

# Exercises – Phylogenetics

Universität Bielefeld, WS 2016/2017,  
M. Sc. Tizian Schulz

<https://gi.cebitec.uni-bielefeld.de/Teaching/2016winter/Phylogenetik>

## Exercise Sheet 12 — 31.01.2017

Due: 07.02.2017

### Task 1 Modeling Amino Acid Replacements. (4 points)

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

$$\begin{aligned} x &= R D A A A D R A R R D D R A R D R D R D \\ y &= R D A R A A A D D D A A D R A A A D D A \end{aligned}$$

- Write down the values for all  $m_{i,j}$ , all  $f_i$  and  $N$ . Use them to calculate the transition matrix  $P$ .
- Obviously (without computing it),  $P$  is not calibrated to 1 PAM. Why?
- Calculate the score matrix  $S$  using  $P$  and  $\pi_i = f_i$ . If you do not have any results from (a) use

$$P = \begin{pmatrix} \frac{4}{15} & \frac{2}{5} & \frac{1}{3} \\ \frac{3}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{5}{11} & \frac{4}{11} & \frac{2}{11} \end{pmatrix} \text{ and } \pi = \left( \frac{3}{8}, \frac{7}{20}, \frac{11}{40} \right).$$

### 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:

$$P = \begin{pmatrix} 1 - \frac{4}{6}t & \frac{1}{6}t & \frac{1}{6}t \\ \frac{1}{6}t & 1 - \frac{4}{6}t & \frac{1}{6}t \\ \frac{1}{6}t & \frac{1}{6}t & 1 - \frac{4}{6}t \end{pmatrix}$$

Determine the time  $t$  such as to maximize the likelihood  $Pr(A, B | t)$ .

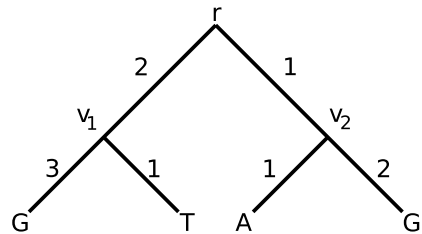
### Task 3 Computing the Likelihood of a Given Tree. (4 points)

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!

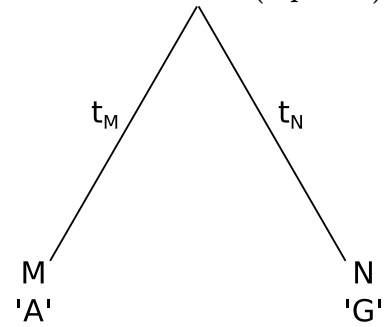
#### Task 4 Pulley Principle.

(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:



- Write down the likelihood in dependence of  $t_M$  and  $t_N$ . (Example: formular on top of page 82 in the lecture notes.)
- Use the assumption that the process of evolution is *reversible* to write the likelihood in a way such that it only contains one  $\pi_i$  (e.g.  $\pi_G$ ).
- 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  $t_M + t_N$ .