

Automatic genomics data processing for database implementation

Su MELSER, Charlotte HÉRICÉ

Elysia Bioscience, Technopole Bordeaux Montesquieu, Martillac

su.melser@elysia-bioscience.com, charlotte.herice@elysia-bioscience.com

Sujet :

Les technologies de séquençage de l'ADN et de l'ARN avancées ont permis aux scientifiques de récolter un vaste ensemble de données. Ces dernières sont étudiées dans le but de mieux comprendre la signification et les conséquences de variations au niveau de l'ADN génomique et des produits de sa transcription. Grâce à ces données, il est possible d'étudier l'expression différentielle des gènes en fonction de différentes conditions expérimentales. Dans les données d'expression différentielle, il est classique de retrouver notamment le code UniProt [1] d'une protéine d'intérêt, ainsi que le ou les codes GO associé(s). Ces codes GO (Gene Ontology [2]) sont issus d'une initiative bio-informatique majeure visant à unifier la représentation des attributs des gènes et des produits génétiques dans toutes les espèces. Ils sont classifiés selon une certaine hiérarchie en fonction des composants cellulaires auxquels ils s'appliquent, des fonctions moléculaires réalisées ou des processus biologiques.

Le but de ce projet est (1) le traitement automatique des données d'expression différentielle en fonction des codes GO pour permettre (2) leur intégration efficace dans une base de données au sein d'Elysia Bioscience. La partie « biologie » du projet sera encadrée par Su MESLER et la partie « bio-informatique » par Charlotte HÉRICÉ.

Langages :

- SQL pour la partie base de données
- Python pour la partie traitement (possibilité de changer suivant les préférences du groupe)

Conditions :

- Autonomie
- Meetings au moins une fois par semaine pour l'avancement du projet
- Confidentialité
- Rédaction du cahier des charges et du rapport en anglais

Bibliographie :

[1] <https://www.uniprot.org>

[2] <http://geneontology.org/docs/ontology-documentation/>