Assemblée Générale « orientée biologistes »
16 avril 2013 - Paris
(réunion préparatoire le 15 avril à Paris)

Membres du CATI présents les deux jours: Corinne Rancurel, Martine Da Rocha, Emeline Deleury (Sophia), Alexandre Dehne-Garcia, Bernhard Gschloessl, Franck Dorkeld (Montpellier), Fabrice Legeai, Anthony Bretaudeau (Rennes), Joseph Tran, Adeline Simon (Versailles), Arnaud Mounier (Dijon), Patrice Baa-Puyoulet (Lyon) (Sébastien Carrere), Emmanuel Courcelle, Ludovic Legrand, Erika Sallet, Ludovic Cottret, Jérôme Gouzy (Toulouse).

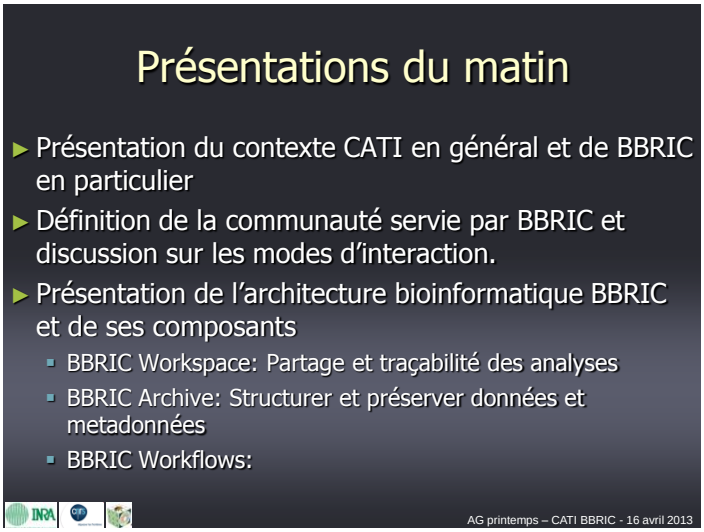
Participants à la réunion du 16 : idem + Thierry Candresse (Responsable Scientifique du CATI, Bordeaux), Sébastien Aubourg (représentant de l'autorité du CATI, représentant BAP cellule bioinformatique, Evry), Hubert Charles, Yvan Rabbé (Lyon), Matthieu Barret (Angers), Stéphanie Daval (Rennes), Carole Kerdelhue (Montpellier), Isabelle Fudal, Muriel Viaud, Thierry Marcel, Fabienne Granier (Versailles), Barbara Pivato, Sébastien Terrat (Dijon)

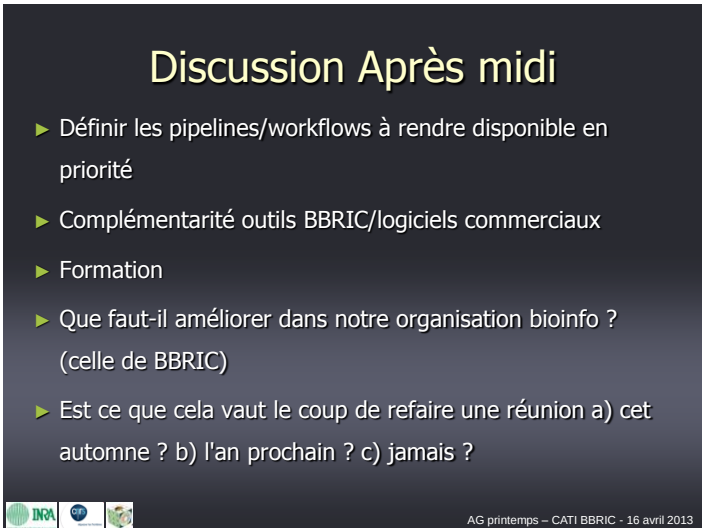
Communication/inscriptions : Arnaud Ridel (dpt SPE)
Annonce : http://cati-bbric.toulouse.inra.fr/lib/exe/fetch.php/bbric-ag-communaute-20130416.pdf
Liste de diffusion : communaute-bbric@liste.inra.fr
Prise de notes : Emmanuel Courcelle, Erika Sallet, Ludovic Cottret

Rédacteur : Jérôme Gouzy
Type du document : hybride, entre compte rendu et relevé de conclusions/perspectives. Ce document reprend quelques éléments de contexte, les présentations sont disponibles sur https://listes.inra.fr/sympa/info/communaute-bbric .

Objectif de la journée : Que les biologistes et bioinformaticiens aient une meilleure connaissance de leurs préoccupations respectives, Interagir et définir ensemble les modalités pour le faire.

Ordre du jour





Définition de la communauté servie par BBRIC et discussion sur les interactions

Le CATI BBRIC a été homologué en 2012 avec des membres SPE et BV et la communauté servie regroupait donc les biologistes de ces deux départements. Suite à la création de BAP par fusion de GAP avec BV dont plusieurs ingénieurs étaient membres de BBRIC, il semble important de reformuler la définition de notre communauté servie en tenant compte de l'existence des 3 CATI bioinformatiques du BAP (chacun doit accomplir ses missions dans son périmètre). La définition de notre communauté devient donc

« Chercheurs/ingénieurs biologistes issus des laboratoires sous tutelle principale SPE et chercheurs/ingénieurs d'autres unités avec un agent affilié au CATI. »

Discussion autour du mode Appel d'offre vs Arbitrage au fil de l'eau

Un mode de fonctionnement sous forme d'appel d'offre annuel présente l'avantage de proposer aux bioinformaticiens une photo du spectre des demandes et doit éviter l'effet de bord « premier arrivé, premier servi ». Ce mode peut faciliter l'arbitrage des tutelles. Il offre également une protection efficace contre la surcharge de projets au prix d'un coût important si l'on souhaite faire les choses dans une vraie transparence. Le mode de fonctionnement par arbitrage au fil de l'eau permet d'établir un vrai dialogue bio/bioinfo et de coller au plus près aux appels à projets fournisseurs de moyens pour la production et/ou l'analyse de données (ANR, département, France Génomique) dont les deadlines s'étalent sur toute l'année. Dans notre CATI, le mode le plus réactif « au fil de l'eau » a été favorisé jusqu'alors et les avis exprimés nous confortent dans ce choix. Nous avons discuté mais n'avons pas retenu la mise en place d'un outil informatique de collecte, le contact direct ou par email est suffisant à la condition que la demande du biologiste et la réponse du bioinformaticien soient formalisées.

Questionnements des biologistes

- Attentes de clarification de l'offre bioinformatique et comment identifier le bon interlocuteur ?

Nous pouvons espérer que la présentation de ce que sont les CATIs et de leur inventaire ait un peu contribué à clarifier la question mais une meilleure communication institutionnelle sur les communautés servies par les différents CATI pourrait combler ce manque.

Il existe des listes destinées aux bioinformaticiens mais elles sont relativement peu actives, on pourrait les ouvrir aux biologistes en espérant qu'ils puissent contribuer à leur animation en posant des questions (*NDLR : pas sur que cela soit plus efficace que les communautés plus larges type seqanswers*) et les utiliser pour mieux identifier les bons interlocuteurs. Pour les biologistes « avertis » (bioanalystes), les PEPI peuvent également contribuer à l'acquisition de connaissances. *NDLR : En outre le site de la cellule bioinformatique va voir le jour très bientôt et la lettre de la bioinformatique INRA propose également un aperçu régulier de la bioinformatique institutionnelle.*

Pour ce qui relève du périmètre de BBRIC, nous avons organisé cette journée à cet effet et avons créé la liste communaute-bbric@listes.inra.fr pour maintenir le contact établi. Lorsque le besoin s'en fait sentir nous proposons que le biologiste contacte dans un premier temps le membre du CATI le plus proche (*cf* http://cati-bbric.toulouse.inra.fr/ pour la liste des membres) ou à défaut contacte directement le responsable du CATI.

- Questionnement par rapport aux relations inter CATI.

L'animation inter CATI est une mission de la cellule bioinformatique de l'INRA, ce n'est pas à BBRIC de traiter cet aspect, par contre BBRIC (et son ancêtre BIPAS) n'ont pas des frontières étanches, nous sommes ouverts aux membres des autres CATI qui travaillent avec nous aussi bien à travers nos échanges (réunions, documents, listes) que dans le cadre de nos projets (merci à Sébastien Aubourg, responsable du CATI de l'URGV pour l'avoir souligné et illustré a travers le projet BIOS).

- Quels sont les critères de décisions utilisés pour la sélection des projets ?

  Dans BBRIC il y a plusieurs plateformes/plateaux bioinformatiques avec une organisation formelle différente mais les facteurs positifs en communs sont les suivants :

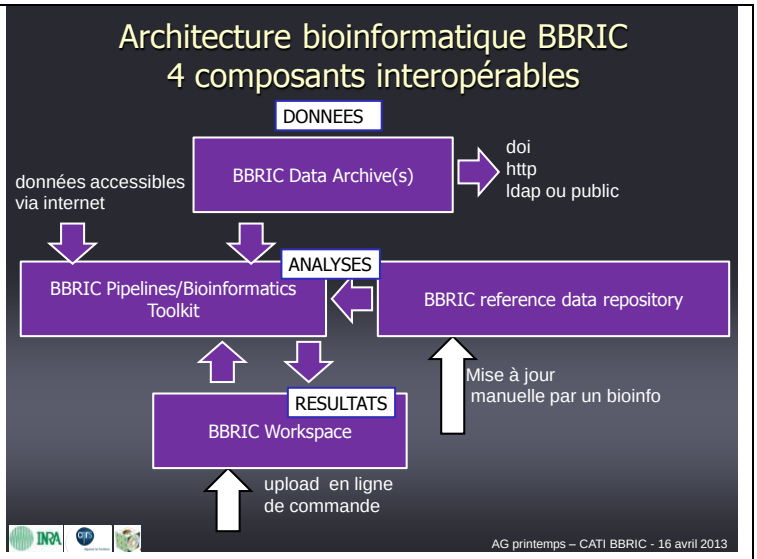
  - Dialogue précoce au moment de la conception du projet pour :
    - Calibrer au mieux la production de données en tenant compte de l'expérience des bioinformaticiens

        - Prévoir éventuellement des ressources complémentaires sur les gros projets
    - Envisager le projet sous la forme d'une collaboration avec la reconnaissance qu'impliquent l'expertise amenée et le travail réalisé
    - Valoriser le précédent projet avant d'en apporter un nouveau pour éviter l'effet de fuite en avant.
        - Un projet qui apporte une nouvelle classe de problème peut être intéressant si le type de problème est amené à se généraliser car cela permet d'anticiper l'acquisition de compétences avant le « rush »
        - Inversement, un projet qui correspond à un problème « classique » ne devrait pas être refusé car l'effort d'automatisation doit avoir été fait.
        - Il faut accepter que le projet candidat soit planifié en fonction des autres projets en cours ou éventuellement en fonction d'autres projets qui pourraient donner les « clefs » techniques ou méthodologiques pour le projet candidat.

Présentation de l'architecture bioinformatique BBRIC et de ses composants

L'architecture bioinformatique répartie que nous sommes en train de développer a pour objectifs :
1) de structurer et préserver les données et les métadonnées produites par notre communauté servie ➔ Composant **Archive**
2) de disposer d'un environnement permettant l'échange, le partage, la traçabilité, le reporting des analyses effectuées dans le cadre d'une activité de service ➔ Composant **Workspace**
3) de proposer un environnement d'analyse intégrant des pipelines d'analyses et les données de références nécessaires
➔ Composant **Pipelines/Workflows**
4) de disposer d'un environnement de gestion et de versionning des données de références (ex : genomes, annotation de génomes, etc.) intégré à l'environnement d'analyse et partageable entre sites ➔ Composant **Reference repository**

L'ensemble des développements proposent des accès web pour les utilisateurs finaux et programmatiques pour les bioinformaticiens.



Les composants **Workspace** et **Archive** sont en production depuis février 2013 à Toulouse (LIPM). L'Archive accueille déjà 2.5To de données (7To en attente d'import) issues de projets impliquant le département SPE et sera prochainement installée sur deux autres sites (Montpellier et Rennes) afin d'étendre la capacité tout en assurant la réplication des données. Une deuxième instance de **Workspace** sera installée à Sophia dans l'année. Une présentation de **Workspace** et une présentation/démo de **Archive** ont été effectuées lors de la réunion. Concernant le composant « **Pipelines/workflows** », nous avons entrepris de collecter et de formaliser nos protocoles d'analyses dans un document que nous appelons les « BBRIC protocols ». Les développements de l'application viennent de débuter et seule la maquette de l'interface utilisateur a été présentée lors de la réunion. L'objectif est de présenter la version opérationnelle lors de la prochaine assemblée générale « orientée biologistes ». Il est à noter que pour les données produites dans le cadre de notre communauté, leur chargement dans **Archive** sera un pré requis pour l'utilisation des workflows d'analyse.

Les premiers composants sont désormais en production, dont le composant critique **Archive**. Nous avons beaucoup investi pour prouver que cela fonctionne mais ne pouvons pas assurer un tel service pendant très longtemps à l'échelle de notre communauté sur des moyens propres de quelques laboratoires. Nous allons donc monter une demande de financement afin d'assurer la pérennité de notre solution (que les biologistes qui liront ces lignes n'hésitent pas à nous soutenir si l'occasion se présente, merci d'avance).

Discussions

Définir les pipelines/workflows d'analyse à rendre disponible en priorité

La plupart des participants sont venus pour chercher de l'information et n'ont pas exprimé d'attentes particulières. Une demande a émergé par rapport à l'analyse de résultats d'assemblages, car il est très difficile d'évaluer la qualité du résultat souvent présenté sous la forme d'un fichier multifasta accompagné de métriques qui ne donnent pas une idée suffisamment précise de la qualité. **Il y a donc une demande sur la mise à disposition d'interfaces graphiques pour proposer aux biologistes une meilleure compréhension des résultats via leur visualisation**. Le constat a également été fait que la problématique de visualisation est critique pour la plupart des problématiques bioinformatiques. Nous en sommes bien conscients et avons plusieurs projets dans ce domaine mais il faut aussi être lucide : le développement d'interfaces permettant une analyse multi niveaux (du détail à la vision globale et intégrée) demande beaucoup de ressources en développement. Nous sommes souvent obligés de nous appuyer sur des outils existants qu'il faut intégrer sachant qu'ils ne couvrent jamais le spectre complet des besoins et qu'ils peuvent contenir des bugs problématiques. Une autre discussion intéressante a eu lieu sur l'analyse de données métagénomique par amplification. La vision naïve est que c'est un problème « facile », des protocoles/outils existent et sont suffisamment simples et documentés pour que tout un chacun puisse avoir un résultat très facilement. La vision experte étant que lorsque l'on évalue ces mêmes protocoles on peut mettre en évidences certains biais/problèmes. En conclusion, l**es bioinformaticiens que nous sommes devons travailler plus encore sur la définition de critères de qualité de nos analyses et des analyses qui sont faites au sein de la communauté.**

Complémentarité outils BBRIC/logiciels commerciaux

Nous rencontrons de plus en plus de cas d'utilisation du logiciel commercial CLC dans les publications mais aussi dans les laboratoires INRA. Nous avons donc trouvé important d'échanger sur ce sujet afin de clarifier notre action et nos priorités. Grâce aux retours des utilisateurs de CLC présents nous pouvons faire une première analyse:

- La version complète semble très satisfaisante sur un nombre important d'analyses NGS classiques
- On peut utiliser CLC sans formation, la documentation est bien faite et l'interface très ergonomique
- L'utilisation de CLC permet au biologiste, lorsqu'il se retrouve bloqué, de discuter avec les bioinformaticiens à un niveau de compréhension des problèmes bien supérieur à celui qu'il peut avoir s'il n'a pas pu travailler/voir lui-même ses données.
- La version NGS complète coûte 3500€ par utilisateur, la version la plus simple 70€ et elle propose déjà de nombreuses fonctionnalités. Il n'est pas impossible que dans le futur, les biologistes incluent le coût de la licence dans leur demande de financement (le prix d'une licence correspond environ au coût d'une lane illumina)
- L'utilisation massive de CLC en mode anarchique risque de poser des problèmes majeurs de sauvegarde des postes de travail.

Notre CATI ne doit par sous estimer l'émergence de CLC dans les laboratoires et il faut que nous trouvions le meilleur niveau de complémentarité entre nos développements et ce que les biologistes peuvent trouver dans le commerce. Cela semble à la fois possible et souhaitable car si nous n'aurons jamais les moyens de développer des interfaces utilisateurs aussi poussées pour faire des choses basiques, nous pouvons développer grâce aux logiciels académiques (galaxy, mobyle) ou à travers des sites web dédiés, des

workflows d'analyses complémentaires de CLC car plus sophistiqués ou plus proches de l'activité de recherche et de l'état de l'art en bioinformatique. Dans l'année qui vient, nous évaluerons la suite logicielle CLC pour avoir un avis argumenté si des chercheurs de la communauté nous questionnent mais aussi pour mieux identifier les manques et les complémentarités.

Formation

Nous avons abordé la question de savoir à quel niveau il est préférable de cibler les formations utilisateurs.

- « Bas niveau » : formations à unix et à l'utilisation de programmes en ligne de commande,
- « Medium » : formations à des outils de workflows pour que les utilisateurs construisent leurs propres workflows d'analyse à partir de briques élémentaires
- « Haut niveau » : formations à l'utilisation de workflows « intelligents » avec interface web. Formations thématiques pour donner les clés aux biologistes pour mieux comprendre et interpréter les résultats fournis par le bioinformaticien ou par un pipeline d'analyse (ex : annotation de génomes)

Suite à la discussion, la priorité sera donnée à l'organisation de formations de « haut niveau » avec un focus sur l'interprétation des résultats et une mise à disposition des détails de la complexité du pipeline seulement pour les utilisateurs les plus curieux.

Pour répondre aux besoins de biologistes familiers avec la ligne de commande unix, nous pensons que continuer à les accueillir dans nos locaux pendant une ou plusieurs semaines est le mode de formation le plus efficace pour former à la résolution d'un problème particulier. Cela a un coût important pour nous mais cela semble efficace et permet de consolider une collaboration car si le biologiste progresse en bioinformatique, le bioinformaticien progresse dans sa compréhension de la question biologique.

Est-ce que cela vaut le coup de refaire une réunion a) cet automne ? b) l'an prochain ? c) jamais ?

La réponse des participants est oui, une fréquence annuelle semble un bon compromis. Les prochaines journées seront thématiques, centrées sur une problématique particulière de traitement de données (analyse de l'expression, du polymorphisme, etc.), sur une problématique générale (comment évaluer la qualité des résultats, etc.), ou couplées à une formation à l'utilisation de l'architecture bioinformatique de BBRIC.

La question de l'ouverture au-delà de BBRIC s'est posée. Bien sûr nos journées « orientées biologistes » seront ouvertes à ceux qui travaillent avec nous, peu importe leur CATI/PEPI/Institut d'origine, mais cela sortirait des prérogatives de BBRIC de prétendre animer au-delà (cf cellule bioinfo, PEPI).