Assignment 4:

Instructions

Complete the following problem set showing your work. Problems may be worked out "by hand" or in "python" or with the assistance of other analytical software (e.g., Mathematica, MatLab). You may use chatGPT to assist in coding.

Solutions must be type written (e.g., in Jupyter, markdown, or latex). Upload PDF solution by question to crowdmark (link will be emailed to you) by **11:59pm** on the **Saturday** of the corresponding week (see syllabus). If you have issues with Crowdmark submission please email solutions to Rebekah Hall (rah11@sfu.ca).

All problems are equally weighted within an assignment. Students in 468 may or may not choose to attempt the challenge question for a bonus pts. Students in 795 are required to complete the challenge question.

Problem Set

1. Molecular evolution under the HKY 84
Consider the following ancestor and descendant viral sequences separated by $\Delta t = 1\text{year}$. Assume that the mutation/substitution of nucleotides in this virus occurs according to the HKY84 model.

Ancestor:	G	C	G	C	C	C	A	G	C	C	A	G	G	G	A	G	G	A	C	G
Descendant:	A	A	C	C	T	C	A	G	C	C	C	C	G	A	A	C	G	A	C	G

Part A:. What is the empirical stationary distribution $\vec{\pi}$?

Part B: What is the emperical (aka observed) probability of transitioning between A, C,G, and T in 1 year?

Part C: Using the HKY84 model and assuming that at most a single mutation occurs at a given base in a year, what is the transition probability matrix?

2. Molecular evolution under the GTR model

Consider a GTR model with the following parameters:**

$$\pi_A = 0.27, \pi_C = 0.25, \pi_G = 0.33, \pi_T = 0.15$$
$$a = 0.1, b = 0.2, c = 0.21, d = 0.12, e = 0.23, f = 0.15$$

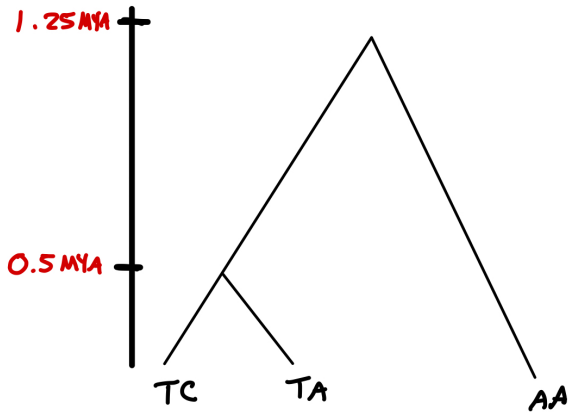
Part A: In the GTR model with these parameters, what is the stationary distribution?

Part B: Using the above model simulate a random ancestral genome with 50 base pairs.

Part C: Using the genome in part B as your initial condition, simulate molecular evolution in this genome for units of time. What is the final genome sequence?

Part D: (Challenge for 795) Repeat the simulation in Part C 20 times. Plot the number of mutations (an integer counter) over time for each of these replicate runs. The rate at which mutations occur in a genome is known as the **molecular clock**, estimate the **molecular clock rate** in this model.

3. Tree likelihoods
Consider the following tree topology and corresponding sequences.



Part A: Using a JC69 model with a mutation rate of $\mu = 1.5 \frac{\text{mutations}}{\text{Million Years}}$ what is the likelihood of this tree given the molecular model?

Part B: What is the most likely ancestral sequence at the root?

Part C (Challenge question for 795): Plot the tree likelihood for $0.05 < \mu < 0.75$. What is the maximum likelihood estimate for the mutation rate?

4. Models of mutation

Part A: In this course, we have considered several different models of mutation including models of mutation in the coalescent and models of nucleotide evolution. List these different models and describe their assumptions. Which are used for the coalescent and which are used for phylogenetic trees?

Part B: Coalescent mutation models are based on the assumption that no two mutations affect the same base in the genome either resulting in different alleles or occurring at different sites. Why might this be a reasonable assumption for a coalescent but not for a phylogenetic tree?