Journal of Discrete Algorithms 25 (2014) 66–78

Contents lists available at ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

# Computing the rooted triplet distance between galled trees by counting triangles ☆

Jesper Jansson [a,*,1], Andrzej Lingas [b,2]

[a] Laboratory of Mathematical Bioinformatics, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[b] Department of Computer Science, Lund University, 22100 Lund, Sweden

ARTICLE INFO

*Article history:*
Available online 11 October 2013

*Keywords:*
Graph algorithm
Rooted triplet distance
Phylogenetic network comparison
Triangle counting
Matrix multiplication

ABSTRACT

We consider a generalization of the rooted triplet distance between two phylogenetic trees to two phylogenetic networks. We show that if each of the two given phylogenetic networks is a so-called galled tree with $n$ leaves then the rooted triplet distance can be computed in $o(n^{2.687})$ time. Our upper bound is obtained by reducing the problem of computing the rooted triplet distance between two galled trees to that of counting monochromatic and almost-monochromatic triangles in an undirected, edge-colored graph. To count different types of colored triangles in a graph efficiently, we extend an existing technique based on matrix multiplication and obtain several new algorithmic results that may be of independent interest: (i) the number of triangles in a connected, undirected, uncolored graph with $m$ edges can be computed in $o(m^{1.408})$ time; (ii) if $G$ is a connected, undirected, edge-colored graph with $n$ vertices and $C$ is a subset of the set of edge colors then the number of monochromatic triangles of $G$ with colors in $C$ can be computed in $o(n^{2.687})$ time; and (iii) if $G$ is a connected, undirected, edge-colored graph with $n$ vertices and $R$ is a binary relation on the colors that is computable in $O(1)$ time then the number of $R$-chromatic triangles in $G$ can be computed in $o(n^{2.687})$ time.



## 1. Introduction

Phylogenetic trees and their generalization to non-treelike structures, phylogenetic networks, are commonly used by scientists to describe evolutionary relationships [10,14,18,19,22]. In certain applications, it is necessary to compare two given phylogenetic trees and measure their (dis-)similarity, for example when evaluating methods for phylogenetic reconstruction [17] or querying phylogenetic databases [2]. Various ways of measuring the dissimilarity of two phylogenetic trees have been proposed and analyzed in the literature; see [2] and the references therein. One such measure is the *rooted triplet distance* [2,3,8,9], which counts the number of substructures (more precisely, subtrees induced by three leaves) that differ between the two trees. Intuitively, if the two trees are "similar" and share a lot of branching structure then this number will be small.

Formally, the rooted triplet distance is defined as follows. A *rooted phylogenetic tree* is an unordered, rooted tree in which every internal node has at least two children and all leaves are distinctly labeled. A rooted phylogenetic tree with three leaves is called a *rooted triplet*. A rooted triplet leaf-labeled by $\{a, b, c\}$ with exactly one internal node is called a *rooted fan*

☆ A preliminary version of this article appeared in Proceedings of the 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012), vol. 7354 of Lecture Notes in Computer Science, pp. 385–398, Springer-Verlag, Berlin, Heidelberg, 2012.

* Corresponding author.
*E-mail addresses:* jj@kuicr.kyoto-u.ac.jp (J. Jansson), Andrzej.Lingas@cs.lth.se (A. Lingas).

[1] Funded by The Hakubi Project and KAKENHI grant number 23700011.
[2] Research supported in part by VR grant 621-2008-4649.

1570-8667/$ – see front matter 
http://dx.doi.org/10.1016/j.jda.2013.10.002

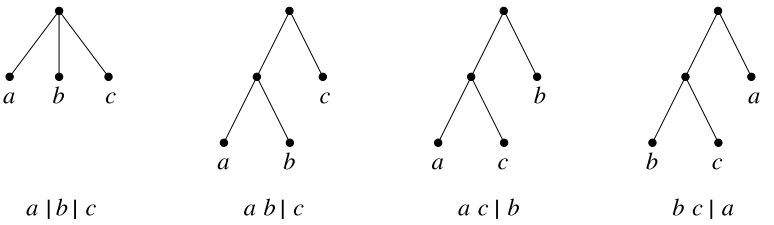


**Fig. 1.** The rooted fan triplet $a|b|c$ and the three rooted binary triplets $ab|c$, $ac|b$, and $bc|a$.

*triplet* and is denoted by $a|b|c$, and a rooted triplet leaf-labeled by $\{a, b, c\}$ with exactly two internal nodes is called a *rooted binary triplet*; in the latter case, there are three possibilities, denoted by $ab|c$, $ac|b$, and $bc|a$, corresponding to the three possible topologies. See Fig. 1 for an illustration. A rooted triplet $t$ is said to be *consistent with* a rooted phylogenetic tree $T$ if $t$ is an embedded subtree of $T$, i.e., $a|b|c$ is consistent with $T$ if $lca^T(a, b) = lca^T(a, c) = lca^T(b, c)$, and $ab|c$ is consistent with $T$ if $lca^T(a, b)$ is a proper descendant of $lca^T(a, c) = lca^T(b, c)$, where for any two leaf labels $x$ and $y$, $lca^T(x, y)$ denotes the lowest common ancestor in $T$ of the leaves labeled by $x$ and $y$. Now, given two rooted phylogenetic trees $T_1, T_2$ with the same set $L$ of leaf labels, the *rooted triplet distance* $d_{rt}(T_1, T_2)$ is the number of rooted triplets over $L$ that are consistent with exactly one of $T_1$ and $T_2$.

The rooted triplet distance was introduced by Dobson [9] in 1975. The naive algorithm for computing $d_{rt}(T_1, T_2)$ between two phylogenetic trees $T_1$ and $T_2$ with a leaf label set of cardinality $n$ runs in $O(n^3)$ time: Just preprocess $T_1$ and $T_2$ in $O(n)$ time so that lowest common ancestor queries can be answered in $O(1)$ time by the method in [13], and then check each of the $O(n^3)$ possible rooted triplets for consistency with $T_1$ and $T_2$ in $O(1)$ time. Critchlow et al. [8] provided a more efficient algorithm for computing the rooted triplet distance between two *binary* phylogenetic trees with $O(n^2)$ running time, and Bansal et al. [2] extended the $O(n^2)$-time upper bound to two phylogenetic trees of *arbitrary* degrees. The current record is held by Brodal et al. [3], who achieved a running time of $O(n \log n)$ for two phylogenetic trees of arbitrary degrees.

Due to the recently increasing popularity of the phylogenetic *network* model and its potential impact on evolutionary biology in the near future (see the two textbooks [14,18]), it is compelling to consider generalizations of the rooted triplet distance to the network case. As observed by Gambette and Huber [11], $d_{rt}$ can be canonically extended by replacing the two trees $T_1$ and $T_2$ in the definition above by two networks. However, for phylogenetic networks, it seems much harder to improve on the naive $O(n^3)$-time algorithm and to derive a subcubic upper bound on the running time. Therefore, one would like to know if any important special classes of phylogenetic networks such as the *galled trees* [12,14] admit fast algorithms. Galled trees are structurally restricted phylogenetic networks in which all underlying cycles are vertex-disjoint; for a formal definition, refer to Section 3.3 below. They constitute one of the simplest classes of phylogenetic networks and are useful in certain settings where reticulation events do occur but are known to be rare [12]. (See, e.g., Figure 9.22 in [14] for an example of a galled tree for a set of strains of *Fusarium graminearum*.) As a consequence, a number of algorithms for building galled trees from different kinds of data have been published [6,12,14–16].

In this article, we focus on the rooted triplet distance and describe how to compute it efficiently when the two input networks are galled trees. Several other measures of the dissimilarity between two phylogenetic networks, including *the Robinson–Foulds distance*, *the tripartitions distance*, *the $\mu$-distance*, *the nodal distance*, and *the split nodal distance*, were investigated for the special case of galled trees by Cardona et al. in [5]. (See [5] for the definitions of these measures and many references to the literature.)

### 1.1. New results

Our main contribution is an $o(n^{2.687})$-time algorithm for computing the rooted triplet distance between two galled trees with $n$ leaves each (Theorem 4). This breaks the natural $O(n^3)$-time barrier for any kind of non-tree phylogenetic networks for the first time. The precise running time of our algorithm is $O(n^{(3+\omega)/2})$, where $\omega$ denotes the exponent in the running time of the fastest existing method for matrix multiplication. It is currently known that $\omega < 2.373$ [25].

Theorem 4 is obtained in part by a reduction to the problem of counting monochromatic and "almost-monochromatic" triangles in an undirected graph with colored edges. To solve the latter problem quickly, we strengthen a technique based on matrix multiplication used in [1] and [24] for *detecting* if a graph contains a triangle to also *count* the number of triangles in the graph. More exactly, we show that:

- The number of triangles in a connected, undirected, uncolored graph with $m$ edges can be computed in $O(m^{\frac{2\omega}{\omega+1}}) \leqslant o(m^{1.408})$ time (Theorem 1).
- If $G$ is a connected, undirected, edge-colored graph with $n$ vertices and $C$ is a subset of the set of edge colors then the number of monochromatic triangles of $G$ with colors in $C$ can be computed in $O(n^{(3+\omega)/2}) \leqslant o(n^{2.687})$ time (Theorem 2).

We also need to relax the concept of a monochromatic triangle to what we call an $R$-chromatic triangle (see Section 2 for the definition), and obtain: