Étude de faisabilité pour l'automatisation de recherche de jeux de données en bioinformatique

Fabien Duchateau, Guillaume Launay
Université Claude Bernard Lyon 1
prénom.nom@univ-lyon1.fr

Description générale : Sujet de TER pour L3 BISM, d'une durée de 4 mois.

Objectif : développer un script Python qui permet d'automatiser la recherche de jeux de données cohérents sur les interactions entre protéines.

Contexte : dans le cadre de l'UE *base de données pour la bioinformatique* enseignée en Master BIOINFO, les étudiant.e.s réalisent un projet qui consiste à intégrer différents jeux de données dans une même base de données, par exemple pour produire ou vérifier des connaissances. Ces jeux de données sont hétérogènes (modèles, formats, vocabulaires différents) et leur intégration s'apparente à un *workflow*. Par exemple, en partant du génome d'un virus (format GenBank converti en SQL), il faut intégrer son protéome (format XML), les interactions avec des protéines humaines (format JSON), puis intégrer des informations supplémentaires sur les protéines humaines (format RDF) et enfin sur le génome humain (requêtes REST). Pour concevoir de nouveaux sujets de projet, l'une des difficultés est la recherche des différents jeux de données pour l'ensemble du *workflow* (e.g., on peut trouver un génome et son protéome, mais aucune interaction, et donc le workflow est incomplet).

Fonctionnalités attendues (liste non exhaustive) :

- Étude de faisabilité (quelles parties sont automatisables, nouvelles sources).
- Détection et affichage de workflows. Les tentatives, quel que soit le résultat (workflow complet, partiel, etc.), seront stockées (fichier JSON ou base de données). Si assez de temps, une interface web pourra permettre de visualiser les workflows testés et de les modifier, notamment après une vérification manuelle.
- Implémentation de script(s) de conversion (e.g., format GenBank vers SQL ou JSON).

Langages et outils :

- Programmation en Python.
- Utilisation de git pour le suivi.

Contact : pour toute question ou pour candidater sur le sujet, nous envoyer un message.