

Algorithms in Computational Biology – Homework Exercise 5

Publication date: Thursday, *January 7*
Due date: Sunday, *January 24 (9pm IST)*

Problem 1:

We proved in class (Lecture #10) a fundamental claim as basis for Fitch's algorithm for maximum parsimony. The claim considered a binary phylogenetic tree T with two principle subtrees T_1 and T_2 and states assigned to all leaves.

Claim: There is a maximum parsimony assignment to nodes in T , s.t. the induced assignments to T_1 and T_2 are also optimal.

- For each of the three claims below (a-c), specify whether it is **true** or **false**.
 - If you argue that a claim is **true**, then prove it using clear and formal arguments.
 - If you argue that a claim is false, then provide an explicit counter example. Your example should include a tree T with assignments to the leaves and a maximum parsimony assignment to the internal nodes of T . Make sure to support your claims that a given assignment is most parsimonious.
- a. In all maximum parsimony assignments to nodes in T , the induced assignments to T_1 and T_2 are also optimal.
 - b. In all maximum parsimony assignments to nodes in T , the induced assignment to either T_1 or T_2 (or both) is also optimal.
 - c. Every maximum parsimony assignment to T_1 and T_2 can be extended to an optimal assignment to T .

- a. Provide a simple example that shows that the UPGMA algorithm is not generally statistically consistent.
- Your example should consist of a pairwise distance matrix, which is consistent with a given unrooted edge-weighted tree T .
 - Show that when UPGMA is applied to this matrix, it reconstructs a tree that does not have the same unrooted topology at T .
 - Note that while UPGMA reconstructs a rooted tree (the root is the final joining point), what we actually care about is the unrooted topology of that tree.
 - Try to think of as simple example as you can (small number of taxa and simple edge weights).
- b. Prove that UPGMA is statistically consistent for the special case of ultrametric trees.
- Assume that the input pairwise distances are consistent with an arbitrary ultrametric tree T . You may assume that T is binary and that all edge weights are strictly positive.
 - Prove that if clusters A, B correspond disjoint subtrees of T , and they are not neighbors (they are not sister subtrees), then there is a subtree C , which is disjoint to both of them and $\min\{d(A, C), d(B, C)\} < d(A, B)$. Distances between clusters are defined as in the UPGMA algorithm.
Tip: you may find it useful to start thinking about subtrees that consist of a single leaf.
 - Use the claim you proved above to prove that the clusters held by UPGMA in every iteration correspond to valid subtrees of T , and use this to prove that the algorithm eventually returns the (rooted) tree topology of T .

Submission Instructions:

- Submit your work on the course **Moodle website** by Sunday, Jan 24 @21:00.
- Type your solutions or write legibly and scan. **If you scan, make sure the scan came out fine.**
- **Submit your work in pairs!** One student should submit a solution file with both of your student IDs specified. The other student should submit a simple text file with your two student ids, to help us match back the grade to both of you.
- If you consult with other pairs on ideas, specify their names clearly on the first page and make sure that they **acknowledge your collaboration** as well.
- You have two weeks to complete the assignment. **Plan your time wisely.** Extensions due to special circumstances, will be granted only upon **request by e-mail at least 48 hours** before the deadline. No last minute extensions!
- Please post any questions that you have on the course Piazza website:
<https://piazza.com/idc.ac.il/fall2020/cs3571/>.