COMP 571: Homework #3

Vi Nguyen (vkn1)

April 6, 2017

On my honor, I have neither given nor received any unauthorized aid on this homework.

1. (a) Simply concatenate A and B as is. X and Y are disjoint sets, so every sequence is strictly in either X or Y. Thus, there is no risk of a sequence having different alignments in A and B since no sequence will be aligned in two MSAs.

   In the case that the aligned sequences in A and B are of different lengths (due to the addition of gaps), simply add columns of gaps at the end of whichever MSA is shorter. These columns of gaps will not change the induced alignments between sequences in the same original MSA, thus preserving the requirements for constructing W.

   (b) Let s^* be the single sequence that exists in both X and Y. Consider how s^* is aligned in A and B. Create a "consensus alignment" of s^* by adding gaps present in both alignments to s^*. For each gap added because it was present in the A alignment of s^*, add a column of gaps to the B MSA, and vice versa.

   Thus, when inducing alignemnts between sequences originally from the same sets, they are identical to the original induced alignments, since they will share the added gaps. This fulfills the requirements for constructing W.

   (c) In this case, X and Y share multiple sequences. If we try to apply the method we did in part (b) for all of the sequences, we will create conflicts when we add gaps shared sequences (in an attempt to "align" another shared sequence with itself). These conflicts will require us to continuously add gaps in order to align all of these shared sequences with themselves but also with the other sequences that are shared.

2. If M is an additive matrix, then M has a distance tree associated with it. In this case, let us consider the subtree for points i, j, k, l. There are two distinct possible unrooted trees for these four points.

$M[i,j] + M[k,l] \leq \max\{M[i,k] + M[j,l], M[i,l] + M[k,j]\}$
$I + J + K + L \leq \max\{I + X + K + J + X + L, I + X + L + K + X + J\}$
$I + J + K + L \leq \max\{I + J + K + L + 2X, I + J + K + L + 2X\}$
$I + J + K + L \leq I + J + K + L + 2X \ 0 \leq 2X$

$0 \leq 2X$ is clearly true, since $X \geq 0$ as all distances in the tree must be non-negative. Now let us consider the second distinct case.

$M[i,j] + M[k,l] \leq \max\{M[i,k] + M[j,l], M[i,l] + M[k,j]\}$
$I + X + J + K + X + L \leq \max\{I + K + J + L, I + X + L + K + X + J\}$
$I + X + J + K + X + L \leq I + X + L + K + X + J$
$0 \leq 0$

$0 \leq 0$ is clearly true.

Therefore, if M is additive, then it satisfies the four-point condition.

3.
$$d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = -\frac{3}{4}\ln(1 - \frac{4}{3}(0.65)) = 1.51118$$

4. (a)
$$d_{S_1,S_2} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = -\frac{3}{4}\ln(1 - \frac{4}{3}(0.10)) = 0.107326$$

(b)

$$d_{S_2,S_3} = -\frac{3}{4}\ln(1-\frac{4}{3}p) = -\frac{3}{4}\ln(1-\frac{4}{3}(0.10)) = 0.107326$$

(c) The combined branch lengths we obtained in (a) and (b) is 0.214652, while

$$d_{S_1,S_3} = -\frac{3}{4}\ln(1-\frac{4}{3}p) = -\frac{3}{4}\ln(1-\frac{4}{3}(0.20)) = 0.232616$$

The combined branch length is slightly smaller, because the Jukes-Cantor model corrects evolutionary distance by accounting for the presence of backward and parallel substitutions. For example, the combined branch length assumes that no backwards/parallel mutations occurred between branches S_1, S_2 and S_2, S_3, while computing the JC distance between S_1, S_3 does take this into account.

5. Performing the Jukes-Cantor correction on the distances between some sequences (such as S_1 and S_3) results in taking the natural log of a non-positive number, which is undefined.

   We could fix this problem by adding a constant to the JC model, like the following

$$d_{JC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}p+c\right)$$

   where c is greater than the absolute value of the smallest $1-\frac{4}{3}p$ for all of the pairs of sequences.

6. Clearly, a key difference is that the infinite alleles model allows a site to mutate multiple times, unlike infinite sites. However, since it always mutates to a unique state, if we are considering maximum parsimony, infinite alleles will act identically to infinite sites. In both cases, if a site mutates it will (appear to have, in the case of infinite alleles) have only mutated once, regardless of how many times it actually mutated under infinite alleles. Thus, for example, when using Fitch's algorithm to determine the parsimony of a tree leaf-labeled by a set of sequences, the two models are functionally the same.

   To test if a set of sequences evolved under the infinite alleles model, we can run Fitch's algorithm and generate a tree that now has internal nodes labelled with sequences under maximum parsimony. Then, for each sequence, we can traverse from the root to its leaf and ensure that when a site "mutates" down the tree that it is a unique allele for that site. If this doesn't hold, then the sequences didn't evolve under the infinite alleles model.

7. The informative sites are columns 3 and 5. Using these informative sites, we will find the maximum parsimony with the tree:

$$((s_1, s_2), (s_3, s_4))$$

This is because columns 3 and 5 favor s_1 and s_2 being clustered together and s_3 and s_4 being clustered together. Columns 1, 2, and 4 are uninformative. Columns 1 and 2 have all unique nucleotides and column 4 has all identical nucleotides.

8. Following is the distance matrix with the Poisson corrected distances.

	A	B	C	D
A	-	0.163	0.916	1.049
B		-	0.916	1.049
C			-	0.598
D				-

See attached paper for solutions and calculations.