JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 12, 2021
Mary Ann Liebert, Inc.
Pp. 1181–1195
DOI: 10.1089/cmb.2021.0342

Research Articles

# The Generalized Robinson-Foulds Distance for Phylogenetic Trees

MERCÈ LLABRÉS,[1,2] FRANCESC ROSSELLÓ,[1,2] and GABRIEL VALIENTE[3,i]

## ABSTRACT

**The Robinson-Foulds (RF) distance, one of the most widely used metrics for comparing phylogenetic trees, has the advantage of being intuitive, with a natural interpretation in terms of common splits, and it can be computed in linear time, but it has a very low resolution, and it may become trivial for phylogenetic trees with overlapping taxa, that is, phylogenetic trees that share some but not all of their leaf labels. In this article, we study the properties of the Generalized Robinson-Foulds (GRF) distance, a recently proposed metric for comparing any structures that can be described by multisets of multisets of labels, when applied to rooted phylogenetic trees with overlapping taxa, which are described by sets of clusters, that is, by sets of sets of labels. We show that the GRF distance has a very high resolution, it can also be computed in linear time, and it is not (uniformly) equivalent to the RF distance.**

**Keywords:** metrics, phylogenetic tree, Robinson-Foulds distance.

## 1. INTRODUCTION

Trees are simple mathematical structures that are used to represent or model a relation on individuals. When those individuals are species and the relation is the sequence similarity of their genomes, the tree-like representation becomes a phylogenetic tree. Phylogenetic trees can also be inferred from metabolic pathways (Forst and Schulten, 2001; Chor and Tuller, 2007), protein–protein interaction networks (Erten et al., 2009), tumor clones of cancer (Miura et al., 2020), languages (Pompei et al., 2011), music (Bomin et al., 2016), etc., with appropriate distance or similarity relations. At the beginning, phylogenetic trees were designed to infer evolutionary relationships based on some species appearance (mainly morphological and physiological traits). However, the explosion of sequencing technologies yielding a vast amount of DNA and RNA sequence data has generated different methodologies to obtain such a phylogenetic tree, such as maximum parsimony and maximum likelihood approaches; see Bruyn et al. (2014) and Kapli et al. (2020) for an overview on these phylogenetic reconstruction methods), Bayesian inference (Rannala and

[1]Department of Mathematics and Computer Science, University of the Balearic Islands, Palma de Mallorca, Spain.
[2]Balearic Islands Health Research Institute (IdISBa), Palma, Spain.
[3]Department of Computer Science, Technical University of Catalonia, Barcelona, Spain.
[i]ORCID ID (https://orcid.org/0000-0001-9194-2703).

Yang, 1996, 1997; Mau et al., 1999; Li et al., 2000), and the neighbor-joining method (Saitou and Nei, 1987; Studier and Keppler, 1988). Each of them reconstructs a phylogenetic tree from a set of sequences; however, those phylogenetic trees may differ from one methodology to another, providing different trees for the same input data.

As a first solution to the disagreement on the different methodologies, consensus trees have been also defined and consensus tree methods have been implemented as well (Jansson et al., 2014, 2016). Nevertheless, one step further is when different experiments are performed, yielding different sequence data. In this case, and to analyze the evolutionary relationship of such sequences, the reconstructed phylogenetic trees must be compared. Reconstructed phylogenetic trees can be of any type: binary or multifurcating, labeled only on the leaves (taxa) or on all the nodes, with no repeated labels (injectively labelled) or with repeated labels, etc. Hence, different experiments may yield different phylogenetic trees of different types, such as phylogenetic trees with overlapping taxa as in the case of the experiment described later, where trees share some of the leaves labels but not all of them. Regarding a comparison of phylogenetic trees, several metrics have been defined as dealing with different types of trees. For phylogenetic trees injectively labeled on a set of taxa, that is, with no repeated labels, the most widely used distance is the Robinson-Foulds (RF) distance, which compares the clades (or clusters) of every node in the trees and counts how many of them are not shared by both trees. Many other metrics on phylogenetic trees have been proposed in the past few years; see Kuhner and Yamato (2015); Wang et al. (2020) for recent reviews.

Unlike phylogenetic trees reconstructed by different experiments on the same input data, phylogenetic trees reconstructed from different input data usually have overlapping taxa and thus, the comparison of phylogenetic trees with overlapping taxa is also of utmost importance. Figure 1 shows two alternative phylogenetic trees with overlapping taxa for several species of tomato (genus *Solanum*): a chloroplast DNA phylogeny, adapted from (Palmer and Zamir, 1982), and a mitochondrial DNA phylogeny, adapted from (McClean and Hanson, 1986); see also Baum and Ragan (2004). They overlap in all taxa but *S. juglandifolium* and *S. rickii*, shown to be highlighted. Both the chloroplast DNA phylogeny and the mitochondrial DNA phylogeny have 19 clusters each. They differ in all clusters but those of their nine common taxa, $\{1\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$, $\{7\}$, $\{8\}$, $\{9\}$, $\{10\}$, and cluster $\{9, 10\}$ and thus, their RF distance is $19+19-2\cdot 10=18$, which, normalized to $[0, 1]$, becomes $9/14 \approx 0.6429$. However, these phylogenies share several similar clusters, such as $\{3, 4, 7\}$ and $\{3, 5, 7\}$, or $\{3, 4, 5, 6, 7, 8, 9, 10\}$ and $\{3, 5, 6, 7, 8, 9, 10\}$. In fact, their Generalized Robinson-Foulds (GRF) distance, defined later in Section 3.1, is $797/266 \approx 2.9962$. Normalized to $[0, 1]$, it is $69,339/131,782 \approx 0.526164$.

## 2. BACKGROUND

### 2.1. The RF distance

Recall that the cluster (also, the clade or the monophyletic group) associated with a node in a phylogenetic tree is the set of descendant leaf labels of the node in the tree, and the cluster representation of a phylogenetic tree (Steel, 2016, §2.2) is the set of clusters for the nodes in the tree.

The RF distance (Robinson and Foulds, 1981), a widely used metric for comparing phylogenetic trees, can be computed in linear time in the size of the trees (Day, 1985; Pattengale et al., 2007). It was originally defined as the cardinality of the symmetric difference between the sets of clusters of the phylogenetic trees. When normalized to the unit interval, it becomes the Jaccard distance on these sets (Levandowsky and Winter, 1971): the one-complement of their Jaccard index (Jaccard, 1912).

Despite the wide acceptance over several decades now, the RF distance has some shortcomings. On the one hand, it has a very low resolution, because it can only take a small number of different values—the total

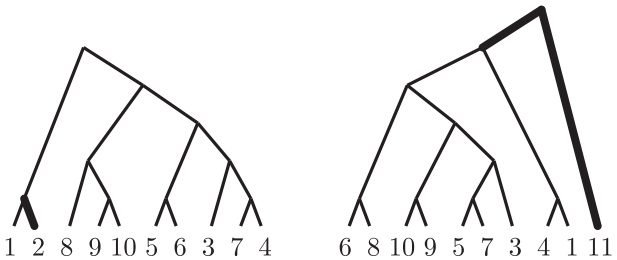


**FIG. 1.** Chloroplast DNA phylogeny (left) and mitochondrial DNA phylogeny (right) of several species of the genus *Solanum*. 1: *S. lycopersicoides*; 2: *S. juglandifolium*; 3: *S. peruvianum*; 4: *S. chilense*; 5: *S. pennellii*; 6: *S. hirsutum*; 7: *S. chmielewskii*; 8: *S. esculentum*; 9: *S. pimpinellifolium*; 10: *S. cheesmaniae*; 11: *S. rickii*.

number of leaves in the pair of compared trees—and it only takes into account whether two clusters are equal or not. Hence, when two clusters differ in only one label, this is equally counted as when they differ in all their labels (Section 3). On the other hand, and as a consequence of the latter, it may become trivial for phylogenetic trees with overlapping taxa, which differ in all their clusters except at most those consisting only of common labels.

### 2.2. Previous generalizations of the RF distance

Several generalizations of the RF distance have been proposed over the past few years, in an attempt to address the shortcomings discussed earlier. One approach consists of considering the distance between two phylogenetic trees as an optimal matching problem on a weighted complete bipartite graph, where the vertices correspond to the clusters of descendant node labels of the two phylogenetic trees. In this setting, in the RF distance (Robinson and Foulds, 1981), the edges are weighted by 1 (for different clusters) or 0 (for identical clusters). In the $\beta$ distance (Boorman and Olivier, 1973), which is basically the same as the matching cluster distance of (Bogdanowicz and Giaro, 2012b), each edge is weighted by the size of the symmetric difference of the pair of clusters it connects.

A related approach consists of matching each cluster in one tree to the most similar cluster in the other tree. In the cluster dissimilarity (CD) (Shuguang and Zhihui, 2015), the edges are also weighted by the size of the symmetric difference of the two clusters, and the distance between two phylogenetic trees is the sum of the minimum edge weights for the clusters of the first tree and the non-trivial clusters of the second tree, averaged with the sum of the minimum edge weights for the clusters of the second tree and the non-trivial clusters of the first tree.

Similar generalizations of the RF distance based on matching have also been proposed for unrooted phylogenetic trees (Boc et al., 2010; Bogdanowicz and Giaro, 2012a; Lin et al., 2012; Shuguang et al., 2014; Smith, 2020), and for trees with labeled internal nodes (Briand et al., 2020; Jahn et al., 2020). Further generalizations based on matching that take into account not only the clusters but also the structure of the phylogenetic trees have been proposed (Böcker et al., 2013; Borozan et al., 2019).

We have presented (Llabrés et al., 2020) a different generalization of the RF distance, based on the distances between sets of sets defined in Fujita (2013) and generalized to distances between multisets of multisets, which is a metric for the clonal trees (Govek et al., 2018; Karpov et al., 2019; DiNardo et al., 2020; Jahn et al., 2020) and the mutation trees (Kim and Simon, 2014; Aguse et al., 2019) that model tumor evolution under perfect phylogeny, phylogenetic trees, and several classes of phylogenetic networks, such as binary galled trees (Cardona et al., 2011), tree-child time-consistent phylogenetic networks (Cardona et al., 2008c, 2009a,b), and semi-binary tree-sibling time-consistent phylogenetic networks (Cardona et al., 2008a).

In this article, we further study the generalization of the RF distance when applied to phylogenetic trees, including theoretical properties, further empirical results, and more efficient algorithms. In comparison to previous generalizations, the latter has the advantage that it can be applied to phylogenetic trees over the same taxa and with overlapping taxa as well, that it has a much higher resolution, and that it, similar to the RF distance, can be computed in linear time.

### 2.3. Notation and basic results

Except when otherwise explicitly stated, all phylogenetic trees considered in this article are rooted phylogenetic trees. For every phylogenetic tree $T$, let: $V(T)$ be its set of nodes; $r_T$ its root; $V_{\text{int}}(T)$ its set of internal nodes; $L(T)$ its set of leaves, which we identify with their labels and hence we also understand $L(T)$ as its set of taxa; $C(T)$ its set of clusters; $C_T(v)$ the cluster of a node $v$; and $\delta_T(v)$ the depth of a node $v$ in $T$.

For every pair of sets $A, B$, their *symmetric difference* is $A \oplus B = (A \setminus B) \cup (B \setminus A)$. Thus, $|A \oplus B| = |A| + |B| - 2|A \cap B|$. Notice that $|A \oplus B| \leqslant |A| + |B|$, and $|A \oplus B| = |A| + |B|$ if, and only if, $A \cap B = \varnothing$.

For every pair of phylogenetic trees $T_1, T_2$ on the same set of taxa, their *RF distance* is

$$\mathrm{RF}(T_1, T_2) = \frac{|C(T_1) \oplus C(T_2)|}{|C(T_1) \cup C(T_2)|}.$$

## 3. THE GRF DISTANCE

In this section, we introduce the GRF distance, and we establish some properties of this distance comparing it with the RF distance.

### 3.1. Definition

In the original formulation, the GRF distance allowed for comparing any structures that can be described by sets or multisets of sets of multisets of labels (Llabrés et al., 2020). When restricted to phylogenetic trees (with possibly different sets of taxa), which are described by sets of sets of taxa, the GRF distance is defined as follows.

**Definition 1.** *Let $T_1$ and $T_2$ be phylogenetic trees, not necessarily on the same set of taxa. The GRF distance between $T_1$ and $T_2$ is given by*

$$\mathrm{GRF}(T_1, T_2) = \frac{\sum_{x \in C(T_1)} \sum_{y \in C(T_2) \setminus C(T_1)} |x \oplus y|}{|C(T_1) \cup C(T_2)| \cdot |C(T_1)|} + \frac{\sum_{x \in C(T_1) \setminus C(T_2)} \sum_{y \in C(T_2)} |x \oplus y|}{|C(T_1) \cup C(T_2)| \cdot |C(T_2)|}.$$

This GRF distance is, indeed, a metric, in the sense, that, for any phylogenetic trees $T_1, T_2, T_3$, the following properties hold:

**Separation** $\mathrm{GRF}(T_1, T_2) = 0$ if, and only if, $T_1 = T_2$,

**Symmetry** $\mathrm{GRF}(T_1, T_2) = \mathrm{GRF}(T_2, T_1)$, and

**Triangular inequality** $\mathrm{GRF}(T_1, T_3) \leqslant \mathrm{GRF}(T_1, T_2) + \mathrm{GRF}(T_2, T_3)$.

The proof of this fact is a simple application of Fujita, 2013, Theorem 1, taking into account that the cardinality of symmetric difference is a metric on sets and that every phylogenetic tree is characterized by its set of clusters.

**Example 1.** *Let $K_n$ be a caterpillar with n leaves, let $w_1, w_2, \ldots, w_{n-1}$ be its path of internal nodes from its root to the parent of its only cherry, and for every $i = 1, \ldots, n-2$ let i be the label of the leaf child of $w_i$; the children of $w_{n-1}$ are labeled with $n-1$ and n (*Fig. 2*). It is clear that*

$$C(K_n) = \{C(w_j) = \{j, j+1, \ldots, n\} | j = 1, \ldots, n-1\} \cup \{\{j\} | j = 1, \ldots, n\}.$$

*Now, take some $i \in \{2, \ldots, n-1\}$ and consider the tree $K_n^{(i)}$ obtained by collapsing the arc $(w_{i-1}, w_i)$* (Fig. 2). *Then, on the one hand,*

$$C(K_n^{(i)}) = C(K_n) \setminus \{\{i, i+1, \ldots, n\}\}$$

*and*

$$\mathrm{RF}(K_n, K_n^{(i)}) = \frac{|C(K_n) \oplus C(K_n^{(i)})|}{|C(K_n) \cup C(K_n^{(i)})|} = \frac{1}{2n-1}$$

*for every $i = 1, \ldots, n-1$. On the other hand,*

$$\mathrm{GRF}(K_n, K_n^{(i)}) = \frac{\sum_{C \in C(K_n^{(i)})} |\{\{i, i+1, \ldots, n\} \oplus C|}{|C(K_n) \cup C(K_n^{(i)})| \cdot |C(K_n^{(i)})|} =$$

$$= \frac{\sum_{j=1, j \neq i}^{n-1} |\{i, i+1, \ldots, n\} \oplus \{j, j+1, \ldots, n\}| + \sum_{j=1}^{n-1} |\{i, i+1, \ldots, n\} \oplus \{j\}|}{(2n-1)(2n-2)} \quad (1)$$

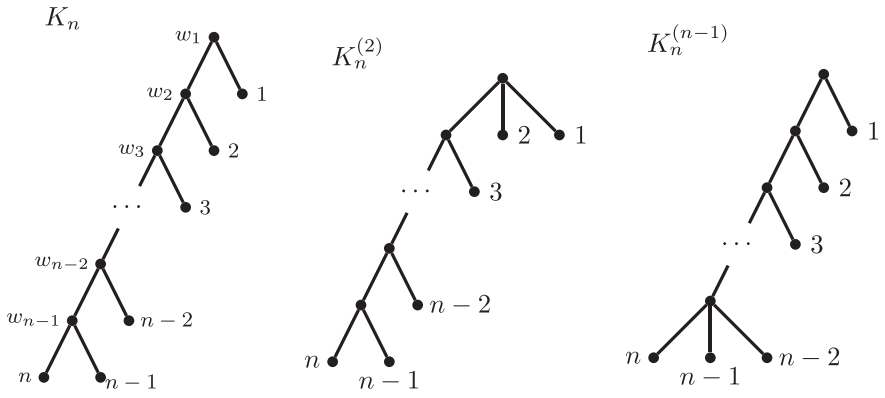


**FIG. 2.** The contraction of edge $e = (w_1, w_2)$ in the caterpillar on the left yields the phylogenetic tree $K_n^{(2)}$ at the center, and the contraction of edge $e' = (w_{n-2}, w_{n-1})$ in the caterpillar on the left yields the phylogenetic tree $K_n^{(n-1)}$ on the right.

*where*

$$|\{i, i+1, \ldots, n\} \oplus \{j, j+1, \ldots, n\}| = \begin{cases} |\{i, \ldots, j-1\}| = j-i & \text{if } j > i \\ |\{j, \ldots, i-1\}| = i-j & \text{if } j < i \end{cases}$$

$$|\{i, i+1, \ldots, n\} \oplus \{j\}| = \begin{cases} |\{i, \ldots, j-1, j+1, \ldots, n\}| = n-i & \text{if } j \geqslant i \\ |\{j, i, i+1, \ldots, n\}| = n-i+2 & \text{if } j < i \end{cases}$$

*So, returning to Eq. (1), the numerator of its right-hand side is*

$$\sum_{j=1}^{i-i} (i-j) + \sum_{j=i+1}^{n-1} (j-i) + \sum_{j=1}^{i-i} (n-i+2) + \sum_{j=i}^{n} (n-i) = \frac{(n-i)(3(n-i)+1)}{2}$$

*and hence, finally,*

$$\mathrm{GRF}(K_n, K_n^{(i)}) = \frac{(n-i)(3(n-i)+1)}{2(2n-1)(2n-2)}$$

*Therefore, contrary to what happens with the RF distance, all phylogenetic trees $K_n^{(i)}$ are at a different distance from $K_n$. Actually,*

$$\mathrm{GRF}(K_n, K_n^{(i+1)}) < \mathrm{GRF}(K_n, K_n^{(i)}), \quad i = 2, \ldots, n-2.$$

**Example 2.** *Consider now the caterpillar $K_{n-1}$ with $n-1$ leaves obtained by replacing the cherry $(n-1, n)$ at the bottom of $K_n$ by a single leaf $n-1$ (*Fig. 3*). If we were to compute the RF distance between $K_n$ and $K_{n-1}$, it would turn out that all internal nodes in both trees have different clusters. Therefore,*

$$C(K_n) \backslash C(K_{n-1}) = \{C_{K_n}(w_1), \ldots, C_{K_n}(w_{n-1}), \{n\}\}$$
$$C(K_{n-1}) \backslash C(K_n) = \{C_{K_{n-1}}(w_1), \ldots, C_{K_{n-1}}(w_{n-2})\}$$

*and hence,* $\mathrm{RF}(K_n, K_{n-1}) = \frac{2n-2}{3n-3} = \frac{2}{3}$.

*Now,*

$$\mathrm{GRF}(K_n, K_{n-1}) = \frac{\sum_{C \in C(K_{n-1}) \backslash C(K_n)} \sum_{C' \in C(K_n)} |C \oplus C'|}{|C(K_n) \cup C(K_{n-1})| \cdot |C(K_n)|} + \frac{\sum_{C' \in C(K_n) \backslash C(K_{n-1})} \sum_{C \in C(K_{n-1})} |C \oplus C'|}{|C(K_n) \cup C(K_{n-1})| \cdot |C(K_{n-1})|} =$$

$$= \frac{1}{(3n-3)(2n-1)} \Big( \sum_{i=1}^{n-2} \sum_{j=1}^{n-1} |C_{K_{n-1}}(w_i) \oplus C_{K_n}(w_j)| + \sum_{i=1}^{n-2} \sum_{j=1}^{n} |C_{K_{n-1}}(w_i) \oplus \{j\}| \Big)$$

$$+ \frac{1}{(3n-3)(2n-3)} \Big( \sum_{j=1}^{n-1} \sum_{i=1}^{n-2} |C_{K_{n-1}}(w_i) \oplus C_{K_n}(w_j)| + \sum_{i=1}^{n-2} |C_{K_{n-1}}(w_i) \oplus \{n\}|$$

$$+ \Big( \sum_{j=1}^{n-1} \sum_{i=1}^{n-1} |C_{K_n}(w_j) \oplus \{i\}| + \sum_{i=1}^{n-1} |\{i\} \oplus \{n\}| \Big)$$

*where*

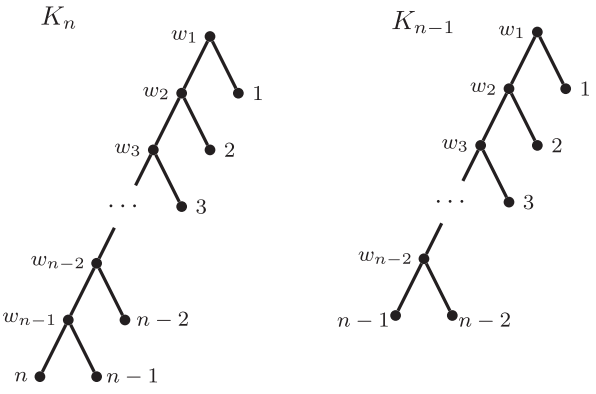


**FIG. 3.** The caterpillars with $n$ and $n-1$ leaves.

$$|C_{K_{n-1}}(w_i) \oplus C_{K_n}(w_j)| =$$

$$= |\{i, \ldots, n-1\} \oplus \{j, \ldots, n\}| = \begin{cases} |\{j, \ldots, i-1, n\}| = i-j+1 & \text{if } j < i \\ |\{n\}| = 1 = i-j+1 & \text{if } j = i \\ |\{i, \ldots, j-1, n\}| = j-i+1 & \text{if } j > i \end{cases}$$

$$|C_{K_{n-1}}(w_i) \oplus \{j\}| =$$

$$= |\{i, \ldots, n-1\} \oplus \{j\}| = \begin{cases} |\{j, i, i+1, \ldots, n-1\}| = n-i+1 & \text{if } j < i \\ |\{i, \ldots, j-1, j+1, \ldots, n-1\}| & \\ \qquad = n-i-1 & \text{if } i \leqslant j < n \\ |\{i, \ldots, n-1, n\}| = n-i+1 & \text{if } j = n \end{cases}$$

$$|C_{K_n}(w_j) \oplus \{i\}| =$$

$$= |\{j, \ldots, n\} \oplus \{i\}| = \begin{cases} |\{j, \ldots, i-1, i+1, \ldots, n\}| = n-j & \text{if } j \leqslant i \\ |\{i, j, j+1, \ldots, n\}| = n-j+2 & \text{if } j > i \end{cases}$$

$$|\{i\} \oplus \{n\}| = 2 \text{ for every } i \leqslant n-1.$$

*Therefore,*

$$\sum_{i=1}^{n-2} \sum_{j=1}^{n-1} |C_{K_{n-1}}(w_i) \oplus C_{K_n}(w_j)| =$$

$$= \sum_{i=1}^{n-2} \left( \sum_{j=1}^{i} (i-j+1) + \sum_{j=i+1}^{n} (j-i+1) \right) = \frac{n^3-n-6}{3}$$

$$\sum_{i=1}^{n-2} \sum_{j=1}^{n} |C_{K_{n-1}}(w_i) \oplus \{j\}| =$$

$$= \sum_{i=1}^{n-2} \left( \sum_{j=1}^{i-1} (n-i+1) + \sum_{j=i}^{n-1} (n-i-1) + n-i+1 \right) = \frac{(n+2)(n-1)(n-2)}{2}$$

$$\sum_{i=1}^{n-2} |C_{K_{n-1}}(w_i) \oplus \{n\}| =$$

$$= \sum_{i=1}^{n-2} (n-i+1) = \frac{(n+3)(n-2)}{2}$$

$$\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} |C_{K_n}(w_j) \oplus \{i\}| =$$

$$= \sum_{i=1}^{n-1} \left( \sum_{j=1}^{i} (n-j) + \sum_{j=i+1}^{n-1} (n-j+2) \right) = \frac{n(n^2+n-4)}{2}$$

$$\sum_{i=1}^{n-1} |\{i\} \oplus \{n\}| = 2(n-1)$$

*and, finally,*

$$\mathrm{GRF}(K_n, K_{n-1}) =$$

$$= \frac{1}{(3n-3)(2n-1)} \left( \frac{n^3-n-6}{3} + \frac{(n+2)(n-1)(n-2)}{2} \right) +$$

$$+ \frac{1}{(3n-3)(2n-3)} \left( \frac{n^3-n-6}{3} + \frac{(n+3)(n-2)}{2} + \frac{n(n^2+n-4)}{2} + 2(n-1) \right) =$$

$$= \frac{20n^4 - 14n^3 - 23n^2 - 43n + 42}{18(n-1)(2n-3)(2n-1)}$$

**Remark 1.** *The GRF distance can be easily generalized to unrooted phylogenetic trees, as follows. Let $T_1$ and $T_2$ be unrooted phylogenetic trees, not necessarily on the same set of taxa, and let $S(T_1), S(T_2)$ be their sets of splits. For each $A_1|B_1 \in S(T_1)$ and $A_2|B_2 \in S(T_2)$, let*

$$D^*(A_1|B_1, A_2|B_2) = \min\{|A_1 \oplus A_2| + |B_1 \oplus B_2|, |A_1 \oplus B_2| + |B_1 \oplus A_2|\}$$

*Then,*

$$\mathrm{GRF}(T_1, T_2) = \frac{\sum_{x \in S(T_1)} \sum_{y \in S(T_2) \setminus S(T_1)} D^*(x, y)}{|S(T_1) \cup S(T_2)| \cdot |S(T_1)|} + \frac{\sum_{x \in S(T_1) \setminus S(T_2)} \sum_{y \in S(T_2)} D^*(x, y)}{|S(T_1) \cup S(T_2)| \cdot |S(T_2)|}$$

*defines a GRF distance for unrooted phylogenetic trees.*

*The proof that this is a metric is again an application of (Fujita, 2013, Thm. 1), stating that every unrooted phylogenetic tree is characterized by its set of splits and that $D^*$ defines a distance on splits. This last assertion can be proved as follows. Since the cardinality of the symmetric difference is a metric on sets,*

$$D_2((A, B), (C, D)) = |A \oplus C| + |B \oplus D|$$

*is a metric on ordered pairs of sets. Now, each split $A|B$ can be understood as the set of ordered pairs $\{(A, B), (B, A)\}$. The distance $D^*$ is then simply (half) the Hausdorff distance between sets of this type induced by $D_2$.*

### *3.2. Theoretical properties*

Our first lemma deals with the metric equivalence between GRF and RF. Recall that two metrics $d_1$ and $d_2$ defined on a space $X$ are equivalent when there exist a pair of non-negative real numbers $\lambda, \mu$ such that, for every $x, y \in X$,

$$d_1(x, y) \leqslant \lambda \cdot d_2(x, y), \;\; d_2(x, y) \leqslant \mu \cdot d_1(x, y).$$

This definition captures the intuitive idea that both metrics define the same notion of ''closeness'' on $X$.

**Lemma 1.** *(a) If $T_1, T_2$ are phylogenetic trees on the same set of taxa, then* $\mathrm{RF}(T_1, T_2) \leqslant \mathrm{GRF}(T_1, T_2)$.
*(b) There is no constant $C \in \mathbb{R}$ such that, for every pair of phylogenetic trees $T_1, T_2$ on the same set of taxa,* $\mathrm{GRF}(T_1, T_2) \leqslant C \cdot \mathrm{RF}(T_1, T_2)$.

*Proof.* (a) Notice that

$$\mathrm{GRF}(T_1, T_2) \geqslant \frac{|V(T_1)| \cdot |C(T_2) \setminus C(T_1)|}{|V(T_1)| \cdot |C(T_1) \cup C(T_2)|} + \frac{|V(T_2)| \cdot |C(T_1) \setminus C(T_2)|}{|V(T_2)| \cdot |C(T_1) \cup C(T_2)|} = \frac{|C(T_1) \oplus C(T_2)|}{|C(T_1) \cup C(T_2)|} = \mathrm{RF}(T_1, T_2)$$

(b) Let $K_n$ be the caterpillar with $n$ leaves and $K_n^{(2)}$ the tree obtained from $K_n$ by collapsing the arc from the root to its internal child. By Example 1,

$$\mathrm{GRF}(K_n, \; K_n^{(2)}) = \frac{3(n^2 - 3n + 4)}{4(2n-1)(n-1)}$$

whereas

$$\mathrm{RF}(K_n, K_n^{(2)}) = \frac{1}{|C(K_n) \cup C(T_w)|} = \frac{1}{2n-1}$$

and there is no $C \in \mathbb{R}$ such that, for every $n \geqslant 3$,

$$\frac{n^2 - 3n + 4}{n - 1} \leqslant C.$$

□

The previous lemma entails that the GRF distance is not equivalent to the RF distance on the whole space of phylogenetic trees with any number of leaves. That is, although for every set of taxa, GRF and RF are equivalent metrics on the space of phylogenetic trees on this set of taxa, because all metrics on a given finite set are equivalent, any factor that transforms RF into a metric greater than GRF must depend on the number of taxa, being impossible to find a real number that works for every number of taxa. This is usually phrased by saying that RF and GRF are not uniformly equivalent (with respect to the number of leaves).

The following results will be used to show that, much like the RF distance, the GRF distance can also be computed in linear time in the size of the phylogenetic trees. Recall that the Sackin index of balance, introduced in Shao and Sokal (1990), for a phylogenetic tree $T$ is

$$S(T) = \sum_{x \in L(T)} \delta_T(x) = \sum_{v \in V_{\text{int}}(T)} |C(v)|$$

and set

$$\widehat{S}(T) = S(T) + |L(T)| = \sum_{x \in L(T)} (\delta_T(x) + 1) = \sum_{v \in V(T)} |C(v)|.$$

**Lemma 2.** *Let $T$ be a phylogenetic tree, and let $X$ be a set of labels. Then,*

$$\sum_{C \in C(T)} |C \oplus X| = \widehat{S}(T) + |V(T)| \cdot |X| - 2 \sum_{x \in X \cap L(T)} (\delta_T(x) + 1).$$

*Proof.* Since $|C \oplus X| = |C| + |X| - |C \cap X|$, we have that

$$\sum_{C \in C(T)} |C \oplus X| = \sum_{C \in C(T)} |C| + |C(T)| \cdot |X| - 2 \sum_{C \in C(T)} |C \cap X|$$

where

$$\begin{aligned} \sum_{C \in C(T)} |C \cap X| &= \sum_{C \in C(T)} \sum_{x \in X \cap L(T)} |C \cap \{x\}| \\ &= \sum_{x \in X \cap L(T)} \sum_{C \in C(T)} |C \cap \{x\}| = \sum_{x \in X \cap L(T)} (\delta_T(x) + 1) \end{aligned}$$

□

**Corollary 1.** *For every pair of phylogenetic trees $T_1, T_2$,*

$$\begin{aligned} \sum_{C_1 \in C(T_1)} \sum_{C_2 \in C(T_2) \setminus C(T_1)} |C_1 \oplus C_2| &= (|V(T_2)| - |C(T_1) \cap C(T_2)|) \widehat{S}(T_1) \\ &+ |V(T_1)| \cdot \widehat{S}(T_2) - |V(T_1)| \sum_{C \in C(T_1) \cap C(T_2)} |C| \\ &- 2 \sum_{x \in L(T_1) \cap L(T_2)} (\delta_{T_1}(x) + 1)(\delta_{T_2}(x) + 1) \\ &+ 2 \sum_{x \in L(T_1) \cap L(T_2)} |\{C \in C(T_1) \cap C(T_2) : x \in C\}| (\delta_{T_1}(x) + 1). \end{aligned}$$

*Proof.* By the previous lemma,

$$\begin{aligned} &\sum_{C_1 \in C(T_1)} \sum_{C_2 \in C(T_2) \setminus C(T_1)} |C_1 \oplus C_2| \\ &= \sum_{C_2 \in C(T_2) \setminus C(T_1)} (\widehat{S}(T_1) + |V(T_1)| \cdot |C_2| - 2 \sum_{x \in C_2 \cap L(T_1)} (\delta_T(x) + 1)) \end{aligned}$$

$$
\begin{aligned}
&= |C(T_2)\backslash C(T_1)| \cdot \widehat{S}(T_1) + |V(T_1)| \sum_{C_2 \in C(T_2)\backslash C(T_1)} |C_2| \\
&\quad - 2 \sum_{x \in L(T_1)} |\{C_2 \in C(T_2)\backslash C(T_1) : x \in C_2\}| (\delta_{T_1}(x) + 1) \\
&= (|V(T_2)| - |C(T_1) \cap C(T_2)|)\widehat{S}(T_1) + |V(T_1)| \cdot \widehat{S}(T_2) - |V(T_1)| \sum_{C \in C(T_1) \cap C(T_2)} |C| \\
&\quad - 2 \sum_{x \in L(T_1) \cap L(T_2)} (|\{C_2 \in C(T_2) : x \in C_2\}| - |\{C \in C(T_1) \cap C(T_2) : x \in C\}|)(\delta_{T_1}(x) + 1) \\
&= (|V(T_2)| - |C(T_1) \cap C(T_2)|)\widehat{S}(T_1) + |V(T_1)| \cdot \widehat{S}(T_2) - |V(T_1)| \sum_{C \in C(T_1) \cap C(T_2)} |C| \\
&\quad - 2 \sum_{x \in L(T_1) \cap L(T_2)} (\delta_{T_1}(x) + 1)(\delta_{T_2}(x) + 1) \\
&\quad + 2 \sum_{x \in L(T_1) \cap L(T_2)} |\{C \in C(T_1) \cap C(T_2) : x \in C\}| (\delta_{T_1}(x) + 1)
\end{aligned}
$$

□

### 3.3. Computation in linear time

Let $T_1$ and $T_2$ be phylogenetic trees, let $m = |L(T_1)|$, and let $n = |L(T_1) \cup L(T_2)|$. The cluster representation of $T_1$ and $T_2$ can be obtained during a postorder traversal of the trees (Llabrés et al., 2020, §2.2).

With a sorted list representation of the clusters, which uses $O(n^2)$ space and can be obtained in $O(n^2)$ time by radix sorting (Davis, 1992), and based on the idea behind the merge algorithm (Mehlhorn and Sanders, 2016, §5.2) of the simultaneous traversal of two sorted lists or arrays, the union and the symmetric difference of two clusters can be computed in $O(n)$ time and thus, the GRF distance can be computed in $O(n^3)$ time using $O(n^2)$ space by a direct implementation of the formula in Definition 1. See Llabrés et al. (2020) for details.

With a bit-vector representation of the clusters, which uses $O(n \lg n)$ space and can be obtained in $O(n \lg n)$ time, the GRF distance can be computed in $O(n^2 \lg n)$ time using $O(n \lg n)$ space by a direct implementation of the formula in Definition 1, where the union and the symmetric difference of two clusters are implemented by the OR and the XOR of the corresponding bit-vectors, respectively.

However, both the cluster representation of the trees and the intersection of the sets of clusters of the trees can be computed in $O(n)$ time with the algorithm of Day (1985), even if the trees have overlapping taxa, that is, when $m \neq n$. Then, it follows from Corollary 1 that the GRF distance can actually be computed in $O(n)$ time.

**Lemma 3.** *Let $T_1$ and $T_2$ be phylogenetic trees, and let $n = |L(T_1) \cup L(T_2)|$. Then,* $\mathrm{GRF}(T_1, T_2)$ *can be computed in $O(n)$ time.*

*Proof.* Let $T_1$ and $T_2$ be phylogenetic trees, let $m = |L(T_1)|$, let $n = |L(T_1) \cup L(T_2)|$, let $k = |L(T_1) \cap L(T_2)|$, and assume, without loss of generality, that $L(T_1) = \{1, \ldots, m\}$, and that $L(T_2) = \{1, \ldots, k, m+1, \ldots, n\}$. During a postorder traversal of $T_1$, both $\widehat{S}(T_1)$ and $(\delta_{T_1}(x) : x \in L(T_1))$ can be computed in $O(m)$ time and, during a postorder traversal of $T_2$, both $\widehat{S}(T_2)$ and $(\delta_{T_2}(y) : y \in L(T_2))$ can be computed in $O(n)$ time. On the other hand, $C(T_1)$, $C(T_2)$, and $C(T_1) \cap C(T_2)$ can be computed in $O(n)$ time, using the algorithm of Day (1985) and thus, both $|C(T_1) \cap C(T_2)|$ and $\sum_{C \in C(T_1) \cap C(T_2)} |C|$, and also $\sum_{i=1}^{k} (\delta_{T_1}(i) + 1)(\delta_{T_2}(i) + 1)$, can be computed in $O(n)$ time. Now, $H = C(T_1) \cap C(T_2)$ are the sets of clusters of a phylogenetic tree on $\{1, \ldots, m\}$ with $O(m)$ nodes and, therefore, $(\delta_H(1), \ldots, \delta_H(m))$ can be computed in $O(m)$ time. Then, $\sum_{x \in L(T_1) \cap L(T_2)} |\{C \in C(T_1) \cap C(T_2) : x \in C\}| (\delta_{T_1}(x) + 1) = \sum_{i=1}^{m} (\delta_H(i) + 1)(\delta_{T_1}(i) + 1)$, which shows that this sum can be computed in $O(m)$ time. Then, by Corollary 1, $\mathrm{GRF}(T_1, T_2)$ can be computed in $O(n)$ time. □

## 4. EXPERIMENTAL RESULTS

To study the resolution of the GRF distance and to compare it with the RF distance, we have performed a series of experiments on phylogenetic trees.

First of all, we have generated the 3 binary phylogenetic trees with 3 labeled leaves, the 15 binary phylogenetic trees with 4 labeled leaves, the 105 binary phylogenetic trees with 5 labeled leaves, and the 945 binary phylogenetic trees with 6 labeled leaves, using the algorithm described in Valiente (2009, §5.3.3), as implemented in Bio::Phylo (Cardona et al., 2008b; Vos et al., 2011).

Further, we have also generated 10,000 pairs of random binary phylogenetic trees with $n$ labeled leaves, for $n = 4, 6, 8, \ldots, 100$, using the algorithm of Quiroz (1989) to obtain Prüfer codes for random uniform phylogenetic trees and decoding them by using the algorithm of Caminiti et al. (2007).

Then, we computed both the RF distance and the GRF distance for every generated pair of phylogenetic trees.

### 4.1. Resolution of the metric

The *resolution* of a distance between phylogenetic trees on a set of labels is the number of different values taken by the distance upon all pairs of phylogenetic trees on that set of labels. When divided by the number of phylogenetic trees on that set of labels, it is the *recognition ratio* defined by Shao and Sokal (1986) for consensus indices between phylogenetic tree shapes.

We have computed the RF distance and the GRF distance between each pair of binary phylogenetic trees with the same number $n = 3, 4, 5, 6$ of labeled leaves, and also for a random uniform sample of 10,000 pairs of binary phylogenetic trees with $n = 4, 6, 8, \ldots, 100$ labeled leaves.

The resolution of the RF distance on binary phylogenetic trees with $n \geqslant 3$ labeled leaves is $n - 1$, even when normalized to the unit interval. The GRF distance, on the other hand, has a much higher resolution. As a matter of fact, there are $\Theta(n^2)$ different values for the GRF distance on binary phylogenetic trees with $n \geqslant 6$ labeled leaves. Table 1 shows the number of different values for the GRF distance on all pairs of binary phylogenetic trees with $n = 3, 4, 5, 6$ labeled leaves, and on a random uniform sample of 10,000 pairs of binary phylogenetic trees with $n = 4, 6, \ldots, 100$ labeled leaves.

Table 1 also shows the number of different values for two previous generalizations of the RF distance based on matching, the $\beta$ distance of Boorman and Olivier (1973), and the CD of Shuguang and Zhihui (2015). Although their resolution is slightly better than the resolution of the RF distance, there are still only $\Theta(n)$ different values for these previous generalizations of the RF distance on binary phylogenetic trees with $n$ labeled leaves.

### 4.2. Refinement of the RF distance

In this second test, we wanted to determine whether the GRF distance is a refinement of the RF distance. Hence, we first checked whether any triplet of phylogenetic trees, such that two of them are at the same GRF distance of the third one, is also at the same RF distance of the third one. We shall call *GRF-equidistant triplets* and *RF-equidistant triplets* those triplets of phylogenetic trees such that two of them are at the same GRF distance and RF distance of the third one, respectively. Thus, for all triplets of binary phylogenetic trees with $n = 3, 4, 5, 6$ labeled leaves, and for a random uniform sample of 10,000 triplets of binary phylogenetic trees with $n = 4, 6, 8, \ldots, 100$ labeled leaves, we computed the ratio of GRF-equidistant triplets to RF-equidistant triplets. That is, the number of GRF-equidistant triplets that are not RF-equidistant over the number of GRF-equidistant triplets, and the number of RF-equidistant triplets that are not GRF-equidistant over the number of RF-equidistant triplets.

As shown in Table 2, the first ratio (i.e., the number of GRF-equidistant triplets that are not RF-equidistant over the number of GRF-equidistant triplets) quickly converges to one as the number of labeled leaves increases, meaning that almost none of the GRF-equidistant triplets are RF-equidistant. This result reinforces the statement in the previous section that the GRF distance has a much higher resolution than the RF distance. On the other hand, the second ratio (i.e., the number of RF-equidistant triplets that are not GRF-equidistant over the number of RF-equidistant triplets) turned out to be zero except for phylogenetic trees with 6 leaves. Indeed, in the example given next, we show a triplet of phylogenetic trees with 6 leaves that is a GRF-equidistant triplet but it is not RF-equidistant. This means that the GRF distance is not a refinement of the RF distance. However, we can also observe in Table 2 that almost all the RF-equidistant triplets are GRF-equidistant as well.

**Example 3.** *The triplet of phylogenetic trees $T_1$, $T_2$, $T_3$ with 6 labeled leaves shown in* Figure 4 *has clusters* $C(T_1) = \{\{1\}, \{1, 2, 3, 4, 5, 6\}, \{1, 3, 4, 5, 6\}, \{1, 4, 5, 6\}, \{1, 5, 6\}, \{1, 6\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$, $C(T_2) = \{\{1\}, \{1, 2, 3, 4, 5, 6\}, \{1, 3, 4, 5\}, \{1, 4\}, \{1, 4, 5\}, \{2\}, \{2, 6\}, \{3\}, \{4\},$

Table 1. Number of Different Values Taken by the $\beta$ Distance, the Cluster Dissimilarity, and the Generalized Robinson-Foulds Distance, on All Pairs of Binary Phylogenetic Trees with $n = 3, 4, 5, 6$ Labeled Leaves (a) and for a Random Uniform Sample of 10,000 Pairs of Binary Phylogenetic Trees with $n = 4, 6, 8, \ldots, 100$ Labeled Leaves (b)

| | *No. of values* | | |
|---|---|---|---|
| *(a) n* | *β* | *CD* | *GRF* |
| 3 | 2 | 2 | 2 |
| 4 | 5 | 3 | 9 |
| 5 | 9 | 7 | 32 |
| 6 | 15 | 11 | 142 |

| | *No. of values* | | |
|---|---|---|---|
| *(b) n* | *β* | *CD* | *GRF* |
| 4 | 5 | 3 | 9 |
| 6 | 15 | 11 | 140 |
| 8 | 28 | 22 | 475 |
| 10 | 36 | 33 | 828 |
| 12 | 44 | 45 | 1245 |
| 14 | 57 | 56 | 1956 |
| 16 | 62 | 65 | 2458 |
| 18 | 76 | 80 | 3504 |
| 20 | 87 | 92 | 4114 |
| 22 | 101 | 100 | 5111 |
| 24 | 115 | 117 | 5711 |
| 26 | 129 | 134 | 6447 |
| 28 | 139 | 141 | 6882 |
| 30 | 158 | 162 | 7360 |
| 32 | 180 | 177 | 7702 |
| 34 | 192 | 195 | 8072 |
| 36 | 209 | 216 | 8276 |
| 38 | 228 | 230 | 8641 |
| 40 | 247 | 248 | 8708 |
| 42 | 268 | 264 | 8911 |
| 44 | 294 | 282 | 8980 |
| 46 | 303 | 312 | 9137 |
| 48 | 319 | 336 | 9256 |
| 50 | 350 | 354 | 9320 |
| 52 | 369 | 375 | 9365 |
| 54 | 384 | 402 | 9429 |
| 56 | 403 | 427 | 9480 |
| 58 | 425 | 456 | 9515 |
| 60 | 437 | 469 | 9605 |
| 62 | 470 | 492 | 9638 |
| 64 | 485 | 510 | 9614 |
| 66 | 523 | 547 | 9670 |
| 68 | 550 | 575 | 9689 |
| 70 | 559 | 609 | 9760 |
| 72 | 588 | 625 | 9738 |
| 74 | 623 | 665 | 9791 |
| 76 | 627 | 669 | 9785 |
| 78 | 671 | 690 | 9803 |
| 80 | 689 | 742 | 9809 |
| 82 | 699 | 762 | 9844 |
| 84 | 723 | 770 | 9802 |
| 86 | 745 | 805 | 9841 |

(*continued*)

Table 1. (Continued)

| | No. of values | | |
|---|---|---|---|
| *(b) n* | *β* | *CD* | *GRF* |
| 88 | 774 | 836 | 9850 |
| 90 | 792 | 857 | 9869 |
| 92 | 825 | 897 | 9888 |
| 94 | 865 | 946 | 9895 |
| 96 | 865 | 959 | 9900 |
| 98 | 890 | 994 | 9894 |
| 100 | 940 | 1,009 | 9903 |

CD, cluster dissimilarity; GRF, generalized Robinson-Foulds.

$\{5\}, \{6\}\}$, *and* $C(T_3) = \{\{1\}, \{1,2,3,4,5,6\}, \{1,6\}, \{2\}, \{2,3\}, \{2,3,4,5\}, \{2,3,5\}, \{3\}, \{4\}, \{5\}, \{6\}\}$. *Then,* $C(T_1)\backslash C(T_2) = \{\{1,3,4,5,6\}, \{1,4,5,6\}, \{1,5,6\}, \{1,6\}\}$, $C(T_2)\backslash C(T_1) = \{\{1,3,4,5\}, \{1,4\}, \{1,4,5\}, \{2,6\}\}$, $|C(T_1) \oplus C(T_2)| = 8$, $|C(T_1) \cup C(T_2)| = 15$, *and* $\mathrm{RF}(T_1, T_2) = 8/15$. *On the other hand,* $C(T_1)\backslash C(T_3) = \{\{1,3,4,5,6\}, \{1,4,5,6\}, \{1,5,6\}\}$, $C(T_3)\backslash C(T_1) = \{\{2,3\}, \{2,3,4,5\}, \{2,3,5\}\}$, $|C(T_1) \oplus C(T_3)| = 6$, $|C(T_1) \cup C(T_3)| = 14$, *and* $\mathrm{RF}(T_1, T_3) = 3/7$. *However,* $\mathrm{GRF}(T_1, T_2) = \mathrm{GRF}(T_1, T_3) = 17/11$. *In fact,*

Table 2. Ratio of Generalized Robinson-Foulds-Equidistant Triplets to Robinson-Foulds-Equidistant Triplets and Vice Versa, for All the Triplets of Binary Phylogenetic Trees with $n = 3, 4, 5, 6$ Labeled Leaves (a) and for a Random Uniform Sample of 10,000 Triplets of Binary Phylogenetic Trees with $n = 4, 6, 8, \ldots, 100$ Labeled Leaves (b)

| *(a) n* | *GRF vs. RF* | *RF vs. GRF* |
|---|---|---|
| 3 | 0.000000 | 0.000000 |
| 4 | 0.697417 | 0.000000 |
| 5 | 0.801518 | 0.000000 |
| 5 | 0.957795 | 0.000228 |
| *(b) n* | *GRF vs. RF* | *RF vs. GRF* |
| 4 | 0.616926 | 0.000000 |
| 6 | 0.950519 | 0.000000 |
| 8 | 0.990442 | 0.000157 |
| 10 | 0.994981 | 0.000000 |
| 12 | 0.997874 | 0.000000 |
| 14 | 0.998600 | 0.000000 |
| 16 | 0.998787 | 0.000000 |
| 18 | 0.999864 | 0.000000 |
| 20 | 1.000000 | 0.000000 |
| 22 | 0.999733 | 0.000000 |
| 24 | 0.999606 | 0.000000 |
| 26 | 0.999474 | 0.000000 |
| 28 | 1.000000 | 0.000000 |
| 30 | 0.999868 | 0.000000 |
| 32 | 1.000000 | 0.000000 |
| 34 | 1.000000 | 0.000000 |
| 36 | 1.000000 | 0.000000 |
| 38 | 0.999870 | 0.000000 |
| 40 | 1.000000 | 0.000000 |
| 42 | 0.999871 | 0.000000 |
| 44 | 1.000000 | 0.000000 |
| 46 | 1.000000 | 0.000000 |
| ... | 1.000000 | 0.000000 |
| 100 | 1.000000 | 0.000000 |

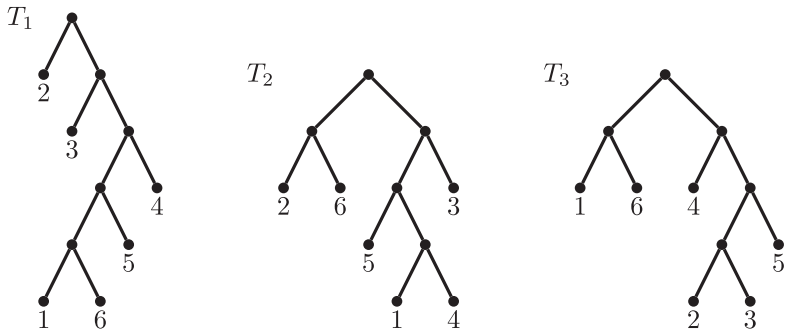


**FIG. 4.** Triplet of phylogenetic trees $T_1,\ T_2,\ T_3$ with $\mathrm{GRF}(T_1,\ T_2) = \mathrm{GRF}(T_1,\ T_3)$ but $\mathrm{RF}(T_1,\ T_2) \neq \mathrm{RF}(T_1,\ T_3)$.

$$\mathrm{GRF}(T_1, T_2) = \frac{123}{15 \cdot 11} + \frac{132}{15 \cdot 11} = \frac{17}{11} = \frac{119}{14 \cdot 11} + \frac{119}{14 \cdot 11} = \mathrm{GRF}(T_1, T_3).$$

## 5. DISCUSSION

Even though the RF distance is the most widely used distance for phylogenetic trees with no repeated labels, it has some drawbacks. First, it is only defined for pairs of trees on the same set of taxa. Second, it counts how many clades are shared by a pair of trees but, for the non-shared clades, it does not take into account how similar they are and, consequently, it has a very low resolution. In contrast to the RF distance, the GRF distance studied in this article happens to solve these shortcomings while keeping the advantages of the former one. First of all, the GRF distance allows for the comparison of any structures that can be described by multisets of multisets of labels. Thus, in the phylogenetic trees setting, phylogenetic trees are not restricted to be defined on the same set of taxa. In addition, for every pair of phylogenetic trees, it considers their shared clades but also, for the non-shared ones, it considers their dissimilarity, thus producing a distance with a high resolution. When restricted to phylogenetic trees on the same set of taxa, the tests presented in this study to compare both distances show that the GRF distance is nearly a refinement of the RF distance and it has a much higher resolution. As it is the case of the RF distance, the GRF distance can be computed in linear time and keeps the advantage of being intuitive, with a natural interpretation in terms of common splits.

Our current agenda involves the analysis of the topological behavior of the GRF distance, such as the metric diameter, that is, those phylogenetic trees at the maximum distance, the phylogenetic trees at minimum distance as well as the effect of elementary edit operations such as contracting an edge or removing a leaf, and rearrangement operations such as nearest-neighbor interchange, subtree pruning and regrafting, and tree bisection and reconnection (Allen and Steel, 2001) on the distance.

## AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

## FUNDING INFORMATION

This research was partially supported by the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund through project PGC2018-096956-B-C43 (FEDER/MICINN/AEI), and by the Agency for Management of University and Research Grants (AGAUR) through grant 2017-SGR-786 (ALBCOM).

## REFERENCES

Aguse, N., Qi, Y., and El-Kebir, M. 2019. Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. *Bioinformatics*. 35, i408–i416.

Allen, B.L., and Steel, M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Comb.* 5, 1–15.

Baum, B.R., and Ragan, M.A. 2004. The MRP method, chapter 1, 17–34. In Bininda-Emonds, O.R.P., ed. *Phylogenetic Supertrees: Combining information to reveal the Tree of Life, Volume 4 of Computational Biology*. Springer, Dordrecht, The Netherlands.

Boc, A., Philippe, H., and Makarenkov, V. 2010. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.* 59, 195–211.

Böcker, S., Canzar, S., and Klau, G.W. 2013. The generalized Robinson-Foulds metric, 156–169. In Darling, A., and Stoye, J., eds. *Proc. 13th Int. Workshop Algorithms in Bioinformatics, Volume 8126 of Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Bogdanowicz, D., and Giaro, K. 2012a. Matching split distance for unrooted binary phylogenetic trees. *IEEE ACM Trans. Comput. Biol.* 9, 150–160.

Bogdanowicz, D., and Giaro, K. 2012b. On a matching distance between rooted phylogenetic trees. *Int. J. Appl. Math. Comp.* 23, 669–684.

Bomin, S.L., Lecointre, G., and Heyer, E. 2016. The evolution of musical diversity: The key role of vertical transmission. *PLoS One*. 11, e0151570.

Boorman, S.A., and Olivier, D.C. 1973. Metrics on spaces of finite trees. *J. Math. Psychol.* 10, 26–59.

Borozan, L., Matijević, D., and Canzar, S. 2019. Properties of the generalized Robinson-Foulds metric, 330–335. In *Proc. 42nd Int. Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE, New York, NY.

Briand, S., Dessimoz, C., El-Mabrouk, N., et al. 2020. A generalized Robinson-Foulds distance for labeled trees. *BMC Genom.* 21(Suppl. 10):779.

Bruyn, A.D., Martin, D.P., and Lefeuvre, P. 2014. Phylogenetic reconstruction methods: An overview, chapter 13, 257–277. In Besse, P., ed. *Molecular Plant Taxonomy: Methods and Protocols, Volume 1115 of Methods in Molecular Biology*. Humana Press, New York, New York, USA.

Caminiti, S., Finocchi, I., and Petreschi, R. 2007. On coding labeled trees. *Theor. Comput. Sci.* 382, 97–108.

Cardona, G., Llabrés, M., Rosselló, F., and Valiente, G. 2009a. Metrics for phylogenetic networks i: Generalizations of the Robinson-Foulds metric. *IEEE ACM Trans. Comput. Biol.* 6, 46–61.

Cardona, G., Llabrés, M., Rosselló, F., and Valiente, G. 2011. Comparison of galled trees. *IEEE ACM Trans. Comput. Biol.* 8, 410–427.

Cardona, G., Rosselló, F., and Valiente, G. 2008a. A distance metric for a class of tree-sibling phylogenetic networks. *Bioinformatics*. 24, 1481–1488.

Cardona, G., Rosselló, F., and Valiente, G. 2008b. A Perl package and an alignment tool for phylogenetic networks. *BMC Bioinformatics*. 9, 175.

Cardona, G., Rosselló, F., and Valiente, G. 2008c. Tripartitions do not always discriminate phylogenetic networks. *Math. Biosci.* 211, 356–370.

Cardona, G., Rosselló, F., and Valiente, G. 2009b. Comparison of tree-child phylogenetic networks. *IEEE ACM Trans. Comput. Biol.* 6, 552–569.

Chor, B., and Tuller, T. 2007. Biological networks: Comparison, conservation, and evolution via relative description length. *J. Comp. Biol.* 14, 817–838.

Davis, I.J. 1992. A fast radix sort. *Comput. J.* 35, 636–642.

Day, W.H.E. 1985. Optimal algorithms for comparing trees with labeled leaves. *J. Classif.* 2, 7–28.

DiNardo, Z., Tomlinson, K., Ritz, A., and Oesper, L. 2020. Distance measures for tumor evolutionary trees. *Bioinformatics*. 36, 2090–2097.

Erten, S., Li, X., Bebek, G., et al. 2009. Phylogenetic analysis of modularity in protein interaction networks. *BMC Bioinformatics*. 10, 333.

Forst, C.V., and Schulten, K. 2001. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* 52, 471–489.

Fujita, O. 2013. Metrics based on average distance between sets. *Jpn. J. Ind. Appl. Math.* 30, 1–19.

Govek, K., Sikes, C., and Oesper, L. 2018. A consensus approach to infer tumor evolutionary histories, 63–72. In *Proc. 2018 ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics*. Association for Computing Machinery, New York, NY.

Jaccard, P. 1912. The distribution of flora in the alpine zone. *New Phytol.* 11, 37–50.

Jahn, K., Beerenwinkel, N., and Zhang, L. 2020. The Bourque distances for mutation trees of cancer, 14:1–14:23. In Pisanti, N. and Kingsford, C., eds. *Proc. 20th Int. Workshop Algorithms in Bioinformatics*, volume 172 of *Leibniz International Proceedings in Informatics*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

Jansson, J., Li, Z., and Sung, W.-K. 2014. On finding the Adams consensus tree. *Inform. Comput.* 256, 334–347.

Jansson, J., Shen, C., and Sung, W.-K. 2016. Improved algorithms for constructing consensus trees. *J. ACM*. 63, 28: 1–28:24.

Kapli, P., Yang, Z., and Telford, M. J. 2020. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444.

Karpov, N., Malikic, S., Rahman, M.K., and Sahinalp, S.C. 2019. A multi-labeled tree dissimilarity measure for comparing ''clonal trees'' of tumor progression. *Algorithms Mol. Biol.* 14, 17.

Kim, K.I., and Simon, R. 2014. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*. 15, 27.

Kuhner, M.K., and Yamato, J. 2015. Practical performance of tree comparison metrics. *Syst. Biol.* 64, 205–214.

Levandowsky, M., and Winter, D. 1971. Distance between sets. *Nature*. 234, 34–35.

Li, S., Pearl, D.K., and Doss, H. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Stat. Assoc.* 95, 493–508.

Lin, Y., Rajan, V., and Moret, B.M.E. 2012. A metric for phylogenetic trees based on matching. *IEEE ACM Trans. Comput. Biol*, 9, 1014–1022.

Llabrés, M., Rosselló, F., and Valiente, G. 2020. A generalized Robinson-Foulds distance for clonal trees, mutation trees, and phylogenetic trees and networks, 13:1–13:10. In *Proc. 11th ACM Int. Conf. Bioinformatics, Computational Biology and Health Informatics*. ACM Press, New York, NY.

Mau, B., Newton, M.A., and Larget, B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*. 55, 1–12.

McClean, P.E., and Hanson, M.R. 1986. Mitochondrial DNA sequence divergence among Lycopersicon and related Solanum species. *Genetics*. 112, 649–667.

Mehlhorn, K., and Sanders, P. 2016. *Algorithms and Data Structures: The Basic Toolbox*. Springer, Berlin, Heidelberg.

Miura, S., Vu, T., Deng, J., et al. 2020. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *Sci. Rep.* 10, 3498.

Palmer, J.D., and Zamir, D. 1982. Chloroplast DNA evolution and phylogenetic relationships in Lycopersicon. *Proc. Natl. Acad. Sci. USA*. 79, 5006–5010.

Pattengale, N.D., Gottlieb, E.J., and Moret, B.M.E. 2007. Efficiently computing the Robinson-Foulds metric. *J. Comput. Biol.* 14, 724–735.

Pompei, S., Loreto, V., and Tria, F. 2011. On the accuracy of language trees. *PLoS One*. 6, e20109.

Quiroz, A.J. 1989. Fast random generation of binary, *t*-ary and other types of trees. *J. Classif.* 6, 223–231.

Rannala, B., and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311.

Rannala, B., and Yang, Z. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14, 717–724.

Robinson, D.F., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Shao, K.-T., and Sokal, R.R. 1986. Significance tests of consensus indices. *Syst. Zool.*, 35, 582–590.

Shao, K.-T., and Sokal, R.R. 1990. Tree balance. *Syst. Zool.* 39, 266–276.

Shuguang, L., Shuying, C., and Mengtian, C. 2014. On the theoretical properties of bipartition dissimilarity measure. *Comput. Model. New Technol.* 18(12A), 322–327.

Shuguang, L., and Zhihui, L. 2015. Algorithms for computing cluster dissimilarity between rooted phylogenetic trees. *Open Cybern. Syst. J.* 9:2218–2223.

Smith, M.R. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics*. 36, 5007–5013.

Steel, M. 2016. *Phylogeny: Discrete and Random Processes in Evolution*. SIAM, Philadelphia, PA.

Studier, J.A., and Keppler, K.J. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* 5, 729–731.

Valiente, G. 2009. *Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R*. Chapman & Hall/CRC, Boca Raton, FL.

Vos, R.A., Caravas, J., Hartmann, K., et al. 2011. Bio::Phylo: Phyloinformatic analysis using Perl. *BMC Bioinformatics*. 12, 63.

Wang, J., Qi, X., Cui, B., and Guo, M. 2020. A survey of metrics measuring difference for rooted phylogenetic trees. *Curr. Bioinform.* 15, 697–702.

Address correspondence to:
*Dr. Gabriel Valiente*
*Department of Computer Science*
*Technical University of Catalonia*
*Barcelona E-08034*
*Spain*

*E-mail:* gabriel.valiente@upc.edu