Seven11 Assignment 5

Due 11:59pm, Friday, Dec 3rd

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

- S.F. Altschul, et al., Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal* (2005) v272: p5101-5109.
  `https://pubmed.ncbi.nlm.nih.gov/16218944/`

and then apply the compositional adjustment to construct a (very small) substitution matrix.

Homework must be submitted by 11:59pm electronically to Canvas.

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with, and cite all additional references you used to complete this assignment:

You wish to construct a substitution matrix for scoring alignments of nucleic acid sequences that have been recoded in the two-symbol alphabet, $\Sigma = \{R, Y\}$, corresponding to purines and pyrimidines.

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 Z. From these alignments, you obtain frequencies for all possible pairs of aligned symbols (RR, RY, YR, YY). RY pairs correspond to a purine in genome X and a pyrimidine in genome Z. YR pairs correspond to a *pyrimidine* in genome X and purine is genome Z.

1. Derive expressions for the following quantities in terms of the pair frequences, q_{RR}, q_{RY}, q_{YR}, and q_{YY}.

   (a) p_R, the frequency of purines (R) in all of the sequences in the training data (i.e., from both genomes).

   (b) p_Y, the frequency of pyrimidines (Y) in all of the sequences in the training data (i.e., from both genomes).

   (c) p_R^X, the frequency of purines (R) in sequences from genome X.

   (d) p_Y^X, the frequency of pyrimidines (Y) in sequences from genome X.

   (e) p_R^Z, the frequency of purines (R) in sequences from genome Z.

   (f) p_Y^Z, the frequency of pyrimidines (Y) in sequences from genome Z.

2. Suppose the pair frequencies in the training data are

$$\begin{aligned} q_{RR} &= 0.15 \\ q_{RY} &= 0.35 \\ q_{YR} &= 0.35 \\ q_{YY} &= 0.15 \end{aligned}$$

(a) Using the equations you derived in Question 1, calculate p_R and p_Y, the frequencies of purines and pyrimidines in all sequences in the training data.

(b) Based on your answer to 2(a), what are the *expected* frequencies of RR, RY, YR and YY?

(c) Calculate the likelihood ratios of observing RR, RY, YR and YY in ungapped alignments of related sequences and ungapped alignments of unrelated sequences, based on your answer to 2(b).

(d) Calculate a log odds scoring matrix with entries S_{RR}, S_{RY}, S_{YR} and S_{YY} 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. Consider the same training data with the same pair frequencies as in Question 2

$$\begin{aligned} q_{RR} &= 0.15 \\ q_{RY} &= 0.35 \\ q_{YR} &= 0.35 \\ q_{YY} &= 0.15 \end{aligned}$$

(a) Calculate p_R^X and p_Y^X, the frequencies of R and Y in sequences from genome X.

(b) Calculate p_R^Z and p_Y^Z, the frequencies of R and Y in sequences from genome Z.

(c) Based on your answers to 3(a) and 3(b), what are the expected frequencies of RR, RY, YR and YY?

(d) Calculate the likelihood ratios of observing RR, RY, YR and YY in ungapped alignments of related sequences and ungapped alignments of unrelated sequences, based on your answers to 3(c).

(e) Calculate a log odds scoring matrix with entries S_{RR}, S_{RY}, S_{YR} and S_{YY} 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?

(f) In order to be a valid scoring matrix, the mean score per position must be negative. Does your matrix satisfy this requirement?

4. Suppose you have a different training data set with the following pair frequencies:

$$\begin{aligned} q_{RR} &= 0.2 \\ q_{RY} &= 0.3 \\ q_{YR} &= 0.1 \\ q_{YY} &= 0.4 \end{aligned}$$

   (a) Calculate p_R and p_Y, the frequencies of purines and pyrimidines in all sequences in the training data.

   (b) Based on your answer to 4(a), what are the *expected* frequencies of RR, RY, YR and YY?

(c) Calculate the likelihood ratios of observing RR, RY, YR and YY in ungapped alignments of related sequences and ungapped alignments of unrelated sequences, based on your answer to 4(b).

(d) Calculate a log odds scoring matrix with entries S_{RR}, S_{RY}, S_{YR} and S_{YY} 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.

5. Consider the same training data with the same pair frequencies as in Question 3

$$\begin{aligned} q_{RR} &= 0.2 \\ q_{RY} &= 0.3 \\ q_{YR} &= 0.1 \\ q_{YY} &= 0.4 \end{aligned}$$

(a) Calculate p_R^X and p_Y^X, the frequencies of R and Y in sequences from genome X.

(b) Calculate p_R^Z and p_Y^Z, the frequencies of R and Y in sequences from genome Z.

(c) Based on your answers to 4(a) and 4(b), what are the expected frequencies of RR, RY, YR and YY?

(d) Calculate the likelihood ratios of observing RR, RY, YR and YY in ungapped alignments of related sequences and ungapped alignments of unrelated sequences, based on your answers to 4(c).

(e) Calculate a log odds scoring matrix with entries S_{RR}, S_{RY}, S_{YR} and S_{YY} 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?

(f) 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.