Stage M2 ou 3A Ingénieur en informatique.
Durée : 5 à 6 mois, débutant entre janvier et avril 2026.

Sujet : Transformation de graphes pour réduire la complexité de l'analyse des pangénomes

L'ADN code l'information génétique, et la modification de certains gènes peut avoir des impacts visibles sur les individus qui les portent. Un des grands défis de la biologie est d'établir des liens entre ces variations génétiques d'un individu et l'ensemble de ses caractères apparents, son phénotype. Par exemple, pour deux variétés distinctes d'une plante cultivée, si elles présentent une réponse différente à la sécheresse, il est désirable d'identifier les variations génétiques à l'origine de ce changement de réponse.

Pour étudier ces variations, il est possible de modéliser la diversité des génomes d'une espèce *via* un **graphe de pangénome [1, 2].** Dans ce graphe dirigé, les nœuds sont labellisés par une sous-chaîne de la séquence d'un génome et les liens indiquent leur contiguïté dans au moins un des génomes (figure 1). Le génome d'un individu est alors un chemin dans ce graphe (figure 1, chemins colorés). Ce modèle permet d'extraire et caractériser les variations génétiques *via* la recherche de structures particulières appelées « **bulles** » (une paire de chemins connectant une source commune et une destination commune, mais autrement disjoints).

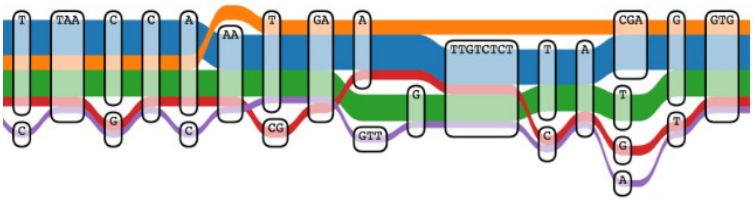

Figure 1. *Graphe de variation : les nœuds représentent des séquences, un chemin représente le génome d'un individu. Si deux chemins passent par le même nœud, alors la séquence est commune entre les individus en question. Les chemins alternatifs (« bulles ») montrent que chaque individu porte un génome propre. En pratique, le graphe peut présenter une topologie plus complexe que ce schéma simplifié.*

L'analyse de ce graphe permet d'identifier des variations génétiques entre génomes connus. Mais il est encore plus utile de comparer de nouveaux individus (porteurs, par exemple, d'une maladie génétique) à ce graphe, afin d'identifier de nouvelles variations fortement associées à une caractéristique intéressante d'un point de vue médical ou agronomique.

Cette recherche est opérée généralement *via* un processus appelé « mapping », qui vise trouver la partie du graphe correspondant à une séquence donnée. Il est complexe car il implique d'identifier un chemin du graphe qui maximise la similarité avec le génome requête (des milliards de caractères à comparer) et le graphe lui-même est de très grande dimension (des centaines de millions de nœuds). Explorer tous les chemins possibles n'est tout simplement pas envisageable. Le mapping sur graphe de pangénome utilise de nombreuses approches d'indexation et d'algorithmique du texte pour accélérer cette tâche, mais le processus reste lourd.

Le stage s'intéressera à explorer une voie alternative pour réduire la complexité de cette analyse : **réduire la recherche à un sous-ensemble de bulles** sur la base d'un **critère établi depuis une métadonnée** supplémentaire en entrée du problème. En effet, les nouveaux génomes sont généralement étiquetés par un état pour un phénotype donné (résistant *vs* non résistant à la sécheresse ; il peut peut être quantitatif ou qualitatif). Sur la base de cette information et en ajoutant l'indexation du graphe, il est possible d'évaluer les bulles et de rechercher celles maximisent la ségrégation de ces états. Le stage a pour objectif de tester différentes approches pour évaluer les bulles d'un graphe de pangénome sur la base de cette métadonnée, et ainsi réduire la recherche à un sous-ensemble de bulles. Il explorera différentes stratégies (empiriques ou probabilistes), le but étant de maintenir dans l'analyse les bulles portant les variations génétiques à l'origine de la variation observée dans les génomes.

Aujourd'hui, une dizaine d'outils de mapping sur graphe de pangénome existent [3] et le stage cherchera à évaluer la nouvelle méthode en terme de sensibilité/sensibilité et de coût computationnel. Le ou la stagiaire pourra s'appuyer sur de grands jeux de données issus de différentes espèces végétales et animales liées à des projets en cours dans l'unité (abricotier, choux, bovins, humain…). Il utilisera des libraires de manipulation de graphe existantes (C++ et python), sur le cluster de calcul de Genotoul-Bioinfo. Au-delà des aspects computationnels, ces travaux sont préliminaires à une thèse d'informatique, qui poursuivra ces développements algorithmiques pour la prédiction de liens entre génomes et phénotypes à de très grandes échelles (bases de données de génomes), en collaboration avec différents partenaires de l'INRIA de Rennes, et des Universités de Bordeaux et Montpellier.

REFERENCES : 1. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9. 2. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. preprint. Bioinformatics; 2023. 3. Andreace F, Lechat P, Dufresne Y, Chikhi R. Construction and representation

of human pangenome graphs. preprint. Bioinformatics; 2023.

Objectifs du stage :

- Acquérir les notions liées au modèle graphe de pangénome.
- Développer un critère de sélection des bulles, en utilisant des librairies existantes.
- Développer un algorithme pour réduire la taille du problème de la recherche de liens génotype-phénotype
- Évaluer l'impact de cette réduction (spécificité/sensibilité, coût CPU/mémoire).
- Implémenter l'algorithme dans un prototype en C++ ou python.

Profil de candidat souhaité :

- Connaissances en théorie des graphes et structures de données avancées (indexation).
- Connaissances en algorithmique du texte non requises, mais bienvenues.
- Une expérience en langage compilé (C, C++, …).
- Intérêt pour les contextes multidisciplinaires et appliqués.
- Autonomie et capacité de travail en équipe, de rédaction, de synthèse.
- Intérêt pour l'international (anglais)

Encadrement :

- Le stage sera encadré par Benjamin Linard (MIAT, INRAE, Toulouse) et Sèverine Bérard (Université de Montpellier), spécialisés dans le développement d'algorithmes (graphes et texte) pour l'analyse des génomes.
- Selon ses performances et son attrait pour la recherche, l'étudiant.e aura l'opportunité de poursuivre ses travaux avec une thèse financée (ED MITT) dans le cadre du projet ANR PanQuest.
- Le ou la stagiaire sera hébergé au sein de l'unité MIAT (Mathématique et Informatique Appliqués de Toulouse), à l'INRAE Occitanie-Toulouse (24, Chemin de Borde Rouge 31320 Auzeville-Tolosane).