Stockholm University
Department of Mathematics
Examiner: Marc Hellmuth

3. Homework "DA7065 Computational Biology"

Exercise 1: Additive Metrics and Ultrametric 5+(5+5) = 15p
Let $D : X \times X \to \mathbb{R}$ be a symmetric map that satisfies $D(x, y) = 0$ precisely if $x = y$.

(a) *Prove the 3-point condition:*
D is an ultrametric if and only if for all $x, y, z \in X$ the two largest elements in $\{D(x, y), D(y, z), D(x, z)\}$ are equal.

(b) *Prove or disprove:*

(i) Every ultrametric is an additive metric.

(ii) Every additive metric is an ultrametric.

Exercise 2: UPGMA and Parsimony 5+5 = 10p
Let us consider the following four "genes"

$$\text{a} = \texttt{TTAA}; \quad \text{b} = \texttt{TCGG}; \quad \text{c} = \texttt{AACT}; \quad \text{d} = \texttt{AATC}$$

Assume, for simplicity, that the evolutionary distances between two genes are given by the respective Hamming distance, which results in a distance matrix D for these four genes.

(a) Apply UPGMA on D and provide the resulting tree T together with respective branch-length.

Given that the evolutionary distances in D are the true distances, do you rely in the respective computes tree? Shortly Explain.

(b) Use the rooted tree T obtained with UPGMA on D and assign ancestral sequences to T such that the parsimony score of T gets minimized.

Exercise 3: BUILD, Compatibilty Graphs and Triples 5+5 = 10p
To recall, for a triple set R and a leaf-set L, the comparabilty graph $G[R, L]$ is an undirected graph with vertex set L and edges $\{x, y\}$ precisely if there is a triple $xy|z \in R$ with $x, y, z \in L$

(a) Determine whether the triple sets $R_1 = \{ab|g, ac|g, de|g, ef|g, df|g\}$ and $R_2 = \{ab|g, ac|g, de|g, ef|g, df|g, cd|g, ec|d, cf|d, fd|e\}$ are compatible. To this end, apply the BUILD-algorithm and give the resulting tree obtained with BUILD, if there is one.

(b) Let R be a compatible triple set and assume that $R' = R \cup \{ab|c\}$ is not compatible. Let $L = \cup_{xy|z \in R'} \{x, y, z\}$.

Show, there is a subset $L' \subseteq L$ with $|L'| \geq 3$ such that $G[R, L']$ has exactly two connected components, one containing a and the other b.

HINT: Recheck the proof for the correctness of BUILD sas provided in the lecture video.

Exercise 4: Orthologs 7.5+7.5 = 15p
Let A, B, C, D be four different species from which we extracted some genetic material, i.e., a set of genes $\mathcal{G} = \{a_1, a_2, b_1, b_2, c_1, d_1\}$ where Each gene $x_i \in \mathcal{G}$ is contained in the particular species $X \in \{A, B, C, D\}$.
Using multiple sequence alignments we obtained the (symmetric) similarity scores for the genes in $\mathcal{G}$ as provided in the following matrix:

	b_1	b_2	c_1	d_1
a_1	4	2	1	1
a_2	2	3	1	1
b_1			1	1
b_2			1	1
c_1				1

(a) Apply the graph-based approach (as explained in lecture – see slide no 10) on the similarity scores and determine the estimated orthology relation $\widehat{R}_{\bullet}$ for the genes in $\mathcal{G}$.

(b) Explain why the estimated orthology relation $\widehat{R}_{\bullet}$ is "feasible" and determine the gene tree T together with its duplication and speciation labels t such (T, t) explains $\widehat{R}_{\bullet}$.

Try to add branch-length to this tree to reflect the similarity scores.

$\star$-***exercises***

Exercise 5***:** 7.5
Let R be a consistent triple set and assume that $R' = R \cup \{ab|c\}$ is not consistent. Let $\mathcal{L} = \cup_{xy|z \in R'} \{x, y, z\}$.
Show, there is a subset $L \subseteq \mathcal{L}$ with $|L| \geq 3$ such that the Ahograph $[R, L]$ has exactly two connected components, one containing a and the other b.

Exercise 6***:** 7.5
Show that the two definitions for cographs are equivalent:

Def 1:
- K_1 is a cograph.
- The disjoint union of two cographs is a cograph.
- The complement of a cograph is a cograph.

Def 2:
- K_1 is a cograph.
- The disjoint union of two cographs is a cograph.
- The join of two cographs is a cograph.

Deadline: ASAP, but before the exams!