Exercises – Phylogenetics

Exercise Sheet 12 — 18.01.2018

Due: 25.01.2018

Task 1 Modeling Amino Acid Replacements. (4 points)

Assume the alphabet of amino acids $\mathcal{A} = \{C, E, S\}$. Consider the following alignment of sequences x and y:

$$x = \text{S C C S E E E C S S E C C C E C S C E C}$$
$$y = \text{S S C S E E S C S S E C C C C S S C E C}$$

(a) Write down the values for all $m_{i,j}$, all f_i and N. Use them to calculate the transition matrix P.

(b) Obviously (without computing it), P is not calibrated to 1 PAM. Why?

(c) Calculate the score matrix S using P and $\pi_i = f_i$. If you do not have any results from (a) use

$$P = \begin{pmatrix} \frac{14}{17} & \frac{1}{17} & \frac{2}{17} \\ \frac{1}{10} & \frac{4}{5} & \frac{1}{10} \\ \frac{2}{13} & \frac{1}{13} & \frac{10}{13} \end{pmatrix} \text{ and } \pi = \left(\frac{17}{40}, \frac{1}{4}, \frac{13}{40} \right) \, .$$

Task 2 Maximum Likelihood Estimation of Evolutionary Distances. (4 points)

Assume the evolution of the sequences x and y from Task 1 follows an EMP with the following transition matrix:

$$P = \begin{pmatrix} 1 - \frac{1}{2}t & \frac{1}{4}t & \frac{1}{4}t \\ \frac{1}{4}t & 1 - \frac{1}{2}t & \frac{1}{4}t \\ \frac{1}{4}t & \frac{1}{4}t & 1 - \frac{1}{2}t \end{pmatrix}$$

Determine the time t such as to maximize the likelihood $Pr(A, B \,|\, t)$.
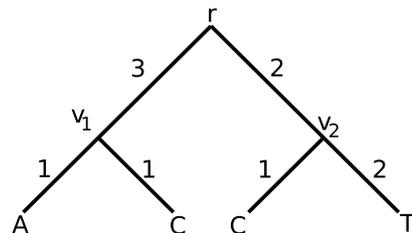
Task 3 Computing the Likelihood of a Given Tree. (4 points)

Consider the given transition probability matrix $P(t)$ and the given tree T. Compute the *likelihood* of T.

$$P(t) = \begin{pmatrix} 1 - 3a_t & a_t & a_t & a_t \\ a_t & 1 - 3a_t & a_t & a_t \\ a_t & a_t & 1 - 3a_t & a_t \\ a_t & a_t & a_t & 1 - 3a_t \end{pmatrix}$$

where

$$a_t = \frac{1 - exp(-4t/30)}{4}$$

Turn over! Bitte wenden!

Task 4 Pulley Principle. (4 points)

The *Pulley Principle* from Joseph Felsenstein says that the likelihood for a phylogenetic tree is independent of the location of the root node.
Consider the simplified scenario of an alphabet $\{A, G\}$ and the given tree.
Show that the likelihood for that tree is independent of the exact location of the root node, i.e., it is only dependent of the **sum** of the lengths: $t_M + t_N$. Proceed as follows:
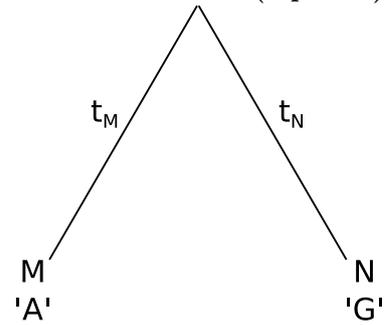
(a) Write down the likelihood in dependence of t_M and t_N. (Example: formular on top of page 82 in the lecture notes.)

(b) Use the assumption that the process of evolution is *reversible* to write the likelihood in a way such that it only contains one π_i (e.g. π_G).

(c) Since P is a stochastic matrix, we can use the *Chapman-Kolmogorov Equation*: $P(t_M + t_N) = P(t_M)P(t_N)$. You can simplify the likelihood with one of the four entries of $P(t_M + t_N)$ such that it is only dependent of the sum $t_M + t_N$.