Projet : Séquençage : Assemblage de Génomes…

Encadrant : Mr A. El Hassouny *elhassounyphd@gmail.com*

Description

D'un point de vue informatique, l'ADN, l'ARN et les protéines peuvent être considérés comme simplement des chaînes qui se composent d'un ensemble fini de lettres (composé de quatre lettres pour l'ADN et l'ARN, et de 20 lettres pour les protéines). les séquences d'ADN peuvent être considérées comme des chaînes sur $\Sigma = \{A, G, C, T\}$ alors que les séquences d'ARN peuvent être considérées comme des chaînes sur $\Sigma = \{A, G, C, U\}$.

Le traitement efficace de ces chaînes est nécessaire pour toutes les techniques de d'assemblage et de correction des erreurs qui sont appliquées aux séquences bioinformatiques. Par conséquent, de telles techniques utilisent largement plusieurs structures de données telles que les arbres, des tableaux de hachage etc.

Pour exploiter les génomes à des fins de traitement des maladies, on y besoin d'un algorithme d'assemblage et une méthode rapide de recherche de fausses souches.
L'assemblage de séquences consiste à aligner et fusionner des fragments d'une chaîne d'ADN plus longue (créées lors de l'étape de séquençage) afin de reconstruire le génome ou la séquence de départ. C'est-à-dire, étant donné une collection de deux ou plusieurs fragments d'ADN qui se chevauchent, vous allez aligner et fusionner ces fragments en choisissant les meilleurs correspondances à chaque étape, avec l'objectif de se terminer par une seule séquence plus longue.

Pour la recherche des fausses souches, dans des nombreuses applications, les arbres peuvent être utilisés efficacement pour rechercher un motif de chaîne spécifique (souche bactérienne) dans une grande collection de chaînes (millions de superpositions de séquences du génome).

Dans ce projet, les défis seront :

1. **Dans un premier temps, de retrouver la séquence originale à partir des fragments donnés en sortie du séquenceur.**
2. **Dans deuxième temps, de rechercher un motif de chaîne spécifique (souche bactérienne) dans une grande collection de chaînes (millions de superpositions de séquences du génome).**

Par exemple, ici, il y a un chevauchement de trois (CAT). Si nous les fusionnions, le résultat serait CGCATGAC:

```
  CGCAT
+ CATGAC
---------
  CGCATGAC
```

Si nous les essayions dans l'autre sens, à quoi ressembleraient le chevauchement et le fragment fusionné?

```
  CATGAC
+ CGCAT
-----------
  CATGACGCAT
```

Contraintes de programmation

Votre programme devra proposer **un menu** comme celui-ci :
(1) Assemblage

-
-

(2) Recherche

-
-

(3) Quitter le programme
Il est également recommandé d'avoir la **sécurité** des accès à l'aide de mots de passe.