CS195Z: Problem Set 1

Due: Wed, 2/4/09
Questions: email Crystal (clkahn)

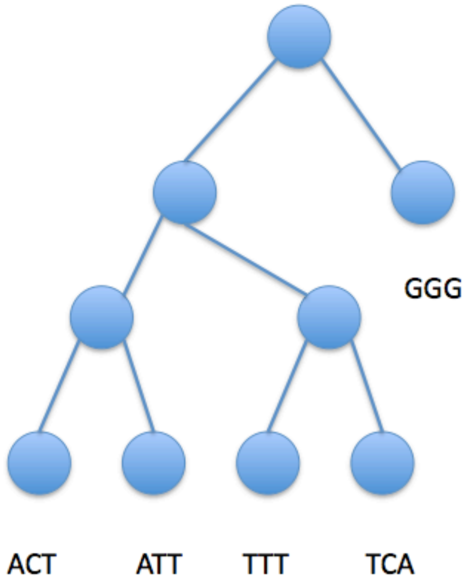
January 28, 2009

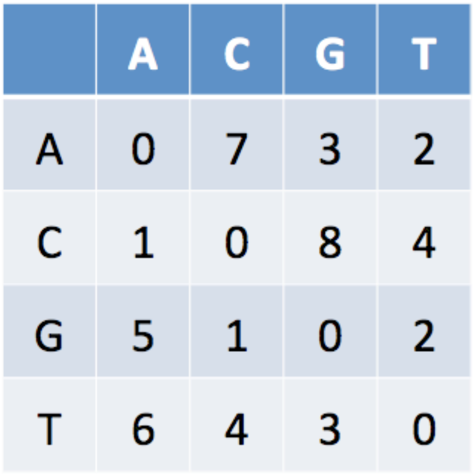
1. A rooted binary tree is a tree in which every vertex has either 0 or 2 children. Prove that for a binary tree with n leaves, there are $n - 1$ internal vertices.

2. An unrooted binary tree is one in which every vertex either has degree 1 or 3. Let $u(n)$ be the number of unrooted binary trees with n leaves. Give a formula for $r(n)$, the number of rooted binary trees with n leaves, in terms of $u(n)$.

3. An unrooted ternary tree is one in which all vertices have either degree 1 or 4. If there are m internal vertices in an unrooted ternary tree, how many leaves are there and how many edges?

4. Using Sankoff's algorithm and the following scoring matrix, find an optimal labeling of the following phylogenetic tree. List the optimal costs on edges.





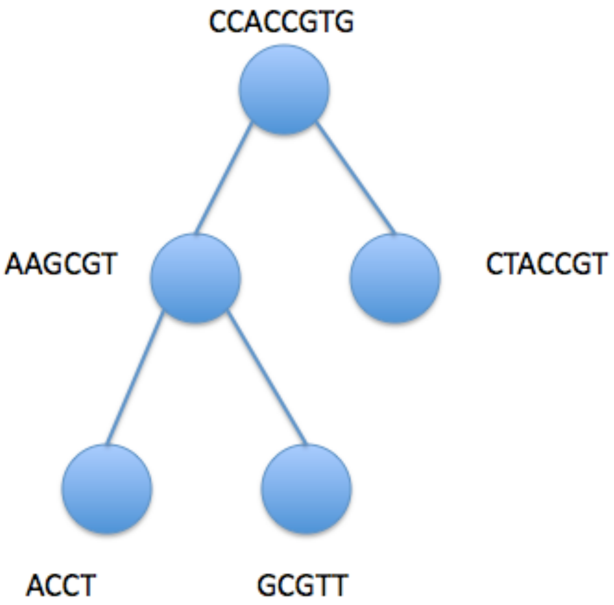
	A	C	G	T
A	0	7	3	2
C	1	0	8	4
G	5	1	0	2
T	6	4	3	0

Note: box (row i, col j) is cost for character i to change to character j

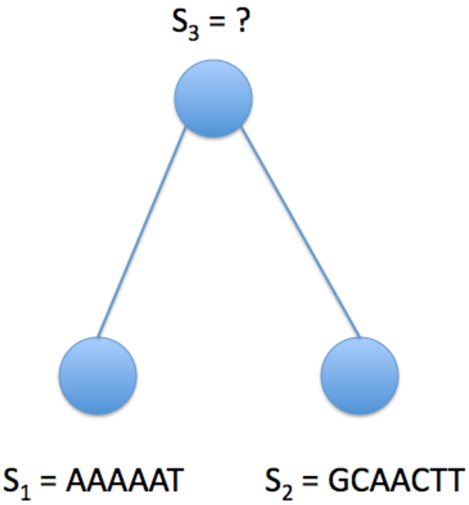
5. Edit distance is a metric between strings. The edit distance between two string is the minimum number of individual character insertions, deletions, and changes needed to transform 1 string into the other. For example, the distance between $S_1 = AAATCCGT$ and $S_2 = ACCGAGG$ is 6. ($S_1 = AAATCCGT$. Delete 'A' twice yields $ATCCGT$. Delete 'T' yields $ACCGT$. Change 'T' to 'A' yields $ACCGA$. Insert 'G' twice yields $ACCGAGG = S_2$.) Edit distance can be used as a measure of

similarity between genes or any other biological sequences. In class, you learned about character-based parsimony methods for computing phylogenetic tress in which every vertex in the tree represents a string of fixed length and the cost on an edge in the tree corresponds to either the total number of character disagreements between the endpoints of the edge or some weighted version of the number of character disagreements (given by a scoring matrix). Alternatively, we could assign costs to edges using a different distance function between strings.

(a) Using edit distance as the cost function, what is the parsimony score of the following tree? Label the edges with costs.



(b) If we are given 2 sequences S_1 and S_2, a *median* sequence with respect to edit distance is a third sequence S_3 that minimizes the sum $edit(S_1, S_3) + edit(S_2, S_3)$. List all optimal median strings S_3 for the two strings given below.



6. (BONUS) In class, we made a distinction between 2 phylogeny problems. The first was: given a tree, a set of leaf vertices, and a cost function, find a labeling of the internal vertices that minimizes the total cost of the tree. The second was: given a set of leaf vertices and a cost function, find a tree and a labeling of the internal vertices that minimizes the total cost of the tree. We mentioned that this second problem is “hard.” How hard is it (in a complexity sense)? A one-sentence proof is sufficient.