Une e-infrastructure nationale en bio-informatique

Christophe BLANCHET

Institut Français de Bioinformatique - IFB
French Institute of Bioinformatics - ELIXIR-FR
CNRS UMS3601 - Gif-sur-Yvette - FRANCE

JDEV 2015
2 juillet 2015, Bordeaux
Since 2004, ReNaBi is the National Network of Bioinformatics platforms with an IBiSA label (Infrastructures in Biology, Health and Agronomy)

In 2010, call of proposals “Infrastructures in Biology and Health” from the “Investments for the Future” initiative.

- Project ReNaBi-IFB accepted in 2012 and endowed with 20m €

Other national infrastructures (NIs)
- France Génomique : sequencing and genotyping NI
- Profi : proteomics NI
- Frisbi : structural biology NI
- etc. (17 NIs all together) + 5 IHUs (Instituts Hospitaliers Universitaires) + 1 IRT (Institut de Recherche Technologique)
IFB, the French distributed infrastructure for life-science information

Mission: to make available core bioinformatics resources to the national/international life science research community.

- To provide support for national biology programs
- To provide an IT infrastructure devoted to management and analysis of biological data
- To act as a middleman between the life science community and the bioinformatics/computer science research community

http://www.france-bioinformatique.fr
CNRS UMS3601. Avenue de la Terrasse, Bât 21. 91190 Gif-sur-Yvette

ELIXIR French Node
- The European distributed infrastructure for life-science information
- To optimize the interactions and coordination between the national level and ELIXIR and other ESFRI infrastructures in biomedical and environmental field,
- To promote consistency and complementarities between the components offered by the ELIXIR French node and those of other European nodes
Experimental data in life sciences (FR)

<table>
<thead>
<tr>
<th>French national platforms (GIS IBISA)</th>
<th>Nb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular imaging</td>
<td>19</td>
</tr>
<tr>
<td>Genomic, Transcriptomic</td>
<td>16</td>
</tr>
<tr>
<td>Proteomic</td>
<td>13</td>
</tr>
<tr>
<td>Structural biology, biophysic</td>
<td>11</td>
</tr>
</tbody>
</table>


Regional centers distribute the load in terms of computing and storage, and provide better interactions with scientists.

Source: omicsmaps.com
A lot of public reference databases

The 2015 *Nucleic Acids Research* Database Issue and Molecular Biology Database Collection

Michael Y. Galperin¹, Daniel J. Rigden² and Xosé M. Fernández-Suárez³

¹National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894, USA, ²Institute of Integrative Biology, University of Liverpool, C UK and ³Thermo Fisher Scientific, Inchinnan Business Park, Paisley, Renfrew PA4 9F

Received November 10, 2014; Accepted November 11, 2014

ABSTRACT

The 2015 *Nucleic Acids Research* Database Issue contains 172 papers that include descriptions of 56 new molecular biology databases, and updates on 115 databases whose descriptions have been previously published in *NAR* or other journals. Following the classification that has been introduced last year in order to simplify navigation of the entire issue, these articles are divided into eight subject categories. This year’s highlights include RNAcentral, an international community portal to various databases on noncoding RNA; ValidatorDB, a

have been previously desc
and 17 contain updates or descriptions have previou
ble 2).

To simplify navigation last year the division of eight sections: (i) nucle transcrip tional regulation, motifs and domains metabolic and signalling protein modification; (iv
fungi; (v) human genor tive genomics; (vi) genon

[Diagrams of nucleotide sequence data, genomes, gene expression data, protein sequence, macromolecular structures, protein families motif and domains]
A lot of bioinformatics tools

ABYSS 1.3.4
ARIA 2.3
Bioconductor 2.11
biomaj
BLAST+ 2.2.27
Blat 35
Bowtie 0.12.8
Bowtie2 2.0.0-beta7
BWA 0.6.2
BWA 0.7.10
CAP3
CD-HIT 4.6.1
Clustal Omega 1.0.3
CLUSTALW 2.1
Cufflinks 2.0.2
Cutadapt 1.2.1
E-SURGE 1.9.0
Exonerate 2.2.0
eXpress 1.5.1
FastA 3.6
FastQC 0.10.1
Galaxy portal
GATK 2.3.4
HMMer 3.0
ImageJ 1.48
khmer 1.1
M-SURGE 1.8.5
MEME 4.7
MMSEQ 0.11.2a
Mobyle
MODAL
MultAlin 5.4.1
MUSCLE 3.8.31
neo4j
Oases 0.2.08
OMSSA 2.1.9
PeptideShaker 0.18.3
phyml 3.1
PREDATOR 2.1.2
proline
python 2.7
R 2.13
R 3.1.1
R 3.1.2
R-studio
Ray 1.3
RSAT
samtools 0.1.18
Samtools 1.1
SearchGUI 1.10.4
SeqClean
Shiny
Stacks
STAR 2.4.0f1
SuMo v1
TGICL
TopHat 2.0.6
trim_galore 0.3.7
Trinity 2.0.4
U-CARE 2.3.2
VCFtools 0.1.11
Velvet 1.2.10
X!tandem
12-10-01-1
XPLOR-NIH 2.30
...
Many interfaces
Use case: deploy a simple bioinformatics application

Cloud
- VMs lifecycle
  - Deploy
  - Cloud web dashboard
  - AuthN/Z +-encrypt.
  - IFB web dashboard
  - Slispsream
  - Docker
  - StrastusLab
  - Cluster
  - Hadoop

SW

HW
- Secure
  - User’s LAN
  - WAN academic + public
  - DC A
  - refdata

User
- log in + data(+)
- refdata
- isolated network

Cloud
- VM
- user's data
- refdata
- Pipeline
  - web interface

Use case

8
Use case: deploy a complex bioinformatics application

- Cloud web dashboard
- VMs/disk lifecycle
- User log in
- AuthN/Z
- Deploy
- Data (scp/sftp)

- Cloud
- Isolated network
- Shared data
- Worker
- Data (viz)

- SW
  - IFB web dashboard
  - Slisstream
  - TRESOR
  - StrastusLab
  - SAMI/XCAML
  - OpenNaaS

- HW
  - User's LAN
  - WAN (academic + public)
  - Isolated network
  - Sync
  - DC A
  - DC B
  - Genomes
  - Refdata(+)

- User
Deploy and operate IFB’s e-infrastructure as a cloud
Mission: to provide core bioinformatics resources to the life science research community.

- To set up a French IT infrastructure devoted to management and analysis of biological data
- To collaborate with international infrastructure (ELIXIR)

Goal: to help scientists and engineers to deploy and use their tools

- e-infrastructure:
  - provide hardware, data collections and bioinformatics tools
- Current resources
  - A national hub: IFB-core hosted at CNRS IDRIS SC center
  - A network of regional centers:
    - 32 bioinformatics platforms - 15,000 cores - 5 PB

Create a federation of clouds for life sciences
Running clouds
- IFB-core
- GenOuest

PoC & experiments
- URGI
- BiLille/Univ.Lille
- BISTRO/IPHC

Next
- Lyon
- ??
# IFB-core’s cloud

<table>
<thead>
<tr>
<th>IFB-core</th>
<th># Compute Cores</th>
<th># TB Storage</th>
<th># TB RAM</th>
<th>Max VM size</th>
<th>Technology</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot</td>
<td>200</td>
<td>50</td>
<td>2</td>
<td>40c 256GB</td>
<td>StratusLab</td>
<td>CNRS-IDRIS, Paris</td>
</tr>
<tr>
<td>2016-S1</td>
<td>3,000</td>
<td>500</td>
<td>-</td>
<td>?144c 3TB?</td>
<td>StratusLab</td>
<td>CNRS-IDRIS, Paris</td>
</tr>
<tr>
<td>2017</td>
<td>10,000</td>
<td>2,000</td>
<td>-</td>
<td>??</td>
<td>StratusLab</td>
<td>CNRS-IDRIS, Paris</td>
</tr>
</tbody>
</table>

**Diagram:**

- **PaaS**
  - NGS, imaging, statistics, ...

- **IaaS**
  - launch jobs
  - Shared FS
  - Workers
  - Master

- **SaaS**
  - Cloud Hypervisors
    - std nodes: 32c 128GB
    - bigMem nodes: 40c 256GB

- **Virtualization Layer**
  - Web portal
  - Pdisk storage
  - iSCSI

**Hosted:** IDRIS CNRS SC-center
Create bioinformatics “appliances”

Tools
- BLAST
- R
- OMSSA
- FastA
- SSearch
- PeptideShaker
- TopHat
- HMMer
- BWA
- samtools
- X!tandem
- Galaxy
- Muscle
- ClustalW2
- Clustal Omega
- Clustal
- Omega
- fastQC

Appliance?
- predefined virtual machine
- Ready to run

Virtual Machines

Bioinformatics Marketplace
- Sequences
- Structures
- Proteomics
- Galaxy
- ...

Create new cloud services

Description
- Title
- Contact (and maintainer !)
- Description (w. controlled voc.)
Docking bioinformatics tools

IFB docker hub (@GenOuest)
http://docker-ui.genouest.org

Registry of images

Bioinformatics images

Docker Image size (MB)

- abyss
- blast
- blast+
- bowtie
- bwa
- clustalw2
- cufflinks
- fasta36
- fastqc
- gor4
- hmm
- meme
- mmseq
- cufflinks

Developer

User
Move VMs rather than data
IFB’s bioinformatics cloud services
A cloud driven through a web dashboard

http://cloud.france-bioinformatique.fr/cloud
Storage for biological data

Upload your data

sftp/http

Get your results

sftp/http

Public Data sources

- UNIPROT
- EMBL
- PDB
- PROSITE

Genomes

User data

PaaS

BLAST, Clustal, etc.

IaaS

Shared FS

Master & Storage
VM ARIA

Workers
VM CNS

Portal

Identity Mgmt

virtual disks

Identity Mgmt

User data

CLI (scp, sftp), GUI (Cyberduck, Transmit, Filezilla, …)

CLI (scp, sftp), GUI (Cyberduck, Transmit, Filezilla, …)

Identity Mgmt

Virtual disks

User data

Upload your data

sftp/http

Get your results

sftp/http

CLI (scp, sftp), GUI (Cyberduck, Transmit, Filezilla, …)
Browse the marketplace and run an App!

IFB's bioinformatics marketplace
RAINBio: Registry of bioinformatics tools and VMs

Prototype

Life science researcher

Query:
- topic ?
- tool ?
- VM ?

Cloud Marketplace
- e.g. IFB

Services Registry
- e.g. ELIXIR

VMs

Tools

Graph DB (Neo4J)

<table>
<thead>
<tr>
<th>topic</th>
<th>VMs</th>
<th>tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence comparison</td>
<td>BIO compute node - 3.1</td>
<td>ClustalW2</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Galaxy MODAL - 1.0</td>
<td>MPAgeneomics</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>BIO compute node - 3.1, Galaxy MODAL - 1.0</td>
<td>ABySS, MPAgeneomics</td>
</tr>
<tr>
<td>Statistics</td>
<td>Galaxy MODAL - 1.0</td>
<td>MPAgeneomics</td>
</tr>
<tr>
<td>Phylogeny</td>
<td>PhyML - 0.2</td>
<td>PhyML</td>
</tr>
<tr>
<td>Sequence search</td>
<td>BIO compute node - 3.1</td>
<td>NCBI BLAST</td>
</tr>
<tr>
<td>SNP</td>
<td>Galaxy MODAL - 1.0</td>
<td>MPAgeneomics</td>
</tr>
<tr>
<td>Data search, query and retrieval</td>
<td>BIO Data - 1.2</td>
<td>BioMAJ</td>
</tr>
<tr>
<td>Sequence assembly</td>
<td>BIO compute node - 3.1, Galaxy MODAL - 1.0</td>
<td>ABySS</td>
</tr>
</tbody>
</table>

RAINBio
Standard bioinformatics node

With pre-installed standard bioinformatics tools
- BLAST, FastA, SSearch, HMM, ...
- ClustalW2, Clustal-Omega, Muscle, ...
- Bowtie(2), BWA, samtools, ...
- MEME, R, etc.

Connected to public reference datasets
- Uniprot, EMBL, genomes, PDB, etc.
- Automatically shared with the VMs

Cluster mode
- turn several instances in a single virtual cluster
- shared file system
- batch scheduling
R software environment for statistical computing and graphics
- include common bioinformatics module
- Biobase, BiocGenerics, BiocInstaller, GenomInfoDb…

RStudio IDE
- integrated development environment (IDE) for R
- features: console, syntax-highlighting editor …

Shiny web framework
- powerful web framework for building web applications using R.
- without requiring HTML, CSS, or JavaScript knowledge.

Contact: Stéphane Delmotte (PRABI-LBBE)
Motivation

- Collaboration with a mass spectroscopy platform
- Running out of space on their local resources

Protein identification tools

- Mass experimental data
- Reference databases: nr, Swiss-Prot
- Reference screening tools: OMSSA, X!
- Tandem

User interface

- Remote Virtual Desktop (NX)
- Reference GUI
  - PeptidShaker
And others ...

RSAT

Galaxy

Ecology of populations

Hadoop
20 bioinformatics appliances already available

- + 10 in progress
- IFB supports different domain-specific developments
  - Microbial Bioinformatics, Evolutionary bioinformatics, Plant bioinformatics, Structural Biology, NGS data processing

Scientific production - 180 users (July 2015)

- opened to members of IFB (standard allocated resources)
- opened to partners, academic and industry, infrastructures and projects: e.g. BioDataCloud, ProFi, MetaboHub, …
- extra resources allocation according to scientific and financial criteria

Training - 70 users

- French tutorials ‘Cloud pour la Biologie’
- tutorial at ECCB’14 about ‘Analysis of Cis-Regulatory Motifs from High-Throughput Sequence Sets’
- Bioinformatics Master in Marseille (2014), in Rouen (2015)
Perspectives

Create more bioinformatics appliances available to the community

• By the experts of the different life sciences domains
• Appliances in progress: BioDataCloud-RNAseq, ProFi, SymBioWatch, Clinical NGS for cancerology (2x), REPET, TriAnnot, Galaxy RAD-seq, Bacterial genomics, iMetAMOS, BioDataCloud-Genomes-Browser…

• Call in 2015-Q1 for proposals to support technological developments and integration of data and tools

Pilots

• Multi-cloud deployment of complex applications
• Secure cloud for human biomedical data analysis
• Live remote cloud processing of sequencing data
• Data registry of multi-cloud datasets
• Docker hub for life sciences
Acknowledgments

• IFB members
  • IFB hub: Patricia, Marie, Jean-François, Mohamed, Quentin ➡ we are hiring!
  • Groupe de réflexion sur les InfraStructures BioInformatiques IFB-GRISBI (co-chair with Olivier Collin)
  • Appliances developers: Samuel Blanck (Inria Lille), Jacques van Helden (TAGC), Stéphane Delmotte (PRABI-Doua), Bruno Spataro (PRABI-Doua)…
• CNRS IDRIS: R. Medeiros, C. Gauthey and staff
• StratusLab members

• IFB is funded by
  • French program PIA INBS 2012
  • EU H2020 EGI-Engage (654142) and CYCLONE (644925) projects

http://www.france-bioinformatique.fr