

StratusLab, Bioinformatics and the Cloud

Christophe Blanchet

IDB - Infrastructure Distributing Biology

IBCP CNRS FR3302 - LYON - FRANCE

christophe.blanchet@ibcp.fr

Bioinformatics Today

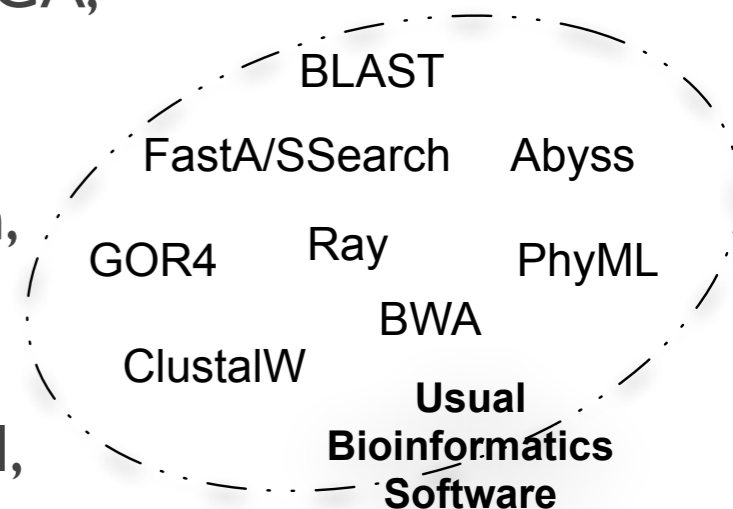
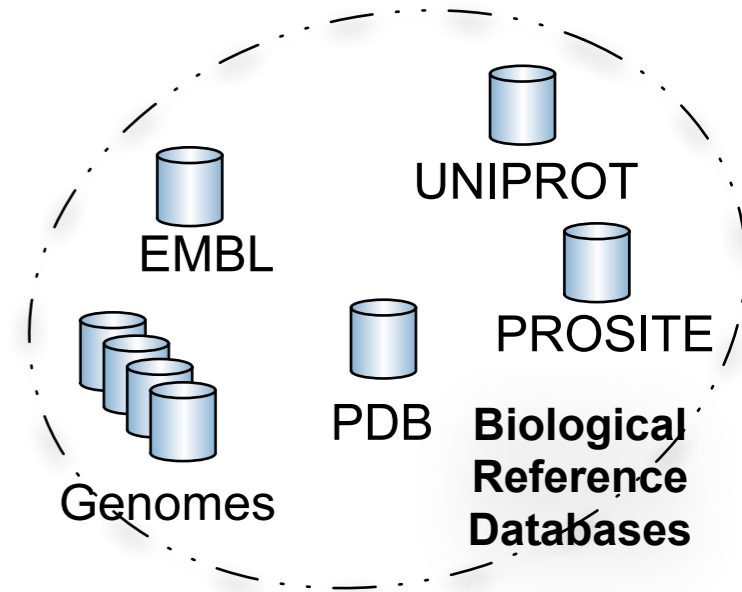
- Size of biological data are tremendous
 - Institut Sanger, UK, 5 PB
 - BGI (Beijing Genome Institute), 4 sites, 10 PB

➔ **Huge data in lot of places**
- How to analyse these data
 - Scale-up of the analyses : from gene/protein to complete genome/proteome; from one metabolic way to systems biology ...
 - Need of more and more computing resources
 - Usual interfaces: portals, Web services, federation,...

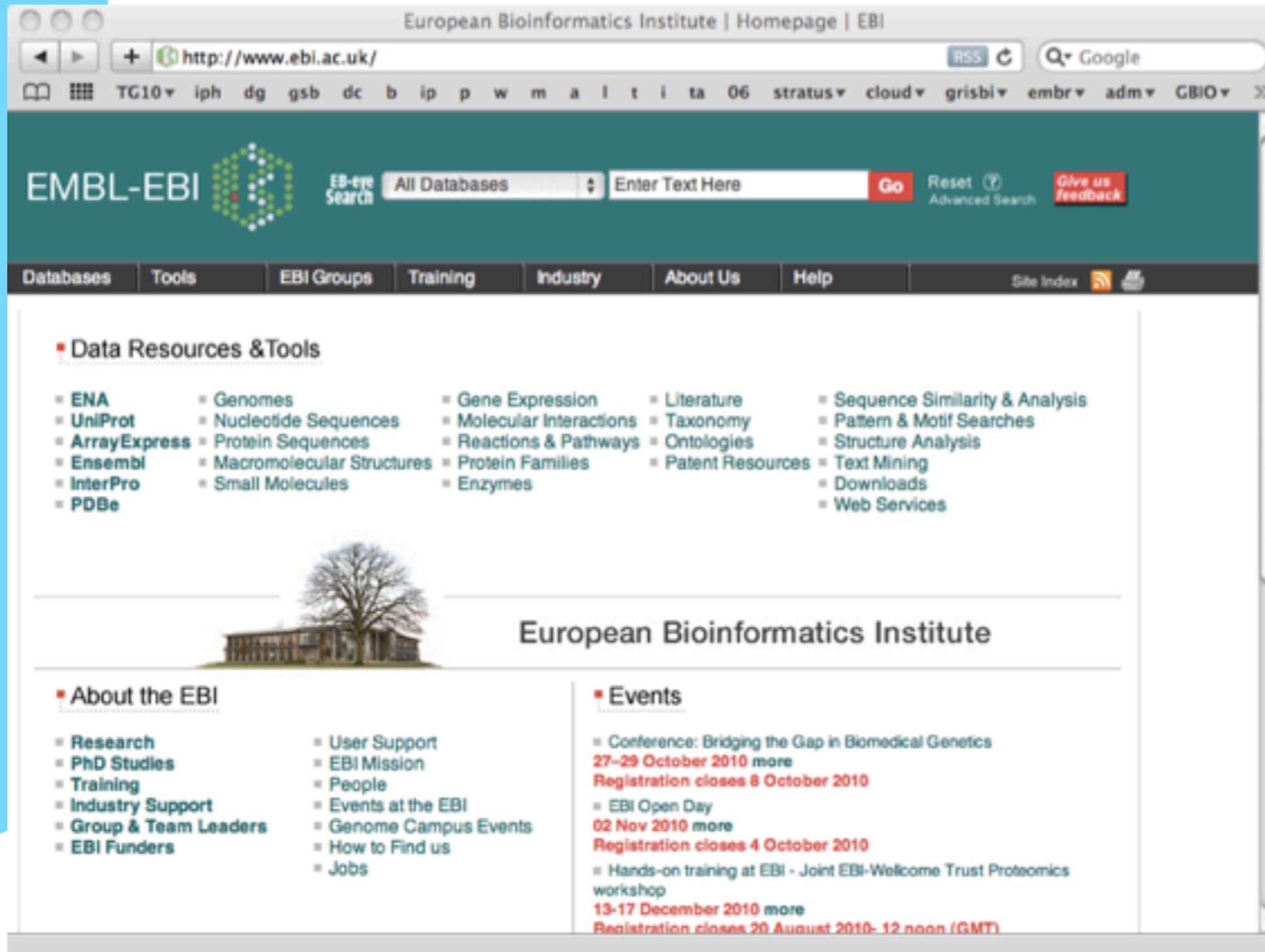
➔ **Require datacenter with ease of access/use**

Biological data & Bioinformatics Tools

- 1330 different reference data sources
 - M.Y. Galperin & G.R. Cochrane, NAR 2011
 - UniProt, Génolevures*, Base, AcNuc (**), GenBank, EMBL, PRODOM*, Ensembl, Hogenom*, Homolens*, PDB, Génomes Complets, TransFac, Nr, SRS (**), SUMO(*), PROSITE, ABC, KEGG, ...
- Thousands of different daily-used 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*, ...



Scientific Gateways



The screenshot shows the homepage of the European Bioinformatics Institute (EBI). The browser address bar displays 'http://www.ebi.ac.uk/'. The page features a search bar with 'All Databases' selected and a search input field containing 'Enter Text Here'. 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 in a grid format. Below this is a section for the 'European Bioinformatics Institute' featuring a photograph of a building. The bottom section is divided into 'About the EBI' and 'Events', each with a list of links and dates.

European Bioinformatics Institute | Homepage | EBI

http://www.ebi.ac.uk/

EMBL-EBI

EB-eye Search

All Databases

Enter Text Here

Go

Reset Advanced Search

Give us feedback

Databases Tools EBI Groups Training Industry About Us Help

Site Index

Data Resources & Tools

- 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
- Web Services

European Bioinformatics Institute

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

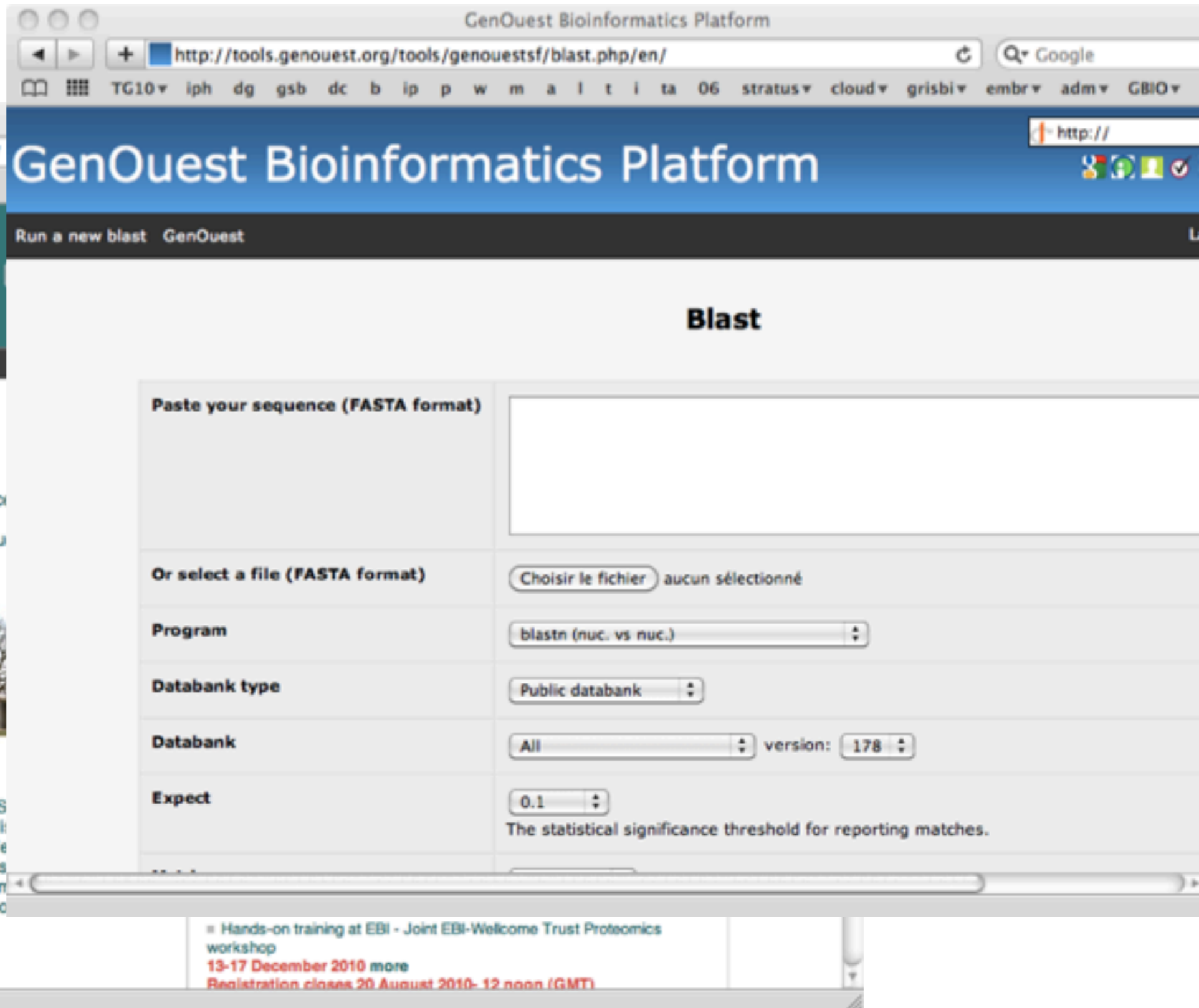
Events

- Conference: Bridging the Gap in Biomedical Genetics
27-29 October 2010 more
Registration closes 8 October 2010
- EBI Open Day
02 Nov 2010 more
Registration closes 4 October 2010
- Hands-on training at EBI - Joint EBI-Wellcome Trust Proteomics workshop
13-17 December 2010 more
Registration closes 20 August 2010- 12 noon (GMT)

Scientific Gateways

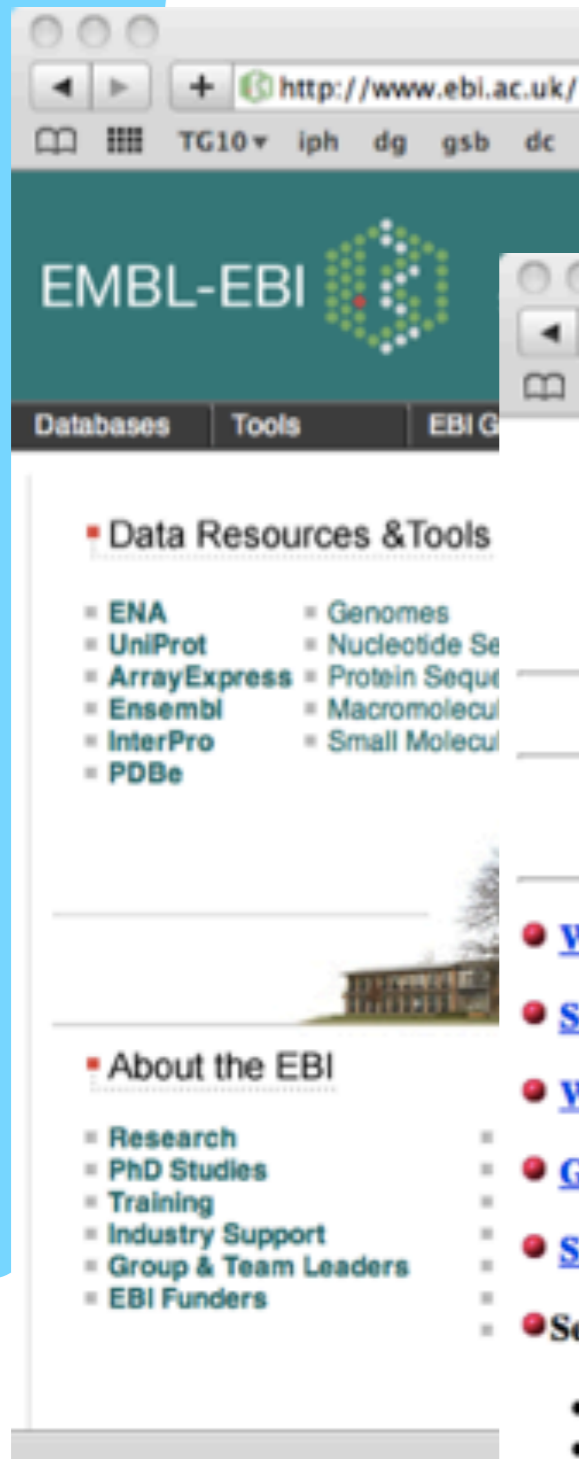


The screenshot shows the EMBL-EBI website homepage. The header includes the EMBL-EBI logo and the text "EB-eye Search". Below the header, there are navigation tabs for "Databases", "Tools", and "EBI Groups". The main content area is titled "Data Resources & Tools" and lists various resources such as ENA, UniProt, ArrayExpress, Ensembl, InterPro, PDBe, Genomes, Nucleotide Sequence, Protein Sequences, Macromolecular Structure, and Small Molecules. There is also a section for "About the EBI" with links to Research, PhD Studies, Training, Industry Support, Group & Team Leaders, EBI Funders, User Support, EBI Meetings, People, Events, Genomes, How to, and Jobs. A small image of a building is visible at the bottom of the page.



The screenshot shows the GenOuest Bioinformatics Platform Blast interface. The browser address bar displays "http://tools.genouest.org/tools/genouestsf/blast.php/en/". The page title is "GenOuest Bioinformatics Platform". Below the title, there is a navigation bar with "Run a new blast" and "GenOuest". The main heading is "Blast". The interface is divided into two columns. The left column contains the following sections: "Paste your sequence (FASTA format)" with a text input field; "Or select a file (FASTA format)" with a "Choisir le fichier" button and "aucun sélectionné" text; "Program" with a dropdown menu set to "blastn (nuc. vs nuc.)"; "Databank type" with a dropdown menu set to "Public databank"; "Databank" with a dropdown menu set to "All" and a "version: 178" dropdown; and "Expect" with a dropdown menu set to "0.1" and a note: "The statistical significance threshold for reporting matches." The right column is currently empty. At the bottom of the page, there is a footer with a logo for "IBCP" and the text "christophe.blanchet@ibcp.fr".

Scientific Gateways



EMBL-EBI

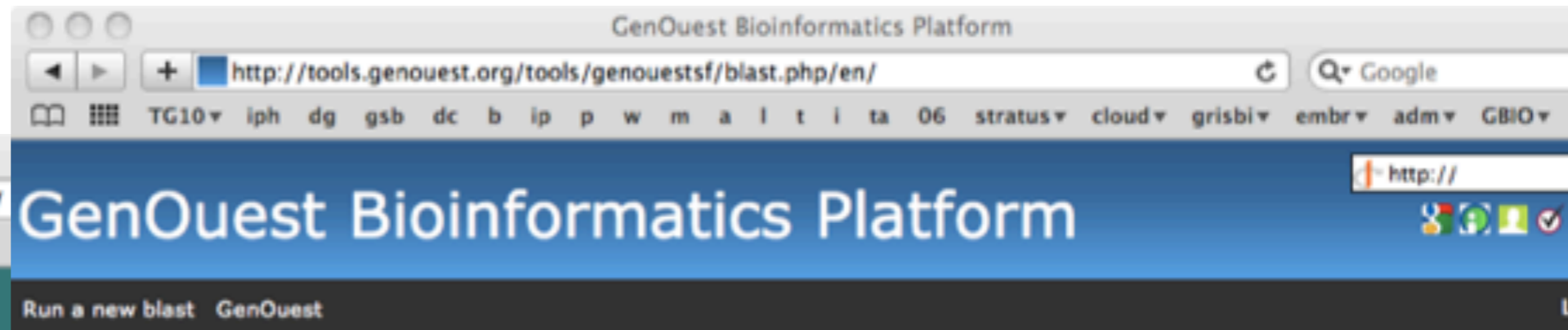
Databases Tools EBI G

Data Resources & Tools

- EN A
- UniProt
- ArrayExpress
- Ensembl
- InterPro
- PDBe
- Genomes
- Nucleotide Se
- Protein Sequ
- Macromolecul
- Small Molecul

About the EBI

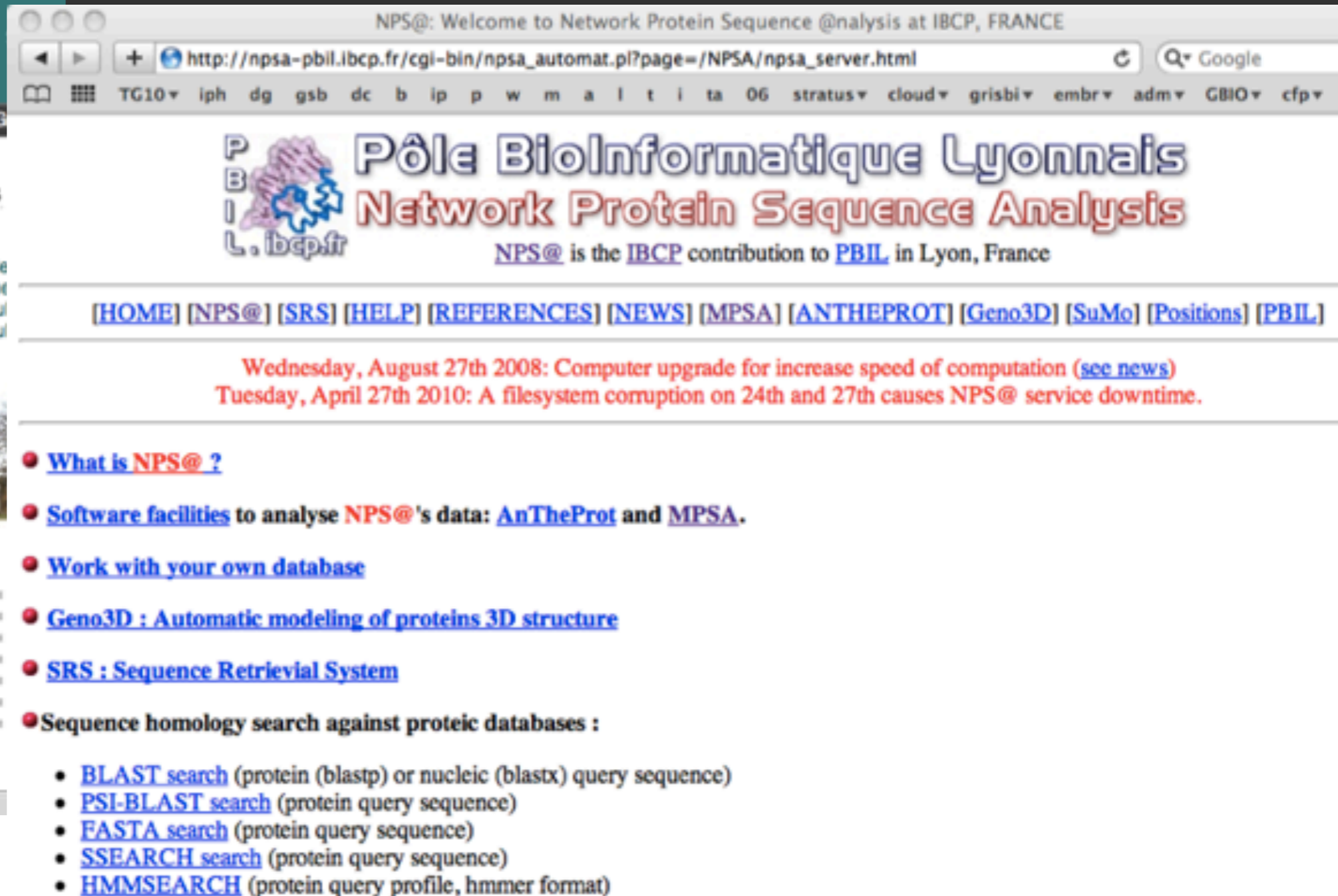
- Research
- PhD Studies
- Training
- Industry Support
- Group & Team Leaders
- EBI Funders



GenOuest Bioinformatics Platform

http://tools.genouest.org/tools/genouestsf/blast.php/en/

Run a new blast GenOuest



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

http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html

P B I L . i b c p . f r

Pôle BioInformatique Lyonnais

Network Protein Sequence Analysis

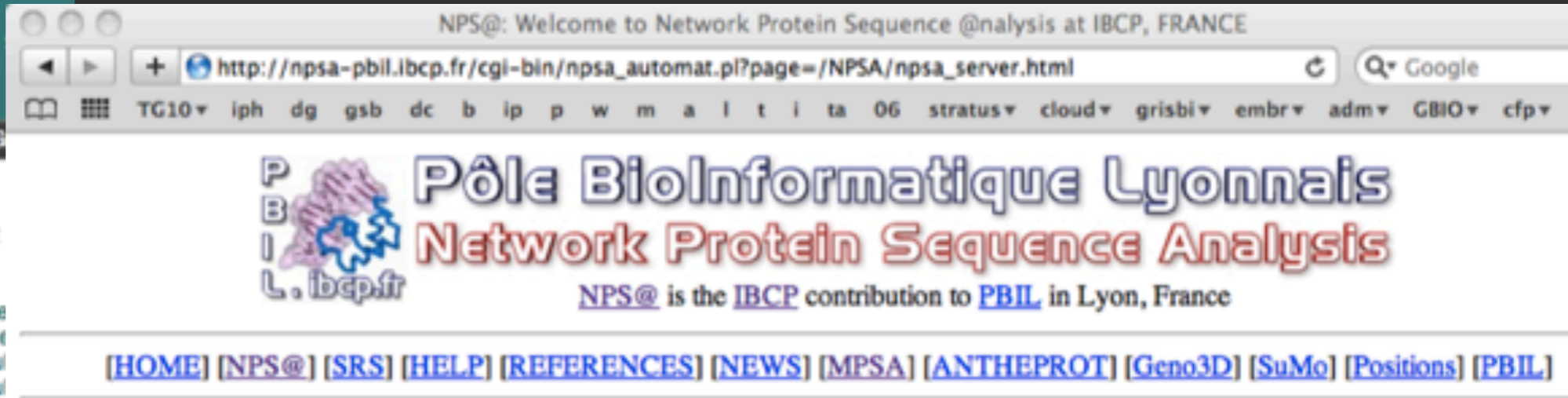
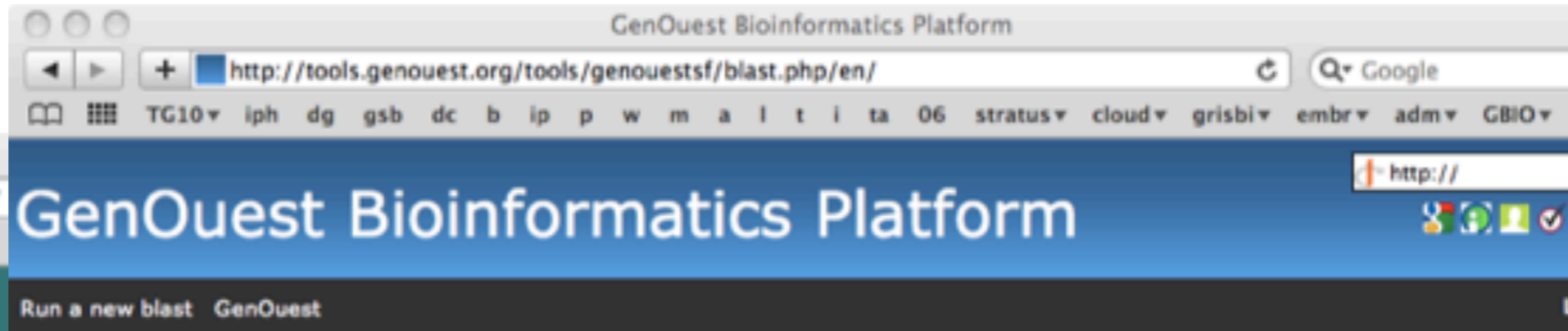
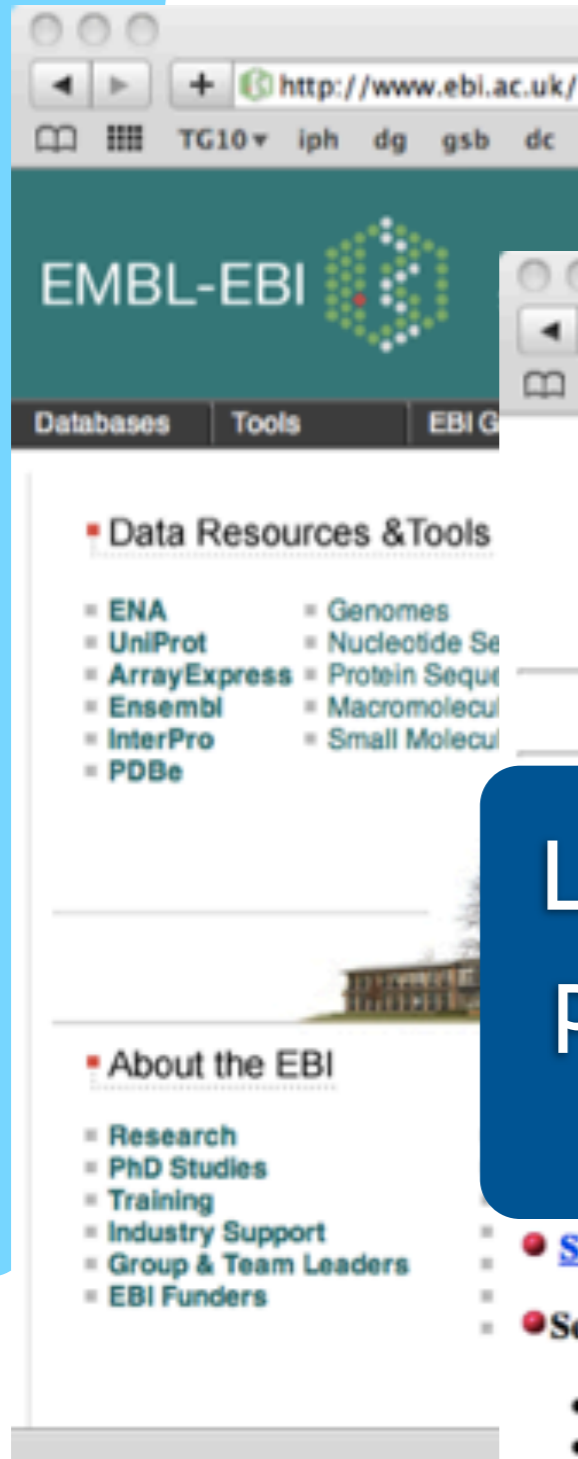
NPS@ is the IBCP contribution to PBIL in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

Wednesday, August 27th 2008: Computer upgrade for increase speed of computation ([see news](#))
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 profile, hmmer format)

Scientific Gateways



Lot of bioinformatics portals providing access to data and tools **without** authentication

- [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 profile, hmmer format)

ation (see news)
service downtime.



Scientific Gateways (2)

Web Services

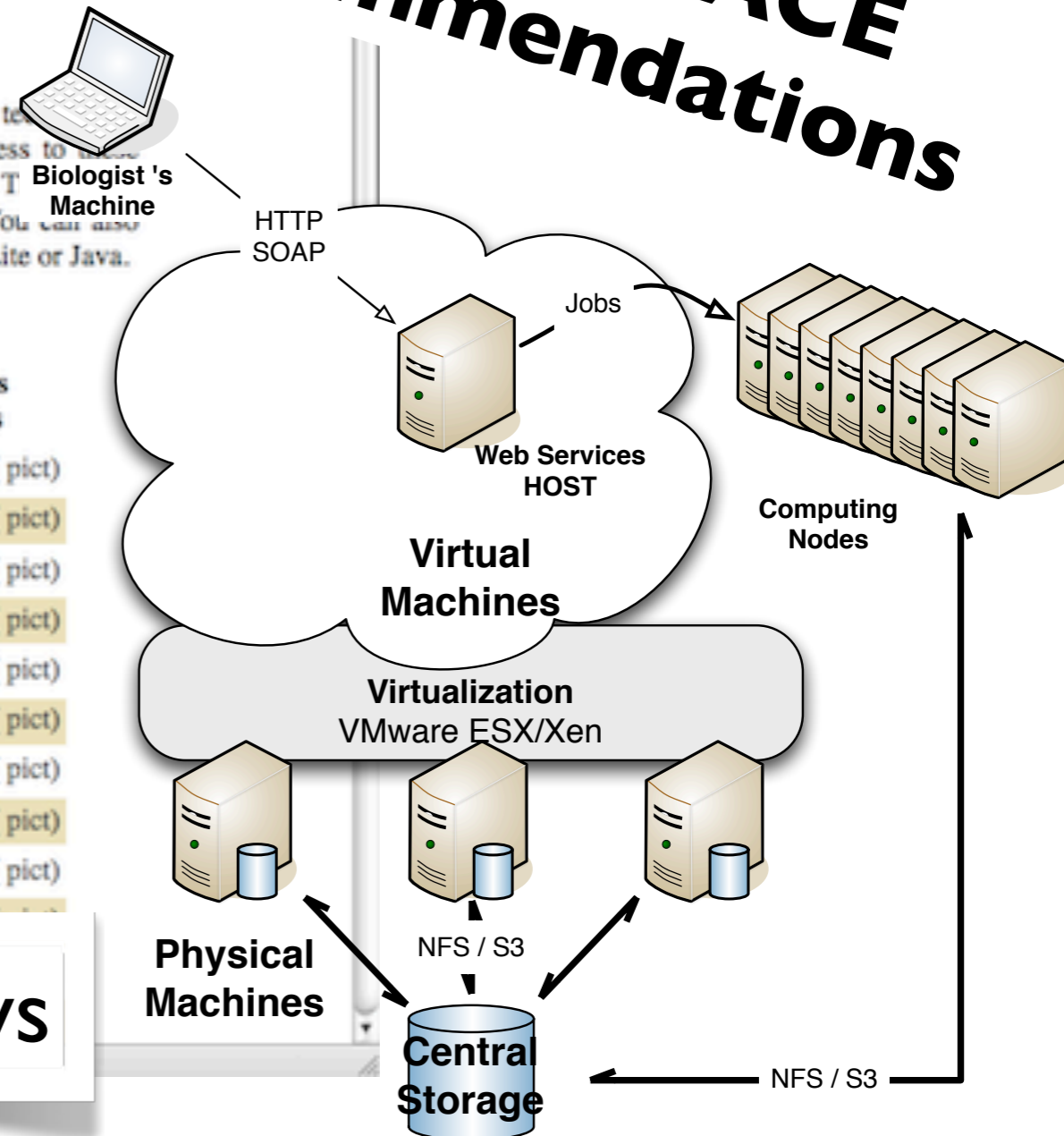
The IBCP have integrated several tools for protein sequence analysis with the Web services to... These Bioinformatics Web services provide scientists and developers with programmatic access to... tools. Our Web services are build upon standards from the W3C like SOAP, WSRF and HTTP. T... can be use remotely through a graphical and integrated SOAP client like Taverna or Triana. You... 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)
SSearch	sequence similarity	wsdl	usage	PyZSI Tav2 (pict)
Dsc	secondary structure of protein	wsdl	usage	PyZSI Tav2 (pict)
Gor I	secondary structure of protein	wsdl	usage	PyZSI Tav2 (pict)
Gor III	secondary structure of protein	wsdl	usage	PyZSI Tav2 (pict)
Gor IV	secondary structure of protein	wsdl </tr		

gbio-pbil.ibcp.fr/ws

Compliant to EU EMBRACE recommendations



Terminé



christophe.blanchet@ibcp.fr



EGI TF, 20 September 2011, Lyon

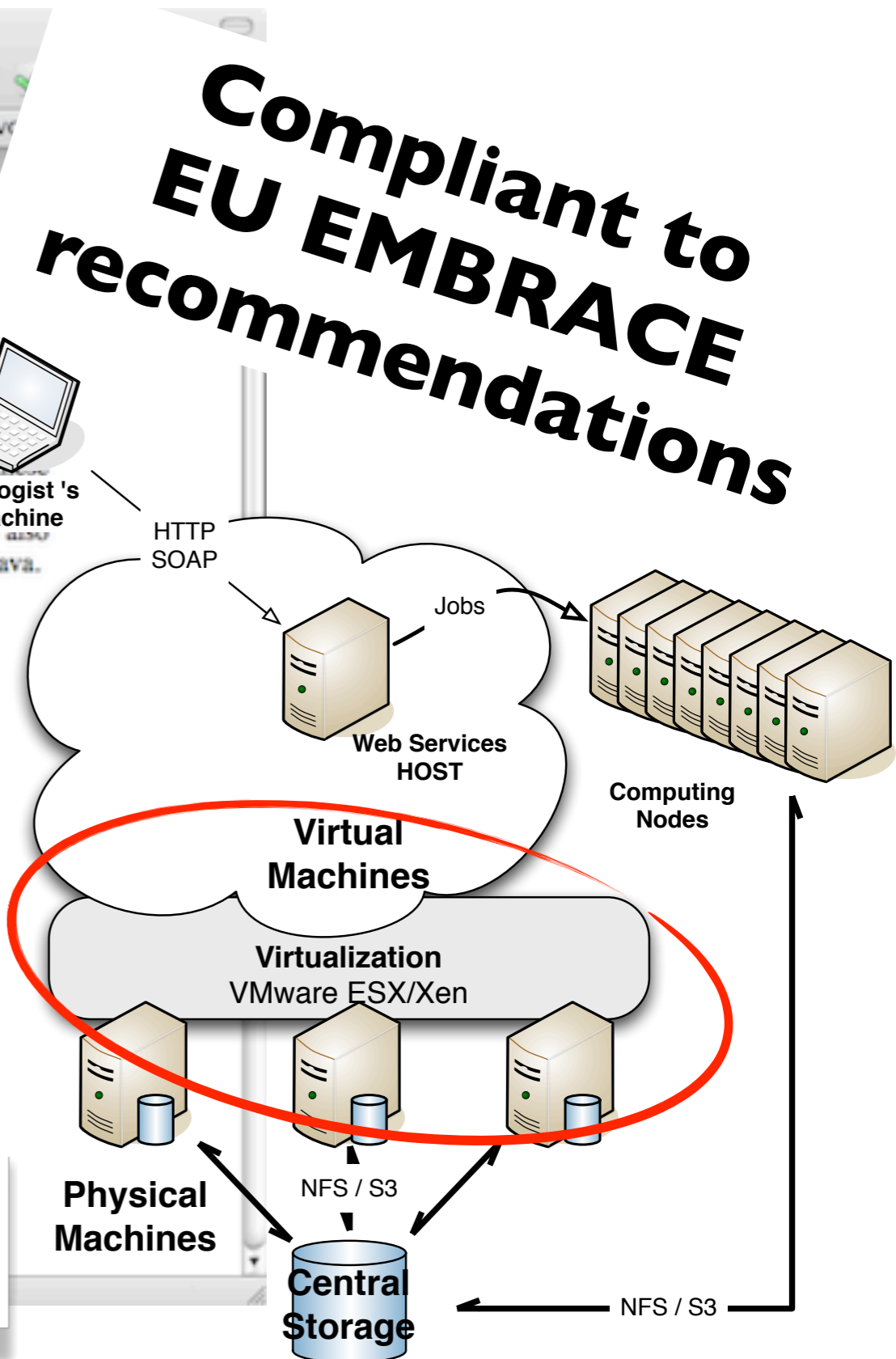
Scientific Gateways (2)

The IBCP have integrated several tools for protein sequence analysis with the Web services to... These Bioinformatics Web services provide scientists and developers with programmatic access to... tools. Our Web services are build upon standards from the W3C like SOAP, WSRF and HTTP. T... can be use remotely through a graphical and integrated SOAP client like Taverna or Triana. You... 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)
SSearch	sequence similarity	wSDL	usage	PyZSI Tav2 (pict)
Dsc	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Gor I	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Gor III	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Gor IV	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Predator	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Simpa96	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)
Protein Secondary	secondary structure of protein	wSDL	usage	PyZSI Tav2 (pict)

gbio-pbil.ibcp.fr/ws



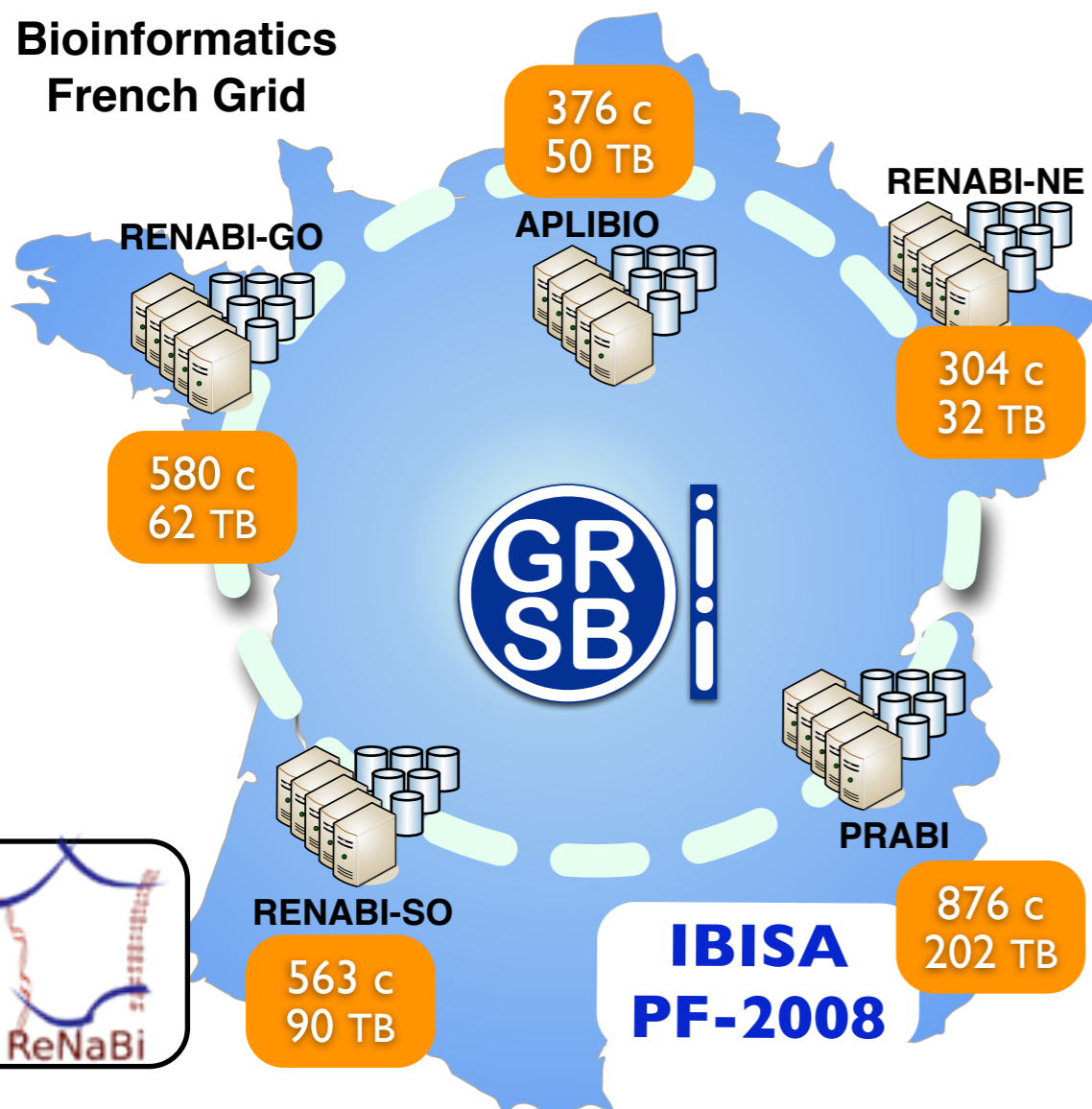
Compliant to EU EMBRACE recommendations



French RENABI GRISBI

Federate Life Science community and provide with IT answers to challenging bioinformatics applications

**- GRISBI -
Bioinformatics
French Grid**



© RENABI GRISBI - www.grisbio.fr

www.grisbio.fr

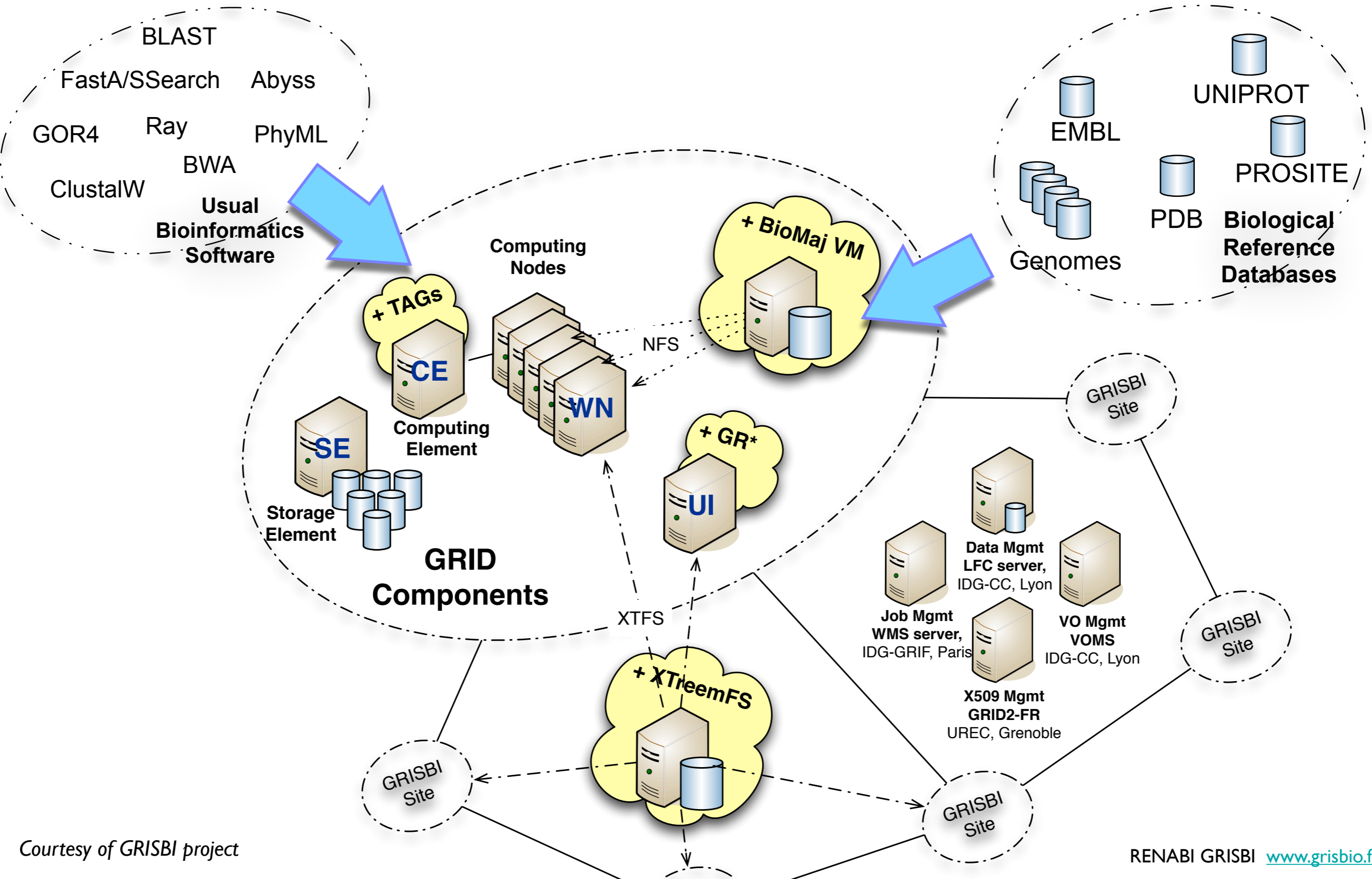
Courtesy of GRISBI project

- Working group on technological aspects:
e.g. gLite, DIET, GridWay, BioMaj, ActiveCircle, Caringo, HDFS, XtremFS, dCache, ...
- Build a distributed infrastructure
 - among the regional centers of RENABI (5)
 - based on the bioinformatics platforms (9)
 - approved as national RIO / IBISA
 - ~70 registered members
 - Computing resources
 - in PFs **2700 cores, 440 TB** storage
 - in grid **860 cores, 30 TB** storage
- Financial support by **RENABI , IBISA 2008-2011, Institut des Grilles 2009-2010**
- In collaboration with the national IT infrastructures: Institut des Grilles, GENCI, Grid5000, regional centers

RENABI GRISBI www.grisbio.fr



Bioinformatics Infrastructure



Goal

- Create comprehensive, open-source, IaaS cloud distribution
- Support a wide range of use cases

Information

- 1 June 2010—31 May 2012 (2 years)
- 6 partners from 5 countries
- Budget : 3.3 M€ (2.3 M€ EC)

Contacts

- Site web: www.stratuslab.eu
- Twitter: @StratusLab
- Support: support@stratuslab.eu



CNRS (FR)



UCM (ES)



GRNET (GR)



SIXSQ (CH)



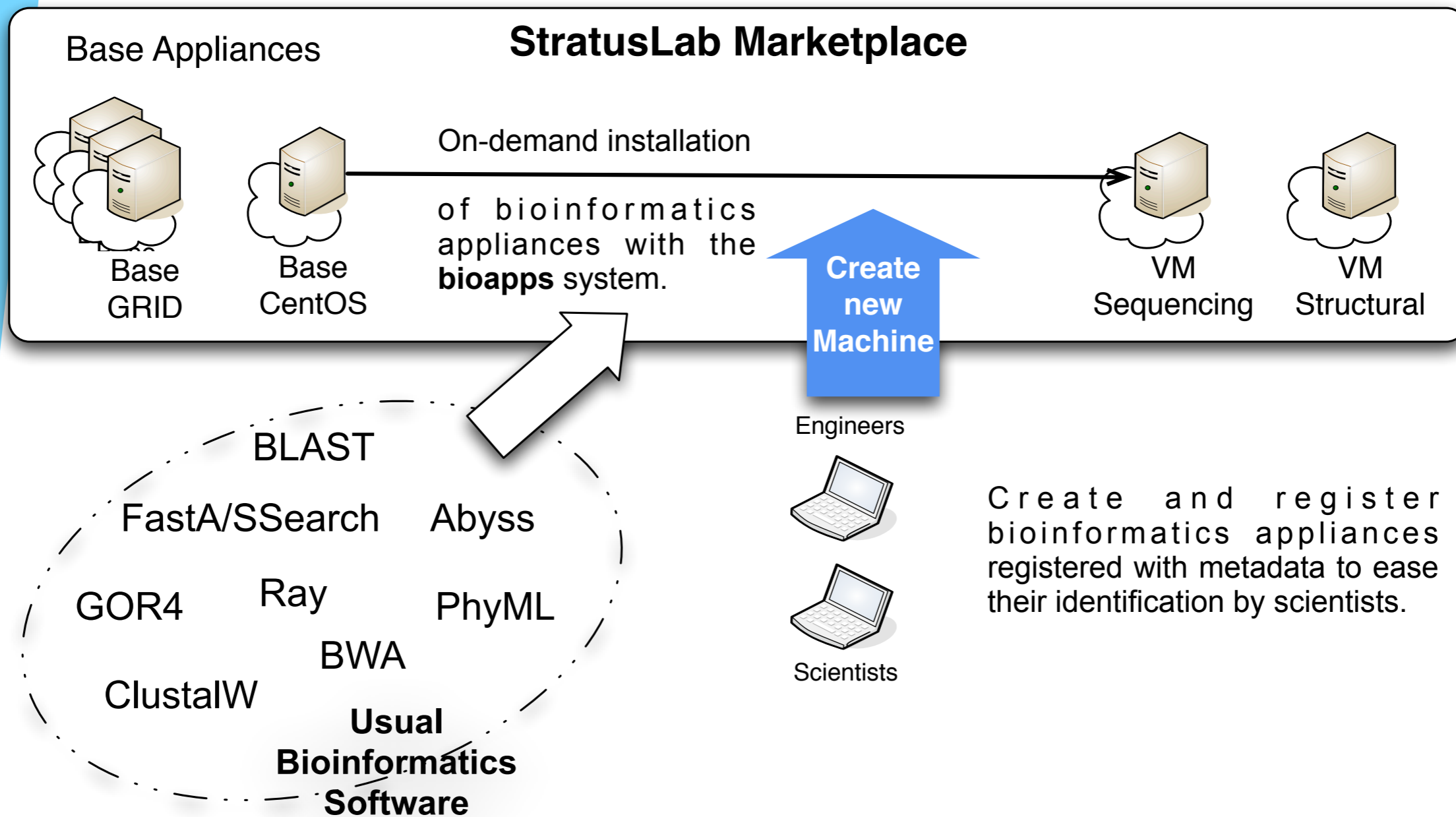
TID (ES)

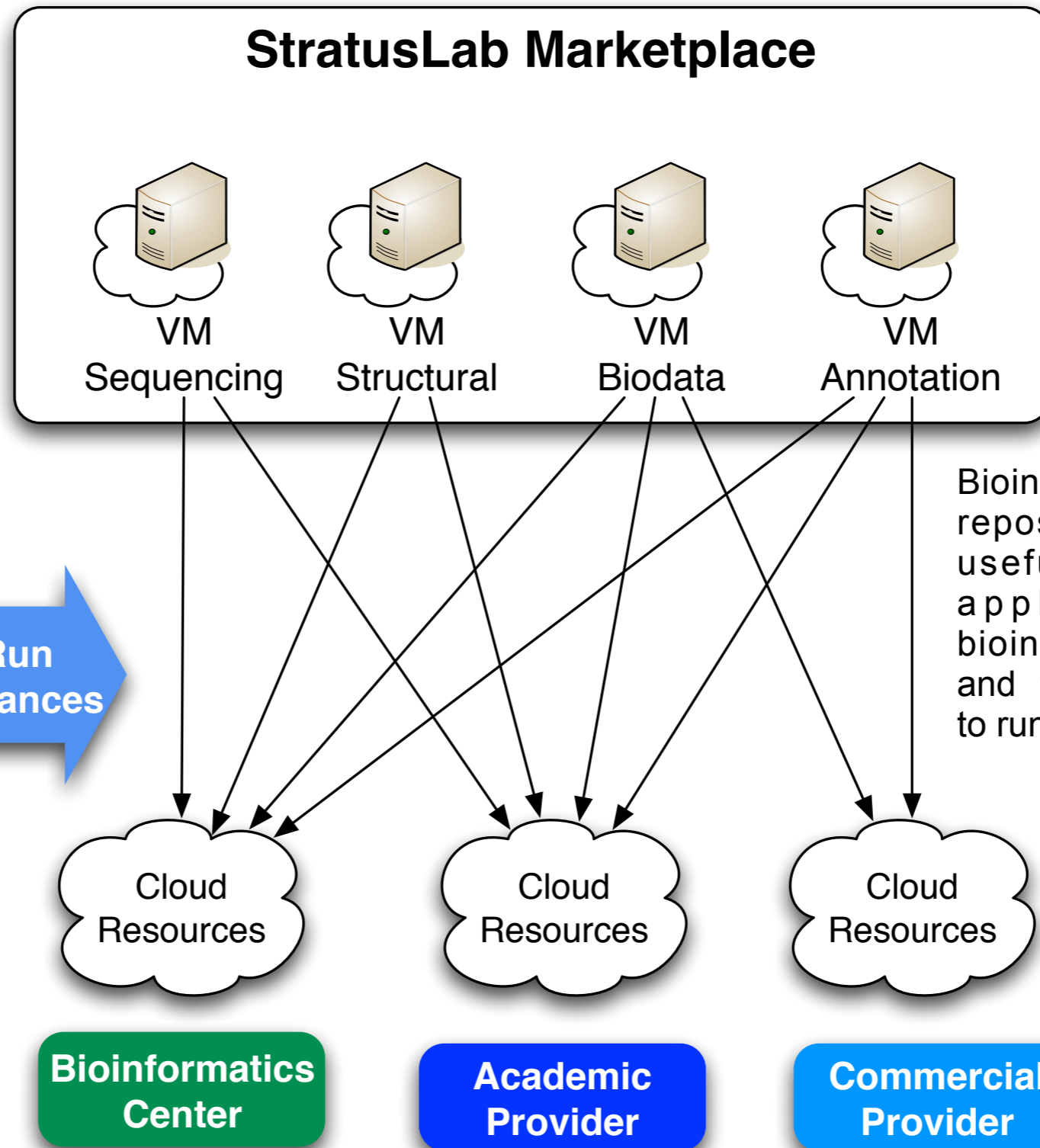


TCD (IE)

- Create **bioinformatics appliances** containing usefull applications that scientists and engineers can **deploy on demand**.
- Provide **POSIX access** (like NFS) to the storage volumes in the cloud repository containing the **biological databases**.
- Biologists and bioinformaticians are regularly **combining multiple software packages** to study their data via analysis pipelines
- Provide them with composable services via **web service** technologies with a **standard** programmatic, public, web service interface (**EMBRACE EU NoE**)

- Provide scientists with bioinformatics appliances to deploy on academic or commercial datacenters, or on their own computer or private cloud.
- Make the cloud infrastructure tightly connected to the storage of the the biological data.
- Ease the procedure of access by using the community's existing authentication methods
 - for example, single sign-on across portals and web services with Shibboleth technology.
- Help bioinformaticians to build and to deploy single machines, clusters, or web service infrastructures to run a complete analysis pipeline.

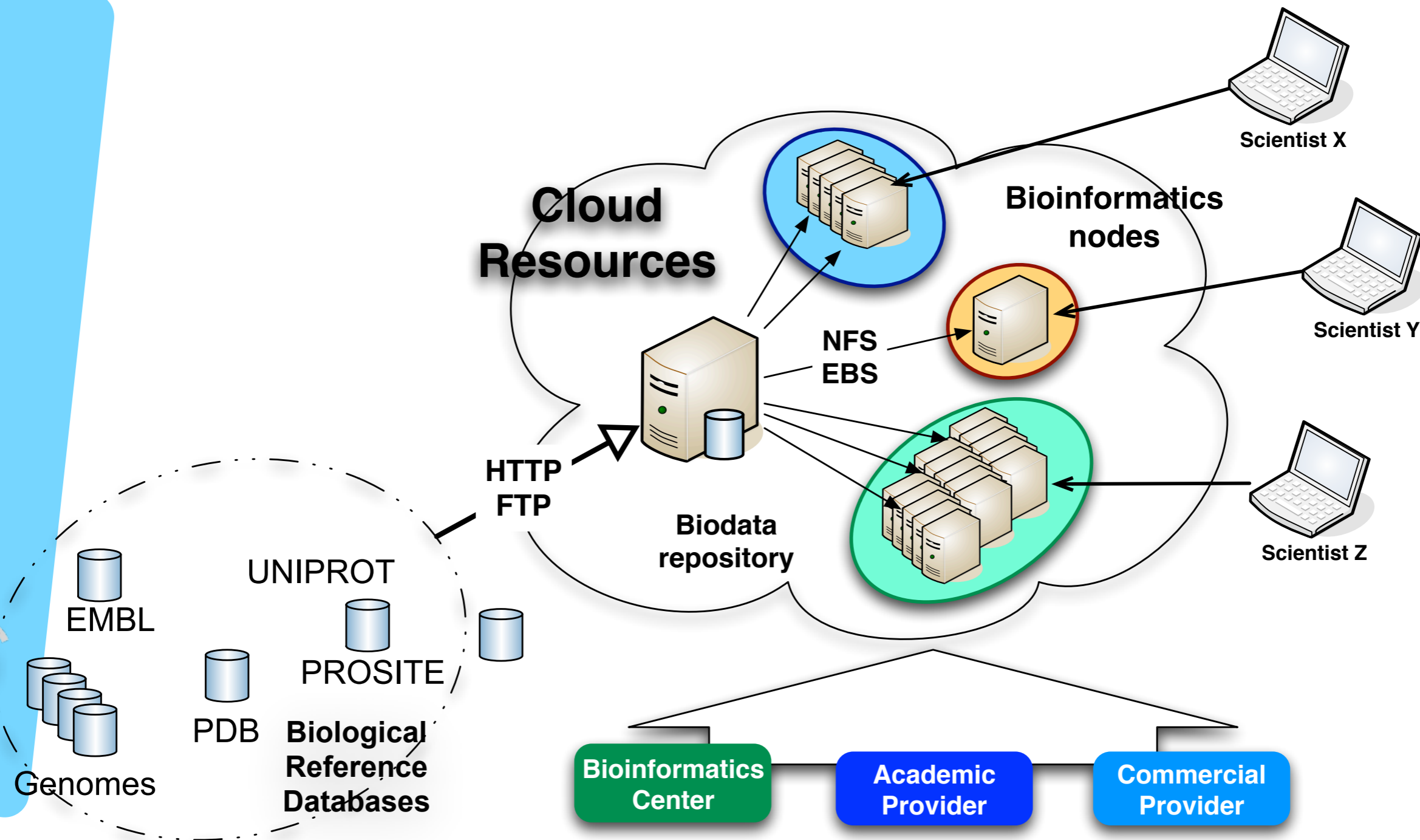




Engineers



Scientists



- Current deployment in CNRS IBCP
 - StratusLab release 1.0
 - 6 cloud nodes (48 cores, 192 GB RAM) + front-end
 - usual constraint of not having enough public IPs
 - port translation on the front to reach the VMs
- Provide bioinformaticians with an experimental platform to evaluate the cloud technology
 - in collaboration with the RENABI GRISBI services
 - will be open to the French community (initially)
 - access based on national Educ-Research federations
 - CNRS GRID2-FR (X509), Renater (shibboleth)
 - provide users with an interactive method to create a new appliance from a base one

What next ?

- Help Bioinformatics centres to provide scientists with services on the cloud
- Platform-as-a-Service
 - Provide relevant Bioinformatics appliances
 - Easily searchable: the marketplace
 - To deploy on academic/commercial datacenters
e.g. RENABI centres, StratusLab clouds (GRNET, LAL)
 - or on their own computer/cloud
- Infrastructure-as-a-Service
 - Complete infrastructures
 - web services and portal with backoffice cluster,
 - multi-nodes applications (e.g. ARIA/ISD),
 - GRISBI grid sites
 - Complete pipelines for bioinformatics analysis

Thanks

- Questions ?

- Acknowledgment:

- C. Gauthey, StratusLab members
- The European Commission through the project
FP7 StratusLab -- 2010-12 -- RI-261552

