Solution Set 4

Due 11:59pm, Friday, October 8th

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. 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 values for α and t, your scoring system will have one score, $M(\alpha, t)$, for all matches and one score, $m(\alpha, t)$, for all mismatches.

   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, $p_x \cdot p_y$.

   (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.

$$\begin{aligned}
\Pr(M|H_A) &= \sum_{x \in \{A,G,C,T\}} \Pr\left(\begin{matrix} x \\ x \end{matrix} \mid t, \alpha\right) \\
\Pr(M|\alpha, t) &= 4 \cdot \frac{1}{16} \left(1 + 3e^{-8\alpha t}\right) \\
&= \tfrac{1}{4}\left(1 + 3e^{-8\alpha t}\right)
\end{aligned}$$

   -OR-

$$\begin{aligned}
\Pr(M|H_A) &= \sum_{x \in \{A,G,C,T\}} \sum_{z \in \{A,G,C,T\}} p_z \cdot p_{zx}(t)^2 \\
\Pr(M|\alpha, t) &= \sum_{x \in \{A,G,C,T\}} \tfrac{1}{4}\left(p_{Ax}(t)^2 + p_{Gx}(t)^2 + p_{Cx}(t)^2 + p_{Tx}(t)^2\right) \\
&= p_{Ax}(t)^2 + p_{Gx}(t)^2 + p_{Cx}(t)^2 + p_{Tx}(t)^2 \\
&= \left[\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right]^2 + 3\left[\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right]^2 \\
&= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}
\end{aligned}$$

(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.

$$\begin{aligned}
\Pr(m|H_A) &= 1 - \Pr(M|H_A) \\
\Pr(m|\alpha, t) &= 1 - \tfrac{1}{4}\left(1 + 3e^{-8\alpha t}\right) \\
&= \frac{3}{4} - \frac{3}{4}e^{-8\alpha t}
\end{aligned}$$

-OR-

$$\begin{aligned}
\Pr(m|H_A) &= \sum_{x \in \{A,G,C,T\}} \sum_{y \neq x} \Pr\left(\begin{matrix} x \\ y \end{matrix} \mid t, \alpha\right) \\
\Pr(m|\alpha, t) &= 12 \cdot \frac{1}{16}\left(1 - e^{-8\alpha t}\right) \\
&= \tfrac{3}{4}\left(1 - e^{-8\alpha t}\right)
\end{aligned}$$

(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.

$$\begin{aligned}
\Pr(M|H_0) &= \Pr\left(\begin{matrix} x \\ x \end{matrix} \text{ randomly}\right) \\
\Pr(M|H_0) &= \sum_x p_x p_x \\
\Pr(M|H_0) &= p_A^2 + p_C^2 + p_G^2 + p_T^2.
\end{aligned}$$

(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.

At steady state, $p_A = p_C = p_G = p_T = \frac{1}{4}$. *Therefore,*

$$\Pr(M|H_0) \;=\; 4 \cdot (\tfrac{1}{4})^2 \;=\; \tfrac{1}{4}.$$

(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.

$$\Pr(m|H_0) \;=\; 1 - \Pr(M|H_0) \;=\; \tfrac{3}{4}.$$

(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.*

$$\begin{aligned} M(\alpha,t) &= \log_2 \frac{\Pr(M|H_A)}{\Pr(M|H_0)} \\ &= \log_2 \frac{\frac{1}{4} + \frac{3}{4}e^{-8\alpha t}}{\frac{1}{4}} \\ &= \log_2[1 + 3e^{-8\alpha t}] \end{aligned}$$

(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.*

$$\begin{aligned} m(\alpha,t) &= \log_2 \frac{\Pr(m|H_A)}{\Pr(m|H_0)} \\ &= \log_2 \frac{\frac{3}{4} - \frac{3}{4}e^{-8\alpha t}}{\frac{3}{4}} \\ &= \log_2[1 - e^{-8\alpha t}] \end{aligned}$$

(h) Suppose that the ungapped alignment of s_1 and s_2 is 100 nucleotides long and contains 55 mismatches. Give an expression for the score of this ungapped alignment in terms of α, and t.

$$\begin{aligned}\mathcal{S} &= \sum_{i=1}^{100} p(s_1[i], s_2[i]) \\ &= 45 \cdot M(\alpha, t) \; + \; 55 \cdot m(\alpha, t) \\ &= 45 \cdot \log_2[1 + 3e^{-8\alpha t}] \; + \; 55 \cdot \log_2[1 - e^{-8\alpha t}]\end{aligned}$$

(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?

$\mathcal{S} = -8.39$. *When $\mathcal{S} < 0$, the null hypothesis is more probable than the alternate hypothesis. The observed similarity is more likely due to chance.*

(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?

$\mathcal{S} = 11.35$. *When $S > 0$, H_A is more likely than H_0. The sequences are more likely to be related.*

Note that interpretation of the data is crucially dependent on the amount of divergence. When $\alpha = 0.002$, after 30 million years a pair of sequences with 45% identity appear to be unrelated; after 60 million years, a pair of sequences with 45% identity is 2000 times more likely to share common ancestry, than chance similarity.

2. Consider a pair of DNA sequences that are evolving according to the Jukes-Cantor model. The residues in these sequences are changing at two different rates: Half the sites change at $\alpha_1 = 8 \cdot 10^{-4}$ substitutions per site per million years, while the other half of the sites are changing five times as fast ($\alpha_2 = 5\alpha_1$).

   (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.

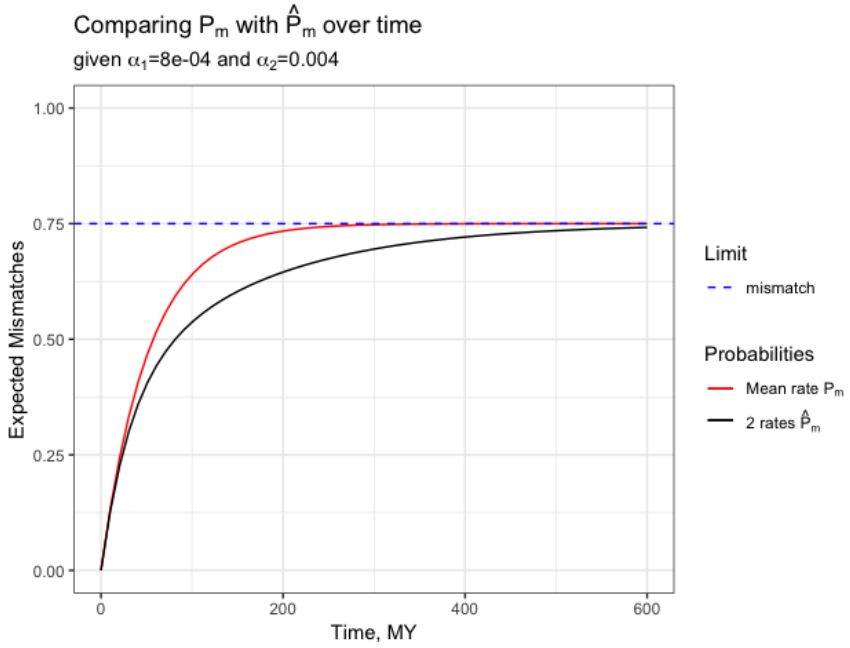
$$\begin{aligned}
\hat{P}_m &= \frac{1}{2} \cdot \left(\frac{3}{4} \cdot (1 - e^{-8\alpha_1 t}) + \frac{3}{4} \cdot (1 - e^{-8\alpha_2 t}) \right) \\
&= \frac{3}{8} \left[(1 - e^{-8\alpha_1 t}) + (1 - e^{-40\alpha_1 t}) \right] \\
&= \frac{3}{8} \left[2 - e^{-8\alpha_1 t} - e^{-40\alpha_1 t} \right]
\end{aligned}$$

   (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.

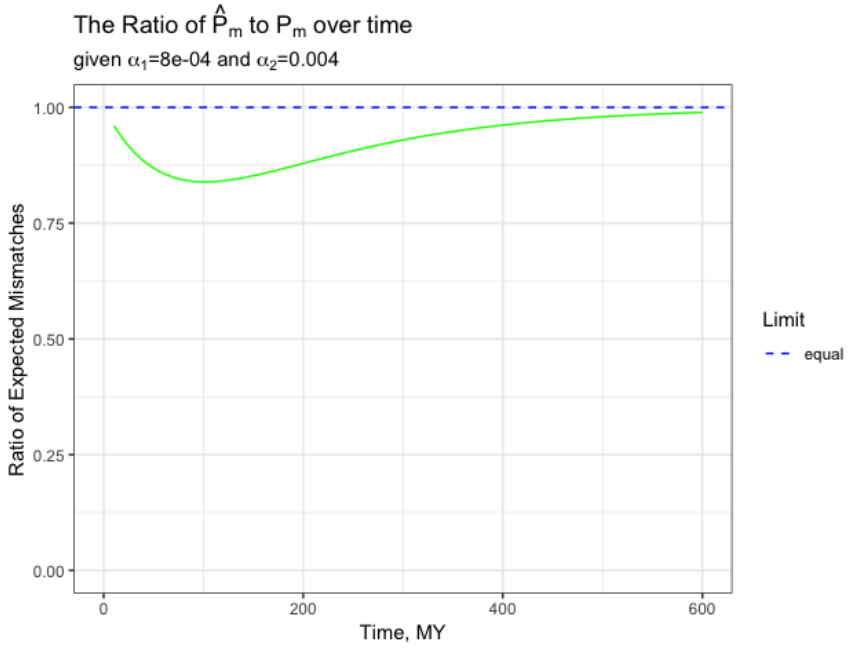
$$\begin{aligned}
\alpha &= \frac{\alpha_1 + \alpha_2}{2} \\
&= \frac{\alpha_1 + 5\alpha_1}{2} \\
&= \frac{6\alpha_1}{2}
\end{aligned}$$

$$\begin{aligned}
P_m &= \frac{3}{4} \cdot (1 - e^{-8\alpha t}) \\
&= \frac{3}{4} \cdot (1 - e^{-24\alpha_1 t})
\end{aligned}$$

(c) Use your favorite plotting program to plot $\hat{P}_m$ and P_m as a function of t, where t, given in units of millions of years, 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 are the consequences of ignoring rate variation and using the average rate instead?

Looking at the equations given in (a) and (b), we see that when t is zero, $\hat{P}_m = P_m = 0$. *Similarly, as t approaches infinity,* $\hat{P}_m = P_m = 3/4$, *because mutations in the sequence have become saturated. Therefore, when t is close to zero or t is very large,* $\hat{P}_m$ *and* P_m *give the same result.*
When t is intermediate, P_m *is bigger than* $\hat{P}_m$*; i.e., using the average rate overestimates the expected number of observed differences.*