Exercises – Lecture 3
Reconstruction of Phylogenetic Trees

Question 1 *(without* `R`*)*

a) How many edges does a rooted phylogenetic binary tree with m leave nodes have?

b) 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 *(without* `R`*)*

Two present day organisms with a binary genetic code (comprising of zero's and ones) have a common ancestor. The substitution process of any locus follows a first order Markov process. The probability of a substitution, irrespective which, equals α.

a) Let $\alpha = 0.10$. One generation separates (i.e. there is no generation inbetween their own generations) the organisms from their common ancestor (i.e. it is only one time step from the common ancestor to the present day organisms). The locus of a present day organism contains a zero. What is the probability that the locus of the common ancestor also contains a zero (ignore the existence of the other present day organism)?

b) All assumptions as with part *a)*. Now consider two independent loci. Both loci of a present day organism contain a zero. What is the probability that the loci of the common ancestor also contain a zero (ignore the existence of the other present day organism)?

c) Let now $\alpha = 0.5$. Assume the number of generations from the present day organisms to the most recent common ancestor unknown. Consider again two independent loci. In one locus the DNA differs between the present day organisms. Give the likelihood for the two observed loci.

Question 3

In Lecture 3 we only discussed a few substitution models. 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:

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

with the following associated stationary distribution:

$$\boldsymbol{\varphi} = \left[\frac{\delta(\beta+\gamma)}{(\gamma+\delta)(\alpha+\beta+2\gamma)}, \frac{\delta(\alpha+\gamma)}{(\gamma+\delta)(\alpha+\beta+2\gamma)}, \frac{\gamma(\alpha+\delta)}{(\gamma+\delta)(\alpha+\beta+2\delta)}, \frac{\gamma(\beta+\delta)}{(\gamma+\delta)(\alpha+\beta+2\delta)}\right]^T.$$

Check that Kimura model is reversible by showing that it satisfies the detailed balance equations. Also assess whether its extension is reversible or not.

Question 4 *(without* `R`*)*
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.

a) Suppose both organisms are only one generation separated from their common ancestor (i.e. it is only one time step from the common ancestor to the present day organisms). 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.

b) Suppose the organisms are only one generation separated from their common ancestor (i.e. it is only one time step from the common ancestor to the present day organisms). Calculate the probability that in both organisms this locus is occupied by a `G`.

c) Suppose the organisms are only one generation separated from their common ancestor (i.e. it is only one time step from the common ancestor to the present day organisms). Calculate the probability that, if organism 2 has a `G` at this locus, so will organism 1.

d) Suppose both organisms are two generations separated from their common ancestor (i.e. it is only two time step from the common ancestor to the present day organisms). 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.

e) Suppose both organisms are two generations separated from their common ancestor (i.e. it is only two time step from the common ancestor to the present day organisms). Calculate the probability that in both organisms this locus is occupied by a `G`.

f) Suppose the organisms are two generations separated from their common ancestor (i.e. it is only two time step from the common ancestor to the present day organisms). Calculate the probability that, if one organism has a `G` at this locus, so will the other.

Question 5 *(without* `R`*)*
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.

a) Use the spectral decomposition of the Jukes-Cantor substitution model, and calculate the probability that the locus in one organism is occupied by an `A`, given that the common ancestor also has an `A` there.

b) 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`.

c) Give the likelihood (i.e. the probability) of observing the 'data' discussed in Question 5a and 5b, now not knowing the nucleotide of the common ancestor.
d) Let $\alpha = 1/8$, where α is the parameter of the Jukes-Cantor substitution model. What time of separation of the two organisms is most likely?
e) Do you expect the parameter of Jukes-Cantor model, α, to be higher in a coding or non-coding region of the DNA?

Question 6* *(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
          TTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAACGGG
Loci 5: TTTTTTTTTTGGTTTTTTTTTTTTTTGGGGGGGGGGGAAAAAACCTTTTT
          TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
```

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 7* *(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:

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