Comparing and Aggregating Partially Resolved Trees [⋆]

Mukul S. Bansal, Jianrong Dong, and David Fernández-Baca

Department of Computer Science, Iowa State University, Ames, IA, USA
{bansal, jrdong, fernande}@cs.iastate.edu

Abstract. We define, analyze, and give efficient algorithms for two kinds of distance measures for rooted and unrooted phylogenies. For rooted trees, our measures are based on the topologies the input trees induce on *triplets*; that is, on three-element subsets of the set of species. For unrooted trees, the measures are based on *quartets* (four-element subsets). Triplet and quartet-based distances provide a robust and fine-grained measure of the similarities between trees. The distinguishing feature of our distance measures relative to traditional quartet and triplet distances is their ability to deal cleanly with the presence of unresolved nodes, also called polytomies. For rooted trees, these are nodes with more than two children; for unrooted trees, they are nodes of degree greater than three.
Our first class of measures are parametric distances, where there is parameter that weighs the difference between an unresolved triplet/quartet topology and a resolved one. Our second class of measures are based on Hausdorff distance. Each tree is viewed as a set of all possible ways in which the tree could be refined to eliminate unresolved nodes. The distance between the original (unresolved) trees is then taken to be the Hausdorff distance between the associated sets of fully resolved trees, where the distance between trees in the sets is the triplet or quartet distance, as appropriate.

1 Introduction

Evolutionary trees, also known as phylogenetic trees or phylogenies, represent the evolutionary history of sets of species. Such trees have uniquely labeled leaves, corresponding to the species, and unlabeled internal nodes, representing hypothetical ancestors. The trees may be rooted, if the evolutionary origin is known, or unrooted, otherwise.

This paper addresses two related questions: (1) How does one measure how close two evolutionary trees are to each other? (2) How does one combine or *aggregate* the phylogenetic information from conflicting trees into a single *consensus tree*? Among the motivations for the first question is the growth of phylogenetic databases, such as TreeBase [19], with the attendant need for sophisticated querying mechanisms and for means to assess the quality of answers to queries. The second question arises from the fact that phylogenetic analyses — e.g., by parsimony — typically produce multiple evolutionary trees (often in the thousands) for the same set of species.

We address the above questions by defining appropriate *distance measures* between trees. While several such measures have been proposed before (see below), ours provide a feature that previous ones do not: The ability to deal elegantly with the presence

[⋆] This work was supported in part by National Science Foundation AToL grant EF-0334832.

of *unresolved* nodes, also called *polytomies*. For rooted trees these are nodes with more than two children; for unrooted trees, they are nodes of degree greater than three. Polytomies cannot simply be ignored, since they arise naturally in phylogenetic analysis. Furthermore, they must be treated with care: A node may be unresolved because it truly must be so or because there is not enough evidence to break it up into resolved nodes — that is, the polytomies are either "hard" or "soft" [17].

Our contributions. We define and analyze two kinds of distance measures for phylogenies. For rooted trees, our measures are based on the topologies the input trees induce on *triplets*; that is, on three-element subsets of the set of species. For unrooted trees, the measures are based on *quartets* (four-element subsets). Our approach is motivated by the observation that triplet and quartet topologies are the basic building blocks of rooted and unrooted trees, in the sense that they are the smallest topological units that completely identify a phylogenetic tree [21]. Triplet and quartet-based distances thus provide a robust and fine-grained measure of the differences and similarities between trees[1]. In contrast with traditional quartet and triplet distances, our two classes of distance measures deal cleanly with the presence of unresolved nodes.

The first kind of measures we propose are *parametric distances*: Given a triplet (quartet) X, we compare the topologies that each of the two input trees induces on X. If they are identical, the contribution of X to the distance is zero. If both topologies are fully resolved but different, then the contribution is one. Otherwise, the topology is resolved in one of the trees, but not the other. In this case, X contributes p to the distance, where p is a real number between 0 and 1. Parameter p allows one to make a smooth transition between hard and soft views of polytomy. At one extreme, if $p = 1$, an unresolved topology is viewed as different from a fully resolved one. At the other, when $p = 0$, unresolved topologies are viewed as identical to resolved ones. Intermediate values of p allow one to adjust for the degree of certainty one has about a polytomy.

The second kind of measures proposed here are based on viewing each tree as a set of all possible fully resolved trees that can be obtained from it by refining its unresolved nodes. The distance between two trees is defined as the Hausdorff distance between the corresponding sets, where the distance between trees in the sets is the triplet or quartet distance, as appropriate.

After defining our distance measures, we proceed to study their mathematical and algorithmic properties. We obtain exact and asymptotic bounds on expected values of parametric triplet distance and parametric quartet distance. We also study for which values of p, parametric triplet and quartet distances are metrics, *near-metrics* (in the sense of [15]), or non-metrics.

Aside from the mathematical elegance that metrics and near-metrics bring to tree comparison, there are also algorithmic benefits. We formulate phylogeny aggregation as a *median* problem, in which the objective is to find a consensus tree whose total distance to the given trees is minimized. We do not know whether finding the median tree relative to parametric (triplet or quartet) distance is NP-hard, but conjecture that it is. This is suggested by the NP-completeness of the maximum triplet compatibility problem (see [8]). However, by the results mentioned above and well-known facts about

[1] Biologically-inspired arguments in favor of triplet-based measures can be found in [11].

the median problem [26], there are simple constant-factor approximation algorithms for the aggregation of rooted and unrooted trees relative to parametric distance: Simply return the input tree with minimum distance to the remaining input trees. We show that there are values of p for which parametric distance is a metric, but the median tree may not be fully resolved even if all the input trees are. However, beyond a threshold, the median tree is guaranteed to be fully resolved if the input trees are fully resolved.

We suspect that computing Hausdorff triplet (quartet) distance between two trees is NP-hard. However, we show that one can partially circumvent the issue by proving that, under a certain density assumption, Hausdorff distance is within a constant factor of parametric distance — that is, the measures are *equivalent* in the sense of [15].

Finally, we present a $O(n^2)$-time algorithm to compute parametric triplet distance and a $O(n^2)$ 2-approximate algorithm for parametric quartet distance. To our knowledge, there was no previous algorithm for computing the parametric triplet distance between two rooted trees, other than by enumerating all $\Theta(n^3)$ triplets. Two algorithms exist that can be directly applied to compute the parametric quartet distance. One runs in time $O(n^2 \min\{d_1, d_2\})$, where, for $i \in \{1, 2\}$, d_i is the maximum degree of a node in T_i [10]; the other takes $O(d^9 n \log n)$ time, where d is the maximum degree of a node in T_1 and T_2 [24].[2] Our faster $O(n^2)$ algorithm offers a 2-approximate solution when an exact value of the parametric quartet distance is not required. Additionally, our algorithm gives the exact answer when $p = \frac{1}{2}$.

Related work. Several other measures for comparing trees have been proposed; we mention a few. A popular class of distances are those based on symmetric distance between sets of *clusters* (that is, on sets of species that descend from the same internal node in a rooted tree) or of *splits* (partitions of the set of species induced by the removal of an edge in an unrooted tree); the latter is the well-known Robinson-Foulds (RF) distance [20]. It is not hard to show that two rooted (unrooted) trees can share many triplet (quartet) topologies but not share a single cluster (split). Cluster- and split-based measures are also coarser than triplet and quartet distances.

One can also measure the distance between two trees by counting the number of *branch-swapping* operations needed to convert one of the trees into the other [2]. However, the associated measures can be hard to compute, and they fail to distinguish between operations that affect many species and those that affect only a few. An alternative to distance measures are *similarity* methods such as maximum agreement subtree (MAST) approach [16]. While there are efficient algorithms for computing the MAST, the measure is coarser than triplet-based distances.

There is an extensive literature on consensus methods for phylogenetic trees. A non-exhaustive list of methods based on splits or clusters includes strict consensus trees [18], majority-rule trees [3], and the Adams consensus [1]. For a thorough survey on the subject, see [9].

The fact that consensus methods tend to produce unresolved trees, with an attendant loss of information, has been observed before. An alternative approach is to cluster the

[2] Note that unresolved nodes seem to complicate distance computation: The quartet distance between a pair of *fully resolved* unrooted trees can be obtained in $O(n \log n)$ time [7].

input trees into groups using some distance measure, each of which is represented by a consensus tree, in such a way as to minimize some measure of information loss [25].

In addition to consensus methods, there are techniques that take as input sets of quartet trees or triplet trees and try to find large compatible subsets or subsets whose removal results in a compatible set [5, 22]. These problems are related to the *supertree problem*, which generalizes the consensus problem by allowing the leaves of the input trees to overlap only partially [6].

The consensus problem on trees exhibits parallels with the *rank aggregation problem* [14, 15]. Here we are given a collection of rankings (that is, permutations) of n objects, and the goal is to find a ranking of minimum total distance to the input rankings. A distance between rankings of particular interest is *Kendall's tau*, defined as the number of pairwise disagreements between the two rankings. Like triplet and quartet distances, Kendall's tau is based on elementary ordering relationships. Rank aggregation under Kendall's tau is NP-complete even for four lists [14].

A permutation is the analog of a fully resolved tree, since every pairwise relationship between elements is given. The analog to a partially-resolved tree is a *partial ranking*, in which the elements are grouped into an ordered list of *buckets*, such that elements in different buckets have known ordering relationships, but elements within a bucket are not ranked [15]. Our definitions of parametric distance and Hausdorff distance are inspired by Fagin et al.'s *Kendall tau with parameter* p and their Hausdorff version of Kendall's tau, respectively [15]. We note, however, that aggregating partial rankings seems computationally easier than the consensus problem on trees. For example, while the Hausdorff version of Kendall's tau is easily computable [15], it is unclear whether the Hausdorff triplet or quartet distances are polynomially-computable for trees.

Organization of the paper. Section 2 reviews basic notions in phylogenetics and distances. Our distance measures and the consensus problem are formally defined in Section 3. The basic properties of parametric distance are proved in Section 4. Section 5 studies the connection between Hausdorff and parametric distances. Section 6 gives efficient algorithms for computing parametric distance.

2 Preliminaries

Phylogenies. By and large, we follow standard terminology (i.e., similar to [21]). We write $[N]$ to denote the set $\{1, 2, \ldots, N\}$, where N is a positive integer.

Let T be a rooted or unrooted tree. We write $\mathcal{V}(T)$, $\mathcal{E}(T)$, and $\mathcal{L}(T)$ to denote, respectively, the node set, edge set, and leaf set of T. A *taxon* (plural *taxa*) is some basic unit of classification; e.g., a species. Let S be a set of taxa. A *phylogenetic tree* or *phylogeny* for S is a tree T such that $\mathcal{L}(T) = S$. Furthermore, if T is rooted, we require that every internal node have at least two children; if T is unrooted, every internal node is required to have degree at least three. We write $RP(n)$ and $P(n)$ to denote, respectively, the sets of all rooted and unrooted phylogenetic trees over $S = [n]$.

An internal node in a *rooted* phylogeny is *resolved* if it has exactly two children; otherwise it is *unresolved*. Similarly, an internal node in an *unrooted* phylogeny is *resolved* if it has degree three, and *unresolved* otherwise. Unresolved nodes in rooted and unrooted trees are also referred to as *polytomies* or *multifurcations*. A phylogeny (rooted or unrooted) is *fully resolved* if all its internal nodes are resolved.

A *contraction* of a phylogeny T is obtained by deleting an internal edge and identifying its endpoints. A phylogeny T_2 *refines* phylogeny T_1 if and only if T_1 can be obtained from T_2 through 0 or more contractions. T_2 is a *full refinement* of T_1 if T_2 is a fully resolved tree that refines T_1. $\mathcal{F}(T)$ denotes the set of all full refinements of T.

Let X be a subset of $\mathcal{L}(T)$ and let $T[X]$ denote the minimal subtree of T having X as its leaf set. The *restriction* of T to X, denoted $T|X$, is the phylogeny for X defined as follows. If T is unrooted, then $T|X$ is the tree obtained from $T[X]$ by suppressing all degree-two nodes. If T is rooted, $T|X$ is obtained from $T[X]$ by suppressing all degree-two nodes except for the root.

A *triplet* is a three-element subset of S; a *quartet* is a four-element subset of S. A triplet (quartet) X is said to be *resolved* in a phylogenetic tree T over S if $T|X$ is fully resolved; otherwise, X is *unresolved*.

Finally, we need some special notation for rooted trees T. We write *rt*(T) to denote the root node of T. Let v be a node in T. Then, *pa*(v) denotes the parent of v in T and *Ch*(v) is the set of children of v. Furthermore, $T(v)$ denotes the subtree of T rooted at v and $\overline{T(v)}$ denotes the tree obtained by deleting $T(v)$ from T, as well as the edge from v to its parent, if such an edge exists.

Distance measures, metrics, and near-metrics. A *distance measure* on a set D is a binary function d on D satisfying the following three conditions: (i) $d(x, y) \geq 0$ for all $x, y \in D$; (ii) $d(x, y) = d(y, x)$ for all $x, y \in D$; and (iii) $d(x, y) = 0$ if and only if $x = y$. Function d is a *metric* if, in addition to being a distance measure, it satisfies the triangle inequality; i.e., $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in D$. Distance measure d is a *near-metric* if there is a constant c, independent of the size of D, such that d satisfies the *relaxed polygonal inequality*: $d(x, z) \leq c(d(x, x_1) + d(x_1, x_2) + \cdots + d(x_{n-1}, z))$ for all $n > 1$ and $x, z, x_1, \ldots, x_{n-1} \in D$ [15]. Two distance measures d and d' with domain D are *equivalent* if there are constants $c_1, c_2 > 0$ such that $c_1 d'(x, y) \leq d(x, y) \leq c_2 d'(x, y)$ for every pair $x, y \in D$ [15].

3 Distance measures for phylogenies

Let T_1 and T_2 be any two rooted (respectively, unrooted) phylogenies over the same taxon set S. We partition the set of triplets (quartets) over S into the following five sets.[3]

1. $\mathcal{S}(T_1, T_2)$: triplets (quartets) X that are resolved in T_1 and T_2, and $T_1|X = T_2|X$.
2. $\mathcal{D}(T_1, T_2)$: triplets (quartets) X that are resolved in T_1 and T_2, but $T_1|X \neq T_2|X$.
3. $\mathcal{R}_1(T_1, T_2)$: triplets (quartets) X that are resolved in T_1, but not in T_2.
4. $\mathcal{R}_2(T_1, T_2)$: triplets (quartets) X that are resolved in T_2, but not in T_1.
5. $\mathcal{U}(T_1, T_2)$: triplets (quartets) X that are unresolved in both T_1 and T_2.

[3] Note that the sets $\mathcal{S}(T_1, T_2)$ and $\mathcal{U}(T_1, T_2)$ are not used in this section, but are needed in later ones.

Let p be a real number in the interval $[0, 1]$. The *parametric triplet (quartet) distance between T_1 and T_2* is defined as

$$d^{(p)}(T_1, T_2) = |\mathcal{D}(T_1, T_2)| + p\left(|\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)|\right). \tag{1}$$

Parameter p allows one to make a smooth transition from soft to hard views of polytomy: When $p = 0$, resolved triplets (quartets) are treated as equal to unresolved ones, while when $p = 1$, they are treated as being completely different. Intermediate values of p allow one to adjust for the amount of evidence required to resolve a polytomy.

Let d be a metric over fully resolved trees. Metric d can be extended to partially resolved trees via *Hausdorff distance*, as follows.

$$d_{\text{Haus}}(T_1, T_2) = \max\left\{\max_{t_1 \in \mathcal{F}(T_1)} \min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2), \max_{t_2 \in \mathcal{F}(T_2)} \min_{t_1 \in \mathcal{F}(T_1)} d(t_1, t_2)\right\} \tag{2}$$

When d is the triplet (quartet) distance, d_{Haus} is called the *Hausdorff triplet (quartet) distance*. Observe that, in Equation (2), $\max_{t_1 \in \mathcal{F}(T_1)} \min_{t_2 \in \mathcal{F}(T_2)} d(t_1, t_2)$ gives the maximum distance between a full refinement of T_1 and the set of full refinements of T_2. Similarly, $\max_{t_2 \in \mathcal{F}(T_2)} \min_{t_1 \in \mathcal{F}(T_1)} d(t_1, t_2)$ is the maximum distance between a full refinement of T_2 and the set of full refinements of T_1. Therefore, T_1 and T_2 are at Hausdorff distance r of each other if every full refinement of T_1 is within distance r of a full refinement of T_2 and vice-versa.

Aggregating phylogenies. Let k be a positive integer and S be a set of taxa. A *profile of length k* (or simply a *profile*) is a mapping $\mathcal{P}$ that assigns each $i \in [k]$ a phylogenetic tree $\mathcal{P}(i)$ over S. We refer to these trees as *input trees*. A *consensus rule* is a function that maps a profile $\mathcal{P}$ to some phylogenetic tree T over S called a *consensus tree*.

Let d be a distance measure whose domain is the set of phylogenies over S. We extend d to define a distance measure from profiles to phylogenies as $d(T, \mathcal{P}) = \sum_{i=1}^{k} d(T, \mathcal{P}(i))$. A consensus rule is a *median rule* for d if for every profile $\mathcal{P}$ it returns a phylogeny T^* of minimum distance to $\mathcal{P}$; such a T^* is called a *median*. The problem of finding a median for a profile with respect to a distance measure d is referred to as the *median problem* (relative d), or as the *aggregation problem*.

4 Properties of parametric distance

In what follows, unless mentioned explicitly, whenever we refer to parametric distance, we mean both its triplet and quartet varieties. We begin with a useful observation.

Proposition 1. *For every p, q such that $p, q \in (0, 1]$, $d^{(p)}$ and $d^{(q)}$ are equivalent.*

The proof of the next theorem is along the lines of an analogous result for aggregating partial rankings by Fagin et al. [15] and is omitted from this extended abstract.

Theorem 1. *(1) For $p = 0$, $d^{(p)}$ is not a distance measure. (2) For $0 < p < 1/2$, $d^{(p)}$ is a near-metric, but not a metric. (3) For $p \geq 1/2$, $d^{(p)}$ is a metric.*

Part (3) of Theorem 1 leads directly to approximation algorithms. Part (2) indicates that the measure degrades nicely, since constant factor approximation ratios are also achievable with near-metrics [15].

The next result establishes a threshold for p beyond which a collection of fully resolved trees give enough evidence to produce a fully resolved tree.

Theorem 2. *Let $\mathcal{P}$ be a profile of length k, such that for all $i \in [k]$, tree $\mathcal{P}(i)$ is fully resolved. Then, if $p \geq 2/3$, there exists median tree T for $\mathcal{P}$ relative to $d^{(p)}$ such that T is fully resolved.*

Proof (sketch). Suppose T is a median tree that contains an unresolved node v. The key idea is to show that there is a way to refine v into two nodes such that the number of input triplet (quartet) topologies with which the resulting tree disagrees is at most twice the number with which it agrees. The theorem follows by applying this refinement step repeatedly, until a fully resolved tree is obtained. □

We can, in fact, show that if $p > 2/3$ and the input trees are fully resolved, the median tree relative to $d^{(p)}$ *must* be fully resolved. On the other hand, it is easy to show that when $p \in [1/2, 2/3)$, there are profiles of fully resolved trees whose median tree is only partially resolved.

It is interesting to compare Theorem 2 with analogous results for partial rankings. Consider the variation of Kendall's tau for partial rankings in which a pair of items that is ordered in one ranking but is in the same bucket in the other contributes p to the distance, where $p \in [0, 1]$. This distance measure is a metric when $p \geq 1/2$ [15]. Furthermore, if $p \geq 1/2$ the median ranking relative to this distance is a full ranking if the input consists of full rankings [4]. In contrast, Proposition 1 and Theorem 2 show that, for $p \in [1/2, 2/3]$, parametric triplet or quartet distance are metrics, but the median tree is not guaranteed to be fully resolved even if the input trees are. This opens up a range of values for p wherein parametric triplet/quartet distance is a metric, but where one can adjust for the degree of evidence needed to resolve a node.

We now consider the expected value of parametric triplet and quartet distances.

Theorem 3. *Let $u(n)$ and $r(n)$ denote the probabilities that a given quartet is, respectively, unresolved or resolved in an unrooted phylogeny chosen uniformly at random from $P(n)$. Then,*

(i) $E(d^{(p)}(T_1, T_2)) = \binom{n}{4} \cdot \left(\frac{2}{3} \cdot r(n)^2 + 2 \cdot p \cdot r(n) \cdot u(n)\right)$, *if T_1 and T_2 are* unrooted *phylogenies chosen uniformly at random with replacement from $P(n)$, and*

(ii) $E(d^{(p)}(T_1, T_2)) = \binom{n}{3} \cdot \left(\frac{2}{3} \cdot r(n+1)^2 + 2 \cdot p \cdot r(n+1) \cdot u(n+1)\right)$, *if T_1 and T_2 are* rooted *phylogenies chosen uniformly at random with replacement from $RP(n)$.*

Part (i) of Theorem 3 follows directly from [13, 23]. Part (ii) follows from part (i) and the relationship between rooted and unrooted trees [21]. Since $u(n) \sim \sqrt{\frac{\pi(2 \ln 2 - 1)}{4n}}$ [23] and $r(n) = 1 - u(n)$, Theorem 3 implies that $E(d^{(p)}(T_1, T_2))$ is asymptotically $\frac{2}{3} \cdot \binom{n}{4}$ for unrooted trees and $\frac{2}{3} \cdot \binom{n}{3}$ for rooted trees.

5 Relationships among the metrics

We do not know whether the Hausdorff triplet or quartet distances are computable in polynomial time. Indeed, we suspect that, unlike its counterpart for partial rankings, this may not be possible. On the positive side, we show here that, in a broad range of cases, it is possible to obtain an approximation to the Hausdorff distance by exploiting its connection with parametric distance. As in the previous section, our results apply to both triplet and quartet distances.

Lemma 1. *For every two phylogenies T_1 and T_2 over S,* $|\mathcal{D}(T_1, T_2)| + \frac{2}{3} \cdot \max\{|\mathcal{R}_1(T_1, T_2)|, |\mathcal{R}_2(T_1, T_2)|\} \leq d_{\mathrm{Haus}}(T_1, T_2) \leq |\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)| + |\mathcal{U}(T_1, T_2)|$.

Proof (sketch). The proof of the lower bound on d_{Haus} is in two steps. We first show that T_1 can be refined so that it disagrees with T_2 in at least two thirds of the triplets (quartets) in $\mathcal{R}_2(T_1, T_2)$. Next, we show the existence of an analogous refinement of T_2. Note that the triplets (quartets) in $\mathcal{D}(T_1, T_2)$ are resolved differently in any refinements of T_1 and T_2. This gives lower bounds for both arguments in the outer $\max$ of the definition of $d_{\mathrm{Haus}}(T_1, T_2)$ (Equation 2) and yields the lemma.

The upper bound follows by assuming that T_1 and T_2 are refined so that the triplets (quartets) in $\mathcal{R}_1(T_1, T_2)$, $\mathcal{R}_2(T_1, T_2)$, and $\mathcal{U}(T_1, T_2)$ are resolved differently. □

It is instructive to compare Lemma 1 with the situation for partial rankings. In the Hausdorff version of Kendall's tau, each partial ranking is viewed as the set of all possible full rankings that can be obtained by refining it (that is, ordering elements within buckets). The distance is then the Hausdorff distance between the two sets, where the distance between two elements is Kendall's tau. Let L_1 and L_2 be two partial rankings. Re-using notation, let $\mathcal{D}(L_1, L_2)$ be the set of all pairs that are ordered differently in L_1 and L_2, $\mathcal{R}_1(L_1, L_2)$ be the set of pairs that are ordered in L_1 but in the same bucket in L_2, and $\mathcal{R}_2(L_1, L_2)$ be the set of pairs that are ordered in L_2 but in the same bucket in L_1. Then, $d_{\text{Haus}}(L_1, L_2) = |\mathcal{D}(L_1, L_2)| + \max\{|\mathcal{R}_1(L_1, L_2)|, |\mathcal{R}_2(L_1, L_2)|\}$ [12, 15]. This expression leads to an efficient way to compute $d_{\text{Haus}}(L_1, L_2)$ and establishes an equivalence with the parametric version of Kendall's tau defined in Section 4 [15]. It seems unlikely that a similar simple expression can be obtained for Hausdorff triplet or quartet distance. There are at least two reasons for this. Let L_1 and L_2 be partial rankings. Then, it is possible to resolve L_1 so that it disagrees with L_2 in any pair in $\mathcal{R}_2(L_1, L_2)$. Similarly, there is a way to resolve L_2 so that it disagrees with L_1 in any pair in $\mathcal{R}_1(L_1, L_2)$. An analog for trees cannot be established for this property; hence, the $\frac{2}{3}$ factor in the lower bound of Lemma 1. The second reason is due to the properties of the set $\mathcal{U}(L_1, L_2)$. It can be shown that is one can refine L_1 and L_2 in such a way that pairs of elements that are unresolved in both rankings are resolved the same way in the refinements. This is, in general, impossible for trees and leads to the presence of $|\mathcal{U}(T_1, T_2)|$ in the upper bound of Lemma 1.

While the above observations are an obstacle to establishing equivalence between d_{Haus} and $d^{(p)}$, we *can* show equivalence when the number of triplets (quartets) that are unresolved in both trees is suitably small. The result below follows from Lemma 1.

Theorem 4. *Let β be a positive real number. Suppose we restrict ourselves to pairs of trees (T_1, T_2) such that $|\mathcal{U}(T_1, T_2)| \leq \beta(|\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| + |\mathcal{R}_2(T_1, T_2)|)$. Then, Hausdorff distance and parametric distance are equivalent.*

6 Computing parametric distance

Let $R(T)$ and $U(T)$ denote the sets of all triplets (quartets) that are, respectively resolved and unresolved in T. We need the following fact, which holds for rooted and unrooted trees.

Proposition 2. *For any two phylogenies T_1, T_2 over the same set of taxa,*

$$\begin{aligned} d^{(p)}(T_1, T_2) = \; & |R(T_1)| - |\mathcal{S}(T_1, T_2)| + p \cdot (|U(T_1)| - |U(T_2)|) \\ & + (2p-1) \cdot |\mathcal{R}_1(T_1, T_2)|. \end{aligned} \tag{3}$$

Proof. It can be shown that $|\mathcal{R}_1(T_1, T_2)| + |\mathcal{U}(T_1, T_2)| = |U(T_2)|$, $|\mathcal{R}_2(T_1, T_2)| + |\mathcal{U}(T_1, T_2)| = |U(T_1)|$, and $|\mathcal{S}(T_1, T_2)| + |\mathcal{D}(T_1, T_2)| + |\mathcal{R}_1(T_1, T_2)| = |R(T_1)|$. These relationships, along with Equation (1), establish Equation (3). □

6.1 Computing the parametric triplet distance

Theorem 5. *The parametric triplet distance $d^{(p)}(T_1, T_2)$ for two rooted phylogenetic trees T_1 and T_2 over the same set of n taxa can be computed in $O(n^2)$ time.*

Proof (sketch). Our algorithm computes $d^{(p)}(T_1, T_2)$ via Equation (3). For this, it needs $|R(T_1)|$, $|U(T_1)|$, $|U(T_2)|$, $|\mathcal{S}(T_1, T_2)|$ and $|\mathcal{R}_1(T_1, T_2)|$. The first three values can easily be obtained in $O(n)$ time. Below we outline an algorithm that computes the remaining two values in $O(n^2)$ time. This gives a $O(n^2)$ parametric triplet distance algorithm.

Our algorithm relies on a preprocessing step that calculates and stores the following four quantities for every pair u, v such that u, v are internal nodes of T_1 and T_2, respectively: $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$, $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(\overline{T_2(v)})|$, $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(T_2(v))|$, and $|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|$. All these $O(n^2)$ values can be computed in $O(n^2)$ time by visiting the pairs according to interleaved postorder traversals of T_1 and T_2, in which the set intersection sizes for each pair of nodes are computed by using the set intersection sizes computed for their children. We omit the details.

We need two definitions. Let T be a rooted phylogenetic tree. Let $X = \{x, y, z\}$ be a triplet. Suppose X is resolved in T. We say that X is *induced* by edge $(\mathit{pa}(v), v)$ in T if x, y are in $\mathcal{L}(T(v))$, and z is in $\mathcal{L}(\overline{T(v)})$. Note that X may be induced by multiple edges in T. Now suppose X is unresolved in T. We say that X is *associated* with the least common ancestor (lca) v of X in T. Observe that node v is unique and that it must be unresolved.

To compute $|\mathcal{S}(T_1, T_2)|$ we enumerate all pairs of internal edges $(\mathit{pa}(u), u) \in \mathcal{E}(T_1)$ and $(\mathit{pa}(v), v) \in \mathcal{E}(T_2)$ according to an order obtained by interleaving postorder traversals of T_1 and T_2. For each pair, we compute the number of common triplet topologies induced by the pair in $O(1)$ time by using the values $|\mathcal{L}(T_1(u)) \cap \mathcal{L}(T_2(v))|$, and

$|\mathcal{L}(\overline{T_1(u)}) \cap \mathcal{L}(\overline{T_2(v)})|$ computed in the preprocessing step. Thus, each identically resolved triplet is counted at least once. Since a triplet may be induced by multiple edges, it is necessary to adjust for over counting. Indeed, among the triplets induced by the edges $(\mathit{pa}(u), u) \in T_1$ and $(\mathit{pa}(v), v) \in T_2$, the ones that have already been counted at an earlier step are exactly those that are either (i) induced by both edges $(\mathit{pa}(u), u)$ and (u, y) in T_1, for some $y \in \mathit{Ch}(u)$, and are induced by the edge $(\mathit{pa}(v), v)$ in T_2, or, (ii) induced by both edges $(\mathit{pa}(v), v)$ and (v, y) in T_2, for some $y \in \mathit{Ch}(v)$, and are induced by the edge $(\mathit{pa}(u), u)$ in T_1. Both the counting and the correction for over counting can be done in $O(|\mathit{Ch}(u)| + |\mathit{Ch}(v)|)$ per pair, for a total of $O(n^2)$ time

To compute the value of $|\mathcal{R}_1(T_1, T_2)|$ we enumerate all pairs formed by picking an edge $e = (\mathit{pa}(u), u) \in \mathcal{E}(T_1)$ and an internal unresolved node $v \in \mathcal{V}(T_2)$ according to interleaved postorder traversals of T_1 and T_2. At each step, we count the number of triplets that are induced by e in T_1 and associated with v in T_2. Such triplets must be resolved in T_1 but unresolved in T_2. Let us say that a triplet X is *relevant* if it is induced by edge $(\mathit{pa}(u), u)$ in T_1, and $T_2[X]$ is a subtree of $T_2(v)$. There are $m = \binom{|P|}{2} \cdot |Q|$ relevant triplets, where $P = \mathcal{L}(T_2(v)) \cap \mathcal{L}(T_1(u))$ and $Q = \mathcal{L}(T_2(v)) \cap \mathcal{L}(\overline{T_1(u)})$. Out of these, we are interested in counting the number of triplets X whose lca in T_2 is v, and X is unresolved in T_2. Any such triplet X falls into one of three categories: (i) the lca of X in T_2 is not v, (ii) the lca of X in T_2 is v, X is resolved in T_1 and T_2, and $T_1|X = T_2|X$, (iii) the lca of X in T_2 is v, X is resolved in T_1 and T_2, but $T_1|X \neq T_2|X$. The sizes of these sets can be obtained in $O(|\mathit{Ch}(u)| \cdot |\mathit{Ch}(v)|)$ time — details are omitted. The number thus computed is then subtracted from m to get the quantity we need. The total time over all pairs u, v is $O(n^2)$. As in the computation of $|\mathcal{S}(T_1, T_2)|$, we must correct for over counting. Indeed any triplet induced by edge $(\mathit{pa}(u), u)$ and edge (u, y) in T_1, for some $y \in \mathit{Ch}(u)$, has already been counted in an earlier step of the interleaved traversals of T_1 and T_2. It can be shown that one can adjust for this over counting while keeping within the required time bound. □

6.2 An approximation algorithm for parametric quartet distance

Theorem 6. *Let T_1 and T_2 be two unrooted phylogenetic trees on the same n leaves. Then, for $p = \frac{1}{2}$, $d^{(p)}(T_1, T_2)$ can be computed in $O(n^2)$ time. For $p \in (\frac{1}{2}, 1]$, a value x such that $d^{(p)}(T_1, T_2) \leq x \leq 2 \cdot d^{(p)}(T_1, T_2)$ can be computed in $O(n^2)$ time.*

Proof (sketch). Our algorithm first computes the values of $|\mathcal{S}(T_1, T_2)|$, $|R(T_1)|$, $|U(T_1)|$, and $|U(T_2)|$ — this can be done in $O(n^2)$ time [10]. If $p = \frac{1}{2}$, these values suffice to obtain $d^{(p)}(T_1, T_2)$ exactly, since the term involving $|\mathcal{R}_1(T_1, T_2)|$ in Equation (3) vanishes. For $p > \frac{1}{2}$, we also use Equation (3), but instead of $|\mathcal{R}_1(T_1, T_2)|$ we use a 2-approximation y to $|\mathcal{R}_1(T_1, T_2)|$; that is, y satisfies $|\mathcal{R}_1(T_1, T_2)| \leq y \leq 2|\mathcal{R}_1(T_1, T_2)|$. Below, we outline how to compute such a y in $O(n^2)$ time. As a result, we obtain a 2-approximation to $d^{(p)}(T_1, T_2)$ in $O(n^2)$ time.

Let (u, v) be an edge in tree T. We denote the subtree of $T - (u, v)$ that contains node u by $T(u \leftarrow v)$, and the other subtree by $T(v \leftarrow u)$. Quartet $\{a, b, c, d\}$ is *induced* by edge (u, v) if $\{a, b\} \in \mathcal{L}(T(u \leftarrow v))$ and $\{c, d\} \in \mathcal{L}(T(v \leftarrow u))$. Every resolved quartet is induced by at least one edge. Quartet $\{a, b, c, d\}$ is *associated* with node v in

T if the paths from v to a, v to b, v to c, and v to d are edge-disjoint. Note that each unresolved quartet is associated with exactly one node in T.

Our algorithm roots T_1 by adding a root node to an arbitrarily chosen edge in T_1. It then enumerates each edge $e = (\mathit{pa}(u), u) \in \mathcal{E}(T_1)$ according to a preorder traversal of T_1 and each internal node $v \in \mathcal{V}(T_2)$ of degree at least 4. For each pair, it counts the number of quartets that are induced by e in T_1 and associated with v in T_2. As in the rooted case (Theorem 5), we do this indirectly. We first obtain the number of *relevant* quartets; namely those induced by $(\mathit{pa}(u), u)$. This can be done efficiently with suitable preprocessing. To find the size of the subset of these quartets that are unresolved and associated with v (which is what we need), we count the number of all other quartets and subtract it from the number of relevant quartets. Each of these other quartets appears in one of the following five configurations in the tree T_2: (i) there exists a neighbor x of v in T_2, such that the quartet is completely contained in $T_2(x \leftarrow v)$, (ii) there exist two neighbors x, y of v in T_2, such that $T_2(x \leftarrow v)$ contains three leaves from the quartet and $T_2(y \leftarrow v)$ contains the other leaf, (iii) there exist two neighbors x, y of v in T_2, such that $T_2(x \leftarrow v)$ contains two leaves from the quartet and $T_2(y \leftarrow v)$ contains the other two leaves, and (iv) there exist three neighbors x, y, z of v in T_2, such that $T_2(x \leftarrow v)$ contains two leaves from the quartet, $T_2(y \leftarrow v)$ contains one leaf of the quartet, and $T_2(z \leftarrow v)$ contains the remaining leaf. Handling cases (i), (ii) and (iii) efficiently is relatively easy, but case (iv) requires computing first a combined value that counts each quartet from case (iii) twice and each quartet from (iv) once, and then deriving the value for case (iv). The time per pair $(\mathit{pa}(u), u) \in \mathcal{E}(T_1), v \in \mathcal{V}(T_2)$ is $O(|\mathit{Ch}(u)| \cdot |\mathit{adj}(v)|)$, for a total of $O(n^2)$ time.

Note that, as described, the above computation over counts some quartets. While, we can correct for this, the best we are able to guarantee while staying within a $O(n^2)$ time bound is that no quartet is counted more than twice. This is the source of the 2-approximation. □

References

1. E. N. Adams III. N-trees as nestings: Complexity, similarity, and consensus. *J. Classification*, 3(2):299–317, 1986.
2. B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–13, 2001.
3. J. P. Barthélemy and F. R. McMorris. The median procedure for n-trees. *Journal of Classification*, 3:329–334, 1986.
4. J. J. Bartholdi, C. A. Tovey, and M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.
5. V. Berry, T. Jiang, P. E. Kearney, M. Li, and H. T. Wareham. Quartet cleaning: Improved algorithms and simulations. In *Proceedings of the 7th Annual European Symposium on Algorithms*, volume 1643 of *LNCS*, pages 313–324. Springer, 1999.
6. O. R. P. Bininda-Emonds, editor. *Phylogenetic supertrees: Combining Information to Reveal the Tree of Life*, volume 4 of *Computational Biology Series*. Springer-Verlag, 2004.
7. G. S. Brodal, R. Fagerberg, and C. N. S. Pedersen. Computing the quartet distance in time $O(n \log n)$. *Algorithmica*, 38(2):377–395, 2003.
8. D. Bryant. *Building trees, hunting for trees, and comparing trees: Theory and methods in phylogenetic analysis*. PhD thesis, Department of Mathematics, University of Canterbury, New Zealand, 1997.

9. D. Bryant. A classification of consensus methods for phylogenetics. In M. Janowitz, F.-J. Lapointe, F. McMorris, B. B. Mirkin, and F. Roberts, editors, *Bioconsensus*, volume 61 of *Discrete Mathematics and Theoretical Computer Science*, pages 163–185. American Mathematical Society, Providence, RI, 2003.
10. C. Christiansen, T. Mailund, C. N. Pedersen, M. Randers, and M. S. Stissing. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1(16), 2006.
11. J. A. Cotton, C. S. Slater, and M. Wilkinson. Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic Biology*, 55(2):345–350, April 2006.
12. D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*, volume 34 of *Lecture Notes in Statist.* Springer-Verlag, Berlin, 1980.
13. W. H. E. Day. Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Systematic Zoology*, 35(3):325–333, Sep. 1986.
14. C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Tenth International World Wide Web Conference*, pages 613–622, Hong Kong, May 2001.
15. R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM J. Discrete Math.*, 20(3):628–648, 2006.
16. C. R. Finden and A. D. Gordon. Obtaining common pruned trees. *J. Classification*, 2(1):225–276, 1985.
17. W. P. Maddison. Reconstructing character evolution on polytomous cladograms. *Cladistics*, 5:365–377, 1989.
18. F. R. McMorris, D. B. Meronk, and D. A. Neumann. A view of some consensus methods for trees. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 122–125. Springer-Verlag, 1983.
19. W. Piel, M. Sanderson, M. Donoghue, and M. Walsh. Treebase. http://www.treebase.org. Last accessed 2 February 2007.
20. D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
21. C. Semple and M. Steel. *Phylogenetics*. Oxford Lecture Series in Mathematics. Oxford University Press, Oxford, 2003.
22. S. Snir and S. Rao. Using max cut to enhance rooted trees consistency. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 3(4):323–333, 2006.
23. M. Steel and D. Penny. Distributions of tree comparison metrics — some new results. *Systematic Biology*, 42(2):126–141, 1993.
24. M. Stissing, C. N. S. Pedersen, T. Mailund, G. S. Brodal, and R. Fagerberg. Computing the quartet distance between evolutionary trees of bounded degree. In D. Sankoff, L. Wang, and F. Chin, editors, *APBC*, volume 5 of *Advances in Bioinformatics and Computational Biology*, pages 101–110. Imperial College Press, 2007.
25. C. Stockham, L.-S. Wang, and T. Warnow. Statistically based postprocessing of phylogenetic analysis by clustering. In *ISMB*, pages 285–293, 2002.
26. V. V. Vazirani. *Approximation Algorithms*. Springer-Verlag, Berlin, 2001.