ELIXIR French Node

J-F. Gibrat

Unité Mixte de Service IFB-core,
CNRS, Gif-sur-Yvette
and
Unité Mathématique, Informatique et Génome,
INRA, Jouy-en-Josas

Transplant-ELIXIR workshop,
Hinxton July 1-2 2014

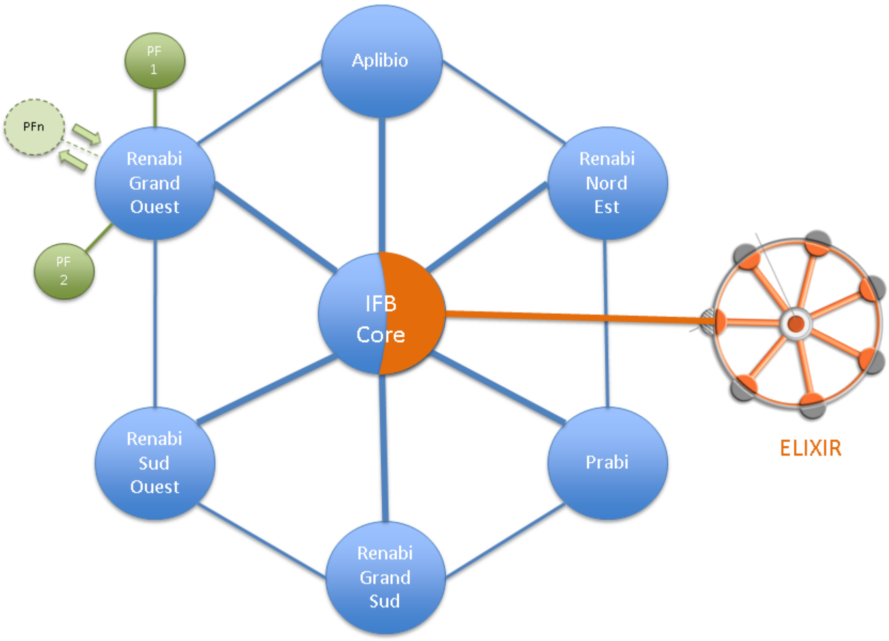
ELIXIR French node

- French Institute of Bioinformatics (IFB) is the ELIXIR French node
- In the context of ELIXIR, IFB missions are :
  - ▷ To coordinate interactions between national level and ELIXIR and other ESFRI (biomedical, environmental fields)
  - ▷ To promote consistency and complementarities between services offered by the French node and other national nodes

IFB structure

IFB consists of :

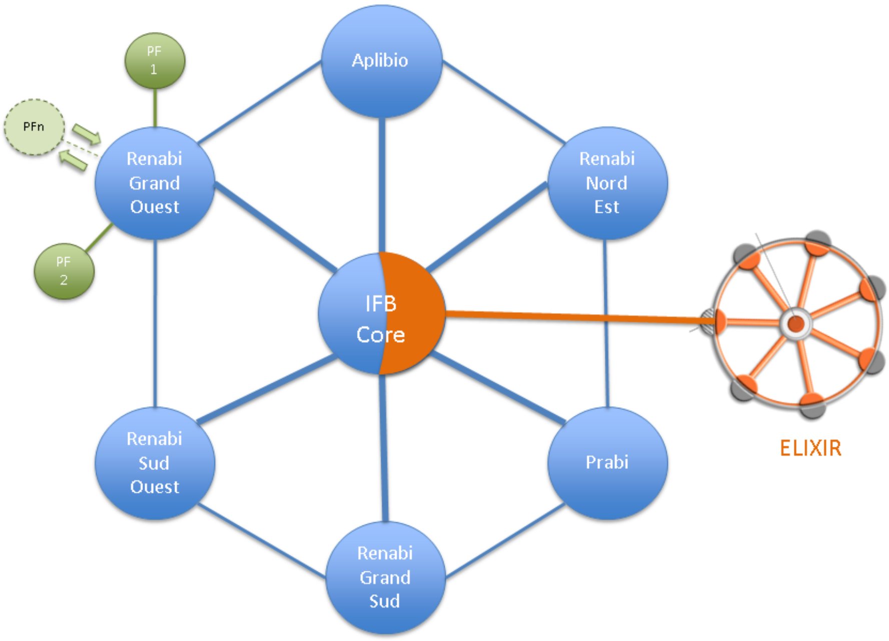
- A network of 6 regional centers (21 PFs)
  ↪ about 110 FTE permanent staff + 70 FTC staff
- A national node : IFB-core
  ↪ about 10 FTE permanent staff + a few FTC staff



IFB structure

IFB consists of :

- A network of 6 regional centers (21 PFs)
  ↪ about 110 FTE permanent staff + 70 FTC staff
- A national node : IFB-core
  ↪ about 10 FTE permanent staff + a few FTC staff



IFB structure

IFB consists of :

- A network of 6 regional centers (21 PFs)
  ↪ about 110 FTE permanent staff + 70 FTC staff
- A national node : IFB-core
  ↪ about 10 FTE permanent staff + a few FTC staff

▷ CNRS : National Centre for Scientific Research

▷ CEA : Alternative Energies and Atomic Energy Commission

▷ INRA : National Institute for Agriculture Research

▷ INRIA : National Institute for Computer Science and Control

▷ INSERM : National Institute for Health and Medical Research

▷ CIRAD : French Agricultural Research Centre for International Development

▷ Universities

▷ Pasteur and Curie Institutes (research foundations)

25% French bioinformatics community involved in provision of service

IFB missions

IFB : national infrastructure of *service* in Bioinformatics

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

- To provide support for biology programs
  - projects’ bolstering
  - users’ training
- To provide an IT infrastructure devoted to management and analysis of biological data
  - material resources : CPUs, disks, etc.
  - availability of biology data collections
  - deployment of bioinformatics tools
- To act as a “middleman” between the life science community and the bioinformatics/computer science research community

IFB missions

IFB : national infrastructure of *service* in Bioinformatics

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

- To provide support for biology programs
  - ▷ projects' bolstering
  - ▷ users' training
- To provide an IT infrastructure devoted to management and analysis of biological data
  - ▷ material resources : CPUs, disks, etc.
  - ▷ availability of biology data collections
  - ▷ deployment of bioinformatics tools
- To act as a "middleman" between the life science community and the bioinformatics/computer science research community

IFB missions

IFB : national infrastructure of *service* in Bioinformatics
Mission : to make available core bioinformatics resources to the national/international life science research community.

- To provide support for biology programs
  - ▷ projects' bolstering
  - ▷ users' training
- To provide an IT infrastructure devoted to management and analysis of biological data
  - ▷ deployment of an academic cloud
  - ▷ portfolio of VMs : CLI, web interface (Galaxy), virtual desktop
    - ↪ encapsulating tools of a given subdomain
    - ↪ encapsulating pipelines/workflows for particular use cases
- To act as a “middleman” between the life science community and the bioinformatics/computer science research community

IFB missions

IFB : national infrastructure of *service* in Bioinformatics

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

- To provide support for biology programs
  - ▷ projects' bolstering
  - ▷ users' training
- To provide an IT infrastructure devoted to management and analysis of biological data
  - ▷ deployment of an academic cloud
  - ▷ portfolio of VMs : CLI, web interface (Galaxy), virtual desktop
    - ↪ encapsulating tools of a given subdomain
    - ↪ encapsulating pipelines/workflows for particular use cases
- To act as a "middleman" between the life science community and the bioinformatics/computer science research community

IFB financial sustainability

- 20 M€ grant from Investments for the Future program (until 2020)
  - 10 M€ expendible endowment
  - 10 M€ loan interests (1.25 M€ /year)
- Four items of expenditure :
  - Equipment (6.8 M€)
  - Hired manpower (6 M€)
  - Operating costs (3.8 M€)
  - French contribution to ELIXIR (3.4 M€)
- Permanent staff and operating costs for the regional PFs will be covered by their respective supervising authorities

Services to be provided

Data : ✓ Compute : ✓ Training : ✓ Tools : ✓ Standards : ✓

- Scientific domains
  - **D-1 Microbial world :** with a special focus on bacterial and viral genomic curation and microbial RNA structures
  - **D-2 Plants :** emphasis on species of agronomic interest (wheat, grapevine, etc.)
  - **D-3 Health :** emphasis on rare diseases, cancer and signaling networks that govern them, immunogenetics
- Transversal fields
  - **F-1 :** Phylogeny and classification resources
  - **F-2 :** Protein sequence and structure resources (structural and functional domains, carbohydrate-active enzyme, immunogenetics)
- Transversal activities
  - **A-1** French academic cloud : large scale storage, computing, networking and security components
  - **A-2** Analysis of metagenomic data :
↪ environment (soil, fresh waters, marine ecosystems)
↪ plants interaction with environment (phyllosphere, rhizosphere)
↪human gut microbiota

Services to be provided

Data : ✓ Compute : ✓ Training : ✓ Tools : ✓ Standards : ✓

- Scientific domains
  - **D-1 Microbial world :** with a special focus on bacterial and viral genomic curation and microbial RNA structures
  - **D-2 Plants :** emphasis on species of agronomic interest (wheat, grapevine, etc.)
  - **D-3 Health :** emphasis on rare diseases, cancer and signaling networks that govern them, immunogenetics
- Transversal fields
  - **F-1 :** Phylogeny and classification resources
  - **F-2 :** Protein sequence and structure resources (structural and functional domains, carbohydrate-active enzyme, immunogenetics)
- Transversal activities
  - **A-1** French academic cloud : large scale storage, computing, networking and security components
  - **A-2** Analysis of metagenomic data :
    ↪ environment (soil, fresh waters, marine ecosystems)
    ↪ plants interaction with environment (phyllosphere, rhizosphere)
    ↪human gut microbiota

French plant bioinformatics landscape

Three main actors :

- INRA "domestic" plants
  - URGI PF (Versailles) – Plants, trees, fungi
  - Genouest PF (Rennes) – Plant pests
  - LIPM PF (Toulouse) – Plants, fungi
  - Plant genetics unit IS (Orsay) – Plants
  - bioinfo@SPIBOC PF (Sophia-Antipolis) – Plant pests
  - GenoToul PF (Toulouse) – Multi-kingdom
  - CBIB PF (Bordeaux) – Trees/metabolomics
  - a number of IS and data production technical platforms (transcritomics, metabolomics, phenotyping)
- CIRAD/IRD "Southern" and Mediterannean plants
  - Southgreen PF (Montpellier) – Plants, trees
- *CEA/Genome Institute, Evry*
  - National centre for sequencing (Genoscope)

Species of interest

- INRA : field crops, fruit trees, vegetables, forestry
  - wheat, maize, rapeseed, leguminous plants, sunflower, solanaceae (tomato, red pepper, potato), brassica, vine, prunus species, apple/pear
  - oak, poplar, douglas pine
  - **model plants :** *A. thaliana*, *M. truncatula*, *Brachypodium distachyon*
  - **phytopathogen fungi :** *Botrytis cinerea*, *Leptosphaeria maculans*, *Microbotryum violaceum*, *Venturia inaequalis*
  - **plant pests :** aphids, lepidopters (*Helicoverpa armigera*, *Spodoptera frugiperda*), nematods
- CIRAD/IRD :
  - banana, citrus, coffee, cocoa
  - sugar cane, African rices, sorghum
- Genoscope (ESTs, genomic sequence)
  - **Model plants :***Arabidopsis thaliana*, *Medicago truncatula*, *Ectocarpus siliculosus* (brown alga), *Oryza sativa*
  - **Plants of agricultural value :** alder, swamp oak, clementine tree, Eucalyptus, walnut, pine tree, poplar, oak, cocoa tree, **wheat**, rice, common bean, **vine**, rosebush
  - **Phytopathogen fungi :** *Botrytis cinerea*, *Hemileia vastatrix*, *Leptosphaeria maculans*, *Melampsora larici-populina*, *Microbotryum violaceum*

EBI and URGI services

- EBI
  - "Generic", comprehensive bio-molecular archives (ENA, ArrayExpress, PRIDE, etc.)
  - Reference datasets on which are based all other analyses (produced by consortia)
  - Ensembl Plants : assembly, annotation, variations, regulations
  - Tools for comparative genomics and genome browser
- URGI
  - Information system for species of agricultural value
  - links with genetic resources and high-throughput phenotype
  - integration of genomic, genetic data and phenotypes
    - gene and repeats, SNPs, SSRs, physical map
    - genetic markers, genetic maps, genetic collections
    - phenotypes

Plant service interoperability

- IFB "plant" bioinformatics nodes : URGI, SouthGreen, Roscoff, Genouest + INRA LIPM
- Maintain information systems that integrate genetics/phenotype and genomics data (>25 species)
- Data are shared with academic users and private companies (21 companies)
- Objective : Strengthen data interoperability and data integration between plant PFs
  - ↬ WP1 : Development of an RDF-based semantic interoperability between plant databases
  - ↬ WP2 : Development of a full-text searching tool to query distributed databases.
  - ↬ WP3 : Development of a Galaxy Plant Server to distribute curated analysis workflows
- IFB provides three 18-months FTC engineers

WP1 : RDF-based semantic interoperability

Objective : develop an RDF-based repository integrating ontologies (CROP), metadata, mappings and allowing web-service accesses to the Plant bioinformatics node databases

- Develop an RDF data model to link data from genetics and genomics (Bioportal)
- Create a repository with RDF data models mapped to relational databases (WebSmatch)
- Develop semantic web services querying the repository to connect distributed data (Biosemantic)

WP2 : Free text search portal

Objective : develop a free text search portal to query simultaneously several database servers located in the different IFB Plant-node platforms.
Tools based on Lucene et Solr used in TransPlant and Wheat IS projects
↬ Addition of semantic layers on top of query searches

- Create a network of index servers based on Solr (distributed text searches)
- Develop a web query portal able to query index servers through web-services
- Develop a tool to build queries using ontologies implemented in the web query portal (semantic layer)

WP3 : Galaxy plant servers

Objectives :

- develop Galaxy Plant Server(s) to distribute and promote curated analysis workflows
- extend Galaxy interoperability functions to allow remote programmatic execution of workflows

▷ workflow for repeat elements detection and annotation

▷ workflow for genetic diversity analyses (linkage disequilibrium, population structure, locus or haplotype frequencies, etc.) and genome wide association mapping (with a focus on data visualization)

▷ workflow for RAD-seq analyses in genome-wide association studies (GWAS)

▷ extend Galaxy API to allow programmatic workflow execution

▷ develop a Galaxy workflow to query WP1 portal

Managing biological information

- Data sustainability
  - ▷ Raw data vs biological information
  - ▷ What should we store and for how long ?
  - ▷ What are the data life cycles ?
  - ▷ Distinguish between legacy data and others
  - ▷ Cost of data sustainability (data management by the biologists)
- Is there a need for a distributed data model ?
  - ▷ distributed Ensembl Plant ?
- Collaboration beyond Europe (e.g., Wheat Initiative)