HAU902I Bioinformatique avancée

TP Assemblage

L'objectif de ce TP est de créer un outil d'assemblage de génomes qui soit capable de réaliser un assemblage sur un jeu de données de taille réduite. L'exemple choisi est le génome de la mitochondrie du varan de Komodo. Les données de séquençage sont disponibles sur l'espace moodle.

Le TP, ainsi que le compte-rendu seront à rendre sous la forme d'une archive sur moodle. Le travail peut se réaliser en groupe de deux ou trois, ou individuellement. Vous indiquerez dans votre compte-rendu le nom des participants, et ne rendrez qu'une seule fois l'archive.

1 Étape de conception

Exercice 1 *Choisir parmi les trois stratégies vues en cours, une stratégie à mettre en place. Vous rappelerez sur le compte-rendu les grands principes de cette stratégie, ses avantages et ses inconvénients.*

Exercice 2 *Indiquer dans un schéma les grandes étapes de votre outil, depuis la donnée initiale jusqu'à la donnée finale, en précisant les formats de fichiers (par exemple, la donnée initiale se présente sous la forme d'un fichier fastq, qu'il va falloir lire...)*

Exercice 3 *Indiquer quel sera le choix du langage pour l'implémentation, en justifiant.*

Exercice 4 *Écrire les algorithmes correspondant à chacune des étapes identifiées dans le schéma de conception, en indiquant les structures de données choisies*

2 Implémentation

Exercice 5 *Implémenter, dans un code propre et commenté, les étapes de votre assembleur.*

Exercice 6 *Testez votre assembleur sur les données de génome de mitochondrie du varan de komodo. Indiquez le temps moyen d'exécution de votre programme sur ces données. Listez les difficultés rencontrées lors de l'étape d'implémentation*

3 Avez-vous bien travaillé ?

Pour vérifier que votre algorithme d'assemblage fonctionne bien, on peut utiliser un outil comme QUAST, qui prend en paramètre le génome de référence et les contigs, et vous donne beaucoup de statistiques intéressantes. Vous pouvez l'utiliser en ligne ici par l'intermédiaire de la plate-forme galaxy : `https://usegalaxy.org/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fiuc%2Fquast%2Fquast%2F5.0.2%20galaxy3`.

Histoire de comparer la qualité de votre assembleur avec l'assembleur minia, un assemblage a été généré (en 0.1s sur une machine i7 4 coeurs, 16Go de RAM) avec la commande :

```
minia reads.fastq 30 3 10000 minia
```

Le résultat (`minia.contigs.fa`) est sur le Moodle, ainsi que le génome de référence `varankomodo_reference.fasta`.

Exercice 7 *Ajouter à votre compte-rendu l'analyse proposée par quast de votre assemblage, comparativement à celle de l'assemblage réalisé par* `minia` *et au génome de référence.*