Solution Set 3

Due 11:59 pm, Saturday, October 2nd

Collaboration is allowed on this homework. You may discuss the problems with your colleagues, but each student must prepare and submit a separate assignment. Please list the names of the people you worked with:

Please provide answers to all questions. In order to obtain full credit, explain your reasoning on each question and show all intermediate steps leading to your solution.

You may submit your assignments in any of the following ways:

- Download the assignment in pdf format and print it out. Write your solutions in the space provided. Scan your handwritten solution and upload it to Canvas. Attach additional pages, if needed.
- Download the assignment in pdf format. Use the commenting features in adobe or a similar tool to enter your solution directly on the pdf. Upload the completed assignment, annotated with your solutions.
- Download the assignment in latex format. Enter your solutions in latex format, compile the assignment, and turn in the resulting pdf.

1. Consider two DNA sequences, s_1 and s_2, that are diverging from a common ancestor that lived t million years ago. Suppose that the sequences are evolving according to the Jukes Cantor model with rate $\alpha = 0.01$ substitutions per site per million years and that s_1 and s_2 diverged $t = 25$ million years ago.

   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 an `A`. The nucleotide at the same site in sequence s_2 is also an `A`. For all questions, show your work, including any equations you used to solve the problem.

   (a) What is the probability of observing `A` aligned with `A` at site i, if the nucleotide at site i in the ancestral sequence was a `C`?

   This is the probability of substituting a C with an A in both sequences. We do not have to consider the probability of obsserving the C in the ancestral sequence because it is a given.

$$\begin{aligned} P_{\text{given } C} &= P_{CA}(\alpha, t) \cdot P_{CA}(\alpha, t) \\ &= \left(\frac{1}{4} - \frac{1}{4}\, e^{-4\alpha t}\right)^2 \\ &= \left(\frac{1}{4} - \frac{1}{4}\, e^{-4 \cdot 25 \cdot 10^{-2}}\right)^2 \\ &\approx 0.025 \end{aligned}$$

   (b) What is the probability of observing `A` aligned with `A` at site i, if the nucleotide at site i in the ancestral sequence was also an `A`?

$$\begin{aligned} P_{\text{given } A} &= P_{AA}(\alpha, t) \cdot P_{AA}(\alpha, t) \\ &= \left(\frac{1}{4} + \frac{3}{4}\, e^{-4\alpha t}\right)^2 \\ &= \left(\frac{1}{4} + \frac{3}{4}\, e^{-1}\right)^2 \\ &\approx 0.277 \end{aligned}$$

(c) What is the probability of observing `A` aligned with `A` at site i?

$$\begin{aligned} P_{\text{all}} &= \sum_{z\in\{A,C,G,T\}} p_z P_{zA}(\alpha,t)\cdot P_{zA}(\alpha,t) = \sum_{z\in\{A,C,G,T\}} p_z P_{zA}(\alpha,t)^2 \\ &= \frac{3}{4}\left(\frac{1}{4}-\frac{1}{4}\,e^{-4\alpha t}\right)^2 + \frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}\,e^{-4\alpha t}\right)^2 \\ &= \frac{3}{4}\left(\frac{1}{4}-\frac{1}{4}\,e^{-4\cdot 25\cdot 10^{-2}}\right)^2 + \frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}\,e^{-4\cdot 25\cdot 10^{-2}}\right)^2 \\ &\approx 0.0879 \end{aligned}$$

(d) What is the probability that the nucleotide at site i in the ancestral sequence was a `C`?

$$\begin{aligned} P_{\text{ancestral } C} &= \frac{p_C P_{\text{given } C}}{P_{\text{all}}} = \frac{p_C P_{CA}(\alpha,t)^2}{\sum_{z\in\{A,C,G,T\}} p_z P_{zA}(\alpha,t)^2} \\ &\approx \frac{\frac{1}{4}\cdot 0.025}{0.0879} \\ &\approx 0.0710 \end{aligned}$$

(e) What is the probability that the nucleotide at site i in the ancestral sequence was an `A`?

$$\begin{aligned} P_{\text{ancestral } A} &= \frac{p_A P_{\text{given } A}}{P_{\text{all}}} = \frac{p_A P_{AA}(\alpha,t)^2}{\sum_{z\in\{A,C,G,T\}} p_z P_{zA}(\alpha,t)^2} \\ &\approx \frac{\frac{1}{4}\cdot 0.277}{0.0879} \\ &\approx 0.787 \end{aligned}$$

(f) Suppose that the divergence rate is ten-fold slower; i.e., $\alpha = 0.001$ substitutions per site per million years. As before, assume that $t = 25$ million years and that s_1 and s_2 are evolving according to the Jukes Cantor model. What is the probability of observing `A` aligned with `A` at site i, if the nucleotide at site i in the ancestral sequence was a `C`?

$$\begin{aligned} P_{\text{given } C} &= P_{CA}(\alpha, t) \cdot P_{CA}(\alpha, t) \\ &= \left(\frac{1}{4} - \frac{1}{4}\, e^{-4\alpha t}\right)^2 \\ &= \left(\frac{1}{4} - \frac{1}{4}\, e^{-4 \cdot 25 \cdot 10^{-3}}\right)^2 = \left(\frac{1}{4} - \frac{1}{4}\, e^{-0.1}\right)^2 \\ &\approx 5.66 \times 10^{-4} \end{aligned}$$

(g) At this slower rate, what is the probability of observing `A` aligned with `A` at site i, if the nucleotide at site i in the ancestral sequence was also an `A`?

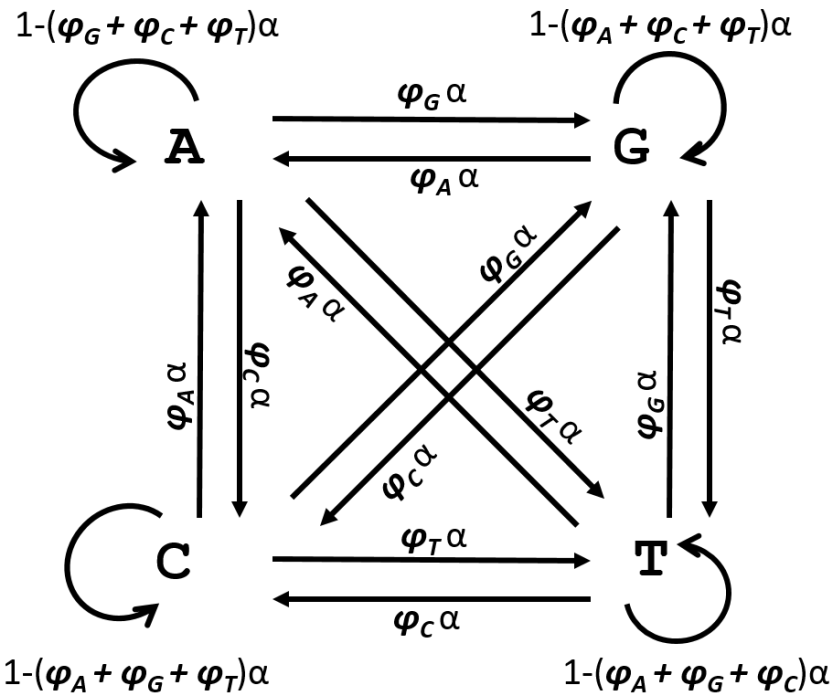
$$\begin{aligned} P_{\text{given } A} &= P_{AA}(\alpha, t) \cdot P_{AA}(\alpha, t) \\ &= \left(\frac{1}{4} + \frac{3}{4}\, e^{-4\alpha t}\right)^2 = \left(\frac{1}{4} + \frac{3}{4}\, e^{-0.1}\right)^2 \\ &\approx 0.862 \end{aligned}$$

(h) At this slower rate, what is the probability of observing `A` aligned with `A` at site i?

$$\begin{aligned} P_{\text{all}} &= \sum_{z \in \{A,C,G,T\}} p_z P_{zA}(\alpha, t) \cdot P_{zA}(\alpha, t) = \sum_{z \in \{A,C,G,T\}} p_z P_{zA}(\alpha, t)^2 \\ &= \frac{3}{4}\left(\frac{1}{4} - \frac{1}{4}\, e^{-4\alpha t}\right)^2 + \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4}\, e^{-4\alpha t}\right)^2 \\ &= \frac{3}{4}\left(\frac{1}{4} - \frac{1}{4}\, e^{-0.1}\right)^2 + \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4}\, e^{-0.1}\right)^2 \\ &\approx 0.216 \end{aligned}$$

2. *The Felsenstein 1981 (F81) substitution model*, described in Section 2.3.3, 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, α. Show your work.

	A	G	C	T
A	0.84	0.05	0.06	0.05
G	0.04	0.85	0.06	0.05
C	0.04	0.05	0.86	0.05
T	0.04	0.05	0.06	0.85

The general Felsentstein matrix is:

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\,(\varphi_C^*+\varphi_G^*+\varphi_T^*) & \alpha\,\varphi_G^* & \alpha\,\varphi_C^* & \alpha\,\varphi_T^* \\ G & \alpha\,\varphi_A^* & 1-\alpha\,(\varphi_A^*+\varphi_C^*+\varphi_T^*) & \alpha\,\varphi_C^* & \alpha\,\varphi_T^* \\ C & \alpha\,\varphi_A^* & \alpha\,\varphi_G^* & 1-\alpha\,(\varphi_A^*+\varphi_G^*+\varphi_T^*) & \alpha\,\varphi_T^* \\ T & \alpha\,\varphi_A^* & \alpha\,\varphi_G^* & \alpha\,\varphi_C^* & 1-\alpha\,(\varphi_A^*+\varphi_C^*+\varphi_G^*) \end{bmatrix}$$

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 $\varphi^* = \{0.2, 0.25, 0.3, 0.25\}$.
Next, use any off-diagonal entry to determine α. For example, once you know the value of φ_A^*, you can solve $\varphi_A^*\alpha = 0.04$ to determine that $\alpha = 0.2$.
Alternatively, you can solve for α first by observing that since

$$\varphi_A^*+\varphi_G^*+\varphi_C^*+\varphi_T^* = 1,$$

then

$$\alpha(\varphi_A^*+\varphi_G^*+\varphi_C^*+\varphi_T^*) = \alpha.$$

Therefore, solving

$$\alpha\varphi_A^*+\alpha\varphi_G^*+\alpha\varphi_C^*+\alpha\varphi_T^* = 0.04+0.05+0.06+0.05$$

yields $\alpha = 0.2$.
Next, you can find the stationary distribution by solving $\varphi_A^* \cdot 0.2 = 0.04$, $\varphi_G^* \cdot 0.2 = 0.05$, $\varphi_C^* \cdot 0.2 = 0.06$, and $\varphi_T^* \cdot 0.2 = 0.05$, yielding $\varphi^* = \{0.2, 0.25, 0.3, 0.25\}$.

(c) Substitute the values of α, φ_A, φ_G, φ_C, and φ_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.

$$\begin{aligned} 1-\alpha\,(\varphi_C^*+\varphi_G^*+\varphi_T^*) &= \\ 1-0.2\cdot(0.25+0.3+0.25) &= 0.84 \end{aligned}$$

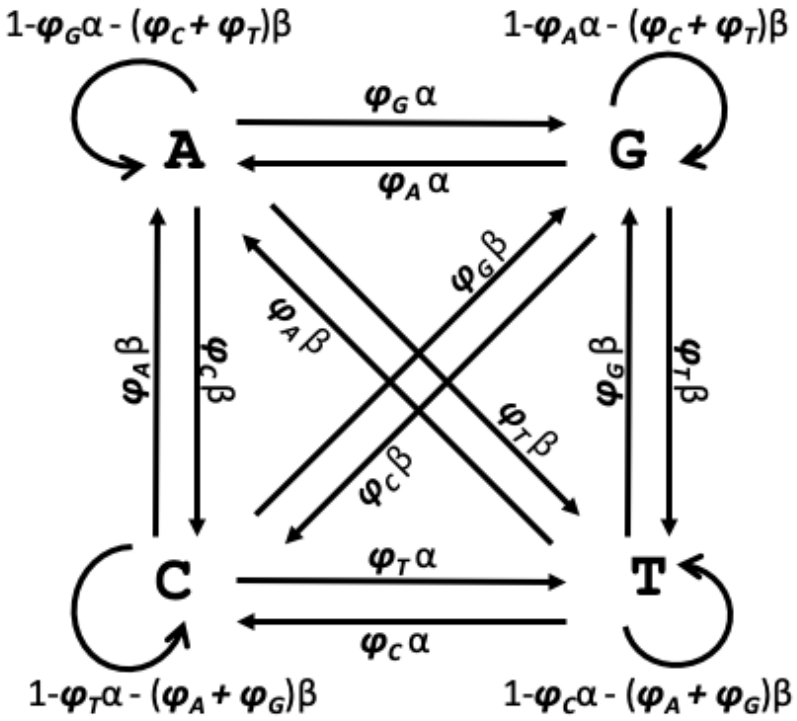
$$\begin{aligned} 1-\alpha\,(\varphi_A^*+\varphi_C^*+\varphi_T^*) &= \\ 1-0.2\cdot(0.2+0.3+0.25) &= 0.85 \end{aligned}$$

$$\begin{aligned} 1-\alpha\,(\varphi_A^*+\varphi_G^*+\varphi_T^*) &= \\ 1-0.2\cdot(0.2+0.25+0.25) &= 0.86 \end{aligned}$$

$$\begin{aligned} 1-\alpha\,(\varphi_A^*+\varphi_C^*+\varphi_G^*) &= \\ 1-0.2\cdot(0.2+0.25+0.3) &= 0.85 \end{aligned}$$

3. *The Hasegawa, Kishino, Yano (HKY) model substitution model*, described in Section 2.3.4, uses different rates for transitions and transversions and allows for unequal base frequencies under the stationary distribution.

   (a) Draw the graphical representation of the HKY 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 HKY model. Calculate the stationary frequencies of the four nucleotides and the values of the parameters α and β. Show your work.

	A	G	C	T
A	0.895	0.05	0.03	0.025
G	0.04	0.905	0.03	0.025
C	0.02	0.025	0.905	0.05
T	0.02	0.025	0.06	0.895

The general HKY matrix is:

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\,\varphi_G^*-\beta\,(\varphi_C^*+\varphi_T^*) & \alpha\,\varphi_G^* & \beta\,\varphi_C^* & \beta\,\varphi_T^* \\ G & \alpha\,\varphi_A^* & 1-\alpha\,\varphi_A^*-\beta\,(\varphi_C^*+\varphi_T^*) & \beta\,\varphi_C^* & \beta\,\varphi_T^* \\ C & \beta\,\varphi_A^* & \beta\,\varphi_G^* & 1-\alpha\,\varphi_T^*-\beta\,(\varphi_A^*+\varphi_G^*) & \alpha\,\varphi_T^* \\ T & \beta\,\varphi_A^* & \beta\,\varphi_G^* & \alpha\,\varphi_C^* & 1-\alpha\,\varphi_C^*-\beta\,(\varphi_A^*+\varphi_G^*) \end{bmatrix}$$

You can solve this problem as for problem 2b. Solving the system of equations, $\varphi^* = \varphi^* P$ *yields the stationary distribution* $\varphi^* = \{0.2, 0.25, 0.3, 0.25\}$. *Then, use a transition entry (A to G or C to T, i.e.) to determine* α, *and a transversion entry to determine* β. *For example, solving* $\varphi_G^*\alpha = 0.05$ *and* $\varphi_C^*\beta = 0.025$ *yields* $\alpha = 0.2$ *and* $\beta = 0.1$.
Alternatively, you can solve for α *and* β *by solving for*

$$\alpha(\varphi_A^*+\varphi_G^*+\varphi_C^*+\varphi_T^*) = \alpha$$

and

$$\beta(\varphi_A^*+\varphi_G^*+\varphi_C^*+\varphi_T^*) = \beta.$$

This yields

$$\alpha\varphi_A^*+\alpha\varphi_G^*+\alpha\varphi_C^*+\alpha\varphi_T^* = 0.04+0.05+0.06+0.05$$

and

$$\beta\varphi_A^*+\beta\varphi_G^*+\beta\varphi_C^*+\beta\varphi_T^* = 0.02+0.025+0.03+0.025$$

solving to $\alpha = 0.2$ *and* $\beta = 0.1$.
Next, you can find the stationary distribution by solving $\varphi_A^* \cdot \alpha = \varphi_A^* \cdot 0.2 = 0.04$, $\varphi_G^* \cdot \alpha = \varphi_G^* \cdot 0.2 = 0.05$, $\varphi_C^* \cdot \beta = \varphi_C^* \cdot 0.1 = 0.03$, *and* $\varphi_T^* \cdot \beta = \varphi_T^* \cdot 0.1 = 0.025$, *yielding* $\varphi^* = \{0.2, 0.25, 0.3, 0.25\}$.

(c) Substitute the values of α, β, φ_A, φ_G, φ_C, and φ_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 HKY model.

$$\begin{aligned} 1-\alpha\,\varphi_G^*-\beta\,(\varphi_C^*+\varphi_T^*) &= \\ 1-0.2\cdot 0.25-0.1\cdot(0.3+0.25) &= 0.895 \end{aligned}$$

$$\begin{aligned} 1-\alpha\,\varphi_A^*-\beta\,(\varphi_C^*+\varphi_T^*) &= \\ 1-0.2\cdot 0.2-0.1\cdot(0.3+0.25) &= 0.905 \end{aligned}$$

$$\begin{aligned} 1-\alpha\,\varphi_T^*-\beta\,(\varphi_A^*+\varphi_G^*) &= \\ 1-0.2\cdot 0.25-0.1\cdot(0.2+0.25) &= 0.905 \end{aligned}$$

$$\begin{aligned} 1-\alpha\,\varphi_C^*-\beta\,(\varphi_A^*+\varphi_G^*) &= \\ 1-0.2\cdot 0.3-0.1\cdot(0.2+0.25) &= 0.895 \end{aligned}$$