# Algorithms in Computational Biology, 2021
# Exercise 4 - Evolution

Due date: 13/01/2021

Please submit a tar file containing the code (Python3) and a PDF containing plots and **typed** answers.

## 1 MLE for Branch Length

Suppose we are given two aligned sequences $a_1, ..., a_n$ and $b_1, ..., b_n$ (assume no gaps). In class we derived the following probabilistic model for two single character sequences:

$$P(a \xrightarrow{t} b) = P(X^{t_0+t} = b | X^{t_0} = a) = [e^{tR}]_{a,b}$$

We want to derive the MLE for branch length for the two sequences, e.g:

$$\hat{t} = \arg\max_t \prod_i \pi_{a_i} [e^{t\mathbf{R}}]_{a_i,b_i}$$

Assume R is the Jukes-Cantor rate matrix introduced in class with $\alpha = 1$:

$$\mathbf{R}_{\mathrm{JC}}(\alpha) = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

with transition probabilities:

$$P_{JC}(a \xrightarrow{t} b) = \begin{cases} \frac{1}{4}(1 + 3e^{-4\alpha t}) & a = b \\ \frac{1}{4}(1 - e^{-4\alpha t}) & a \neq b \end{cases}$$

1. Define the sufficient statistics for this likelihood and derive equation for calculating the maximum-likelihood estimator $\hat{t}$ . Assume independence between different positions in the alignment.

2. **Building a sampler for a branch**

   (a) Given distance $t$ and character $a$ build a procedure that samples $b$ from $P_{JC}(a \xrightarrow{t} b)$.

   (b) Use the procedure to generate $N$ samples of $b$ and compare between the actual frequency of $b$ and the predicted frequency (based on the JC model). Repeat the process for $t = 0.04, 0.1, 0.3$ and $N = 10, 100, 1000, 10000$.

   (c) Discuss your results. What are your conclusions about the sampling process?

   Submit a table/graph that summarizes the comparison and your code.

3. **Estimating the evolutionary distance**

   (a) Now you will use the sampler that you wrote to sample a pair of nucleotides $(a, b)$. Write a procedure that given $t$, samples $a$ from a stationary distribution $\pi$(uniform in the case of Jukes Cantor) $b$ using the sampler.

   (b) Use the procedure from the previous item to sample a pair of sequences of length $N$ nucleotides that are distance $t$ from each other (i.e., apply the procedure above $N$ times).

   Repeat this procedure $M$ times, and examine the relation between the "real" $t$ you used in generating the data and the MLE estimate.

   For each value of $t$, visualize the distribution of the estimated distances. Plot the box-plot of estimates for each value of $t = 0.04, 0.1, 0.3$ with $N = 500$ and $M = 100$.

   (c) What are your conclusions about the branch length estimate? How will it affect distance-based methods for tree reconstruction?

# 2 Choosing rate matrices

1. Write a rate matrix $R$ that converges into a **uniform** stationary distribution $\pi$ but is **not reversible**.

2. Write a rate matrix $R$ that converges into a **non-uniform** stationary distribution $\pi$ but is **reversible**.

Explain your rationale in one sentence, and demonstrate by calculating $e^{tR}$, and using the Detailed Balance property.

$$\forall a, b, t : \pi_a P(a \xrightarrow{t} b) = \pi_b P(b \xrightarrow{t} a)$$

# 3 Probabilistic Model of Evolution

Suppose we are given a binary tree T with leaves labeled 1...n and n-1 internal nodes. Assume 2n-1 is the root of the tree and let $\tau = \{t_{ij} | (i \xrightarrow{t_{ij}} j) \in T\}$ be the set of branch lengths.

Let $X_1, X_2, ...X_n, ...X_{2n-1}$ be random variables representing the sequences in the tree. For simplicity, let's assume that all the sequences are of length 1.

We would like to show that if R generates reversible matrices P, then we are allowed to reroot the tree without changing the likelihood.

We can calculate the likelihood using:

$P(X_1, X_2, ...X_{2n-1}) = P(X_{2n-1}) \prod_{(i \to j) \in T} P(X_i \xrightarrow{t_{ij}} X_j)$

(We sample the root and then sample each node given its parent)

where $P(a \xrightarrow{t} b) = [e^{tR}]_{a,b}$ and $P(X_{2n-1} = a) = \pi_a$ is the stationary distribution.

1. Please show that the likelihood can also be written as:

$$P(X_1, X_2, ...X_{2n-1}) = \left[\prod_i \pi_{X_i}\right] \prod_{(i \to j) \in T} \frac{[e^{t_{ij}R}]_{X_i X_j}}{\pi_{X_j}}$$

2. Now assume the underlying Markov process is reversible, and show that re-rooting the tree will not change the joint distribution.