B363: Bioinformatics algorithms

HW5 (Due: **Dec. 5 Monday 5pm**)

http://darwin.informatics.indiana.edu/col/courses/B363-16

(You do not need to write computer programs for the following questions.)

1. (10 pts) Given the following matrix of pairwise distances among the genes from four species (A-D), construct the evolution tree of these four species using the neighbor-joining algorithm. Describe each step in your construction. Is the distance matrix *additive*?

	A	B	C	D
A	0	1.0	0.5	1.3
B	1.0	0	1.1	1.9
C	0.5	1.1	0	1.2
D	1.3	1.9	1.2	0

2. (10 pts) Dr. Smart makes the following claim: given a distance matrix $\{D_{ij}\}$ on n species (leaves), if there exist two species (leaves) i and j, the sum of their limb-lengths is equal to their distance: i.e., D_{ij} = LimbLength(i) + LimbLength(j), then these two leaves must be *neighbors* in the evolutionary tree (i.e., they share the immediate common ancestor); if such pair does not exist, then the input distance matrix is not *additive*. Is Dr. Smart right? If so, explain why. Otherwise, provide a counterexample.

3. (10 pts) Dr. Smart goes ahead, based on his claim, to devise the following algorithm to construct an evolutionary tree from an input additive distance matrix, and report it if the input matrix is not additive. What is the runtime of the algorithm, in comparison with the neighbor-joining algorithm?

```
Input: Distance matrix $D_{ij}$ on n species (leaves).
SmartAdditivePhylogeny(D, n)
For each leaf i
        Compute LimbLength(i)
For each pair of leaves i and j
        if $D_{ij}$ = LimbLength(i) + LimbLength(j)
                Create the node m as the common ancestor of i and j
                Compute the distance from m to each leaf k ($\neq$ i or j), $D_{mk} \leftarrow D_{ik}$- LimbLength(i)
                Construct D' (with n-1 leaves) by adding m and removing i and j from D
                Tree' $\leftarrow$ SmartAdditivePhylogeny(D', n-1)
                Join i and j with m in Tree', and form Tree
                Return(Tree)
Output "D is not an additive matrix."
```

4. (10 pts) Mr. Fuzzy comes up the following algorithm to solve the small parsimony problem (page 37, vol II) for an input rooted tree T with each leaf labeled by a character c.

```
FuzzySmallParsimony(T)
        N ← the size alphabet  //alphabet represent all putative characters
        for i←1 to N
                Count[i] ← 1
        Root ← the root of T
        Left ← the left child of the root
        if Left is not a leaf
                T' ← the subtree rooted by Left
                FuzzySmallParsimony(T')
        else
                k ← the character labeled on Left
                Count[k] ← Count[k] + 1
        Right ← the right child of the root
        if Right is not a leaf
                T' ← the subtree rooted by Right
                CountCharacter(T')
        else
                k ← the character labeled on Right
                Count[k] ← Count[k] + 1
        Max_k ← the character k with the maximum count Count[k] for all characters
        Assign Max_k to Root
```

Is the algorithm correct? Justify your answer.

5. (10 pts) Design a polynomial time algorithm to check whether there exist a solution to the good clustering problem (page 76, vol II) for a given set of n points and cluster number k.

6. (10 pts) Let *Centers* be the set of centers returned by FARTHESTFIRSTTRAVERL algorithm (page 80, vol II), and let $Center_{opt}$ be a set of centers corresponding to an optimal solution of the *k*-Center Clustering Problem (page 79, vol II). Prove that MAXDISTANCE(Data, Centers) ≤ 2 × MAXDISTANCE(Data, Centers^{opt}).

7. (10 pts) Although the k-Means Clustering Problem is NP-hard for k > 1, it can be solved in polynomial time for any value of k in the case of clustering in on dimensional space, i.e., when all data points fall on a line. Design an algorithm for solving the k-Means Clustering Problem in this case.

8. (10 pts) Given the parameters=(θ_A, θ_B)=(0.6, 0.5) and the sequence of coin flips "HTHHHHHTHH", what are the responsibilities for coin A and B to generate the sequence?

9. (10 pts) Mr. Fuzzy claims, if we defines the distance between two clusters as the smallest distance between any pair of elements from these clusters, the

hierarchical clustering will result in the clusters in which every pair of elements with the distance below the distance threshold must fall into the same cluster. Is he right? Justify your answer.

10. (10 pts) Construct the suffix tree and the suffix array for the following string “mississippi”.