

# Cloud Ready for Bioinformatics ?

**C. Blanchet and C. Gauthey**

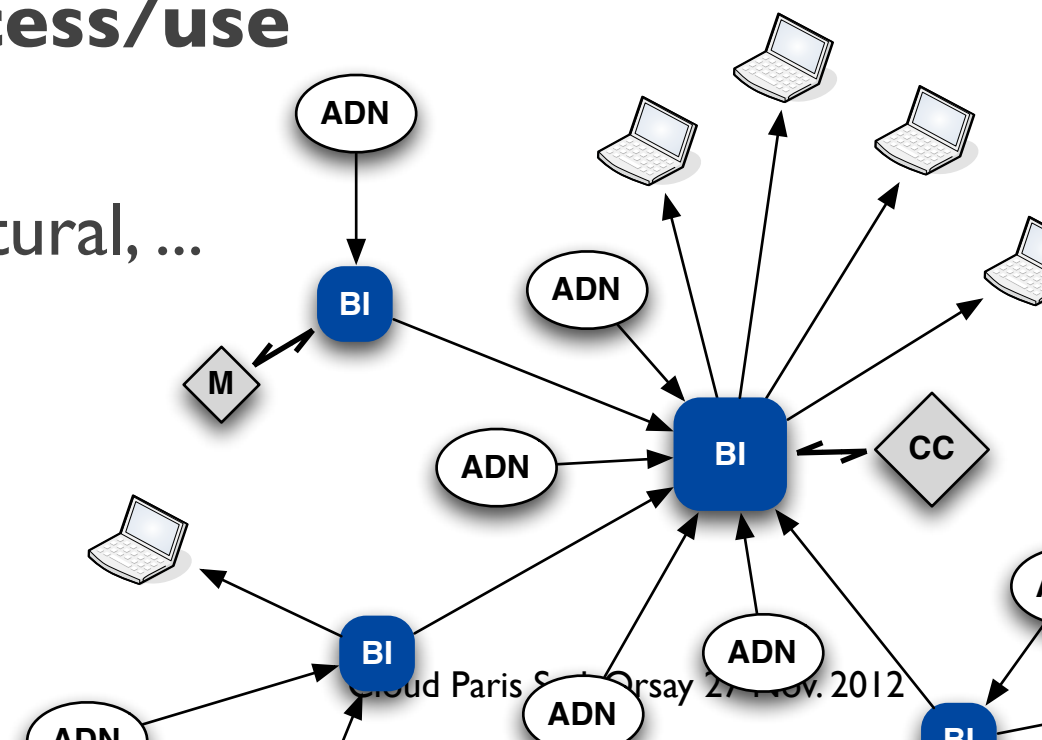
**Plateforme 'Infrastructure Distribuée pour la Biologie'**

Journée 'Clouds pour le Calcul Scientifique, Paris Sud' (Orsay, France)

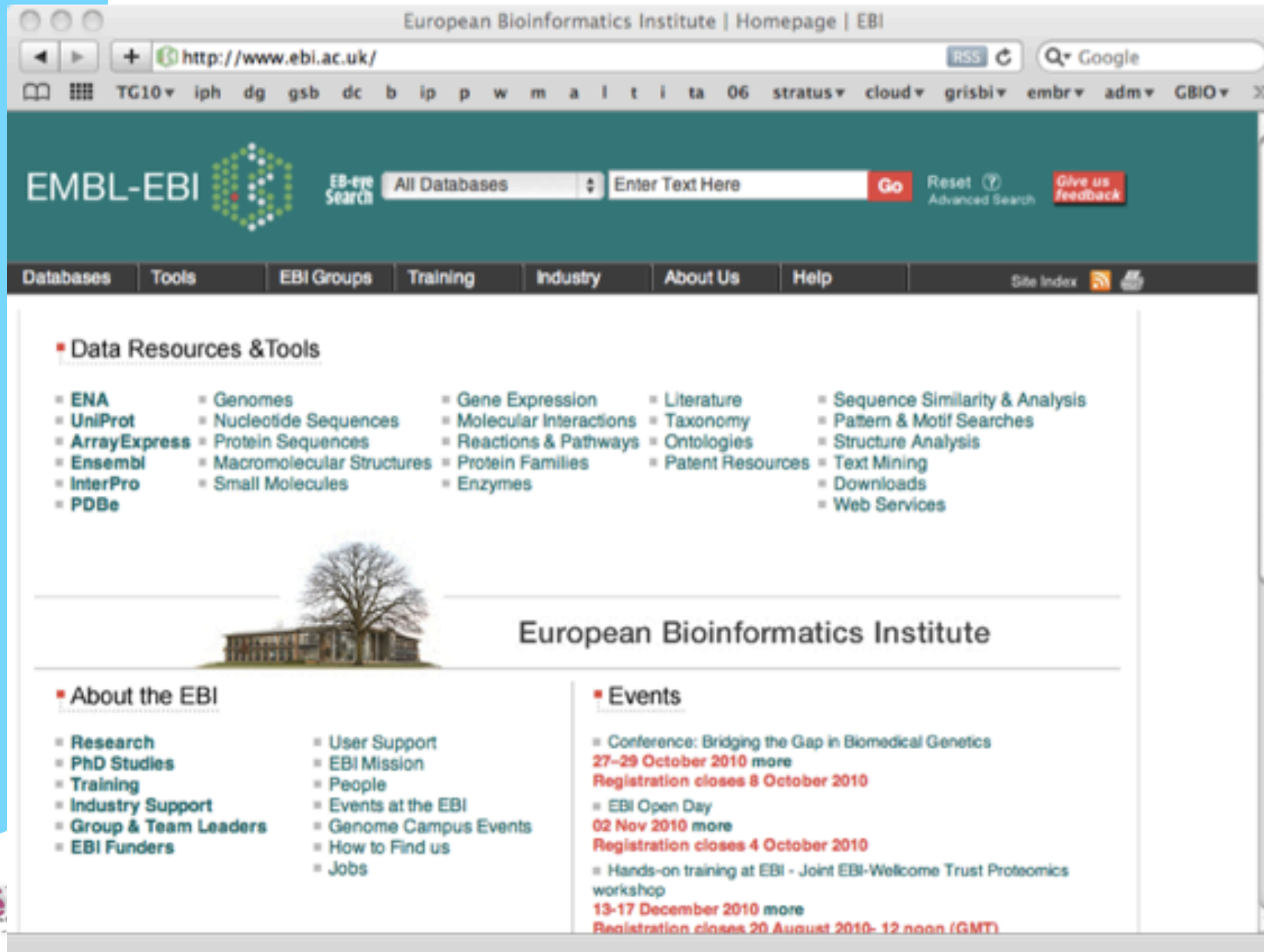
27 novembre 2012

# Bioinformatics Today

- Size of biological data are tremendous
  - Institut Sanger, UK, 5 PB
  - Beijing Genome Institute, China, 4 sites, 10 PB
- ➔ **Huge data in lot of places**
- Analysing such data became difficult
  - Scale-up of the analyses : gene/protein to complete genome/proteome, ...
  - Lot of different daily-used tools
  - That need to be combined in workflows
  - Usual interfaces: portals, Web services, federation,...
- ➔ **Datacenters with ease of access/use**
- Distributed resources
  - Experimental platforms: NGS, structural, ...
  - Bioinformatics platform
- ➔ **Federation**



# Infrastructure in Biology



The screenshot shows the homepage of the European Bioinformatics Institute (EMBL-EBI). The browser address bar displays "http://www.ebi.ac.uk/". The page features a search bar with "All Databases" selected and a "Go" button. Below the search bar is a navigation menu with categories: Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. The main content area is titled "Data Resources & Tools" and lists various resources such as ENA, UniProt, ArrayExpress, Ensembl, InterPro, PDBe, Genomes, Nucleotide Sequences, Protein Sequences, Macromolecular Structures, Small Molecules, Gene Expression, Molecular Interactions, Reactions & Pathways, Protein Families, Enzymes, Literature, Taxonomy, Ontologies, Patent Resources, Sequence Similarity & Analysis, Pattern & Motif Searches, Structure Analysis, Text Mining, Downloads, and Web Services. Below this is a section for the "European Bioinformatics Institute" with a photograph of a building. Further down, there are sections for "About the EBI" (Research, PhD Studies, Training, Industry Support, Group & Team Leaders, EBI Funders, User Support, EBI Mission, People, Events at the EBI, Genome Campus Events, How to Find us, Jobs) and "Events" (Conference: Bridging the Gap in Biomedical Genetics, EBI Open Day, Hands-on training at EBI - Joint EBI-Wellcome Trust Proteomics workshop).



# Infrastructure in Biology

The screenshot displays the Galaxy web interface. At the top, the browser address bar shows `http://idb-cloud.ibcp.fr:20007/`. The main navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. On the left, a 'Tools' sidebar lists categories like 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', and 'FASTA manipulation'. Below this, there are sections for 'Data Resources' (listing ENA, UniProt, ArrayExpress, Ensembl, InterPro, PDB) and 'About Us' (listing Research, PhD Students, Training, Industry Support, Group & Team Leaders, EBI Funders, Events at the EBI, Genome Campus Events, How to Find us, Jobs, EBI Open Day, Hands-on training at EBI).

The central workspace shows a workflow diagram titled 'WWFSMD? grow noodly appendages...'. The workflow includes steps such as 'Input dataset', 'Filter', 'Join', 'Group', 'Sort', 'Join two Queries', and 'Select first'. A green notification box at the top of the workspace says 'Hello world! It's running... To customize this page edit `static/welcome.html`'. Below the workflow, the text 'usegalaxy.org' is visible. At the bottom of the workspace, a message states: 'This project is supported in part by [NSE](#), [NHGRI](#), and [the Huck Institutes of the Life Sciences](#).'

On the right side, a 'History' panel shows '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.





# Infrastructure in Biology

The image displays two overlapping web browser windows. The background window is the GenOuest Bioinformatics Platform, showing a 'Blast' tool interface. The foreground window is the Galaxy web interface, showing a 'Tools' menu with various bioinformatics options.

**Galaxy Tools Menu:**

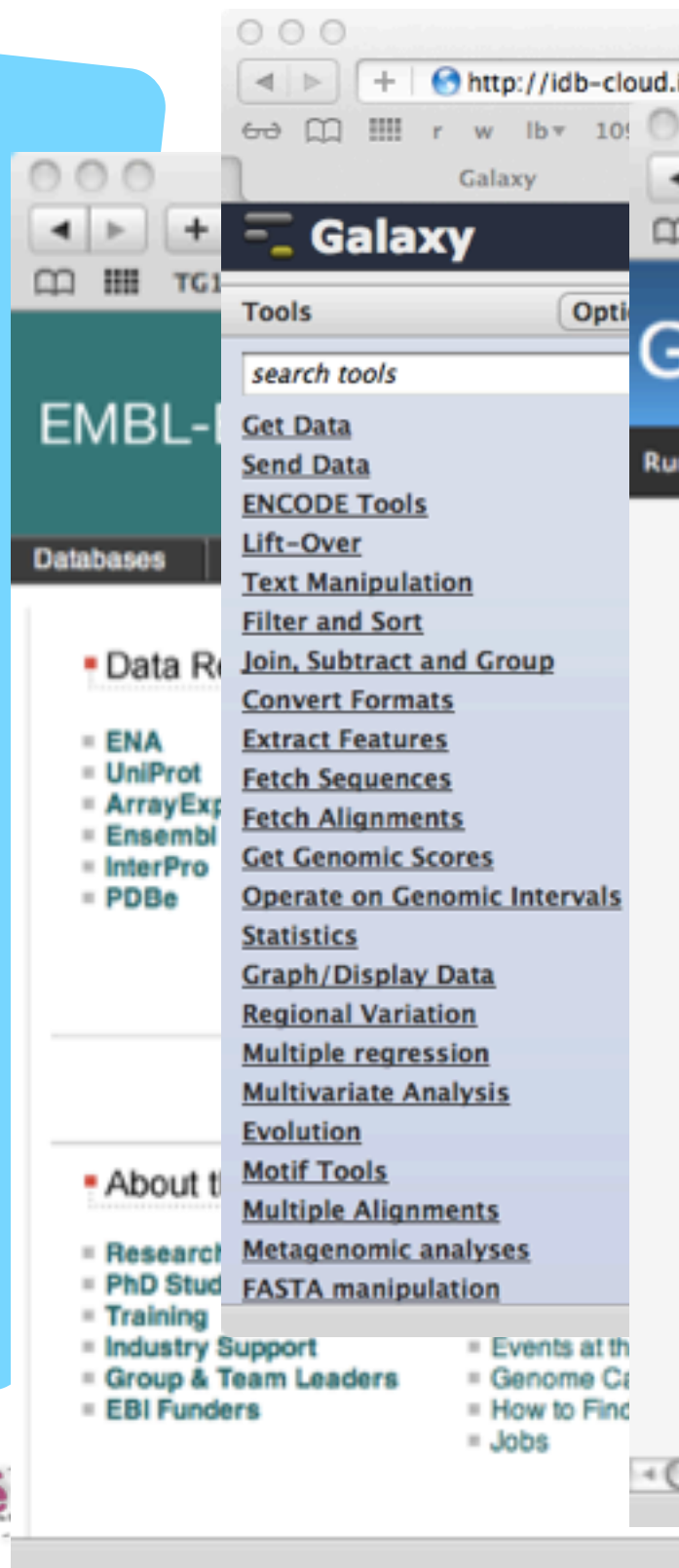
- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation

**GenOuest Bioinformatics Platform - Blast Interface:**

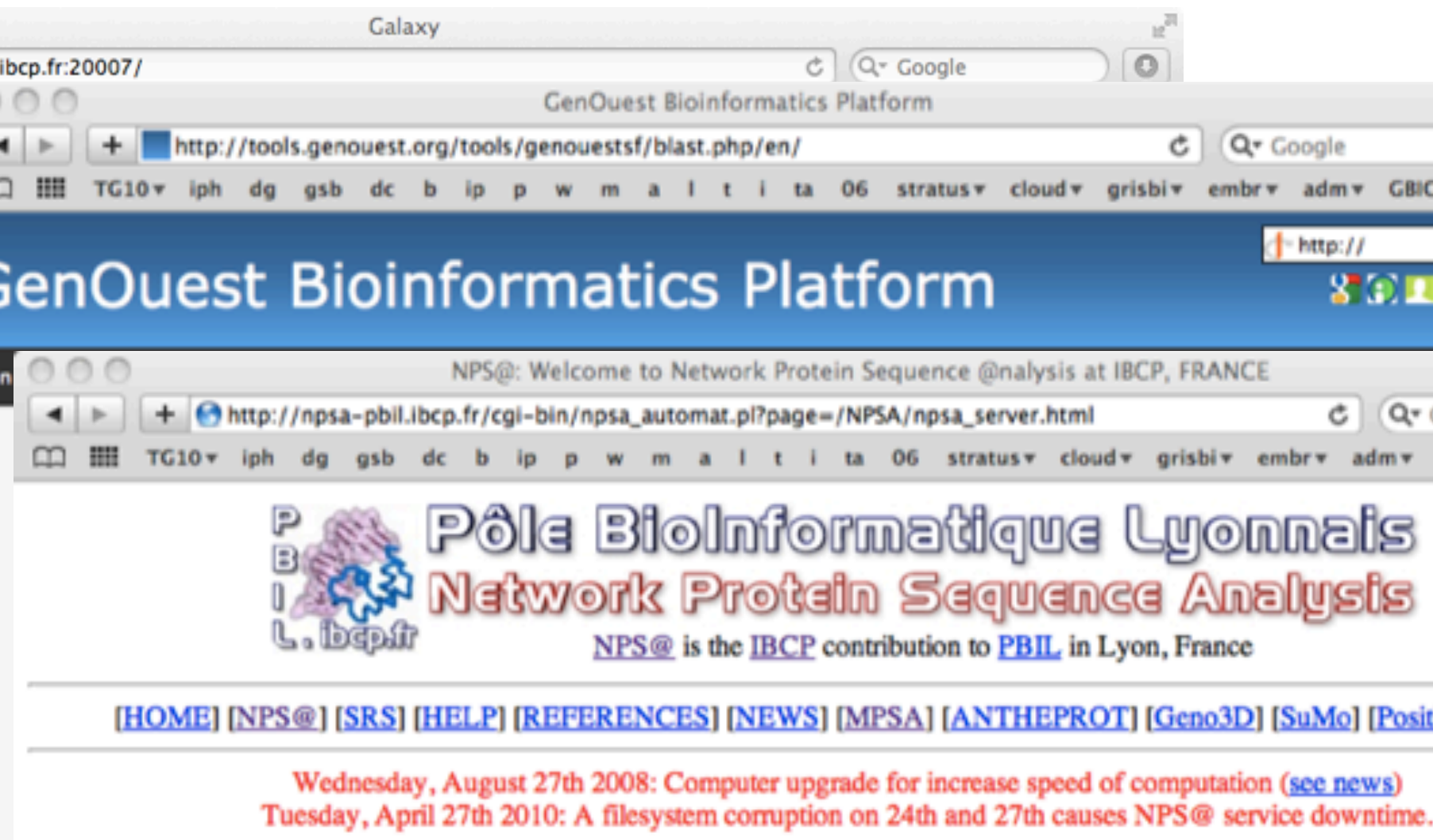
- URL: <http://tools.genouest.org/tools/genouestsf/blast.php/en/>
- Section: **Blast**
- Form fields:
  - Paste your sequence (FASTA format): [Empty text area]
  - Or select a file (FASTA format):  aucun sélectionné
  - Program:
  - Databank type:
  - Databank:  version:
  - Expect:  The statistical significance threshold for reporting matches.



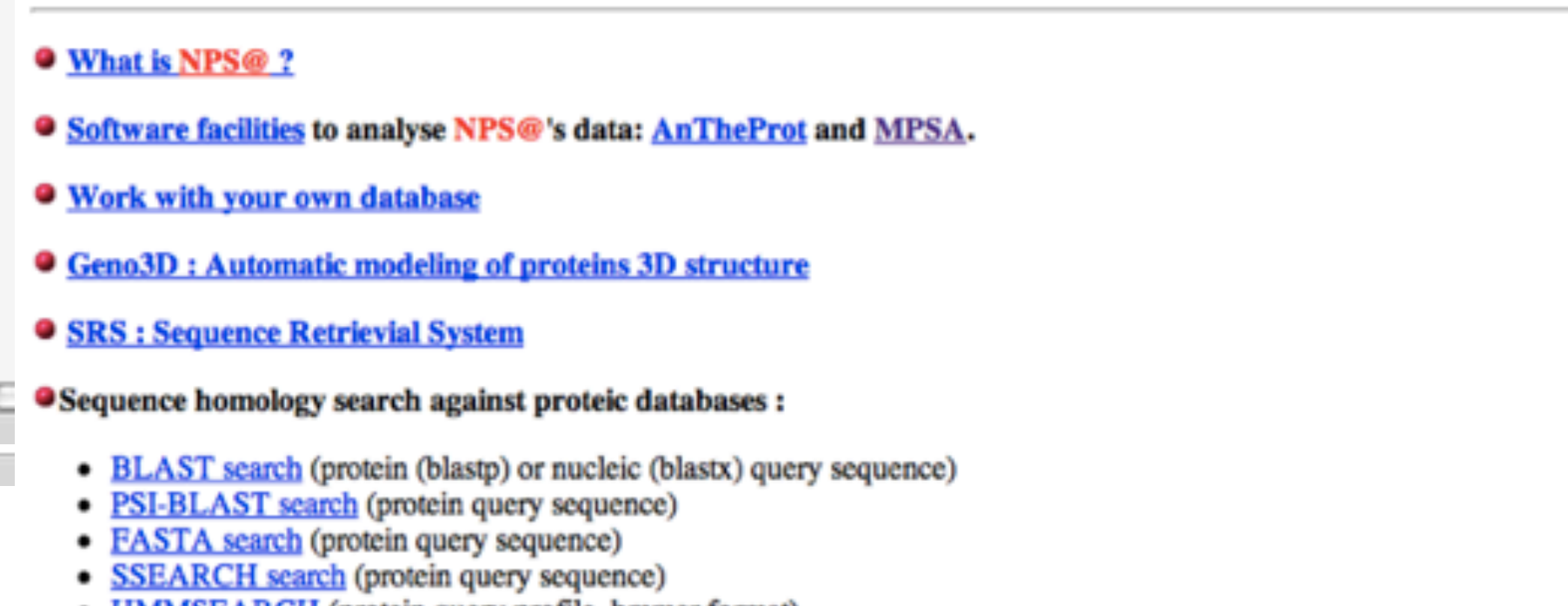
# Infrastructure in Biology



The screenshot shows the Galaxy web interface. On the left, there is a sidebar with 'Tools' and 'Databases' sections. The 'Tools' section is expanded to show a list of tools including 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', and 'FASTA manipulation'. The 'Databases' section includes 'Data Resources' with links to ENA, UniProt, ArrayExpress, Ensembl, InterPro, and PDBe. Below the sidebar, there are sections for 'About the Galaxy Project' and 'Events at the IBCP'.



The screenshot shows the GenOuest Bioinformatics Platform website. The header includes the text 'GenOuest Bioinformatics Platform' and a navigation menu with links like 'TG10', 'iph', 'dg', 'gsb', 'dc', 'b', 'ip', 'p', 'w', 'm', 'a', 'l', 't', 'i', 'ta', '06', 'stratus', 'cloud', 'grisbi', 'embr', 'adm', 'GBIC'. The main content area features the logo for 'Pôle BioInformatique Lyonnais' and 'Network Protein Sequence Analysis'. Below the logo, there is a navigation menu with links: [HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positional Cloning]. A news section contains two entries: 'Wednesday, August 27th 2008: Computer upgrade for increase speed of computation (see news)' and 'Tuesday, April 27th 2010: A filesystem corruption on 24th and 27th causes NPS@ service downtime.'



The screenshot shows the NPS@ website. The header includes the text 'NPS@: Welcome to Network Protein Sequence @analysis at IBCP, FRANCE' and the URL 'http://npsa-pbil.ibcp.fr/cgi-bin/npsa\_automat.pl?page=/NPSA/npsa\_server.html'. The main content area features the logo for 'Pôle BioInformatique Lyonnais' and 'Network Protein Sequence Analysis'. Below the logo, there is a navigation menu with links: [HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positional Cloning]. A news section contains two entries: 'Wednesday, August 27th 2008: Computer upgrade for increase speed of computation (see news)' and 'Tuesday, April 27th 2010: A filesystem corruption on 24th and 27th causes NPS@ service downtime.'

- [What is NPS@ ?](#)
- [Software facilities](#) to analyse NPS@'s data: [AnTheProt](#) and [MPSA](#).
- [Work with your own database](#)
- [Geno3D : Automatic modeling of proteins 3D structure](#)
- [SRS : Sequence Retrieval System](#)
- **Sequence homology search against proteic databases :**
  - [BLAST search](#) (protein (blastp) or nucleic (blastx) query sequence)
  - [PSI-BLAST search](#) (protein query sequence)
  - [FASTA search](#) (protein query sequence)
  - [SSEARCH search](#) (protein query sequence)
  - [HMMSEARCH](#) (protein query sequence)





# Infrastructure in Biology

The image displays several overlapping browser windows related to bioinformatics infrastructure:

- Galaxy**: A web-based platform for genomic data analysis, showing a search for tools and a list of categories like 'Get Data', 'Send Data', and 'ENCODE Tools'.
- GenOuest Bioinformatics Platform**: A central hub for bioinformatics tools, with a URL of <http://tools.genouest.org/tools/genoueststf/blast.php/en/>.
- NPS@**: Network Protein Sequence analysis, with a URL of [http://npsa-pbil.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_server.html](http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html).
- gBIO - Web Services**: A page titled 'Web Services' that provides programmatic access to various tools. It includes a logo and a table of available tools.

**Web Services**

The IBCP have integrated several tools for protein sequence analysis with the Web services technology. These Bioinformatics Web services provide scientists and developers with programmatic access to these tools. Our Web services are build upon standards from the W3C like SOAP, WSRF and HTTP. These tools can be use remotely through a graphical and integrated SOAP client like Taverna or Triana. You can also write your own SOAP client with languages such as Python & ZSI, C/C++ gSOAP, perl SOAP::Lite or Java.

**Bioinformatics Tools available**

	Type of analysis	Description	Documentation	Examples of clients
ClustalW	multiple alignment	wsdl	usage	PyZSI Tav2 ( pict)
Multalin	multiple alignment	wsdl	usage	PyZSI Tav2 ( pict)
BLAST	sequence similarity	wsdl	usage	PyZSI Tav2 ( pict)
FastA	sequence similarity	wsdl	usage	PyZSI Tav2 ( pict)

Additional logos visible in the bottom left corner include **idé**, **IBCP**, and **CPRS**.

# Infrastructure in Biology

Galaxy

GenOuest Bioinformatics Platform

NPS@: Welcome to Network Protein Sequence @analysis at IBCP, FRANCE

gBIO - Web Services

**Lot of tools and web services to treat and visualize lot of data**

	Type of analysis	Description	Documentation	Examples of clients
ClustalW	multiple alignment	wsdl	usage	PyZSI Tav2 ( pict)
Multalin	multiple alignment	wsdl	usage	PyZSI Tav2 ( pict)
BLAST	sequence similarity	wsdl	usage	PyZSI Tav2 ( pict)
FastA	sequence similarity	wsdl	usage	PyZSI Tav2 ( pict)

idé

IBCP

CNRS

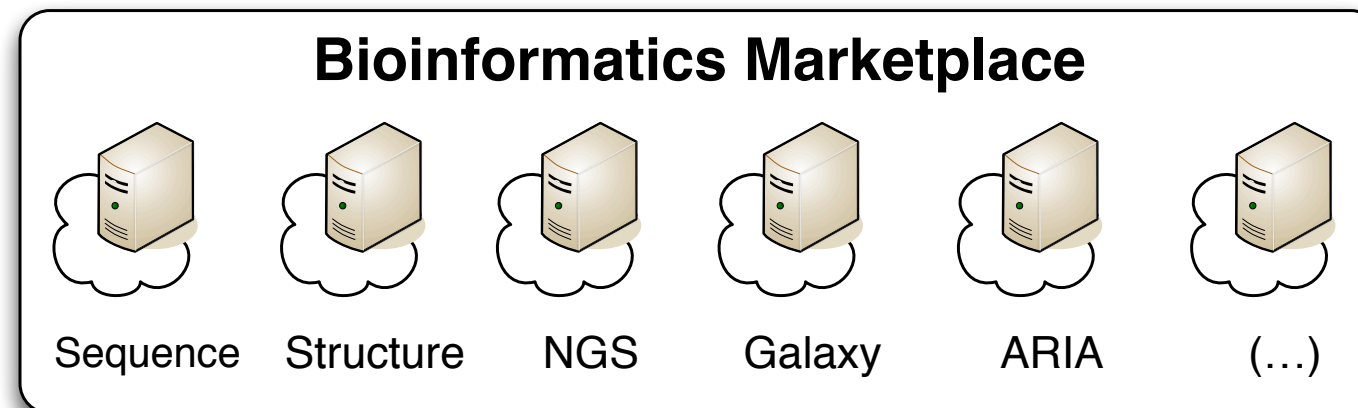
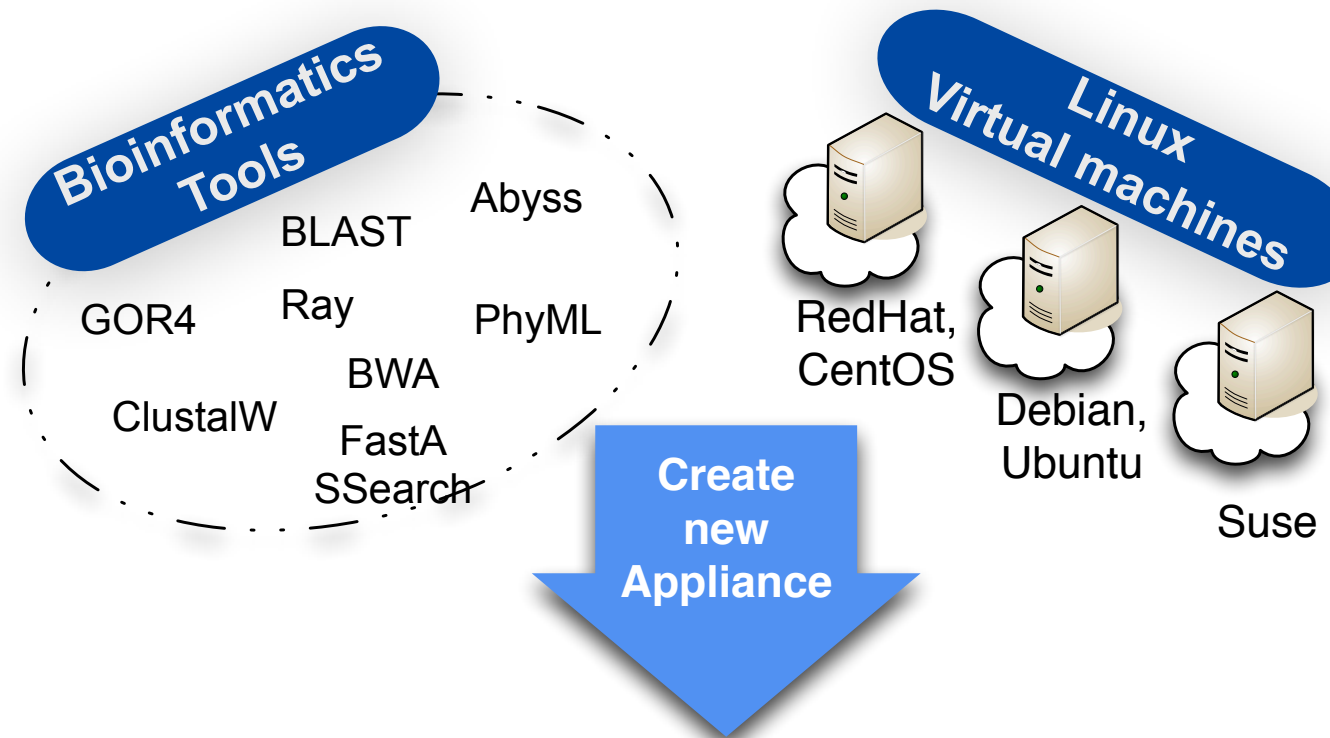


# The scene

- Core facility providers
  - Is it easy to deploy lot of (incompatible) tools ?
  - To make them connected to public databases ?
  - To limit transfer of huge data ?
  - To provide users with their own computing resources ?
  - With their own isolated storage ?
- Scientific users
  - Is it easy to access/use these tools ?
  - To adapt to your usage ?
  - To get your/other tools deployed on a datacenter ?
  - To combine them ?



# Integrate Bioinformatics Tools in Cloud



- Predefined virtual machines
  - small : few GB, easy to convert in most virtualization formats
- Installed and pre-configured with common bioinformatics tools
  - e.g. BLAST, Clustalw, ARIA, MEME, HMMer, TopHat, BWA, Samtools, etc.

# IDB's Cloud Workbench

- Demonstrate usefulness of cloud for Biology
- 13 turnkey bioinformatics appliances (as of Nov. 2012)
- Running since Sept. 2011, opened to Biology community
  - 184 cores, 536 GB RAM, 36 TB de stockage
  - Powered by Toolkit StratusLab (EU FP7 INFSO-RI-261552) Rel. 1.4
- Specific developments of a Web interface ready for biologists

You are signed in as cblanchet | [Settings](#) | [Home](#) | [Help](#) | [Sign out](#)

**idéeB**  
infrastructure distribuée pour la Biologie

**Bioinformatics cloud**

Powered by **StratusLab**

Instance

Shutdown Go

<input type="checkbox"/>	ID	Name	State	Appliance	CPU%	CPU	Mem. (GB)	Storage	Port translation
<input type="checkbox"/>	1149	test	Running	BioData	1%	1	4		<a href="#">http</a>
<input type="checkbox"/>	1199	upg	Running	ARIA2.3	4%	4	16		ssh
<input type="checkbox"/>	1239	qr7	Running	BioCompute	10%	2	8		ssh <a href="#">http</a>
<input type="checkbox"/>	1258	cos6	Running	CentOS 6.2	0%	2	8		ssh

New Instance Refresh

**Room for VMs**

xsmall	245 / 284
small	142 / 160
medium	68 / 76
large	32 / 34
xlarge	15 / 16
bigmem	2 / 2
xxl	9 / 16
htc	6 / 8

Cpu



# Bioinformatics Appliances

## ARIA 2.3

Endorser: [christophe.blanchet@ibcp.fr](mailto:christophe.blanchet@ibcp.fr)  
Identifier: [N\\_zDsconV86gvkjZtt7D-ePv4M6](#)  
Created: 2012-10-11T14:01:38Z

This appliance is part of the StratusLab bioinformatics usecase TOSCANI (TOWards StruCTural Assignment Improvement). The goal is to improve the determination of protein...

[More...](#)

## CentOS 6

Endorser: [christophe.blanchet@ibcp.fr](mailto:christophe.blanchet@ibcp.fr)  
Identifier: [B18HibMjBB6uu231adQUkqyGtnl](#)  
Created: 2012-09-27T12:04:15Z

A minimal installation for CentOS 6.x. Only root account configured. Firewall enabled with SSH and HTTP port open. SELinux disabled. Enhanced StratusLab contextualization used...

[More...](#)

## BIO data

Endorser: [christophe.blanchet@ibcp.fr](mailto:christophe.blanchet@ibcp.fr)  
Identifier: [FtCJFZ7xO5uxKyzThGRX9Ex5cqR](#)  
Created: 2012-09-24T08:56:26Z



Biological databases repository appliance built by IDB-IBCP (CNRS, Lyon, France. <http://idee-b.ibcp.fr>). The following databases are installed and available: SwissProt,...

[More...](#)

## Mobyle

Endorser: [christophe.blanchet@ibcp.fr](mailto:christophe.blanchet@ibcp.fr)  
Identifier: [NaCGZfy9NxClc3ISU158RHrG0ik](#)  
Created: 2012-09-07T14:20:38Z

This appliance provides cloud users with a fully functional Mobyle portal. Mobyle is a framework and web portal



# Bioinformatics Appliances (2)

## Protein identification

*Endorser:* christophe.blanchet@ibcp.fr  
*Identifier:* H6KPqxYIZRdlhPhs2ZKIENiiVyx  
*Created:* 2012-10-23T14:05:07Z

Bioinformatics virtual appliance for protein identification from mass spectrometry data. Contains OMSSA, X!Tandem, PeptideShaker and SearchGUI tools. Constructed by IDB...

[More...](#)

## Galaxy portal

*Endorser:* christophe.blanchet@ibcp.fr  
*Identifier:* OqucGN3bQD9FdlenGRlqZ4ZNNHW  
*Created:* 2012-10-11T15:11:59Z

Bioinformatics gateway appliance configured with the GALAXY portal, built by CNRS IBCP-IDB. You have also access to the pre-installed bioinformatics tools through the web...

[More...](#)

## Hadoop MapReduce

*Endorser:* clement.gauthey@ibcp.fr  
*Identifier:* PEIfkAp5mOwULVh1KLsprFcji0s  
*Created:* 2012-10-11T14:42:36Z

This appliance provides an easy way to deploy an Hadoop MapReduce cluster. You just need to run the bash script `hadoop-create-cluster` with a nodes list and an username in parameters and wait few...

[More...](#)

## BIO compute node

*Endorser:* christophe.blanchet@ibcp.fr



# Select your bioinformatics tools

https://idb-cloudweb.ibcp.fr/cloud/

You are signed in as cblanchet | [Settings](#) | [Home](#) | [Help](#) | [Sign out](#)

### Run Instance

Choose the appliance  
Select ? **ARIA2.3**

Filter by:

- thematic fields
- tools

- Genomics tools
- Molecular structural analysis**
- Multiple Sequence Alignment
- Nucleotide and Protein sequence searching
- Public databases
- Sequence analysis

Configure your virtual machine

Name ?

Type ? **small (1 CPU, 4GB RAM)**

Number ? **1**

Storage ?

**Run** **Cancel**

Powered by **StratusLab**

Refresh **Room for VMs**

xsmall	247 / 284
small	145 / 160
medium	69 / 76
large	32 / 34
xlarge	15 / 16
bigmem	2 / 2
xxl	10 / 16
htc	6 / 8

Instance them (25.00%)

Cpu

Memory

Une erreur lors de l'ouverture de la page. Pour en savoir plus, choisissez Fenêtre > Activité.



# Select your bioinformatics tools

Run Instance

Choose the appliance

Filter by:

- thematic fields
- tools

Configure your machines

small (1 CPU, 4GB RAM)

1

Run Cancel

Run Instance

Choose the appliance

Select ? BioCompute

Filter by:

- thematic fields
- tools

Configure your machines

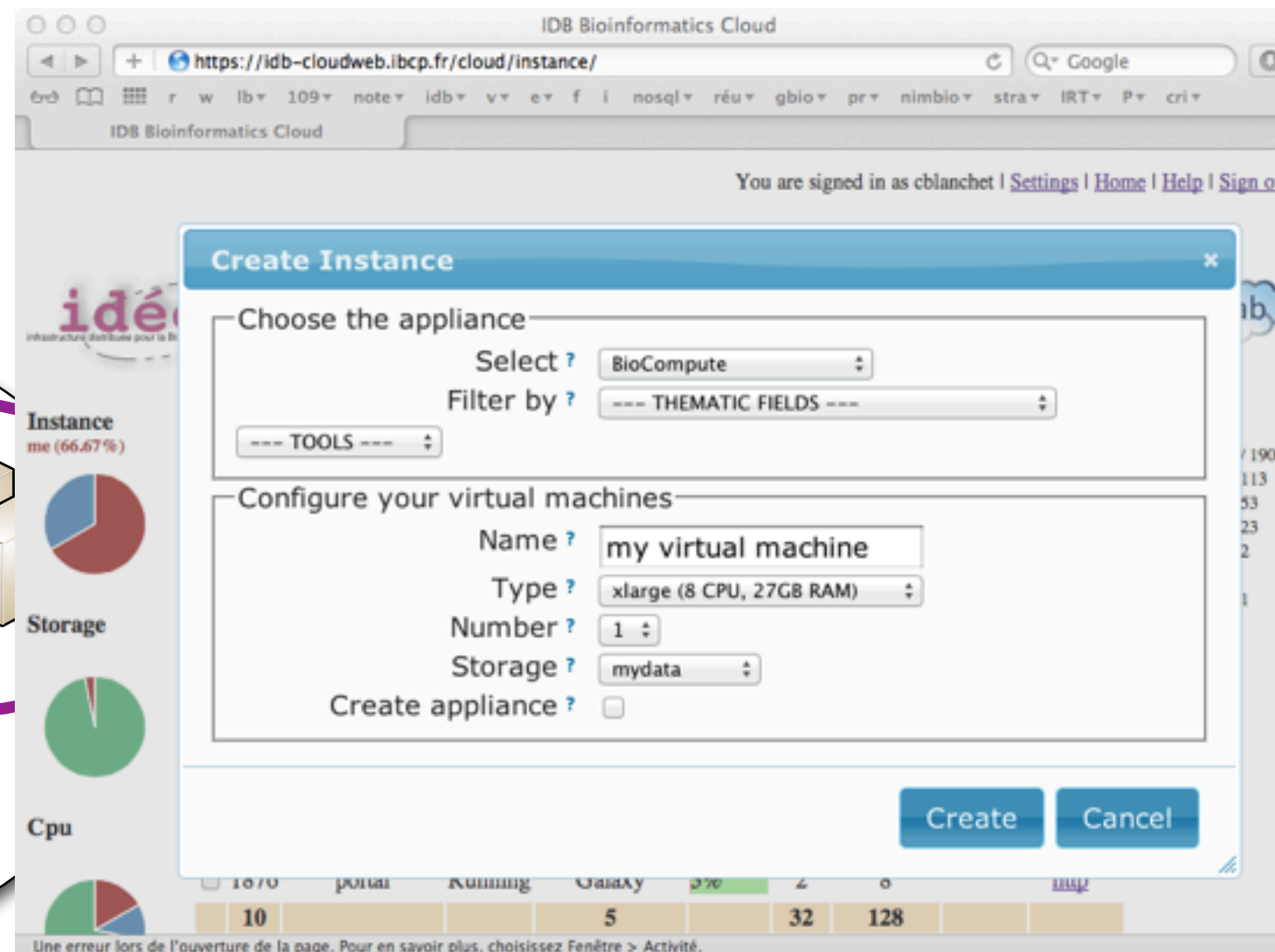
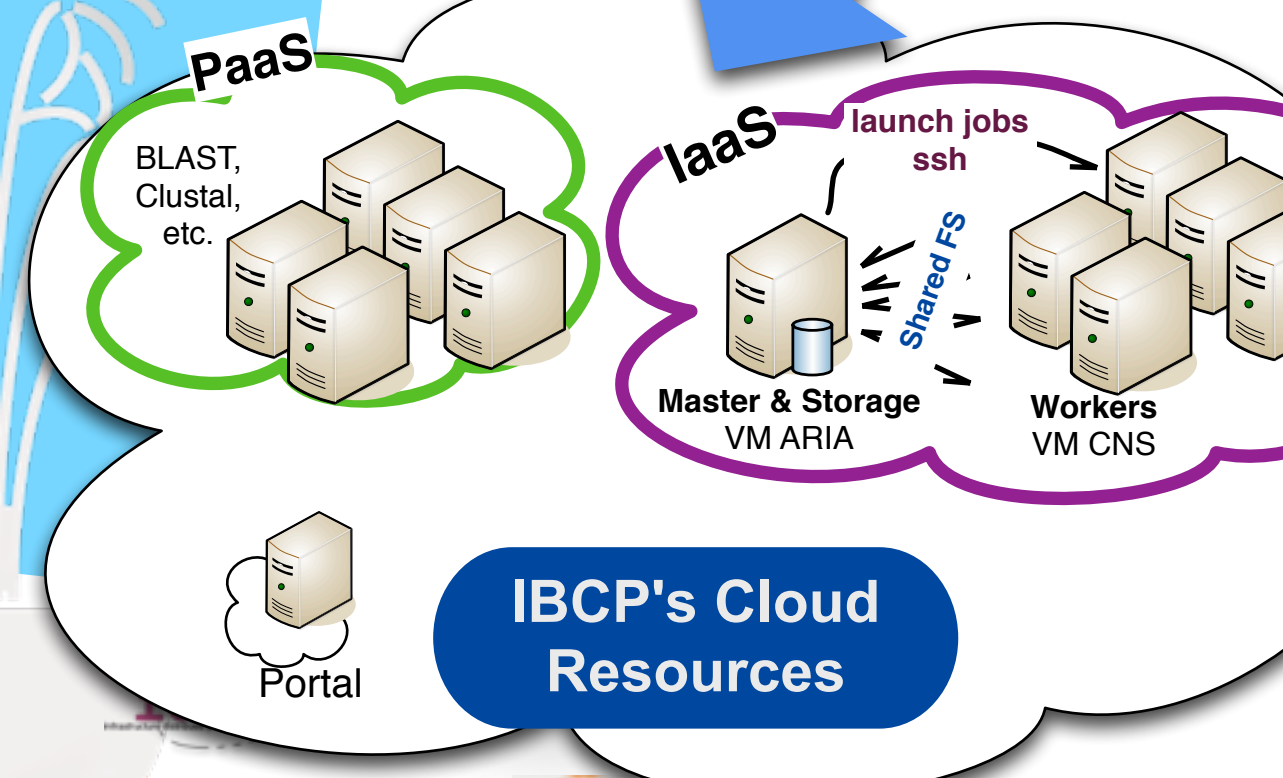
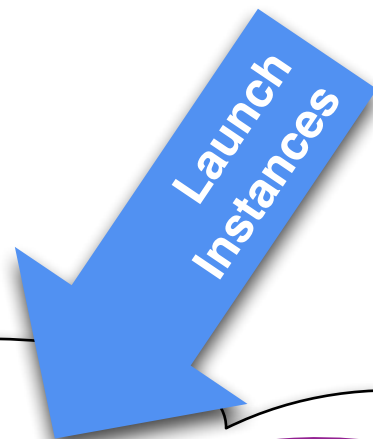
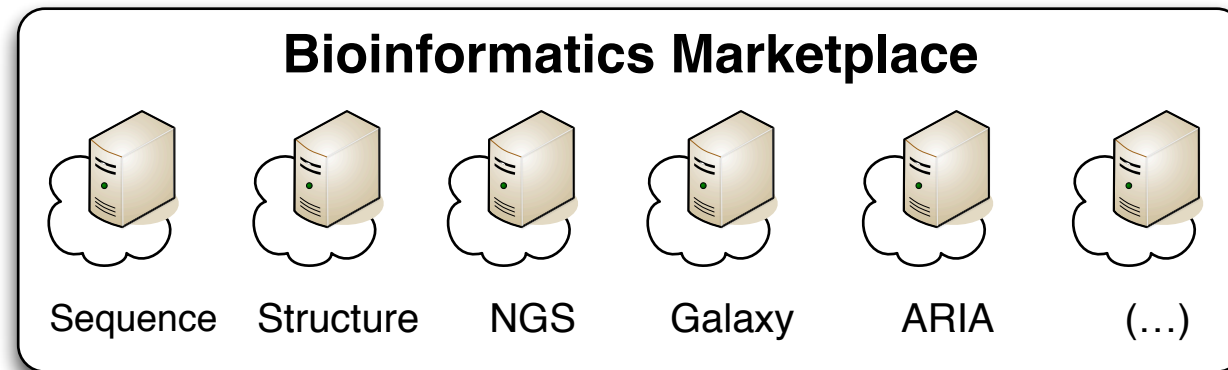
small (1 CPU, 4GB RAM)

1

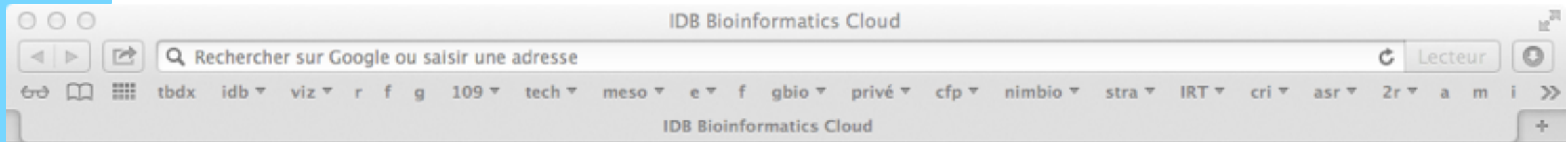
Run Cancel

- ABYSS 1.3.2
- ARIA 2.3
- BLAST 2.2.25**
- BWA 0.5.8c
- CAP3
- CLUSTALW 2.1
- FastA 3.6
- HMMer 3.0
- MEME 4.7
- PREDATOR 2.1.2
- Ray 1.3
- XPLOR-NIH 2.30
- biomaj
- galaxy

# Run Bioinformatics Cloud Instances



# Manage your Cloud Instances



You are signed in as cblanchet | [Settings](#) | [Home](#) | [Help](#) | [Sign out](#)



## Bioinformatics cloud



### Instance

Shutdown Go Get IPs

New Instance New Storage Show Instances Show Storages

Showing 1 to 10 of 10 entries

Search:

<input type="checkbox"/>	ID	Name	State	Appliance	CPU%	CPU	Mem.	Storage	Access	<input type="checkbox"/>
<input type="checkbox"/>	2729	omssa mas115	Running	Protein Identification	0%	8	27		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2737	struct det	Running	ARIA2.3	1%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2738	struct det	Running	ARIA2.3	1%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2739	struct det	Running	ARIA2.3	1%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2740	struct det	Running	ARIA2.3	1%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2741	struct det	Running	ARIA2.3	2%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2742	struct det	Running	ARIA2.3	0%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2743	struct det	Running	ARIA2.3	1%	2	8		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2744	prot seq	Running	BIO compute node	0%	4	16		ssh	<input type="checkbox"/>
<input type="checkbox"/>	2746	CCMP	Running	Protein Identification	47%	24	48	omssa nr	ssh	<input type="checkbox"/>
		10			3	50	147			

Show 25 entries

First Previous 1 Next Last

### Room for VMs

xsmall	251 / 341
small	134 / 202
medium	66 / 101
large	29 / 49
xlarge	16 / 23
bigmem	4 / 4
xxl	12 / 21
htc	9 / 10
htc+	2 / 2
g201-half	18 / 26
g201-full	8 / 9

### Storage

### Cpu

### Memory





# Data in cloud

**Upload your data**

scp http

**PaaS**

BLAST,  
Clustal,  
etc.

**IaaS**

launch jobs  
ssh

Master & Storage  
VM ARIA

Workers  
VM CNS

Shared FS

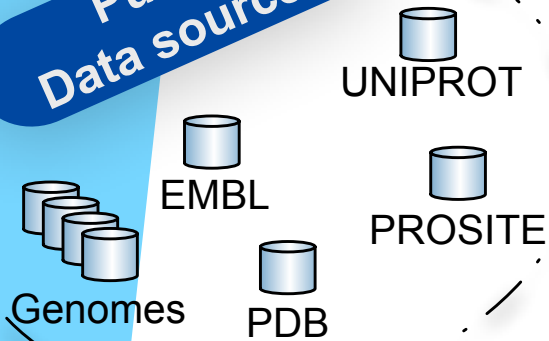
Portal

**Bioinformatics  
Cloud**

scp http

**Get your results**

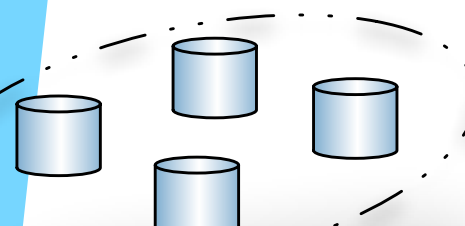
**Public  
Data sources**



shared  
(NFS)

pdisk  
(iSCSI)

**User  
Persistent data**



# Examples

- Structural Biology: the TOSCANI usecase
- Galaxy portal for NGS analyses
- Proteomics

# Structural Biology

- **TOWARDS STRUCTURAL ASSIGNMENT IMPROVEMENT**
  - To improve the determination of protein structures based on Nuclear Magnetic Resonance (NMR) information with ARIA software
  - Large computational needs.
  - A NMR laboratory will not specially invest in building a cluster of about 100 nodes to be able to run such NMR structure calculations.
  - Flexibility of the cloud to deploy the different required bioinformatics tools can accelerate such a procedure.
  - Commercial interest in providing such tools to structural biologists on a “pay as you go” basis.

- *Endorsers:*  
*Institut Pasteur Paris*  
*and CNRS IBCP*

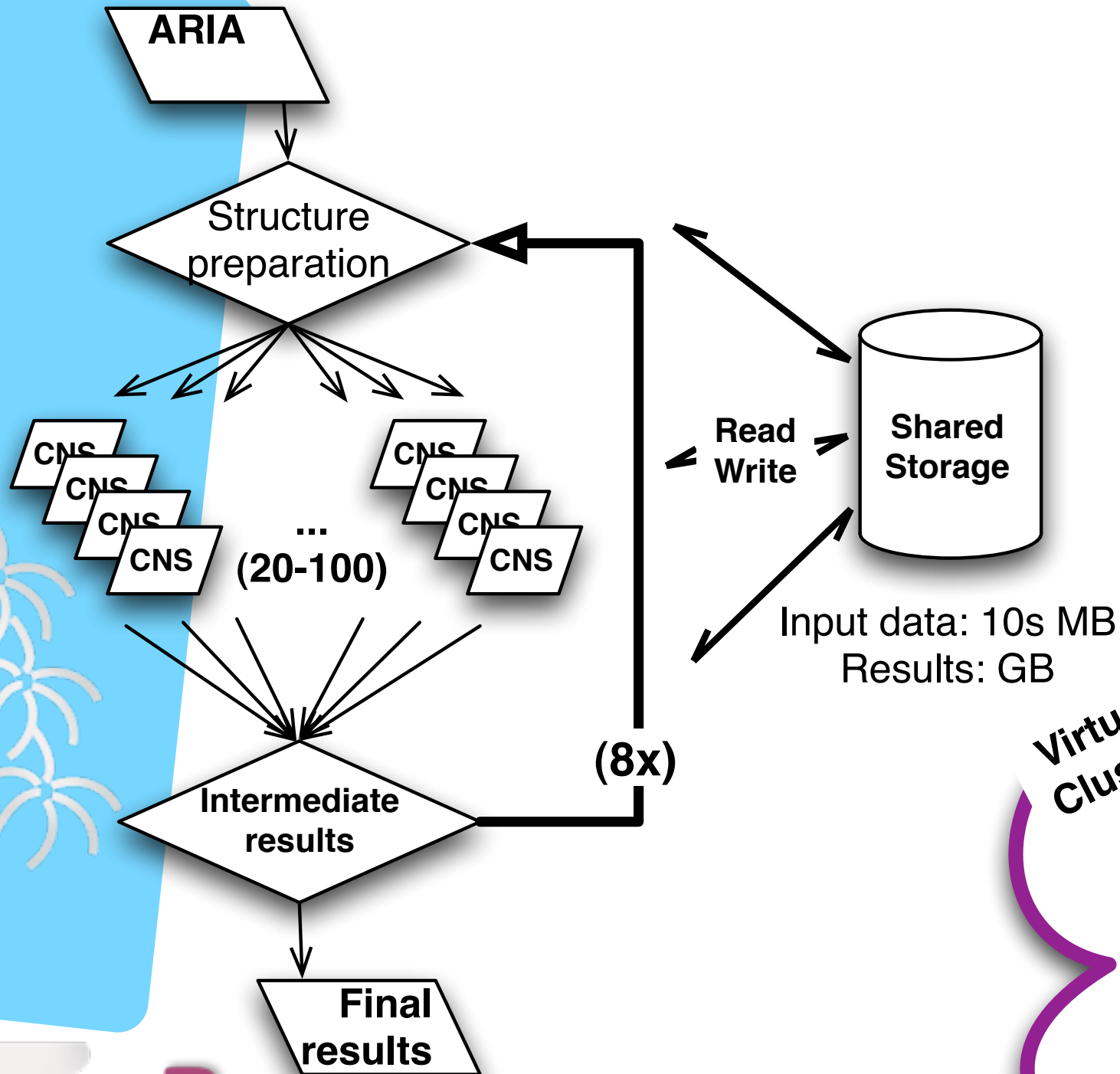


A screenshot of the ARIA website's home page. At the top, there is a navigation menu with links for Home, Documentation, Download, Related links, News, FAQ, Developers, Example files, and Patch files. Below the menu, a breadcrumb trail reads "You are here: Home". On the left side, there is a "Navigation" sidebar with a list of links: Home, Documentation, Download, Related links, and News. The main content area features a "Welcome to ARIA" section with a brief description of the software's purpose: "Our computer program ARIA (Ambiguous Restraints for Iterative Assignment) is a software for automated NOE assignment and NMR structure calculation. It speeds up and automatizes the assignment process through the use of an iterative structure calculation scheme. Additionally, a refinement in explicit water improves the quality of the calculated structures, validation tests help spectroscopists to judge the quality of the final structures, and the support of the CCPN data model simplifies the exchange of information with other NMR software packages."

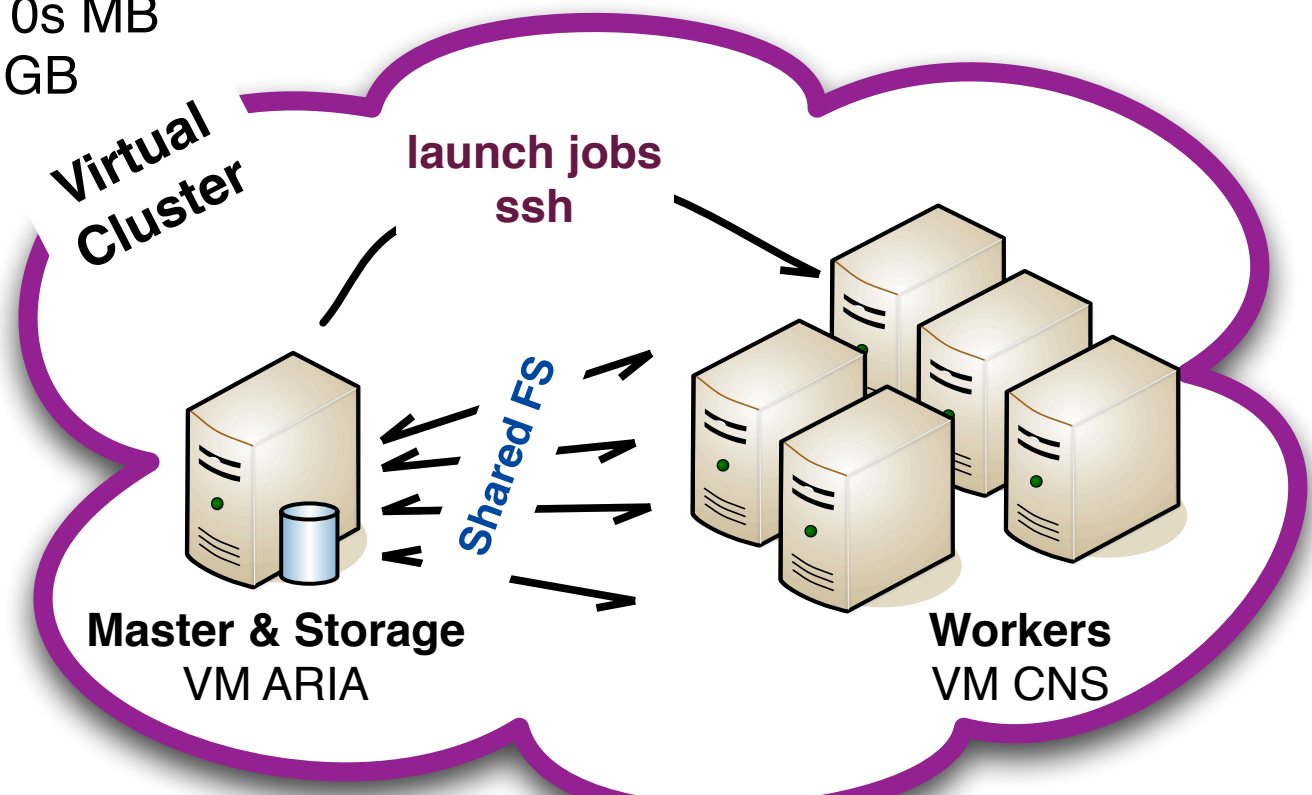




# IaaS deployment of ARIA



Significant increase in the number of calculated protein conformations improves the statistics on the NMR conformations and can help to overcome the ambiguity bottleneck.



# Galaxy portal for NGS analyses

<input type="checkbox"/>	1874	blast	Running	BioCompute	0%	2	8	ssh <a href="#">http</a>
<input type="checkbox"/>	1875	blast	Running	BioCompute	2%	2	8	ssh <a href="#">http</a>
<input type="checkbox"/>	1876	portal	Running	Galaxy	2%	2	8	<a href="#">http</a>
<input type="checkbox"/>	1877	blast machine	Running	BioCompute	0%	4	16	data ssh <a href="#">http</a>



# Galaxy portal for NGS analyses

<input type="checkbox"/>	1874	blast	Running	BioCompute	2%	2	8	<a href="#">ssh http</a>
<input type="checkbox"/>	1875	blast	Running	BioCompute	2%	2	8	<a href="#">ssh http</a>
<input type="checkbox"/>	1876	portal	Running	Galaxy	2%	2	8	<a href="#">http</a>
<input type="checkbox"/>	1877	blast machine	Running	BioCompute	2%	4	16	<a href="#">data ssh</a>



# Galaxy portal for NGS analyses

<input type="checkbox"/>	1874	blast	Running	BioCompute	0%	2	8	<a href="#">ssh</a> <a href="#">http</a>
<input type="checkbox"/>	1875	blast	Running	BioCompute	2%	2	8	<a href="#">ssh</a> <a href="#">http</a>
<input type="checkbox"/>	1876	portal	Running	Galaxy	2%	2	8	<a href="#">http</a>
<input type="checkbox"/>	1877	blast machine	Running	BioCompute	0%	4	16	<a href="#">data</a> <a href="#">ssh</a>

Galaxy  
 Analyze Data Workflow Shared Data Help User

Tools Options

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Batmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server

Upload File

File Format: Auto-detect

Which format? See help below

File: Choisir le fichier **fewSeqs.fasta**

TIP: Due to browser limitations, uploading files large 2GB is guaranteed to fail. To upload large files, use the method (below) or FTP (if enabled by the site administrator)

URL/Text:

Here you may specify a list of URLs (one per line) or the contents of a file.

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand

Genome: Click to Search or Select

**Execute**

Auto-detect

Galaxy  
 Analyze Data Workflow Shared Data Help User

Tools Options

Get Data

- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Abstract and Group
- Open Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- ClustalW multiple sequence alignment program for DNA or proteins
- Metagenomic analyses
- FASTA manipulation

ClustalW

Fasta File: 1: fewSeqs.fasta

Name for output files to make it easy to remember what you did: Clustal\_run

Data Type: DNA nucleotide sequences

Output alignment format: Native Clustal output format

Show residue numbers in clustal format output: no

Output Order: aligned

Output complete alignment (or specify part to fetch): complete alignment

**Execute**

Note

This tool allows you to run a multiple sequence alignment using ClustalW2 (see ClustalW) using the default options

History Options

2.5 Kb

1: fewSeqs.fasta

6 sequences  
 format: fasta, database: ?  
 Info: uploaded fasta file

Galaxy  
 Analyze Data Workflow Shared Data Help User

Options

CLUSTAL 2.1 multiple sequence alignment

```

sp|P0ACP4|FRUR_SHIFL
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP1|FRUR_ECOLI
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP2|FRUR_ECOL6
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP3|FRUR_ECO57
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0A2P8|FRUR_SALTY
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0A2P9|FRUR_SALTI
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
  
```

# Galaxy portal for NGS analyses

<input type="checkbox"/>	1874	blast	Running	BioCompute	0%	2	8	<a href="#">ssh</a> <a href="#">http</a>
<input type="checkbox"/>	1875	blast	Running	BioCompute	2%	2	8	<a href="#">ssh</a> <a href="#">http</a>
<input type="checkbox"/>	1876	portal	Running	Galaxy	2%	2	8	<a href="#">http</a>
<input type="checkbox"/>	1877	blast machine	Running	BioCompute	0%	4	16	<a href="#">data</a> <a href="#">ssh</a>

Galaxy  
http://idb-cloud.ibcp.fr:20007/

Tools  
search tools

Get Data  

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Batmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server

Upload File  
 File Format: Auto-detect  
 Which format? See help below  
 File: Choisir le fichier **fewSeqs.fasta**  
 URL/Text:  
 Convert spaces to tabs:  
 Yes  
 Genome: Click to Search or Select

Galaxy  
http://idb-cloud.ibcp.fr:20007/

Tools  

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Abstract and Group
- Open Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
  - ClustalW multiple sequence alignment program for DNA or
- Metagenomic analyses
- FASTA manipulation

ClustalW  
 Fasta File: 1: fewSeqs.fasta  
 Clustal\_run  
 Data Type: DNA nucleotide sequences  
 Output alignment format: Native Clustal output format  
 Show residue numbers in clustal format output: no  
 Output Order: aligned  
 Output complete alignment (or specify part to): complete alignment

Galaxy  
http://idb-cloud.ibcp.fr:20007/

CLUSTAL 2.1 multiple sequence alignment

```

sp|P0ACP4|FRUR_SHIFL
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP1|FRUR_ECOLI
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP2|FRUR_ECOL6
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0ACP3|FRUR_ECO57
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0A2P8|FRUR_SALTY
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
sp|P0A2P9|FRUR_SALTI
MKLDEIARLAGVSRRTTASYVINGKAKQYRVSDKTVEKMAVVREHNYHPNAVAAGL
    
```

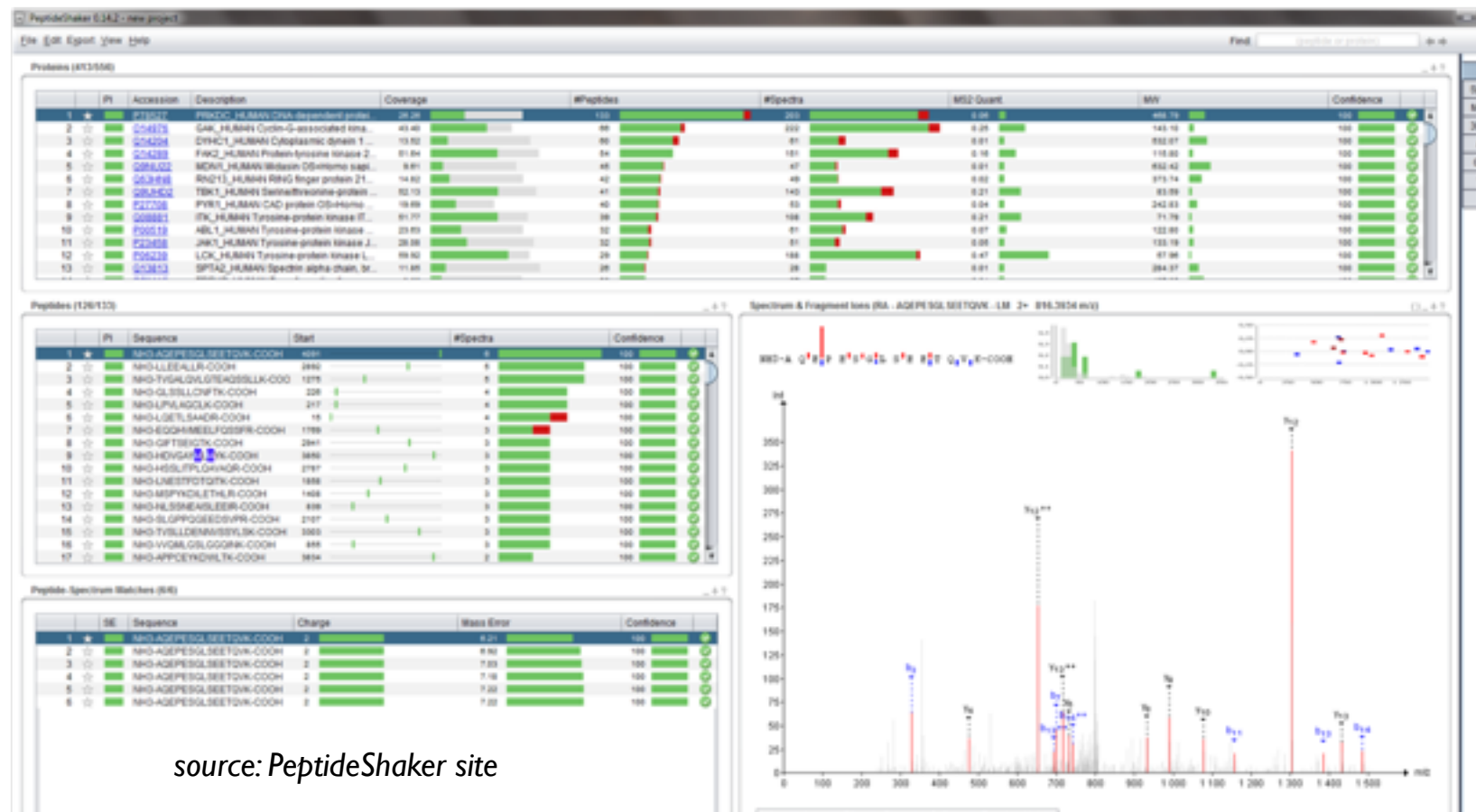


# Proteomics

- Motivation
  - Mass spectrometry platform
  - Running out of space on their local resources
- Protein identification
  - Mass spectrometry experimental data
  - reference databases : nr, Swiss-Prot
  - Screening tools: OMSSA, X!Tandem
- User interface
  - Remote display
  - Common GUIs
    - SearchGUI
    - PeptidShaker

OMSSA

X!



source: PeptideShaker site





# Conclusion

- Provide turnkey bioinformatics cloud services
  - Standard tools and pipelines
  - Ready to run on core facilities and commercial datacenters
  - Easier to transfer appliances than data (GB vs TB)
- Cloud infrastructure tightly connected to existing bioinformatics infrastructure
  - Public IDB's bioinformatics cloud
  - Linked to public biological databases
  - Collaboration with national Research Infrastructures like the French RENABI GRISBI
- Ease the access
  - Web interface for cloud management
  - Usual scientific gateways
  - Persistent and large ubiquitous storage



# What's next ?

- Help bioinformatics centres to provide academic and commercial community with bioinformatics services!
  - Pre-defined bioinformatics appliances, webservices and portal (PaaS)
  - Multi-nodes applications, e.g. ARIA, or comprehensive pipelines (IaaS)
  - Referenced in a bioinformatics marketplace
- Ready to deploy on the future cloud infrastructure of the French Bioinformatics Institute (IA ReNaBi-IFB)

# Questions ?

Platform ' Infrastructure Distributed for Biology - IDB ' acknowledges

Institut Pasteur: M Nilges, T Malliavin, F. Mareuil  
StratusLab partners

co-funding by the European Community's Seventh Framework Programme (INFSO-RI-261552) and the French National Research Agency's Arpege Programme (ANR-10-SEGI-001)

<http://idee-b.ibcp.fr>

