Problem set 2

This problem set covers the content from week 2: sequence alignment and phylogenetics.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking from you is not just the correct answer but the rationale for your answer.
- Please provide evidence of how you think and what sources of information you used.
- AI such as ChatGPT may be used. You can also work together with friends. But you must write the answer in your own words.
- Any incidence of plagiarism and copying of another student's work will be reported to the Graduate Affairs.

Applications of sequence alignment

Q1: What are the key differences in the algorithms of MEGABLAST and BLASTN that make them yield alignment results with different levels of similarities? Explain how.

```
Query 213 GCTACCTCAGTTttcttcttcttctcttccttcagattcttccagttctctGCGCTGTCA  272
          |||||  ||  ||||||||  |||||  ||||||||  ||  ||||||||||  |||||  ||  |||
Sbjct 650 GCTACTTCGGTTTTCTTGTTCTTGTCTTCCTTTAGTTTCTTCCAGTCGTCTGCACTATCA  591
```

Q2: In the above alignment result, there are some nucleotides in **lower case letters** which indicate a low-complexity region. Explain what they are and how they impact the alignment.

Q3: Perform dynamic programming to determine the **best global alignment** for ATAGC and ATGCAGC under the following scoring scheme: **Match +1, Mismatch –1, Gap –2**. Indicate the best alignment.

	-	A	T	G	C	A	G	C
-	0							
A								
T								
A								
G								
C								

Q4: From your work in Q3, what would be your guess for the **best local alignment**?

You discovered the following transcript from an **RNA-sequencing** of an indigenous Thai plant:

>CL2742.contig2
GGCAGATACAGTCATAAAAGAAGTAACCAGAATCAGAAAGAATAAAGGTTGATTTG
GATTTAAGCCGACAGAACGATATAGCTACAATACAAAATTAACTTACAGTACAACA
GTATGATTTTAAATTTGCTGATAAACGAAAGGCCATTAACTTGAGAAACTACTAACT
CTTACATTGTTTTTTATCCTTTTAAGAAGCAGCAGTAGTAGTGGCTACCTCAGTTTTC
TTCTTCTTCTCTTCCTTCAGATTCTTCCAGTTCTCTGCGCTGTCAATAATTTCGCCAGT
TTTTATGAGGTAACGAACTCGTTTCCCATTGTCGAGGACTTTGTGACCCACCCGGCT
AACTACATCCTTTTCTTTAGAGTAGAGCATCACATTTGAACTATGAATAGGCGCTTC
GATCTTGATAATCTGGCCAGGTTCTCCTTCTTCTCTGCTCTTCATATGCTTTGTCTTCA
AATTTATTTCACCGACTACCACAGTGCTGTTGTGCTTAAAGATTTTTGTAACTTCACC
GACTTTCCCTTTATCATCTCCAGCTATCACTTTAACTGTGTCTCCTACTTTAACATGC
ATTTTATGTAAAACTGGAAGGCTGTTTGGTTTACATTCCTTCCGCTCCCACCGCTTAA
TCTGCAAGGGGGGAAGAATTGAGTAAATGGAATATAGGTTCATCAGACCCAAAATG
ACTTCAAA

Use this transcript to answer Q5-Q9.

Q5: Which BLAST algorithms would you use to investigate the biological function of this transcript? Why? *Try to come up with more than one answer!*

Q6: Perform a BLAST search. Summarize the top two hits.

Q7: What do you think is the likely function of this transcript?

Q8: What is the likely taxonomic group for this plant?

Q9: Identify the likely **open reading frame (ORF)** for this transcript. Explain your work.

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 tool, but **MEGA** (https://www.megasoftware.net) will be used as example here.

First, let's try to understand the data.

Q10: What information is contained in this FASTA file? Are they nucleotide or amino acid sequences? Are they coding or genomic DNA sequences?

Q11: 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]

Explain what "NC_002645.1", "NP_073551.1", and "918758" mean.

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

Q12: Given that these sequences are coding sequences, which type of alignment would you perform, *nucleotide* alignment, *codon* alignment, or *protein* alignment? Why?

Q13: Perform the alignment of your choice. Show a screenshot of the result.

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

Q14: What is the key assumption or hypothesis behind these approaches?

Next, we will identify a good substitution model for this dataset. 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

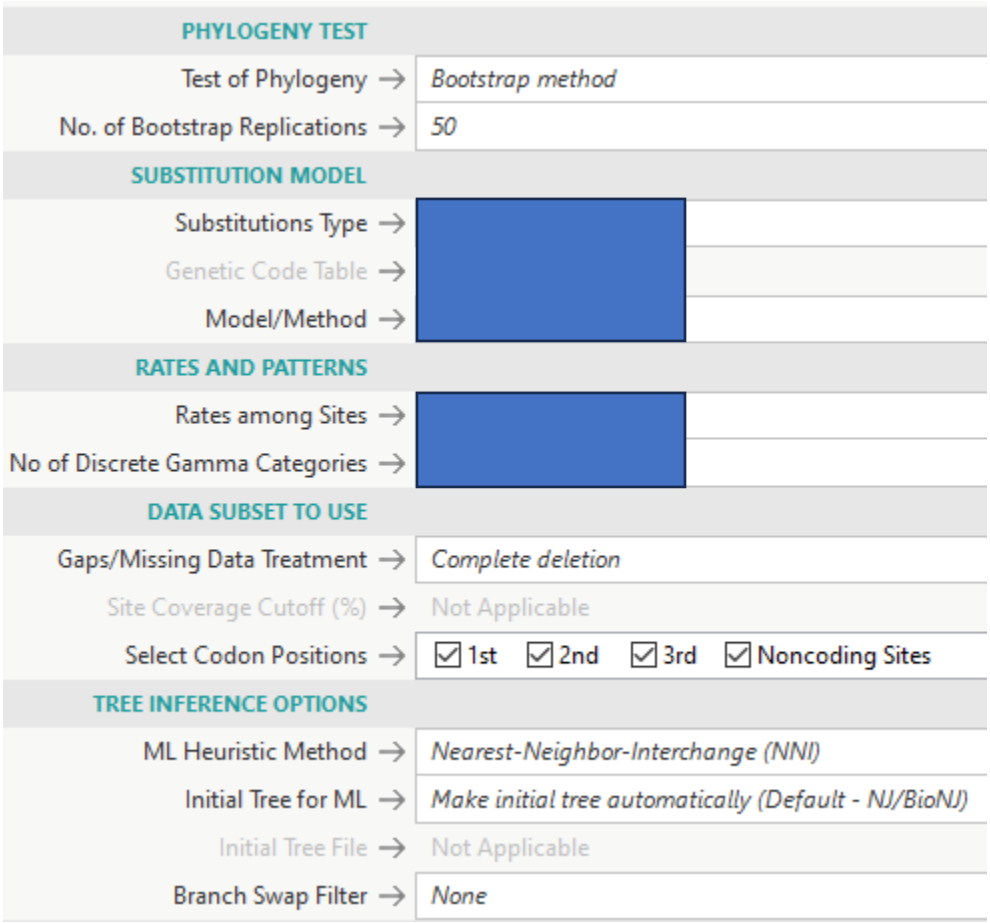
Q15: Judging the models by only the likelihood (*lnL*), which is the best model? Why?

Q16: Noticing that the likelihoods for models with +G are exactly equal to the likelihoods of the corresponding models with +G+I (such as JC+G and JC+G+I), what is your conclusion?

Q17: Explain what Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are. How can they be used to identify the best substitution model?

Q18: Based on BIC, which is the best substitution model?

Finally, let us build phylogenetic trees using the maximum likelihood approach with the model selected in **Q18**. Here is a screenshot from MEGA for reference.



You may decrease the number of bootstraps from 50 if your computer is slow.

Q19: Show your setting for the Substitution Type, Model/Method, and Rates among Sites (that were hidden above).

Q20: Show the resulting phylogenetic tree, with the bootstrap support values on the branches.

Q21: Discuss whether you are confident about this phylogenetic tree.