Charles-Antoine Lanoix 20 novembre 2019  
Félix Adam

IFT3295 – TP3

2. Questions sur la mise en pratique

**1. Donnez l’output de votre programme pour chacune des séquences du fichier unknown.fasta qui contient des séquences d’ARNts dont la nature est inconnue. On vous demande d’utiliser les paramètres par défaut (-ss = 0:001, -E = 4, seed = ’11111111111’).**

Exécuter la dernière cellule du notebook avec seqNum de 1 à 4.

**2. En déduire, si possible la nature de chacune des séquences. Notez que l’identifiant des séquences de la banque de données est sous la forme : Amino-Acide|Anticodon|Espèce. Vous pouvez au besoin jouer avec le seuil de signification (-ss) pour affiner votre recherche.**

* 1. Sans équivoque, on voit que la nature est “cat”.
  2. “tcg” et “tct” donnent les deux des e-value et pourcentage d’identité semblables, il est difficile de déterminer la nature exacte de la séquence.
  3. Les deux matchs les plus parlant pointent vers “tgg”.
  4. Même si les résultats ocillent entre “tct”, “tcg” et “tgg”, le meilleur match, et de loin, est celui ayant comme nature “tct”. Nous sommes donc portés à croire qu’il s’agit de la bonne réponse

3. **Vérifiez vos résultats en vous servant du véritable outil BLAST pour nucléotide (BLASTN) disponible sur NCBI ==> https://blast.ncbi.nlm.nih.gov/Blast.cgi . On vous demande de comparez les deux résultats.**

* 1. Les résultats concordent pour les deux meilleurs matchs trouvés, matchs qui se retrouvent aussi dans la liste de nBLAST.
  2. On retrouve Marchantia polymorphia avec le même pourcentage d’identité, mais on voit clairement que Nephroselmis olivacea aurait été le meilleur match globalement (il n’est pas dans notre petite BD).
  3. Pas les mêmes résultats. La séquence de la BD permettant le meilleur alignement local n’est pas dans la liste produite par nBLAST, probablement en raison de sa faible taille. Pour les alignements plus long, le pourcentage d’identité descend autour de 75%, ce qui n’est clairement pas suffisant.

Le match mixant meilleur e-value et meilleur pourcentage d’identité, Marchantia polymorpha, se retrouve dans la liste de nBLAST avec le même résultat.

**4. Qu’arrive t’il lorsque vous utilisez des graines plus longues ? plus courtes ? (on vous demande l’impact sur la vitesse, la précision et la sensibilité de PLAST)**

Les tests se feront en utilisant la séquence CGTAGTCGGCTAACGCATACGCTTGATAAGCGTAAGAGCCC sur tRNAs.fasta avec un premier seuil de -5 et un second seuil de 10.

Avec une graine de taille 11, nous avons un temps d’exécution de 0,445 seconde et 1 résultat retenu. Le meilleur appariement possède un score de 20 avec 81,25% d’identité. Le score du meilleur HSP est de 78 avec un bitscore de 24 et e-value de 1.81e-04.

Avec une graine de taille 7, nous avons un temps d’exécution de 0,755 secondes et 24 résultats retenus. Le meilleur appariement possède un score de 8 avec 90% d’identité. Le score du meilleur HSP est de 31 avec un bitscore de 11 et e-value de 1.38

Avec une graine de taille 4, nous avons un temps d’exécution de 3,13 secondes et 240 résultats retenus. Le meilleur appariement possède un score de 8 avec 90% d’identité. Le score du meilleur HSP est de 31 avec un bitscore de 11 et e-value de 1.38

Avec une graine de taille de plus de 11, la recherche ne retourne aucun résultat mais se fait rapidement (0,413 seconde pour graine de taille 12, 0,371 seconde pour graine de taille 15).

Pour résumer, l’utilisation de graines plus courtes aura un impact négatif sur la performance de l’algorithme. Le nombre de résultats retenus augmentera de façon significative, au détriment de la précision des appariements.

Inversement, l’utilisation de graines plus longues aura un impact positif sur la performance de l’algorithme. Le nombre de résultats retenus diminuera mais la précision des appariements sera augmenté.

**5. PatternHunter, un autre algorithme de recherche, utilise des graines espacés par des positions "don’t care" comme "111010010100110111". Adaptez votre algorithme aux graines de PatternHunter et comparez ses perfomances à celui de l’algorithme original (utiliser un seuil de signifiance (-ss) moins stringent tel que 0.01)**

Pour ces tests, nous utilisons les mêmes paramètres que dans le numéro précédent excepté pour le seuil de signifiance qui sera paramétré à 0.01.

Avec « 11001110011 » comme graine, on obtient un temps d’exécution de 0,45 seconde rapportant 1 résultat. L’appariement en question a un score de 17 et un pourcentage d’identité de 77,42%. Le score du meilleur HSP est de 74 avec un bitscore de 23 et un e-value de 3,57e-04

Avec une alternance de uns et de zéros de longueur 11, on obtient un temps d’exécution de 0,5 seconde. Le seul appariement sortant est le même que sur l’exemple du TP3, avec un score de 20 et un 81,25% d’identité.

Une observation est le fait qu’il est possible d’avoir des graines de longueur supérieure à 11 avec des « don’t care » qui trouveront des appariements.

Une trop grande quantité de « don’t care » dans la graine utilisée fera en sorte que l’algorithme prend un temps considérable.