CS195Z: Problem Set 2

Due: Wed, 3/4/09
Questions: email Crystal (clkahn)

February 23, 2009

1. Find the compatibility graph and derive the compatibility tree for the species with the following character data set.

Species	1	2	3	4	5	6
Platypus	1	1	0	1	1	1
Elephant	1	1	0	0	0	0
Tiger	1	0	0	1	1	0
Horse	0	0	1	1	1	0
Guinea Pig	0	0	0	0	0	0
Cat	0	0	1	0	0	0

2. Note that the compatibility tree you provided in the previous question was not the same as the most-parsimonious tree for the same set of species. Show that, however, with 5 or fewer species and 0/1 data with unknown ancestral states, the parsimony and compatibility trees will always be the same.

3. Prove that a tree-derived distance satisfies the following 4 properties. Let S be a set of points. For all points x and y in S:

   (a) $d(x, y) \geq 0$

   (b) $d(x, y) = 0$ if and only if $x = y$

   (c) $d(x, y) = d(y, x)$

   (d) for all x, y, and z in S, $d(x, y) \leq d(x, z) + d(z, y)$

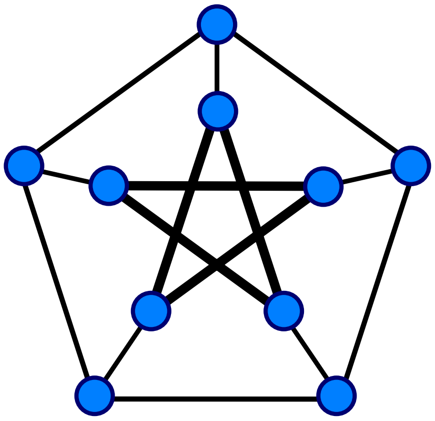
4. Give a method for computing the trimming parameter δ from the additive phylogeny algorithm presented in class.

5. Prove that following statement. If an n x n distance matrix is ultrametric, then it is additive.

6. Consider a character χ on a set of species. We say χ is *trivial* if there is at most one state of the character that is assigned to two or more species. Otherwise, χ is *non-trivial.*

   (a) How many non-trivial binary characters are there on a set of size n?

   (b) Let $X = \{A, B, C, D, E\}$. Show that the compatibility graph of non-trivial binary characters on X is isomorphic to the Petersen graph given below.



7. For five species a, b, c, d, and e with distances given by

	a	b	c	d	e
a	0	9	8	7	8
b		0	3	6	7
c			0	5	6
d				0	3
e					0

reconstruct the tree using the neighbor joining algorithm and the UPGMA algorithm. Compare your answers.

8. Suppose we have two nucleotide sequences:

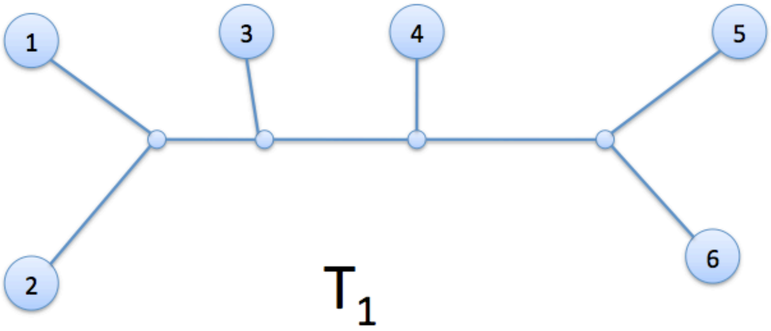
CCGGCCGCGCG
CGGGCCGGCCG

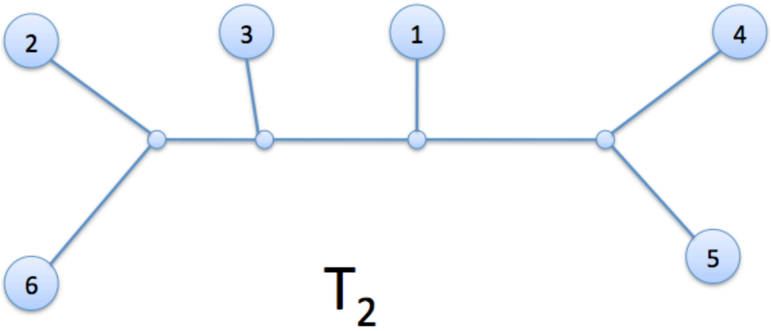
Using the Jukes-Cantor substitution probabilities ($r_t = \frac{1}{4}(1+3e^{-4\alpha t})$ is the probability that a character does not change in time t, and $s_t = \frac{1}{4}(1 - e^{-4\alpha t})$ is the probability of a change to any other character in time t), show that the maximum likelihood solution is given by

$$t_1 + t_2 = -\frac{3}{4} \ln \frac{3n_1 - n_2}{3n_1 + 3n_2}, \tag{1}$$

where t_1 and t_2 are the maximum likelihood edge lengths, n_1 is the number of sites where the residues in the two sequences are identical and n_2 is the number of sites where a substitution occurs. (Recall that for two sequences, there is only one possible tree, namely the one with two branches and a root node which represents the hypothetical common ancestor.)

9. Find the split distance and nearest neighbor interchange distance between the trees





10. Find two trees $T, S \in \mathcal{T}_n$ such that the splits metric is $\rho(T, S) = 2n - 6$.