Homework 6: Phylogenetics

1. GTR vs JC69 Tree likelihood

Two possible molecular models for nucleotide evolution are the 1-parameter JC69 model (parameter: π) and another is the six parameter General Time Reversible Model "GTR" (parameters: $a, b, \ldots, f$) . The substitution rate matrices for these two model are given by:

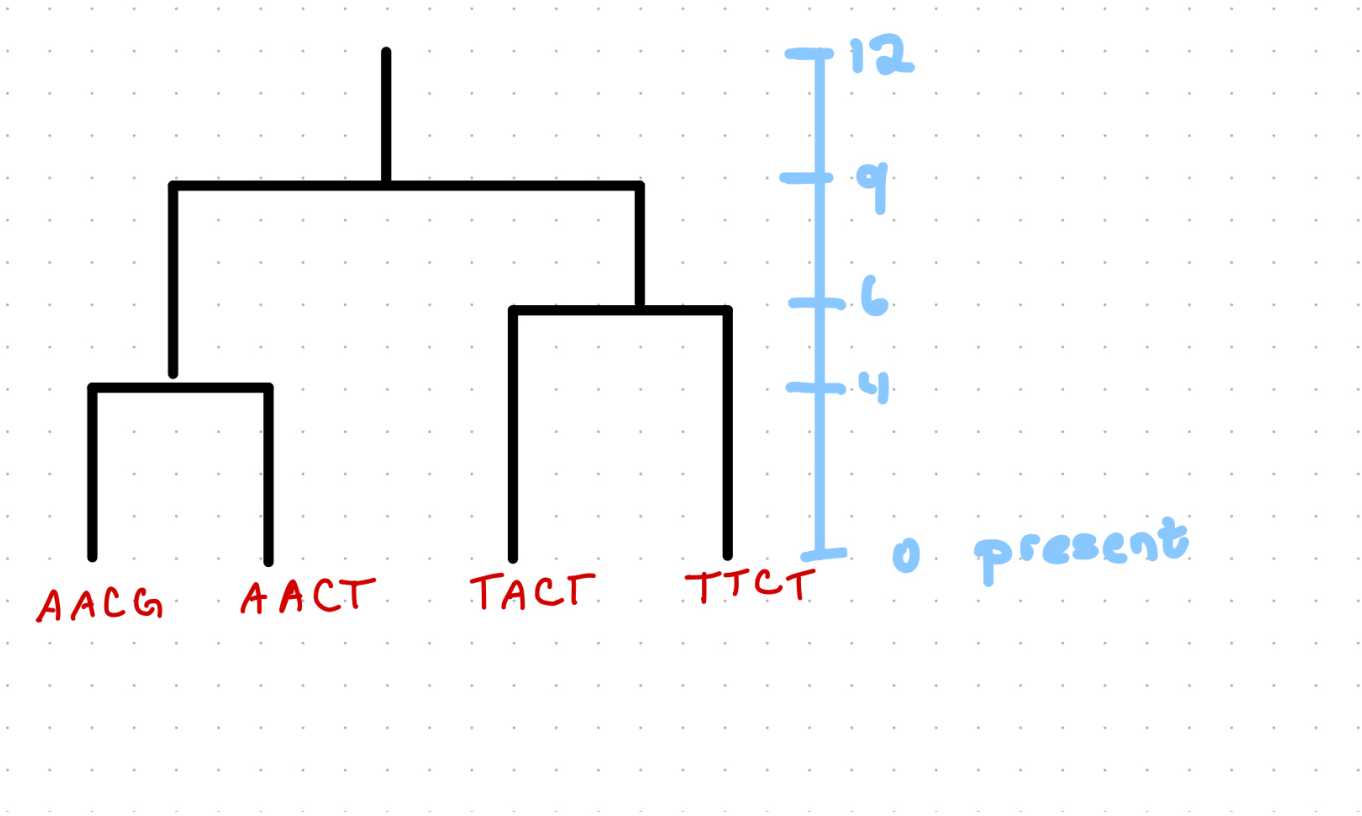
$$Q_{JC} = \begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}$$

and

$$Q_{GTR} = \begin{bmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_G + f\pi_G) \end{bmatrix}$$

where π_i is the "equilibrium" observed frequency of nucleotide i.

For the question below consider the following tree and sequence data.



A. What is $\vec{\pi}$ from the data? Assume $a = b = \cdots = f$ (the units of μ are measured in the time scale over which the sequences are collected). Show that π is the stationary distribution of the GTR matrix. What is the stationary distribution of the JC69 Matrix?

B. Calculate the transition probability matrix for the GTR model as a function of the model parameters and evaluate it for two parameter sets:

Parameter set 1: $a = b = c = d = e = f = 0.01$.
Parameter set 2: $a = b = c = 0.015$ & $d = e = f = 0.005$.

C. Calculate the likelihood of the tree under the GTR model two parameter sets. Which one is more likely? If this is all the information that you have what does this tell you about molecular evolution in this system?

2. Yule Model

The Yule model is a tree prior where only speciation occurs according to an poisson process (e.g., the waiting time between speciation events is exponentially distributed).

A. One way to simulate a tree is to simulate the "Cophenetic Similarity Matrix", $\mathbf{C}$ in which element c_{ij} gives the amount of time over which lineages i and j share a common ancestor. What is the cophenetic similarity matrix for the tree in question 1?

B. Consider a Yule tree prior where the rate at which speciation events occur is given by :

$$\lambda = \frac{1}{\text{million years}}$$

Simulate a tree with 10 species at the present day under the Yule model.

C. Repeat the simulation in A, 20 times and plot the distribution of time to the common ancestor T_{MRCA} of the tree.