Algorithms for the Majority Rule (+) Consensus Tree and the Frequency Difference Consensus Tree

Jesper Jansson[1,⋆], Chuanqi Shen[2], and Wing-Kin Sung[3,4]

[1] Laboratory of Mathematical Bioinformatics (Akutsu Laboratory), Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
jj@kuicr.kyoto-u.ac.jp
[2] Stanford University, 450 Serra Mall, Stanford, CA 94305-2004, USA
shencq@stanford.edu
[3] School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417
ksung@comp.nus.edu.sg
[4] Genome Institute of Singapore, 60 Biopolis Street, Genome, Singapore 138672

Abstract. This paper presents two new deterministic algorithms for constructing consensus trees. Given an input of k phylogenetic trees with identical leaf label sets and n leaves each, the first algorithm constructs the *majority rule (+) consensus tree* in $O(kn)$ time, which is optimal since the input size is $\Omega(kn)$, and the second one constructs the *frequency difference consensus tree* in $\min\{O(kn^2), O(kn(k + \log^2 n))\}$ time.

1 Introduction

A *consensus tree* is a phylogenetic tree that summarizes a given collection of phylogenetic trees having the same leaf labels but different branching structures. Consensus trees are used to resolve structural differences between two or more existing phylogenetic trees arising from conflicts in the raw data, to find strongly supported groupings, and to summarize large sets of candidate trees obtained by bootstrapping when trying to infer a new phylogenetic tree accurately [2, 10, 12, 27].

Since the first type of consensus tree was proposed by Adams III [1] in 1972, many others have been defined and analyzed. See, e.g., [5], Chapter 30 in [12], or Chapter 8.4 in [27] for some surveys. Which particular type of consensus tree to use in practice depends on the context. For example, the *strict consensus tree* [25] is very intuitive and easy to compute [9] and may be sufficient when there is not so much disagreement in the data, the *majority rule consensus tree* [21] is "the optimal tree to report if we view the cost of reporting an estimate of the phylogeny to be a linear function of the number of incorrect clades in the estimate and the number of true clades that are missing from the estimate and we

⋆ Funded by The Hakubi Project and KAKENHI grant number 23700011.

A. Darling and J. Stoye (Eds.): WABI 2013, LNBI 8126, pp. 141–155, 2013.

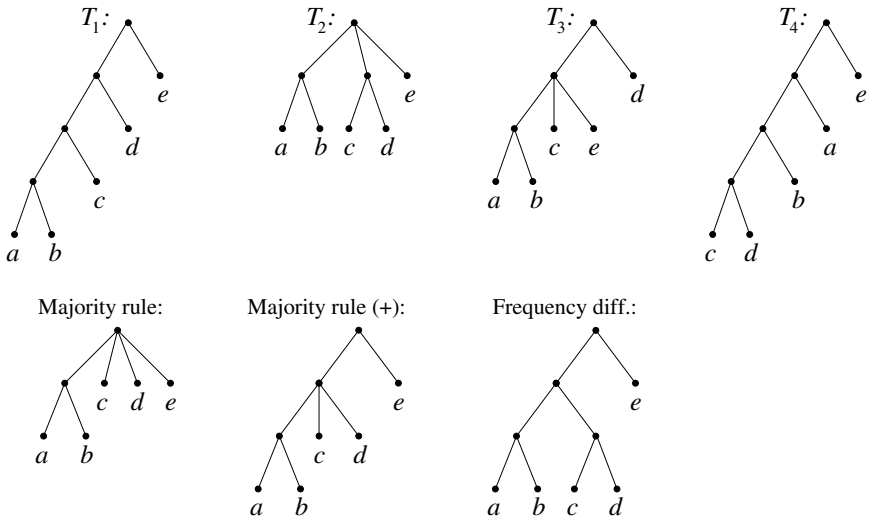

Fig. 1. Let $\mathcal{S} = \{T_1, T_2, T_3, T_4\}$ as shown above with $L = \Lambda(T_1) = \Lambda(T_2) = \Lambda(T_3) = \Lambda(T_4) = \{a, b, c, d, e\}$. The only non-trivial majority cluster of $\mathcal{S}$ is $\{a, b\}$, the non-trivial majority (+) clusters of $\mathcal{S}$ are $\{a, b\}$ and $\{a, b, c, d\}$, and the non-trivial frequency difference clusters of $\mathcal{S}$ are $\{a, b\}$, $\{a, b, c, d\}$, and $\{c, d\}$. The majority rule, majority rule (+), and frequency difference consensus trees of $\mathcal{S}$ are displayed.

view the reporting of an incorrect grouping as a more serious error than missing a clade" [16], and the *R* consensus tree* [5] provides a statistically consistent estimator of the species tree topology when combining gene trees [10]. Therefore, scientists need efficient algorithms for constructing a broad range of different consensus trees.

In a recent series of papers [8, 17–19], we have developed fast algorithms for computing the *majority rule consensus tree* [21], the *loose consensus tree* [4] (also known in the literature as the *combinable component consensus tree* or the *semi-strict consensus tree*), a *greedy consensus tree* [5, 13], the *R* consensus tree* [5], and consensus trees for so-called *multi-labeled phylogenetic trees (MUL-trees)* [20]. In this paper, we study two relatively new types of consensus trees called the *majority rule (+) consensus tree* [7, 11] and the *frequency difference consensus tree* [14], and give algorithms for constructing them efficiently.

1.1 Definitions and Notation

We shall use the following basic definitions. A *phylogenetic tree* is a rooted, unordered, leaf-labeled tree in which every internal node has at least two children and all leaves have different labels. (Below, phylogenetic trees are referred to as "trees" for short). For any tree T, the set of all nodes in T is denoted by $V(T)$ and the set of all leaf labels in T by $\Lambda(T)$. Any nonempty subset C of $\Lambda(T)$ is called a *cluster* of $\Lambda(T)$; if $|C| = 1$ or $C = \Lambda(T)$ then C is *trivial*, and otherwise,