# IFB-Biosphère, Services Cloud pour l'Analyse des Données des Sciences de la Vie

Christophe Blanchet, Olivier Collin, Matéo Boudet, Stéphane Delmotte, Hervé Gilquin, Jean-François Guillaume, Efflam Lemaillet, Jonathan Lorenzo, Olivier Sallou, Bruno Spataro, et al.

# IFB-Biosphère : Services cloud pour l'analyse des données des sciences de la vie


### Christophe Blanchet
CNRS, UMS 3601, Institut Français de Bioinformatique, IFB-core
2 rue Gaston Crémieux
F-91000 Evry, France

### Olivier Collin
Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA
263 avenue Général Lerclerc
F-35000 Rennes, France

### Matéo Boudet
Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA
263 avenue Général Lerclerc
F-35000 Rennes, France

### Stéphane Delmotte
CNRS, UMR 5558, LBBE - Biométrie et Biologie évolutive, UCB Lyon 1
43 bd du 11 novembre 1918
F-69100 Villeurbanne, France

### Hervé Gilquin
CNRS, UMR 5669, PSMN (Pôle Scientifique de Modélisation Numérique) ENS de Lyon
46 Allée d'Italie
F-69007 Lyon, France

### Jean-François Guillaume
BiRD, UMR_S 1087/UMR_C 6291, Unité de Recherche de l'Institut du Thorax, LS2N UMR 6004, Université de Nantes
8 quai Moncousu
F-44007 Nantes, France

### Efflam Lemaillet
Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA
263 avenue Général Lerclerc
F-35000 Rennes, France

### Jonathan Lorenzo
CNRS, UMS 3601, Institut Français de Bioinformatique, IFB-core
2 rue Gaston Crémieux
F-91000 Evry, France

### Jérôme Pansanel
Université de Strasbourg, IPHC, CNRS, UMR7178
23 rue du Loess
F-67037 Strasbourg, France

**Olivier Sallou**
Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA
263 avenue Général Lerclerc
F-35000 Rennes, France

**Bruno Spataro**
CNRS, UMR 5558, LBBE - Biométrie et Biologie évolutive, UCB Lyon 1
43 bd du 11 novembre 1918
F-69100 Villeurbanne, France

## Résumé

*L'Institut Français de Bioinformatique (IFB) propose différents services pour le traitement des données des sciences de la vie, en partie basés sur une fédération de clouds académiques. Le portail Biosphère (https://biosphere.france-bioinformatique.fr) fournit plusieurs interfaces pour simplifier l'usage du cloud de l'IFB : le catalogue RAINBio des environnements modèles (appliances), un tableau de bord pour gérer les déploiements et un registre des données publiques disponibles.*

*La fédération IFB-Biosphère, initiée fin 2016, comporte 5 400 cœurs et 27 téraoctets de mémoire, répartis entre 6 sites basés sur Openstack, fédérés avec le système Nuvla. En plus des composants de base, d'autres plus spécifiques comme Manila pour la fourniture de volumes partagés en mode fichier, sont requis pour la majorité des applications bioinformatiques. La gestion des utilisateurs repose sur les identifiants institutionnels de la fédération d'identités eduGAIN, avec un proxy "keycloack" et des clients OpenID Connect.*

*Les appliances bioinformatiques proposent de nombreux outils courants pour l'analyse de données biologiques, 33 sont actuellement publiées dans le catalogue RAINBio. Ces environnements fournissent des outils comme "conda", "docker" ou "ansible"; des interfaces scientifiques de haut-niveau (portails web Rstudio ou Jupyter Notebook), ou un bureau graphique à distance. Certains environnements comprennent plusieurs composants reposant sur autant de machines virtuelles ou conteneurs. Le quota de base, extensible, permet de déployer des VMs, avec jusqu'à 128 cœurs et 3 To de RAM.*

*Le cloud IFB-Biosphère est utilisé pour des analyses scientifiques pouvant être intensives (4 000 cœurs), et par de nombreuses sessions de formation, écoles scientifiques, cursus de masters universitaires, workshops ou hackathons.*

## Mots-clefs

*Sciences de la vie, Bioinformatique, Calcul scientifique, Traitement des données scientifiques, Cloud computing*

## 1 Introduction

L'Institut Français de Bioinformatique (IFB, www.france-bioinformatique.fr) propose différents services pour le traitement des données des sciences de la vie. Ces services s'appuient sur une infrastructure distribuée entre les plateformes régionales membres de l'IFB proposant deux types d'environnements de calculs et traitements, suivant un modèle de cluster ou celui de cloud computing. Une partie de l'offre de services de l'IFB est ainsi basée sur une fédération de clouds académiques (détails en ligne) . Cette infrastructure IFB-Biosphère est distribuée entre les plates-formes participantes sous la forme d'une fédération de clouds, fournissant des services standards et personnalisables.

Il faut garder à l'esprit que la biologie et la bioinformatique présentent une caractéristique forte, à savoir, une profusion de logiciels et de données. Un nombre important d'acteurs ont développé des milliers de bases de données et d'outils, au service d'un domaine dynamique couvrant la biologie, la biotechnologie et la médecine. Les chercheurs doivent composer avec des données biologiques intrinsèquement complexes, intégrées dans des centaines de formats pour être analysées par un grand nombre de logiciels via diverses interfaces sur des infrastructures variées. Les développements des outils bioinformatiques sont souvent ponctuels et, en l'absence d'une source d'information unifiée, il n'est pas aisé d'évaluer la portée et la compatibilité des nouvelles ressources dans le contexte des offres académiques. Par exemple, les logiciels peuvent ne pas avoir de description officielle de leur fonction scientifique et technique, et l'absence d'identificateurs d'outils uniques et persistants nuit à la fiabilité des citations et à la reproductibilité des analyses. Il existe des obstacles importants pour trouver et interconnecter les bons outils parmi une multitude de possibilités, ce qui rend le travail du bioinformaticien - développer des workflows pratiques pour la découverte scientifique - loin d'être négligeable. A titre d'exemple, la ressource bio.tools (https://bio.tools) recense environ 13 000 outils. La proposition d'environnements prêts à l'emploi sous forme de machines virtuelles (VM) est par conséquent très intéressante pour l'utilisateur.

Les environnements virtuels bioinformatiques disponibles fournissent de nombreux outils courants pour l'analyse de données biologiques, 33 sont actuellement publiés dans le catalogue RAINBio. Ces environnements fournissent des outils système comme 'conda', 'docker' ou 'ansible'; des interfaces scientifiques de haut-niveau comme les portails web Rstudio ou Jupyter, ou un bureau graphique à distance.

# 2 L'institut Français de Bioinformatique

L'Institut Français de Bioinformatique (IFB) est l'infrastructure nationale de service en bioinformatique créée dans le cadre du programme national des «Investissements d'Avenir» (ANR-11-INBS-0013). Elle mutualise, soutient et coordonne le développement des ressources et des activités de support à la recherche de plateformes de bioinformatique dépendant des organismes publics de recherche CNRS - INRA - INRIA - CEA - INSERM, des universités, du CIRAD, et des Instituts Pasteur et Curie.

## 2.1 Missions

L'IFB a pour mission d'offrir aux communautés des sciences de la vie et de la bioinformatique, du monde académique et privé, un accès aux services qui sont vitaux pour leur recherche, un accompagnement de projets reposant sur un fort niveau d'expertise, et la possibilité de participer à des projets ambitieux au niveau national et international. Afin de maintenir la recherche française au plus haut niveau de compétitivité et de performance dans l'analyse bioinformatique, l'IFB anticipe les futurs besoins du domaine et participe aux innovations méthodologiques, en particulier pour répondre aux challenges de la bioinformatique intégrative. L'IFB est le noeud français de l'infrastructure européenne de bioinformatique ELIXIR (ESFRI).

## 2.2 Organisation

L'Institut Français de Bioinformatique est composé d'une UMS - unité mixte de services, avec cinq organismes de tutelle (CNRS UMS3601, INRA UMS1385, INSERM US2, CEA, INRIA) et de 32 plateformes bioinformatiques régionales (*cf.* la Figure 1 présentant leur localisation géographique). Son organisation est centrée sur des interactions continues avec la communauté scientifique, les tutelles et avec les autres infrastructures nationales et internationales des sciences de la vie. La gouvernance (détails en ligne) repose sur plusieurs comités (collège de direction, conseil de direction, comité en charge de la stratégie et de l'orientation scientifique (CCSO), comité d'éthique, comité de conseil industriel) et plusieurs cellules opérationnelles (valorisation et de communication, interaction plateformes régionales IFB-core, responsables d'actions et groupes de travail).

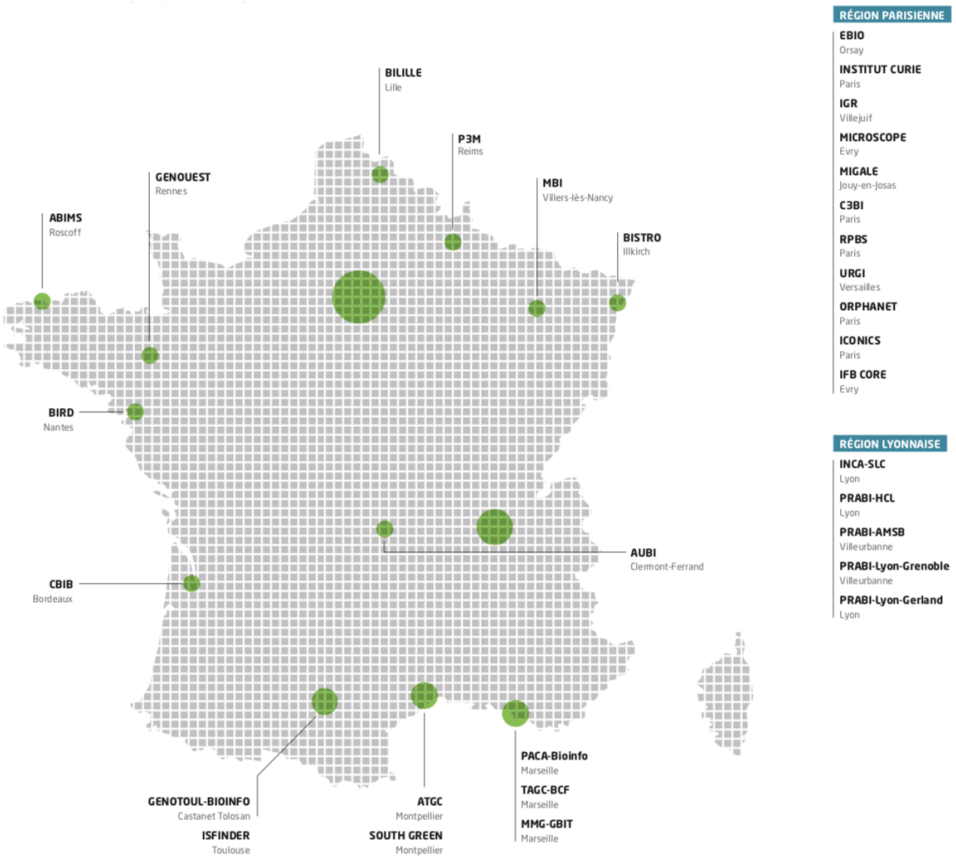


*Figure 1 - Organisation géographique de l'Institut Français de Bioinformatique*

## 2.3 Feuille de route 2018-21

Pour relever les défis techniques tout en apportant le service le plus adapté aux différentes communautés partenaires, l'IFB a proposé une feuille de route, validée par le M.E.S.R.I, actuellement en cours d'exécution. L'ambition est de proposer une infrastructure qui permettra de contribuer au développement de projets d'envergure pour les sciences biologiques et médicales.

Les actions de l'IFB pour la période 2018-21 ont été structurées en différents groupes de travail et tâches dont les grandes lignes sont les suivantes (détails en ligne) :

- WP1. Un environnement distribué de services en bioinformatique
  - A1.1 Réseau national de ressources informatiques (NNCR)
  - A1.2 Environnements logiciels et données
  - A1.3 Support aux bases de données
  - A1.4 Catalogue des ressources
  - A1.5 Accès aux usagers
  - A1.6 Mutualisation des services inter-infra
  - A1.7 Guichet d'orientation et de consultation
- WP2. Innovation: bioinformatique intégrative
  - A2.1 Projets pilotes inter-infrastructures
  - A2.2 Appel à défis: lever les verrous scientifiques et technologiques
  - A2.3 Interopérabilité entre ressources

- WP3. Ouverture internationale + industrie
    - A3.1 IFB, nœud français d'ELIXIR
    - A3.2 Partenariat avec l'industrie
- WP4. Formation et diffusion
    - A4.1 Formation
    - A4.2 Actions jointes avec SFBI + GDR BIM
    - A4.3 Communication & Valorisation
- WP5. Gouvernance
    - A5.1 Structures de gouvernance et de coordination
    - A5.2 Système de gestion de qualité
    - A5.3 Modèle économique

# 3 L'infrastructure cloud IFB-Biosphère

L'IFB opère une infrastructure informatique et bioinformatique distribuée entre les plateformes régionales membres de l'IFB proposant deux types d'environnements de calculs et traitements, suivant le modèle de cluster ou celui de cloud computing (détails en ligne). Une partie de l'offre de services de l'IFB est ainsi basée sur une fédération de clouds académiques dont il est question dans cet article.

La fédération de clouds IFB-Biosphère a été initiée à la fin 2016, et comporte actuellement plus de 5 400 cœurs de calcul et 27 téraoctets (To) de mémoire. Ces ressources sont réparties entre 6 sites : GenOuest, PRABI-LBBE, BiRD, BIstrO, Bilille et le nœud national IFB-core. Certains de ces clouds fonctionnent depuis le début des années 2010. Et 4 autres plates-formes bioinformatiques de l'IFB souhaitent raccorder leur cloud existant à la fédération IFB-Biosphère, ou installent un cloud sur leurs ressources pour le raccorder.

L'infrastructure cloud IFB-Biosphère est accessible à l'ensemble de la communauté des sciences de la vie, avec un quota de ressources de base, extensible selon différents critères. Les scientifiques peuvent ainsi déployer en un clic leur propre environnement d'analyse avec des ressources qui leur sont réservées. Ces environnements modulaires peuvent aller de 1 cœur de CPU avec 2 Go de mémoire à 128 cœurs avec 3 To de RAM pour une seule machine virtuelle, et jusqu'à des centaines ou milliers de cœurs avec des centaines de Go ou plusieurs To de mémoire dans de nombreuses machines virtuelles (cf. Figure 2).

## 3.1 Le portail Biosphère

Le portail Biosphère (https://biosphere.france-bioinformatique.fr) fournit plusieurs interfaces pour simplifier l'usage de l'infrastructure cloud distribuée de l'IFB :

- le catalogue RAINBio des appliances cloud, qui référence les environnements basés sur des machines virtuelles, et dans certains cas associées à des conteneurs, prêts à être déployés en un clic, dimensionnés pour différentes tâches bioinformatiques,
- un tableau de bord qui permet à chaque usager de gérer ses déploiements dans le cloud IFB-Biosphère, qu'ils reposent sur une seule ou plusieurs machines virtuelles,
- un centre de données qui recense les banques de données publiques disponibles dans les clouds IFB-Biosphère, et les volumes partagés des utilisateurs. Ces données, accessibles en mode fichier, sont montées directement dans les machines virtuelles des utilisateurs.

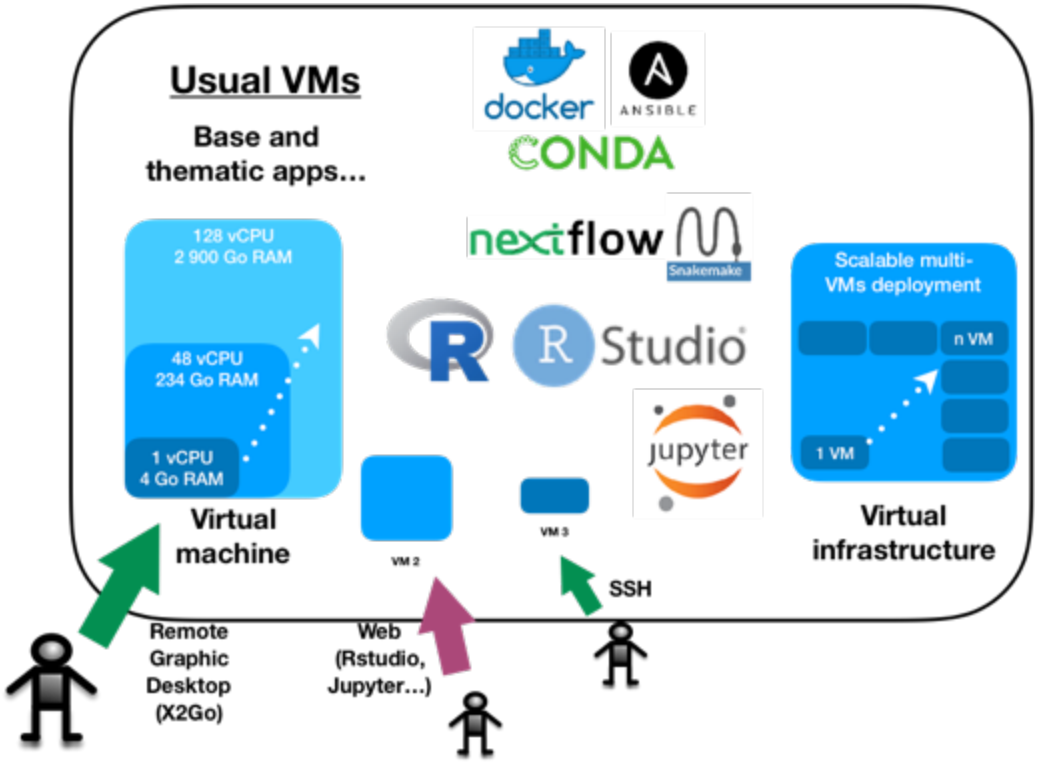


*Figure 2 - Le cloud IFB-Biosphère fournit des environnements modulaires allant d'une seule machine virtuelle à plusieurs dizaines, jusqu'à 48 vCPU et 250 Go RAM par VM pour les environnements standards, et une taille mémoire jusqu'à 3 To pour les environnements à grosse capacité mémoire*

## 3.2 Les sites clouds

Tous les clouds IFB sont basés sur le système logiciel Openstack pour fournir les services cloud. Les composants utilisés sont ceux de base (`keystone`, `nova`, `glance`, `cinder`, `swift`) et ceux plus spécifiques comme `manila` pour la fourniture de stockages partagés en mode fichier (en cours de prototypage), qui sont requis pour la grande majorité des applications en sciences de la vie.

Les services des différents clouds sont fédérés avec le système Nuvla (détails en ligne). Celui-ci propose des connecteurs pour les grands types de clouds du marché et permet de gérer différents sites de façon uniforme, tant pour la gestion des machines virtuelles que pour les configurations à leur appliquer (cf. Figure 3).

La gestion des utilisateurs dans la fédération IFB-Biosphère s'appuie sur la fédération d'identités eduGAIN, un proxy de fédération basé sur le logiciel 'keycloack', et des clients OIDC (OpenID Connect) dans les différents services Biosphère.

Le portail Biosphère permet ainsi :

- aux scientifiques d'utiliser les différents clouds d'une manière uniformisée et simplifiée avec leurs identifiants institutionnels,
- aux développeurs de construire de nouveaux environnements de traitement avec les outils système communément utilisés (apt/yum, pip, conda, docker, ansible, git…),
- et propose aux administrateurs de site des outils de gestion avancés et complémentaires des outils courants Openstack.

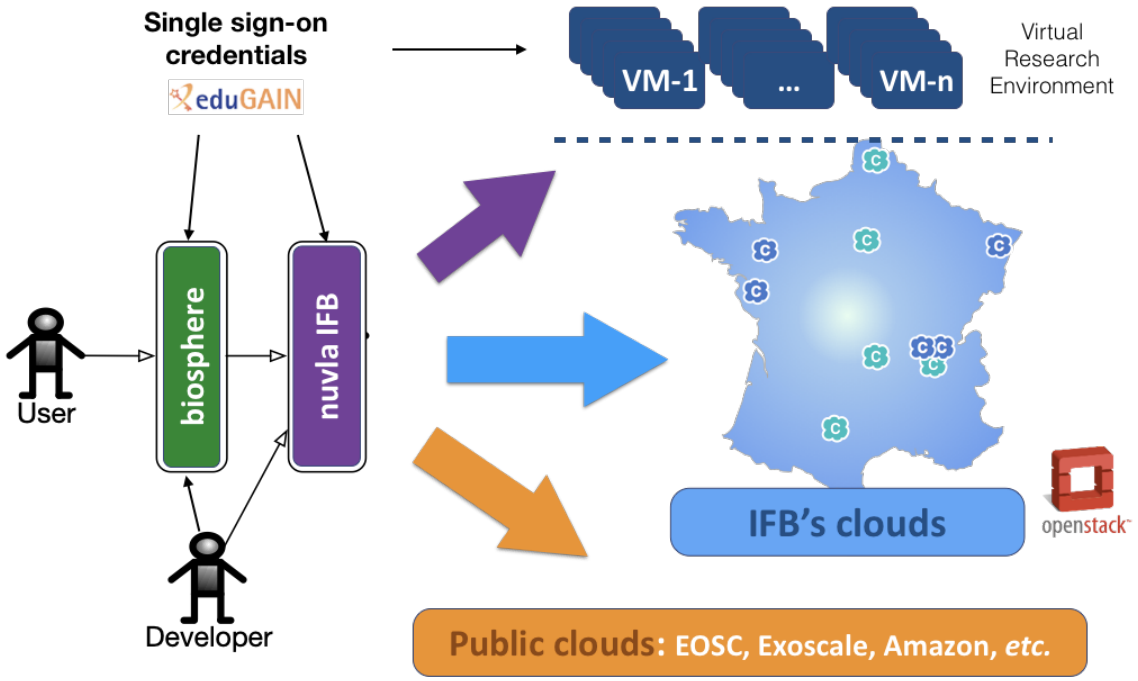


*Figure 3 - Architecture de la fédération de clouds IFB-Biosphère*

## 3.3 Les données biologiques

Avec la production de données, la gestion, le stockage et, par conséquent, l'extraction, l'analyse et l'interprétation des données sont au cœur de tout projet de recherche biologique. De plus, la capacité d'avoir accès aux ensembles de données de référence est souvent cruciale pour la bonne réalisation du projet. Les environnements virtuels proposés sur l'infrastructure de l'IFB offrent donc divers volumes de stockage pour répondre aux besoins de la communauté scientifique.

Les banques de données biologiques sont des entrepôts dédiés au stockage de l'information biologique. La présence de ces données de référence est un pré-requis sur chaque plate-forme bioinformatique pour l'analyse des données expérimentales. La revue *Nucleic Acids Research* publie annuellement un numéro spécial sur les banques de données biologiques et recense la liste complète des banques actives. Le dernier numéro présente une liste de 1613 banques de données [2].

La gestion des données de référence est dévolue à BioMAJ [3], un outil open source pour la gestion des banques de données biologique. Il est couramment utilisé dans de nombreuses installations de bioinformatique pour fournir à tous les utilisateurs les principales données biologiques publiques. BioMAJ est un moteur de workflow dédié à la synchronisation et au traitement des données. Il automatise le cycle de mise à jour, la transformation et la supervision du référentiel de la banque de données en miroir local. Il a également été intégré au portail bioinformatique Galaxy. BioMAJ offre ainsi le moyen de créer et de maintenir plusieurs ensembles de données provenant de différents endroits sans avoir à gérer des tâches complexes.

Différents types de stockage (détails dans le Tableau 1) sont disponibles sur les sites clouds :

- les banques publiques de référence en science de la vie, qui sont accessibles depuis toutes les VMs avec le montage du répertoire */ifb/data/public* ,
- les données d'un utilisateur, disponibles dans un volume partagé entre toutes ses VMs d'un même site,

- les données d'un projet, disponibles dans un volume partagé entre toutes les VMs des membres de ce projet sur un même site,

- un disque local, directement sur le serveur hôte, sur les sites le proposant.

Tableau 1 – Les différents types de stockage IFB-Biosphère et leurs fonctionnalités.

| **Stockage** | **Droits** | **Disponibilité** | **Chemin** | **Mise en oeuvre** |
|---|---|---|---|---|
| **Banques publiques** | *Lecture* | Toutes VMs | `/ifb/data/public` | BioMaJ |
| **Ephémère** | *Ecriture* | Interne à la VM | `/ifb/data/mydatalocal` | Disque local |
| **Partage usager** | *Ecriture* | Toutes VMs de l'usager | `/ifb/data/mydatashare` | Manila |
| **Partage projet** | *Ecriture* | Toutes VMs des membres du projet | `/ifb/data/<nom_projet>` | Manila |

# 4 Services scientifiques

## 4.1 Les appliances cloud – environnements virtuels de recherche pour l'analyse des données biologiques

Les appliances bioinformatiques du cloud IFB-Biosphère sont disponibles en différents formats pour différentes thématiques, permettant aux scientifiques, biologistes et bioinformaticiens, de choisir le plus approprié pour leurs analyses. Il y a actuellement 33 environnements modèles, développés par les membres de l'IFB, référencés dans le catalogue RAINBio comme présenté sur la vue générale de la Figure 4.

Les appliances bioinformatiques de l'IFB proposent de nombreux outils bioinformatique (260+) et modules R (pour les statistiques), couramment utilisés pour l'analyse de données dans différents domaines comme en génomique, bio-imagerie, réseaux métaboliques, écologie microbienne, protéomique ou métabolomique. Ces environnements virtuels de recherche se déploient avec la configuration type définie par leurs développeurs (cf. Figure 5), mais tous peuvent être adaptés par l'utilisateur suivant ses besoins sans interférer avec les autres usagers.

Tous les environnements incluent des outils technologiques comme `pip`, `conda` (avec les canaux *bioconda*, *R* et *conda-forge* pré-configurés), `docker` pour les conteneurs, ou `ansible` pour le déploiement automatisé de logiciels. D'autres environnements proposent des interfaces scientifiques de haut-niveau reposant sur des portails web (comme `Rstudio` ou `Jupyter Notebook/Lab`) ou des interfaces graphiques (GUI) à travers un bureau virtuel à distance. Enfin, certains environnements comprennent plusieurs composants reposant sur autant de machines virtuelles ou conteneurs, comme des clusters de calcul (`SGE` ou `SLURM`) ou des environnements d'exécution de workflows bioinformatiques (`Nextflow`, `Snakemake` ou `CWLtool`).

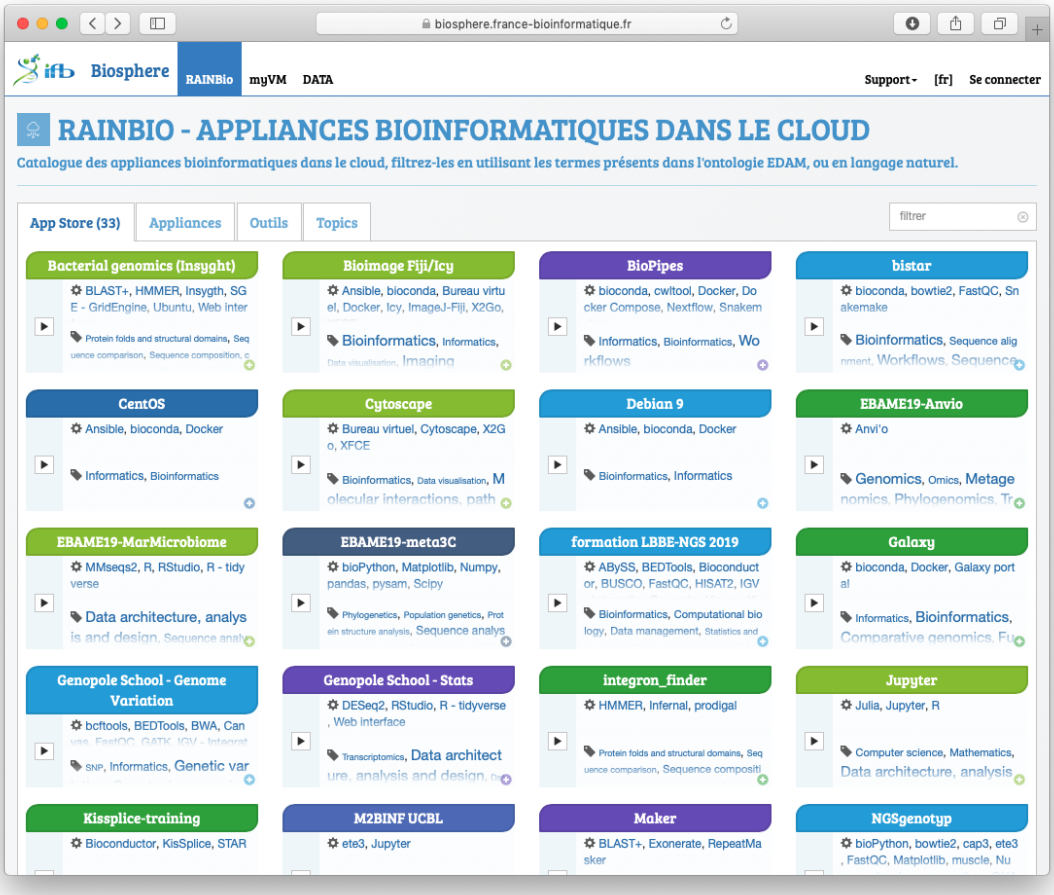


*Figure 4 - Vue principale du catalogue RAINBio avec différents environnements bioinformatiques disponibles.*

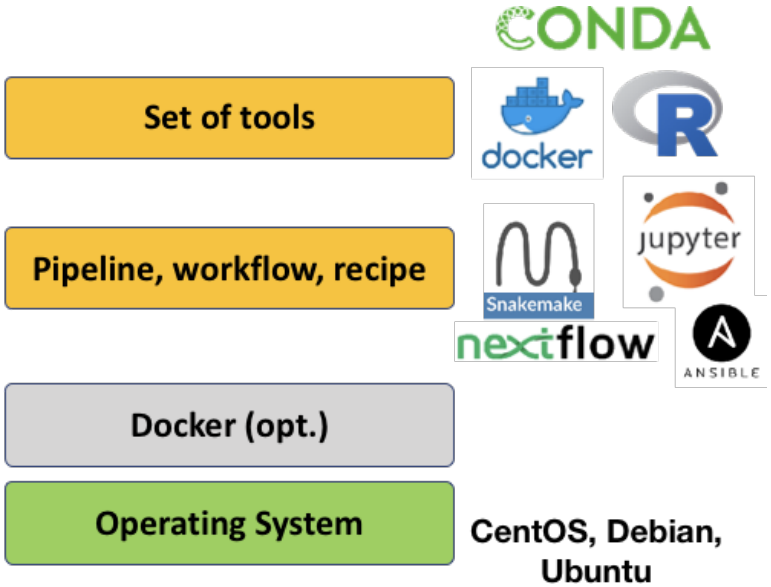


*Figure 5 - Principe d'intégration des logiciels bioinformatiques avec les différentes techniques logicielles et système disponibles.*

## 4.2 Le soutien aux formations et événement scientifiques

L'IFB propose de nombreuses formations scientifiques en biologie (détails en ligne), ainsi que des formations à l'utilisation du cloud avec les modules IBI (détails en ligne), et les « Ateliers du Cloud IFB-Biosphère : Usage Avancé et Développement d'Appliances » (détails en ligne). Ces derniers notamment, traitent sur deux journées de l'architecture du cloud IFB-Biosphère et son utilisation courante, ainsi que de la présentation des bonnes pratiques pour une utilisation avancée. Les points abordés vont de l'installation de nouveaux outils, l'utilisation de pipelines et workflows, à l'intégration avancée des outils bioinformatiques pour la création d'appliances. La seconde journée est consacrée à un hackathon pour l'intégration d'outils et pipelines apportés par des participants volontaires, en collaboration avec les autres participants et formateurs cloud présents.

L'infrastructure cloud IFB-Biosphère est également utilisée par de nombreuses sessions de formation, écoles scientifiques, cursus de masters universitaires, workshops ou hackathons, dont certains depuis plusieurs années, dont quelques exemples sont proposés dans le Tableau 2. Ces différents évènements et formations représentent plusieurs centaines de participants, ayant bénéficié d'environnements bioinformatiques existants ou développés spécifiquement par les formateurs.

Tableau 2 – Exemples d'évènements et formations scientifiques supportés par le cloud IFB-Biosphère pour leurs sessions pratiques.

| **Date** | **Intitulé** | **Organisateur** | **Partici-pants** | **Ancienneté sur le cloud** |
|---|---|---|---|---|
| Sept. 2018 | Elixir Training: Plant Genome Assembly and Annotation | IFB-core, CIRAD | 25 | - |
| Juin 2019 | Summer school Genopole | Genoscope, Evry | 20 | 3 ans |
| Sept. 2019 | Bioinformatique pour le traitement de données de séquençage | LBBE, Lyon<br>& CNRS Entreprise | 12 | 2 ans |
| Oct. 2019 | EBAME Workshop on Computational Microbial Ecogenomics | IUEM, Brest | 50 | 2 ans |
| Oct. 2019 | Master AMI2B | Université Paris Saclay | 15 | 5 ans |
| Nov. 2019 | ReproHackathon-3 (GDR MaDICS) | LRI, IFB-core, CIRAD | 20 | 3 ans |
| 2019-20 | Master Biologie ENS-Lyon - Practicals in next generation sequencing | ENS-Lyon | 20 | 4 ans |
| 2019-20 | Master 2 de bioinformatique | Université Claude Bernard Lyon1 | 20 | - |

## 4.3 Exemple d'utilisation

L'IFB a porté sur le cloud l'application d'analyse de génomes bactériens Insyght, afin de tirer parti des avantages du cloud pour le déploiement d'environnements virtuels de recherche personnalisés et adaptés aux grandes échelles d'analyse.

L'application Insyght, développée par la plate-forme Migale (INRA MIA), permet l'analyse de la co-localisation des gènes de génomes bactériens à différentes échelles de proximité, en particulier la

conservation du voisinage des gènes dans différents génomes. L'information extraite de la co-localisation des gènes est complémentaire de celle, classique, tirée de la recherche d'homologie et est utilisée pour améliorer l'annotation des nouveaux génomes séquencés.

Le portage Cloud de l'application Insyght a été réalisé dans le cadre du projet européen CYCLONE (H2020 644925) en collaboration avec la société SixSq (Genève), éditrice de la solution SlipStream. Cette solution logicielle facilite le déploiement multi-cloud d'environnements virtuels de recherche. Dans le cas d'Insyght, cela permet sa diffusion et son utilisation intensive en s'affranchissant des limites de ressources locales. Un scientifique est ainsi capable de déployer en quelques clics un environnement complet d'Insyght sur plusieurs clouds avec les ressources requises.

Un exemple est l'analyse de 5 663 génomes bactériens (plus de 16 millions de comparaisons 2 à 2) qui a été déployée sur 4 100 processeurs du cloud de l'IFB. Cette expérience demandant normalement 52 000 heures de calcul a été réalisée en 13 heures et a produit 2,4 téraoctets (To) de données.

# 5 Conclusion

L'Institut Français de Bioinformatique (IFB) propose différents services pour le traitement des données des sciences de la vie. Une partie de ceux-ci sont basés sur la fédération de clouds académiques IFB-Biosphère, réparties entre 6 sites, utilisant Openstack et fédérés avec le système Nuvla. Les environnements cloud proposent de nombreux outils bioinformatiques courants pour l'analyse de données biologiques, et sont publiés dans le catalogue RAINBio. Certains environnements comprennent plusieurs composants reposant sur autant de machines virtuelles ou conteneurs.

Le cloud IFB-Biosphère fournit ainsi aux scientifiques :

- Un identifiant unique, utilisant les identifiants académiques des organismes nationaux et internationaux avec la fédération d'identité eduGAIN.
- Un portail web unifié pour déployer les environnements virtuels de recherche sur tous les clouds de l'infrastructure
- Plus de 5 400 vCPU et 27 To RAM
- Des environnements cloud modulaires allant d'une seule machine virtuelle à plusieurs dizaines, avec jusqu'à 48 vCPU et 250 Go RAM par VM pour les environnements standards
- Une taille mémoire jusqu'à 3 To pour les environnements à grosse capacité mémoire
- Des environnements bioinformatiques pré-définis, déployables en un clic depuis le catalogue RAINBio
- Les droits d'administration de leurs VMs pour installer d'autres outils bioinformatiques et les configurer
- Une haute-disponibilité des ressources grâce à l'unification des différents sites de la fédération
- Les banques publiques de référence en science de la vie
- Le soutien des formations, ateliers et écoles scientifiques (CPU, RAM, stockage, experts IFB)

L'infrastructure mise en place dans le cadre d'IFB-Biosphère est un des moyens de promouvoir une reproductibilité des résultats scientifiques puisque les systèmes virtualisés dans le cloud sont réplicables aisément. Les images système pré-configurées avec des logiciels et outils personnalisés de façon normalisée facilitent les tâches courantes dans un domaine scientifique (*e.g.* l'assemblage de séquences de génomes à partir des données produites par un séquenceur d'ADN) en garantissant une reproductibilité.

Ces images pré-configurées, de conteneurs et machines virtuelles, peuvent alors être partagées comme ressources publiques pour diffuser un logiciel ou une méthode.

L'infrastructure est utilisée pour des analyses scientifiques intensives (jusqu'à 4 000 cœurs de calcul) et par de nombreuses sessions de formation, écoles scientifiques, cursus de masters universitaires, workshops ou hackathons, dont certaines depuis plusieurs années.

## Bibliographie


[1] J. Ison et al., The bio.tools registry of software tools and data resources for the life sciences. Genome Biol., 2019, 20 (1): 164. https://doi.org/10.1186/s13059-019-1772-6

[2] D. J. Rigden and X. M. Fernández. The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. Nucleic Acids Research, 2019, 47, D1–D7. https://doi.org/10.1093/nar/gky1267

[3] O. Filangi et al., BioMAJ: a flexible framework for databanks synchronization and processing. Bioinformatics, 2008, 24(16): 1823–1825. https://dx.doi.org/10.1093/bioinformatics/btn325