Exercise 1: Designing an mRNA vaccine against SARS-CoV-2 variants

Hello! You were recruited by the Pfizer/BioNTech Bioinformatics team. As you know, the vaccine operation is going great but we are concerned that the vaccine is not working well for the BA.2 variant. Your mission is to update the mRNA vaccine based on the BA.2 Spike!

Prerequisite:

Step 1:

Install the BioPython package.

Step 2:

Download the mRNA design of the Pfizer and Moderna vaccines from: Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273.

Step 3:

Download the sequence of “Wuhan” SARS-CoV-2 sequence from NCBI Genbank accession number NC_045512.2.

Step 4:

The NCBI Virus database has millions of complete sequences of SARS-CoV-2. Download all complete sequences of SARS-CoV-2 BA.2 in a FASTA format. You don’t need these sequences until question #15 (hint: to download only BA.2 use the Pango lineage field on the left side).

Questions

Basics

1. Create a function that takes mRNA as an input and returns the translated coding region (Hint 1: you need to find the initiation codon and the stop codon. Hint 2: use BioPython!).

2. For each vaccine:
    1. What’s the position of the initiation codon (“AUG”)?[1]
    2. What’s the position of the termination codon?
    3. What sequence did each company use for the stop codon?
    4. Please suggest another stop codon sequence.
    5. How would you call the area between the first nucleotide and the “AUG” site?

[1] All positions should be reported as zero based

    6. How would you call the area between the termination codon and the last nucleotide?

3. Using your function, translate the coding area into protein. What’s the length of each coding sequence?

4. Are the coding regions of the vaccines identical in the protein level?

5. Are the coding regions of the vaccines identical in the nucleotide level?

Codon usage

6. Create a histogram of amino-acids usage for the Pfizer vaccine

7. Create the same histogram but color each codon within the amino-acid (i.e. create a stacked histogram)

Comparison to the Wuhan strain:

8. What’s the length of the Wuhan strain genome?

9. Report the nucleotide sequence of the Spike region of the Wuhan strain from the initiation codon to the termination codon (hint: use the information in the NCBI website about the starting and stopping position of Spike).

10. Report the GC content of the Spike region of the Wuhan strain and compare it to the GC content of the Spike region of the Pfizer vaccine.

11. Translate the sequence into amino acids and report the sequence of the protein

12. Compare between the Spike region of the Pfizer vaccine to the Spike of the Wuhan strain in the protein level. Are there any differences and where?
    a. Bonus: explain the differences!

13. For the amino acids that appear in the vaccine and the Wuhan strain, create a look-up table that maps between the Wuhan strain codon to the Pfizer codon

14. Can you deduce any rule about the look-up table that Pfizer uses internally?
    a. Bonus: explain the differences!

The Variants:

15. The BA.2 genomes that you have are NOT aligned to the Wuhan strain. For example:

```
Position:                                  11111 11111 2
                               01234 56789 01234 56789 0
Wuhan strain:                  ACCGT GCAAT TGGCT AAAAA
A genome in your collection:   ACCGT GCTAT TGGGC TAAAA A
```

    Notice that a nucleotide 13th in the genome Wuhan strain corresponds to nucleotide 14th in the genome in your collection. So you cannot simply use positions from the Wuhan strain to find positions in your collection.

    Unfortunately, the only information that you have about the BA.2 variant is based on the Wuhan coordinates. There are multiple strategies to identify the Spike region but using the same coordinates of Wuhan is the wrong strategy.

16. Report all entries in your collection that you think match the BA.2 variant. Where were they collected? Does it make sense?

17. Develop a method to extract the Spike sequence of one of the entries, translate, and report the protein sequence (hint: the BA.2 spike is not very different than the Wuhan strain and must start with AUG).

18. Use everything you learned so far, including the look up table of Pfizer and the addition of special nucleotides to design a South African mRNA vaccine. Don’t forget the add the 5’UTR and 3’UTR. Report the vaccine sequence.

19. Repeat the same process for the BA.1 genomes. How many amino-acid differences are there between the genomes of BA.1 and BA.2?