Sujet : Qualité de l'information dans des scénarios d'annotation fonctionnelle

Ce sujet peut être orienté plus Pro ou plus recherche.

Encadrants :
Sarah Cohen-Boulakia (Maître de Conférences), cohen@lri.fr
Christine Froidevaux (Professeur), chris@lri.fr

Lieu du stage :
Equipe Bioinformatique, Laboratoire de Recherche en Informatique, Université Paris-Sud, Orsay, France

Collaboration : annotateurs de l'INRA utilisateurs de la plate-forme AGMIAL, bioinformaticiens de MIG

Contexte :
L'annotation fonctionnelle est la tâche d'assigner à une protéine, une ou plusieurs fonctions. Le travail proposé dans ce stage s'inscrit dans la continuité du projet d'annotation fonctionnelle semi-automatique ACI IMPBio RAFALE, qui a été effectué en collaboration entre le LRI (CNRS et Université Paris Sud, Orsay) et le laboratoire MIG à l'INRA à Jouy-en-Josas. Il s'appuie sur l'annotation d'un pool de protéines réalisée par des annotateurs de l'INRA sur plusieurs génomes bactériens, à l'aide de la plate-forme d'annotation AGMIAL. AGMIAL est une chaîne d'annotation de génomes microbiens développée à MIG, qui a été utilisée pour l'annotation de Lactobacillus sakei, Lactobacillus bulgaricus et Flavobacterium psychrophilum. Plus d'une dizaine de génomes microbiens, actuellement séquencés par des équipes de l'INRA, sont en cours d'annotation à l'aide de cette plateforme.

Travail :
Après interview auprès des annotateurs, un certain nombre de scénarios d'annotation ont été recueillis, et une liste de tâches a été répertoriée. Les différents scénarios ont conduit à l'élaboration de workflows d'annotation. Les étapes du travail demandé sont les suivantes :

a) Construction d'un workflow générique, rassemblant les différentes stratégies des annotateurs
b) Spécification et organisation des différentes tâches liées à l'annotation, par exemple sous la forme d'une ontologie
c) Eventuellement, implémentation et évaluation des workflows sous Taverna
d) Proposition de critères pour évaluer la fiabilité du résultat d'annotation obtenu à la fin de l'exécution du workflow pour une protéine donnée
e) Proposition d'une fonction de qualité qui tienne compte des outils utilisés, des bases de données consultées, et du cheminement dans le workflow (provenance).