Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Testing the Agreement of Trees with Internal Labels

David Fernández-Baca
Iowa State University https://orcid.org/0000-0002-8563-3637
Lei Liu (lliu@iastate.edu)
Iowa State University College of Liberal Arts and Sciences https://orcid.org/0000-0002-8566-6391

Research Article

Keywords: Phylogenetic tree, Taxonomy, Agreement, Algorithm

Posted Date: May 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-454322/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Algorithms for Molecular Biology on December 1st, 2021. See the published version at https://doi.org/10.1186/s13015-021-00201-9.

Abstract

Background: A semi-labeled tree is a tree where all leaves as well as, possibly, some internal nodes are labeled

with taxa. Semi-labeled trees encompass ordinary phylogenetic trees and taxonomies. Suppose we are given a collection P = {T_1, T_2, . . . , T_k} of semi-labeled trees, called input trees, over partially overlapping sets of taxa. The agreement problem asks whether there exists a tree T , called an agreement tree, whose taxon set is the union of the taxon sets of the input trees such that the restriction of T to the taxon set of T_i is isomorphic to $_i$, for each i ε 1, 2, . . . , k . The agreement problems is a special case of the supertree problem, the problem of synthesizing a collection of phylogenetic trees with partially overlapping taxon sets into a single supertree that represents the information in the input trees. An obstacle to building large phylogenetic supertrees is the limited amount of taxonomic overlap among the phylogenetic studies from which the input trees are obtained. Incorporating taxonomies into supertree analyses can alleviate this issue.

Results: We give a O(nk(i∈[k]d_i + log^2(nk))) algorithm for the agreement problem, where n is the total number of distinct taxa in P, k is the number of trees in P, and di is the maximum number of children of a node in Ti. Our computational experience with the algorithm suggests that its performance in practice is much better than its worst-case bound indicates.

Full Text

This preprint is available for download as a PDF.

Figures

Figure 1

A prole P = {T1; T2; T3; T4}. The letters are the original labels; grey numbers are labels added to make the trees fully labeled. We use this prole as a running example throughout the paper.

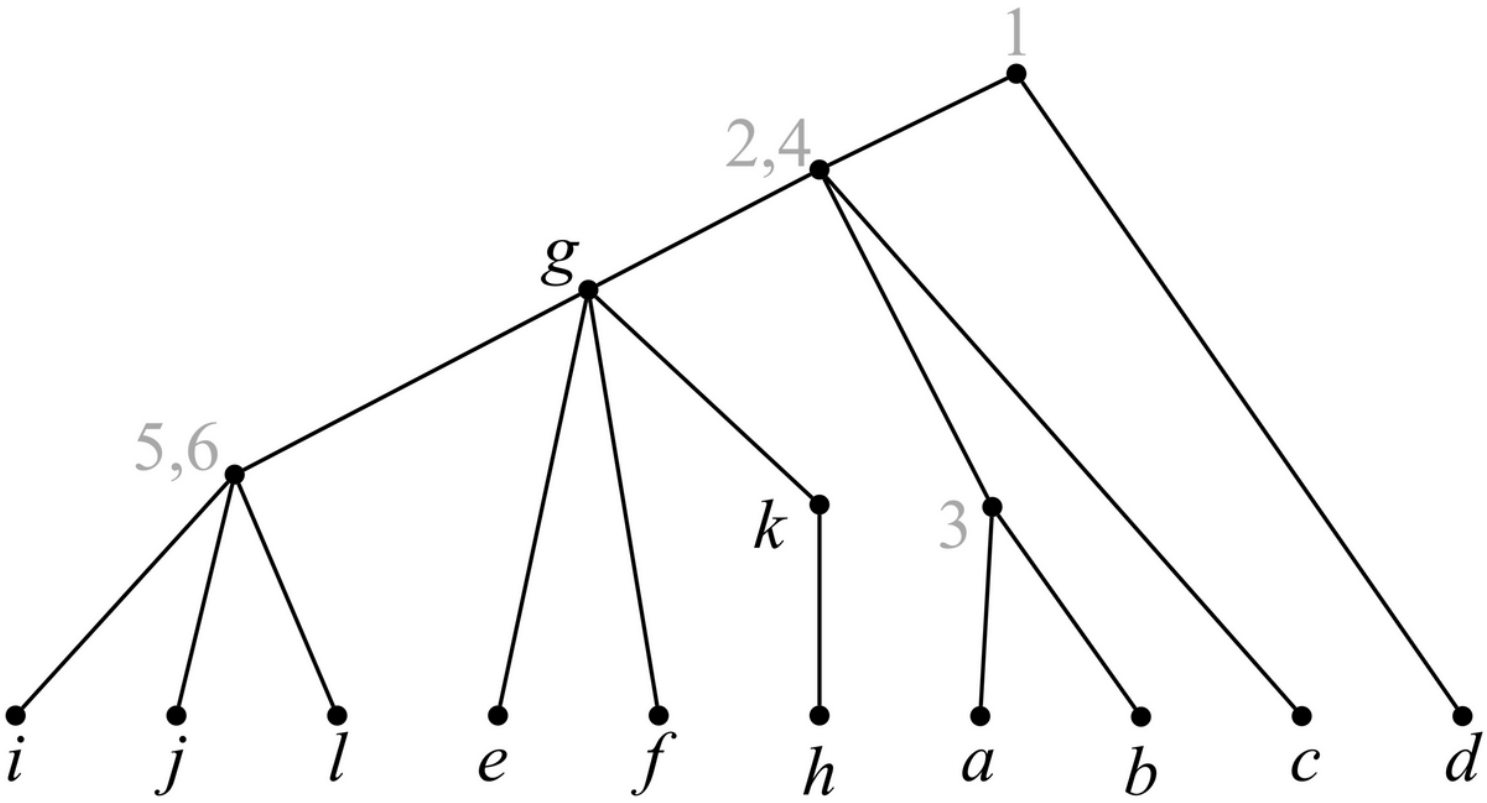

Figure 2

An agreement tree for P. Although all input trees are singularly labeled, the agreement tree is not.

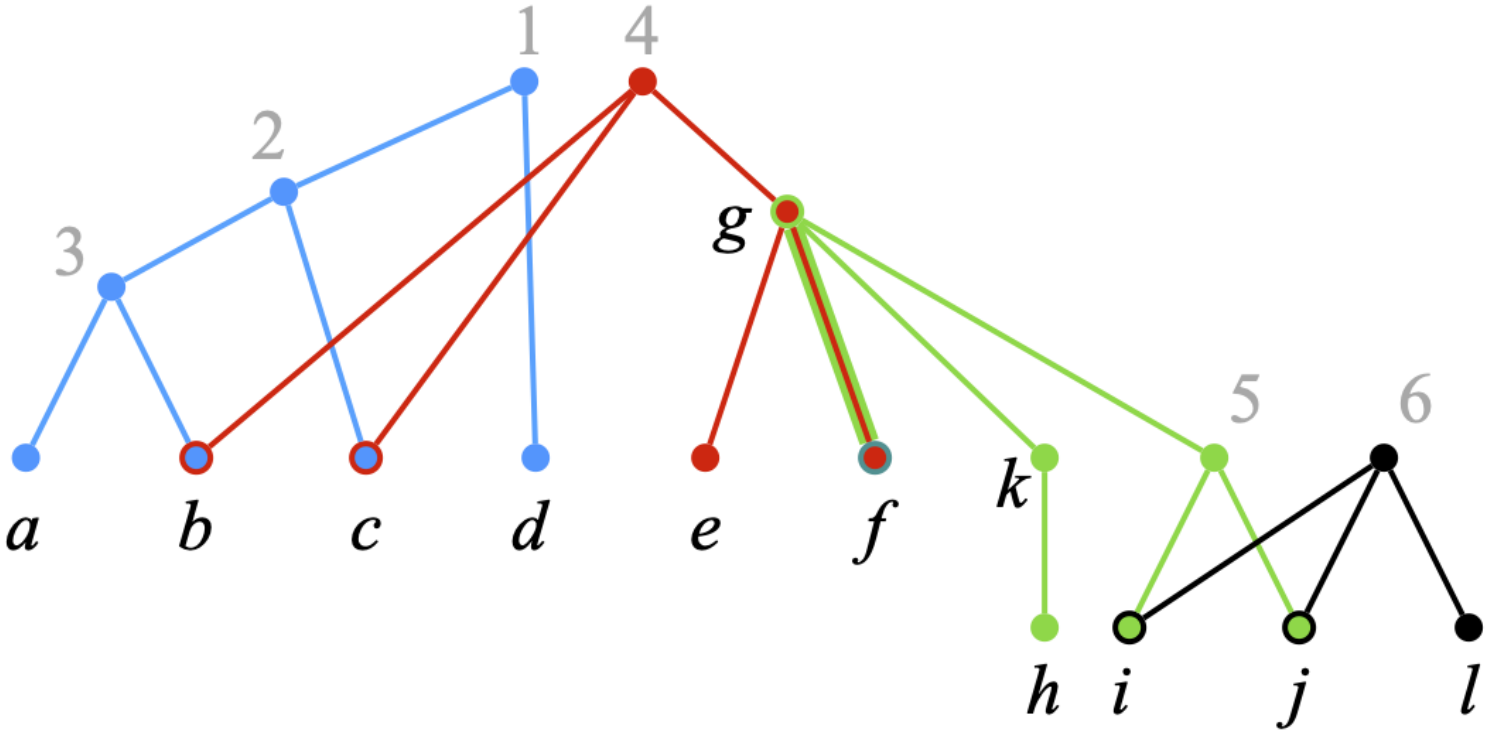

Figure 3

See legend for figure 3 in the manuscript file.

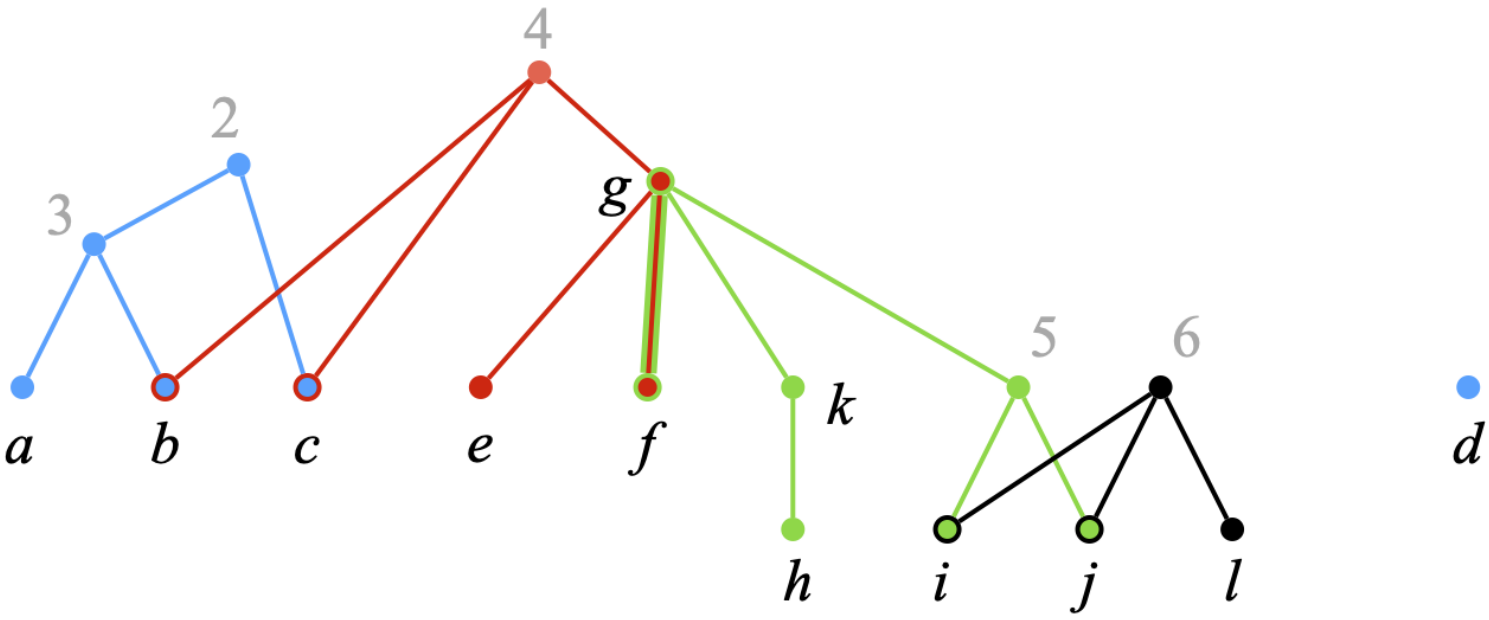

Figure 4

See legend for figure 4 in the manuscript file.

Figure 5

See legend for figure 5 in the manuscript file.

Figure 6

See legend for figure 6 in the manuscript file.

Figure 7

Runtime curves. Runtime curves for degree D = 2; 3; 10 and number of input trees k = 100.

Figure 8

Theoretical runtime against practical runtime Comparison of runtime among theoretical time bound, practical runtime with and without edge promotions of trees with degree D = 3, k = 100.

Figure 9

Runtime curves. Running times for proles of degrees 2; 3; and 10, with k varying from 20 to 200.

Figure 10

Runtime curves. Running times with and without edge promotion for trees of degree 3, with k varying from 20 to 300.