Exercises – Lecture 3
Reconstruction of Phylogenetic Trees

Question 1
Question 1a (without `R`*)*
How many edges does a rooted phylogenetic binary tree with m leave nodes have?

Question 1b (without `R`*)*
How many hidden nodes (that is, nodes representing ancestors of present day species) does a rooted phylogenetic binary tree with m leave nodes have?

Question 2
In Lecture 3 we only discuss the Jukes-Cantor substition model. Many more exist. For instance, the so-called Kimura model distinguishes two types of substitutions *transition* and *transversion*. A *transition* is the substitution of one purine by the other (e.g., of `A` by `G`) or of one pyrimidine by the other. A *transversion* refers to the replacement of a purine by a pyrimidine or of a pyrimidine by a purine. The empirical observation that transitions are more likely than transversions motivates the Kimura model. The Kimura substitution model is described by the following transition matrix:

$$\mathbf{P}_{K80} = \begin{pmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{pmatrix}.$$

Note in the rows and columns are now ordered in accordance with `A, G, C, T`. An extension of the Kimura model that is sometimes considered is described by the following transition matrix:

$$\mathbf{P}_{K80ext} = \begin{pmatrix} 1-\alpha-2\gamma & \alpha & \gamma & \gamma \\ \beta & 1-\beta-2\gamma & \gamma & \gamma \\ \delta & \delta & 1-\beta-2\delta & \beta \\ \delta & \delta & \alpha & 1-\alpha-2\delta \end{pmatrix}.$$

Question 2a (without `R`*)*
Check that Kimura model is reversible by showing that it satisfies the detailed balance equations, but its extension is not reversible.

Question 2b (without `R`*)*
Point out that the Jukes-Cantor model is a special case of the Kimura model.

Question 3
Consider a single locus in the DNA of two organisms that share a common ancestor. Assume the Jukes-Cantor substitution model applies to the evolution of this locus.

Question 3a (without `R`*)*
Suppose both organisms are only one generation separated from their common ancestor. Calculate the probability that in both organisms this locus is occupied by a `G`, while the common ancestor also has a `G` at this position.

Question 3b (without `R`*)*
Suppose the organisms are only one generation separated from their common ancestor. Calculate the probability that in both organisms this locus is occupied by a `G`.

Question 3c (without `R`*)*
Suppose the organisms are only one generation separated from their common ancestor. Calculate the probability that, if one organism has a `G` at this locus, so will the other.

Question 3d (without `R`*)*
Suppose both organisms are two generations separated from their common ancestor. Calculate the probability that in both organisms this locus is occupied by a `G`, while the common ancestor also has a `G` at this position.

Question 3e (without `R`*)*
Suppose both organisms are two generations separated from their common ancestor. Calculate the probability that in both organisms this locus is occupied by a `G`.

Question 3f (without `R`*)*
Suppose the organisms are two generations separated from their common ancestor. Calculate the probability that, if one organism has a `G` at this locus, so will the other.

Question 4
Consider a single locus in the DNA of two organisms that share a common ancestor which lived t generations ago. Assume the Jukes-Cantor substitution model applies to the evolution of this locus.

Question 4a (without `R`*)*
Use the spectral decomposition of the Jukes-Cantor substitution model, and calculate the probability that the locus in one organism and the common ancestor is occupied by an `A`.

Question 4b (without `R`*)*
Use the spectral decomposition of the Jukes-Cantor substitution model, and calculate the probability that the locus in the other organism is occupied by a `T`, while in the common ancestor it is occupied by an `A`.

Question 4c (without `R`*)*
Let $\alpha = 1/10000$, where α is the parameter of the Jukes-Cantor substitution model. Give the likelihood of observing the 'data' discussed in Question 4a and 4b, now not knowing the nucleotide of the common ancestor. Maximum this likelihood w.r.t. t to obtain an estimate of the time of separation, t.

Question 4d (without `R`*)*
Do you expect the parameter of Jukes-Cantor model, α, to be higher in a coding or non-coding region of the DNA?

Question 5 *(beginning without, but end with* `R`*)*
Consider a 100 generation experiment of the C. Elegans. Then, we can study the evolution of the DNA sequence of C.Elegans at multiple loci over the generations. Each generation represents a discrete time point.

Consider the following five sequences at five loci over 100 generations:

```
Loci 1: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAA
          AAAAAAAAAAAAAAGGGGGGAAAAAAAAAAAAAAAAACGGGGGGGGGGGGG
Loci 2: TTTTTTTTTTTTTAAAAAAAAAAAAAAGGGGGTAAAAAAAAAAAAAAAAAAA
          AACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTTTTTTTTTTT
Loci 3: GGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAACCCCCCCC
          CCCGGGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
Loci 4: TTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCAAGGGGGGGGGGGGTTTT
          TTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAACGGG
Loci 5: TTTTTTTTTTTGGTTTTTTTTTTTTTTGGGGGGGGGGGAAAAAACCTTTTT
          TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
```

Assume that these loci are independent and follow the same substitution model (apart from the initial condition). Write down the likelihood for the Jukes-Cantor model. First per loci, then the joint likelihood. Hint: the loci are independent. Simplify this expression as much as possible. Maximize the joint likelihood of the data above with respect to the parameters. Hint: recall the exercises on Markov chains of the Lectures 1 and 2.

Question 6 *(without* `R`*)*
In the lecture the likelihood of a phylogenetic tree for two species is discussed. There, the Jukes-Cantor substitution model is assumed. Derive the likelihood and the corresponding estimates for this case, now assuming the Kimura substition model. Hereby use the spectral decomposition of the Kimura model:

$$\mathbf{P}^t_{K80} = \frac{1}{4} + \frac{1}{4}(1-4\,\beta)^t \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix} + \frac{1}{2}(1-2\,\alpha-2\,\beta)^t \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$