Solutions to Homework 4

Collaboration is allowed on this homework. You must hand in homework assignments individually. You may not use an alignment program to calculate the alignments on this problem set.

Your name:

List the names of the people you worked with.

*Homework must be submitted by 3pm in MI646.*
1. Consider two DNA sequences, $s_1$ and $s_2$, that are diverging from a common ancestor that lived $t = 2.5$ million years ago. Suppose that the sequences are evolving according to the Jukes Cantor model with rate $\alpha = 1 \cdot 10^{-4}$ substitutions per site per million years.

(a) In a pairwise ungapped alignment of $s_1$ and $s_2$, you observe that the base in sequence $s_1$ at site $i$ in the alignment is a $G$. The nucleotide at the same site in sequence $s_2$ is an $A$. What is the probability of this event? Show your work, including any equations you used to solve this problem.

This is the probability of observing $C$ aligned with $G$, considering all possible nucleotides at site $i$ in the ancestral sequence.

$$P_{all} = \sum_{z \in \{A,C,G,T\}} p_z P_{zG}(\alpha,t) \cdot P_{zA}(\alpha,t)$$

$$= 2 \cdot \frac{1}{4} \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right)^2 + 2 \cdot \frac{1}{4} \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \cdot \left( \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right)$$

$$= 1.25 \cdot 10^{-4}$$

(b) What is the probability that the nucleotide at site $i$ in the ancestral sequence was a $C$? Show your work, including any equations you used to solve this problem.

$$P_C = \frac{p_C P_{CG}(\alpha,t) \cdot P_{CA}(\alpha,t)}{\sum_{z \in \{A,C,G,T\}} p_z P_{zG}(\alpha,t) \cdot P_{zA}(\alpha,t)}$$

$$= \frac{\left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right)^2}{2 \cdot \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right)^2 + 2 \cdot \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \cdot \left( \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right) + 1.56 \cdot 10^{-8}}$$

$$= \frac{1.25 \cdot 10^{-4}}{1.25 \cdot 10^{-4}}$$

$$\approx 1.25 \cdot 10^{-4}$$
(c) What is the probability that the ancestral nucleotide at site $i$ was a $G$? Show your work, including any equations you used to solve this problem.

\[
P_G = \frac{p_G P_{GA}(\alpha, t) \cdot P_{GG}(\alpha, t)}{\sum_{z \in \{A,C,G,T\}} p_z P_{zG}(\alpha, t) \cdot P_{zA}(\alpha, t)}
\]

\[
= \frac{(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t}) \cdot (\frac{1}{4} + \frac{3}{4} e^{-4\alpha t})}{2 \cdot (\frac{1}{4} - \frac{1}{4} e^{-4\alpha t})^2 + 2 \cdot (\frac{1}{4} - \frac{1}{4} e^{-4\alpha t}) \cdot (\frac{1}{4} + \frac{3}{4} e^{-4\alpha t})}
\]

\[
= \frac{6.24 \cdot 10^{-5}}{1.25 \cdot 10^{-4}}
\]

\[
\approx 0.5
\]

(d) Suppose you select a site at random in the alignment of $s_1$ and $s_2$. What is the probability of seeing two different nucleotides at that site in the alignment? Show your work, including any equations you used to solve this problem.

\[
P_m(\alpha, t) = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}
\]

\[
P_m(0.0001, 2.5) = 1.5 \cdot 10^{-3}
\]
2. *Estimating substitution distance 1:* In class, we discussed how to use the Jukes-Cantor model to estimate the number of substitutions that occurred during the divergence of a pair of related sequences, given the number of mismatches we observe in the alignment of those sequences. The number of substitutions can also be estimated using the Kimura-2-parameter model, which accounts for two types of substitutions, transitions and transversions. The equations for this estimate are given in the class notes.

You are given an ungapped alignment of two sequences, each of length 1400, with 110 transitions and 30 transversions.

(a) Using the Kimura-2-parameter model, estimate the expected number of substitutions (of all types) that occurred since the divergence of these sequences. Show your work, including any equations you used to solve this problem.

Given that \( n = 1400 \), \( m_s = 110 \), and \( m_v = 30 \),

\[
\begin{align*}
  d &= -1400 \cdot \left[ 0.5 \ln\left(1 - 2 \cdot \frac{110}{1400} - \frac{30}{1400}\right) + 0.25 \ln\left(1 - 2 \cdot \frac{30}{1400}\right) \right] \\
  &= 153.028
\end{align*}
\]

*The expected number of substitutions is 153.*

(b) Suppose you use the Jukes-Cantor model instead of the Kimura-2-parameter model to estimate the substitution distance. Calculate the expected number of substitutions that occurred since the divergence of these sequences. Show your work, including any equations you used to solve this problem.

The substitution distance between two sequences according to the Jukes-Cantor model is

\[
  d = n \cdot -0.75 \ln\left(1 - \frac{4m}{3n}\right)
\]

where \( m \) is the number of mismatches. Therefore,

\[
\begin{align*}
  d &= 1400 \cdot -0.75 \ln\left(1 - \frac{0.4}{3}\right) \\
  &= 150.256
\end{align*}
\]

*The expected number of substitutions estimated with the Jukes-Cantor model is 150. Compared to the Kimura 2-parameter model, the Jukes-Cantor model underestimates the distance between the sequences by 1.8%.*
3. *Estimating substitution distance 2*: Suppose, instead, that you are given an alignment of two sequences, each of length 1400, with 220 transitions and 60 transversions.

(a) Using the Kimura 2-parameter model, estimate the expected number of substitutions (of all types) that occurred since the divergence of these sequences. Show your work, including any equations you used to solve this problem.

\[
d = -1400 \cdot \left[ 0.5 \ln\left(1 - 2 \frac{220}{1400} - \frac{60}{1400}\right) + 0.25 \ln\left(1 - 2 \frac{60}{1400}\right) \right] \\
= 340.65
\]

*The expected number of substitutions is 341.*

(b) Using the Jukes-Cantor model instead of the Kimura-2-parameter model, calculate the expected number of substitutions that occurred since the divergence of these sequences. Show your work, including any equations you used to solve this problem.

*The expected number of substitutions estimated by the Jukes-Cantor model is*

\[
d = 1400 \cdot -0.75 \ln\left(1 - \frac{0.8}{3}\right) \\
= 325.66 \\
\approx 326.
\]

*The Jukes-Cantor model underestimates the substitution distance between the sequences by 4.4%.*
(c) What is the impact of using the Jukes-Cantor model, instead of the Kimura-2-parameter model, on the accuracy of the estimate of the substitution distance? How does the accuracy of the estimate depend on the amount of change that has occurred since divergence of $s_1$ and $s_2$? Why?

The K2P model accounts for different substitution rates for transitions and transversions, while the Jukes-Cantor model is based on the assumption that all substitutions occur at the same rate. In these examples, there are many fewer transversions than transitions, suggesting that the rates are different. In this case, the Jukes-Cantor model underestimates the substitution distance.

Mismatches account for 10% of sites in Problem 2 and 20% of sites (i.e., twice as many) in Problem 3. The underestimate is greater in Problem 3, suggesting that as the number of substitutions increases, so does the importance of modeling separate rates for transitions and transversions.

4. The Felsenstein 1981 (F81) substitution model, described in Section 2.3.2, allows for unequal base frequencies under the stationary distribution.

(a) Draw the graphical representation of the F81 Markov model. Show the states and arcs, including self-arcs. Label each arc with an expression for its transition probability, given in terms of the parameters of the model.
(b) The following matrix is an instance of the F81 model. Calculate the stationary frequencies of the four nucleotides and the value of the parameter, \( \alpha \). Show your work.

\[
\begin{array}{cccc}
A & G & C & T \\
0.675 & 0.1 & 0.1 & 0.125 \\
0.175 & 0.6 & 0.1 & 0.125 \\
0.175 & 0.1 & 0.6 & 0.125 \\
0.175 & 0.1 & 0.1 & 0.625 \\
\end{array}
\]

The general Felsenstein matrix is:

\[
\begin{bmatrix}
A & 1 - \alpha (\varphi^*_C + \varphi^*_G + \varphi^*_T) \\
G & \varphi^*_A \alpha & \varphi^*_A \alpha & \varphi^*_A \alpha \\
C & \varphi^*_A \alpha & 1 - \alpha (\varphi^*_A + \varphi^*_C + \varphi^*_T) & \varphi^*_C \alpha & \varphi^*_C \alpha \\
T & \varphi^*_A \alpha & \varphi^*_G \alpha & \varphi^*_T \alpha & 1 - \alpha (\varphi^*_A + \varphi^*_C + \varphi^*_G) \\
\end{bmatrix}
\]

You can solve this problem in two ways. One approach is to solve the system of equations, \( \varphi^* = \varphi^*P \), to find the stationary frequencies. The stationary distribution is \( \varphi^* = \{0.35, 0.2, 0.2, 0.25\} \).

Next, use any off-diagonal entry to determine \( \alpha \). For example, once you know the value of \( \varphi^*_A \), you can solve \( \varphi^*_A \alpha = 0.175 \) to determine that \( \alpha = 0.5 \).

Alternatively, you can solve for \( \alpha \) first by observing that

\[
\alpha \varphi^*_A + \alpha \varphi^*_G + \alpha \varphi^*_C + \alpha \varphi^*_T = 0.175 + 0.1 + 0.1 + 0.125
\]

However, since \( \varphi^*_A + \varphi^*_G + \varphi^*_C + \varphi^*_T = 1 \), this yields

\[
\alpha = 0.175 + 0.1 + 0.1 + 0.125 = 0.5.
\]

Now you can find the stationary distribution by solving \( \varphi^*_A \cdot 0.5 = 0.175 \) to determine the value of \( \varphi^*_A \). The values of \( \varphi^*_G, \varphi^*_C \), and \( \varphi^*_T \) can be determined similarly.
(c) Substitute the values of $\alpha$, $\varphi_A$, $\varphi_G$, $\varphi_C$, and $\varphi_T$ that you derived in part (b) into the expressions on the self-arcs in the model in part (a). Verify that the values on the main diagonal of the above matrix are consistent with the specification of the Felsenstein model.

$$1 - \alpha (\varphi_C^* + \varphi_G^* + \varphi_T^*) =$$
$$1 - 0.5 \cdot (0.2 + 0.2 + 0.25) = 0.675$$

$$1 - \alpha (\varphi_A^* + \varphi_C^* + \varphi_T^*) =$$
$$1 - 0.5 \cdot (0.35 + 0.2 + 0.25) = 0.6$$

$$1 - \alpha (\varphi_A^* + \varphi_G^* + \varphi_T^*) =$$
$$1 - 0.5 \cdot (0.35 + 0.2 + 0.25) = 0.6$$

$$1 - \alpha (\varphi_A^* + \varphi_C^* + \varphi_G^*) =$$
$$1 - 0.5 \cdot (0.35 + 0.2 + 0.2) = 0.625$$