Homework 2

STAT 517 - Winter 2023

1. The Tamura and Nei (1993) model (referred to from now on as TN93) is a very general model for DNA evolution that is widely used in practice. The model posits that each site evolves (independently) according to a continuous Markov chain with highly structured transition probabilities.

   Some background in biology is needed to understand the model. There are four DNA bases, two purines (adenine, A, and guanine, G) and two pyrimidines (cytosine, C, and thymine, T). Replacements of a purine by another purine are called *transitions*, while replacements of a purine by a pyrimidine (or viceversa) are called transversions. In TN93, the matrix of transition rates between the states (A, G, C, T) is given by

$$\begin{pmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \gamma\pi_T \\ \beta\pi_A & \beta\pi_G & \gamma\pi_C & - \end{pmatrix},$$

   where $\beta = [1 - \alpha(\pi_A + \pi_G) - \gamma(\pi_C + \pi_T)]/2$ and $\alpha, \gamma \in [0, 1)$. The probabilities π_A, π_G, π_C and π_T correspond to the average frequencies for each one of the four types of nucleotides (and, therefore, $\pi_A + \pi_G + \pi_C + \pi_T = 1$), and α, β and γ represent purine-to-purine, pyrimidine-to-pyrimidine and purine-to-pyrimidine/pyrimidine-to-purine transition rates, respectively.

   Consider the phylogenetic tree in Figure 1, where the root node represents the original ancestor, nodes are lower levels represent descendants of the node above they are connected to, and the values associated with each branch represent the times elapsed until an speciation event happens.

   (a) Show that the stationary distribution of the continuous time Markov chain is indeed $(\pi_A, \pi_G, \pi_C, \pi_T)$.

   (b) Is the chain time reversible? Does this make sense in terms of the application?

   (c) Assume that the sequence for the root node is:

   ATTCTGACTATTCGCGTAATCTAGGTATCGACCTACTGGATCCGTAT

   Use the TN93 model with parameters $\pi_m = 1/4$ for all m, $\alpha = 0.6$ and $\gamma = 0.73$ to generate descendants for each of the nodes in the graph in Figure 1.

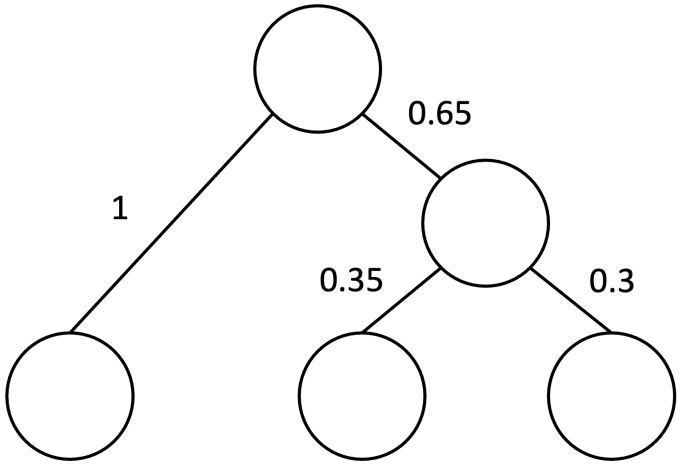

Figure 1: A simple evolutionary tree.

(d) Write an algorithm to compute the the maximum likelihood estimator of the parameters π_A, π_G, π_C, π_T, α, γ given the sequences at each node and the lengths of the branches. Using the sequences that you just generated as your data, apply the algorithm to compute the estimates.

(e) The previous version of the problem is a major oversimplification. One way in which that is true is that we typically only observed the sequences at the very bottom of the tree (in the leafs), which corresponds to "current" species. Discuss how you would extend the previous inference procedure to this setting. (You do not need to write any code for this, just discuss how the previous algorithm could be extended to this more complicated setting. You can still assume that you know the tree and length of the branches of its branches).

References

Tamura, K. and M. Nei. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526