

Titre : A quoi ressemblent les workflows scientifiques ?
What do scientific workflows look like ?

Encadrant(s) : Sarah Cohen-Boulakia et Christine Froidevaux (cohen@lri.fr, chris@lri.fr)

Lieu/Equipe : LRI, équipe Bioinformatique

En collaboration avec : Biologistes et bioinformaticiens de l'INRA.

Des contacts avec P. Missier Université de Manchester et S. Davidson, Université de Pennsylvanie pourront avoir lieu durant le stage.

Pré requis : Une curiosité pour la biologie est attendue. Une bonne connaissance de la théorie des graphes est nécessaire. Une bonne connaissance de l'anglais écrit et parlé est souhaitable.

Description :

Les systèmes de workflow scientifiques sont devenus de plus en plus utilisés pour la gestion des expériences scientifiques où de nombreuses tâches bioinformatiques sont chaînées les unes aux autres. Un grand nombre d'outils disponibles pour manipuler ces workflows (stockage, interrogation, visualisation etc.) existe aujourd'hui. Chaque système représente un workflow sous la forme d'un graphe et de plus en plus de systèmes choisissent de se restreindre à des classes de graphes particulières pour optimiser le traitement/stockage des informations. On retrouve ainsi notamment des modèles très linéaires, d'autres à base de DAG (pas de boucles), ou encore des SP-Graphes (graphes où les nœuds sont en parallèle ou en série uniquement) ou SPFL-Graphes (SP-graphes auxquels on ajoute une forme restreinte de boucle). On retrouve d'autres formes de restriction structurelle comme par exemple le fait d'imposer qu'on ne considère qu'un seul nœud d'entrée (source) et un seul de sortie (puits).

L'objectif général de ce stage est d'étudier les structures des workflows scientifiques existant, à la fois ceux qui sont présents dans de grands entrepôts comme *myExperiments* et des workflows conçus par nos partenaires bioinformaticiens de l'INRA.

Plus précisément, le travail comportera 3 tâches majeures

- a) Dégager des informations statistiques sur les workflows scientifiques collectés : les motifs présents dans leurs structures (des boucles, des branches parallèles, etc.) et la proportion de chacun de ces motifs. L'adéquation des modèles les plus souvent choisis par les systèmes aux workflows existants découlera de cette étape.
- b) Proposer des algorithmes pour transformer un workflow sous une forme de graphe en un workflow *équivalent* (effectuant les mêmes tâches) ayant la structure souhaitée. L'impact de ces changements sur le flot de données devra être évalué.
- c) Etudier la présence de patterns plus complexes (composition de patterns élémentaires) dans les workflows collectés et comparer ces patterns à ceux étudiés dans le contexte des workflows de types business : <http://www.workflowpatterns.com/>