Reunion GUGGO 2

Seconde Réunion Groupe des Utilisateurs de Galaxy du Grand Ouest (GUGGO)

04 juin – 14h/17h – IRISA / INRIA

Présents

Jeanne Cambefort (IE CNRS, GenScale), Laure Quintric (Ingénieur PCIM Ifremer Plouzané), Cédric Mendosa (stagiaire PCIM Ifremer Plouzané), Grégory Carrier (Postdoctorant Ifremer Nantes), Edouard Hirchaud (Plateforme Bio-informatique BRID Nantes), Audrey Bihouée (Plateforme Bioinformatique BRID Nantes), Stéphanie Mottier (IE IGDR), Cyril Falentin (INRA IGEPP), Aurélien Le Roult (Administrateur système INSERM GenOuest et PF Protéomique HD), Olivier Sallou (Responsable développement Université Rennes1 GenOuest), Philippe Vanderkoornhuyse (Professeur Université Rennes1 OSUR Rennes), Alexan Andrieux (Ingénieur INRIA Genscale), Olivier Quénez (Ingénieur INRIA GensCale), Sivasangari Nandy (Ingénieur GenOuest), Anthony Bretaudeau (IE INRA), Mathieu Bahin (IE GenOuest/PF Séquençage environnemental), Cyril Monjeaud (Ingénieur GenOuest) , Yvan Le Bras (CNRS /GenOuest)

Excusés

Olivier Collin (CNRS/GenOuest), Gilles Lassale (Ingénieur INRA IGEPP), Marc Aubry (IR Université Rennes1 PF Séquençage Santé)

Introduction

Suite à la création du groupe des utilisateurs de Galaxy du Grand Ouest, une réunion s'était tenue en avril 2012 afin de préparer la mise en place d'instances de Galaxy au niveau des différentes plateformes de Bio-informatique de Bretagne et Pays de la Loire. Nous nous réunissons ce 4 juin dans les locaux de l'IRISA / INRIA pour une seconde réunion GUGGO afin de faire un point sur le développement des différentes instances.

Point sur les différentes instances

La plateforme GenOuest IRISA/INRIA

L'instance Galaxy de la plateforme GenOuest est le fruit du travail de 6 personnes :

- Administration : Aurélien Roult
- Développement : Yvan Le Bras, Cyril Monjeaud, Olivier Quenez

(aujourd'hui sur un autre projet) et Mathieu Bahin
- Supervision : Olivier Collin

De septembre à novembre 2012, une première instance a été mise en place avec l'intégration d'outils issus de Symbiose principalement. Cependant, des difficultés (conflits) ont été rencontrées lors de la mise à jour à partir du serveur principal de Galaxy (main.g2.bx.psu.edu), c'est pourquoi une vraie stratégie a été mise en place. Pendant un temps, il a été envisagé de faire une instance par domaine (NGS, protéomique, etc.) mais il a rapidement été conclus que cela serait compliqué à maintenir et inadapté car les utilisateurs peuvent être amenés à faire des études transversales. De ce fait, il est important de veiller à une bonne organisation des outils.

Depuis novembre 2012, il existe deux instances, une de développement et une de production. L'instance de développement est mise à jour à partir du dépôt du main Galaxy (ce qui est parfois dangereux car les versions proposées ne sont pas toujours très stables). Cependant, l'équipe Galaxy est réactive pour corriger les erreurs lorsqu'elles sont soulevées. Les développements sont réalisés sur cette instance avant d'être basculés en production. Tout cela est géré à l'aide d'un dépôt Git. L'instance de production est mise à jour plus d'une fois par mois.

En terme de configuration, la connexion se fait via le LDAP, il n'y a pas d'accès sans un compte sur le cluster GenOuest. Les jobs sont exécutés sur le Genocluster SGE sur lequel un demi nœud est affecté à chacune des deux instances (8 cores / 144 Go de RAM / 11 To d'espace). La base de données est sous PostgreSQL. Une file spéciale a été créée pour répondre aux demandes dans le système de gestion de tickets de la plateforme GenOuest basé sur OTRS.

De nombreux outils issus de des équipes Genscale et Dyliss, de la plateforme GenOuest et autres logiciels développés ou utilisés par la communauté Biogenouest ont été intégrés. De plus, la mise à jour des banques de données sont gérées via BioMAJ. L'instance se veut non spécialisée dans un domaine d'application en particulier. L'équipe bénéficie de l'expertise de plusieurs collègues sur Mobyle, dont principalement Olivier Sallou.

Actuellement, l'utilisation d'un toolshed privé est en test. L'objectif est d'intégrer de nombreux outils à ce toolshed et posséder un certain recul avant de proposer la mise en place d'un « toolshed du Grand Ouest ».

L'instance compte environ 60 utilisateurs à ce jour.

3 formations ont été dispensées autour de l'instance de la plateforme GenOuest :

- 04/12/12 : Formation test avec le groupe Symbiose (2h)
- 08/01/13 : Formation pour les utilisateurs de GenOuest (1 journée)
- 15/02/13 : Formation pour les utilisateurs de l'OSUR (1 journée)

Les retours sur ces formations ont été positifs et de nouvelles formations devraient être proposées.

Une 3ème instance est en cours de création, elle contiendra en particuliers l'outil « toolfactory » qui permet de créer et tester des outils à la volée. Cela permettra à des utilisateurs avertis de produire des outils avant de les proposer à l'équipe afin de les intégrer à l'instance GenOuest. Cette nouvelle instance est quasiment prête et devrait voir le jour prochainement. Une réflexion est également menée sur la possibilité de mettre les outils développés à disposition de tous par simple téléchargement (fichier xml et éventuel script appelé).

La plateforme ABiMS, Station Biologique de Roscoff

L'équipe de Roscoff dispose de 3 instances : une de formation, une de développement et une de production.

L'instance de production regroupe 4-5 workflows et quelques outils « maison ».

L'objectif est de monter une formation à la rentrée.

L'équipe sera présente à Oslo en juin pour le grand meeting Galaxy où elle présentera un poster et elle participe au groupe de travail IFB Galaxy France.

Le Pôle de calcul intensif pour la mer (PCIM),
U.B.O./Ifremer/IUEM/SHOM/IRD/ENSTA, Plouzané

Le projet est géré par 2 personnes : Laure Quintric et Grégory Carrier. Entre avril et juin 2013, un stagiaire, Cédric Mendoza, aide au développement.

3 instances sont en interaction :

- instance de l'Ifremer pour l'utilisation de pipeline NGS et des bases de données nationales
- instance de PBA (Physiologie & Biotechnologies des Algues) pour les outils bioinfo du quotidien, des outils spécifiques de l'étude des micro-algues et l'accès aux bases de données du laboratoire
- une troisième instance pour l'utilisation de pipeline NGS

L'instance de Brest fonctionne sur un serveur web intranet et avec une base de données PostgreSQL. Les jobs sont envoyés sur le calculateur et gérés par PBSpro. L'accès se fait via le LDAP.

Divers outils ont été intégrés notamment Velvet pour l'assemblage, des outils de nettoyage/qualité NGS et Qiime pour la métagénomique. Des workflows ont également été développés. Plusieurs projets ont été réalisés à l'aide de Galaxy en assemblage, annotation, etc. Dans la plupart des cas, il y avait utilisation d'un outil intégré à Galaxy puis communication avec une base de données et production de données au format HTML. Des projets et des collaborations sont en cours ou à venir.

Laure Quintric assistera à la conférence à Oslo.

INRA BIPAA, Rennes

Une instance a été créée (clone à partir de l'instance de la plateforme GenOuest) par Anthony Bretaudeau et Fabrice Legeai dans le cadre de projets INRA.

L'instance est en production depuis janvier 2013, elle fonctionne par un accès LDAP et est ouverte à une quinzaine de personnes. Elle dispose d'accès restreints à quelques libraires, d'outils spécifiques et d'un quota de base plus élevé que l'instance de GenOuest.

Les objectifs sont de fournir aux utilisateurs une autonomie pour l'analyse des données, de produire quelques workflows standards et d'avoir un lien avec la gestion de méta-données. Une forte demande de formation a été ressentie.

Plateforme BIRD, Université de Nantes

L'instance est encore en cours de développement, Audrey Bihouée et Edouard Hirchaud s'en occupent. Elle tourne actuellement sur un server interne avec une base de données PostgreSQL et un serveur FTP (demande d'expertise auprès de la plateforme GenOuest qui a déjà réalisé cette opération).

Les outils intégrés seront principalement MadTools, des outils d'analyse de données de puce (en cours) et de données NGS (fin 2013). Cette instance sera a priori très axée sur l'analyse des données de puces.

Il est prévu d'envoyer les jobs sur un cluster SGE quand il sera installé et de passer par le LDAP pour l'identification.

Étant donné que les autres instances ont un peu d'avance sur le plateforme BIRD, ils n'hésiteront pas à demander des conseils pour bénéficier de l'expérience sur les autres instances.

Retours des utilisateurs

Utilisation de Galaxy

Certains utilisateurs largement avertis sur un outil peuvent être déçu par l'implémentation « simpliste » faite dans Galaxy. En effet, tous les paramètres disponibles en ligne de commande ne le sont pas toujours via Galaxy. De plus, il peut être regretté de ne pas pouvoir suivre réellement le déroulement/état d'avancement d'un job (quand il tourne, en jaune dans le panneau de droite). Cependant, cela reste très pratique pour la partage de données, la collaboration et l'amélioration du lien entre les communautés bio et info qui est aujourd'hui essentiel. L'aspect workflow simplifié représente également une bonne raison de passer à Galaxy.

Dans un futur proche, il n'est pas impossible que des enchainements d'outils soient proposés automatiquement (avec un apprentissage des enchainements régulièrement réalisés par les utilisateurs).

Infrastructure d’accueil

Des inquiétudes ont été soulevées quant au nombre grandissant d’utilisateurs et à la masse grandissante de données à stocker et à traiter. Les différentes instances seront-elles capable de faire face à la demande ?

Sauvegarde des données

Le problème de la sauvegarde des données a été évoqué. Dans le cadre de la plateforme GenOuest, les données peuvent être sauvegardées sur le home des utilisateurs sur le cluster mais pas sur Galaxy car ce n’est pas un espace de stockage (et il n’est pas souhaitable de dupliquer les données). Pour les utilisateurs, il est important de pouvoir avoir accès aux analyses faites dans le passé.

Maintien des outils/workflows

Le maintien des outils pose un réel problème (de la même manière que sur le cluster) car il n’est pas souhaitable de voir de trop de version d’un même outil sur Galaxy. Cela pose également un problème pour les workflows car dès qu’un outil subit une mise à jour, l’ensemble des workflows l’utilisant sont sucseptible de ne plus fonctionner correctement. Le fait de figer la version d’un outil est également problématique car s’il est utilisé dans plusieurs workflows, les auteurs et utilisateurs des différents workflows peuvent ne pas souhaiter figer la même version.

Paramétrage des outils dans les workflows

Certains paramètres d’outils ne peuvent être paramétrés dans le cadre de workflows. A priori, certaines balisent xml dans les wrappers ne permettent pas d’accéder à l’option « set at runtime » dans le design des workflows. Ce serait le cas notamment pour la balise « conditionnal ». Problème avec les conditions pour les workflows.

Développement, maintenance et perspectives

Un communiqué sera prochainement proposé à la communauté GUGGO pour diffusion sur la liste de diffusion Galaxy France. Il serait intéressant d’envoyer un communiqué global en collaboration avec les autres instances de Galaxy Grand Ouest pour faire une annonce générale de l’existence du GUGGO et de ses instances associées.

Formations : retours et perspective

La possibilité de monter des formations en commun a été évoquée. Cependant, cela implique une réflexion commune en amont afin d’avoir les mêmes outils et les mêmes jeux de données en partage.

Réflexion sur un toolshed commun

La plateforme GenOuest est actuellement en train de faire des tests sur le

développement d’un toolshed privé. Elle propose, à terme, d’héberger le toolshed du Grand Ouest. Cyril Falentin a également proposé de se pencher un peu sur la question et notamment de tenter d’intégrer 2 packages R dans un toolshed commun.

Ce toolshed commun pourrait tout d’abord permettre d’y répertorier tous les outils présents sur les différentes instances du Grand Ouest. Ainsi, les utilisateurs pourraient devenir contributeurs en téléchargeant les fichiers concernant un outil (.xml, .py ou autre) et en le modifiant s’il trouve par exemple que les options proposées ne lui correspondent pas.

Parallèlement, ce toolshed permettraient aux contributeurs utilisant l’instance galaxy-contrib d’uploader et donc partager leurs nouveaux outils développés à partir de simples scripts (Python, Perl, R ou autre).

Enfin, ce dépôt faciliterait l’échange d’outils entre instances distantes, et l’ajout d’outils provenant de contributeurs variés.

Réflexion sur les visualisations

Alexan Andrieux, qui développe actuellement un outil de visualisation de réseaux (notamment pour l’assemblage), a évoqué la possibilité d’intégrer son outil de manière plus ou moins interne à Galaxy.

Pour le reste, Trackster semble être un bon visualisateur associé à Galaxy. Mais de nombreuses choses restent encore à faire dans ce domaine.

Il existe une demande de la part des utilisateurs pour avoir plus de possibilités de visualisation dans Galaxy.