HAI109H – Bioinfo 1

PROJET

Date limite de rendu : vendredi 10 décembre 2021 - 18h

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

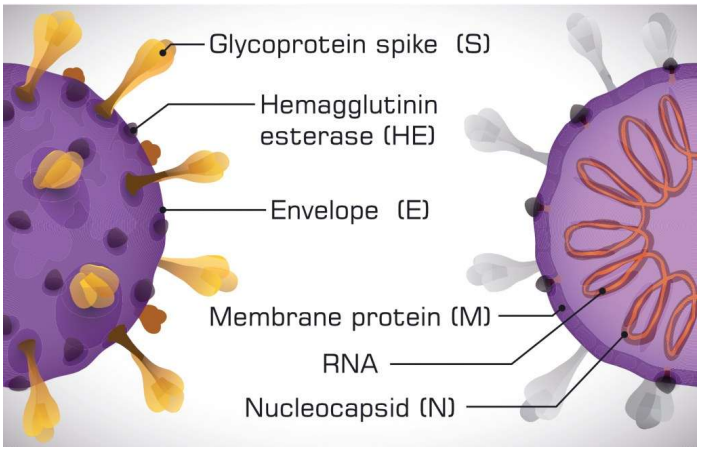
1. SUJET

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 pour comparer 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. <u>RENDU avant le VENDREDI 10 décembre à 18h</u>

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 10 décembre 2021 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” contenant un petit rapport expliquant ce que vous avez fait et comment exécuter votre code (si besoin spécifique). Vous indiquerez également la constitution du groupe dans le rapport.

3. <u>DÉTAILS DU PROJET</u>

Pour réaliser le projet vous devrez coder chacune des étapes suivantes :

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

   Requête pour l’ Homme : “SARS-Cov2 [ORGN] AND srcdb_refseq [PROP]”
   Requête pour la Chauve souris : “SARSr-CoV RaTG13”
   Requête pour le Pangolin : “Pangolin coronavirus isolate MP789 MT121216”

B. Utiliser les données présentes dans le fichier Genbank 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.

C. 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.

D. 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) et vous devrez également afficher ce résultat à l'écran.

Exemple de résultat attendu à l'étape D :

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

E. 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 différentes pour le virus de la Chauve-Souris et celui du Pangolin par rapport à celui de l'Homme.

F. Vous devrez automatiser votre code pour permettre d'exécuter votre analyse (étapes B à E) sur n'importe quelle protéine dont on connaît le nom du gène correspondant. Vous exécuterez alors votre analyse (étapes B à E) sur les autres protéines (codées par les gènes E, M et N).

G. 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 ?