Les besoins de la communauté des sciences de la vie en termes d'infrastructure informatique

J-F. Gibrat

Unité Mathématique, Informatique et Génome,
INRA, Jouy-en-Josas

Séminaire IDRIS,
Orsay, 14 novembre 2013

Données produites par les sciences de la vie

- Changement d'échelle en biologie depuis une dizaine d'années
- Révolution des techniques : développement de technologies (très) haut-débit
- Permet de collecter des données à l'échelle de la cellule entière
- Démarche encyclopédique : tous les gènes, tous les transcrits, toutes les protéines, toutes les interactions, tous les métabolites et les flux, etc.
- Changement de perspective en génétique : on part du génome pour aller vers les propriétés biologiques de l'organisme
- Les biologistes sont noyés sous avalanche de données hétérogènes.
- 25% du temps à générer les données 75% du temps à les analyser

Génomes et séquençage

- Génome : séquence d'ADN composée de 4 nucléotides A, C, G, T
- Plus petit génome (non viral) connu : *Carsonella ruddii* 0.16 Mbp
- Plus grand génome connu : *Amoeba dubia* 670 Gbp
- Taille max. des *lectures* :
  - 1ère génération : Sanger 900 bp
  - 2ème génération : 35 bp $\longrightarrow$ 2 * 100 bp
  - 3ème génération : Pac Bio min. 500, médiane 3100, max. 27000 bp
- Nécessité de découper le génome en millions de fragments (shotgun sequencing)

Principe du séquençage

ATTAGTGGTCATCCATGGCTATCGCCCGATGAGTGAGG

TCCATGGCTATCG

TCATCCATGGCTA

GGTCATCCATG

GCTATCGCCCGATGAG

CCGATGAGTGAGG

ATTAGTGGTCATCCA

GGCTATCGCCCGA

TAGTGGTCATCCA

ATTAG?????ATCCA

TAGTG???ATCCA

GGTC??CCATG

TCCATGGCTATCG

GCTAT??????ATGAG

TCATC???GGCTA

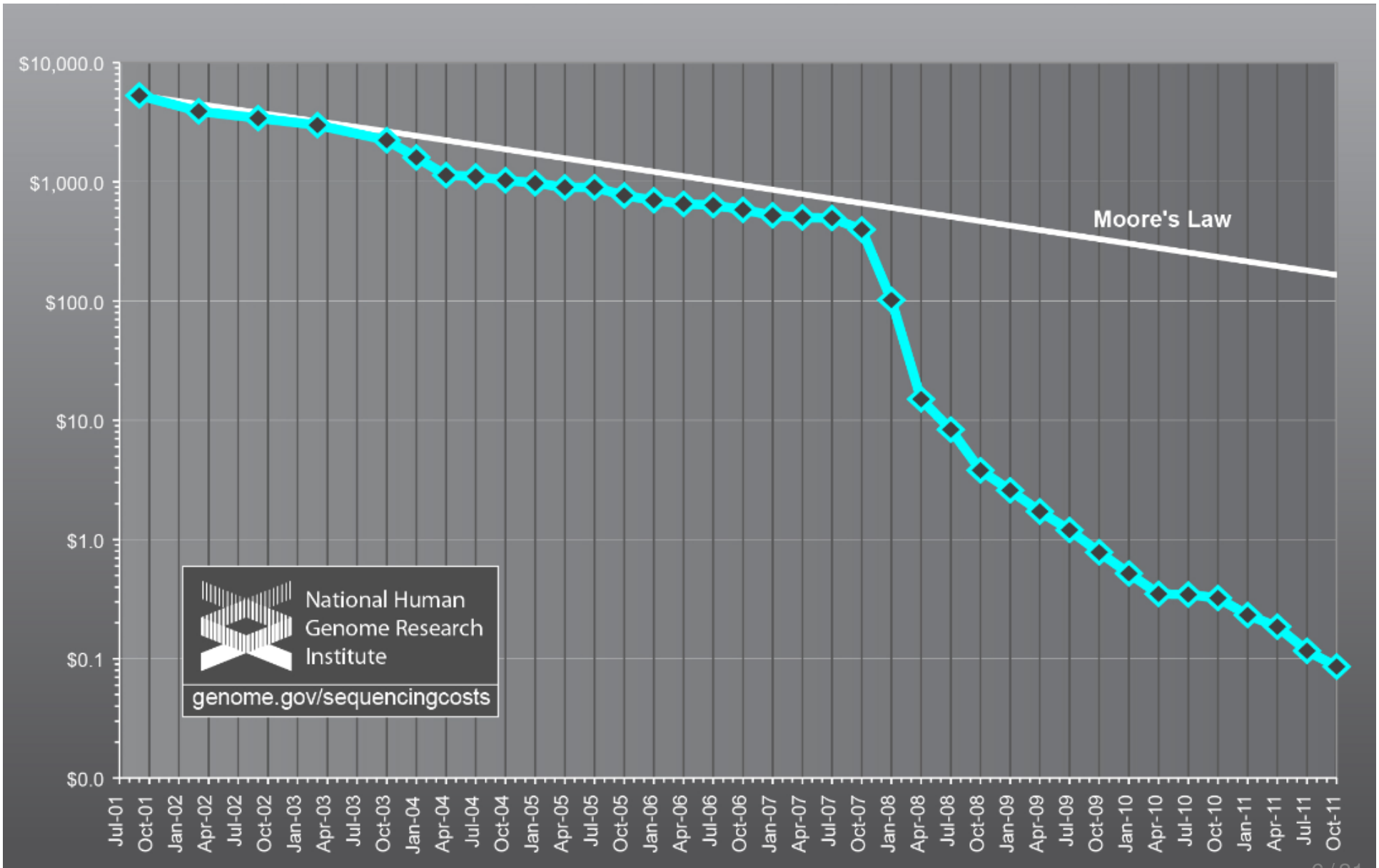
GGCTA???CCCGA

CCGAT???TGAGG

Les nouvelles technologies de séquençage (NGS)

- Depuis 2007 développement de nouvelles technologies de séquençage très haut débit (NGS)
- Un « run » produit 3 milliards lectures appariées 2 x 100 bp
  600 Gbp $\sim$ 4.8 To « brutes » $\sim$ 10 To avec méta-données
- Coûts de séquençage sont passés de 10 000 $ à 0.03 $ par million nucléotides séquencés
- Séquençage et loi de Moore

Loi de Moore


$10,000.0
$1,000.0
$100.0
$10.0
$1.0
$0.1
$0.0
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11
Apr-11
Jul-11
Oct-11

Le séquençage : pour quoi faire ?

- On s'intéresse à un génome non encore séquencé
  - ▷ Assemblage des lectures et annotation du génome
    - séquençage *de novo*
    - re-séquençage
    - métagénomique
    - réarrangements chromosomiques
- On dispose d'un génome de référence
  - ▷ Alignement (mapping) des lectures sur le génome
    - Détection de variants génomiques (SNPs)
    - RNA-seq (expression des gènes)
    - ChIP-seq (régulation de l'expression des gènes)
    - Réarrangements chromosomiques, variation du nombre de copies des gènes
    - Détection de petits ARN non-codants
    - séquençage de l'exome
    - métagénomique

Le séquençage : pour quoi faire ?

- On s'intéresse à un génome non encore séquencé
  - ▷ Assemblage des lectures et annotation du génome
    Algorithmes basés sur chevauchements de lectures
    - reconstruction de la plus courte super-chaîne commune
    - chemin hamiltonien dans un graphe
    - chemin eulérien dans un graphe de *de Bruijn*

    Besoin de beaucoup de RAM (1To)

- On dispose d'un génome de référence
  - ▷ Alignement (mapping) des lectures sur le génome
    - algorithme exact de programmation dynamique
    - heuristiques basées sur du hachage
    - heuristiques basées sur arbres et tableaux de suffixes et transformée de Burrows-Wheeler

La bioinformatique

- Bioinformatique : ensemble des méthodes informatiques pour gérer, organiser et analyser les données biologiques.
  - Communauté issue de l'analyse des séquences (ADN, ARN, protéines), de leurs structures et interactions.
  - Analyse d'images est une communauté différente.
- Rôle bioinformatique :
  - Gérer les données produites en masse par la biologie
  - Organiser ces données
  - Extraire des connaissances biologiques à partir des données brutes

La bioinformatique

- Bioinformatique : ensemble des méthodes informatiques pour gérer, organiser et analyser les données biologiques.
  - Communauté issue de l'analyse des séquences (ADN, ARN, protéines), de leurs structures et interactions.
  - Analyse d'images est une communauté différente.
- Rôle bioinformatique :
  - Gérer les données produites en masse par la biologie
  - Organiser ces données
  - Extraire des connaissances biologiques à partir des données brutes

Caractéristiques des analyses bioinformatiques ... 1

- Beaucoup d'analyses sont distribuables (parallélisables par les données)
- Les analyses nécessitent des enchaînements de traitements différents (pipelines, workflows) et l'intégration de données hétérogènes
- Différents langages utilisés (perl, python, java, ...)
- Les logiciels ont souvent beaucoup de dépendances (versions)
- Foisonnement de logiciels pour ces traitements (98 logiciels d'alignement de lectures sur un génome)
- Typiquement, une PF bioinformatique met à disposition plusieurs centaines de logiciels différents

Caractéristiques des analyses bioinformatiques ... 2

- Nécessité de développer des interfaces conviviales pour les utilisateurs (Apache Tomcat)
- Utilisation de nombreuses collections de données qui sont mises à jour régulièrement
- Utilisation de bases de données relationnelles (SGBD : mySQL, postgreSQL, ...) et noSQL (BDD de graphes)
- Nécessité de suivre l'évolution très rapide des technologies de production des données

Infrastructures PF, mésocentres et centres du GENCI

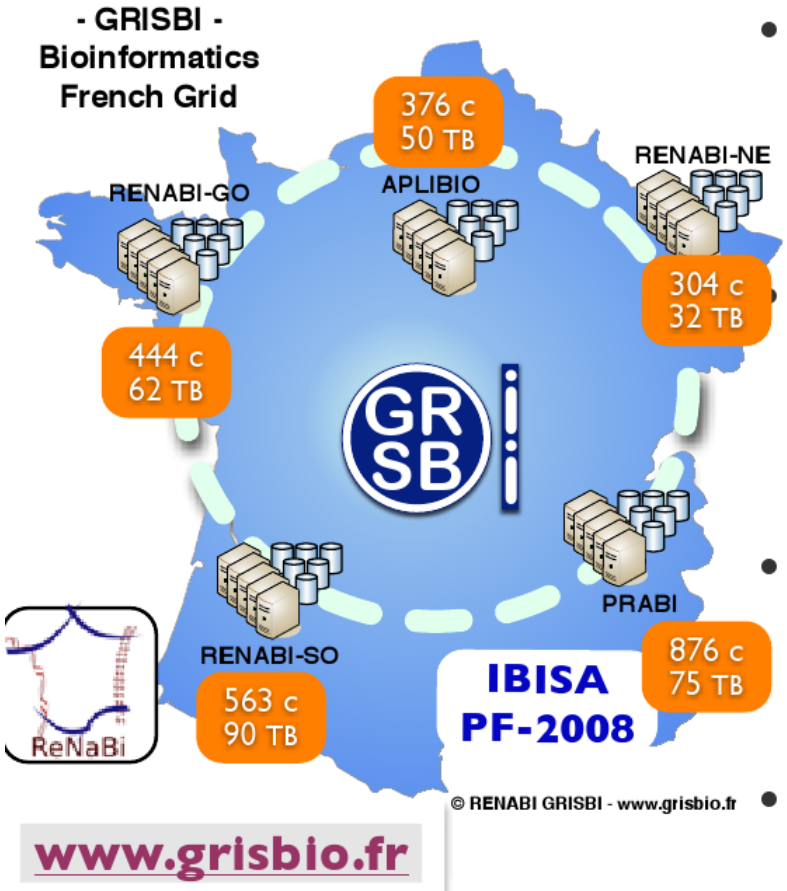
- Les PF ont eu tendance à développer leur propre infrastructure informatique
- Faible utilisation des centres de calcul HPC du GENCI (excepté simulations de dynamique moléculaire)
- Utilisation modérée des mésocentres régionaux
- Passage à l'échelle difficile pour les infrastructures des PF
- Le projet GRISBI : test d'utilisation de la grille

Infrastructures PF, mésocentres et centres du GENCI

- Les PF ont eu tendance à développer leur propre infrastructure informatique
- Faible utilisation des centres de calcul HPC du GENCI (excepté simulations de dynamique moléculaire)
- Utilisation modérée des mésocentres régionaux
- Passage à l'échelle difficile pour les infrastructures des PF
- Le projet GRISBI : test d'utilisation de la grille

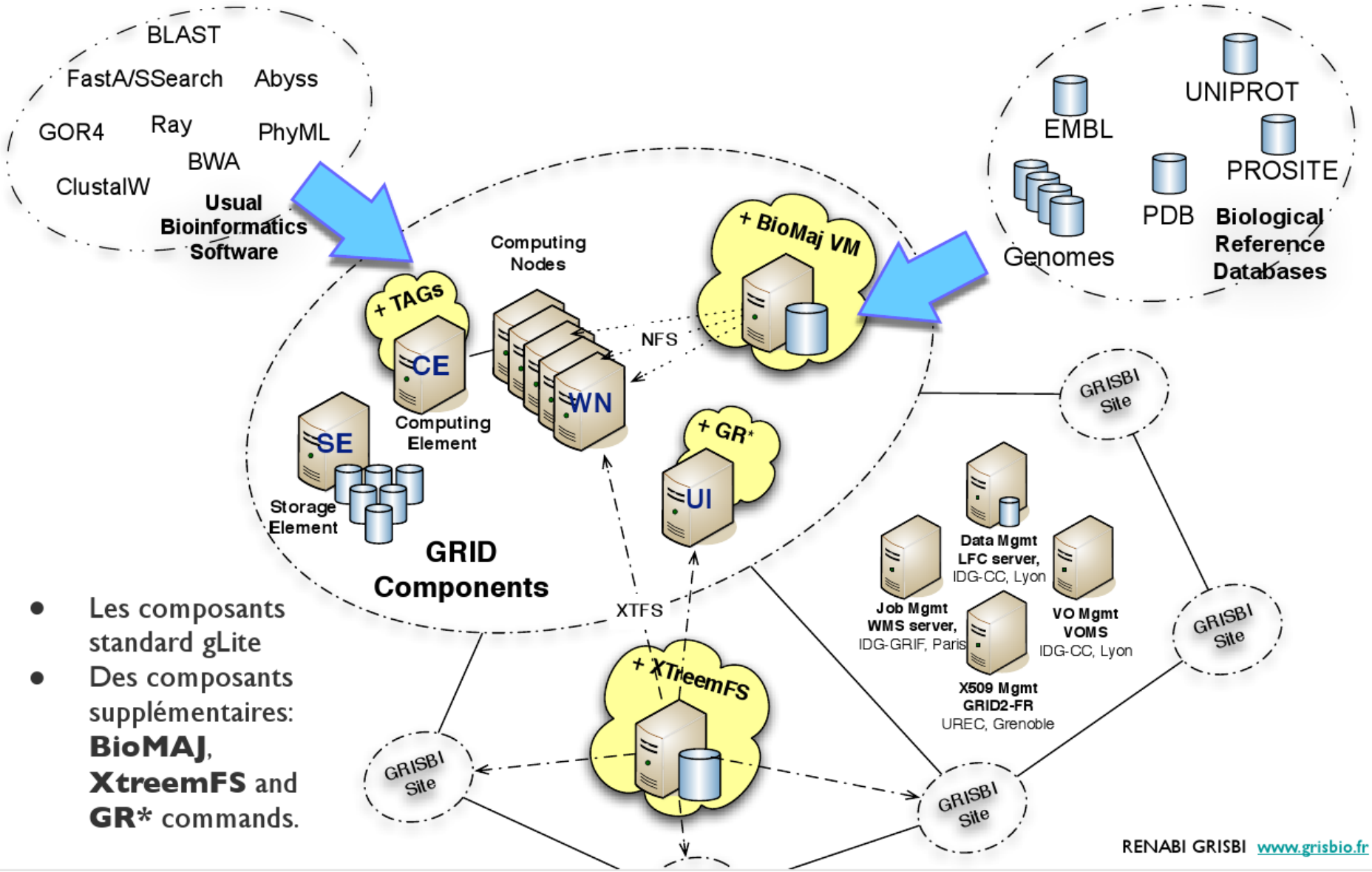
Projet GRISBI

GRISBI : Groupe de réflexion sur les infrastructures en bioinformatique (coordinateur : C. Blanchet)



- Cinq centres régionaux ReNaBi (7 PF)
- Définir l'organisation et les technologies pour fédérer les PF bioinformatiques nationales e.g. gLite, DIET, BioMaj, ActiveCircle, Caringo, HDFS, XtreemFS, ...
- Collaboration avec les infrastructures informatiques nationales : Institut des Grilles, GRID5000, GENCI, etc.

Infrastructure GRISBI



- Les composants standard gLite
- Des composants supplémentaires: **BioMAJ**, **XtreemFS** and **GR*** commands.

Besoins de la communauté des sciences de la vie

	gLite	GRISBI
Banques internationales	~ oui	biomaj NFS
Espace personnel	~ oui	XtreemFS ?
Espace commun	~ oui	
Accès simple au stockage	non	XtreemFS ?
Distribution des calculs	WMS	
Intégration cluster l'existant	~ oui	CE-gateway
Déploiement des logiciels	SWAREA	++ temps humain
Workflow/pipeline	~ DAG	
Gestion des identités et accès	vo.renabi.fr	Shibboleth/LDAP
Interface facile à utiliser	~ CLI	« commandes GR »
Interface publique: accès anonyme sur portail et web services	non	? certificats robot, myproxy ?

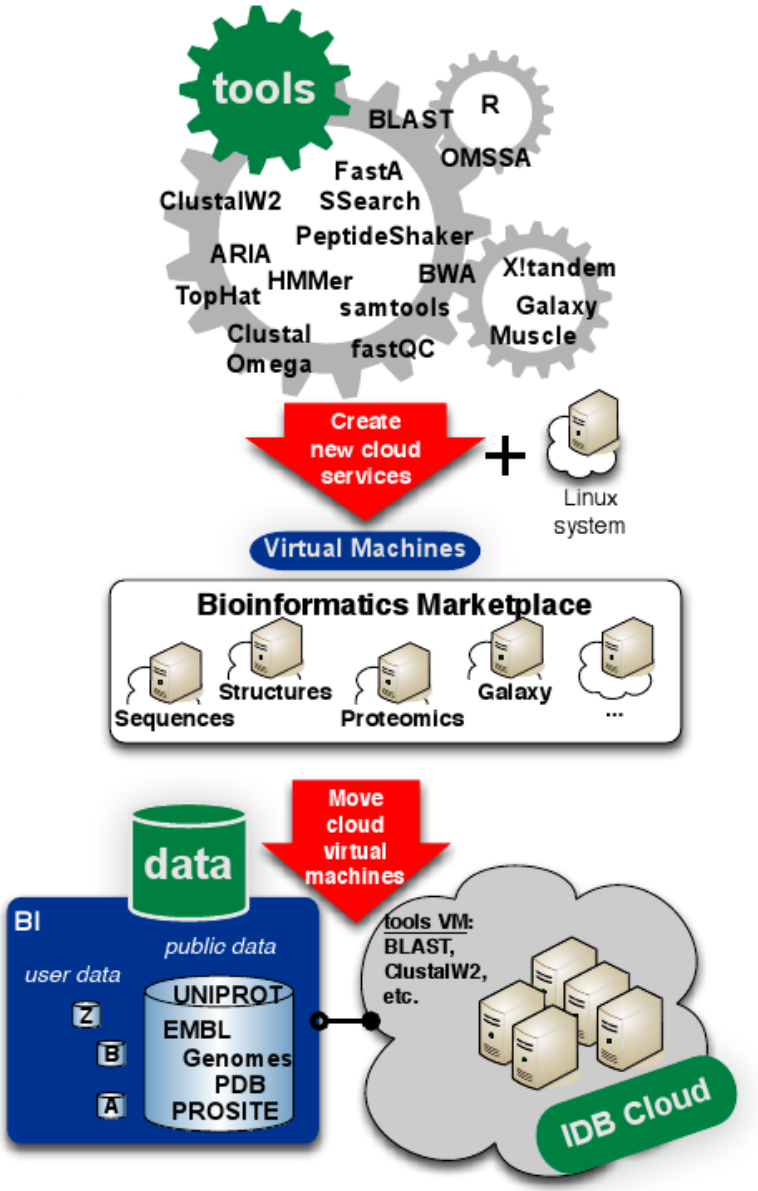
- Intergiciel gLite répond au besoin en puissance de calcul
- Modes d'accès et de gestion des données sont moins adaptés aux usages de la communauté

Infrastructures PF, mésocentres et centres du GENCI

- Les PF ont eu tendance à développer leur propre infrastructure informatique
- Faible utilisation des centres de calcul HPC du GENCI (excepté simulations de dynamique moléculaire)
- Utilisation modérée des mésocentres régionaux
- Passage à l'échelle difficile pour les infrastructures des PF
- Le projet GRISBI : test d'utilisation de la grille
- Développements de Clouds académiques (IBCP Lyon, C. Blanchet ; GenOuest Rennes, O. Collin)

Cloud IDB

Infrastructure distribuée pour la biologie, IBCP Lyon.



- Banc d'essai d'un Cloud pour la biologie
  - ouvert à la communauté des sciences de la vie sept. 2011
  - 14 MV dédiées (*appliances*)
  - quarantaine d'utilisateurs
  - MV jusqu'à 32 c et 768 Go RAM
- Infrastructure
  - Calcul : 900 cœurs et 4 To RAM
    - nœuds standard : 32c-128Go
    - nœuds grande mémoire : 64c-768Go
  - Stockage 250 To : disques virtuels et stockage objet (S3)
  - Basé sur les outils StratusLab

Contexte du projet IFB

- 2010, appel à propositions « Infrastructures en Biologie et Santé » du programme « Investissements d'Avenir »
- ReNaBi (2004) : Réseau national des plates-formes de bioinformatique (label IBiSA[1])
- Projet ReNaBi-IFB accepté en 2012 et doté de 20 M€ (jusqu'en 2020)
- Autres infrastructures nationales du PIA
  - France Génomique : séquençage et génotypage
  - Profi : protéomique
  - Frisbi : biologie structurale
  - 17 autres infrastructures + 5 IHU (Instituts Hospitaliers Universitaires) + 1 IRT (Institut de Recherche Technologique)

[1] Infrastructures en Biologie, Santé et Agronomie

Missions de l'IFB

IFB : infrastructure nationale de *service* en bioinformatique

Mission générale : fournir des ressources de base en bioinformatique à la communauté des sciences de la vie.

- Fournir un appui aux programmes de la communauté des sciences du vivant
  - soutien aux projets de recherche
  - formation des utilisateurs
- Mettre à disposition une infrastructure informatique dédiée à la gestion et l'analyse des données biologiques
  - ressources matérielles : calcul, stockage, RAM, etc.
  - accès aux collections de données biologiques
  - déploiement des outils bioinformatiques
- Agir comme un « intermédiaire » entre la communauté des sciences du vivant et celle de la recherche en (bio)informatique

Missions de l'IFB

IFB : infrastructure nationale de *service* en bioinformatique

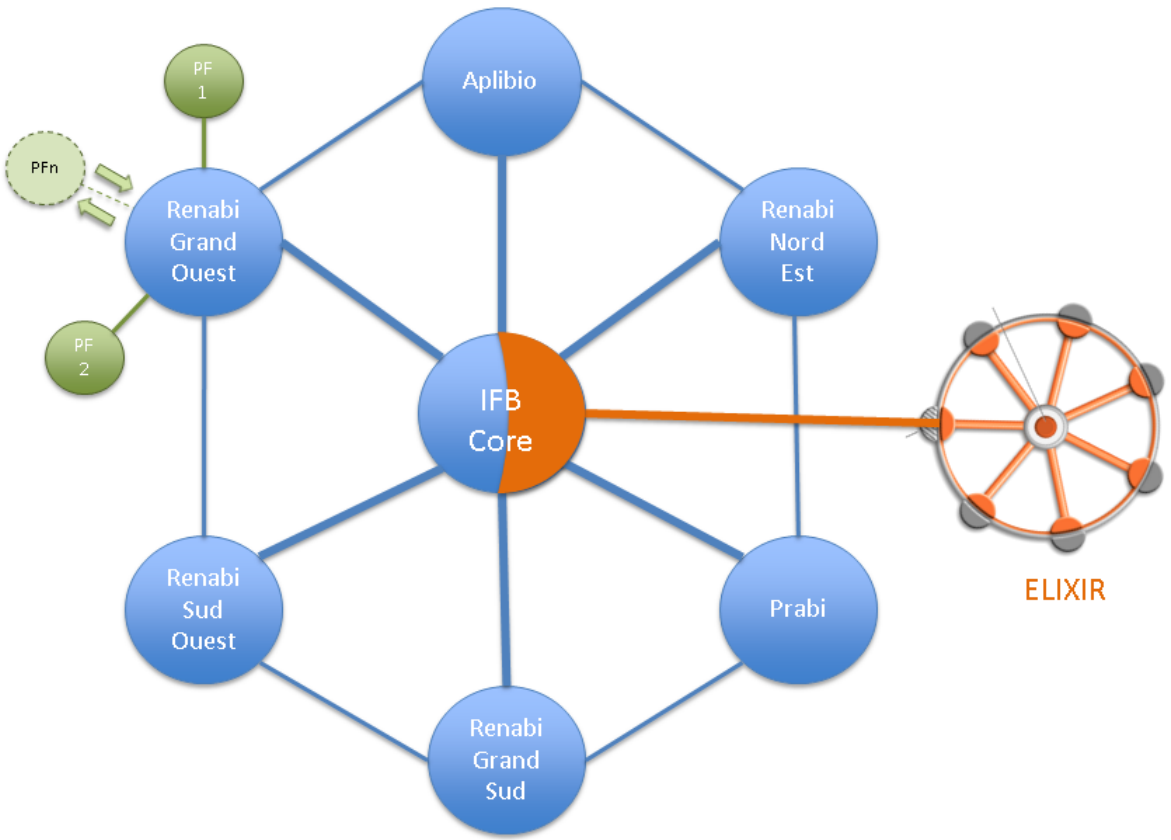
Mission générale : fournir des ressources de base en bioinformatique à la communauté des sciences de la vie.

- Fournir un appui aux programmes de la communauté des sciences du vivant
  - soutien aux projets de recherche
  - formation des utilisateurs
- Mettre à disposition une infrastructure informatique dédiée à la gestion et l'analyse des données biologiques
  - ressources matérielles : calcul, stockage, RAM, etc.
  - accès aux collections de données biologiques
  - déploiement des outils bioinformatiques
- Agir comme un « intermédiaire » entre la communauté des sciences du vivant et celle de la recherche en (bio)informatique

Structure de l'IFB

L'IFB est constitué :

- un noeud national : IFB-core
  ↪ environ 10 permanents + 6 CDD
- un réseau de 6 centres régionaux (21 PF)
  ↪ environ 100 ETP permanents + 50 CDD



Structure de l'IFB

L'IFB est constitué :

- un noeud national : IFB-core
  ↪ environ 10 permanents + 6 CDD
- un réseau de 6 centres régionaux (21 PF)
  ↪ environ 100 ETP permanents + 50 CDD

IFB-core = Unité Mixte de Service (UMS3601)

IFB-core est chargé des questions administratives et techniques pour l'IFB (entre autres, la gestion des fonds de l'ANR).

Budget prévisionnel de l'IFB

- 10 M€ consommables
- 10 M€ intérêts d'emprunt (1.25 M€/année)

Year	2012	2013	2014	2015	2016	2017	2018	2019
Loan interests	0.86	1.25	1.25	1.25	1 .25	1.25	1.25	1.64
Expendible endowment		2.70	1.90	1.40	1.80	1.20		1.00
IFB-core hired staff		0.20	0.20	0.20	0.14	0.20		
Project hired staff		1.00	1.00	1.00	1.00	1.00		
National infrastructure eq.		1.50	0.50			0.40	0.40	1.00
Regional PF infrastructure eq.		1.00	0.40			0.30	0.30	1.00
IFB operating costs	*0.10*	0.10	0.10	0.10	0.10	0.15	0.15	0.24
Hosting of the IFB-core IT infrastructure		0.40	0.40	0.40	0.40	0.40	0.40	0.40
ELIXIR Hub	*0.17*	0.31	0.56	0.95	1.43			

figures are in M€

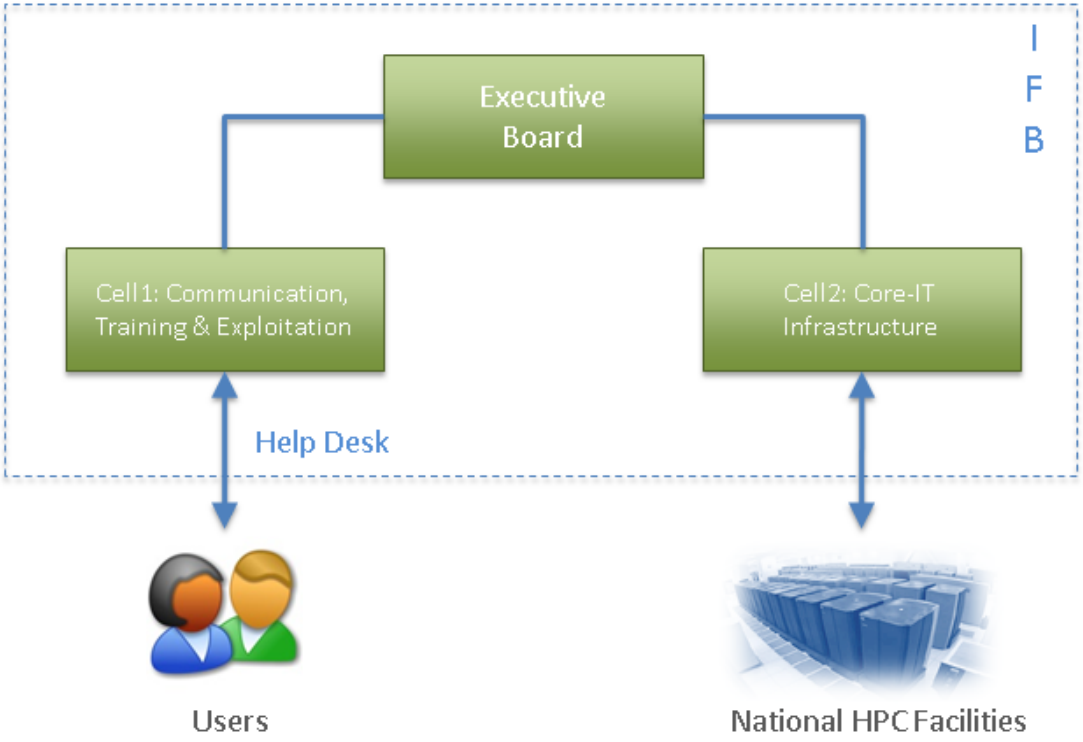
Quatre postes de dépenses :

- Équipement (6.8 M€)
- CDD (6 M€)
- Coûts de fonctionnement (3.8 M€)
- Fonds dédiés aux coûts de fonctionnement du Hub ELIXIR (3.4 M€)

IFB-core et core-IT

IFB-core (UMS3601) aura sa propre infrastructure informatique (l'infrastructure *nationale* de l'IFB nommée core-IT) et son propre personnel. Elle fournira une appui technique et administratif à l'IFB. Elle est constituée de deux cellules :

- La cellule *Communication, Formation et Valorisation*
- La cellule *Infrastructure* en charge de déployer un Cloud académique sur l'infrastructure nationale de l'IFB.



Infrastructure nationale et Cloud académique

- Dévelopment d'un Cloud académique dédié à la gestion et l'analyse des données produites par les sciences du vivant :
  - Infrastructure nationale de l'IFB
  - Infrastructures régionales des PF
- Les PF régionales fournissent un accès gratuit à leurs infrastructures informatiques :
- Infrastructure nationale de l'IFB (core-IT)
  - ▷ hébergée à l'IDRIS
  - ▷ 10 000 cœurs et 1 Po stockage $\sim$ machine ADA à l'IDRIS
- Infrastructure PF régionales
  - ▷ essayer de convaincre les PF de s'appuyer sur des mésocentres régionaux

Infrastructure nationale et Cloud académique

- Dévelopment d'un Cloud académique dédié à la gestion et l'analyse des données produites par les sciences du vivant :
  - Infrastructure nationale de l'IFB
  - Infrastructures régionales des PF
- Les PF régionales fournissent un accès gratuit à leurs infrastructures informatiques :
  - ▷ ∼6000 cœurs et ∼1Po stockage
- Infrastructure nationale de l'IFB (core-IT)
  - ▷ hébergée à l'IDRIS
  - ▷ 10 000 cœurs et 1 Po stockage ∼ machine ADA à l'IDRIS
- Infrastructure PF régionales
  - ▷ essayer de convaincre les PF de s'appuyer sur des mésocentres régionaux

Infrastructure nationale et Cloud académique

- Dévelopment d'un Cloud académique dédié à la gestion et l'analyse des données produites par les sciences du vivant :
  - Infrastructure nationale de l'IFB
  - Infrastructures régionales des PF
- Les PF régionales fournissent un accès gratuit à leurs infrastructures informatiques :
  - ▷ ∼6000 cœurs et ∼1Po stockage ... mais >20 localisations
- Infrastructure nationale de l'IFB (core-IT)
  - ▷ hébergée à l'IDRIS
  - ▷ 10 000 cœurs et 1 Po stockage ∼ machine ADA à l'IDRIS
- Infrastructure PF régionales
  - ▷ essayer de convaincre les PF de s'appuyer sur des mésocentres régionaux

Infrastructure nationale et Cloud académique

- Dévelopment d'un Cloud académique dédié à la gestion et l'analyse des données produites par les sciences du vivant :
  - Infrastructure nationale de l'IFB
  - Infrastructures régionales des PF
- Les PF régionales fournissent un accès gratuit à leurs infrastructures informatiques :
  - ▷ $\sim$6000 cœurs et $\sim$1Po stockage ... mais >20 localisations
- Infrastructure nationale de l'IFB (core-IT)
  - ▷ hébergée à l'IDRIS
  - ▷ 10 000 cœurs et 1 Po stockage $\sim$ machine ADA à l'IDRIS
- Infrastructure PF régionales
  - ▷ essayer de convaincre les PF de s'appuyer sur des mésocentres régionaux

Merci de votre attention