Solution Set 4

Due 4pm, Friday, October 4th

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with:

Homework must be submitted by 4pm in MI650 or electronically to mstolzer@andrew.cmu.edu.
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 $\alpha$ 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 $\alpha$. According to your null hypothesis, $H_0$, the probability of observing $x$ aligned with $y$ is the product of their background frequencies, $p_x \cdot p_y$.

1. 15 pts. Give an expression in terms of $\alpha$ and $t$ for $\Pr(M|H_A)$, the likelihood of observing a match in a pair of related sequences.

$$\Pr(M|H_A) = \sum_{x \in \{A,G,C,T\}} \Pr(x \mid t, \alpha)$$

$$\Pr(M|\alpha, t) = 4 \cdot \frac{1}{16} \left(1 + 3e^{-8\alpha t}\right)$$

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

-OR-

$$\Pr(M|H_A) = \sum_{x \in \{A,G,C,T\}} \sum_{z \in \{A,G,C,T\}} p_z \cdot p_{zx}(t)^2$$

$$\Pr(M|\alpha, t) = \sum_{x \in \{A,G,C,T\}} \frac{1}{4} \left(p_{Ax}(t)^2 + p_{Gx}(t)^2 + p_{Cx}(t)^2 + p_{Tx}(t)^2\right)$$

$$= p_{Ax}(t)^2 + p_{Gx}(t)^2 + p_{Cx}(t)^2 + p_{Tx}(t)^2$$

$$= \left[\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right]^2 + 3 \left[\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right]^2$$

$$= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$
2. 10 pts. Give an expression in terms of $\alpha$ and $t$ for $\Pr(m|H_A)$, the likelihood of observing a mismatch in a pair of related sequences.

\[
\Pr(m|H_A) = 1 - \Pr(M|H_A)
\]

\[
\Pr(m|\alpha, t) = 1 - \frac{1}{4} (1 + 3e^{-8\alpha t})
\]

\[
= \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}
\]

-OR-

\[
\Pr(m|H_A) = \sum_{x \in \{A,G,C,T\}} \sum_{y \neq x} \Pr \left( \frac{x}{y} \mid t, \alpha \right)
\]

\[
\Pr(m|\alpha, t) = 12 \cdot \frac{1}{16} (1 - e^{-8\alpha t})
\]

\[
= \frac{3}{4} (1 - e^{-8\alpha t})
\]

3. 10 pts. Give an expression for $\Pr(M|H_0)$, the likelihood of observing a match by chance, in terms of the background frequencies $p_A$, $p_C$, $p_G$, and $p_T$.

\[
\Pr(M|H_0) = \Pr \left( \frac{x}{x} \text{ randomly} \right)
\]

\[
\Pr(M|H_0) = \sum_x p_x^2
\]

\[
\Pr(M|H_0) = p_A^2 + p_C^2 + p_G^2 + p_T^2.
\]
4. 5 pts. Calculate the numerical value of \( \Pr(M|H_0) \), 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,

\[
\Pr(M|H_0) = 4 \cdot \left(\frac{1}{4}\right)^2 = \frac{1}{4}.
\]

5. 5 pts. Calculate the numerical value of \( \Pr(m|H_0) \), 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.

\[
\Pr(m|H_0) = 1 - \Pr(M|H_0) = \frac{3}{4}.
\]

6. 10 pts. 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.

\[
M(\alpha, t) = \log_2 \frac{\Pr(M|H_A)}{\Pr(M|H_0)}
\]

\[
= \log_2 \frac{\frac{1}{4} + \frac{3}{4} e^{-8\alpha t}}{\frac{1}{4}}
\]

\[
= \log_2 [1 + 3 e^{-8\alpha t}]
\]

7. 10 pts. 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.

\[
m(\alpha, t) = \log_2 \frac{\Pr(m|H_A)}{\Pr(m|H_0)}
\]

\[
= \log_2 \frac{\frac{3}{4} - \frac{3}{4} e^{-8\alpha t}}{\frac{3}{4}}
\]

\[
= \log_2 [1 - e^{-8\alpha t}]
\]
8. 10 pts. 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 $\alpha$, and $t$.

$$S = \sum_{i=1}^{100} p(s_1[i], s_2[i])$$

$$= 45 \cdot M(\alpha, t) + 55 \cdot m(\alpha, t)$$

$$= 45 \cdot \log_2[1 + 3e^{-8\alpha t}] + 55 \cdot \log_2[1 - e^{-8\alpha t}]$$

9. 10 pts. 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?

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

10. 10 pts. 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?

$$S = 11.35.$$ 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.