Stage : Construction et interrogation de graphes de connaissances pour l'intégration de données de génomique fonctionnelle

Contexte du projet :

La compréhension de la régulation de l'expression des gènes est un enjeu fondamental avec des conséquences en agronomie et en santé humaine. Cette régulation passe par différents mécanismes souvent médiés par une grande diversité d'acteurs (micro ARNs, ARNs longs non codants, facteurs de transcription, etc). L'essor des techniques de génomique fonctionnelle permet de mieux appréhender ces mécanismes en fournissant des informations sur l'expression des gènes (mRNA-seq, smRNA-seq), l'accessibilité de la chromatine (ATAC-seq) ou la conformation tridimensionnelle de la chromatine (Hi-C).

Pour intégrer et interroger ces données, nous proposons une approche basée sur la construction d'un multi-graphe coloré (graphe de connaissances) utilisant pour noeuds des molécules ou des régions génomiques d'intérêt (gènes, régions régulatrices, etc) et comme arrêtes des propriétés issues de données fonctionnelles (corrélations d'expression, proximités génomiques, etc) ou de connaissances *a priori.* Cette structure de données pourra ensuite être exploitée pour rechercher différents mécanismes de régulation, caractérisés par des motifs à identifier dans le graphe.

Au cours de ce stage, nous proposons de construire des graphes de ce type, et d'évaluer la possibilité d'y rechercher des patrons de régulation, dans un premier temps grâce à des systèmes de bases de données tels que `Neo4j/Cypher`. Nous proposons aussi de travailler d'abord sur les interactions miRNA-mRNA.

Pour développer cette approche, nous travaillerons dans un premier temps sur des données générées dans le cadre du projet européen GENE-SWitCH, qui ont l'avantage de réunir différentes techniques de génomique fonctionnelle appliquées aux mêmes échantillons dans deux espèces différentes (poulet et porc).

Nous souhaitons que la personne recrutée poursuive en thèse (concours de l'école doctorale).

Vous serez plus particulièrement en charge de :

- Vous approprier des données de séquençage haut débit (RNA-seq, smRNA-seq)
- Définir les entités et les relations qui permettront de construire un graphe de connaissances pour intégrer ces données à l'aide de l'outil `Neo4j`
- Utiliser l'outil `Cypher` pour interroger ce graphe dans le cadre de la recherche d'interactions miRNA-mRNA

Environnement d'accueil

Vous serez accueilli·e au sein de l'unité GenPhySE (Génétique, Physiologie, et Systèmes d'élevages), une Unité Mixte de Recherche (UMR) INRAE-Université de Toulouse-ENVT située au sud de Toulouse. GenPhySE regroupe plus de 150 personnes travaillant sur des espèces animales d'élevages (porc, mouton, chèvre, lapin, caille, abeille). Vous serez plus particulièrement intégré·e à l'équipe Dynagen (Dynamique évolutive des génomes). Vous serez encadré·e par un ingénieur de recherche de l'équipe et une chercheuse INSERM.

Le profil que nous recherchons

- Diplôme minimum requis : Master/Ingénieur (Bac+5)
- Formation recommandée : Bioinformatique
- Connaissances souhaitées :
    - Linux et Bash, Python et/ou R
    - Travail sur cluster de calcul
    - Bon niveau d'anglait à l'écrit (lecture d'articles scientifiques)

Conditions d'accueil

- Début du stage : janvier 2024
- Durée : 6 mois
- Rémunération : 4.05 €/heure (35h par semaine)

Pour faire une candidature, merci d'envoyer CV et lettre de motivation à sarah.djebali@inserm.fr et cervin.guyomar@inrae.fr avant le 15 octobre 2023.