Exercise Sheet 5 for
Computational Biology (Part 2), SS 14

Hand In: *Until Tuesday, 08.07.2014, 10:00 am, email to* `wild@cs...` *or in lecture.*

Problem 13
2 + 2 points

In this exercise, we consider again the algorithm for finding a *compact additive tree* for a metric δ on a set of taxa A.

You are expected to precisely argue why your examples have the desired properties.

a) Give an example for a metric δ whose associated weighted graph has several minimal spanning trees (MSTs). Show that none of them is a (compact) additive tree for δ by giving a pair of taxa for each MST that is not assigned the correct distance in the tree.

b) Give an example for a metric δ whose associated graph has a unique MST, but that nevertheless does not allow a compact additive tree.

   This shows that the converse of Theorem 24 of the German lecture notes does not hold.

Problem 14
3 points

Design an algorithm that computes a *perfect phylogenetic tree* — or shows that none exists — for n taxa and m binary characters[1] in time $\mathcal{O}(m \cdot n)$. Your solution should comprise

- a clear description of the algorithm (pseudocode or prose is sufficient),

- a proof of its correctness

- and an analysis showing that the algorithm satisfies the given time bound.

You can build upon the algorithm sketched in lecture, but make sure to argue precisely how your algorithm stays in $\mathcal{O}(m \cdot n)$.

---

[1] deutsch: Merkmale

Problem 15 3 points

Let $\mathcal{T} = \{T_1, \ldots, T_m\}$ be a set of trees obtained by repeated execution of steps $2-4$ of the *quartet puzzling* algorithm (page 182 in the German lecture notes). We *root* all trees in $\mathcal{T}$ by the following procedure: First, we pick an arbitrary, but fixed taxon a. Then in each T_i, we insert a new *artificial* node in the middle of the (unique) egde to the leaf corresponding to a and declare this new node the root of this tree.

Now, we label the resulting trees bottom-up: A leaf for taxon a_i is assigned label $\{a_i\}$ and each inner node's label is the union of the labels of its children. Note that these labels fully describe the topology of a tree, so it suffices to consider the set of labels for each tree T_i.

From these m label sets, we now pick all labels that occur in $> \frac{m}{2}$ label sets. Show that the resulting collection of labels always encodes a (not necessarily binary) tree T whose leaves exactly correspond to the taxa.

This means, T is a valid phylogenetic tree for the given taxa. It is called the *majority consensus tree* for $\mathcal{T}$.