Ayaana Patel Sikora

Bi 181 PS4

Question 1

	A	B	C	D	E
A	0	2	6	4	7
B	2	0	6	4	7
C	6	6	0	6	9
D	4	4	6	0	5
E	7	7	9	5	0

We can call this matrix D_1. Here, $D_1(A, B) = 2$ is the smallest value of D_1, so we join elements A and B. Let u denote the node to which A and B are now connected. To ensure the elements A and B are equidistant from u, we can set their branch lengths to be $D_1(A, B)/2 = 1$.
Let us now update the distance matrix with the clustering of A and B. Bold values in D_2 denote the new distance calculated by averaging distance between the first cluster and each of the remaining elements:

$D_2((A, B), C) = (D_1(A, C) + D_1(B, C))/2 = (6 + 6)/2 = 6$
$D_2((A, B, D) = (D_1(A, D) + D_1(B, D))/2 = (4 + 4)/2 = 4$
$D_2((A, B), E) = (D_1(A, E) + D_1(B, E))/2 = (7 + 7)/2 = 7$

This gives us D_2:

	(A,B)	C	D	E
(A,B)	0	**6**	**4**	**7**
C	**6**	0	6	9
D	**4**	6	0	5
E	**7**	9	5	0

Here, $D_2((A,B), D) = 4$ is the smallest value of D_2, so we join cluster (A, B) and element D.
Let v denote the node to which (A, B) and D are now connected. The branches joining A or B to v and D to v are equal and have the following length: 4/2 = 2. Again, let us update the distance matrix with the clustering of (A, B) and D. Bold values in D_3 denote the new distance calculated by averaging distance:

$D_3(((A, B), D), C) = (D_2((A, B), C) \times 2 + D_2(D, C))/3 = (6 \times 2 + 6)/3 = 6$
$D_3(((A, B), D), E) = (D_2((A, B), E) \times 2 + D_2(D, E))/3 = (7 \times 2 + 5)/3 = 19/3 = 6.33$

This gives us D_3:

	((A, B), D)	C	E

((A, B), D)	0	**6**	**6.33**
C	**6**	0	9
E	**6.33**	9	0

Here, D_3(C, ((A, B), D)) = 6 is the smallest value of D_3, so we can join elements C and ((A, B), D). Let w denote the node to which C and ((A, B), D) are now connected. The branches joining C and ((A, B), D) to w then have lengths 6/2 = 3.

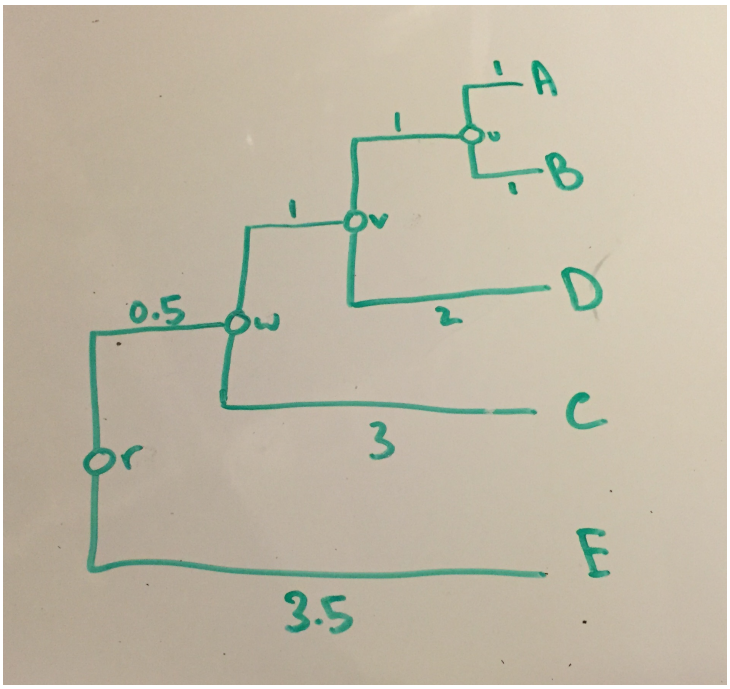
D_4 ((((A, B), D), C), E) = (D_3(((A,B), D), E) x 3 + D_3(C, E) x 1)/4 = (6.33 x 3 + 9)/4 = 7

This gives us D_4:

	(((A, B), D), C)	E
(((A, B), D), C)	0	**7**
E	**7**	0

So, we join clusters (((A, B), D), C) and E. The branch lengths would be 7/2 = 3.5

Thus, we have our branch lengths and clusters:



Question 2

1.

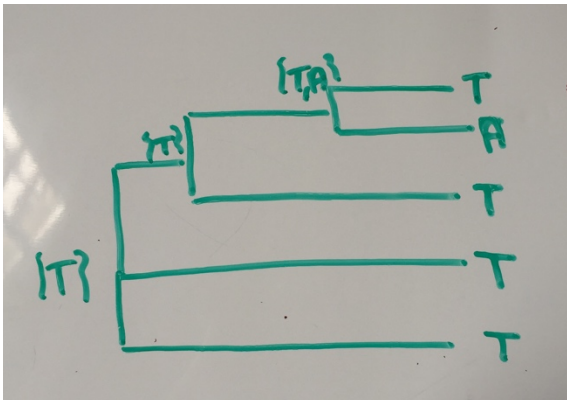
$$MI(i,j) = \sum\sum p_{ij} log \frac{p_{ij}}{p_i p_j} = p_{TA} log \frac{p_{TA}}{p_T p_A} + p_{AT} log \frac{p_{AT}}{p_A p_T}$$
$$= \frac{4}{5} log \frac{\frac{4}{5}}{\frac{4}{5} * \frac{4}{5}} + \frac{1}{5} log \frac{\frac{1}{5}}{\frac{1}{5} * \frac{1}{5}} = 0.72192$$

2. 0.72192 tells us that the amount of correlation between the two columns is pretty high. In other words, knowing the identity of the first position greatly reduces the uncertainty about the identity of the second position.

3. Another pair of positions with a similar mutual information value is GGGAG (column 43) and AAAGA (column 61).
   This is because the probability of G and A are synonymous with the probabilities of T and A in the sequences looked at before (the sequences in the red boxes).

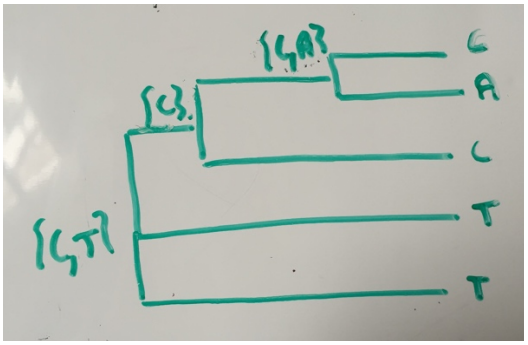
4. We will consider each column separately:

   Column 40: TTTAT



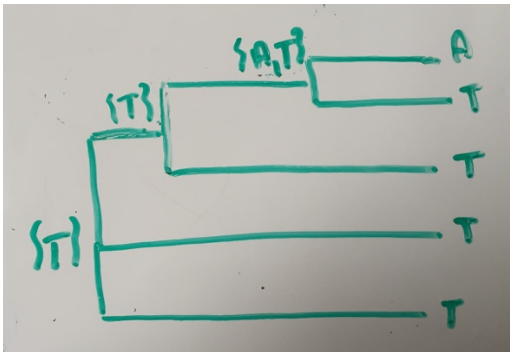
   Thus, the score for this is 1 since only one union is needed. The best choice is picking T at every vertex.

   Column 41: CCTAT



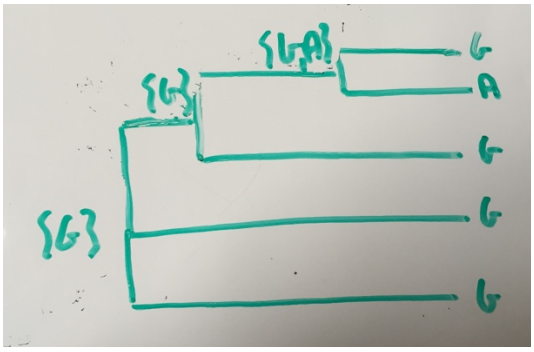
   The score for this is 2 since two unions are needed.

Column 42: ATTTT



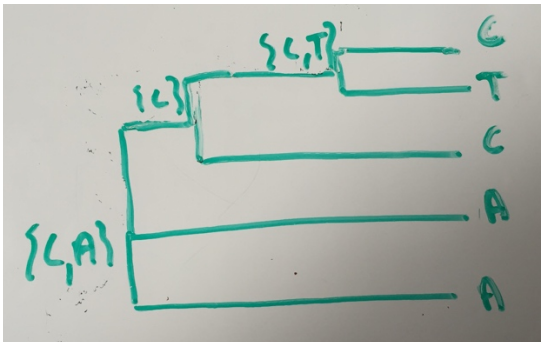
The score for this is 1 since only one union is needed.

Column 43: GGGAG



The score for this is also 1 since just one union is needed.

Column 44: CCATA



The score for this is 2 since two unions are required.

Thus, since the best tree is the one with the lowest score, we can use either the tree resulting from column 40 (TTTAT), 42 (ATTTT) or 43 (GGGAG) since all of these have a score of 1.

Question 3

A. Observed number of differences: 22
Proportion: 0.2650602
See the attached code entitled question3.py, which shows how this was computed.

B. Formula: $d_{ij} = t = -\frac{1}{4} \propto^{-1} \log\left(1 - \frac{4}{3} f\right)$ where f is the fraction of total sites in which the two sequences are different (computed above to be 0.2650602) and ∝ = 0.001. Thus, d_{ij} = 109.01.

Question 4

1. The Markov model assumes that all individuals collaborate with us and give birth at the same time every time. This allows us to model the evolutionary process in discrete steps. The model also assumes that there are no absorving nucleotides and that all substitutions are possible. This orderly population reproduction assumption allows us to easily model the evolutionary process.

2. We know that degeneracy of codons is the redundancy of the genetic code that results because there are more codons than encodable amino acids. A 4-fold degenerate site is where only 4 of 4 possible nucleotides at this position specify the same amino acid. Therefore, a nucleotide substitution a fourfold degenerate site is a synonymous nucleotide substitution, which means using a 4-fold degenerate site would result in a neutral evolutionary matrix. This can lead to unrepresentative evolutionary matrices since it shouldn't always be neutral. This is because some sites are more/less evolutionary stable, but this would not be seen in the matrix resulting from using 4-fold degenerate codons.

3. The scalar ω models the slowing or accelerating of overall rate at which the process operates depending on regional variation. Thus, to find elements or regions with different substitution patterns, we can "fit" ω for every base or region in the genome. Regions in which the fit explains the observed alignment are candidate regions under selection.

Question 5

See the attached code entitled question5.py, which shows the generation of the trees and the algorithms to find the ancestral sequence.

Ancestral Sequence:
[c/t]ta[a/g]ct[c/g]g[c/g]tctt[a/t][a/t][a/g][a/t]g[a/t]a[a/c]ctg[a/g]a[a/c][a/t]c[t/g]