Appel à projets Jeunes Chercheurs et Animations Scientifiques (2014)

Institut de biologie computationnelle

Projet Jeunes Chercheurs X **Animation Scientifique** ☐

ProvRecFlow: Réconciliation entre scripts et workflows par la provenance

Porteur(s):

Cohen-Boulakia Sarah Mdc INRIA (VirtualPlants et Zénith), sarah.cohen_boulakia@inria.fr

Pradal Christophe ingénieur-chercheur CIRAD

Participants	
Christian Fournier	INRA
Pierre Fernique, Yann Guédon, Christophe Godin	INRIA VirtualPlants
Patrick Valduriez	INRIA Zénith

Mots clés: Analyses bioinformatiques ; Provenance ; workflows scientifiques ; scripts ; notebooks ; OpenAlea

Résumé (12 lignes max)

Acquérir de nouvelles connaissances en biologie passe par le développement d'analyses de données complexes. Réaliser ces analyses peut se faire par le développement de scripts, de workflows scientifiques et plus récemment de Notebooks. Chaque approche a ses atouts et ses limites et regroupe une communauté spécifique d'utilisateurs et de chercheurs. L'objectif de ce projet est de construire des ponts forts et transparents pour l'utilisateur entre ces différentes formes d'une même analyse en exploitant la provenance comme moyen d'établir ces ponts. La conception de transformations *provenance-équivalentes* entre scripts, workflows et Notebooks permettrons de tirer parti des nombreux travaux de recherche conçus dans chaque domaine indépendamment des autres. Ce projet se place à l'interface du WP4 et 5 de l'IBC et prendra comme cas d'utilisation les analyses phénotypiques de plantes faites dans le cadre de la plateforme INRA Phenome.

Budget Prévisionnel : Nature et montant des moyens demandés (pas de salaires)

- 2 stagiaires (10 à 12 mois gratification)~€5000; Sujet 1 : Transformations *Provenance-equivalentes* entre le NoteBook IPython et le système de workflow Open Alea; Sujet 2 : Optimisation conjointe de scripts et de workflows.
- Missions : 2 missions Galaxy Community Conference (UK) : € 2000
  Frais de publication (conférences et/ou coût open access) :€ 3 000

COMPLEMENTS (3 pages maximum)

1) Contexte et projet scientifique

Les données biologiques sont particulièrement nombreuses, fortement hétérogènes et distribuées dans des sources très diverses. Acquérir de nouvelles connaissances passe nécessairement par l'analyse de données et la conception de modèles qui peuvent nécessiter d'exploiter ces masses de données disponibles (*big data*). Deux types de possibilités existent alors. La première, et la plus utilisée, est l'utilisation de **scripts**, *ie* de code (Java, Python...) pour implémenter une analyse. Néanmoins, échanger et partager cette analyse, maintenir ou faire évoluer une portion du code sont des tâches reconnues pour être particulièrement délicates. Les **workflows scientifiques** [CL11] se sont alors développés pour répondre à ces problèmes et offrent une seconde façon d'analyser les données. Les étapes de l'analyse y sont représentées par des modules qui encapsulent les tâches les plus fréquentes. Les workflows sont définis de façon visuelle (*visual programming*) en enchaînant des modules préexistant. C'est le système de gestion de workflows qui contrôle tous les aspects relatifs aux exécutions, depuis l'éventuelle parallélisation (ordonnancement) jusqu'au fait de garder la trace exacte des outils, paramètres et données utilisées et générées (provenance). La provenance joue un rôle particulièrement important en bioinformatique [DF09], permettant de reproduire une analyse mais aussi mieux comprendre les résultats obtenus.

Pour chacun de ces paradigmes (script ou workflow), on retrouve une communauté dédiée (*Bioperl* et *Biopython* versus *Galaxy*, *Taverna* ou *Open Alea*). Les problématiques de modularité (du code ou du workflow) et d'optimisation y sont présentes avec des terminologies similaires : *(anti-)patterns*, *refactoring* (cf. [CFC12, CCM+14])... Mais, alors que des techniques fondées sur la compilation sont utilisées pour les scripts, ce sont des techniques plus algorithmiques (graphes) qui sont déployées pour les workflows avec la considération de préférences de l'utilisateur. Très récemment, une troisième solution, intermédiaire, qui rencontre déjà un vif succès a vu le jour : les Notebooks interactifs [Sh14]. Ces environnements sont disponibles sous forme d'applications web et se composent de pages avec des cellules contenant le script de l'analyse et les données produites. L'utilisateur peut visualiser les résultats obtenus, modifier les cellules... Néanmoins, comme pour les scripts, aucun mécanisme de traçabilité n'est proposé : il est impossible de garder une provenance fine des variantes de cellules testées et des données ainsi générées.

L'objectif de ce projet est construire un pont fort entre les paradigmes scripts et workflow, permettant à l'utilisateur de passer de l'un à l'autre, de façon transparente. La provenance de l'analyse sera le moyen d'établir ce pont, que l'analyse soit représentée sous la forme d'un script, d'un workflow ou d'un Notebook.

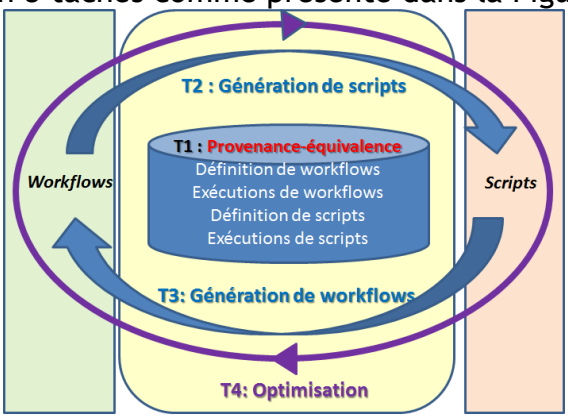
Nous travaillerons dans le cadre d'analyses phénotypiques de plantes développées dans la plateforme INRA Phenome[1]. Des workflows OpenAlea [PDB+08] et de scripts Python y ont été développés pour reconstruire l'évolution de la structure et de la forme des plantes à partir d'images (2D+t). En particulier, notre preuve de concept prendra appui sur les travaux récents de Virtual Plants en phénotypage dans lesquels 100 000 cellules ont été extraites d'épidermes de feuilles, 20 000 images de feuilles utilisées dans le cadre de scripts pour estimer 2000 cinétiques et agréger l'ensemble des cellules et une partie des cinétiques ainsi que des mesures morphologiques dans un seul modèle.

Nous illustrerons aussi notre approche à travers le Notebook IPython (équipe de Berkeley avec laquelle l'EPI VirtualPlants est en contact).

[1] https://www.phenome-fppn.fr/Plates-formes/Montpellier-Controle

2) Mise en œuvre du projet et objectifs

Ce projet se décompose en 5 tâches comme présenté dans la Figure 1.



T1 **:** Définition de la *provenance-équivalence* entre un workflow et un script (possiblement dans une cellule de NoteBook) relatifs à la même analyse. Cette tâche inclut la conception d'une base noSQL de provenance contenant les variantes de workflows, scripts (Notebook) et leurs exécutions.

Positionnement : Les standards de provenance (D-Prov, Prov-DM…) seront suivis dans ce contexte. Le concept de *provenance-équivalence* introduit dans [CFC12] sera généralisé au cas original script-workflow.

T2**:** Développement d'un algorithme *provenance-equivalent* de transformation (compilation) d'un workflow vers un script. Il sera implementé pour transformer tout workflow d'OpenAlea en un Notebook IPython.

Positionnement **:** Taverna [WG14] et VisTrails [MBC+14] ont récemment proposés des prototypes pour exporter leurs workflows en IPython. Néanmoins ces projets se limitent à une transformation unidirectionnelle sans considérer l'évolution de l'analyse.

T3: Conception de l'inférence *provenance-équivalente* d'un script à partir d'un Notebook (ie. toute exécution du Notebook aura même provenance que celle du workflow inféré). Chaque cellule du Notebook sera convertie en un module après une analyse syntaxique inférant les variables d'entrées/sorties. Un workflow sera calculé à partir des relations entre les variables des différentes cellules du Notebook.

Positionnement **:** Galaxy propose d'exécuter un notebook IPython comme un élément du workflow[2] et d'intégrer les résultats obtenus dans son historique. Néanmoins, aucune fonctionnalité d'inférence de workflow *équivalent* à un Notebook n'est offerte.

T4: Exploitation du passage script-workflow pour s'adapter au paradigme de préférence de l'utilisateur. On utilisera des techniques de refactoring, simplification/factorisation pour rendre les workflows et scripts plus lisibles, modulaires et échangeables... Ces opérations seront *provenance-équivalentes*.

T5: Extension des résultats à des scripts plus généraux. Il conviendra de déterminer les contraintes (fonctions, objets, variables globales) que devront respecter les scripts qui pourront automatiquement être réécrits sous la forme de workflows[3].

Positionnement : Les approches transformant un workflow en script ou un script en workflow cités en T3-4 ne considèrent que de scripts python fonctionnels. Notre originalité est d'exploiter les aspects de flot de contrôle et orienté objet d'OpenAlea.

Les objectifs de notre projet sont les suivants :

- Réconciliation des formes d'une même analyse (workflow, script, ou notebook);

[2] https://www.youtube.com/watch?v=jQDyTuYnn1k

[3] 80% des scripts actuellement collecté sont directement transformables en Open Alea.

- Prise en compte de la provenance des variantes des analyses sous toutes ses formes et des exécutions (reproductibilité et compréhension des résultats);
- Aide à la conception, au débogage et à un partage simplifié et donc plus important d'analyses bioinformatiques.

3) Description des partenaires et de leur complémentarité

Les co-porteurs ont une expertise internationale reconnue dans les workflows scientifiques. SCB est membre des équipes VirtualPlants et Zénith et travaille depuis 8 ans dans le domaine de la provenance et de la transformation de workflows (avec Univ. Manchester/Taverna, UPenn, Berlin...). CP est le principal concepteur du système OpenAlea (logiciel le plus téléchargé sur la gforge de l'INRIA), il collabore activement avec la communauté iPython et travaille sur le projet OpenAleaLab avec son équipe. CF est responsable du projet INRA Alinea. Avec CP, ils développent et utilisent la plateforme pour animer de grands projets dans différents domaines (phénotypage haut-débit (Phenome), impact des pesticides sur l'environnement (ECHAP), ...).

Intégration du projet dans IBC**.** Les cas d'utilisation sont au cœur du WP4 : workflows OpenAlea pour l'analyse d'images (2D ou 3D +t) issues d'une plateforme de phénotypage de plantes. L'objet principal de ce projet, le workflow, est central à WP5. Une partie des résultats pourra s'étendre à Galaxy, utilisé dans la plateforme *SouthGreen*. Notre approche pourra être considérée dans de nouveaux contextes (annotation fonctionnelle, analyse haut-debit ou phylogénie) où des scripts et des workflows existent, renforçant les liens entre les WPs 1-2-3 et WP4-5.

4) Justification du budget demandé

- 2* 5 à 6 mois de gratification (€5000) : Deux stagiaires (M2) (i) développement de transformations *Provenance-equivalentes* entre le NoteBook IPython et le système Open Alea et (ii) conception de méthodes d'optimisation exploitant conjointement la forme script et workflow d'une analyse.
- Publications (€5000): Galaxy Community Conf (3j UK, *2, €2000); conférence informatique (€1500) et frais article prototype revue bioinfo (€3000).

Références

[CCM+14] Cohen-Boulakia, S., Chen, J., Missier, P., Goble, C., Williams, A. R., & Froidevaux, C. (2014). Distilling structure in Taverna scientific workflows: a refactoring approach. *BMC bioinformatics*, *15*(Suppl 1), S12.

[CFC12] Cohen-Boulakia, S., Froidevaux, C., & Chen, J. (2012, October). Scientific workflow rewriting while preserving provenance. In Proc. of *IEEE e-Science, 2012*

[CL11] Cohen-Boulakia, S., & Leser, U. (2011). Search, adapt, and reuse: the future of scientific workflows. *ACM SIGMOD Record, 40*(2), 6-16.

[DF08] Davidson, S. B., & Freire, J. (2008, June). Provenance and scientific workflows: challenges and opportunities. In *Proc of SIGMOD 2008* (pp. 1345-1350).

[MBC+14] Murta, L., Braganholo, V., Chirigati, F., Koop, D., & Freire, J. noWorkflow: Capturing and Anal*yzing Pro*venance of Scripts. In Proc of. IPAW 2014.

[PDB+08] Pradal, C., Dufour-Kowalski, S., Boudon, F., Fournier, C., & Godin, C. (2008). OpenAlea: a visual programming and component-based software platform for plant modelling. *Functional plant biology*, *35*(10), 751-760.

[Sh14] Shen H (2014). Interactive notebooks: Sharing the code. Nature. 515(7525):151-2

[WPG14] Williams A., Pawlik A., Goble, C., Running Taverna Workflows within IPython Notebook (Univ of Manchester, internal report, 2014)