Titre du sujet : Comparaison d'assemblages de génomes par l'approche de k-mer

Encadrant.e.s : Anne-Muriel Arigon (Anne-Muriel.Arigon@umontpellier.fr), Sèverine Bérard (Severine.Berard@umontpellier.fr) et Annie Chateau (annie.chateau@lirmm.fr)

Equipe: LIRMM-MAB et ISEM-PEM

Mots clefs : algorithmique, structure de données, k-mers, assemblage de génomes

Résumé :(10 à 20 lignes avec des objectifs clairement définis)

L'assemblage de génome consiste à produire, à partir de données de séquençage (des millions de mots qui se chevauchent), un ensemble de séquences représentant les chromosomes des organismes vivants.
Plusieurs méthodes co-existent, et peuvent produire, sur un même jeu de données, des résultats très différents. Le projet consiste à comparer les assemblages produits par différentes méthodes, sur un même jeu de données.
Une méthode consiste à comparer le contenu des différents assemblages découpés en k-mers, c'est-à-dire en mots de taille k. À partir de la comparaison de ces k-mers, on pourra déterminer des zones qui sont communes à tous ces assemblages pour pouvoir reconstruire un génome "consensus". La comparaison des k-mers passe par une étape d'indexation de ces k-mers dans une structure de données optimisée permettant de connaître le nombre et la position des occurrences des k-mers dans les différents génomes. Une application déjà existante permet d'effectuer cette première étape : RedOak (développé au LIRMM).
L'objectif général de ce projet est donc de concevoir un programme intégrant une première étape d'indexation de k-mers des différents assemblages à comparer , puis d'analyser les résultats de cette 1ere étape afin d'identifier les zones communes et les zones divergentes entre les assemblages à comparer. Ce travail s'appuiera sur un travail préliminaire de L3 CMI 2021-2022.

L'objectif du projet est donc de :

- Comprendre les grandes lignes de l'assemblage de génomes,
- Comprendre l'indexation de k-mers et les grandes lignes de l'outil RedOak
- Comprendre le travail préliminaire déjà réalisé
- Préparer un jeu de données de test avec des données réelles
- Exécuter l'outil RedOak sur ce jeu de données
- Concevoir et implémenter un programme permettant d'obtenir les zones communes et les zones divergentes de plusieurs assemblages.

Bibliographie

- Assemblage de génome: https://fr.wikipedia.org/wiki/Assemblage_(bio-informatique)

- RedOak : RedOak: a reference-free and alignment-free structure for indexing a collection of similar genomes. Clément Agret, Annie Chateau, Gaetan Droc, Gautier Sarah, Manuel Ruiz, Alban Mancheron. bioRxiv 2020.12.19.423583; doi: https://doi.org/10.1101/2020.12.19.423583

- Méta-assemblage : rapport bibliographique d'Elisa Henrion-Gueneau

- Travail préliminaire : rapport de L3 CMI info de Julien Blanco