Solution Set 3

Due 4:00pm, Friday, Oct 30

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

Homework must be submitted by 4pm in MI646 or electronically to comp-bio@cs.cmu.edu.

1. Your goal is to devise a scoring system 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 $\sigma$ and $\tau$, your alternate hypothesis, $H_A$, is that $\sigma$ and $\tau$ diverged from a common ancestor $t$ million years ago and are evolving according to the Jukes Cantor model with parameter $\alpha$. Assume that a single time step in the discrete Jukes Cantor model corresponds to one million years. 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 $\alpha$ and $t$ for $P(\text{match}|H_A)$, the likelihood of observing a match in a pair of related sequences.

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

(b) Give an expression in terms of $\alpha$ and $t$ for $P(\text{mismatch}|H_A)$, the likelihood of observing a mismatch in a pair of related sequences.

\[
P(m|\alpha, t) = \frac{3}{4} - \frac{3}{4}e^{-8\alpha t}
\]
(c) Give an expression for $P(\text{match}|H_0)$, the likelihood of observing a match by chance, in terms of the background frequencies $p_A$, $p_C$, $p_G$, and $p_T$.

$$P(M|H_0) = p_A^2 + p_C^2 + p_G^2 + p_T^2$$

(d) 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. At steady state, $p_A = p_C = p_G = p_T = \frac{1}{4}$. Therefore, $P(M|H_0) = 4 \cdot \left(\frac{1}{4}\right)^2 = \frac{1}{4}$.

$$\log_2 \frac{P(M|H_A)}{P(M|H_0)} = \log_2 \frac{\frac{1}{4} + \frac{3}{4}e^{-8\alpha t}}{\frac{1}{4}} = \log_2 \left[1 + 3e^{-8\alpha t}\right]$$

(e) 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. At steady state,

$$P(m|H_0) = 1 - P(M|H_0) = \frac{3}{4}.$$  

$$\log_2 \frac{P(m|H_A)}{P(m|H_0)} = \log_2 \frac{\frac{3}{4} - \frac{3}{4}e^{-8\alpha t}}{\frac{3}{4}} = \log_2 \left[1 - e^{-8\alpha t}\right]$$
(f) Suppose that the ungapped alignment of $\sigma$ and $\tau$ is 100 nucleotides long and contains 55 mismatches. Give an expression for the score of this ungapped alignment in terms of $\alpha$, and $t$?

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

(g) What is the score of this ungapped alignment if $\alpha = 0.002$ 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?

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

(h) What is the score of this ungapped alignment if $\alpha = 0.002$ 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?

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

*Note that interpretation of the data is crucially dependent on the amount of divergence. When } \alpha = 0.002, \text{ 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 64 times more likely to share common ancestry, than chance similarity.*
(i) Suppose the sequences have been diverging for 30 million years and $\alpha = 0.003$. What is the alignment for score in this case? Based on this score, do you think the similarity of the sequences indicates common ancestry or chance similarity? Why?

$S = 5.52$. Common ancestry is more likely than chance.

(j) Suppose you have an alignment that has a score of 10 with this scoring system. How much more likely is the alternate hypothesis than the null hypothesis given this an alignment?

$$S = \sum_{i=1}^{n} \log_2 \frac{Pr(A[i]|H_A)}{Pr(A[i]|H_0)}$$

$$= \log_2 \prod_i \frac{Pr(A[i]|H_A)}{Pr(A[i]|H_0)}$$

$$= \log_2 \frac{\prod_i Pr(A[i]|H_A)}{\prod_i Pr(A[i]|H_0)}$$

$$= \log_2 \frac{Pr(A|H_A)}{Pr(A|H_0)}.$$  

When $S = 10$,

$$\log_2 \frac{Pr(A|H_A)}{Pr(A|H_0)} = 10$$

$$\frac{Pr(A|H_A)}{Pr(A|H_0)} = 1024$$

The alternate hypothesis is more than 1000 times more likely than the null hypothesis.
2. (a) Is the PAM-1 transition matrix symmetric? Justify your answer algebraically.

Suppose the PAM transition matrix is symmetric. Then, \( P_{jk} = P_{kj} \), which expands to

\[
\frac{0.01}{p_j} \sum_h \sum_{l \neq h} A_{hl} = \frac{0.01}{p_k} \sum_h \sum_{l \neq h} A_{hl}.
\]

Since the procedure for counting pairs in the PAM framework ensures that \( A_{jk} = A_{kj} \), the second term on the left hand side is equal to the second term on the right hand side, yielding

\[
\frac{0.01}{p_j} = \frac{0.01}{p_k}
\]

It is not generally true that \( p_j = p_k \), so the proposition is false: The PAM transition matrix is not symmetric.
(b) Is the PAM-1 log odds scoring matrix symmetric? Justify your answer algebraically.

\[
S[j, k] = \lambda \log \frac{q[j, k]}{p_j p_k} = \lambda \log \frac{p_j P[j, k]}{p_j p_k} = \lambda \log \frac{P[j, k]}{p_k}
\]

and

\[
S[k, j] = \lambda \log \frac{q[k, j]}{p_j p_k} = \lambda \log \frac{p_k P[k, j]}{p_j p_k} = \lambda \log \frac{P[k, j]}{p_j}
\]

So,

\[
S[j, k] = S[k, j]
\]

if and only if

\[
\frac{P[k, j]}{p_j} = \frac{P[j, k]}{p_k}
\]

Note that

\[
\frac{P[j, k]}{p_k} = \frac{m_j A_{jk}}{p_k \sum_{i \neq j} A_{ji}}
\]

\[
= \frac{1}{100p_j p_k} \sum_h \sum_{l \neq h} A_{jh} A_{jk} \sum_{i \neq j} A_{ji}
\]

\[
= \frac{1}{100p_j p_k} \sum_h \sum_{l \neq h} A_{hl} A_{kj}
\]

\[
= \frac{P[k, j]}{p_j}
\]

The PAM-1 log odds scoring matrix is symmetric.
(c) Given a pair of sequences with one PAM of divergence, 99 out of 100 positions should be identical. Verify that \( \sum_i p_i P_1[i, i] = 0.99 \).

We know that \( P_1[i, i] = 1 - m_i \),

\[
\sum_i p_i P_1[i, i] = \sum_i p_i (1 - m_i) = \sum_i p_i - \sum_i p_i m_i = 1 - \sum_i p_i \frac{1}{100 p_i} \frac{\sum_{j \neq i} A_{ij}}{\sum_h \sum_{j \neq h} A_{hj}}
\]

\[
= 1 - \frac{1}{100} \sum_i \left( \frac{\sum_{j \neq i} A_{ij}}{\sum_h \sum_{j \neq h} A_{hj}} \right) = 1 - \frac{1}{100} = 0.99
\]

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

We know that the probability that amino acid \( j \) will be replaced by amino acid \( k \) in sequences separated by one PAM of evolutionary distance is

\[
P_1[j, k] = m_j \frac{A_{jk}}{\sum_{i \neq j} A_{ji}}, j \neq k
\]

\[
P_1[j, j] = 1 - m_j
\]

Summing a row in the PAM 1 matrix we get

\[
\sum_k P_1[j, k] = P_1[j, j] + \sum_{k \neq j} P_1[j, k]
\]

\[
= 1 - m_j + m_j \frac{\sum_{k \neq j} A_{jk}}{\sum_{i \neq j} A_{ji}}
\]

\[
= 1 - m_j + m_j \frac{\sum_{k \neq j} A_{jk}}{\sum_{i \neq j} A_{ji}} = 1 - m_j + m_j = 1
\]
3. Serine and threonine (S and T) are small, hydrophilic amino acids; asparagine, aspartic acid, glutamic acid, and glutamine (N, D, E, and Q) are large, hydrophilic amino acids; and methionine, isoleucine, leucine and valine (M, I, L, and V) are small, hydrophobic amino acids.

(a) Based on the entries in the PAM250 matrix (available on the class web site), which of the following amino acid replacements are you more likely to observe in highly diverged sequences?

- a small, hydrophilic amino acid with a small, hydrophobic amino acid.
- a small, hydrophilic amino acid with a large, hydrophilic amino acid.

Show the evidence on which you base your answer.

The average of the $S_{250}$ scores for aligning M, I, L or V with either S or T is -1.3. This suggests that, on average, the observation of a small hydrophilic amino acid aligned with a small hydrophobic amino acid is less likely in related sequences with 250 PAMs divergence than expected by chance.

The average of the $S_{250}$ scores for aligning N, D, E, or Q with either S or T is -0.1. In other words, the replacement of a small hydrophilic amino acid aligned with a large hydrophilic amino acid occurs slightly less frequently than chance.

(b) Which property do you think is more important to protein structure: size or hydrophobicity?

Comparison of the two averages in part (a) indicates that replacement of a small hydrophilic amino acid with a small hydrophobic amino acid is less likely than replacement of a small hydrophilic amino acid with a large hydrophilic amino acid.

This suggests that there is greater selective pressure to maintain the hydrophobicity of the residue at a particular site than to maintain a residue of the same size at a particular site.
4. Both the PAM and the BLOSUM substitution matrix families are parametrized by evolutionary divergence.

(a) Which represents a greater degree of divergence, BLOSUM80 or BLOSUM62? Why?

The BLOSUM $x$ matrices are constructed by clustering sequences such that each sequence in a cluster is at least $x\%$ identical to some other sequence in the cluster. When estimating amino acid substitution frequencies, only sequences from different clusters are compared. This means that the BLOSUM $x$ matrix is derived from sequences that are as much as $(x - 1)\%$ identical. Therefore, smaller values of $x$ correspond to more sequence divergence. In particular, BLOSUM62 represents greater divergence than BLOSUM80.

(b) Which represents a greater degree of divergence, BLOSUM62 or PAM40? Why?

BLOSUM62 corresponds to roughly 30% sequence identity, whereas PAM40 represents more than 60% identity, so BLOSUM62 represents greater divergence.