Seven11 Assignment 2

Due 11:59pm, Tuesday, November 2nd

Your name:

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

Amino acid substitution matrices are derived from amino acid pair counts that are obtained from training data. These statistics can be estimated in various ways. For the PAM matrices, Dayhoff *et al.* counted amino acid pairs on a tree to correct for sample bias. In this assignment, we further explore the impact of counting pairs on a tree versus counting pairs in columns of a multiple sequence alignment.

In the first question, you are asked to infer ancestral sequences using the parsimony criterion. An introduction to this approach is given in the following reading:

- Page and Holmes, Molecular Evolution: A Phylogenetic Approach. Wiley-Blackwell, 1st ed. (1998), pp. 187-189

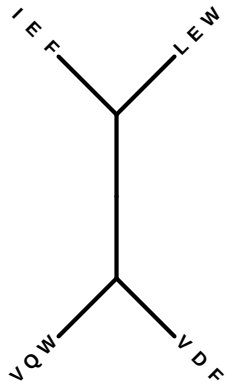
  `http://www.cs.cmu.edu/~durand/03-711/2021/Homework/Page-parsimony.pdf`

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

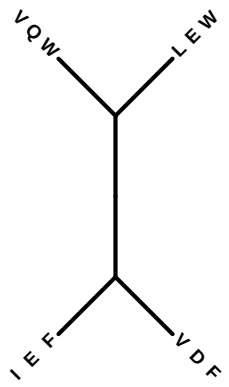
Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with, and cite all additional references you used to complete this assignment:

1. The relationships between k homologous sequences can be expressed as an unrooted tree with k leaves. There are three possible unrooted trees when $k = 4$. The following figure shows a multiple sequence alignment and the unrooted trees representing the three possible hypotheses:

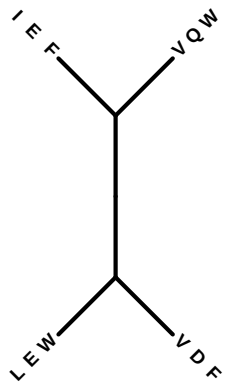
	1	2	3
s_1:	I	E	F
s_2:	L	E	W
s_3:	V	Q	W
s_4:	V	D	F



Tree 1



Tree 2



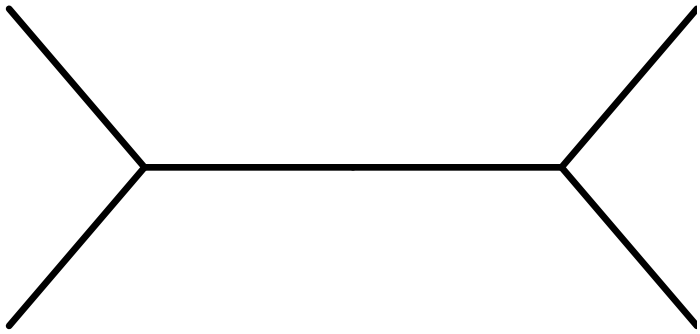
Tree 3

(a) For each of these trees, calculate the parsimony score for each column in the alignment, as well as the total score, following the procedure described in Page and Holmes.

Site	1	2	3	Total
Tree 1				
Tree 2				
Tree 3				

(b) Which tree is the most parsimonious tree?

(c) Annotate the tree shown below with the leaf labels of the most parsimonious tree you obtained in part (b). Annotate the internal nodes with ancestral sequences that require the fewest changes. If there is more than one set of most parsimonious ancestral sequences, just show one.



(d) For the labeled tree you obtained in part (c), use the Dayhoff method to calculate the amino acid pair counts, A_{xy}. Initialize A_{xy} to zero for all x and y (including the case where $y = x$). For every site, if an edge in the tree has labels x and y, increment A_{xy} by one and the A_{yx} by one. If an edge is labeled with the same amino acid at both ends (e.g., x), then increment A_{xx} by two. Enter the values of A_{xx} and A_{xy} in the table given below:

	D	E	F	I	L	Q	V	W
D								
E								
F								
I								
L								
Q								
V								
W								

(e) What is the total count, considering all amino acid pairs?

(f) The frequency, q_{xy}, of an amino acid pair is the pair count A_{xy}, normalized by the total count across all pairs. Calculate the values of q_{xy} and enter them in the table below:

	D	E	F	I	L	Q	V	W
D								
E								
F								
I								
L								
Q								
V								
W								

2. For comparison, we will count amino acid pairs directly from the alignment, which is reproduced here for your convenience.

	1	2	3
s_1:	I	E	F
s_2:	L	E	W
s_3:	V	Q	W
s_4:	V	D	F

(a) For each pair of amino acids, calculate the number of times that pair occurs in the same column of the alignment. For example, if a column of the alignment was `VPVL`, pair `V,P` would appear twice (VPVL and VPVL). Count pairs in both directions; i.e., for every instance of x and y in the same column, increment A_{xy} by one and A_{yx} by one. To be consistent with the Dayhoff counting strategy, for every xx pair in the same column, increment A_{xx} by two. Enter your counts in the table given below:

	D	**E**	**F**	**I**	**L**	**Q**	**V**	**W**
D								
E								
F								
I								
L								
Q								
V								
W								

(b) What is the total count, considering all amino acid pairs?

(c) Calculate the corresponding amino acid pair frequencies by normalizing by the total count of pairs.

	D	E	F	I	L	Q	V	W
D								
E								
F								
I								
L								
Q								
V								
W								

3. Compare the amino acid pair frequencies you calculated using the Dayhoff method on a tree and the all-pairs method using the MSA.

   (a) Are the amino acid pairs with non-zero frequencies obtained with the two methods the same? If not, what pairs have zero frequencies with the tree method, but not the MSA method? What pairs have zero frequencies with the MSA method, but not the tree method?

   (b) For the Dayhoff frequency matrix that you calculated in Question 1, sum the values on the main diagonal (i.e., sum the q_{xx} values). What fraction of the total pair frequency is associated with identical matches in this matrix?

   (c) For the alignment-based the frequency matrix that you calculated in Question 2, sum the values on the main diagonal (i.e., sum the q_{xx} values). What fraction of the total pair frequency is associated with identical matches in this matrix?