



2. Under the maximum parsimony criterion, we say a column, or site, in a multiple sequence alignment is informative, if it favors one tree topology over another. If the parsimony score at a given site in the alignment is the same for all topologies, then the site is uninformative.

- (a) For each site in the following alignment of sequences from four taxa,

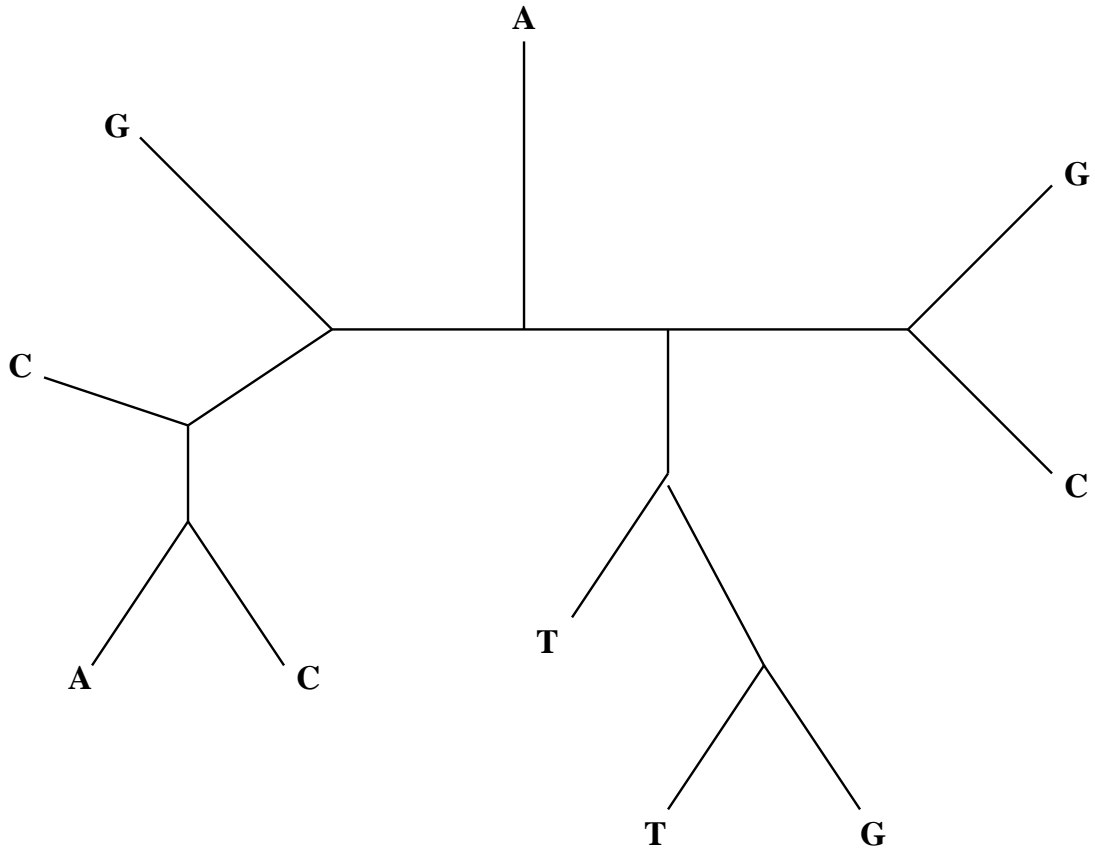
	1	2	3	4	5	6	7	8	9
X.	C	C	G	T	A	G	G	A	C
Y.	A	C	C	T	G	T	G	T	C
Z.	A	G	A	T	G	T	G	C	C
W.	A	G	T	T	A	G	G	C	C

state

- i. if it is an informative site
  - ii. if so, which of the possible tree topologies for four taxa does it favor?
  - iii. if not, what is the parsimony score for this site?
- (b) Show the most parsimonious tree(s).

- (c) What is the maximum parsimony score for this data set?

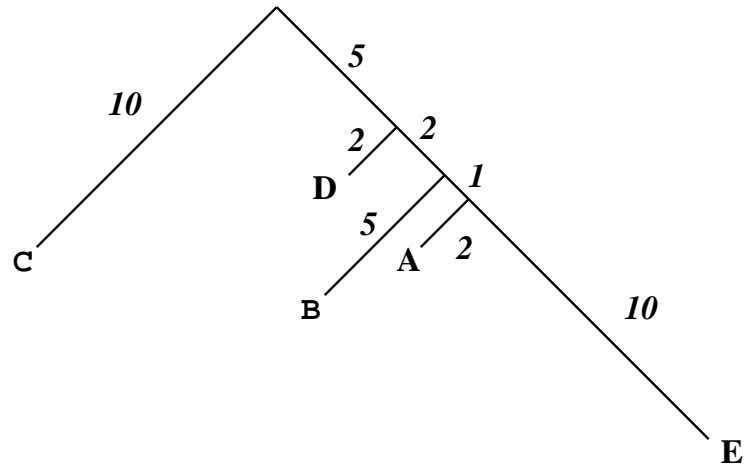
3. What is the parsimony score of the following tree? Show your work.



4. Consider a pair of DNA sequence that are changing according to the Jukes-Cantor model. The residues in these sequences are changing at two different rates: Half the sites change at  $\alpha_1 = 8 \cdot 10^{-4}$  changes per site per million years, while the other half of the sites are changing five times as fast ( $\alpha_2 = 5\alpha_1$ ). Unfortunately, we don't know which of the sites are changing at the fast rate.
- (a) Give an expression for  $P_{\text{mismatch}}$ , the expected fraction of sites with observable differences between two sequences, as a function of  $\alpha_1$ ,  $\alpha_2$  and  $t$ .
- (b) Suppose we incorrectly assume that all sites are following a Jukes-Cantor model of change with the same parameter  $\alpha = (\alpha_1 + \alpha_2)/2$ . Give an expression for  $\hat{P}_{\text{mismatch}}$ , the expected fraction of observable differences as a function of  $\alpha$  and  $t$  in this case.
- (c) Use your favorite plotting program, plot  $\hat{P}_{\text{mismatch}}$  and  $P_{\text{mismatch}}$  as a function of  $t$ , where  $t$ , given in units of millions of years, varies from 0 to 3000.
- (d) What are the consequences of ignoring rate variation and using the average rate instead? How does this effect vary with  $t$ ? Can you explain why the impact is higher at some values of  $t$  than at others?

5. This problem illustrates the strengths and weaknesses of different distance-based methods for tree reconstruction. Remember that while in this case you know the tree from which the distance matrix was derived, when analyzing real data, you will only have the matrix, not the tree.

(a) Compute the distance matrix for the following rooted tree: A, B, C, D and E:



- (b) UPGMA and Neighbor Joining (NJ) are greedy algorithms that build trees by iteratively identifying neighboring nodes and merging them to form a new subtree. How does the UPGMA algorithm select the taxa that will be neighbors in the next iteration?
- (c) *Working only from the distance matrix* (forget you know the tree), which nodes will UPGMA select as the first pair of neighbors? Why?
- (d) Draw the first subtree with an intermediate node  $x$  linking the two neighbors. What are the branch lengths in this subtree?

- (e) How does the NJ algorithm select the taxa that will be neighbors in the next iteration?
- (f) *Working only from the distance matrix*, which nodes will Neighbor Joining select as the first pair of neighbors? Show your calculations for the first iteration of the algorithm.
- (g) Draw the first subtree constructed by the Neighbor Joining algorithm. What are the branch lengths in this subtree?
- (h) Which algorithm is a better choice for reconstructing this tree? Why?

6. (a) Consider the following matrix of observed distances between four taxa, A, B, C and D:

	B	C	D
A	21	6	42
B		26	61
C			45

You wish to reconstruct a tree for the above matrix. Which algorithm would you use to reconstruct the topology and why? How would you determine the location of the root?

- (b) Consider the following matrix of observed distances between four taxa, A, B, C and D:

	B	C	D
A	5	8	35
B		12	26
C			22

You wish to reconstruct a tree for the above matrix. Which algorithm would you use to reconstruct the topology and why? How would you determine the location of the root?

- (c) Consider the following matrix of observed distances between four taxa, A, B, C and D:

	B	C	D
A	9	18	19
B		19	20
C			5

You wish to reconstruct a tree for the above matrix. Which algorithm would you use to reconstruct the topology and why? How would you determine the location of the root?