CS195Z: Take-home final

Due: Monday, May 11, 5pm
Questions: email Crystal (clkahn)

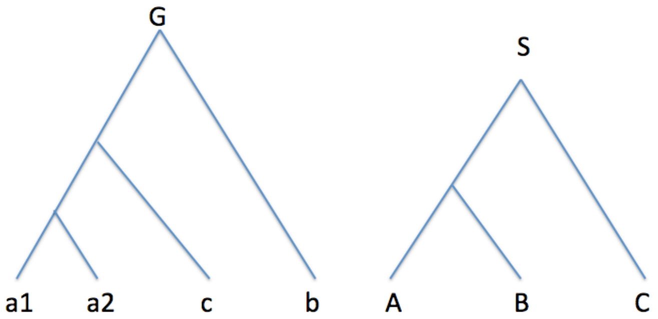
April 30, 2009

1. Suppose d_{ab}, d_{ac}, and d_{bc} are distances that form an additive tree for the unrooted tree with leaves $\{a, b, c\}$. (There is only one such tree.) The tree has length x for the leaf a, length y for the leaf b and length z for the leaf c. Derive a formula for x, y, and z in terms of d_{ab}, d_{ac}, and d_{bc}.

2. Using the results from the previous problem, find the unique tree with distances given by

	a	b	c	d
a	0	3	6	5
b		0	7	6
c			0	3
d				0

3. Use the reconciliation algorithm of Zmasek and Eddy (2001) [described in lecture on March 2] to reconcile the species tree and gene trees below. Please show your work.



4. Use Aho's BUILD algorithm to find the Adams consensus tree of the three rival trees shown in Figure 1.

5. One day your friend the biologist asks for your help with her research. She has spent weeks in the lab collecting gene expression data for 200 human samples. In each sample, there are expression values for 20000 human genes. Using your favorite clustering algorithm, you partition the samples into groups according to gene expression. The result is two clusters consisting of 150 and 50 samples. You show your friend these results and she responds "I forgot to mention: half the samples were from normal patients and the other half are from cancer patients." Looking at the results of your clustering, you see that one of your clusters contains 75 normal and 75 cancer samples, while the other cluster contains 25 of each.

   (1) Assuming that the normal/cancer labels are the correct clustering, evaluate your clustering by computing the Jaccard and Minkowski coefficients.

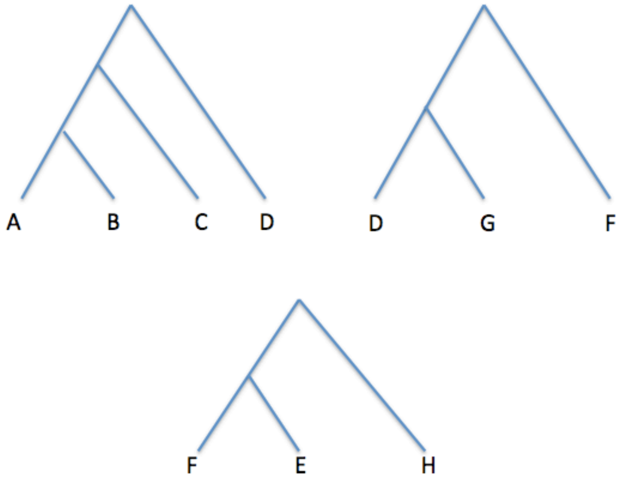


Figure 1:

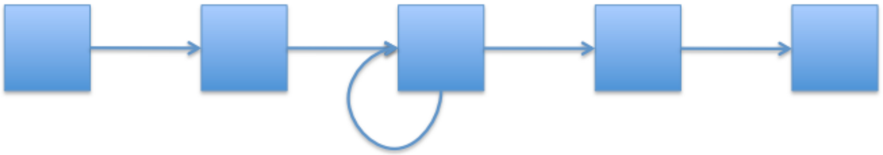
(2) Now you take the 100 cancer samples and cluster the genes in these samples using your favorite clustering algorithm once more. You obtain 30 clusters, the smallest of which contains 50 genes. Looking at what is known about these 50 genes, you discover that 10 of them are annotated as “cell proliferation”. There are 600 human genes total annotated with the “cell proliferation”’ label.

How suprised are you by this result? Express your answer as a p-value according to an appropriate null model.

6. When modeling a phenomenon with an HMM, there is a fixed probability of entering or leaving a state. For example, the probability of entering a state might be p and the probability of leaving is thus $1-p$. The probability of staying in the state for exactly l steps is:

$$P(l\ steps) = (1-p)p^{l-1} \tag{1}$$

Consequently, the distribution of the lengths of “runs” in the same state will be geometric. This might be an inappropriate assumption in some applications. For example, when using an HMM to model aCGH data, there is no reason to assume that the lengths of amplifications/deletions in the genome follows the geometric distribution. One way to model more complex length distributions is to duplicate a state multiple times and chain these copies together as in the following figure:



In the model above, we remain in the duplicated state for at least 5 steps. Suppose we create a chain of n states, for an arbitrary n, each with a transition to itself of probability p and a transition to the next of probability $1-p$ (as in Figure 2):

The minimum number of steps that we remain in the duplicated state is n. Suppose $l > n$.

(a) For any given path of length l through the model, what is the probability of this path (i.e. the product of the probability of all its transitions)?

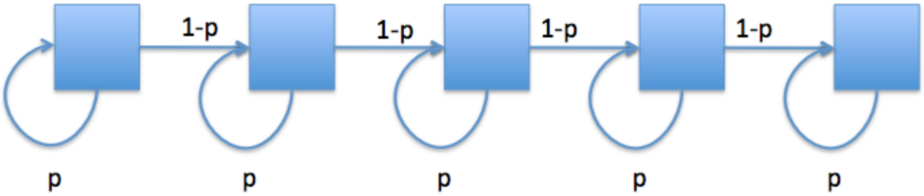


Figure 2:

(b) What is the number of possible paths of length l through the duplicated state?

(c) What is the total probability $P(l)$ over all possible paths of length l ?

7. Given a graph $G = (V, E)$ and set I of start vertices, we described the color-coding algorithm for finding min-weight simple paths of length k that start at a vertex in I. Briefly, this algorithm assigns a color $c(v)$ to each vertex $v \in V$. Colors are drawn uniformly at random from the set $\{1, 2, \ldots, k\}$. A path is *colorful* if it contains exactly one vertex of each color. We seek a min-weight colorful path from I to each vertex v. For each nonempty set $S \subseteq \{1, 2, \ldots, k\}$ and each vertex v such that $c(v) \in S$, let $W(v, S)$ be the minimum weight of a simple path of length $|S|$ that starts at some vertex in I, visits each color in S, and ends at v. For $|S| > 1$, we gave a dynamic programming algorithm for computing $W(v, S)$ based on the following recurrence:

$$W(v, S) = \min_{u:c(u)\in(S\setminus\{c(v)\})} W(u, S \setminus \{c(v)\}) + w(u, v) \tag{2}$$

The colorful property ensures that the path we find is simple. And over many randomized colorings of our graph, we find the min-weight simple path with high probability.

Now suppose there is special set of vertices $T \subseteq V$. For example, in biological applications, T might be the set of receptors or transcription factors. Extend the dynamic program above to enforce a constraint that the computed path contains at least a and at most b vertices from the set T. Be sure to state the initial conditions.

8. Suppose now that we wish to count the number of colorful paths of length k that end at a particular vertex v on a randomly colored graph $G = (V, E)$ (refer to problem 7 for the definition of "colorful"). Let the color set be $\{1, \ldots, k\}$. Again, let $c(v)$ denote the color of v. For each vertex $v \in V$ and each subset S of the color set, let $C(v, S)$ be the number of colorful paths for which one of the endpoints is v. Given a color l, for all $v \in V$, $C(v, \{l\}) = 1$ if $c(v) = l$ and 0 otherwise. For each vertex v and color subset S, where $|S| > 1$, we have:

$$C(v, S) = \sum_{u:(u,v)\in E} C(u, S \setminus c(v)). \tag{3}$$

Note that the number of simple colorful paths of length k would be:

$$\frac{1}{2} \sum_{v} C(v, \{1, \ldots, k\}). \tag{4}$$

Now suppose that we wish to count the number of colorful subgraphs of G that are isomorphic to a given unrooted tree T with k vertices. To get started, let us pick an arbitrary vertex ρ of T and set it to be the root. Denote this rooted tree by $\tau(\rho)$. For each $v \in G$, let $c(v, \tau(\rho), \{1, \ldots, k\})$ be the number of k-colorful rooted subtrees in G with root v that are isomorphic to $\tau(\rho)$.

(a) Suppose we know this value, $c(v, \tau(\rho), \{1, \ldots, k\})$, for every vertex v. Derive a formula for the number of k-colorful occurrences of T in G.

(b) Derive a dynamic programming routine to compute the value $c(v, \tau(\rho), k)$ as defined above. Hint: consider splitting $\tau(\rho)$ by removing a single edge (ρ, ρ').