1. The following table shows the frequencies of bases at corresponding sites of two 1000 site sequences of DNA from two different taxa.

$S_1 \backslash S_0$	A	G	C	T
A	212	36	6	5
G	41	204	9	8
C	7	4	181	40
T	10	6	34	197

(a) What fraction of sites have undergone a mutation? What fraction of sites have had a transition? What fraction have had a transversion?
There are $212 + 204 + 181 + 197 = 794$ sites that did not mutate so the fraction that mutated is $(1000 - 794)/1000 = 0.206$. Of the 206 that mutated $36 + 41 + 40 + 34 = 151$ had transitions, so the fraction that had transitions is 0.151, and the fraction that had transversions is $(206 - 151)/1000 = 0.054$.

(b) Use the appropriate fractions from above to find the Jukes Cantor distance between these two sequences.
Assuming Jukes-Cantor we take $p = .206$ and get $d_{jc} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = 0.2408$.

(c) Use the appropriate fractions above to find the Kimura 2-parameter distance between these two sequences.
Assuming Kimura 2-parameter we take $p_1 = .151$ and $p_2 = 0.054$.
So $d_{k2} = -\frac{1}{2}\ln(1 - 2p_1 - p_2) - \frac{1}{4}\ln(1 - 2p_2) = 0.2475$.

(d) Explain why one distance is larger than the other by explaining why one model would more accurately take into account back mutations than the other.
The Kimura 2-parameter model gives a higher value since it assumes a higher mutation rate for transitions (0.151 as opposed to $p/3 = 0.206/3 = 0.079$). The higher the rate of mutation the more likely there is to be a back mutation or a double mutation at a site.

2. The Jukes Cantor model, where the probability that a site will change from one generation to the next is α and for which each possible base substitutions occur with equal probability $\alpha/3$, can be modeled with a 2×2 Markov matrix by considering the fraction of sites that are the same after t generations and the fraction of sites that different. Let q_t be the fraction of sites are the same after t generations and p_t be the fraction of sites that are different. Our task is to find how how these fractions change from one generation to the next.

   (a) If a site is unchanged at time t what is the probability that it is still unchanged at time $t+1$.
   This is just $1-\alpha$.

   (b) If is a site is changed at time t what is the probability that it appears to be unchanged at time $t+1$ (Hint: if a site was originally an A at time $t=0$, and is not an A at some time t, what is the probability that it will be an A at time $t+1$.)
   The site would need to change back to a specific base that it started out at. The probability of changing to a specific base is $\alpha/3$. (The probability of changing to any base is α).

   (c) If a site is unchanged at time t what is the probability that is is changed at time $t+1$?
   This is just the probability of a change, given by α.

   (d) If a site is changed at time t what is the probability that it is still changed at time $t+1$. (Hint: if a site was originally an A at time $t=0$, and is not an A at some time t, what is the probability that it will still not be an A at time $t+1$. Remember if it could stay the same or change from something that is not an A to something else that is still not an A.)
   The easiest way of thinking about this is that it is the probability of not changing back to the specific base it started out as. So the answer is $1-(\alpha/3)$. Alternatively it is the probability of staying the same plus the probability of changing to one of the two other bases, which is $(1-\alpha)+(2\alpha/3) = 1-(\alpha/3)$.

   (e) Put your answers to the above together to find the appropriate matrix M for the matrix model.

$$\begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix} = M \begin{pmatrix} q_t \\ p_t \end{pmatrix}.$$

   Based on the above

$$\begin{aligned} q_{t+1} &= (1-\alpha)q_t + \frac{\alpha}{3}p_t \\ p_{t+1} &= \alpha q_t + (1-\frac{\alpha}{3})p_t \end{aligned}$$

   So the matrix is $M = \begin{pmatrix} 1-\alpha & \frac{1}{3}\alpha \\ \alpha & 1-\frac{1}{3}\alpha \end{pmatrix}$

(f) Show that $\vec{v_1} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\vec{v_2} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ are eigenvectors of this matrix, and find the corresponding eigenvalues.

$$M\vec{v}_1 = \begin{pmatrix} 1-\alpha & \frac{1}{3}\alpha \\ \alpha & 1-\frac{1}{3}\alpha \end{pmatrix}\begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1-\alpha+3\frac{1}{3}\alpha \\ \alpha+3(1-\frac{1}{3}\alpha) \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

So $\lambda_1 = 1$.

$$M\vec{v}_2 = \begin{pmatrix} 1-\alpha & \frac{1}{3}\alpha \\ \alpha & 1-\frac{1}{3}\alpha \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1-\alpha-\frac{1}{3}\alpha \\ \alpha-(1-\frac{1}{3}\alpha) \end{pmatrix} = \begin{pmatrix} 1-\frac{4}{3}\alpha \\ \frac{4}{3}\alpha-1 \end{pmatrix} = (1-\tfrac{4}{3}\alpha)\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

So $\lambda_2 = 1 - \frac{4}{3}\alpha$.

(g) Explain why the initial vector for this matrix model is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and express this vector as a linear combination of the eigenvectors.

Initially, none of the sites have changed so $q_t = 1$ and $p_t = 0$. We want to solve the equation

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = c_1\begin{pmatrix} 1 \\ 3 \end{pmatrix} + c_2\begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

which means that

$$\begin{aligned} 1 &= c_1 + c_2 \\ 0 &= 3c_1 - c_2 \end{aligned}$$

Adding the two equations together gives $1+0 = c_1+3c_1+c_2-c_2 \Rightarrow 4c_1 = 1 \Rightarrow c_1 = \frac{1}{4}$. Thus, based on the second equation, $c_2 = \frac{3}{4}$, so that

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 \\ 3 \end{pmatrix} + \frac{3}{4}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

(h) Write down a solution to the matrix models in terms of eigenvectors, and hence find an expression for p_t, the fraction of sites that have changed. This expression should be the same as the one we described in class.

The solution in vector form is

$$\begin{pmatrix} q_t \\ p_t \end{pmatrix} = \tfrac{1}{4}(1)^t\begin{pmatrix} 1 \\ 3 \end{pmatrix} + \tfrac{3}{4}(1-\tfrac{4}{3}\alpha)^t\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Thus, $q_t = \frac{1}{4} + \frac{3}{4}(1-\frac{4}{3}\alpha)^t$ and $p_t = \frac{3}{4} - \frac{3}{4}(1-\frac{4}{3}\alpha)^t$. Note that $q_t + p_t = 1$ as expected since a site is either changed or not changed!

3. Given the three aligned sequences corresponding to three different taxa

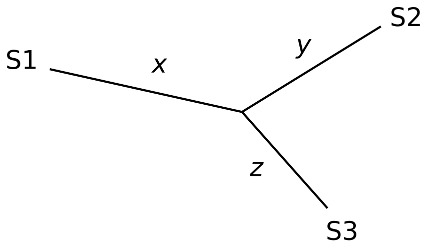
$$\begin{aligned} S1: &\quad GCGCGTTACC \\ S2: &\quad GCGACTTAGG \\ S3: &\quad GCTGTGTACG \end{aligned}$$

(a) Calculate the phylogenetic distances $d_{jc}(S1, S2)$, $d_{jc}(S1, S3)$ and $d_{jc}(S2, S3)$

For $d_{jc}(S1, S2)$, there are 4 differences out of ten so $p = 0.4$. Hence $d = -\frac{3}{4}\ln(1 - \frac{4}{3}(0.4)) = 0.57$. For $d_{jc}(S1, S3)$, there are 6 differences out of ten so $p = 0.6$. Hence $d = -\frac{3}{4}\ln(1 - \frac{4}{3}(0.6)) = 1.21$. For $d_{jc}(S2, S3)$, there are 5 differences out of ten so $p = 0.5$. Hence $d = -\frac{3}{4}\ln(1 - \frac{4}{3}(0.5)) = 0.82$.

(b) Based on these distances, which two tax are most closely related?
Sequences 1 and 2 have the shortest phyllogenetic distance so we take them as being most closely related.

(c) Below is an unrooted phylogenetic tree connecting these taxa. Your task is to compute the phylogenetic distance x, y and z on each branch, based on the distances between the taxa that you calculated in part (a).



$$\begin{aligned} x + y &= d_{jc}(S1, S2) = 0.57 \quad (1) \\ x + z &= d_{jc}(S1, S3) = 1.21 \quad (2) \\ y + z &= d_{jc}(S2, S3) = 0.82 \quad (3) \end{aligned}$$

If we take equation (2)-(1) we get $z - y = 1.21 - 0.57 = 0.64$. Now if we add this equation to equation (3) we get $2z = 0.64 + 0.82 = 1.46$, so that $z = 0.73$. Then equation (3) gives us $y + 0.73 = 0.82$, so that $y = 0.09$. Equation (1) or (2) both lead to $x = 0.48$.