Problem Set 3

Due 11:5-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`?

   (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`?

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

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

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

(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`?

(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`?

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

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

(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.

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

(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.