CS 590 Introduction to Bioinformatics
Homework Assignment (6)

Objectives

- Query/read/Analyze DNA sequence data.
- Create and use score matrices.
- Score and analyze sequence alignments

Description

Consider two sequences for the Human corona virus stored in the GenBank under accession numbers "AY884001" and "MH940245". *Write a R code to perform each of the following tasks :*

1. **(10 points)** Retrieve the two sequences and store them in two different objects.
2. **(20 points)** for each sequence, compute the frequency of each amino acid and plot them as a pie chart. *Generate only one figure for both sequences.*
3. **(15 points)** create a dot plot for the first ORF in each of the sequences. *Comment on the result.*
4. **(20 points)** Find the optimal global alignment between the two sequences and print the alignment for the first 20 nucleotides. *Use a score +2 for a match, -1 for a mismatch, and the gap penalty = 2.*
5. **(15 points)** Is the global alignment statistically significant? *Explain your answer.*
6. **(20 points)** What is the score of the optimal local alignment between the two sequences? What is the length of the aligned segments? *Use a score +3 for a match, -2 for a mismatch, the gap penalty = 4, and gap extension = 2.*

Notes:

- Handwritten answers are not allowed!
- Use Rmarkdown (https://rmarkdown.rstudio.com/) and provide a neatly formatted "pdf" file showing both code and output.
- Include your name as a comment at the beginning of the script file.