Technische Universität
Kaiserslautern

Department of Computer Science
Algorithms & Complexity Group

Issue Date: 03.06.2014
Version: 2014-06-26 15:02

Exercise Sheet 4 for Computational Biology (Part 2), SS 14

Hand In: *Until Tuesday,* ***24.06.2014****, 10:00 am, email to* `wild@cs...` *or in lecture.*

Problem 9

3 points

We consider the *WPGMA (weighted pair group method with arithmetic mean)* algorithm, which is essentially the UPGMA algorithm as introduced in the lecture notes on page 162.[1] However in step 3, new distances between W and all $X \in \Gamma \setminus W$ are computed as follows:

$$dist(W,X) \;=\; dist(X,W) \;=\; \frac{|R_1| \cdot dist(R_1,X) + |R_2| \cdot dist(R_2,X)}{|R_1| + |R_2|}$$

Let $dist(X,Y)$ be the (final) distance between two nodes X and Y of the tree computed by the WPGMA algorithm. Show that

$$dist(X,Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} \delta(x,y)\,,$$

where δ is the input metric.

Problem 10

2 points

Let $T = (V,E)$ be an arbitrary tree (not necessarily binary and possibly unrooted) and let $A \subseteq V$ be a subset of the nodes of T. Further assume that the edges have positive weights $d : E \to \mathbb{R}_{>0}$. We define the *distance* $dist_T(x,y)$ of two nodes $x, y \in A$ as the sum of edges weights on the simple path from x to y in T.

Show that $dist_T(x,y)$ is well-defined for all $x, y \in A$ and that $dist_T$ is a metric on A.

[1] Unfortunately, the names UPGMA and WPGMA are used the other way round in parts of the scientific literature ...
Always check the update formula to be sure!

Problem 11

4 points

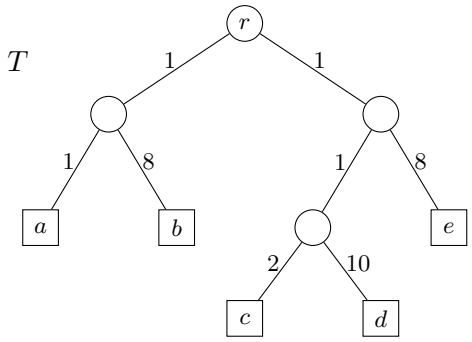
Design an implementation of the UPGMA algorithm that runs in $o(n^3)$ time.

Problem 12

2 + 4 points

We consider additive, binary and rooted phylogenetic trees.

a) Let T be the following additive phylogenetic tree for taxa $A = \{a, b, c, d, e\}$:



Explicitly write down the distance matrix δ encoded by T. Show that δ is *not* an ultrametric. Then, apply the UPGMA algorithm on A and δ — even though δ is not ultrametric.

b) Let T be any additive, binary and rooted phylogenetic tree for taxa A where the taxa correspond exactly to the *leaves* of T. Let r be the root of T and call $dist_T(x, y)$ the length of the unique path from x to y in T.

Now consider the following distances for all $x, y \in A$, $x \neq y$

$$\delta'(x, y) \quad := \quad \hat{\delta}_r \; + \; \frac{dist_T(x, y) - dist_T(x, r) - dist_T(y, r)}{2},$$

where $\hat{\delta}_r := \max_{a \in A} dist_T(a, r)$ is the height of T. Of course, we set $\delta'(x, x) := 0$.

(i) Compute δ' for the tree T from a) and apply the UPGMA algorithm on δ'. Compare the resulting ultrametric tree to T.

(ii) Show that δ' is always an ultrametric.

Note: This statement is equivalent to the fact that the UPGMA algorithm on δ' always reconstructs the topology of the underlying additive phylogenetic tree T.