Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 389 (2007) 44–55
Theoretical Computer Science
www.elsevier.com/locate/tcs

On the hardness of inferring phylogenies from triplet-dissimilarities

Ilan Gronau, Shlomo Moran*

Department of Computer Science, Technion-Israel Institute of Technology, Technion-city, 32000 Haifa, Israel

Received 19 September 2006; received in revised form 26 July 2007; accepted 27 July 2007

Communicated by A. Apostolico

Abstract

This work considers the problem of reconstructing a phylogenetic tree from *triplet-dissimilarities*, which are dissimilarities defined over taxon-triplets. Triplet-dissimilarities are possibly the simplest generalization of pairwise dissimilarities, and were used for phylogenetic reconstructions in the past few years. We study the hardness of finding a tree best fitting a given triplet-dissimilarity table under the ℓ_∞ norm. We show that the corresponding decision problem is NP-hard and that the corresponding optimization problem cannot be approximated in polynomial time within a constant multiplicative factor smaller than 1.4. On the positive side, we present a polynomial time constant-rate approximation algorithm for this problem. We also address the issue of best-fit under *maximal distortion*, which corresponds to the largest *ratio* between matching entries in two triplet-dissimilarity tables. We show that it is NP-hard to approximate the corresponding optimization problem within any constant multiplicative factor.

Keywords: NP hardness; Hardness of approximation; Phylogenetic trees; Distance based reconstruction algorithms

1. Introduction

Phylogenetic reconstruction methods attempt to find the evolutionary history of a given set of extant species (taxa). This history is usually described by an edge-weighted tree whose internal vertices represent past speciation events (extinct species) and whose leaves correspond to the given set of taxa. The amount of evolutionary change between two subsequent speciation events is indicated by the weight of the edge connecting them. It is usually assumed (for uniqueness of representation) that internal edges[1] have strictly positive weights. Distance-based phylogenetic reconstruction methods typically try to reconstruct this evolutionary tree from estimates of distances (sum of weights) along edges in this tree.

Most common distance-based reconstruction algorithms receive as input a *dissimilarity matrix* D, where $D(i, j)$ is an estimate of the distance between taxa i and j. A dissimilarity matrix is said to be *additive* if it can be realized by distances along the edges of a tree whose leaves are the elements of S [3]. There are numerous algorithms which reconstruct a tree given its additive metric, the earliest of which appeared in [3,13,14]. However, in reality we are

* Corresponding author.
E-mail address: moran@cs.technion.ac.il (S. Moran).

[1] An edge is *external* if it is adjacent to a leaf, and is *internal* otherwise.

doi:10.1016/j.tcs.2007.07.048

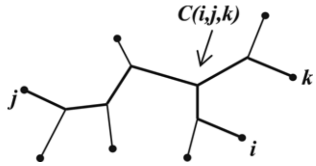

Fig. 1. $C(i, j, k)$ is the inner vertex of degree 3 in the claw spanned by i, j, k. $D_T(i; jk)$ is the weight of the path connecting i and $C(i, j, k)$.

unable to obtain accurate distance estimates, and the input dissimilarity matrix is rarely additive. In such a case, a natural goal is to reconstruct a tree fitting the input matrix in some way. One approach is to return a tree whose implied metric is 'close' to the input under a certain distance norm. Unfortunately, finding a tree closest to a given dissimilarity matrix was shown to be NP-hard under the ℓ_1 and ℓ_2 norms in [4], and under the ℓ_∞ norm in [1]. [1] also presents a 3-approximation algorithm for the problem of finding the tree closest, under ℓ_∞, to an arbitrary metric; another 3-approximation algorithm for this problem was presented later in [8].[2]

In this paper we study the problem of reconstructing a phylogenetic tree based on estimates of *triplet-distances*. Given an edge-weighted tree T and three taxa i, j, k, we denote by $C(i, j, k)$ the inner vertex of degree 3 in the claw spanned by i, j, k (see Fig. 1), and by $D_T(i; jk)$ the weight of the path connecting i and $C(i, j, k)$. Note that for all $k \in S$, $D_T(i; jk) + D_T(j; ik) = D_T(i, j)$, and in particular $D_T(i; jj) = D_T(i, j)$. Hence, triplet-distances generalize the classical notion of pairwise-distances.

A *triplet-dissimilarity table* contains estimates of all triplet-distances over a given taxon-set. A function $\tau : S \times S \times S \to \mathbb{R}^+$ is a valid triplet-dissimilarity table iff it satisfies the following properties:

1. $\tau(i, i, j) = 0$
2. $\tau(i, j, k) = \tau(i, k, j)$
3. $\tau(i, j, j) = \tau(j, i, i)$.

For such a function we denote: $\tau(i; jk) \stackrel{\triangle}{=} \tau(i, j, k)$ and $\tau(i, j) \stackrel{\triangle}{=} \tau(i, j, j)$.

There are several previous works which propose algorithms for reconstructing trees from triplet-dissimilarity tables. In [11], triplet-dissimilarities are used to obtain more accurate estimates of pairwise-distances for Saitou and Nei's Neighbor Joining algorithm (commonly referred to as NJ) [12]. [10] generalizes NJ to receive as input *m-dissimilarity maps*, which contain the *total weights* of all subtrees spanned by subsets of m taxa. In [8] we present a family of algorithms (DLCA) which construct trees from estimates of triplet-distances from a single root-taxon r, meaning that the input is a symmetric matrix L_r, where $L_r(i, j)$ is an estimate of $D_T(r; ij)$. We show there that a tree whose triplet-distances $\{D_T(r; ij) : i, j \in S\}$ are closest to L_r under ℓ_∞ can be constructed in $O(n^2)$ time. In this paper we show that it is NP-hard to find an edge-weighted tree T whose *entire* triplet-distance table $\{D_T(i; jk) : i, j, k \in S\}$ is closest to a given triplet-dissimilarity table under ℓ_∞.

The ℓ_∞ norm measures the maximal *difference* between corresponding entries in two triplet-dissimilarity tables:

$$||\tau_1, \tau_2||_\infty \stackrel{\triangle}{=} \max_{i,j,k \in S} \{|\tau_1(i; jk) - \tau_2(i; jk)|\}.$$

Another distance measure we refer to is *maximal distortion* [2], which is related to the maximal *ratio* between such entries:

$$MaxDist(\tau_1, \tau_2) \stackrel{\triangle}{=} \max_{i,j,k \in S} \left\{ \frac{\tau_1(i; jk)}{\tau_2(i; jk)} \right\} \cdot \max_{i,j,k \in S} \left\{ \frac{\tau_2(i; jk)}{\tau_1(i; jk)} \right\} \quad (\text{where } 0/0 \stackrel{\triangle}{=} 1).$$

We note that maximal distortion seems to be the most relevant criterion for the evolutionary models assumed in [5,6] and numerous subsequent works.

Consider the decision version of the 'best-fit to triplet-dissimilarities' problem: given a triplet-dissimilarity table τ and a non-negative number K, is there a tree T such that $||D_T, \tau||_\infty \le K$? In Section 2 this decision problem is shown to be NP-hard by a polynomial reduction from 3SAT. In Section 3 we refine the analysis of the reduction to show that it is NP-hard to find a tree whose distance to the input under ℓ_∞ is less than 1.4 times that of the closest

[2] The 3-approximation ratio of the algorithms in [1,8] is proved under the assumption that the input dissimilarity matrix is a *metric*, meaning that it satisfies the triangle inequality $[\mathcal{D}(x, y) + \mathcal{D}(y, z) \ge \mathcal{D}(x, z)]$.

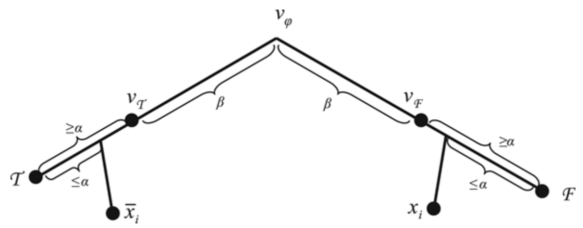

Fig. 2. The topology of a tree satisfying **A1–2**.

tree. In Section 4 we present few other related hardness results implied by our reduction, including the NP-hardness of approximating maximal distortion for triplet-dissimilarities by any multiplicative constant. In Section 5 we give an upper bound on the approximation ratio of this problem by showing that a constant-rate approximation for the closest tree to a dissimilarity matrix implies also a constant-rate approximation for the closest tree to a triplet-dissimilarity table. We conclude with a short discussion of some relevant open questions.

2. A reduction from 3SAT to the *'best-fit to triplets under ℓ_∞'* problem

In this section we present a reduction from 3SAT to the decision version of the 'best-fit to triplets under ℓ_∞' problem. This reduction transforms a 3CNF formula φ into a valid triplet-dissimilarity table τ_φ satisfying three requirements (where Δ is a positive constant independent of φ):

POLY τ_φ can be computed in polynomial time given φ.
SAT If φ is satisfiable, then there is a tree T s.t. $||D_T, \tau_\varphi||_\infty \leq \Delta$.
UNSAT If φ is unsatisfiable, then for every tree T, $||D_T, \tau_\varphi||_\infty > \Delta$.

Similar to the reduction presented in [1] for the problem of fitting trees to dissimilarity matrices, we first transform the formula φ into a set of upper and lower bounds on *some* triplet-distances of a tree T (see **A1–B3** and Fig. 2). **UNSAT** is proven by showing that a tree satisfying all these bounds implies a satisfying assignment to φ (Lemma 2.5). These bounds are enforced by the triplet-dissimilarity table τ_φ in the following way: A bound $D_T(i; jk) \leq \omega_{ijk}$ is enforced by $\tau_\varphi(i; jk) = \omega_{ijk} - \Delta$, and a bound $D_T(i; jk) \geq \omega_{ijk}$ is enforced by $\tau_\varphi(i; jk) = \omega_{ijk} + \Delta$. Clearly, a tree T satisfying $||D_T, \tau_\varphi||_\infty \leq \Delta$ is guaranteed to obey all bounds.[3]

Requirement **POLY** will be obvious from the description of the transformation. To prove **SAT** we show that a satisfying assignment to φ implies a tree satisfying all bounds (Lemma 2.6). In this tree, triplet-distances corresponding to entries restricted by these bounds are set to satisfy the bounds *with equality*. Other triplet-dissimilarities (not restricted by any bound) are undetermined, however each such dissimilarity falls within one of the two intervals: $[r - \Delta, r + \Delta]$ or $[s - \Delta, s + \Delta]$. Entries of τ_φ corresponding to these dissimilarities are set to the mid-point of the appropriate interval (r or s).

Let us start with some notations: A 3CNF formula φ over a set of variables $\{x_1, x_2, \ldots, x_n\}$ is a conjunction of m clauses $\varphi = c_1 \wedge c_2 \wedge \cdots \wedge c_m$, s.t. $\forall j = 1..m : c_j = (l_1^j \vee l_2^j \vee l_3^j)$, where l_1^j, l_2^j, l_3^j are literals (a literal is variable x_i or its negation $\bar{x}_i$). For such a formula, we define a set of taxa:

$$S_\varphi = \{\mathcal{T}, \mathcal{F}\} \cup \{x_i, \bar{x}_i : i = 1..n\} \cup \{y_1^j, y_2^j, y_3^j : j = 1..m\}.$$

We define the following set of bounds on triplet-dissimilarities over S_φ with parameters $\alpha, \beta > 0$ (Fig. 2 can be helpful at this point):

A1 $D_T(\mathcal{T}, \mathcal{F}) \geq 2\alpha + 2\beta$
A2 $\forall i = 1..n : D_T(\mathcal{F}; x_i\bar{x}_i) \leq \alpha$; $D_T(\mathcal{T}; x_i\bar{x}_i) \leq \alpha$
B1 $\forall j = 1..m : D_T(y_1^j; l_2^j l_3^j) \leq \alpha$; $D_T(y_2^j; l_1^j l_3^j) \leq \alpha$; $D_T(y_3^j; l_1^j l_2^j) \leq \alpha$
B2 $\forall j = 1..m : D_T(y_1^j; \mathcal{T}\mathcal{F}) \geq \alpha$; $D_T(y_2^j; \mathcal{T}\mathcal{F}) \geq \alpha$; $D_T(y_3^j; \mathcal{T}\mathcal{F}) \geq \alpha$
B3 $\forall j = 1..m : D_T(\mathcal{T}; y_1^j y_2^j) \leq \alpha$; $D_T(\mathcal{T}; y_1^j y_3^j) \leq \alpha$; $D_T(\mathcal{T}; y_2^j y_3^j) \leq \alpha$.

[3] Note that in order to keep all entries of τ_φ non-negative, we need that $\Delta \leq \omega_{ijk}$ whenever ω_{ijk} is an upper bound on the corresponding entry.

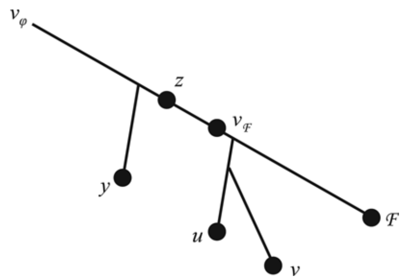

Fig. 3. Proof of Lemma 2.1(2).

Let T be a tree satisfying **A1–B3** above. Denote the mid-point of the path connecting $\mathcal{T}$ and $\mathcal{F}$ in this tree by v_φ. Note that restriction **A1** implies that $\mathcal{T}$ and $\mathcal{F}$ are at distance of at least $\alpha + \beta$ from v_φ. Denote by $v_\mathcal{T}$ and $v_\mathcal{F}$ the points whose distance from v_φ, on the paths leading to $\mathcal{T}$ and $\mathcal{F}$ respectively, is *exactly* β. For the sake of the analysis below, we treat the three points $v_\varphi, v_\mathcal{T}, v_\mathcal{F}$ as vertices in the tree (possibly of degree 2), and assume that T is rooted at v_φ (see Fig. 2).

We now describe and prove the topological restrictions implied by these bounds. Our proof is based on two simple connections between distances and topological properties of quartets (subtrees spanned by four taxa), which we bring next. For vertices x, y in T, denote by $path(x, y)$ the path in T connecting x and y.

Lemma 2.1. *For all taxa u, v, y in T, we have*

1. *If $D_T(\mathcal{F}; uv) \le \alpha$ and $D_T(\mathcal{T}; uv) \le \alpha$, then either u is a descendant of $v_\mathcal{T}$ and v is a descendant of $v_\mathcal{F}$ or vice versa.*
2. *If both u and v are descendants of $v_\mathcal{F}$, and $D_T(y; uv) \le D_T(y; \mathcal{T}\mathcal{F})$ then y is also a descendant of $v_\mathcal{F}$.*

Proof. 1. Since $D_T(\mathcal{F}; uv) + D_T(T; uv) \le 2\alpha < 2\alpha + 2\beta \le D_T(\mathcal{F}, \mathcal{T})$ (by the assumption and bound **A1**), we must have that $C(u, \mathcal{T}, \mathcal{F})$ and $C(v, \mathcal{T}, \mathcal{F})$ are distinct vertices on $path(\mathcal{F}, \mathcal{T})$. In addition, the assumption also implies that one of them is at a distance at most α from $\mathcal{T}$ and the other is at a distance at most α from $\mathcal{F}$, which proves the claim.

2. Let z be the father of $v_\mathcal{F}$ (possibly $z = v_\varphi$). Notice that the edge $(z, v_\mathcal{F})$ is in $path(\mathcal{T}, \mathcal{F})$ (see Fig. 3). Since both u and v are descendants of $v_\mathcal{F}$, we have that if y is not a descendant of $v_\mathcal{F}$, then the path from y to $path(u, v)$ must contain the edge $(z, v_\mathcal{F})$, and hence $D_T(y; uv) \ge D_T(y; \mathcal{T}\mathcal{F}) + w(z, v_\mathcal{F}) > D_T(y; \mathcal{T}\mathcal{F})$, a contradiction. □

As a direct consequence of **A2** and Lemma 2.1(1) above, we have the following:

Corollary 2.2. *For each $i = 1..n$, one of the vertices $x_i, \bar{x}_i$ is a descendant of $v_\mathcal{T}$ and the other is a descendant of $v_\mathcal{F}$ (see Fig. 2).*

The above corollary leads to a natural transformation between trees satisfying **A1–2** and truth-assignments to the variables $x_1..x_n$: $\sigma_T(x_i) = TRUE$ if x_i is a descendant of $v_\mathcal{T}$ and $\sigma_T(x_i) = FALSE$ otherwise. The consistency of this assignment is guaranteed by Corollary 2.2. Lemma 2.1 (2), and **B1-2** lead to the following corollaries:

Corollary 2.3. *Let $j \in \{1, \ldots, m\}$, and let $\{a, b, c\} = \{1, 2, 3\}$. If l_a^j and l_b^j are descendants of $v_\mathcal{F}$, then y_c^j is also a descendant of $v_\mathcal{F}$.*

Corollary 2.4. *If for some $j = 1..m$, l_1^j, l_2^j, l_3^j are* all *descendants of $v_\mathcal{F}$, then the bounds in* **B3** *cannot hold.*

Proof. By Corollary 2.3, if l_1^j, l_2^j, l_3^j are *all* descendants of $v_\mathcal{F}$, then so are y_1^j, y_2^j, y_3^j. This implies, for instance, that $C(\mathcal{T}, y_1^j, y_2^j)$ is a descendant of $v_\mathcal{F}$ as well, so $D_T(\mathcal{T}; y_1^j y_2^j) \ge D_T(\mathcal{T}, v_\mathcal{F}) \ge 2\beta + \alpha > \alpha$, contradicting **B3** . □

Note the slackness (of 2β) we have in the contradiction concluding the proof. This slackness is used to prove hardness of approximation in Section 3. The following lemma concludes the discussion of unsatisfiable formulae:

Lemma 2.5. *If T is an edge-weighted tree over the set of taxa S_φ satisfying all bounds in* **A1–B3***, then the assignment σ_T satisfies the formula φ.*

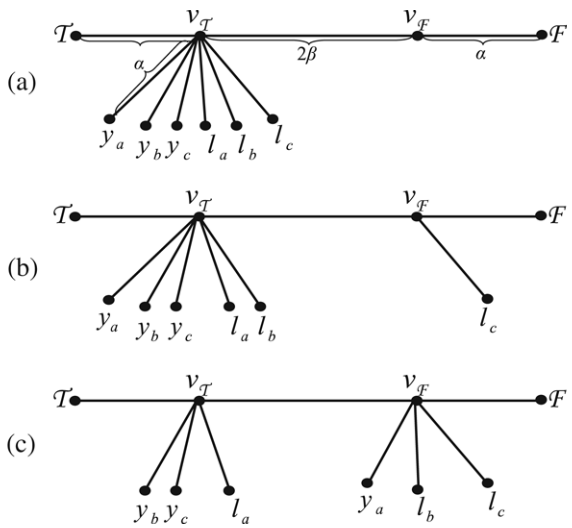

Fig. 4. Construction of a tree given a satisfying assignment. The figure illustrates how to connect the y-taxa for each type of satisfied clause: (a) All literals are satisfied (assigned *TRUE*). (b) Two literals (l_a, l_b) are satisfied. (c) One literal (l_a) is satisfied.

Proof. Assume, to the contrary, that σ_T does not satisfy some clause c_j of φ. Then, by definition of σ_T, the taxa l_1^j, l_2^j, l_3^j are all descendants of $v_{\mathcal{F}}$, and so by Corollary 2.4 the bounds in **B3** cannot hold for T. □

Lemma 2.5 is later used to ensure requirement **UNSAT**. To show that **SAT** holds we first prove the following lemma:

Lemma 2.6. *If the formula φ is satisfiable, then there exists a tree T over the set of taxa S_φ, satisfying* **A1–B3** *with equality.*

Proof. Let σ be a satisfying assignment of φ. We will construct a tree T with only two internal vertices $v_{\mathcal{T}}$, $v_{\mathcal{F}}$, and one internal edge of weight 2β connecting $v_{\mathcal{T}}$ and $v_{\mathcal{F}}$. All external edges are of weight α, and all taxa are either connected to $v_{\mathcal{T}}$ or $v_{\mathcal{F}}$. $\mathcal{T}$ is connected to $v_{\mathcal{T}}$, $\mathcal{F}$ is connected to $v_{\mathcal{F}}$, and a literal-taxon l_a^j is connected to $v_{\mathcal{T}}$ if $\sigma(l_a^j) = TRUE$, and to $v_{\mathcal{F}}$ otherwise. It is easy to see that the bounds in **A1–2** are satisfied by such a tree.

Taxa of the form y_a^j are connected according to the following scheme (see Fig. 4): if l_a^j is **the only** literal in clause c_j satisfied by σ, connect y_a^j to $v_{\mathcal{F}}$; otherwise connect it to $v_{\mathcal{T}}$. **B2** is clearly satisfied by this construction. **B3** is satisfied, since at most one y-taxon is connected to $v_{\mathcal{F}}$ for each clause. **B1** is satisfied, since for $\{a, b, c\} = \{1, 2, 3\}$ the following holds: If l_a^j, l_b^j are connected to $v_{\mathcal{F}}$ ($v_{\mathcal{T}}$) then y_c^j is connected to $v_{\mathcal{F}}$ ($v_{\mathcal{T}}$ respectively) as well. □

We now describe the reduction of the formula φ to a triplet-dissimilarity table τ_φ: Entries of τ_φ corresponding to distances bounded in **A1–B3** are set to enforce the corresponding bounds, as discussed in the beginning of this section. The rest of the entries in τ_φ, and the constant Δ, are set so that $|D_T(i; jk) - \tau_\varphi(i; jk)| \le \Delta$ will hold for *all* taxon-triplets in the tree T described in the proof of Lemma 2.6, as follows. First, for all distinct $i, j, k \in S_\varphi$ we have $D_T(i; jk) \in [\alpha, \alpha + 2\beta]$ (since all external edges are of length α, and the single internal edge is of length 2β). So we set the corresponding entries of τ_φ (which do not appear in **A1–B3**) to $\alpha + \beta$, and we set $\Delta = \beta$. This guarantees that $|D_T(i; jk) - \tau_\varphi(i; jk)| \le \Delta$ for the corresponding entries. Similarly, for all distinct $i, j \in S_\varphi$ we have $D_T(i; jj) = D_T(i, j) \in [2\alpha, 2\alpha + 2\beta]$, so we set corresponding entries of τ_φ to $2\alpha + \beta$. Thus, the entries of the triplet-dissimilarity table τ_φ are defined according to the following rules:

- $\tau_\varphi(\mathcal{T}; \mathcal{F}\mathcal{F}) = \tau_\varphi(\mathcal{F}; \mathcal{T}\mathcal{T}) = 2\alpha + 3\beta$ (**A1**)
- $\forall i = 1..n : \tau_\varphi(\mathcal{F}; x_i\bar{x}_i) = \tau_\varphi(\mathcal{T}; x_i\bar{x}_i) = \alpha - \beta$ (**A2**)
- $\forall j = 1..m : \tau_\varphi(y_1^j; l_2^j l_3^j) = \tau_\varphi(y_2^j; l_1^j l_3^j) = \tau_\varphi(y_3^j; l_1^j l_2^j) = \alpha - \beta$ (**B1**)
- $\forall j = 1..m : \tau_\varphi(\mathcal{T}; y_1^j y_2^j) = \tau_\varphi(\mathcal{T}; y_1^j y_3^j) = \tau_\varphi(\mathcal{T}; y_2^j y_3^j) = \alpha - \beta$ (**B3**)

- $\forall \{s, t\} (\neq \{\mathcal{T}, \mathcal{F}\}) \subseteq S_\varphi : \tau_\varphi(s; tt) = \tau_\varphi(t; ss) = 2\alpha + \beta$ (arbitrary pairwise-distances)
- For all other entries : $\tau_\varphi(s; tu) = \alpha + \beta$ (arbitrary triplet-distances and **B2**).

We conclude with the following lemma:

Lemma 2.7. *Let φ be a satisfiable formula, and let τ_φ be the triplet-dissimilarity table as defined above, using any values for α, β s.t. $\alpha \geq \beta > 0$. Then there exists a tree T over the set of taxa S_φ, such that $||D_T, \tau_\varphi||_\infty \leq \beta$.*

Proof. The tree T corresponding to an assignment σ satisfying φ (as described in the proof of Lemma 2.6) fulfills this requirement. The proof follows directly from the above discussion. □

Theorem 2.8. *The decision version of the problem of finding a tree best fitting a given triplet-dissimilarity table under the ℓ_∞ norm is NP-Hard.*

Proof. By the polynomial time reduction from 3SAT described above. The reduction $\varphi \mapsto (\tau_\varphi, \Delta)$ is clearly polynomial (requirement **POLY**). By Lemma 2.7, if φ is satisfiable then there exists a tree T, s.t. $||D_T, \tau_\varphi||_\infty \leq \Delta$ (**SAT**). If, on the other hand, φ is unsatisfiable, then by Lemma 2.5 there is no tree satisfying **A1–B3**. Due to the construction of τ_φ, this means that there is *no* tree T, s.t. $||D_T, \tau_\varphi||_\infty \leq \Delta$ (**UNSAT**). □

3. Hardness of approximation of the *'best-fit to triplets under ℓ_∞'* problem

We prove hardness of approximation of this problem by showing that the reduction described in Section 2 satisfies stronger requirements:

SAT' If φ is satisfiable, then there is a tree T s.t. $||D_T, \tau_\varphi||_\infty \leq \beta$.
UNSAT' If φ is unsatisfiable, then for every tree T, $||D_T, \tau_\varphi||_\infty \geq 1.4\beta$.

The first requirement is exactly **SAT** as phrased in the previous section, and so it follows from Lemma 2.7. **UNSAT'** requires proving a stronger version of Lemma 2.5, for a δ-relaxed version of inequalities **A1–B3**, for some positive δ which will be defined soon.

A'1 $D_T(\mathcal{T}, \mathcal{F}) \geq 2\alpha + 2\beta - \delta$
A'2 $\forall i = 1..n : D_T(\mathcal{F}; x_i \bar{x}_i) \leq \alpha + \delta$; $D_T(\mathcal{T}; x_i \bar{x}_i) \leq \alpha + \delta$
B'1 $\forall j = 1..m : D_T(y_1^j; l_2^j l_3^j) \leq \alpha + \delta$; $D_T(y_2^j; l_1^j l_3^j) \leq \alpha + \delta$; $D_T(y_3^j; l_1^j l_2^j) \leq \alpha + \delta$
B'2 $\forall j = 1..m : D_T(y_1^j; \mathcal{T}\mathcal{F}) \geq \alpha - \delta$; $D_T(y_2^j; \mathcal{T}\mathcal{F}) \geq \alpha - \delta$; $D_T(y_3^j; \mathcal{T}\mathcal{F}) \geq \alpha - \delta$
B'3 $\forall j = 1..m : D_T(\mathcal{T}; y_1^j y_2^j) \leq \alpha + \delta$; $D_T(\mathcal{T}; y_1^j y_3^j) \leq \alpha + \delta$; $D_T(\mathcal{T}; y_2^j y_3^j) \leq \alpha + \delta$.

Let T be a tree satisfying **A'1–B'3** above for some $\delta < \frac{2\beta}{5}$, and let v_φ be the mid-point of $path(\mathcal{F}, \mathcal{T})$. Let $v_\mathcal{T}$ and $v_\mathcal{F}$ be the points whose distance from v_φ is **exactly** $\beta - 1.5\delta$ on the paths to $\mathcal{T}$ and $\mathcal{F}$ respectively. Note that by **A'1**, $D_T(\mathcal{F}, v_\mathcal{F}) \geq \alpha + \delta$ and $D_T(\mathcal{T}, v_\mathcal{T}) \geq \alpha + \delta$ (see Fig. 5). Using this, we prove a stronger version of Lemma 2.1:

Lemma 3.1. *For all taxa u, v, y in T, we have*

1. *If $D_T(\mathcal{F}; uv) \leq \alpha + \delta$ and $D_T(\mathcal{T}; uv) \leq \alpha + \delta$, then either u is a descendant of $v_\mathcal{T}$ and v is a descendant of $v_\mathcal{F}$ or vice versa.*
2. *If both u and v are descendants of $v_\mathcal{F}$, and $D_T(y; uv) < D_T(y; \mathcal{T}\mathcal{F}) + D_T(v_\mathcal{T}, v_\mathcal{F})$ then y is not a descendant of $v_\mathcal{T}$.*

Proof. 1. As in the proof of Lemma 2.1(1), since $D_T(\mathcal{F}; uv) + D_T(\mathcal{T}; uv) < D_T(\mathcal{F}, \mathcal{T})$ we have that $C(u, \mathcal{T}, \mathcal{F})$ and $C(v, \mathcal{T}, \mathcal{F})$ are distinct vertices on $path(\mathcal{F}, \mathcal{T})$, one at distance at most $\alpha + \delta$ from $\mathcal{F}$ and the other at distance at most $\alpha + \delta$ from $\mathcal{T}$.
2. Assume, to the contrary, that y is a descendant of $v_\mathcal{T}$. Since both u and v are descendants of $v_\mathcal{F}$, the path from y to $path(u, v)$ must contain both $v_\mathcal{T}$ and $v_\mathcal{F}$, which are both on $path(\mathcal{F}, \mathcal{T})$ (see Fig. 6). Thus we must have that $D_T(y; uv) \geq D_T(y; \mathcal{T}\mathcal{F}) + D_T(v_\mathcal{T}, v_\mathcal{F})$, contradicting the assumption. □

The following corollaries follow from Lemma 3.1 and **A'1–B'3**:

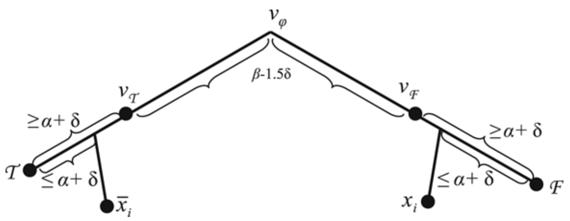

Fig. 5. The topology of a tree satisfying **A'1–2**.

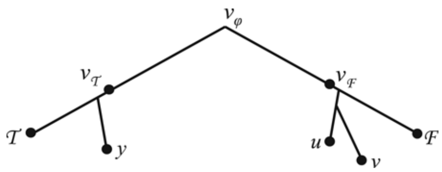

Fig. 6. Proof of Lemma 3.1(2).

Corollary 3.2. *Assume that* $\delta < \frac{2\beta}{3}$. *Then for each* $i = 1..n$, *one of the vertices* $x_i, \bar{x_i}$ *is a descendant of* $v_\mathcal{T}$ *and the other is a descendant of* $v_\mathcal{F}$.

Corollary 3.3. *Assume that* $\delta < \frac{2\beta}{5}$. *Let* j *be in* $\{1, \ldots, m\}$, *and let* $\{a, b, c\} = \{1, 2, 3\}$. *If* l_a^j *and* l_b^j *are descendants of* $v_\mathcal{F}$, *then* y_c^j *is not a descendant of* $v_\mathcal{T}$.

Notice that the relaxation of the bounds prevents us from proving (as in Corollary 2.3) that the y-taxa are descendants of $v_\mathcal{F}$. However, the weaker claim in Corollary 3.3 is sufficient to contradict the bounds in **B'3**, due to the slackness we had in the proof of Corollary 2.4:

Corollary 3.4. *If for some* $j = 1..m$, l_1^j, l_2^j, l_3^j *are* all *descendants of* $v_\mathcal{F}$, *then the bounds in* **B'3** *cannot hold.*

Proof. By Corollary 3.3, if l_1^j, l_2^j, l_3^j are *all* descendants of $v_\mathcal{F}$, then none of y_1^j, y_2^j, y_3^j are descendants of $v_\mathcal{T}$. This implies, for instance, that $C(\mathcal{T}, y_1^j, y_2^j)$ is not a descendant of $v_\mathcal{T}$ as well, so $D_T(\mathcal{T}; y_1^j y_2^j) > D_T(\mathcal{T}, v_\mathcal{T}) \geq \alpha + \delta$ (by definition of $v_\mathcal{T}$), contradicting **B'3** . □

This corollary leads us to the following:

Lemma 3.5. *If there is an edge-weighted tree* T *over the set of taxa* S_φ *satisfying* $||D_T, \tau_\varphi||_\infty < 1.4\beta$, *then the formula* φ *is satisfiable.*

Proof. A tree satisfying $||D_T, \tau_\varphi||_\infty \leq \beta + \delta$ satisfies the δ-relaxed bounds in **A'1–B'3** as well. So if $||D_T, \tau_\varphi||_\infty < 1.4\beta$, then T satisfies the δ-relaxed bounds for $\delta = ||D_T, \tau_\varphi||_\infty - \beta < \frac{2}{5}\beta$. Now since **A'1–2** hold, the assignment σ_T is well defined (Corollary 3.2). Assume that σ_T does not satisfy some clause c_j of φ. Then, by definition of σ_T, l_1^j, l_2^j, l_3^j are all descendants of $v_\mathcal{F}$, and by Corollary 3.3 the bounds in **B'3** cannot hold — a contradiction. □

For a distance table τ, let $OPT(\tau)$ be the minimal value k, for which there is a tree T s.t. $||D_T, \tau||_\infty \leq k$. By Lemma 2.6, if φ is satisfiable then $OPT(\tau_\varphi) \leq \beta$, and by Lemma 3.5, if φ is unsatisfiable then $OPT(\tau_\varphi) \geq 1.4\beta$. Thus if there was a polynomial time algorithm $\mathcal{A}$ which is guaranteed to approximate $OPT(\tau)$ within a factor *smaller* than 1.4, then satisfiability of a formula φ could be determined by executing $\mathcal{A}$ on τ_φ and obtaining $k = ||\tau_\varphi, \mathcal{A}(\tau_\varphi)||_\infty$. If $k < 1.4\beta$ then φ must be satisfiable, and if $k \geq 1.4\beta$ then φ is unsatisfiable. Hence it is NP-hard to find a tree which approximates the optimal ℓ_∞ distance to a given triplet-dissimilarity table by a ratio smaller than 1.4.

4. Hardness of approximation of maximal distortion and other implied results

We now use the reductions presented in the previous sections to obtain several related hardness results. As the constructions are similar to those in previous sections, most proofs in this section are only sketched.

4.1. Hardness of approximation of maximal distortion

Recall the *maximal distortion* between two triplet-dissimilarity tables:

$$MaxDist(\tau_1, \tau_2) \triangleq \max_{i,j,k\in S}\left\{\frac{\tau_1(i;\,jk)}{\tau_2(i;\,jk)}\right\}\cdot \max_{i,j,k\in S}\left\{\frac{\tau_2(i;\,jk)}{\tau_1(i;\,jk)}\right\} \quad (\text{where } 0/0 \triangleq 1).$$

We use a reduction similar to the one in Section 2 to prove that *MaxDist* of the closest tree cannot be approximated by any multiplicative factor. First, note that scaling a tree by a multiplicative factor does not affect its *MaxDist* from a given triplet-dissimilarity table. In other words, $MaxDist(\tau, D_T) = MaxDist(\tau, D_{[\gamma T]})$, where γT is the weighted tree obtained by multiplying edge weights of T by the positive constant γ. This means that if there is a tree T s.t. $MaxDist(\tau, D_T) \leq \rho$, then there is a tree T' (obtained by re-scaling T) s.t.

$$\max\left\{\max_{i,j,k\in S}\left\{\frac{\tau(i;\,jk)}{D_{T'}(i;\,jk)}\right\},\ \max_{i,j,k\in S}\left\{\frac{D_{T'}(i;\,jk)}{\tau(i;\,jk)}\right\}\right\} \leq \sqrt{\rho}.$$

A 3 CNF formula φ is translated to a triplet-dissimilarity table $\tilde{\tau}_\varphi$ which enforces the inequalities in **A1–B3** through bounds on maximal distortion as follows: An upper bound $D_T(i;\,jk) \leq \omega$ is enforced by setting $\tilde{\tau}_\varphi(i;\,jk) = \frac{\omega}{\sqrt{\rho}}$, and a lower bound $D_T(i;\,jk) \geq \omega$ is enforced by setting $\tilde{\tau}_\varphi(i;\,jk) = \sqrt{\rho}\omega$, where $\rho \geq 1$ will soon be defined. By the argument raised above, a tree whose *MaxDist* from $\tilde{\tau}_\varphi$ is at most ρ implies a tree satisfying all bounds. We now show how to set ρ and fill in the rest of the entries of $\tilde{\tau}_\varphi$, such that the tree T described in the proof of Lemma 2.6 satisfies $MaxDist(\tilde{\tau}_\varphi, D_T) \leq \rho$: Recall that in such a tree, triplet-dissimilarities not mentioned in **A1–B3** fall within the interval $[\alpha, \alpha + 2\beta]$ for distinct-taxa triplets, and within the interval $[2\alpha, 2\alpha + 2\beta]$ for taxon-pairs (see discussion following Lemma 2.6). In order to allow triplet-dissimilarities within these intervals, we set $\rho = \max\{\frac{\alpha+2\beta}{\alpha}, \frac{2\alpha+2\beta}{2\alpha}\} = 1 + 2\frac{\beta}{\alpha}$, and set the relevant entries of $\tilde{\tau}_\varphi$ to $\sqrt{\rho}\cdot\alpha$ and $\sqrt{\rho}\cdot 2\alpha$ (corresponding to distinct-taxa triplets and taxon-pairs respectively). The following lemma ensures that $\tilde{\tau}_\varphi$ and ρ have the desired properties:

Lemma 4.1. *Let $\alpha, \beta > 0$ be given, and let $\rho = 1 + 2\frac{\beta}{\alpha}$. Further, let $\tilde{\tau}_\varphi$ be the triplet-dissimilarity table defined by α, β and ρ as described above, then:*

SAT” *If φ is satisfiable, then there exists a tree T s.t. $MaxDist(D_T, \tilde{\tau}_\varphi) \leq \rho$.*
UNSAT” *If φ is unsatisfiable, then for every tree T, $MaxDist(D_T, \tilde{\tau}_\varphi) \geq \rho(1 + \frac{2\beta}{3\alpha})$.*

Proof (*An Outline*). **SAT”** is guaranteed by the tree construction described in the proof of Lemma 2.6 and by the value we chose for ρ, as discussed above. **UNSAT”** is proved by adjusting the proof in Section 3. First, we define a set of bounds **A”1–B”3**, obtained by a relaxation of **A1–B3** by a *multiplicative* factor of $\delta > 1$ as follows:

A”1 $D_T(\mathcal{T}, \mathcal{F}) \geq (2\alpha + 2\beta)/\delta$
A”2 $\forall i = 1..n : D_T(\mathcal{F}; x_i\bar{x}_i) \leq \alpha\delta$; $D_T(\mathcal{T}; x_i\bar{x}_i) \leq \alpha\delta$
B”1 $\forall j = 1..m : D_T(y_1^j; l_2^j l_3^j) \leq \alpha\delta$; $D_T(y_2^j; l_1^j l_3^j) \leq \alpha\delta$; $D_T(y_3^j; l_1^j l_2^j) \leq \alpha\delta$
B”2 $\forall j = 1..m : D_T(y_1^j; \mathcal{T}\mathcal{F}) \geq \alpha/\delta$; $D_T(y_2^j; \mathcal{T}\mathcal{F}) \geq \alpha/\delta$; $D_T(y_3^j; \mathcal{T}\mathcal{F}) \geq \alpha/\delta$
B”3 $\forall j = 1..m : D_T(\mathcal{T}; y_1^j y_2^j) \leq \alpha\delta$; $D_T(\mathcal{T}; y_1^j y_3^j) \leq \alpha\delta$; $D_T(\mathcal{T}; y_2^j y_3^j) \leq \alpha\delta$.

Next, we consider a tree satisfying the relaxed bounds, and define the internal points $v_{\mathcal{T}}, v_{\mathcal{F}}$ to be at distance $\frac{\alpha+\beta}{\delta} - \delta\alpha$ from v_φ. To prove the analogue of Lemma 3.1(1), it is required that $D_T(\mathcal{F}, \mathcal{T})$ be strictly larger than $2\alpha\delta$, which by **A”1** reduces to $\frac{\alpha+\beta}{\delta} - \delta\alpha > 0$, i.e. $\delta < \sqrt{1 + \frac{\beta}{\alpha}}$. Corollary 3.3 is proven by the analogue of Lemma 3.1(2). For this we require $\alpha/\delta + D_T(v_{\mathcal{F}}, v_{\mathcal{T}}) > \alpha\delta$. Since $D_T(v_{\mathcal{F}}, v_{\mathcal{T}}) = 2(\frac{\alpha+\beta}{\delta} - \delta\alpha)$, this is equivalent to $\delta < \sqrt{1 + \frac{2\beta}{3\alpha}}$. This latter upper bound on δ implies also the previous one, and hence if φ is unsatisfiable, there is no tree satisfying the δ-relaxed bounds in **A”1–B”3** for $\delta < \sqrt{1 + \frac{2\beta}{3\alpha}}$. In other words, there is no tree T satisfying:

$$\max\left\{\max_{i,j,k\in S_\varphi}\left\{\frac{\tilde{\tau}_\varphi(i;\,jk)}{D_T(i;\,jk)}\right\},\ \max_{i,j,k\in S_\varphi}\left\{\frac{D_T(i;\,jk)}{\tilde{\tau}_\varphi(i;\,jk)}\right\}\right\} < \sqrt{\rho\left(1 + \frac{2\beta}{3\alpha}\right)}.$$

This means that there is no tree whose *MaxDist* from $\tilde{\tau}_\varphi$ is less than $\rho(1 + \frac{2\beta}{3\alpha})$, as claimed. □

Now, assume that there was a polynomial time algorithm $\mathcal{A}$ which given a triplet-dissimilarity table τ, was guaranteed to return a tree whose *MaxDist* from τ is at most K-times the *MaxDist* of the closest tree to τ, for some constant K. Such an algorithm may be used to efficiently deduce whether a formula φ is satisfiable in the following way: given a formula φ, calculate $\tilde{\tau}_\varphi$ with parameters α, β s.t. $K < 1 + \frac{2}{3}\frac{\beta}{\alpha}$. Execute algorithm $\mathcal{A}$ on this triplet-dissimilarity table to receive a tree T, and calculate $r = MaxDist(\tilde{\tau}_\varphi, D_T)$. Now, if $r \leq K\rho$ (where $\rho = 1 + 2\frac{\beta}{\alpha}$ as previously defined), then φ must be satisfiable due to **UNSAT"**. If, on the other hand, $r > K\rho$, then since $\mathcal{A}$ guarantees a K-approximation, there is *no* tree whose *MaxDist* from $\tilde{\tau}_\varphi$ is at most ρ. From **SAT"** it follows that φ is unsatisfiable.

4.2. Fitting distances of distinct-taxa triplets

Triplet-distance tables, as we defined them, contain entries corresponding to distinct-taxa triplets as well as entries corresponding to taxon-pairs (i.e. $\tau(i; jj)$). In some scenarios it is more natural to separately address pairwise dissimilarities and triplet-dissimilarities. Therefore, we are interested in the problem of finding a best-fit tree to a triplet-dissimilarity table τ, considering entries corresponding only to *distinct-taxa triplets*. The best-fit analysis can be done under any of the ℓ_p norms or *MaxDist*. Results similar to the ones presented above apply in this case as well. The only modification required in order to adapt the reduction to this case is changing the bounds in **A1**, which correspond to the pairwise-distance between $\mathcal{T}$ and $\mathcal{F}$. To ensure a similar bound, we introduce an additional taxon into S_φ: $\mathcal{F}'$, and replace **A1** by:

$\overline{\textbf{A1}}$ $D_T(\mathcal{T}; \mathcal{F}\mathcal{F}') \geq 2\alpha + 2\beta$.

It is easy to see that this new bound implies the desired lower bound on the distance between $\mathcal{T}$ and $\mathcal{F}$ (i.e. **A1**). The original set of bounds is, therefore, equivalent to this one, and all claims proven for it apply here as well. The tree described in the proof of Lemma 2.6 is adapted to the introduction of $\mathcal{F}'$, by turning the original taxon $\mathcal{F}$ into an internal vertex, and adding two zero-weight edges from this vertex to $\mathcal{F}, \mathcal{F}'$. All triplet-dissimilarities concerning $\mathcal{F}'$ are set to be equal to their counterparts concerning $\mathcal{F}$. The analysis done in previous sections is easily adjusted to accommodate this modification of the reduction.

4.3. Best-fit ultrametric

It is possible to generalize all hardness results shown in this paper for ultrametrics as well. A weighted tree is called *ultrametric* if it contains a point which is equidistant from all leaves; this point may be an internal vertex, or a degenerate (degree 2) vertex situated on one of the edges. The problem of finding a best-fit ultrametric to a given *dissimilarity matrix* under ℓ_∞ (and *MaxDist*) was shown to have a polynomial time algorithm in [9,7].

In the case of triplet-dissimilarities, the same reductions presented in Sections 3 and 4.1 imply that it is NP-hard to find (and to approximate) a best-fit ultrametric under the ℓ_∞ norm, as well as *MaxDist*. To see this, observe that if φ is unsatisfiable, then the lower bounds proved for **UNSAT'** in Lemma 3.5 and for **UNSAT"** in Lemma 4.1 (for ℓ_∞ and *MaxDist* respectively) are clearly valid when the trees are restricted to be ultrametrics. We are left to show that if φ is satisfiable then there is an *ultrametric* tree satisfying all bounds. This follows from the fact that the construction described in the proof of Lemma 2.6 yields an ultrametric tree, since the internal point v_φ is at the same distance $(\alpha + \beta)$ from all taxa.

5. A constant-rate approximation scheme

In this section we present a constant-rate approximation algorithm for the problem of finding a closest tree under ℓ_∞ to a given triplet-dissimilarity table. Our algorithm is based on an approximation algorithm for the corresponding problem concerning pairwise-dissimilarities. The main result is stated in the following theorem.

Theorem 5.1. *A polynomial time r-approximation algorithm for finding a tree closest under ℓ_∞ to a given* dissimilarity matrix *implies a polynomial time $(\frac{3}{2}r + 6)$-approximation algorithm for finding a tree closest under ℓ_∞ to a given* triplet-dissimilarity table.

Our approximation algorithm, $\mathcal{APP}$, consists of two stages:

$\mathcal{APP}1$. Given a triplet-dissimilarity table τ over taxon-set S, calculate a dissimilarity matrix D^τ over S as follows: $\forall i, j \in S : D^\tau(i, j) = \tau(i; jj)$.

$\mathcal{APP}2$. Execute the r-approximation algorithm on D^τ to obtain an edge-weighted tree T^{out}.

To analyze the approximation ratio of the above algorithm, we start with some notations. For an arbitrary taxon-pair $i, j \in S$, denote $D^\tau_{\min}(i,j) = \min_{k\in S}\{\tau(i; jk) + \tau(j; ik)\}$, and similarly $D^\tau_{\max}(i,j) = \max_{k\in S}\{\tau(i; jk) + \tau(j; ik)\}$. Furthermore, denote by $I^\tau = \max_{i,j\in S}\{D^\tau_{\max}(i,j) - D^\tau_{\min}(i,j)\}$ the maximum difference between $D^\tau_{\max}$ and $D^\tau_{\min}$. The following lemma contains two basic inequalities required for the proof of our approximation result.

Lemma 5.2. *Let τ be a triplet-dissimilarity table, and D^τ be the corresponding dissimilarity matrix defined in $\mathcal{APP}1$. Further let T be an edge-weighted tree with corresponding additive distance matrix D_T and triplet-dissimilarity table τ_T. Then we have the following:*

$$\frac{1}{4} I^\tau \le ||\tau, \tau_T||_\infty \le \frac{3}{2}\left(I^\tau + ||D^\tau, D_T||_\infty\right).$$

Proof. First we prove that $\frac{1}{4}I^\tau \le ||\tau,\tau_T||_\infty$. Let i, j be a taxon-pair s.t. $D^\tau_{\max}(i,j) - D^\tau_{\min}(i,j) = I^\tau$, and let $k_{\max}$ be a taxon s.t. $D^\tau_{\max}(i,j) = \tau(i; jk_{\max}) + \tau(j; ik_{\max})$. Since $\tau_T(i; jk) + \tau_T(j; ik) = D_T(i,j)$, for all $k \in S$, then:

$$\begin{aligned} D^\tau_{\max}(i,j) - D_T(i,j) &= [\tau(i; jk_{\max}) + \tau(j; ik_{\max})] - [\tau_T(i; jk_{\max}) + \tau_T(j; ik_{\max})] \\ &= [\tau(i; jk_{\max}) - \tau_T(i; jk_{\max})] + [\tau(j; ik_{\max}) - \tau_T(j; ik_{\max})] \\ &\le 2||\tau, \tau_T||_\infty. \end{aligned} \tag{1}$$

Similarly, if $k_{\min}$ is a taxon s.t. $D^\tau_{\min}(i,j) = \tau(i; jk_{\min}) + \tau(j; ik_{\min})$, then:

$$D^\tau_{\min}(i,j) - D_T(i,j) \ge -2||\tau, \tau_T||_\infty. \tag{2}$$

Now, since $D^\tau_{\max}(i,j) - D^\tau_{\min}(i,j) = I^\tau$, then by subtracting (2) from (1) we get $I^\tau \le 4||\tau,\tau_T||_\infty$, thus proving the left inequality.

We now turn to prove the right inequality of the lemma. Given an arbitrary taxon-triplet $i, j, k \in S$, denote $\varepsilon(i; jk) = \tau(i; jk) - \tau_T(i; jk)$. We will show that $|\varepsilon(i; jk)| \le \frac{3}{2}(I^\tau + ||D^\tau, D_T||_\infty)$. First,

$$\begin{aligned} |\varepsilon(i; jk) + \varepsilon(j; ik)| &= |[\tau(i; jk) - \tau_T(i; jk)] + [\tau(j; ik) - \tau_T(j; ik)]| \\ &= |[\tau(i; jk) + \tau(j; ik)] - [\tau_T(i; jk) + \tau_T(j; ik)]| \\ &= |\tau(i; jk) + \tau(j; ik) - D_T(i,j)| \\ &\le \left|\tau(i; jk) + \tau(j; ik) - D^\tau(i,j)\right| + \left|D^\tau(i,j) - D_T(i,j)\right| \\ &\le \left|\tau(i; jk) + \tau(j; ik) - D^\tau(i,j)\right| + ||D^\tau, D_T||_\infty \\ &\le I^\tau + ||D^\tau, D_T||_\infty. \end{aligned}$$

The last inequality follows from the fact that $D^\tau(i,j) = \tau(i; jj) + \tau(j; ij)$. Using a similar line of argument we get $|\varepsilon(i; kj) + \varepsilon(k; ij)|, |\varepsilon(j; ki) + \varepsilon(k; ji)| \le I^\tau + ||D^\tau, D_T||_\infty$ as well. This is used to obtain the desired bound as follows:

$$\begin{aligned} |\varepsilon(i; jk)| &= \frac{1}{2}\,|[\varepsilon(i; jk) + \varepsilon(j; ik)] + [\varepsilon(i; kj) + \varepsilon(k; ij)] - [\varepsilon(j; ki) + \varepsilon(k; ji)]| \\ &\le \frac{1}{2}\,(|\varepsilon(i; jk) + \varepsilon(j; ik)| + |\varepsilon(i; kj) + \varepsilon(k; ij)| + |\varepsilon(j; ki) + \varepsilon(k; ji)|) \\ &\le \frac{3}{2}\left(I^\tau + ||D^\tau, D_T||_\infty\right). \quad \square \end{aligned}$$

Our main result (Theorem 5.1) is directly implied by the following lemma:

Lemma 5.3. *Given a triplet-dissimilarity table τ, denote by τ^{out} the triplet-dissimilarity table induced by the output tree T^{out} returned by the algorithm $\mathcal{APP}$. Then for every triplet-dissimilarity table τ_T induced by an arbitrary edge-weighted tree T, we have:*

$$||\tau, \tau^{\text{out}}||_\infty \le \left(\frac{3}{2}r + 6\right)||\tau, \tau_T||_\infty.$$

Proof. Denote by D^τ the dissimilarity matrix computed in $\mathcal{APP}1$, and by D^{out} and D_T the metrics induced over the leaves of T^{out} and T, respectively. The lemma is proved by the following sequence of inequalities:

$$||\tau, \tau^{\text{out}}||_\infty \leq \frac{3}{2}\left(I^\tau + ||D^\tau, D^{\text{out}}||_\infty\right) \quad (3)$$

$$\leq \frac{3}{2}\left(4||\tau, \tau_T||_\infty + ||D^\tau, D^{\text{out}}||_\infty\right) \quad (4)$$

$$\leq \frac{3}{2}\left(4||\tau, \tau_T||_\infty + r||D^\tau, D_T||_\infty\right) \quad (5)$$

$$\leq \left(\frac{3}{2}r + 6\right)||\tau, \tau_T||_\infty. \quad (6)$$

(3) and (4) follow from the right and left inequalities of Lemma 5.2, respectively. The approximation ratio of the algorithm executed during $\mathcal{APP}2$ implies (5). (6) follows from the fact that $||D^\tau, D_T||_\infty \leq ||\tau, \tau_T||_\infty$, which holds since $D^\tau(i, j) = \tau(i; jj)$ and $D_T(i, j) = \tau_T(i; jj)$ for every taxon-pair $i, j \in S$. □

By Theorem 5.1, the 3-approximation algorithms for pairwise dissimilarities presented in [1,8] imply a $10\frac{1}{2}$-approximation algorithm for triplet-dissimilarities. However, the 3-approximation ratio of the algorithms in [1,8] is proved under the assumption that the input dissimilarity matrix is a *distance metric*. Therefore, this bound (of $10\frac{1}{2}$) is valid only if the matrix D^τ computed in $\mathcal{APP}1$ satisfies the triangle inequality. When the triangle inequality is not assumed, the analysis in [1,8] can be modified to yield a 6-approximation ratio, rather than the original 3-approximation. This 6-approximation algorithm leads, by Theorem 5.1, to a 15-approximation of the closest tree to an arbitrary triplet-dissimilarity table under ℓ_∞.

6. Discussion

In this paper we discussed the hardness of several problems of fitting a phylogenetic tree to a given triplet-dissimilarity table. This question is motivated by several recent works which reconstruct trees using triplet-dissimilarities [11,10,8]. The optimization criteria considered in this paper are the ℓ_∞ norm and *MaxDist*, which measure the maximum discrepancy (difference and ratio respectively) between the input dissimilarities and the ones induced by the desired tree. It is interesting to find out whether similar hardness results apply also for other common distance measures such as the ℓ_1 and ℓ_2 norms.

The construction in Lemma 2.6 which implies our basic NP-hardness result yields a tree containing two vertices of very high degree. Common models for phylogenetic trees assume a binary tree (meaning that all internal vertices have degree 3). Furthermore, edge weights are assumed to lie within an interval $[w_{\min}, w_{\max}]$, where $w_{\min}$ and $w_{\max}$ are strictly positive constants independent of the size of the tree. It is interesting to find out whether our NP-hardness results apply also when introducing these assumptions on the desired tree, and specifically what is the smallest ratio between $w_{\max}$ and $w_{\min}$ mentioned above which still gives similar hardness results. Can this ratio be a constant independent on n? Does the NP-hardness result apply also for binary trees with *uniform* edge weights?

Another question relates to the approximation ratio given in Section 5. Possibly, a better approximation ratio may be obtained by a closer analysis of the algorithms in [1,8].

Acknowledgement

We thank the anonymous referees for their insightful comments.

References

[1] R. Agarwala, V. Bafna, M. Farach, M. Paterson, M. Thorup, On the approximability of numerical taxonomy (fitting distances by tree metrics), SIAM J. Comput. 28 (3) (1999) 1073–1085.
[2] Y. Bartal, N. Linial, M. Mendel, A. Naor, Low dimensional embeddings of ultrametrics, European J. Combin. 25 (1) (2004) 87–92.
[3] P. Buneman, The recovery of trees from measures of dissimilarity, in: Mathematics in the Archeological and Historical Sciences, 1971, pp. 387–395.
[4] W. Day, Computational complexity of inferring phylogenies from dissimilarity matrices, Bull. Math. Biol. 49 (4) (1987) 461–467.
[5] P. Erdos, M. Steel, L. Szekely, T. Warnow, A few logs suffice to build (almost) all trees (I), Random Struct. Algorithms 14 (1999) 153–184.

[6] P. Erdos, M. Steel, L. Szekely, T. Warnow, A few logs suffice to build (almost) all trees (II), Theoret. Comput. Sci. 221 (1999) 77–118.
[7] M. Farach, S. Kannan, T. Warnow, A robust model for finding optimal evolutionary trees, Algorithmica 13 (1–2) (1995) 155–179.
[8] I. Gronau, S. Moran, Neighbor joining algorithms for inferring phylogenies via LCA-distances, J. Comput. Biol. 14 (1) (2007) 1–15.
[9] M. Křivánek, The complexity of ultrametric partitions on graphs, Inform. Process. Lett. 27 (1988) 265–270.
[10] D. Levy, R. Yoshida, L. Pachter, Beyond pairwise distances: Neighbor-joining with phylogenetic diversity estimates, Mol. Biol. Evol. 23 (3) (2006) 491–498.
[11] V. Ranwez, O. Gascuel, Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets, Mol. Biol. Evol. 19 (11) (2002) 1952–1963.
[12] N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[13] S. Sattath, A. Tversky, Additive similarity trees, Psychometrica 42 (3) (1977) 319–345.
[14] M. Waterman, T. Smith, M. Singh, W. Beyer, Additive evolutionary trees, J. Theoret. Biol. 64 (2) (1977) 199–213.