/GRAAL Team**
  - Managing the CPU resource of the grid, the DIET middleware.
- **IGBMC – Strasbourg**
  - Managing of the data resources of the grid, the BIRD database systems.

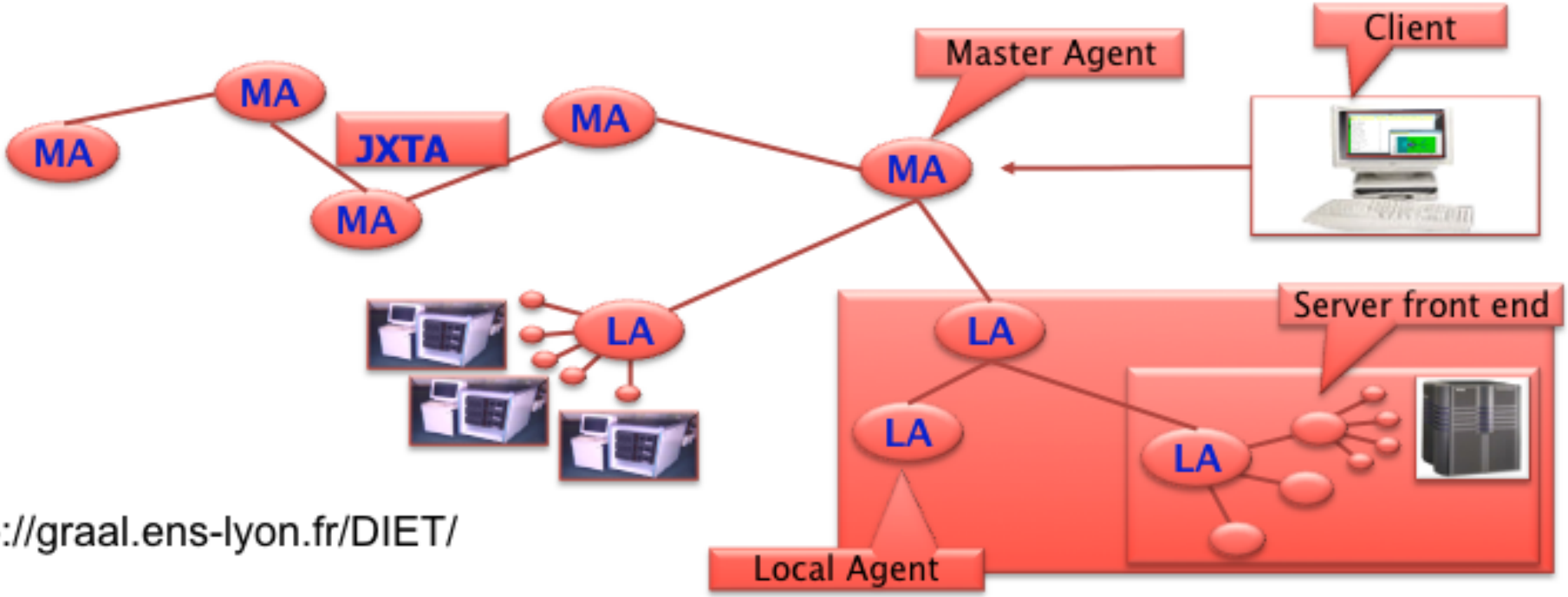
The servers are located in the **universities**, they were given by the **IBM** partner.



- The National Network of Telecommunication for Teaching and Research (RENATER) is connecting all the resources.

DIET Middleware :

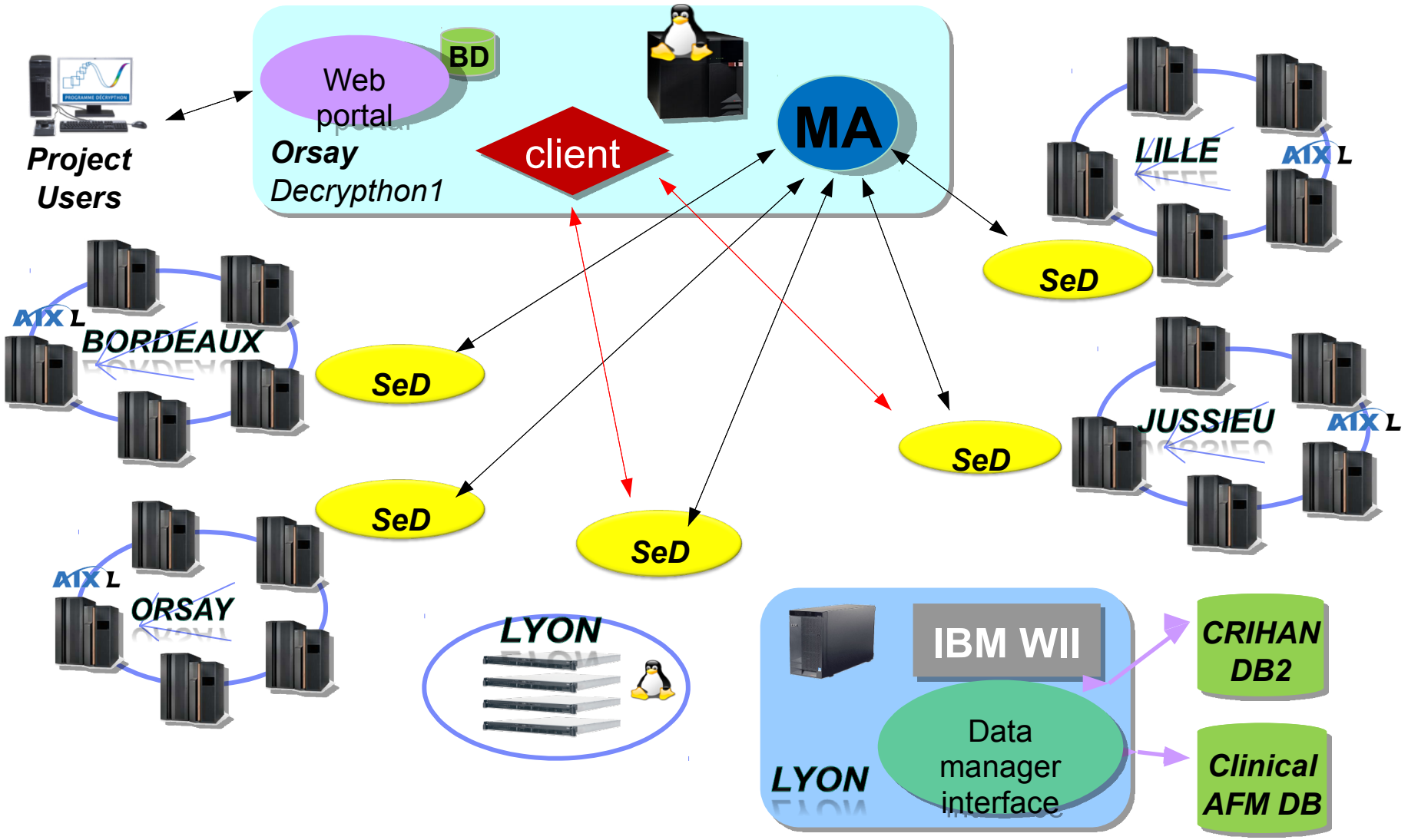
- GridRPC API
- Network Enabled Servers paradigm
- Some features
  - Distributed (hierarchical) scheduling
  - Plugin schedulers
  - Data management
  - Workflow management
  - Sequential and parallel task (through batch schedulers)



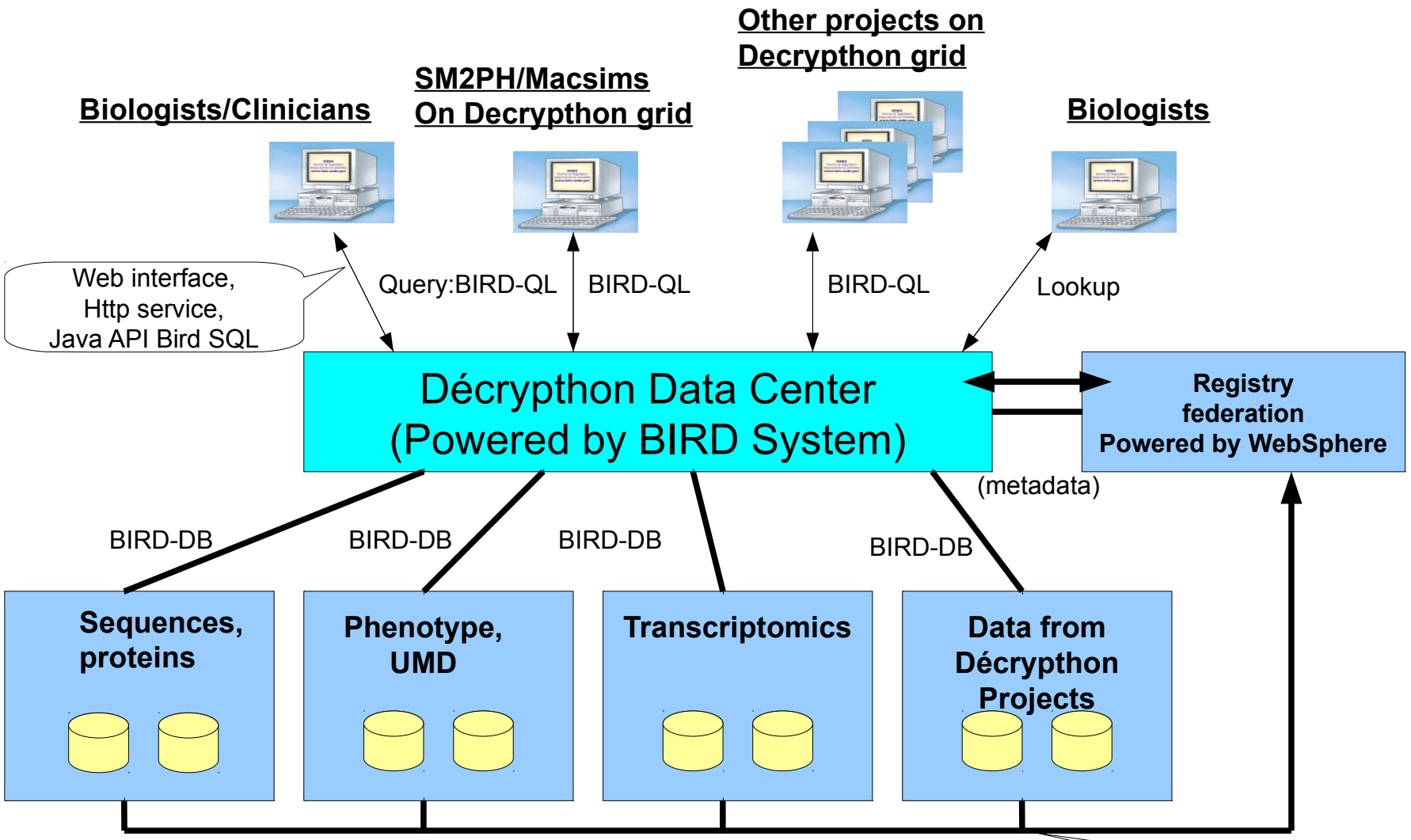
http://graal.ens-lyon.fr/DIET/

Deployment example with Universities

Sed = Server Daemon, installed on any server running Loadleveler. Note that we can define rescue SeD.
MA = master agent, coordinates Jobs. We can define rescue or multiple Master Agent.
WN = worker node



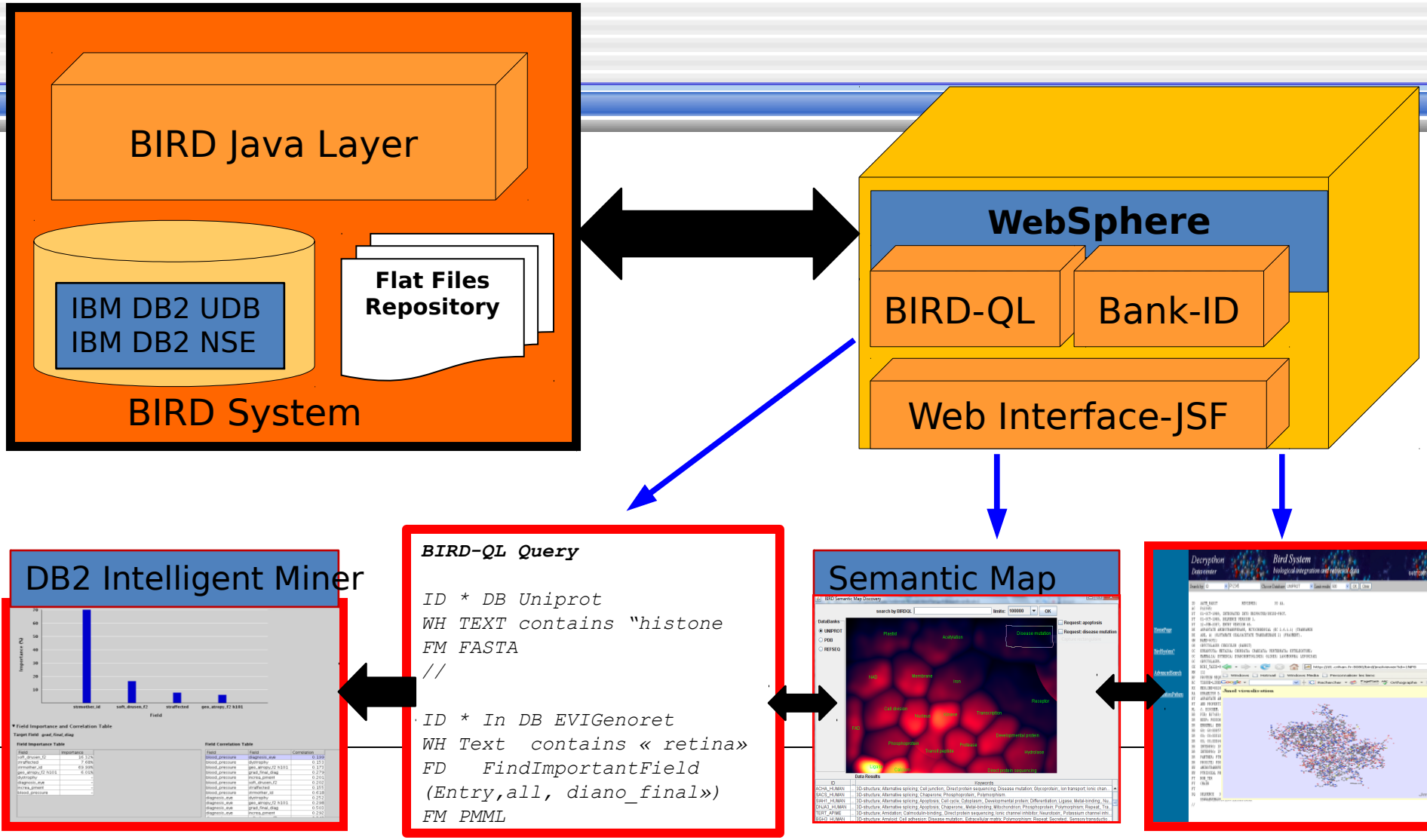
Data management



BIRD-DBs: Database instances generated by BIRD System
Using the query language BIRD-QL you don't have to upload the whole dataset on the grid.

Décrypthon Data Center

BIRD Architecture for Décrypthon



Post-genomic data integration

Data set	Size	Heterogeneities	Integration-Update	Access
Genbank	800 GB	Flat Genbank	Relational schema-60 days	Public
RefSeq	50 GB	Flat Genbank	// -60 days	Public
Uniprot,Uniref	100 GB	Flat Embl	//-20 days	Public
PDP	50 GB	PDB	// -7 days	Public
OMIM	1 G	OMIM	//	Public
Protein Interaction	100 GB	XML	//	Public or Decrypthon
Macsims	> 2TG GB	XML	//	Decrypthon
Genome UCSC	10 GB	Table	// -6 months	Public
Taxonomy	1 GB	Table	// -7 days	Public
Phenotypes* (AFM)			Virtual schema	Decrypthon
Mutation UMD*		UMD	Virtual schema	Decrypthon
GO, SO*, Pato Ontology*,	10 GB	OBO	Service Ontology Lookup EBI	Public

Total : > 5 terabytes of biological data federated in the Decrypthon Data Center

Philosophy of the Décrypthon grid

Transparency : An interface as close as the one users are used to, to submit, to monitor and to download the results of jobs.

Reactivity : new algorithms are ported as quickly as possible on the grid to be close to the practical use case.

Project Name

PipeAlign Job Using web interface

Project annotation

Pipe Align Job, submited with web interface. Date : 23/6/2010

PipeAlign v1.0				**(c)2005 LBGI**
File :		Parcourir...		**Steps**
Blast				☑ : Blast
max expect	5000	max alignments	5000	☑ : Ballast
max seq search	5000	Databank	protall	☑ : Filter
k value	0	f value	9	☑ : Clustal
☑ : gapped	☑ : F value	m value :	0	☑ : Normd
Filter				☑ : Rascal
expect threshold	0.001	max seq kept	500	☑ : Normd
max seq length	3000	Filtering method	none	☑ : Leon
add ballast seq	☐	remove fragments	☑	☑ : Normd
Clustal				☑ : Cluspack
use motifs	☑	propagate	☑	☑ : Macsim
Use project name :	NULL	debug info	☑	☐ : Conservation

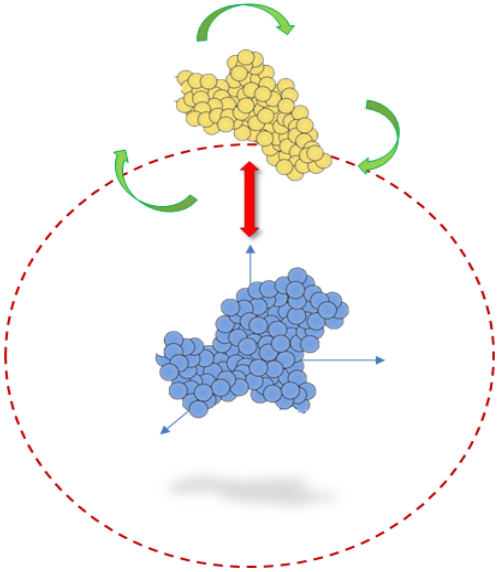
submit PipeAlign job | reset default

Example projects on the Décrypthon grid

▪**The Docking project**

Also known as « Help cure muscular dystrophy » the aim is a large scale investigation of protein-protein, DNA-protein and protein-ligand interactions in the search of new therapeutic targets. It is done by developing computing tools able to locate on the surface of proteins, interaction sites with DNA, ligands, and other proteins.

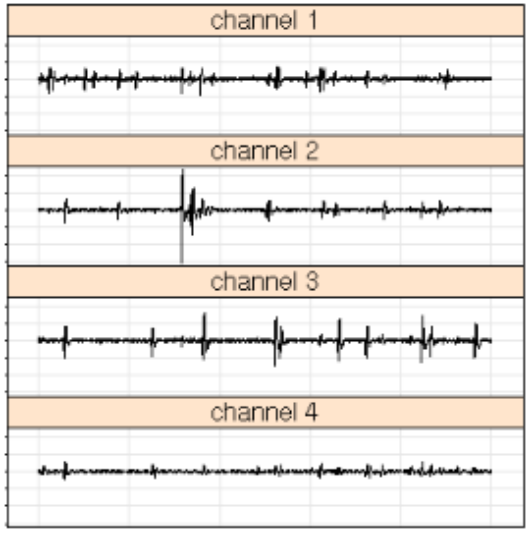
Coordinated by Alessandra Carbone from the "université Pierre et Marie Curie" , Paris (Inserm U511 - Immunologie cellulaire et moléculaire des infections parasitaires - Genomique analytique)



▪**The SpikeOMatic project**

The aim is to be able to sort the electric peaks of multiple neurons, to enable extracellular recording which is less invasive, and the development of an analytic tool for neurosciences and neuromuscular diseases.

Automated statistical spike sorting.
Coordinated by Christophe Pouzat de l'universite Rene Descartes, Paris V (CNRS UMR 8118 - Laboratoire de physiologie cerebrale)



SM2PH a pilot project and a success story

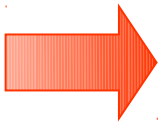
SM2PH : from Structural Mutation to Pathology Phenotypes in Human

(Friedrich A et al. Human Mutation 2010 Feb;31(2):127-35)

Goal: estimate the structural impact of a mutation and correlate with human genotype, phenotype and pathology

First step: introduce **SStElSy (Sequence/STucture/Evolution Inference in SYstems)** reasoning in the context of human monogenic diseases

- Infrastructure: - BIRD database (sequences, genomical data, alignments, 3D structures or models, pathology descriptions, analysis results…)

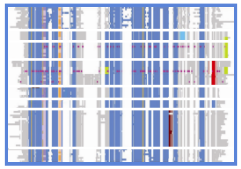


- Genotype/Phenotype description (UMD, LOVD)
- WEB server, interactive analysis interface

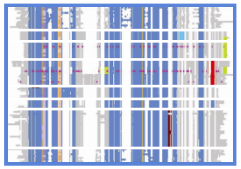
- Software and algorithms: scoring function for prediction of mutation consequences in structural and variability contexts, correlation between mutation and phenotype severity…

SM2PH-db

- **SM2PH-db** (http://decrypthon.igbmc.fr/sm2ph/)
  - entry : 2 296 proteins involved in human monogenic diseases
  - for each entry: wide range of information involved in the genotype – phenotype relationship
    - Evolutionary view : Multiple Alignment of Complete Sequences



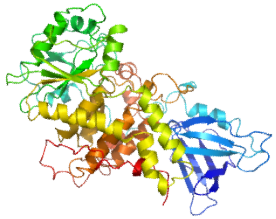
Eukaryote Filter



Structural sampling

    - Structural view : 3D models: 1596 wild type proteins

10245 mutant proteins

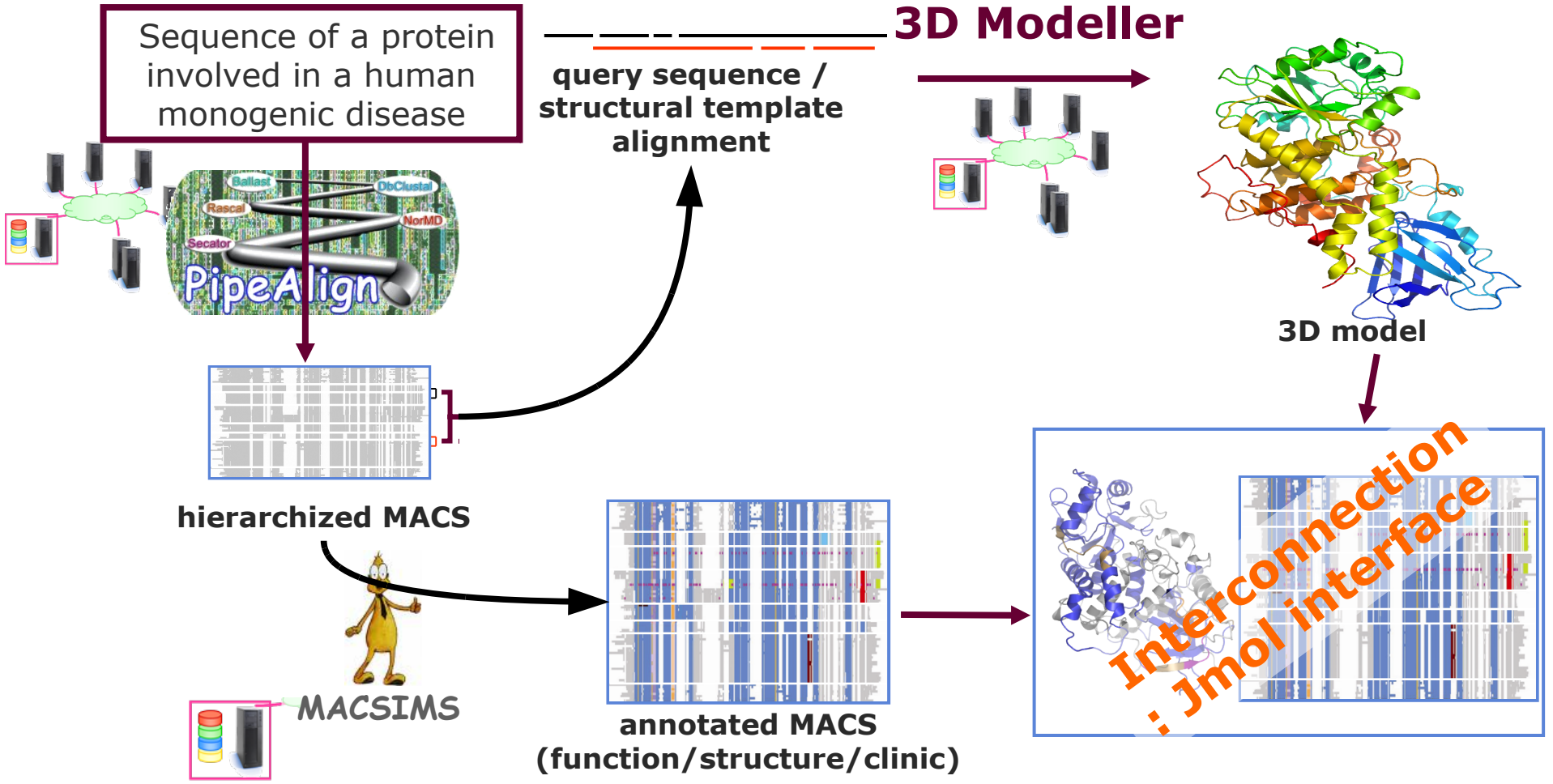


    - informational view :
      - structural and functional annotations (UniProt, Pfam, Prosite, Interpro, GO)
      - mutation and phenotypic data (24 962 missense mutations)

Automatic update every 2 months, current version : 9

Complex interconnected programs

On a single data, up to 25 programs are applied, in cascade or in parallel.



Conclusion and Future Work

- The Décrypthon Grid is a simple and efficient Grid
- Efficiency and transparency of resource management through DIET and a simple portal
- Data management is one of the most important issue of large scale applications over the grid
  - It has to be linked with request scheduling to get the best performance

- **Future work**
  - Work toward Cloud Computing platforms (Décrypthon-Cloud?)
  - Attract new users and new applications
    - COMPARE/MYOBASE (FP 6) , OrthoInspector, EvolHHuPro…
    - SM2PH-KB (version 2.0 of SM2PH-DB) : Knowlegde base for neuromuscular diseases
      - Complex data integration
      - Data mining integrative

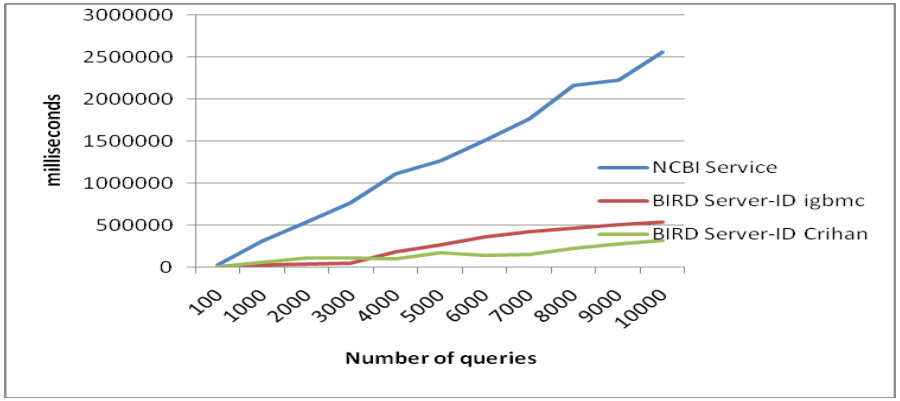
Thank you for your attention
Questions ?

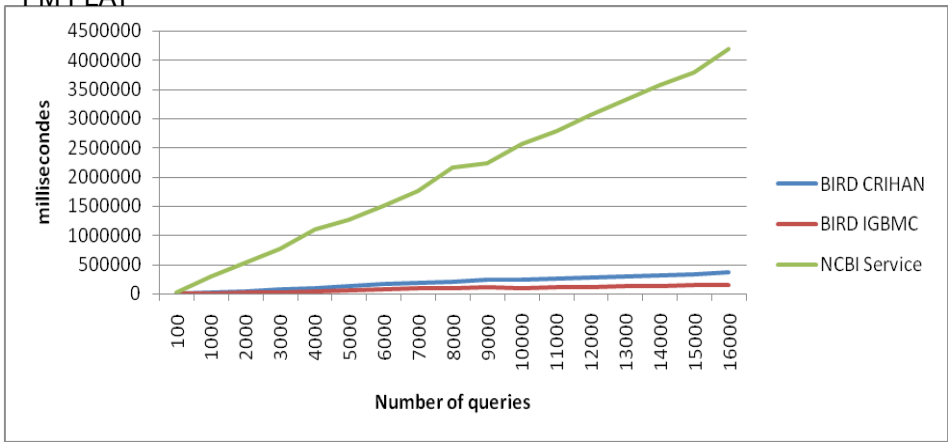
Why we develop the Decrypthon Data Center?

Bank-ID Service Evaluation:
We h**B**ave sent (100 , 1000, ..,20000) queries to BIRD and NCBI. Each query was used to get one full Genbank sequence.

A single query was sent to BIRD and NCBI in order to get (100 , 1000, ..,20000) full Genbank sequences. These queries (get full sequence) below were used for this evaluation.

Nucleotic
WH TEXT cID * DB ontains “histone”
LM 100
FM FLAT
//
…..
//
ID * DB Nucleotic
WH TEXT contains “histone”
LM 20000
FM FLAT





Why: ???