Problem Set 3

Due 11:59pm, October 24th

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.

Homework must be submitted electronically by midnight

1. Your goal is to devise a scoring system for ungapped alignments that will allow you to distinguish between pairs of sequences that are related and pairs of sequences that have chance similarities. You propose to derive a scoring scheme that is parameterized by evolutionary divergence, using the likelihood ratio framework we discussed in class.

   Given DNA sequences s_1 and s_2 of length n, your alternate hypothesis, H_A, is that s_1 and s_2 diverged from a common ancestor t million years ago and are evolving according to the Jukes Cantor model with parameter α. According to your null hypothesis, H_0, the probability of observing x aligned with y is the product of their background frequencies, $p_x \cdot p_y$. In this framework, your scoring system will have one score, $M(\alpha, t)$, for all matches and one score, $m(\alpha, t)$, for all mismatches.

   (a) Give an expression in terms of α and t for $P(M|H_A)$, the likelihood of observing a match in a pair of related sequences.

   (b) Give an expression in terms of α and t for $P(m|H_A)$, the likelihood of observing a mismatch in a pair of related sequences.

(c) Give an expression for $P(M|H_0)$, the likelihood of observing a match by chance, in terms of the background frequencies p_A, p_C, p_G, and p_T.

(d) Calculate the numerical value of $P(M|H_0)$, the probability of observing a match by chance, under the assumption that the nucleotide frequencies in the genomes of interest correspond to the stationary distribution of the Jukes Cantor model.

(e) Calculate the numerical value of $P(m|H_0)$, the probability of observing a mismatch by chance, under the assumption that the nucleotide frequencies in the genomes of interest correspond to the stationary distribution of the Jukes Cantor model.

(f) You define your match score, $M(\alpha, t)$, to be the log of the ratio of the probabilities of a match under the alternate and null hypotheses. Give an expression for $M(\alpha, t)$ under the assumption that the nucleotide frequencies in the genomes of interest correspond to the stationary distribution of the Jukes Cantor model. *Use log base 2.*

(g) You define your mismatch score, $m(\alpha, t)$, to be the log of the ratio of the probabilities of a mismatch under the alternate and null hypotheses. Give an expression for $m(\alpha, t)$ under the assumption that the nucleotide frequencies in the genomes of interest correspond to the stationary distribution of the Jukes Cantor model. *Use log base 2.*

(h) Suppose that the ungapped alignment of s_1 and s_2 is 100 nucleotides long and contains 45 matches. Give an expression for the score of this ungapped alignment in terms of α, and t.

(i) What is the numerical score of this ungapped alignment if $\alpha = 0.002$ per site per million years and the sequences diverged 30 million years ago? Based on this score, do you think the similarity of the sequences indicates common ancestry or chance similarity? Why?

(j) What is the numerical score of this ungapped alignment if $\alpha = 0.002$ per site per million years and the sequences diverged 60 million years ago? Based on this score, do you think the similarity of the sequences indicates common ancestry or chance similarity? Why?

2. In many sequences, different sites evolve at different rates. This question demonstrates the nuances of modeling rate variation.

   Consider the ungapped alignment of a pair of DNA sequences, s_1 and s_2, of the same length that are are descended from a common ancestor that lived t million years ago.

   Suppose that half of the sites in this alignment are changing at a rate of $\alpha_1 = 8 \cdot 10^{-4}$ substitutions per site per million years, while the other half of the sites are changing at a rate of $\alpha_2 = 5\alpha_1$ substitutions per site per million years.

   (a) In this sequence pair, the frequency of mismatches, $\hat{P}_m$, is the average of the frequency of mismatches at the fast sites and the frequency of mismatches at the slow sites. Give an expression for $\hat{P}_m$ as a function of α_1 and t.

   (b) Suppose we incorrectly assume that all sites are following a Jukes-Cantor model of change with the same parameter $\alpha = (\alpha_1 + \alpha_2)/2$. Give an expression for P_m, the frequency of mismatches under this assumption, as a function of α_1 and t.

(c) Use your favorite plotting program to plot $\hat{P}_m$ and P_m as a function of t in units of millions of years, where t varies from 0 to 600.

(d) Plot the ratio $\hat{P}_m/P_m$ as a function of t, where t varies from 0 to 600.

(e) What is the impact on your estmate of the mismatch frequency of ignoring rate variation and using the average rate instead? How does this effect vary with t? Why is the impact higher at some values of t than at others?

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

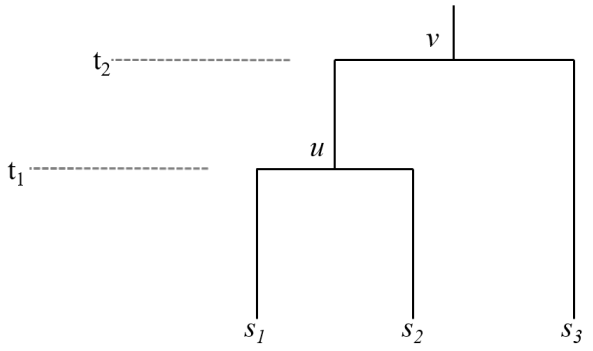
   (a) Draw the graphical representation of the HKY Markov model. Show the states and arcs, including self-arcs. Label each arc with an expression for its transition probability, given in terms of the parameters of the model.

(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.

4. Consider three DNA sequences, s_1, s_2, and s_3. Sequences s_1 and s_2 diverged from sequence u in a common ancestor that lived t_1 million years ago. Sequence s_3 and u diverged from sequence v in a common ancestor that lived t_2 million years ago. As illustrated in the figure below. These sequences are evolving according to the Kimura 2 parameter model with rate transition rate α and transversion rate β substitutions per site per million years.



Derive an algebraic expression for the probability of the event that the bases at site i in sequences s_1, s_2 and s_3 are `C`, `G`, and `A`, respectively, and the bases at site i in both ancestral sequences are `A`. You do not need to simplify this expression algebraically.

5. PAM theory

   (a) Is the PAM-1 transition matrix symmetric? Justify your answer algebraically.

   (b) Is the PAM-1 log odds scoring matrix symmetric? Justify your answer algebraically.

(c) Given a pair of sequences with one PAM of divergence, 99 out of 100 positions should be identical. Verify that $\sum_i p_i P^1[i,i] = 0.99$

(d) Verify that the rows of the PAM-1 transition matrix sum to one.