Exercise 1: Assume an ancestral sequence and a derived sequence are given, and we aim to estimate the evolutionary distance between the sequences. For this, we assume a Jukes-Cantor model with rate $\lambda = 1$. Let n be the length of the sequences, and k be the number of segregating sites. (There are no gaps in the alignment or the alignment is known and positions with gaps are not counted.)

(a) Calculate the function $f(t)$, which is the expected number of segregating sites if t is the true time distance between the sequences (assuming the Jukes Cantor model and sequence length n).

(b) Calculate the log-likelihood function $\ell_k(t)$, which is the log of the probability to observe k segregating sites if t is the right time (assuming Jukes Cantor etc.)

(c) The so-called *moment estimator* for t is the $\widetilde{t}$ such that $f(\widetilde{t}) = k$. How does it depend on the observed k? (It is called moment estimator because the expectation value is also called the first moment of a distribution.)

(d) As you know, the ML estimator $\hat{t}$ is the t that maximizes $\ell_k(t)$. How does $\hat{t}$ depend on k?

(e) Compare $\widetilde{t}$ to $\hat{t}$. Can you find an obvious relationship between the two? And if so, does it also hold for other substitution models that assume independence between the sites?

Exercise 2: How much do the results of Beast and RAxML depend on the choice of DNA substitution models. Explore this with the primates data and for the lizards data. For the latter, also explore the effect of partioning the data.

Exercise 3: Let nucleotide frequencies $(\pi_A, \pi_C, \pi_G, \pi_T)$ and mutation rates α and β of the HKY model be given. Under which conditions can you find rates λ and μ of the Felsenstein 84 model, such that the transition matrices of the two models are the same, and which are the appropriate values for λ and μ? Discuss also the opposite direction. The F84 rate matrix is:

$$\begin{pmatrix} -\lambda(1-\pi_A) - \frac{\mu\pi_G}{\pi_A+\pi_G} & \lambda\pi_C & \lambda\pi_G + \frac{\mu\pi_G}{\pi_A+\pi_G} & \lambda\pi_T \\ \lambda\pi_A & -\lambda(1-\pi_C) - \frac{\mu\pi_T}{\pi_C+\pi_T} & \lambda\pi_G & \lambda\pi_T + \frac{\mu\pi_T}{\pi_C+\pi_T} \\ \lambda\pi_A + \frac{\mu\pi_A}{\pi_A+\pi_G} & \lambda\pi_C & -\lambda(1-\pi_G) - \frac{\mu\pi_A}{\pi_A+\pi_G} & \lambda\pi_T \\ \lambda\pi_A & \lambda\pi_C + \frac{\mu\pi_C}{\pi_C+\pi_T} & \lambda\pi_G & -\lambda(1-\pi_T) - \frac{\mu\pi_C}{\pi_C+\pi_T} \end{pmatrix}$$

You can insert positive values for the F84 parameters.

Exercise 4 (mainly for biologists): Find publications about at least two different genera, where fossils have been used to time-calibrate phylogenetic trees. Find out in detail which traits of the fossils were used for their taxonomic classification—you may have to trace this back in other publications—and how this was used in the phylogeny analysis.

Exercise 5 (mainly for bioinformaticians): In the lecture we will discuss four different methods to compute a substitution matrix $S(t)$ from a rate matrix R. Investigate for the PAM rate matrix for amino acids, for the pfold rate matrix for RNA stem basepairs and for the following F84 rate matrix how efficient and how accurate and numerically stable these methods are. The PAM rate matrix and the pfold rate matrix will be available from the website of the lecture. (Note that the conventions about matrix notation are not always strictly followed in Bioinformatics. Sometimes transposed matrices are given, i.e. the roles of rows and columns are interchanged. You should always check for this with any matrix that is given to you!)

Exercise 6: Compute the stationary distributions of the pfold rate matrix and the PAM substitution rate matrix. Are these evolutionary dynamics (almost) reversible? (You can use R or a similar program to solve this exercise. In R the function eigen may be helpful.)