RESEARCH ARTICLE

A strict upper bound for the partition distance and the cluster distance of phylogenetic trees for each fixed pair of topological trees

Martin Middendorf *, Nicolas Wieseke

Swarm Intelligence and Complex Systems Group, Faculty of Mathematics and Computer Science, University of Leipzig, Leipzig, Germany

* middendorf@informatik.uni-leipzig.de

Abstract

For each given pair of (rooted or unrooted) topological trees with the same number of leaves a strict upper bound is shown for the tree partition distance (also called symmetric difference metric and Robinson-Foulds distance)—in case of unrooted trees—and for the cluster distance (also called Robinson-Foulds distance)—in case of rooted trees—of corresponding phylogenetic trees. In particular, it is shown that there exist assignments of labels (e.g., species) to the leaves of both topological tree where each label is assigned to exactly one leaf in each tree such that: i) in the unrooted case, the tree partition distance between the corresponding phylogenetic trees equals the number of internal edges in both trees minus the number of nodes with degree 2 in both trees, ii) in the rooted case, the cluster distance between any two corresponding phylogenetic trees equals the number of internal edges in both trees minus the number of nodes with degree 2 in both trees, and iii) the values in (i) and (ii) are also the maximum values with respect to all possible assignments. The shown strict worst case bounds are needed as normalization factor to compute a normalized version of the respective tree partition metrics.

OPEN ACCESS

Citation: Middendorf M, Wieseke N (2018) A strict upper bound for the partition distance and the cluster distance of phylogenetic trees for each fixed pair of topological trees. PLoS ONE 13(9): e0204907. <https://doi.org/10.1371/journal.pone.0204907>

Editor: Arndt von Haeseler, Max F Perutz Laboratories GmbH, AUSTRIA

Received: July 5, 2018

Accepted: September 10, 2018

Published: September 28, 2018

Copyright: © 2018 Middendorf, Wieseke. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Tree distance measures are used in many applications to compare trees. The most widely used difference measure between (unrooted) phylogenetic trees (i.e., trees where the leaves are labelled with species) is the tree partition distance (TPD). The TPD was introduced by Bourque [1] and is also called symmetric difference metric or Robinson-Foulds metric [2] (Note, that the latter name has also been used for another tree metric that was proposed in [3]). The TDP measures the size of the symmetric difference between the two sets of bi-partitions (one set for each tree) of all labels that are obtained when for each inner edge of a tree the following is done: the edge is removed and for each of the two emerging connected components the set of labels that are assigned to its leaves is one set of the bipartition. Several authors have studied the TPD (e.g., [4–7]).

For comparing the distances between pairs of phylogenetic trees of different sizes normalized versions of tree distance measures are used. The most often used normalization factors

are strict worst case bounds. Normalized versions of the TPD have been used in several evolutionary studies (e.g., [8, 9]). To define them let $d(T_1, T_2)$ be the TPD between two phylogenetic trees T_1 and T_2 which have the same number of leaves. One normalized version of the TPD (used, e.g., in [9, 10]) is then to divide the value $d(T_1, T_2)$ by the maximum TPD for phylogenetic trees with the same number of leaves as T_1 and T_2 . It is easy to show that maximum TPD for trees with n leaves is $2n - 6$ (e.g., [6]). Instead of using a strict worst case bound as normalization factor another possibility is to use the average TPD over pairs of trees with size n . Specifically, in [11] the normalized TPD of trees T_1 and T_2 is computed as $(d_{rand} - d(T_1, T_2))/d_{rand}$ where d_{rand} is the average TPD computed empirically over 1000 pairs of random trees of size n (in [11] the trees were generated with the Yule model according to a general proposal for computing normalized tree distance metrics from [12]). It should be mentioned that sometimes also $TPD/2$ is called normalized Robinson-Foulds distance (e.g., in [13]). Observe, that in all these definitions of normalized versions of TPD the normalization factor does not depend on the topologies of phylogenetic trees T_1 and T_2 .

In [8] it was argued, however, that the normalization factor should consider the topologies of phylogenetic trees since not all possible phylogenetic trees are biologically relevant. Thus, a new variant of normalized TPD was proposed in [8]. In this version—denoted nTPD—value $d(T_1, T_2)$ is divided by the worst case TPD between any two phylogenetic trees that have the same topology as T_1 and T_2 (but a possibly different assignment of labels to the leaves). Let $w(T_1, T_2)$ denote this worst case value. Clearly, $w(T_1, T_2) \leq 2n - 6$ holds. Unfortunately and to the best of our knowledge, no explicit formula for $w(T_1, T_2)$ is known and it is not feasible for larger n to check all possible assignments of labels to the leaves of trees with the same topology as T_1 and T_2 . Therefore, the computation of the nTPD in the NELSI [14] R package (function `dist.topo.normalised`) is (only) approximated. This is done as follows. First the TPD is computed for several pairs of phylogenetic trees that have been obtained from the two given topological trees by randomizing the assignments of labels to the leaves for one of the trees. Then the maximum TPD over all randomized pairs of phylogenetic trees is used as an approximation for the worst case TPD. In [8] the approximated nTPD was used to compare co-phylogenetic systems of different sizes and it was argued that the maximum over randomized 1000 pairs should give a good approximation. In particular, it was shown empirically for pairs of trees with sizes up to $n = 142$ that the maximum over 1000 randomizations is stable (i.e., fewer randomizations had already given the same maximum value).

In this paper we present an explicit formula to compute $w(T_1, T_2)$. It should be mentioned that a $O(n^5)$ time method is given in [4] to compute for a fully resolved (i.e., each inner node has degree three) phylogenetic tree T with n leaves for all values $m \in [0: 2n - 6]$ the number of phylogenetic trees that have TPD m to T . By taking the maximum value m for which this number is not zero one obtains the worst case TPD for T (with respect to all other fully resolved trees with n leaves). But this is different from the worst case bound shown in this paper where both topological trees are fixed. Moreover, we also consider the case of trees that are not fully resolved.

In addition to the case of unrooted phylogenetic trees we also consider the case of rooted phylogenetic trees. In the unrooted case the corresponding distance measure is called cluster distance (CD) or Robinson-Foulds distance for rooted trees. Recently a Robinson-Foulds metric has been proposed in [15] to compare an unrooted phylogenetic tree T_1 with a rooted phylogenetic tree T_2 when both trees have n leaves. The idea of this measure is to root the unrooted tree T_1 optimally in the sense that the obtained rooted tree T'_1 has a minimal CD distance to T_2 . Similar, as for TPD we show a strict worst case bound for CD and for the Robinson-Foulds distance between an unrooted phylogenetic tree and a rooted phylogenetic tree. Our results imply that the corresponding normalized distances can be computed efficiently.

Basic definitions

An (unrooted) *tree* is a connected graph $T = (V, E)$ with $n = |V|$ nodes and $n - 1 = |E|$ edges. A rooted tree is a tree which has one distinguished node with degree ≥ 2 that is called root. For a tree $T = (V, E)$ a node $v \in V$ is *connected* to a node $w \in V$ when $\{v, w\} \in E$. A *leaf* is a node in a (rooted or unrooted) tree with degree one, a *leaf-edge* is an edge that is incident to a leaf and all other edges are *internal edges*. For an unrooted tree T a node that is not a leaf is an *internal node*. For a rooted tree T a node that is neither a leaf nor the root is an *internal node*. A *proper* (rooted or unrooted) tree is a tree T where each internal node has degree ≥ 3 . Let $\mathcal{T}(n, m)$ ($\mathcal{T}_r(n, m)$) be the set of all unrooted (respectively, rooted) trees with n leaves and m internal edges, $n \geq 3$. For a proper unrooted tree $T \in \mathcal{T}(n, m)$ it holds that $0 \leq m \leq n - 3$ and for a proper rooted tree $T \in \mathcal{T}_r(n, m)$ it holds that $0 \leq m \leq n - 2$.

Let L be a set of n labels. In this paper we assume w.l.o.g. $L = \{1, 2, \dots, n\}$. A *phylogenetic tree* on L is a tree T with n leaves and where each leaf is labelled with exactly one element from L such that for each label $l \in L$ there exists a leaf with label l . For a phylogenetic tree the underlying tree T is also called the *topological tree*, i.e., the topological tree is the phylogenetic tree ignoring the labels of the leaves. A rooted or unrooted phylogenetic tree is *proper* if its corresponding rooted, respectively, unrooted topological tree is proper. If the context is clear notation T is used for a phylogenetic tree and also for its corresponding topological tree.

The removal of an edge e from an unrooted phylogenetic tree T on L induces a two set partition of L —denoted by $\pi(T, e)$ —where each set of the partition corresponds to the labels of all nodes of one of the two connected components of $T - e$. Observe that for each leaf-edge e of T one set of the partition $\pi(T, e)$ is a singleton that contains the label of the corresponding leaf. A two set partition of $\{1, 2, \dots, n\}$ where one set is a singleton is called *trivial partition*. For each internal edge e of a proper unrooted phylogenetic tree T each set of the partition $\pi(T, e)$ has at least two elements. Let $P(T)$ be the set of all non trivial two set partitions of an unrooted phylogenetic tree T . For two unrooted phylogenetic trees T_1 and T_2 with n leaves the *tree partition distance* (TPD) between T_1 and T_2 —denoted by $d(T_1, T_2)$ —is the size of the symmetric difference between $P(T_1)$ and $P(T_2)$, i.e., $d(T_1, T_2) = |P(T_1) \cup P(T_2) - (P(T_1) \cap P(T_2))|$.

For each node v of a rooted phylogenetic tree T let $T(v)$ be the subtree of T with root v and let $cl(T, v)$ be the subset of L that contains all labels of the leaves of $T(v)$. Set $cl(T, v)$ is called the *cluster* of v . Observe that for each leaf v the cluster $cl(T, v)$ is a singleton that contains the label of v . If v is the root of T then $cl(T, v) = L$. A cluster that is a singleton or equals L is called a *trivial cluster*. For each internal node v of a proper rooted phylogenetic tree T the cluster of v has at least two elements. Let $Cl(T)$ be the set of all non trivial clusters of a rooted phylogenetic tree T . For two rooted phylogenetic trees T_1 and T_2 with n leaves the *cluster distance* (CD) between T_1 and T_2 —denoted by $d_r(T_1, T_2)$ —is the size of the symmetric difference between $Cl(T_1)$ and $Cl(T_2)$, i.e., $d_r(T_1, T_2) = |Cl(T_1) \cup Cl(T_2) - (Cl(T_1) \cap Cl(T_2))|$.

Let T be an unrooted phylogenetic tree. Let $E(T)$ be the edge set of T . A rooting of T is defined by choosing an edge $e = \{n, n'\}$ of T on which the root is to be placed, i.e., the edge is removed from T , a new node n'' that is the root is added to T , and n'' is connected to n and to n' . The obtained rooted tree is denoted by T_e . According to [15] for an unrooted phylogenetic tree T_1 and a rooted phylogenetic tree T_2 both with n leaves the *unrooted cluster distance* (urCD) between T_1 and T_2 —denoted by $d_{ur}(T_1, T_2)$ —is defined as $d_{ur}(T_1, T_2) = \min_{e \in E(T_1)} |Cl(T_{1,e}) \cup Cl(T_2) - (Cl(T_{1,e}) \cap Cl(T_2))|$.

Results and discussion

For each two proper unrooted topological trees $T_1 \in \mathcal{T}(n, m_1)$ and $T_2 \in \mathcal{T}(n, m_2)$ the following upper bound on the TPD of two corresponding phylogenetic trees holds: $d(T_1, T_2) \leq m_1 + m_2$.

This is clear because each internal edge can lead to at most one two set partition. If both trees are not necessarily proper and $T_i, i \in \{1, 2\}$ has k_i nodes with degree 2 then $d(T_1, T_2) \leq m_1 - k_1 + m_2 - k_2$ holds. This result follows from the upper bound for proper trees and the fact that both edges that are incident to a node of degree 2 lead to the same two set partition of L when they are removed. Similarly, for two rooted topological trees $T_1 \in \mathcal{T}_r(n, m_1)$ and $T_2 \in \mathcal{T}_r(n, m_2)$ where k_i is the number of nodes of degree 2 in $T_i, i \in \{1, 2\}$ the following upper bound on the CD of two corresponding phylogenetic trees holds: $d_r(T_1, T_2) \leq m_1 - k_1 + m_2 - k_2$. In the rest of this section we show that these upper bounds are all strict for each two unrooted (respectively rooted) topological trees T_1 and T_2 in the sense that there exist labelings of the leaves of T_1 and T_2 such that the TPD (respectively CD) of the corresponding phylogenetic trees equals the upper bound.

Theorem 1. *For each two proper unrooted topological trees $T_1 \in \mathcal{T}(n, m_1)$ and $T_2 \in \mathcal{T}(n, m_2)$ there exist $\{1, 2, \dots, n\}$ -labelings of the leaves of both trees such that $d(T_1, T_2) = m_1 + m_2$ with $n \geq 3, 0 \leq m_1 \leq n - 3, 0 \leq m_2 \leq n - 3$.*

Proof. For the proof we show that there exist labelings for the leaves of T_1 and T_2 such that for each internal edge e_1 of T_1 the partition $T_1 - e_1$ is not contained in $P(T_2)$ and for each internal edge e_2 of T_2 the partition $T_2 - e_2$ is not contained in $P(T_1)$. The proof is done by induction on n .

Base case $n = 3$. Since each internal node has degree three there exists exactly one internal node in each of the trees T_1 and T_2 . Thus, $m_1 = m_2 = 0$ holds and the result follows immediately.

For the inductive step consider two trees T_1 and T_2 with $n \geq 4$ leaves and the following three cases:

1. (At least) one of the trees has only one internal node and therefore has no internal edge.
2. Both trees have (at least) one internal edge and (at least) one internal node that is connected to at least 3 leaves.
3. Both trees have (at least) one internal edge and for (at least) one of the trees each internal node is connected to at most two leaves.

Proof for case (1). W.l.o.g. let T_1 be a tree that has only one internal node. Then, for any labelings of the leaves $P(T_1) = \emptyset$ and it is clear that for each internal edge e of T_2 the partition $\pi(T_2, e)$ is not in $P(T_1)$ and the theorem holds.

Proof for case (2). Let $u(x)$ be an internal node in T_1 (respectively T_2) that is connected to at least 3 leaves. From each tree remove one of the leaves connected to u , respectively x . This does not change the number of internal edges and the resulting trees T'_1 and T'_2 are both proper and have $n - 1$ leaves and m_1 (respectively m_2) internal edges. By the induction hypothesis there exist $\{1, 2, \dots, n - 1\}$ -labelings for T'_1 and T'_2 such that $d(T'_1, T'_2) = m_1 + m_2$. Extend these labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning the label n to both removed leaves. Clearly, then $d(T_1, T_2) = m_1 + m_2$ because the number of internal edges has not changed and the bipartitions of T_1 (T_2) are obtained from the bipartitions of T'_1 (respectively T'_2) by adding n to one of the sets of each bipartition. Hence, bipartitions of T'_1 and T'_2 that are different are also different after adding element n .

It remains to prove case (3). W.l.o.g. assume that in T_1 each internal node is connected to at most two leaves. Since T_1 is a proper tree there exists an internal node u of T_1 that is neighbour to exactly two leaves u_1 and u_2 . Since $n \geq 4$ node u has a neighbour v that is an internal node. Thus, $\{u, v\}$ is an internal edge. Consider the tree T'_1 that is obtained from T_1 by removing nodes u_1, u and edges $\{u_1, u\}, \{u, v\}$ and by connecting node u_2 to v . Then $T'_1 \in \mathcal{T}(n - 1, m_1 - 1)$.

First assume that T_2 has a node x that is connected to at least three leaves x_1, x_2 , and x_3 . Construct T'_2 by removing leaf x_1 . Then, $T'_2 \in \mathcal{T}(n-1, m_2)$ holds. By the induction hypothesis there exist $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 such that $d(T'_1, T'_2) = m_1 - 1 + m_2$. Extend these labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning leaves u_1 and x_1 the label n . Clearly, no bipartition of $P(T_2)$ is equal to a bipartition $\pi(T_1, e)$ where e is an internal edge with $e \neq \{u, v\}$. Also, $\pi(T_1, \{u, v\})$ is not in $P(T_2)$ because one set of $\pi(T_1, \{u, v\})$ equals $\{n, i\}$ for an $i \in [1, n-1]$ and all sets of bipartitions in $P(T_2)$ that include n have at least three elements. Thus, $d(T_1, T_2) = m_1 + m_2$ holds.

It remains to consider the case that all nodes in T_2 are connected to at most two leaves. Then there must exist an internal node x in T_2 that is connected to two leaves x_1 and x_2 . Clearly, x is connected to an internal edge $\{x, y\}$. Similar as for T_1 , create a phylogenetic tree T'_2 by removing nodes x_1, x and edges $\{x_1, x\}, \{x, y\}$ from T_2 and by connecting x_2 to y . By construction, T'_2 has $n-1$ nodes and m_2-1 internal edges. By the induction hypothesis there exist $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 such that $d(T'_1, T'_2) = m_1 - 1 + m_2 - 1$. Now consider four subcases.

Case a: Both nodes v and y have at least 2 neighbouring leaves in T'_1 respectively T'_2 . Then, we can assume w.l.o.g. that leaf u_2 has label i in T'_1 and leaf x_2 has label j in T'_2 with $j \neq i$ (Because, otherwise, the label of u_2 can be exchanged with the label of another leaf that is connected to v). Now, extend the labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning both leaves u_1 and x_1 label n . Clearly, partition $\pi(T_1, \{u, v\})$ is not in $P(T_2)$ and partition $\pi(T_2, \{x, y\})$ is not in $P(T_1)$. Hence, it is not hard to see that $d(T_1, T_2) = m_1 + m_2$ holds.

Case b: Node v has at least two neighbouring leaves (one leaf is u_2 and let v_1 be the other leaf) in T'_1 and node y has only the neighbouring leaf (i.e., leaf x_2) in T'_2 . Assume that j is the label of x_2 . Then at least one u_2 and v_1 has a label $i \neq j$. Assume first, that node u_2 has label i and extend the labelings of T'_1 and T'_2 to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then the partition $\pi(T_1, \{u, v\})$ is not in $P(T_2)$ and the partition $\pi(T_2, \{x, y\})$ is not in $P(T_1)$. It follows easily that $d(T_1, T_2) = m_1 + m_2$ holds. It remains to consider the case that node v_1 has label i . Then exchange the labels of u_2 and v_1 in T'_1 . Clearly, for this labeling $d(T'_1, T'_2) = m_1 - 1 + m_2 - 1$ holds. Now, proceed as before to show the result.

Case c: Node v has only one neighbouring leaf in T'_1 and node y has at least 2 neighbouring leaves in T'_2 . This case is symmetric to Case (b) and the proof is analogously.

Case d: Nodes v and y have only one neighbouring leaf in T'_1 respectively T'_2 . For the chosen $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 with $d(T'_1, T'_2) = m_1 - 1 + m_2 - 1$ let i and j be the labels of u_2 respectively x_2 . Note, that $i = j$ is possible. If $i \neq j$ extend the labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then, partition $\pi(T_1, \{u, v\})$ is not in $P(T_2)$ and partition $\pi(T_2, \{x, y\})$ is not in $P(T_1)$. Now, it is not hard to show that $d(T_1, T_2) = m_1 + m_2$ holds. It remains to consider the case $i = j$. Let $k \in \{1, 2, \dots, n-1\}$ be a label with $k \neq i$. First, extend the labelings of T'_1 and T'_2 to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then change the labeling for T_1 by exchanging the labels of the leaves with labels k and n . Now, partition $\pi(T_1, \{u, v\})$ has one set $\{k, i\}$ and is therefore not in $P(T_2)$. Similarly, partition $\pi(T_2, \{x, y\})$ has one set $\{n, i\}$ and is therefore not in $P(T_1)$. To show that no other partition of $P(T_1)$ can be in $P(T_2)$ and vice versa, assume the contrary, i.e., assume $P(T_1) \cap P(T_2) \neq \emptyset$. Let $\pi(T_1, e_1) = \pi(T_2, e_2)$ be a partition in $P(T_1) \cap P(T_2)$. By construction i and k must be in the same set of the partition because the corresponding leaves are connected to the same internal node in T_1 . Similarly, by construction i and n must be in the same set of the partition because the corresponding leaves are connected to the same internal node in T_2 . Hence $e_1 \neq \{u, v\}$ and $e_2 \neq \{x, y\}$. Altogether it follows that $\pi(T'_1, e_1) = \pi(T'_2, e_2)$ which contradicts the inductive hypothesis that $d(T'_1, T'_2) = m_1 - 1 + m_2 - 1$.

A special case of the theorem is when both trees T_1 and T_2 are binary trees, i.e., $T_1, T_2 \in \mathcal{T}(n, n-3)$. In this case there exist $\{1, 2, \dots, n\}$ -labelings of the leaves such that $d(T_1, T_2) = 2n - 6$ for $n \geq 3$.

For unrooted trees T_1 and T_2 that are not necessarily proper, i.e., where it is possible that internal nodes have degree 2, Theorem 1 implies the following corollary.

Corollary 1. *For each two unrooted trees $T_i \in \mathcal{T}(n, m_i)$, $n \geq 3$, $i \in \{1, 2\}$ where k_i is the number of internal nodes with degree 2 in T_i there exist $\{1, 2, \dots, n\}$ -labelings of the leaves of both trees such that for the corresponding phylogenetic trees $d(T_1, T_2) = m_1 - k_1 + m_2 - k_2$.*

Proof. To see that the corollary holds consider the case that in one of the trees T_i there exists a path of maximal length with internal nodes $n_1, n_2, \dots, n_j$ that all have degree 2. Then, there exist nodes n_0 and n_{k+1} not in the path such that n_0 is connected to n_1 and n_k is connected to n_{k+1} . For each two edges e, e' that are incident to (at least) one of the nodes in the path $\pi(T, e) = \pi(T, e')$ holds. Hence, if the path $n_1, n_2, \dots, n_j$ is removed from T_i and exchanged by a single edge (i.e., n_0 is connected to n_{k+1}) for the resulting tree T' the equality $P(T') = P(T)$ holds. Iteratively, apply this procedure until all $k_1 + k_2$ internal nodes with degree 2 have been removed in both trees and apply Theorem 1 to the resulting trees.

For rooted topological trees we show in the following a theorem that is analogous to Theorem 1 for unrooted trees and gives a bound on the cluster distance of two corresponding phylogenetic trees.

Theorem 2. *For each two proper rooted topological trees $T_1 \in \mathcal{T}_r(n, m_1)$ and $T_2 \in \mathcal{T}_r(n, m_2)$ there exist $\{1, 2, \dots, n\}$ -labelings of the leaves of both trees such that $d_r(T_1, T_2) = m_1 + m_2$ with $n \geq 3$, $0 \leq m_1 \leq n-2$, $0 \leq m_2 \leq n-2$.*

Proof. The proof is similar to the proof of Theorem 1. We show that there exist labelings for the leaves of T_1 and T_2 such that for each internal node v_1 of T_1 the cluster $T_1(v_1)$ is not contained in $Cl(T_2)$ and for each internal node v_2 of T_2 the cluster $T_2(v_2)$ is not contained in $P(T_1)$. The proof is done by induction on n .

Base case $n = 3$. In this case each tree has either no internal node (then all three leaves are connected to the root) or it has one internal node that is connected to two leaves and the other leaf is connected to the root. In the first case a tree has only trivial clusters and in the second case it has exactly one non-trivial cluster that contains the labels of the two leaves that are connected to the internal node. If the second case holds for both trees then the two leaves that are connected to the internal node can get labels 1 and 2 for T_1 respectively labels 1 and 3 for T_2 and the theorem holds. Otherwise, the theorem holds for all $\{1, 2, 3\}$ -labelings of T_1 and T_2 .

For the inductive step consider two trees T_1 and T_2 with $n \geq 4$ leaves and the following three cases:

1. One of the trees has no internal node.
2. Both trees have (at least) one internal node that is connected to at least 3 leaves.
3. Both trees have (at least) one internal node and for (at least) one of the trees each internal node is connected to at most two leaves.

Proof for case (1). W.l.o.g. let T_1 be a tree that has no internal node and therefore $Cl(T_1)$ contains only trivial clusters. Then, for each labeling of the leaves of T_2 and any internal node v of T_2 a non-trivial $cl(T_2, v)$ is not in $Cl(T_1)$ and the theorem holds.

Proof for case (2). Let $u(x)$ be an internal node in T_1 (respectively T_2) that is connected to at least 3 leaves. From each tree remove one of the leaves connected to u , respectively x . The resulting trees T'_1 and T'_2 are both proper and have $n-1$ leaves and m_1 (respectively m_2) internal nodes. By the induction hypothesis there exist $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 such that $d_r(T'_1, T'_2) = m_1 + m_2$. Extend these labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by

assigning the label n to both removed leaves. Since for each two internal nodes v of T'_1 and y of T'_2 the clusters $cl(T'_1, v)$ and $cl(T'_2, y)$ are different the clusters $cl(T_1, v)$ and $cl(T_2, y)$ are also different. Hence, $d_r(T_1, T_2) = m_1 + m_2$.

It remains to prove case (3). W.l.o.g. assume that in T_1 each internal node is connected to at most two leaves. Since T_1 is a proper tree there exists an internal node u of T_1 that is neighbour to exactly two leaves u_1 and u_2 and has exactly one other neighbour v (which is an internal node or the root). Consider the tree T'_1 that is obtained from T_1 by removing nodes u_1, u and edges $\{u_1, u\}, \{u, v\}$ and by connecting node u_2 to v . Then T_1 is proper and in $\mathcal{T}_r(n-1, m_1-1)$.

First assume that T_2 has a node x that is connected to at least three leaves x_1, x_2 , and x_3 . Construct T'_2 by removing leaf x_1 . Then, T'_2 is proper and in $\mathcal{T}_r(n-1, m_2)$. By the induction hypothesis there exist $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 such that $d_r(T'_1, T'_2) = m_1 - 1 + m_2$. Extend these labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning leaves u_1 and x_1 the label n . Clearly, no cluster of $Cl(T_2)$ is equal to a cluster $cl(T_1, w)$ where w is an internal node with $w \neq u$. Also, $cl(T_1, u) = \{n, i\}$ for an $i \in [1, n-1]$ is not in $Cl(T_2)$ because all cluster in $Cl(T_2)$ that include n have at least three elements. Thus, $d(T_1, T_2) = m_1 + m_2$ holds.

It remains to consider the case that all internal nodes in T_2 are connected to at most two leaves. Then there must exist an internal node x in T_2 that is connected to two leaves x_1 and x_2 and to exactly one other node y that is an internal node or the root. Similar as for T_1 , create a phylogenetic tree T'_2 by removing nodes x_1, x and edges $\{x_1, x\}, \{x, y\}$ from T_2 and by connecting x_2 to y . By construction, T'_2 is proper and in $\mathcal{T}_r(n-1, m_2-1)$. By the induction hypothesis there exist $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 such that $d_r(T'_1, T'_2) = m_1 - 1 + m_2 - 1$. Now consider four subcases.

Case a: Both nodes v and y have at least 2 neighbouring leaves in T'_1 respectively T'_2 . Then, we can assume w.l.o.g. that leaf u_2 has label i in T'_1 and leaf x_2 has label j in T'_2 with $j \neq i$ (Because, otherwise, the label of u_2 can be exchanged with the label of another leaf that is connected to v). Now, extend the labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning both leaves u_1 and x_1 label n . Clearly, cluster $cl(T_1, u)$ is not in $Cl(T_2)$ and cluster $cl(T_2, x)$ is not in $Cl(T_1)$. Hence, $d_r(T_1, T_2) = m_1 + m_2$ easily follows.

Case b: Node v has at least two neighbouring leaves (one leaf is u_2 and let v_1 be the other leaf) in T'_1 and node y has only one neighbouring leaf (i.e., leaf x_2) in T'_2 . Assume that j is the label of x_2 . Then at least one u_2 and v_1 has a label $i \neq j$. Assume first, that node u_2 has label i and extend the labelings of T'_1 and T'_2 to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then the cluster $cl(T_1, u)$ is not in $Cl(T_2)$ and the cluster $cl(T_2, x)$ is not in $Cl(T_1)$. It follows easily that $d_r(T_1, T_2) = m_1 + m_2$ holds. It remains to consider the case that node v_1 has label i . Then exchange the labels of u_2 and v_1 in T'_1 . Clearly, for this labeling $d_r(T'_1, T'_2) = m_1 - 1 + m_2 - 1$ holds. Now, proceed as before to show the result.

Case c: Node v has only one neighbouring leaf in T'_1 and node y has at least 2 neighbouring leaves in T'_2 . This case is symmetric to Case (b) and the proof is analogously.

Case d: Nodes v and y have only one neighbouring leaf in T'_1 respectively T'_2 . For the chosen $\{1, 2, \dots, n-1\}$ -labelings for T'_1 and T'_2 with $d_r(T'_1, T'_2) = m_1 - 1 + m_2 - 1$ let i and j be the labels of u_2 respectively x_2 . If $i \neq j$ extend the labelings to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then, cluster $cl(T_1, u)$ is not in $Cl(T_2)$ and cluster $cl(T_2, x)$ is not in $Cl(T_1)$. Now, it is not hard to show that $d_r(T_1, T_2) = m_1 + m_2$ holds.

It remains to consider the case $i = j$. Let $k \in \{1, 2, \dots, n-1\}$ be a label with $k \neq i$. First, extend the labelings of T'_1 and T'_2 to $\{1, 2, \dots, n\}$ -labelings for T_1 and T_2 by assigning label n to leaves u_1 and x_1 . Then change the labeling for T_1 by exchanging the labels of the leaves with labels k and n and for T'_1 by exchanging the label of the leaf with label k by n . Now, cluster $cl(T_1, u) = \{k, i\}$ is not in $Cl(T_2)$. Similarly, cluster $cl(T_2, x) = \{n, i\}$ is not in $Cl(T_1)$. By the construction it holds

for each cluster in $Cl(T_1)$ that it contains either i and k or none of them. Similarly, it holds for each cluster in $Cl(T_2)$ that it contains either i and n or none of them. Hence, for each internal node $w \neq u$ in T_1 it holds: either $i \notin cl(T_1, w)$ and $cl(T_1, w) = cl(T'_1, w)$ or $i, k \in cl(T_1, w)$ and $cl(T_1, w) - \{k\} = cl(T'_1, w)$. Similarly, for each internal node $z \neq x$ in T_2 it holds: either $i \notin cl(T_2, z)$ and $cl(T_2, z) = cl(T'_2, z)$ or $i, n \in cl(T_2, z)$ and $cl(T_2, z) - \{n\} = cl(T'_2, z)$.

Now, it remains to show that for each $cl \in Cl(T_1)$, $cl \neq \{k, i\}$ implies $cl \notin Cl(T_2)$ and, vice versa, for each $cl \in Cl(T_2)$, $cl \neq \{n, i\}$ implies $cl \notin Cl(T_1)$. To show the first statement, let $cl \in Cl(T_1)$, $cl \neq \{k, i\}$. There exist four cases:

- Case i) $i, n \notin cl$. Then $k \notin cl$ and $cl \in Cl(T'_1)$. Therefore, $cl \notin Cl(T'_2)$. By the construction it follows that $cl \notin Cl(T_2)$.
- Case ii) $i \notin cl, n \in cl$. Then $cl \notin Cl(T_2)$ because every cluster in $Cl(T_2)$ contains either i and n or none of them.
- Case iii) $i \in cl, n \notin cl$. Similar as in case (ii) it follows that $cl \notin Cl(T_2)$.
- Case iv) $i, n \in cl$. Then $k \in cl$ and $cl - \{k\} \in Cl(T'_1)$. Therefore, $cl - \{n\} \notin Cl(T'_2)$. By the construction it follows that $cl \notin Cl(T_2)$.

The second statement, i.e., $cl \in Cl(T_2)$, $cl \neq \{n, i\}$ implies $cl \notin Cl(T_1)$, can be shown by symmetric arguments. Thus, the theorem holds.

Similarly, as for unrooted trees the following corollary can be shown for rooted trees T_1 and T_2 that are not necessarily proper, i.e., where it is possible that internal nodes have degree 2. Theorem 2 implies the following corollary.

Corollary 2. For each two rooted trees $T_i \in \mathcal{T}_r(n, m_i)$, $n \geq 3$, $i \in \{1, 2\}$ where k_i is the number of internal nodes with degree 2 in T_i there exist $\{1, 2, \dots, n\}$ -labelings of the leaves of both trees such that for the corresponding phylogenetic trees $d_r(T_1, T_2) = m_1 - k_1 + m_2 - k_2$.

Proof. To see that the corollary holds consider the case that in one of the trees T_i exists a path $n_1, n_2, \dots, n_k$ of maximal length such that all nodes n_j of the path have degree 2. Let n_k be the node of the path that is farthest away from the root and let n_{k+1} be a node that is connected with n_k and is not on the path. Clearly such a node n_{k+1} must exist and n_{k+1} is a leaf or an internal node with degree ≥ 3 . Let n_0 be a node that is connected to n_1 and is not on the path. Clearly such a node n_0 must exist and n_0 is the root or an internal node with degree ≥ 3 . For each two node n_j , $1 \leq j \leq k$ holds $cl(T, n_j) = cl(T, n_{k+1})$. Hence, if the path $n_1, n_2, \dots, n_k$ is removed from T_i and exchanged by a single edge that connects n_0 and n_{k+1} for the resulting tree T'_i the equality $Cl(T_i) = Cl(T'_i)$ holds. Iteratively, apply this procedure until all $k_1 + k_2$ internal nodes with degree 2 have been removed from both trees and apply Theorem 1 to the resulting trees.

From Corollary 2 we obtain the following corollary on the worst case of the unrooted CD distance between an unrooted phylogenetic tree and a rooted phylogenetic tree both with n leaves.

Corollary 3. For each unrooted tree $T_1 \in \mathcal{T}(n, m_1)$ and rooted tree $T_2 \in \mathcal{T}_r(n, m_2)$, $n \geq 3$ where k_i is the number of internal nodes with degree 2 in T_i , $i \in \{1, 2\}$ there exist $\{1, 2, \dots, n\}$ -labelings of the leaves of both trees such that for the corresponding phylogenetic trees $d_{ur}(T_1, T_2) = m_1 - k_1 + m_2 - k_2$.

Proof. For an edge e of T_1 consider the rooted tree $T_{1,e}$. By definition of $T_{1,e}$ it holds that $T_{1,e} \in \mathcal{T}_r(n, m_1)$ and $T_{1,e}$ has k_1 internal node that have degree 2. To see this, recall that the root is not an inner node. Now, the corollary follows immediately from Corollary 2.

Note, that the proof of Corollary 3 has shown that for any edge e of T_1 it holds that there exists $\{1, 2, \dots, n\}$ -labelings of the leaves of trees $T_{1,e}$ and T_2 such that $d_r(T_{1,e}, T_2) = m_1 - k_1 + m_2 - k_2$.

Conclusion

It was shown that for two topological trees T_1 and T_2 with n leaves, m_i internal edges in tree T_i , and k_i nodes of degree 2 in T_i , $i \in 1, 2$ there exists assignments of labels $\{1, 2, \dots, n\}$ to the leaves of each tree such that the tree partition distance (TPD; also called Robinson-Foulds distance for unrooted trees) between the corresponding unrooted phylogenetic trees is $m_1 - k_1 + m_2 - k_2$. In addition, this number is an upper bound, i.e., there does not exist assignments of labels $\{1, 2, \dots, n\}$ to the leaves such that the TPD between both trees is larger than $m_1 - k_1 + m_2 - k_2$. Moreover, it was shown that analogous results hold for the cluster distance (CD; also called Robinson-Foulds distance for rooted trees) of two rooted trees and for the unrooted cluster distance (urCD) of an unrooted tree and a rooted tree. Our results can be used to compute a normalized version of the corresponding distance measures.

Author Contributions

Conceptualization: Martin Middendorf, Nicolas Wieseke.

Formal analysis: Martin Middendorf, Nicolas Wieseke.

Investigation: Martin Middendorf, Nicolas Wieseke.

Methodology: Martin Middendorf, Nicolas Wieseke.

Writing – original draft: Martin Middendorf, Nicolas Wieseke.

Writing – review & editing: Martin Middendorf, Nicolas Wieseke.

References

1. Bourque M. Arbres de Steiner et reseaux dont certains sommets sont a localisation variable [dissertation]. Montreal: Université de Montreal; 1978.
2. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math. Biosci.* 1981; 53:131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2)
3. Robinson DF, Foulds LR. Comparison of weighted labeled trees. In: Horadam AF, Wallis WD, editors. *Combinatorial Mathematics VI, Lecture Notes in Mathematics*, vol 748. Berlin, Heidelberg: Springer; 1979; 748:119–126.
4. Bryant D, Steel M. Computing the Distribution of a Tree Metric. *IEEE/ACM Trans Comput Biol Bioinform.* 2009; 6(3):420–426. <https://doi.org/10.1109/TCBB.2009.32> PMID: 19644170
5. Penny D, Hendy MD. The Use of Tree Comparison Metrics. *Syst. Zool.* 1985; 34(1):75–82. <https://doi.org/10.2307/2413347>
6. Hendy MD, Little CHC, Penny D. Comparing Trees with Pendant Vertices Labelled. *SIAM Journal on Applied Mathematics.* 1984; 44(5):1054–1065. <https://doi.org/10.1137/0144075>
7. Steel MA. Distribution of the Symmetric Difference Metric on Phylogenetic Trees. *SIAM J. Discrete Math.* 1988; 1(4):541–551. <https://doi.org/10.1137/0401050>
8. Geoghegan JL, Duchêne S, Holmes EC. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens.* 2017; 13(2): e1006215. <https://doi.org/10.1371/journal.ppat.1006215> PMID: 28178344
9. Kupczok A, Schmidt HA, von Haeseler A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms for Molecular Biology.* 2010; 5(37):17pp
10. Steel MA, Penny D. Distributions of Tree Comparison Metrics—Some New Results. *Syst. Biol.* 1993; 42(2):126–141. <https://doi.org/10.1093/sysbio/42.2.126>
11. Guillerme T, Cooper N. Effects of missing data on topological inference using a Total Evidence approach. *Molecular Phylogenetics and Evolution.* 2016; 94:146–158. <https://doi.org/10.1016/j.ympev.2015.08.023> PMID: 26335040
12. Bogdanowicz D, Giaro K, Wrobel B. TreeCmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics.* 2012; 8:475–487. <https://doi.org/10.4137/EBO.S9657>
13. Weyenberg G, Yoshida R. Phylogenetic Tree Distances. In: Kliman RM, editor. *The Encyclopedia of Evolutionary Biology*, Oxford: Academic Press; 2016; 3:285–290.

14. Ho SYW, Duchêne S, Duchêne D. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular Ecology Resources*. 2015; 15(4):688–996. <https://doi.org/10.1111/1755-0998.12320> PMID: 25155426
15. Górecki P, Eulenstein O. A Robinson-Foulds Measure to Compare Unrooted Trees with Rooted Trees. In: Bleris L, Mandoiu I, Schwartz R, Wang J, editors. *Proc. 8th International Symposium on Bioinformatics Research and Applications (ISBRA 2012)*. Berlin: Springer. 2012; LNCS 7292:115-126.