Examen d'algorithmique des séquences

N. Maillet - nicolas.maillet@pasteur.fr - rendu le 22/01/2023 à 23h59

2022-2023

Le barème est à titre indicatif et prend en compte plusieurs facteurs (utilisation de structures adaptées, résultat correct, commentaires, efficacité du code, etc).

Les librairies externes ne sont pas acceptées (**numpy est interdit**).

Q 1. *(4pts)* À partir de votre algorithme de Needleman-Wunsch, codez la distance de Levenshtein.

Q 2. *(6pts)* À partir de votre algorithme de Needleman-Wunsch, codez l'algorithme de Smith-Waterman et utilisez la distance de Kimura suivante :

	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

Détection de reads mal séquencés

Le fichier `reads.fasta` contient 130 465 reads de 100 paires de base (pb). Toutes ces séquences sont issues d'un séquençage (de type Illumina) d'un petit génome de 100 000 pb. La **couverture** est donc d'environ 130. Autrement dit, chaque nucléotide dans le génome initiale est, en moyenne, présent dans 130 reads, pas forcément identiques.
Le séquençage n'est pas une technique infaillible. Dans ce jeu de données certains reads comportent une substitution d'un nucléotide en un autre. **Le but est d'identifier ces séquences**.
Pour cela, nous allons utiliser une approche de type k-mers. En analysant tous les k-mers et leur fréquence d'apparition, on va pouvoir identifier les "mauvais" reads.

Q 3. *(5pts)* Créez un compteur de k-mers (k=30). *Découpez tous les reads en k-mers et comptez, pour chaque k-mers, combien de fois il apparaît dans le jeu de données.*

Q 4. *(5pts)* Utilisez ce compteur pour identifier les reads mal séquencés. Affichez les **identifiants** des séquences "fausses" du jeu de données. *On va fixer la limite à 2 occurrences : un k-mers apparaissant trois fois dans le jeu de données est considéré comme valide, un k-mer n'apparaissant qu'une ou deux fois sera considéré comme comportant une erreur de séquençage.*

Q 5. Bonus *(2pts)* Trouvez une autre méthode pour identifier ces séquences en allant environ 2 fois plus vite (ou encore plus vite, mais une solution assez simple existe pour 2x :p) *Si vous avez déjà implémenté cette solution dans les deux questions précédentes, vous aurez les points bonus. Et oui, ceci est un indice.*