7Eleven Assignment 4
Due 3pm, Friday, Nov. 30

Your name:

This assignment is required for students taking the 12-unit course only.

Substitution matrices use log-odds scores that are based on the ratio of the probability of observing \( a \) aligned with \( b \) in related sequences \( (q_{ab}) \) and the expected frequency of \( a \) aligned with \( b \) in chance alignments \( (p_a p_b) \). However, the application of these scoring matrices is predicated on the assumption that \( a \) and \( b \) have the same frequency in the query sequence and the matching sequence, which often is not the case.

In order to correct for variations in the underlying frequencies of residues in the query and matching sequences, the present-day BLAST program applies various compositional adjustments when calculating E values. The basis of these compositional adjustments is described in a series of papers by Stephen Altschul and his collaborators.

In this assignment, you are asked to read a review article summarizing that work and then solve a set of problems that illustrate the impact of compositional variation on scoring matrices.

Reading: Protein database searches using compositionally adjusted substitution matrices.

Read this article and answer the following questions. You may read additional materials, if you wish. If you do, you must cite your sources.

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 3pm in MI646.

1. You wish to construct a substitution matrix for scoring alignments of amino acid sequences that have been recoded in the two-symbol alphabet, \( \Sigma = \{H, L\} \), corresponding to hydrophobic and hydrophillic residues. Suppose that you have training data consisting of ungapped pairwise alignments. Each of the alignments in your training data consists of a sequence from genome \( X \) aligned with a sequence from genome \( Y \). From these alignments, you obtain frequencies for all possible pairs of aligned symbols \( (HH, HL, LH, LL) \). \( HL \) pairs correspond to a hydrophobic residue in genome \( X \) and a hydrophillic residue in genome \( Y \). \( LH \) pairs correspond to a hydrophillic residue in genome \( X \) and hydrophobic residue is genome \( Y \).
Derive expressions for the following quantities in terms of the pair frequencies, \( q_{HH}, q_{HL}, q_{LH} \) and \( q_{LL} \).

(a) \( p_H \), the frequency of hydrophobic (\( H \)) residues in all of the sequences in the training data (i.e., from both genomes).

(b) \( p_L \), the frequency of hydrophillic (\( L \)) residues in all of the sequences in the training data (i.e., from both genomes).

(c) \( p_{HX}^X \), the frequency of hydrophobic (\( H \)) residues in sequences from genome \( X \).

(d) \( p_{LX}^X \), the frequency of hydrophillic (\( L \)) residues in sequences from genome \( X \).

(e) \( p_{HY}^Y \), the frequency of hydrophobic (\( H \)) residues in sequences from genome \( Y \).

(f) \( p_{LY}^Y \), the frequency of hydrophillic (\( L \)) residues in sequences from genome \( Y \).
2. Suppose the pair frequencies in the training data are

\[ q_{HH} = 0.1 \]
\[ q_{HL} = 0.3 \]
\[ q_{LH} = 0.3 \]
\[ q_{LL} = 0.3 \]

(a) Using the equations you derived in Question 1, calculate \( p_H \) and \( p_L \), the frequencies of hydrophobic and hydrophillic residues in all sequences in the training data.

(b) Based on your answer to (a), what are the expected frequencies of \( HH \), \( HL \), \( LH \) and \( LL \)?

(c) Calculate the ratios of the likelihoods of observing \( HH \), \( HL \), \( LH \) and \( LL \) in ungapped alignments of related sequences and and ungapped alignments of unrelated sequences.
(d) Calculate a log odds scoring matrix with entries $S_{HH}$, $S_{HL}$, $S_{LH}$ and $S_{LL}$ using log base 2. Scale your matrix by multiplying each entry by 10 and then round the entries to the nearest integer.

(e) In order to be a valid scoring matrix, the mean score per position must be negative. Does your matrix satisfy this requirement? Show your calculation.
3. Given the same training data with the same pair frequencies as in Question 2,
   (a) calculate $p_H^X$ and $p_L^X$, the frequencies of $H$ and $L$ in sequences from genome $X$.

   (b) Calculate $p_H^Y$ and $p_L^Y$, the frequencies of $H$ and $L$ in sequences from genome $Y$.

   (c) Based on your answers to (a) and (b), what are the expected frequencies of $HH$, $HL$, $LH$ and $LL$?
(d) Calculate the ratios of the likelihoods of observing $HH$, $HL$, $LH$ and $LL$ in ungapped alignments of related sequences and and ungapped alignments of unrelated sequences.

(e) Calculate a log odds scoring matrix with entries $S_{HH}$, $S_{HL}$, $S_{LH}$ and $S_{LL}$ using log base 2. Scale your matrix by multiplying each entry by 10 and then round the entries to the nearest integer. Is your matrix different from the matrix you obtained in Question 2? Why or why not?
4. Suppose you have a different training data set with the following pair frequencies:

\[
\begin{align*}
q_{HH} &= 0.15 \\
q_{HL} &= 0.2 \\
q_{LH} &= 0.4 \\
q_{LL} &= 0.25
\end{align*}
\]

(a) Calculate \( p_H \) and \( p_L \), the frequencies of hydrophobic and hydrophillic residues in the alignments.

(b) Based on your answer to (a), what are the expected frequencies of \( HH \), \( HL \), \( LH \) and \( LL \)?
(c) Calculate the ratios of the likelihoods of observing $HH$, $HL$, $LH$ and $LL$ in ungapped alignments of related sequences and and ungapped alignments of unrelated sequences.

(d) Calculate a log odds scoring matrix with entries $S_{HH}$, $S_{HL}$, $S_{LH}$ and $S_{LL}$ using log base 2. Scale your matrix by multiplying each entry by 10 and then round the entries to the nearest integer.
5. Given the same training data with the same pair frequencies as in Question 4,

(a) calculate $p_H^X$ and $p_L^X$, the frequencies of $H$ and $L$ in sequences from genome $X$.

(b) Calculate $p_H^Y$ and $p_L^Y$, the frequencies of $H$ and $L$ in sequences from genome $Y$.

(c) Based on your answers to (a) and (b), what are the expected frequencies of $HH$, $HL$, $LH$ and $LL$?
(d) Calculate the ratio of the likelihoods of observing \( HH \), \( HL \), \( LH \) and \( LL \) in related sequences and unrelated sequences.

(e) Calculate a log odds scoring matrix with entries \( S_{HH} \), \( S_{HL} \), \( S_{LH} \) and \( S_{LL} \) using log base 2. Scale your matrix by multiplying each entry by 10 and then round the entries to the nearest integer. Is your matrix different from the matrix you obtained in Question 4? Why or why not?