HAI109H – Bioinfo 1

PROJET

Date limite de rendu du projet : vendredi 9 décembre 2022 – 18h

Contrôle en TP sur le projet : lundi 12 décembre 2022 – 9h45 à 11h15

Vous devez inscrire votre groupe sur le fichier partagé des groupes projet présent sur Moodle (section Projet) : 2 à 3 étudiants par groupe.

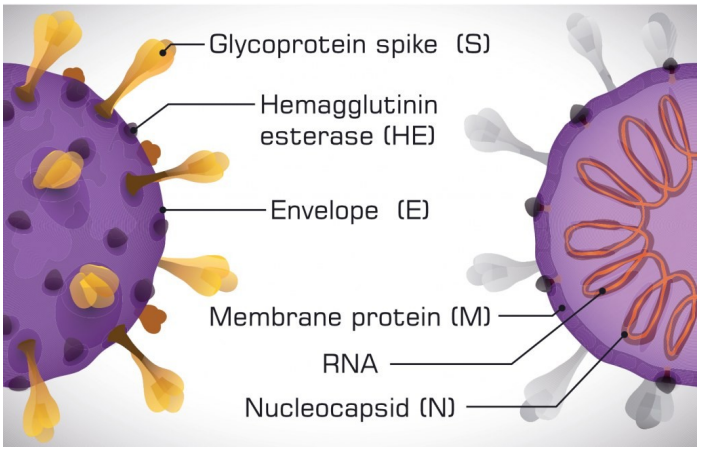
1. SUJET

Le coronavirus SARS-CoV-2 qui infecte l'Homme est proche de deux coronavirus qui infectent d'autres espèces : le coronavirus de chauve-souris RaTG13 (SARSr-CoV RaTG13) et le coronavirus du Pangolin (souche MP789).

Le projet consiste à étudier ces 3 virus (SARS-CoV-2 de l'Homme, SARSr-CoV RaTG13 de la Chauve-Souris et MP789 du Pangolin) et d'observer les ressemblances et les différences en étudiants certaines protéines présentes dans les 3 virus.

Vous devrez réaliser un programme BioPython qui analyse puis compare les différentes versions de certaines protéines de chacun des virus. Les protéines étudiées seront :

- La protéine d'enveloppe (*envelope protein*) codée par le gène "E"
- La protéine de pointe (*spike glycoprotein*) codée par le gène "S"
- La protéine membranaire (*membrane glycoprotein*) codée par le gène "M"
- La protéine de la nucléocapside (*nucleocapsid phosphoprotein*) codée par le gène "N"



Source : https://www.clinisciences.com/achat/cat-sars-cov-2-antigenes-proteines-5102.html

Les étapes à coder dans votre programme sont détaillées par la suite.

2. RENDU avant le VENDREDI 9 décembre à 18h

Pour rendre votre projet, vous devez déposer sur Moodle dans la section Projet un fichier zip dont le nom sera "GroupeNuméro.zip" **avant le vendredi 9 décembre 2022 à 18h**. Vous trouverez le numéro de votre groupe dans le fichier partagé des groupes projet sur Moodle (exemple : Groupe1.zip).

Ce fichier doit contenir :

- Un fichier jupyter notebook “numeroGroupe.ipynb” ou un/des fichiers python “numeroGroupe.py” contenant votre code bien commenté ainsi que les noms des personnes du groupe. Si vous avez plusieurs fichiers, vous créerez une archive contenant tous les fichiers de code.
  Attention, les commentaires devront permettre de comprendre votre code !

- Un fichier “rapport-NumeroGroupe.pdf” correspondant un petit rapport qui contiendra :
  - la constitution du groupe dans le rapport,
  - comment exécuter votre code (jupyter notebook ou besoin spécifique)
  - pour les étapes de base :
    - une explication de votre code pour les étapes A à G,
    - les réponses aux question de l’étape H
  - pour les étapes avancées :
    - une explication de votre code pour l’étape I,
    - une explication détaillée de l’algorithme d’alignement par paire de l’étape J
    - une explication de l’étape K : explications des modifications effectuées dans le code des étapes E, F et G, afin d’utiliser votre algorithme d’alignement dans le programme principal

3. DÉTAILS DU PROJET

Pour réaliser le projet vous devrez coder chacune des étapes suivantes. Les étapes du programme se décompose en 2 ensembles : les étapes de base qui doivent être faites successivement dans l’ordre donné et qui sont dépendantes les unes des autres, et les étapes avancées qui sont indépendantes l’une de l’autre.

[ÉTAPES DE BASE]

A. Récupérer sur la banque « Nucleotide » du NCBI la séquence de référence (RefSeq) du génome du SARS-CoV-2 de l’Homme et les séquences des génomes du SARSr-CoV RaTG13 de la Chauve-souris et de la souche de coronavirus MP789 MT121216 du Pangolin, puis créer un fichier Genbank contenant les 3 séquences (“seq_covid.gb”).

B. A partir du fichier généré précédemment “seq_covid.gb”, vous créerez un fichier texte “info_seq_covid.txt” contenant un résumé des informations importantes présentes dans l’en-tête Genbank de chaque génome. Dans ce fichier, dont la mise en forme devra être soigné, il faudra retrouver pour chaque génome :
    - Le nom de l’organisme dont provient la donnée ainsi que la taxonomie correspondante
    - le numéro d’accession de la donnée Genbank
    - la date de création de la donnée Genbank
    - le nombre de gènes présents sur la séquence
    - la liste des noms de gènes présents avec pour chacun la position de début et de fin du CDS correspondant, ainsi que l’identifiant de la protéine codée
    - le % de GC de la séquence du génome

C. Utiliser les données présentes dans le fichier Genbank “seq_covid.gb” pour trouver la protéine SPIKE (nom de gène “S”) de chaque virus / espèce et faire un fichier multi-fasta (“spike.fasta”) contenant les séquences protéiques de Spike pour chaque virus / espèce.

D. Utiliser le script BioPython disponible sur Moodle (section Projet) pour aligner les séquences présentes dans le fichier spike.fasta. L’exécution du script permet d’obtenir un fichier “aln-spike.fasta” des séquences alignées.

E. A partir du résultat obtenu par l’exécution du code donné à l’étape précédente (“aln-spike.fasta”), comparer les séquences, en donnant les positions où vous observez des différences et en indiquant à chaque fois les lettres correspondantes pour chaque virus / espèce. Le résultat devra être stocké dans un fichier (“resultatComparaison_geneS.txt”).

Exemple de résultat attendu à l’étape D :

POSITION	HOMME	CHAUVE-SOURIS	PANGOLIN
5	A	A	T
43	G	C	C
...			

F. A partir des résultats précédents, vous calculerez le taux de conservation de la protéine Spike dans les trois virus / espèces en prenant comme référence l’Homme : pour cela, indiquez le pourcentage de lettres identiques pour le virus de la Chauve-Souris et celui du Pangolin par rapport à celui de l’Homme.

G. Vous devrez automatiser votre code pour permettre d’exécuter votre analyse (étapes C à F) sur n’importe quelle protéine dont l’utilisateur donne le nom

du gène correspondant. Par exemple, il faudra que votre programme à l'étape C demande le nom du gène d'intérêt à l'utilisateur et que votre programme continue l'analyse pour le gène donné par l'utilisateur (gène "S" par exemple). Vous exécuterez alors votre programme (étapes C à F) sur les autres protéines (codées par les gènes "S", "M" et "N"), en plus de la protéine codée par le gène "S".

H. En observant le résultat obtenu précédemment, que pouvez-vous en conclure sur la conservation / la ressemblance des différentes protéines présentes dans les 3 coronavirus ? Pouvez-vous en conclure que les trois virus se ressemblent, et si oui, quels virus semblent être les plus proches ?

[ÉTAPES AVANCÉES]

I. Vous créerez un fichier au format GenBank contenant les données des 3 séquences protéiques du gène choisi par l'utilisateur à l'étape C pour les 3 génomes. Ce fichier devra contenir les informations disponibles sur le portail du NCBI pour la protéine choisie. Par exemple, si l'utilisateur choisit le gène "S", vous devrez créer le fichier contenant les données de la protéine "Spike" au format GenBank pour les 3 génomes. Pour cela, vous pourrez vous utiliser les identifiants de la protéine correspondant au gène d'intérêt présent dans les données du fichier GenBank des génomes et récupérer sur la banque «Protein» du NCBI les données GenBank des séquences correspondantes.

J. L'étape D d'alignement des séquences a été faite à l'aide d'un script fourni, qui utilise un programme d'alignement. Vous proposerez un programme indépendant que vous aurez vous même implémenter, sans utiliser de programme existant, permettant d'aligner 2 séquences (alignement par paire). Votre programme devra implémenter le principe de programmation dynamique vu en cours. Il devra, à partir d'un fichier contenant plusieurs séquences au format FASTA, créer un fichier contenant tous les alignements par paires des séquences au format FASTA.

K. Vous créerez une 2eme version du programme général implémentant les étape A à G en remplaçant le script d'alignement fourni pour le projet par votre programme d'alignement par paire (étape D). Vous adapterez les étapes E à G en fonction.