Towards a cyber Galaxy ?

Christophe C. Caron, Wilfried Carre, Alexandre Cormier, Sandra S. Derozier, Franck Giacomoni, Olivier Inizan, Gildas Le Corguillé, Alban Lermine, Sarah Maman Haddad, Pierre Pericard, et al.

Toward a cyber Galaxy ?

Christophe CARON[1], Wilfrid CARRE[1], Alexandre CORMIER[1], Sandra DEROZIER[2], Franck GIACOMONI[3], Olivier INIZAN[4], Gildas LE CORGUILLE[1], Alban LERMINE[5], Sarah MAMAN[6], Pierre PERICARD[1] and Franck SAMSON[2]

[1] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
{christophe.caron, wilfrid.carre, alexandre.cormier, gildas.lecorguille, pierre.pericard}@sb-roscoff.fr

[2] INRA, UR1077, MIGALE, Centre de Jouy-en-Josas, 78352, Jouy-en-Josas, France
{sandra.derozier, franck.samson}@jouy.inra.fr

[3] PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Theix, 63122, Saint Genes Champanelle, France
franck.giacomoni@clermont.inra.fr

[4] INRA, UR1164,, Route de St Cyr, Versailles, France
olivier.inizan@versailles.inra.fr

[5] Institut Curie, INSERM, U900, Bioinformatics and Computational Systems Biology of Cancer, 75248 Paris, France
alban.lermine@curie.fr

[6] INRA, UMR444, Laboratoire de Génétique Cellulaire, Centre de Toulouse Auzeville, 24 Chemin de Bordé Rouge, 31320 Auzeville-Tolosane, France
sarah.maman@toulouse.inra.fr

Abstract *The success of the open web based platform "Galaxy" is growing among diverse scientific communities. The French Institute of Bioinformatics - IFB wish to initiate a collaborative work dedicated to scientific workflows and especially to the platform Galaxy. We report here the main items on which future collaborations could be build: (i) software and hardware architecture, (ii) tools integration and (iii) training.*

Keywords Galaxy, training, workflow, NGS, tools integration, data sharing

Vers une Cyber Galaxy ?

Résumé *Le portail Galaxy dédié à l'activité de bio-analyse connaît un succès croissant au sein de multiples communautés scientifiques. L'Institut Français de Bioinformatique (IFB) souhaite mener une action transversale dédiée aux workflows d'analyse de données et en particulier à la plateforme Galaxy. Nous présentons ici les axes majeurs de cette action en termes d'architecture logicielle et matérielle, d'intégration d'outils et de formations.*

Mots-clés Galaxy, formation, workflow, NGS, intégration d'outils, partage de données.

1 Introduction

L'Institut Français de Bioinformatique (IFB) a pour mission la coordination de la communauté bio-informatique nationale. Dans un contexte où l'analyse des données à haut-débit modifie considérablement la façon de mener des analyses avec la mobilisation de nouvelles infrastructures, de nouveaux outils et de nouvelles compétences, l'IFB a décidé de mener une action transversale autour des solutions dédiées aux « workflows » d'analyse de données, avec en particulier la plateforme Galaxy [1].

Galaxy est une plateforme scientifique web, libre et « open source », permettant la mise à disposition d'outils d'analyses orientés principalement pour la bioinformatique (NGS, Métabolomique, etc…) et les

statistiques à un large panel d'utilisateurs. L'émergence de telles plateformes est liée au fait qu'une grande majorité des outils couramment utilisés, le sont via la ligne de commande, en limitant l'accès aux seuls spécialistes. Ainsi, Galaxy propose une méthodologie d'intégration d'outils issus de sources multiples et dispose d'un gestionnaire de workflows intuitif, d'un gestionnaire d'historique assurant la reproductibilité des analyses et d'un environnement de partage des données, des résultats, des outils, des worflows et des méthodologies d'analyses. L'un des objectifs du projet Galaxy est, via l'interfaçage web unifié d'outils et de fonctionnalité, de rendre accessible à des non-bioinformaticiens la réalisation d'analyses *in silico* évoluées.

Il existe autour de la plateforme scientifique Galaxy une communauté très active d'utilisateurs et de développeurs qui couplée aux fonctionnalités de transfert d'outils d'une instance d'un laboratoire à une autre fait que cet « environnement de travail » favorise le partage des développements et des collaborations entre les producteurs d'algorithme et les analystes.

Dans ce contexte, de nombreuses initiatives ont vu le jour depuis quelques mois autour de la plateforme Galaxy : communauté Galaxy-France, école bio-informatique Aviesan, Groupe Galaxy Aplibio. Afin de fédérer et structurer ces actions au niveau national, un groupe de travail a été constitué autour de plusieurs plateformes IFB (ABiMS, Curie, Genotoul/SIGENAE, MIGALE, URGI), et d'une plateforme de l'infrastructure nationale MetaboHub (PFEM). L'objet principal de cet article est d'exposer les axes de travail de cette action transversale en mettant l'accent sur les premières actions structurantes autour des bonnes pratiques de développements et de déploiement, ainsi que des formations qui verront le jour en 2013.

2 Axes de travail

Avec l'avènement des infrastructures réparties (localisation des données, multiplicité des outils, etc.), le déploiement des composants nécessaires au partage des données et aux traitements peut rapidement devenir une réelle difficulté, limitant ainsi les performances ou la qualité du dispositif. Avec la généralisation des infrastructures reposant sur Galaxy, il nous est apparu important de valoriser les retours d'expérience des plateformes qui ont été confrontées à différents verrous. Il s'agit donc pour le groupe de travail de proposer et partager des schémas organisationnels et techniques, en vue de proposer une infrastructure Galaxy cohérente avec la stratégie de l'IFB, notamment dans un contexte de Cloud académique.

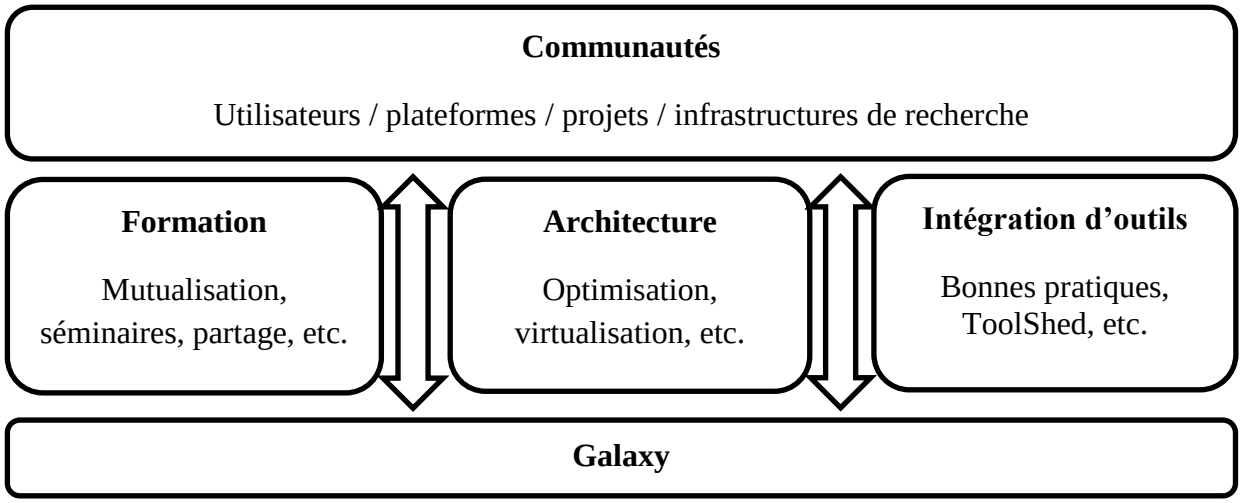

Figure 1. Axes autour du framework Galaxy travail

2.1 Architecture logicielle et matérielle

L'environnement Galaxy est constitué de plusieurs composants comme par exemple le gestionnaire de « jobs », le frontal web et la partie bases de données, dédiés à des tâches bien définies. Son déploiement et sa mise en œuvre peuvent présenter différentes finalités (formation, phases de prototypage de développement et

d'exploitation, etc.), ce qui peut ainsi induire un niveau de qualité de service variable en termes de disponibilité et de scalabilité.

Ces différents contextes d'exécution font souvent appel à différentes techniques : virtualisation des instances, installation sur poste de travail, partage des données plus ou moins évolué, etc. Un des premiers résultats a consisté à valider la capacité de Galaxy à tenir la montée en charge tout au long d'une école thématique (Ecole Bio-informatique Aviesan, janvier 2013 - http://galaxy-ecole.sb-roscoff.fr/).

Les modifications et les optimisations apportées ont permis de supporter un grand nombre de connexions simultanées, dans un contexte d'hétérogénéité à plusieurs niveaux : typologie des analyses, volumétrie, diversité des codes parallèles, niveau de complexité des workflows, etc. Il a ainsi été possible de proposer un environnement supportant plusieurs dizaines d'utilisateurs simultanés sur des workflows d'analyses pouvant atteindre une centaine d'étapes. En continuant à explorer les différentes pistes d'optimisation, un des objectifs sera de proposer différents scénarios d'implémentation tenant compte du contexte d'usage, en puisant dans les retours d'expérience de la communauté.

Plus largement, les objectifs de ce volet consisteront à aborder les problématiques d'exploitation en environnement de production, les possibilités d'automatisation des tâches de déploiement et les solutions de monitoring d'instances de serveurs Galaxy.

2.2 Intégration d'outils

Les analyses de données à haut débit font appel à de nombreuses applications réparties sur un ou plusieurs sites. L'intégration de ces applications peut dès lors se révéler fastidieux. Ceci est d'autant plus vrai dans le cadre de déploiement de workflows, où s'enchaînent de nombreuses étapes et traitements. La plateforme Galaxy apporte des solutions concrètes à ces situations en centralisant tous les outils nécessaires à l'analyse et en proposant un modèle d'intégration simple pour ces outils [2].

L'objectif de cet axe est de proposer des outils et des processus facilitant l'interfaçage des outils dans une instance Galaxy. Nous pourrons aborder des situations comme (i) la montée en version d'outils, (ii) la maintenance d'outils « maison » en relation avec les montées en version de l'instance Galaxy, (iii) le test d'outils et de workflows nouvellement intégrés ou mis à jour.

Cette approche peut aussi bien passer par la rédaction d'un guide des bonnes pratiques, que par une veille technologique autour des méthodes proposées par la communauté.
Afin d'optimiser, en termes de facilité et de rapidité, l'intégration d'outils dans les instances locales de Galaxy, un des objectifs sera également de réfléchir sur l'intérêt du déploiement d'un « ToolShed » Galaxy propre à l'IFB prônant un haut niveau de qualité du descriptif et du contenu des outils partagés.

2.3 Diffusion, valorisation et formations

Les différentes plateformes du groupe de travail sont aujourd'hui amenées à proposer des formations Galaxy pour favoriser son apprentissage (Curie, Genotoul, ABiMS, MIGALE, URGI). Elles utilisent également Galaxy comme support de formation dans les principaux domaines de recherche en bio-informatique : analyses de données de transcriptome, détection de SNP, ChIP-seq, etc. Certaines de ces formations sont aussi proposées lors de programmes internationaux comme la conférence Galaxy 2013 à Oslo (ChIP-seq – Institut Curie). Ce mouvement d'appropriation de Galaxy par la communauté démontre une fois de plus son caractère essentiel comme outil dans le transfert de compétences. La plateforme Genotoul/Sigeneae propose également depuis plusieurs mois, des autoformations en ligne ainsi qu'une FAQ ouvertes à l'ensemble des utilisateurs. Ces modules d'autoformation (Initiation à Galaxy, etc.) assurent une prise en main du « workbench » Galaxy et une transmission des bonnes pratiques d'utilisation.

Au final, ces formations constituent un point d'entrée idéal pour les biologistes et les bio-analystes pour l'apprentissage des bonnes pratiques d'analyses. Elles sont aussi un moyen de démocratisation dans l'usage des concepts autour des workflows.

Un des objectifs du groupe sera, en coordination avec l'ensemble du dispositif IFB, de favoriser la mutualisation des actions de formations au niveau national, et d'assurer une transition vers de nouveaux usages comme le E-learning, en utilisant notamment les compétences acquises par le pôle toulousain.

Conclusion & Perspectives

Les premiers travaux présentés ont permis de fédérer une première communauté qui avait déjà initié un certain nombre de réflexions autour de Galaxy. Les prochains mois vont permettre d'avancer sur les différents axes présentés, avec notamment la mise en ligne d'un site documentaire.

Il s'agira aussi de lier cette action d'animation avec les autres infrastructures nationales de recherche (MetaboHub, EMBRC-France, etc.) financées par le programme ANR des Investissements d'Avenir. L'utilisation de cette action structurante autour de Galaxy pourrait être une première étape vers un partage plus important des usages et bonnes pratiques en termes de développement, mais aussi vers la création de passerelles entre ces différents projets. Des collaborations scientifiques actuelles illustrent de belle manière que l'environnement « Galaxy » peut devenir un excellent média et une passerelle entre les communautés que sont les bio-informaticiens et les biologistes.

Enfin, afin de présenter les résultats des développements en cours, mais aussi d'enrichir le partage de connaissances, nous organiserons un premier séminaire ouvert à la communauté IFB à l'automne. Ce séminaire de deux jours proposera une session avec des retours d'expérience Galaxy, et une formation aux bonnes pratiques (intégration d'outils, E-learning, etc.). Cette animation a pour vocation d'être élargie à d'autres entités (plateformes, communautés, etc.) dans le cadre de projets collaboratifs soutenus par l'IFB.

Acknowledgements

This work was supported by ANR programs 'Investissements d'Avenir': MetaboHub, EMBRC-France, France Génomique, 'Institut Français de Bioinformatique'.

References

[1] B. Giardine, C.Riemer, R.C.Hardison, R.Burhans, L.Elnitski, P.Shah, Y.Zhang, D.Blankenberg, I.Albert, J.Taylor, W.Miller, W.J.Kent and A.Nekrutenko, Galaxy: A platform for interactive large-scale genome analysis. Nucleic Acids Res., 15: 1451-1455, 2005.

[2] J.Goecks, A .Nekrutenko, J.Taylor and The Galaxy Team, Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computaional research in the life sciences. Genome Biology, 11:R86, 2010.