Problem set 3

This problem set will be short and mostly theoretical, because real metagenomics data are too difficult for us to handle at this point in the course. But we will at least do some phylogenetic reconstructions.

Phylogenetics and molecular evolution

The following series of questions refer to a phylogenetic analysis of viral surface glycoproteins. The FASTA file (Surface-glycoprotein.fasta) is provided on the course website. You may use any phylogenetic tool, but MEGA (https://www.megasoftware.net) will be used as example here.

First, let's try to understand the data.

Q1: What information is contained in this FASTA file? Are they nucleotide or amino acid sequences? Are they coding or genomic DNA sequences?

Q2: This is the first entry in the FASTA file:

>lcl|NC_002645.1_cds_NP_073551.1_1 [gene=S] [locus_tag=HCoV229Egp2] [db_xref=GeneID:918758] [protein=surface glycoprotein] [protein_id=NP_073551.1] [location=20570..24091] [gbkey=CDS]

Can you make sense of this header information? Specifically, what do "NC_002645.1", "NP_073551.1", and "918758" refer to?

Now, we will start building the phylogenetic tree. The first step is to align these sequences together using **multiple sequence alignment**.

Q3: Which type of alignment would you perform, nucleotide alignment, codon alignment, or protein alignment? Why?

Q4: Perform the alignment of your choice. Show a screenshot of the resulting alignment.

Before jumping into the maximum likelihood method, let's review simpler ways to build a phylogenetic tree: **maximum parsimony** and **minimum evolution**.

Q5: What is the key assumption or hypothesis behind maximum parsimony and minimum evolution approaches?

Q6: Under which scenario would the assumption behind maximum parsimony and minimum evolution approaches be violated? *Hint: Think about what can happen over a long period of time in terms of mutations.*

Next, we will identify a good substitution model for this data. In class, we learned that this can be done by comparing the likelihood between a simpler model and a more complex model using a procedure called **nested model testing**. A result from MEGA is provided here for the Juke-Cantor (JC), Kimura-2-parameter (K2), and Tamura-3-parameter (T92) models.

Model	Parameters	BIC	AICc	*lnL*
JC+G	60	2478.570	2174.169	-1024.135
K2+G	61	2485.186	2175.815	-1023.858
T92+G	62	2485.304	2170.966	-1020.330
JC+G+I	61	2485.741	2176.370	-1024.135
K2+G+I	62	2492.358	2178.020	-1023.858
T92+G+I	63	2492.475	2173.175	-1020.330
JC	59	2496.850	2197.423	-1036.861

Q7: Judging the models by only the likelihood (*lnL*), which would be the best one and why?

Q8: You may notice that the likelihoods for models with +G are exactly equal to the likelihoods of the corresponding models with +G+I. Can you explain what this means?

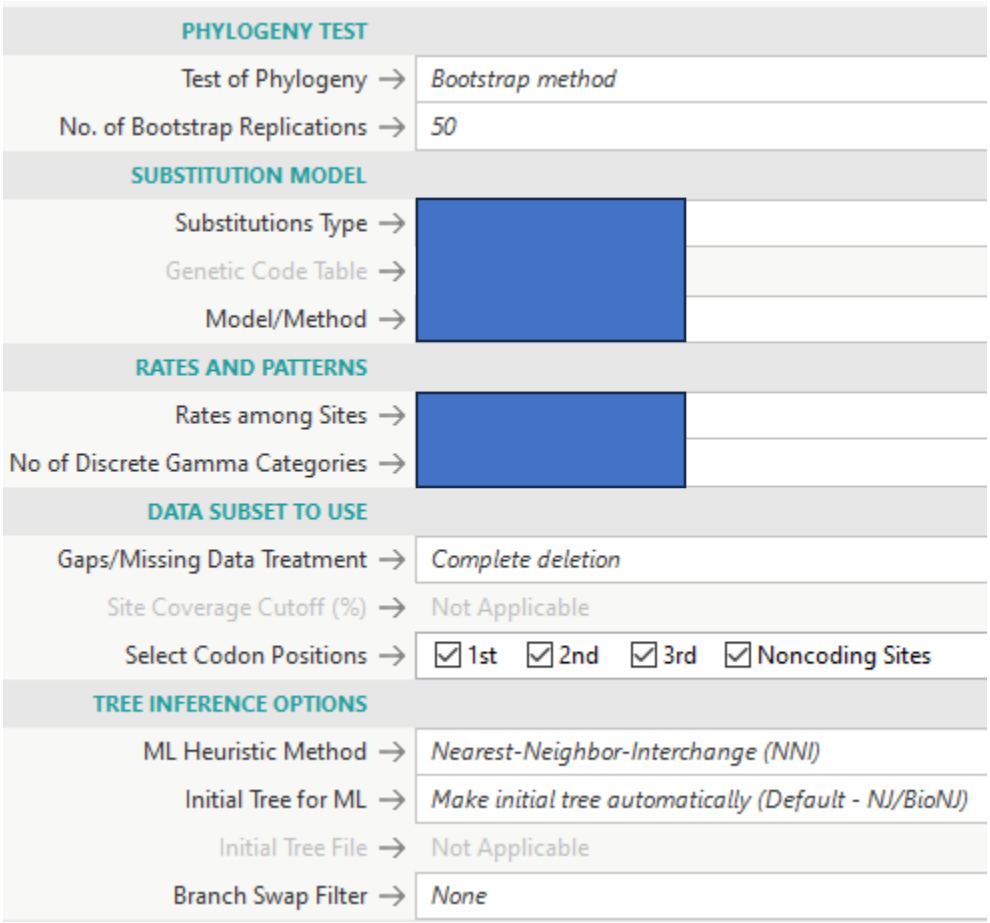
Q9: If you were to perform a nested model testing, which model pair would you considered first for the testing? Why?

Here, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) scores are provided.

Q10: Explain what they are and why they can be used to identify the best substitution model.

Q11: Based on BIC, which is the best model?

Finally, let us build phylogenetic trees using the maximum likelihood approach with the model selected in **Q11**. Here is a screenshot from MEGA for reference.



You may increase the number of bootstraps to 100 or 200 depending on how powerful your computer is.

Q12: What should be the setting for Substitution Type, Model/Method, and Rates among Sites?

Q13: Perform this analysis and show a screen shot of the phylogenetic tree that you get, with the bootstrap support values on the branches.

Q14: Explain what bootstrap support values are. Discuss whether you are confident about this phylogenetic tree.

Metagenomics

Q15: Explain the **rarefaction** procedure. Provide two applications of **rarefaction**.

Q16: Explain the principles behind read **binning** in metagenomics. Explain one algorithm that can be used to **bin** reads.

Q17: Compare the pros and cons of 16S rRNA amplicon sequencing and shotgun metagenomics in the context of what scientific questions we can and cannot answer.

	16S rRNA sequencing	Shotgun metagenomics
Pros		
Cons		

Q18: Visualize the graph of Simpson's index $D = \frac{1}{\sum {p_i}^2}$ for a system with two taxa, where the abundance of one taxon is p and the other is $1 - p$, with p ranging from 0 to 1.

Q19: Compute the 2-mer profile for the sequence AAAGCTAGGATCAGTCAGACT.

Q20: What are the main differences between **Markov Model** and **Hidden Markov Model** (HMM)? Why is HMM appropriate for gene prediction algorithm while the regular Markov Model is not? *Hint: Try thinking about what data can be observed and how it can fit into the model.*