SUJET DE STAGE

NOM, prénom de la personne proposant le stage :

Francillonne Nicolas

Johann Confais

Adresse Professionnelle :

INRAE, Centre de recherche de Versailles, bat.18 RD10, Route de Saint Cyr

78026 Versailles France

Adresse électronique :

johann.confais@inrae.fr et nicolas.francillonne@inrae.fr

Entreprise/Unité d'appartenance :

Institut National de recherche pour l'agriculture, l'alimentation et l'environnement INRAE

Unité de Recherche en Génomique-Info (URGI) - UR 1164 INRAE Versailles

Domaine d'expertise de l'entreprise / laboratoire :

Développement d'outils et acquisition de connaissances sur la structure, l'évolution et le fonctionnement du génome.

Titre du stage : Intégration et exploitation de données hétérogènes en lien avec l'adaptation à des conditions environnementales changeantes dans une base de données orientée graphe.

Mots clés :

Génomique, pangénomique, éléments transposables, système d'information, base graphe, traitement de données, pipeline

Description du sujet (1 page maximum)

Contexte : De nombreuses ressources génétiques, génomiques et -omiques sont disponibles chez plusieurs espèces d'importance agronomique. Pour autant il n'est pas aisé de croiser ces données pour identifier les mécanismes régulateurs de gènes et réseaux de gènes d'intérêts. Il existe plusieurs génomes assemblés de novo et issus de différents environnements pour une espèce. Ces nouveaux jeux de données ouvrent de nouvelles perspectives dans le décryptage des mécanismes d'adaptation à différents environnements.

Croiser les connaissances fournies par les séquences des génomes, avec celles fournies par des approches de génétique quantitative, de détection de polymorphismes (CNV), de données d'annotations (ETs, TFBS, facteur de transcription) et de transcriptomique, pourrait permettre de mettre en évidence les déterminants génétiques et moléculaires régulant des caractères d'intérêt.

Il y a actuellement un réel besoin de développement d'outils qui permettent (1) d'interroger et de croiser les données acquises en génétique et en -omiques d'espèces végétales de manière intelligente et efficiente et (2) d'explorer les limites entre synténie structurale et fonctionnelle.

Ces outils pourront servir à l'amélioration variétale qui doit répondre à de nouveaux enjeux comme le réchauffement climatique et la transition agro-écologique. Une base de données centré sur Arabidopsis thaliana et Brachypodium distachyon, 2 espèces modèles respectivement monocotylédone et dicotylédone a déjà été développé dans de précédents stages. La partie intégration du stage s'articulera autour de 2 objectifs, 1°) modéliser et intégrer des données d'expression (RNAseq, Méthylomes) 2°) intégrer des données d'*Oryza sativa* plus proche de *B.distachyon* pour tester des approches de biologie translationnelle autour de traits d'intérêts.

La partie exploitation se portera sur l'utilisation de la base pour identifier des relations entre les éléments transposables et des motifs régulateurs et prédire quel peut être leur impact fonctionnel.

Ce stage s'inscrit dans cette dynamique et plus particulièrement sur le rôle des éléments transposables dans l'adaptation de leur hôte à des conditions environnementales changeantes.

Objectifs : Intégration des données hétérogènes, dans une base de données de type « graphe » (Neo4j).

Ces données générées au laboratoire et issues de bases publiques devront être traitées pour être insérées dans une base pilote sur la thématique de l'adaptation aux conditions environnementales fluctuantes.

Le(a) candidat(e) devra enfin pouvoir proposer une automatisation de l'insertion des données en base et des visualisations permettant une interrogation accessible et reproductible.

Travail demandé :

Le(a) candidat(e) devra analyser des données -omiques disponibles pour notamment extraire des informations de co-localisation entre différentes sources de données.

Il(elle) travaillera à les insérer dans une base graphe et mettre en place un pipeline d'analyse et de formatage pour automatiser cette démarche pour d'autres organismes.

Le(a) stagiaire acquerra des compétences en développement de pipelines et d'annotation des génomes dans un contexte « big data » où plusieurs génomes sont à annoter simultanément

ainsi que sur les systèmes de gestion de données basé sur les graphes.

Compétences techniques recherchées :

• Maîtrise des commandes UNIX (shell) et de la programmation python.

• Connaissance en SGBD souhaitable notamment NoSQL.

• Connaissance de la technologie Docker souhaitable

Ce sujet constitue un premier pas vers un travail de thèse : Non

Date de début du stage et durée estimée du stage :

A partir de début 2024 de 6 mois.

Montant (brut mensuel) de la rémunération proposée :

Indemnité de stage selon barème en vigueur (environ 550 euros net par mois)

Date de la proposition de stage et date limite de candidature :

Proposition de stage fin septembre 2023

Date limite fin décembre 2023