On the complexity of computing the temporal hybridization number for two phylogenies

Peter J. Humphries[a], Simone Linz[b], Charles Semple[c]

[a]*Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand. Email: pjhumphries@gmail.com.*

[b]*Center for Bioinformatics, University of Tübingen, Tübingen, Germany. Email: linz@informatik.uni-tuebingen.de*

[c]*Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand. Email: charles.semple@canterbury.ac.nz*

Abstract

Phylogenetic networks are now frequently used to explain the evolutionary history of a set of species for which a collection of gene trees, reconstructed from genetic material of different parts of the species' genomes, reveal inconsistencies. However, in the context of hybridization, the reconstructed networks are often not temporal. If a hybridization network is temporal, then it satisfies the time constraint of instantaneously occurring hybridization events; i.e. all species that are involved in such an event coexist in time. Furthermore, although a collection of phylogenetic trees can often be merged into a hybridization network that is temporal, many algorithms do not necessarily find such a network since their primary optimization objective is to minimize the number of hybridization events. In this paper, we present a characterization for when two rooted binary phylogenetic trees admit a temporal hybridzation network. Furthermore, we show that the underlying optimization problem is APX-hard and, therefore, NP-hard. Thus, unless P=NP, it is unlikely that there are efficient algorithms for either computing an exact solution or approximating it within a ratio arbitrarily close to one.

Keywords: agreement forest, APX-hard, phylogenetic network, phylogenetic tree, temporal network

MSC: 05C05, 92D15

1. Introduction

In the process of slowly drifting away from the tradition of representing evolution by means of a phylogenetic tree, phylogenetic networks are now becoming increasingly important for investigating the evolutionary history of a set of species. This change in concept consequently necessitates the development of algorithms that analyze non-tree-like reticulation processes, such as horizontal gene transfer, hybridization, and recombination, in a way that is best suited to the many biological constraints. Recently, a number of tools have become available that reconstruct a phylogenetic network from a collection of phylogenetic trees, clusters, or rooted triplets so as to quantify the extent of reticulation (e.g. see [1, 6, 8, 11, 10], and [9] for an analysis of the relationship among these approaches). In particular, in terms of two phylogenetic trees, the following optimization problem has attracted considerable interest. Suppose that we are given two rooted phylogenetic trees that correctly represent the evolutionary history of a set of present-day species for two distinct genetic markers. What is the minimum number of hybridization events that is needed to explain the evolution of the species under consideration? We call this problem Minimum Hybridization.

Despite the NP-hardness of Minimum Hybridization, several exact algorithms exist that solve many biological problem instances reasonably quickly [1, 6, 8]. Instead of directly minimizing the number of hybridization events over all hybridization networks that explain two phylogenetic trees, most of these algorithms make use of the concept of agreement forests [2]. Roughly speaking, an agreement forest for two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on the same set of species is a smallest collection of non-overlapping subtrees that are common to $\mathcal{T}$ and $\mathcal{T}'$. Having calculated an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, the polynomial-time algorithm HybridPhylogeny [3] can then be applied to reconstruct a hybridization network that explains $\mathcal{T}$ and $\mathcal{T}'$, where the number of hybridization events is at most the size of the forest minus one (precise definitions are given in the next section). In such a network, each vertex that has more than one incoming arc is referred to as a hybridization vertex and represents a hybridization event. As an example,

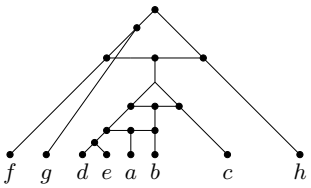

Figure 1: A hybridization network $\mathcal{N}$ with three hybridization vertices and hybridization arcs drawn horizontally.

Figure 1 shows a hybridization network with three hybridization vertices.

However, as pointed out in [13, 14], although a hybridization network might explain conflicting signals in a data set, there may be no process with instantaneously occurring hybridization events that realizes this network. This is due to the fact that the three species that are involved in a hybridization event, i.e. a hybrid species and its two parental species, do not coexist in time. We say that a hybridization network is *temporal* if each hybridization event can be realized between coexisting species. Baroni et al. [3] and Moret et al. [14] showed that a non-temporal hybridization network can always be transformed into a temporal one by adding extinct or unsampled species. Linz et al. [12] showed that determining the minimum number of such extinct or unsampled species is an NP-hard task. Nevertheless, with or without this hardness result, it is natural to reconstruct a temporal hybridization network for a given data set without allowing for additional species. An immediate question is the following. Suppose that we are given two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ that correctly represent the evolutionary history of a set of present-day species for two distinct genetic markers. Does there exist a temporal hybridization network that explains $\mathcal{T}$ and $\mathcal{T}'$? In this paper, we present a characterization to answer this question in terms of so-called *temporal forests* and, subsequently, investigate the problem of minimizing the number of hybridization events needed to merge two rooted phylogenetic trees into a temporal hybridization network if such a network exists. We refer to the latter optimization problem as Minimum Temporal Hybridization. While it is a restricted version of Minimum Hybridization, we will see that it remains computationally hard; in particular, it is APX-hard. Note that if two rooted phylogenetic trees admit a temporal hybridization network, then the minimum number of hybridization

events is always at least as big as the solution to Minimum Hybridization for the same instance. Moreover, while each pair of rooted phylogenetic trees admits a hybridization network it does not necessarily also admit a temporal hybridization network. An example for two trees that cannot be merged into the latter type of network is shown in Figure 2.

The paper is organized as follows. The next section contains some notation and terminology that is used throughout this paper. In Section 3, we present a characterization for when two rooted binary phylogenetic trees admit a temporal hybridization network and, if so, show how to calculate the minimum number of hybridization vertices in such a network. Using this characterization, we show in Section 4 that Minimum Temporal Hybridization is APX-hard and, thus, NP-hard. Unless otherwise stated, the notation and terminology in this paper follow [17].

We end this section with two remarks. First, we expect the characterization of Minimum Temporal Hybridization in terms of temporal agreement forests to become equally important for the development of 'efficient' algorithms as the characterization of Minimum Hybridization in terms of agreement forests. In particular, we will show in a forthcoming paper, that temporal agreement forests are a useful tool to show that Minimum Temporal Hybridization is fixed-parameter tractable. Thus, although being NP-hard, the problem is likely to be tractable for many biological data sets even for a large set of taxa. Second, the results presented in this paper cannot be directly applied to the reconstructing of a phylogenetic network that explains two gene trees whose evolutionary past is likely to include horizontal gene transfer events instead of hybridization events. The reason for this is that, for the former type of event to be temporal, only two species, one of which is the reticulate species, need to coexist in time (as opposed to three for hybridization). For example, the two trees shown in Figure 2 can be explained by invoking two temporal horizontal gene transfer events.

2. Phylogenetic Trees, Networks, and Agreement Forests

This section provides preliminary definitions which are used throughout the rest of the paper.

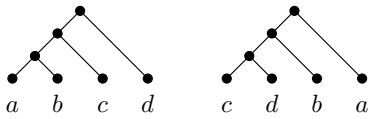

Figure 2: Two rooted binary phylogenetic trees that do not admit a temporal hybridization network.

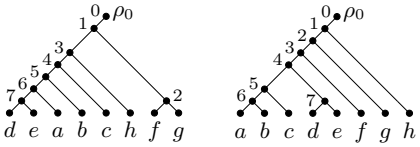

Figure 3: Two planted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$.

Phylogenetic trees. Let X be a finite set. A *rooted binary phylogenetic X-tree* $\mathcal{T}$ is a rooted tree with leaf set X and, apart from the root which has degree two, all interior vertices have degree three. The set X is often referred to as the *label set* of $\mathcal{T}$ and denoted by $\mathcal{L}(\mathcal{T})$. For technical reasons, we frequently view the root of $\mathcal{T}$ as a vertex that is adjoined to the original root via a new pendant edge, in which case, $\mathcal{T}$ is a *planted binary phylogenetic X-tree*. Ignoring the labels of the internal vertices for the moment, Figure 3 shows two planted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ with root label ρ_0 and $\mathcal{L}(\mathcal{T}) = \mathcal{L}(\mathcal{T}') = \{a, b, \ldots, h\}$.

Now, let $\mathcal{T}$ be a rooted phylogenetic X-tree, and let U be a subset of its vertex set, i.e. $U \subseteq V(\mathcal{T})$. The minimal rooted subtree of $\mathcal{T}$ that connects all vertices in U is denoted by $\mathcal{T}(U)$. Furthermore, the rooted tree obtained from $\mathcal{T}(U)$ by contracting all non-root degree-2 vertices is the *restriction of $\mathcal{T}$ to U* and is denoted by $\mathcal{T}|U$. Typically, U is a subset of X. Lastly, a rooted phylogenetic tree $\mathcal{S}$ is *pendant* in $\mathcal{T}$ if $\mathcal{S}$ can be detached from $\mathcal{T}$ by deleting a single edge.

Now, let $\mathcal{T}$ be a rooted phylogenetic X-tree, and let X' be a subset of X. We call X' a *cluster* of $\mathcal{T}$ if there is a vertex v in $\mathcal{T}$ whose set of descendants in X is precisely X'. We denote this cluster by $\mathcal{C}_\mathcal{T}(v)$. Furthermore, the *most recent common ancestor* of X' is the vertex v in $\mathcal{T}$ with $X' \subseteq \mathcal{C}_\mathcal{T}(v)$ such that there exists no vertex v' with $X \subseteq \mathcal{C}_\mathcal{T}(v')$ and $\mathcal{C}_\mathcal{T}(v') \subset \mathcal{C}_\mathcal{T}(v)$. We

denote v by $\mathrm{mrca}_{\mathcal{T}}(X')$.

Temporal hybridization networks. A *hybridization network* $\mathcal{N}$ on a finite set X is a rooted acyclic digraph with the following properties:

(i) the *root* has in-degree 0 and out-degree 2;

(ii) X is the set of *leaves* of the network, that is, the vertices with out-degree 0 and in-degree 1;

(iii) all remaining vertices are *interior vertices*, and each such vertex either has in-degree 1 and out-degree 2 or is a *hybridization vertex* that has in-degree 2 and out-degree 1;

(iv) arcs ending in a hybridization vertex are *hybridization arcs*, while all other arcs in the network are *tree arcs*; and

(v) every interior vertex has at least one outgoing tree arc.

Similar to the definition of a phylogenetic tree, the set X is often referred to as the *label set* of $\mathcal{N}$ and denoted by $\mathcal{L}(\mathcal{N})$.

We remark that the above definition of a hybridization network coincides with that of a so-called tree-child phylogenetic network, introduced by Cardona *et al.* [5]. Furthermore, property (v) in the definition of a hybridization network guarantees that a species that arises from either a speciation or a hybridization event exists for a certain amount of time before possibly going extinct. Hence, assuming that each hybrid species and its two parents coexist in time, no ancestral species yields two new hybrid species and simultaneously becomes extinct.

Now, let $\mathcal{N}$ be a hybridization network on X, and let $\mathcal{T}$ be a rooted binary phylogenetic X'-tree with $X' \subseteq X$. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{N}$ by a sequence of arc and vertex deletions, and degree-2 vertex contractions. Intuitively, if $\mathcal{N}$ displays $\mathcal{T}$, then all of the ancestral relationships visualized by $\mathcal{T}$ are visualized by $\mathcal{N}$.

Again, let $\mathcal{N}$ be a hybridization network on X, and let V be the set of vertices of $\mathcal{N}$. Let $t : V \to \mathbb{R}^+$ be a map such that, for all $u, v \in V$, we have

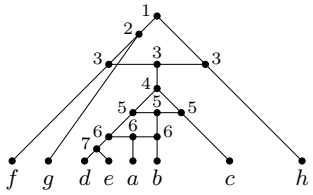

Figure 4: A temporal hybridization network $\mathcal{N}$. For each internal vertex v a temporal labeling $t(v)$ is given next to v. Note that $\mathcal{N}$ is a minimum temporal hybridization network for the two phylogenetic trees shown in Figure 3.

$t(u) = t(v)$ whenever (u, v) is a hybridization arc, and $t(u) < t(v)$ whenever there is a directed path from u to v that contains a tree arc. Then t is a *temporal labeling* of $\mathcal{N}$, in which case, $\mathcal{N}$ is said to be *temporal*. An example of a temporal hybridization network is shown in Figure 4, where arcs directed horizontally are hybridization arcs while arcs directed downwards are tree arcs.

For a temporal hybridization network $\mathcal{N}$, we denote by $h_t(\mathcal{N})$ the number of hybridization vertices of $\mathcal{N}$. Furthermore, if $\mathcal{T}$ and $\mathcal{T}'$ are rooted binary phylogenetic X-trees, we set

$$h_t(\mathcal{T}, \mathcal{T}') = \min\{h_t(\mathcal{N}) : \mathcal{N} \text{ is a temporal hybridization network on } X \text{ that displays } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

A temporal hybridization network $\mathcal{N}$ on X that displays two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ and has the property $h_t(\mathcal{N}) = h_t(\mathcal{T}, \mathcal{T}')$ is said to be a *minimum temporal hybridization network* for $\mathcal{T}$ and $\mathcal{T}'$.

Temporal-agreement forests. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. For the purposes of defining an agreement forest, we view $\mathcal{T}$ and $\mathcal{T}'$ as planted with root vertex ρ_0. An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\mathcal{F} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_k\}$ of planted binary phylogenetic trees with root labels $\rho_0, \rho_1, \ldots, \rho_k$ and label sets $\mathcal{L}_0, \mathcal{L}_1, \ldots, \mathcal{L}_k$, respectively, such that the following properties are satisfied:

(i) the label sets $\mathcal{L}_0, \mathcal{L}_1, \ldots, \mathcal{L}_k$ partition X;

(ii) for all $i \in \{0, 1, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$; and

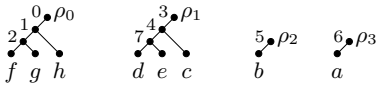

Figure 5: A temporal-agreement forest $\mathcal{F}$ for the two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ that are shown in Figure 3.

(iii) there are one-to-one maps

$$\psi : \{\rho_0, \rho_1, \ldots, \rho_k\} \to V(\mathcal{T}) \text{ and } \psi' : \{\rho_0, \rho_1, \ldots, \rho_k\} \to V(\mathcal{T}')$$

so that the trees in

$$\{\mathcal{T}(\mathcal{L}_i \cup \{\psi(\rho_i)\}) : i \in \{0, 1, \ldots, k\}\}$$

and

$$\{\mathcal{T}'(\mathcal{L}_i \cup \{\psi'(\rho_i)\}) : i \in \{0, 1, \ldots, k\}\}$$

are edge-disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

Ignoring the integer labels of the internal vertices, Figure 5 shows an agreement forest for the two phylogenetic trees depicted in Figure 3.

It is a straightforward consequence of the definition that if $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, then $\{E_0, E_1, \ldots, E_k\}$ and $\{E'_0, E'_1, \ldots, E'_k\}$ are partitions of the edge sets of $\mathcal{T}$ and $\mathcal{T}'$, respectively, where E_i is the edge set of $\mathcal{T}(\mathcal{L}_i \cup \{\psi(\rho_i)\})$ and E'_i is the edge set of $\mathcal{T}'(\mathcal{L}_i \cup \{\psi'(\rho_i)\})$ for all $i \in \{0, 1, \ldots, k\}$. Furthermore, there are natural bijections θ and θ' from the union $V(\mathcal{F})$ of the vertex sets of the trees in $\mathcal{F}$ to the vertex sets of each of $\mathcal{T}$ and $\mathcal{T}'$, respectively. In particular, define $\theta : V(\mathcal{F}) \to V(\mathcal{T})$ as follows. If $v \in X$, set $\theta(v) = v$, while if $v = \rho_i$ for some $i \in \{0, 1, \ldots, k\}$, set $\theta(v) = \psi(\rho_i)$. Otherwise, $v = \text{mrca}_{\mathcal{T}_i}(a, b)$ for some unique $i \in \{0, 1, \ldots, k\}$, where $a, b \in X$ and $a \neq b$, set $\theta(v) = \text{mrca}_{\mathcal{T}}(a, b)$. The bijection θ' is defined analogously. Both θ and θ' are well-defined.

Let $\mathcal{F} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_k\}$ be an agreement forest for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. We say that $\mathcal{F}$ is a *temporal-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ if there exists, for each $i \in \{0, 1, \ldots, k\}$, a temporal labeling t_i of $\mathcal{T}_i$ such that the map $t : V(\mathcal{T}) \to \mathbb{R}^+$ defined by setting $t(u)$ to be the

temporal labeling of $\theta^{-1}(u)$ induces a temporal labeling of $\mathcal{T}$, and the map $t' : V(\mathcal{T}') \to \mathbb{R}^+$ defined by setting $t'(u)$ to be the temporal labeling of $\theta'^{-1}(u)$ induces a temporal labeling of $\mathcal{T}'$. In this case, we refer to $\{t_0, t_1, \ldots, t_k\}$ as a *verification set* for $\mathcal{F}$, the maps t and t' as the temporal labelings of $\mathcal{T}$ and $\mathcal{T}'$ *induced* by $\mathcal{F}$, respectively, and the maps θ and θ' as *temporal embeddings* of $\mathcal{F}$ in $\mathcal{T}$ and $\mathcal{T}'$, respectively. Moreover, if $\mathcal{F}$ contains the smallest number of components amongst all temporal-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, we call $\mathcal{F}$ a *maximum-temporal-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$, in which case, we denote the value of k by $m_t(\mathcal{T}, \mathcal{T}')$. As an example, a temporal-agreement forest $\mathcal{F}$ for the two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Figure 3 is shown in Figure 5. In fact, it is easily checked that $\mathcal{F}$ is a maximum-temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, the temporal embeddings of $\mathcal{F}$ in $\mathcal{T}$ and $\mathcal{T}'$ are shown in Figure 3.

Remark. For readers familiar with acyclic-agreement forests, we end this section with the following note. If $\mathcal{F}$ is a temporal-agreement forest for two rooted binary phylogenetic X-trees, then it is easily seen that $\mathcal{F}$ is also an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. However, the converse does not hold.

3. Characterizing the temporal hybridization number

In this section, we state and prove a characterization for when two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ admit a temporal hybridization network on X that displays $\mathcal{T}$ and $\mathcal{T}'$. This characterization is stated in terms of agreement forests, and as a consequence we show that $h_t(\mathcal{T}, \mathcal{T}') = m_t(\mathcal{T}, \mathcal{T}')$.

Let $\mathcal{N}$ be a hybridization network on X, and view the root ρ_0 of $\mathcal{N}$ as a vertex adjoined to the original root via a new arc. Let $\mathcal{F}$ be a forest of planted binary phylogenetic trees obtained from $\mathcal{N}$ by labeling each hybridization vertex with a distinct ρ_i, where $\rho_i \neq \rho_0$ for all i, deleting each hybridization arc of $\mathcal{N}$, and then suppressing all resulting degree-two vertices. As the forest is unique up to assigning root labels to the roots of each of the trees in $\mathcal{F}$, we refer to $\mathcal{F}$ as the *forest induced by* $\mathcal{N}$.

Proposition 3.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and suppose that $\mathcal{N}$ is a temporal hybridization network on X that displays*

$\mathcal{T}$ and $\mathcal{T}'$. Then the forest induced by $\mathcal{N}$ is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

PROOF. Let $\mathcal{N}$ be a temporal hybridization network on X that displays $\mathcal{T}$ and $\mathcal{T}'$, and let t denote the temporal labeling of $\mathcal{N}$. The proof is by induction on $h_t(\mathcal{N})$. If $h_t(\mathcal{N}) = 0$, then, up to the assignment of temporal labels, $\mathcal{N}$ is isomorphic to each of $\mathcal{T}$ and $\mathcal{T}'$. Thus the forest induced by $\mathcal{N}$ consists of $\mathcal{N}$ itself, and so the proposition holds. Now suppose that the proposition holds for all pairs of rooted binary phylogenetic X'-trees with a temporal hybridization network $\mathcal{N}'$ on X' that displays both trees and for which $h_t(\mathcal{N}') < h_t(\mathcal{N})$.

Let v denote a hybridization vertex of $\mathcal{N}$ for which the assigned temporal labeling $t(v)$ is maximized. Let u and u' denote the parents of v in $\mathcal{N}$. Observe that the arc directed into u and the arc directed out of u (other than (u, v)) are both tree arcs. An analogous observation can also be made for u'. Furthermore, by the maximality of $t(v)$ and as $\mathcal{N}$ is temporal, no hybridization vertex is a descendant of v, u, or u'. Let $\mathcal{L}_k$ denote the subset of X that contains precisely the descendants of v, and let $X' = X - \mathcal{L}_k$. Let $\mathcal{N}'$ be the temporal hybridization network on X' that is obtained from $\mathcal{N}$ by deleting the arcs (u, v) and (u', v), suppressing the resulting degree-two vertices, and ignoring the component, $\mathcal{T}_k$ say, containing v. Since $\mathcal{N}$ displays $\mathcal{T}$ and $\mathcal{T}'$ and since $\mathcal{T}_k$ is a pendant subtree of $\mathcal{T}$ and $\mathcal{T}'$, it follows by the maximality of $t(v)$ that $\mathcal{N}'$ is a temporal hybridization network that displays $\mathcal{T}|X'$ and $\mathcal{T}'|X'$. Therefore, as $h_t(\mathcal{N}') < h_t(\mathcal{N})$, it follows by the induction assumption that the forest $\mathcal{F}' = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{k-1}\}$ induced by $\mathcal{N}'$ is a temporal-agreement forest for $\mathcal{T}|X'$ and $\mathcal{T}'|X'$. Labeling v, the root vertex of $\mathcal{T}_k$ appropriately, let $\mathcal{F} = \mathcal{F}' \cup \{\mathcal{T}_k\}$. Note that $\mathcal{F}$ is the forest induced by $\mathcal{N}$. Clearly, $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. We complete the proof by showing that $\mathcal{F}$ is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

Let (w_1, u) and (u, w_2), where $w_2 \neq v$, and (w_1', u') and (u', w_2'), where $w_2' \neq v$, be arcs in $\mathcal{N}$. By the observations in the previous paragraph, both of the arcs (w_1, w_2) and (w_1', w_2') must be used in displaying each of $\mathcal{T}|X'$ and $\mathcal{T}'|X'$ in $\mathcal{N}'$ (with u and u' being suppressed). Thus, in displaying each of $\mathcal{T}$ and $\mathcal{T}'$ in $\mathcal{N}$ all of the arcs (w_1, u), (u, w_2), (w_1', u'), and (u', w_2'), and exactly one of (u, v) and (u', v), are used. Note that it is possible in displaying $\mathcal{T}$ and $\mathcal{T}'$ that only one of the arcs (u, v) and (u', v) is used. It now follows that

the temporal labeling of $\mathcal{N}$ induces a temporal labeling of $\mathcal{T}_k$ so that $\mathcal{F}$ is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. This completes the proof of the proposition. □

Proposition 3.1 provides one direction of the characterization (Theorem 3.3). The other direction is given by the next proposition. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two temporal-agreement forests for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, let t_1 and t_2 be the temporal labelings of $\mathcal{T}$ induced by the temporal embeddings θ_1 and θ_2 of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively, in $\mathcal{T}$. Similarly, let t'_1 and t'_2 be the temporal labelings of $\mathcal{T}'$ induced by the temporal embeddings θ'_1 and θ'_2 of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively, in $\mathcal{T}'$. Then $\mathcal{F}_2$ is a *refinement* of $\mathcal{F}_1$ precisely if $t_1 = t_2$, $t'_1 = t'_2$, and for each $\mathcal{T}_i \in \mathcal{F}_2$ there exists a $\mathcal{T}_j \in \mathcal{F}_1$ such that $\mathcal{L}(\mathcal{T}_i) \subseteq \mathcal{L}(\mathcal{T}_j)$.

Proposition 3.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and suppose that $\mathcal{F}$ is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then there exists a temporal hybridization network $\mathcal{N}$ on X that displays $\mathcal{T}$ and $\mathcal{T}'$ such that $\mathcal{F}$ is a refinement of the forest induced by $\mathcal{N}$.*

Proof. Let $\mathcal{F} = \{\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. The proof is by induction on k. If $k = 0$, then, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical, and so $\mathcal{T}_0$ is temporal hybridization network with the desired properties. Now suppose that the proposition holds for all pairs of rooted binary phylogenetic trees on the same label sets with a temporal-agreement forest of size at most k.

Without loss of generality, we may assume that, amongst all temporal labelings of the root vertices of the trees in $\mathcal{F}$, the temporal labeling of the root vertex of $\mathcal{T}_k$ is maximized. Thus $\mathcal{T}_k$ is a pendant subtree of $\mathcal{T}$ and $\mathcal{T}'$. Let $X' = X - \mathcal{L}(\mathcal{T}_k)$, and let $\mathcal{F}' = \mathcal{F} - \{\mathcal{T}_k\}$. Then it is easily seen that $\mathcal{F}'$ is a temporal-agreement forest for $\mathcal{T}|X'$ and $\mathcal{T}'|X'$. Therefore, by the induction assumption, there is a temporal hybridization network $\mathcal{N}'$ that displays $\mathcal{T}|X'$ and $\mathcal{T}'|X'$ such that $\mathcal{F}'$ is a refinement of the forest induced by $\mathcal{N}'$. We next construct a network on X with the desired properties.

Let $\mathcal{N}$ be the network on X that is obtained from $\mathcal{N}'$ as follows. Since $\mathcal{T}_k$ is a pendant subtree of $\mathcal{T}$, we can obtain $\mathcal{T}$ from $\mathcal{T}|X'$ by subdividing an

(unique) edge, (u_k, v_k) say, of $\mathcal{T}|X'$ and identifying the root of $\mathcal{T}_k$, labeled ρ_k, with the new vertex. The edge (u_k, v_k) corresponds to an edge of a tree in $\mathcal{F}'$ in which, under the temporal labeling of $\mathcal{T}|X'$ induced by $\mathcal{F}'$, the end vertices are assigned temporal labelings $t(u_k)$ and $t(v_k)$, respectively, where t is the temporal map of $\mathcal{T}$ that is induced by $\mathcal{F}$. Since $\mathcal{N}'$ displays $\mathcal{T}|X'$ and $\mathcal{F}'$ is a refinement of the forest induced by $\mathcal{N}'$, there is a tree arc of $\mathcal{N}'$ whose end vertices are assigned temporal labelings $t(u_k)$ and $t(v_k)$. Subdividing this tree arc and adding a new arc joining the root ρ_k of $\mathcal{T}_k$ with the new vertex assigned temporal labeling $t_k(\rho_k)$, where t_k is the temporal labeling of $\mathcal{T}_k$ in $\mathcal{F}$, the resulting network is a temporal hybridization network on X that displays $\mathcal{T}$. Repeating this construction for $\mathcal{T}'$, let $\mathcal{N}$ denote the resulting temporal hybridization network on X. Note that, as the forest induced by $\mathcal{N}'$ is a refinement of $\mathcal{F}'$, all descendant arcs of the two vertices created by the two subdivisions are tree arcs. By construction, the forest induced by $\mathcal{N}$ is a refinement of $\mathcal{F}$. This completes the proof of the proposition. □

The next theorem combines Propositions 3.1 and 3.2 and completes the characterization.

Theorem 3.3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then there exists a temporal hybridization network on X that displays $\mathcal{T}$ and $\mathcal{T}'$ if and only if there exists a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, in which case,*

$$h_t(\mathcal{T}, \mathcal{T}') = m_t(\mathcal{T}, \mathcal{T}').$$

Proof. If there exists a temporal hybridization network on X that displays $\mathcal{T}$ and $\mathcal{T}'$, then, by Proposition 3.1, there is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, by Proposition 3.1, $h_t(\mathcal{T}, \mathcal{T}') \geq m_t(\mathcal{T}, \mathcal{T}')$. On the other hand, if there is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, then, by Proposition 3.2, there is a temporal hybridization network on X that displays $\mathcal{T}$ and $\mathcal{T}'$. Moreover, it follows by Proposition 3.2 that $h_t(\mathcal{T}, \mathcal{T}') \leq m_t(\mathcal{T}, \mathcal{T}')$. In particular,

$$h_t(\mathcal{T}, \mathcal{T}') = m_t(\mathcal{T}, \mathcal{T}').$$

□

4. Minimum Temporal Hybridization is APX-Hard

In this section, we show that the minimization problem Minimum Temporal Hybridization is APX-hard and thus also NP-hard. In particular, this means that, unless P=NP, there is some fixed constant c strictly bigger than 1 for which there is no polynomial-time approximation algorithm such that, for each instance, the output of a feasible solution is at most c times the size of its optimal solution.

Minimum Temporal Hybridization
Instance: A finite set X, and two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$.
Goal: Find a temporal hybridization network $\mathcal{N}$ on X that displays $\mathcal{T}$ and $\mathcal{T}'$ with minimum hybridization number if such a network exists.
Measure: The value of $h_t(\mathcal{N})$.

We obtain the above hardness result in two steps. For the first step, we show that the following optimization problem is APX-hard using an L-reduction from a restricted version of Maximum 4-Dimensional Matching. Briefly, L-reductions were introduced by Papadimitriou and Yannakakis [16] and preserve approximability. They play a similar role in the study of approximability of optimization problems as polynomial-time reductions in the study of decision problems and their complexity. For more details, we refer the interested reader to [15, 16]. Having established an L-reduction to Minimum Temporal Hybridization, the second step is an almost immediate consequence of Theorem 3.3.

Maximum-Temporal-Agreement Forest
Instance: A finite set X, and two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$.
Goal: Find a maximum-temporal-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ if such a forest exists.
Measure: The number of components in $\mathcal{F}$ minus one.

Maximum 4-Dimensional Matching (Max-4DM)
Instance: 4 disjoint sets W, X, Y, Z, and a subset Q of $W \times X \times Y \times Z$.
Goal: Find a maximum-sized subset M of Q with the property that no members of M agree in any coordinate.

Measure: The cardinality of M.

Chlebík and Chlebíková [7] proved an explicit inapproximability ratio for the restricted version of MAX-4DM when each element of $W \cup X \cup Y \cup Z$ appears in exactly two 4-tuples in Q. Denoting this restriction by MAX-4DM-2, we will show that there is an L-reduction from MAX-4DM-2 to MAXIMUM-TEMPORAL-AGREEMENT FOREST.

The construction and key lemma, Lemma 4.1, used to established this L-reduction is similar to that used by Bordewich and Semple [4] to show that the non-temporal version of MINIMUM TEMPORAL HYBRIDIZATION is APX-hard. However, the temporal constraints mean that modifications and additions are required for the construction. For this reason, we have included the details of the construction and lemma in full, but the details of the remaining part of the L-reduction have been suppressed.

Let W, X, Y, Z and $Q \subseteq W \times X \times Y \times Z$ be an instance I of Max-4DM-2. Let $|W| = p$. Since each element of $W \cup X \cup Y \cup Z$ appears in exactly two members of Q, we have

$$p = |W| = |X| = |Y| = |Z| = |Q|/2.$$

We next construct an instance of MAXIMUM-TEMPORAL-AGREEMENT FOREST based on I.

Let

$$Q = \{(w_1, x_1, y_1, z_1), (w_2, x_2, y_2, z_2), \ldots, (w_{2p}, x_{2p}, y_{2p}, z_{2p})\}.$$

Let $\mathcal{T}$ and $\mathcal{T}'$ be the two rooted binary phylogenetic X-trees shown in Figure 6 and Figure 7, respectively, where, for the moment, we ignore the (temporal) labelings assigned to the interior vertices. In Figure 6, each subtree A_i of $\mathcal{T}$, where $i \in \{1, 2, \ldots, 2p\}$, corresponds to exactly one tuple in Q. In Figure 7, each subtree B_r of $\mathcal{T}'$, where $r \in \{1, 2, \ldots, 4p\}$, corresponds to exactly one element r in $W \cup X \cup Y \cup Z$, while each subtree C_i of $\mathcal{T}'$, where $i \in \{1, 2, \ldots, 2p\}$, corresponds to exactly one tuple in Q. Note that the edge incident with the pendant root in A_i is not part of the subtree A_i (and is only shown in Figure 6 for convenience) but in fact the edge joining A_i to the rest of $\mathcal{T}$. Analogous comments hold for the subtrees B_r and C_i that are shown in Figures 7, 8, and 9. Furthermore, for each B_r, the indices i and

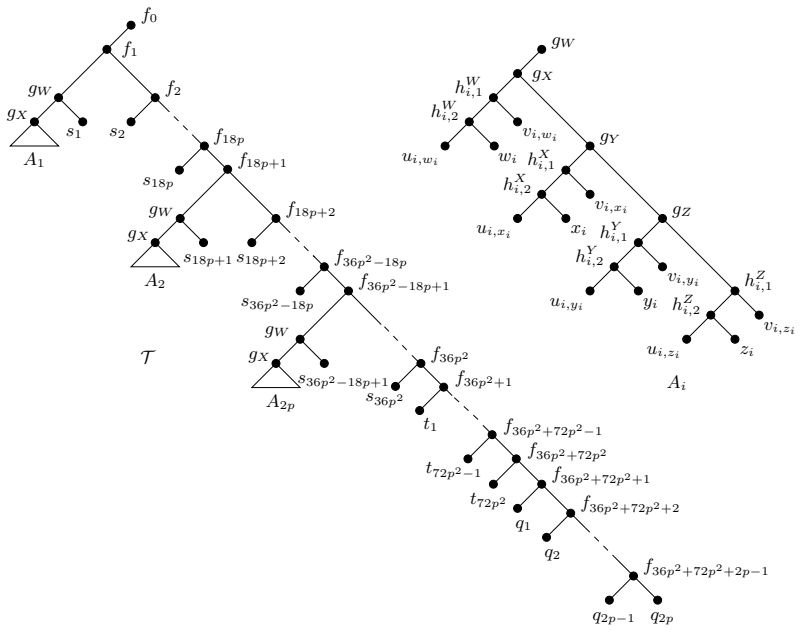

Figure 6: The tree $\mathcal{T}$ and its subtrees A_i that are used in the proof of Lemma 4.1.

j are used to identify the two tuples in Q in which r occurs. Throughout the statement and proof of Lemma 4.1, $\mathcal{T}$ and $\mathcal{T}'$ refer to the rooted binary phylogenetic trees shown in Figures 6 and 7.

Before stating and proving the key lemma, Lemma 4.1, we describe an equivalent way of showing that two rooted binary phylogenetic X-trees have a temporal-agreement forest. This equivalence is used in the proof of Lemma 4.1. Let $\mathcal{T}$ be a planted phylogenetic X-tree with root ρ_0, and let t be a temporal labeling of $\mathcal{T}$. Let $\mathcal{S}$ be a pendant subtree of $\mathcal{T}$, and let $e = (u_1, u_2)$ be the edge of $\mathcal{T}$ attaching $\mathcal{S}$ to $\mathcal{T}$. The *edge separation of* e is performed by deleting e, contracting the resulting non-root degree-two vertex, adjoining a pendant root to u_2, and assigning the new root the temporal label $t(u_1)$. Now let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Viewing $\mathcal{T}$ and $\mathcal{T}'$ as planted with a root vertex ρ_0, it is easily seen that $\mathcal{T}$ and $\mathcal{T}'$ have a temporal-agreement forest of size $k+1$ if and only if there are temporal labelings t and t' of $\mathcal{T}$ and $\mathcal{T}'$, respectively such that there is a sequence of k edge separations in each of $\mathcal{T}$ and $\mathcal{T}'$ that result in the same forest.

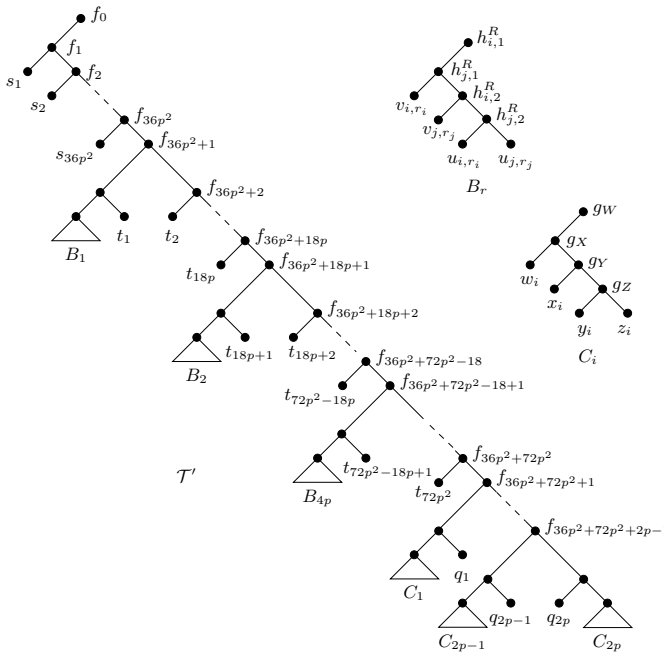

Figure 7: The tree $\mathcal{T}'$ and its subtrees B_r and C_i that are used in the proof of Lemma 4.1.

Lemma 4.1. *The set Q contains a 4-dimensional matching of size k if and only if there is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of size*

$$1 + 8k + 9(2p - k) = 18p - k + 1.$$

In particular, $m_t(\mathcal{T}, \mathcal{T}') = 18p - \mathrm{opt}(Q)$, where $\mathrm{opt}(Q)$ denotes the size of an optimal 4-dimensional matching of Q.

Proof. First suppose that Q contains a 4-dimensional matching M of size k. To show that $\mathcal{T}$ and $\mathcal{T}'$ have a temporal-agreement forest $\mathcal{F}$ of size $18p - k + 1$, we first give a temporal labeling of $\mathcal{T}$ and $\mathcal{T}'$, and then specify the edge separations in both $\mathcal{T}$ and $\mathcal{T}'$ that result in a temporal-agreement forest of this size. Observe that it suffices to assign temporal labels to only the interior vertices of $\mathcal{T}$ and $\mathcal{T}'$.

Let $t : \mathring{V}(\mathcal{T}) \to \mathbb{R}^+$ be the map that assigns positive real values to the

interior vertices of $\mathcal{T}$ as shown in Figure 6, where

$$f_0 < f_1 < \cdots < f_{36p^2+72p^2+2p-1} < g_W < g_X < g_Y < g_Z \\ < h^W_{i,1} < h^W_{j,1} < h^W_{i,2} < h^W_{j,2} < h^X_{i,1} < h^X_{j,1} < \cdots < h^Z_{i,2} < h^Z_{j,2}.$$

For clarification, consider an element in W. This element appears in exactly two 4-tuples, corresponding to A_i and A_j say. For A_i, we assign the temporal labelings $h^W_{i,1}$ and $h^W_{i,2}$ and, for A_j, we assign the temporal labelings $h^W_{j,1}$ and $h^W_{j,2}$ appropriately. This extends analogously to all elements in $W \cup X \cup Y \cup Z$. Clearly, t is a temporal labeling of $\mathcal{T}$. Now let $t' : \mathring{V}(\mathcal{T}') \to \mathbb{R}^+$ be the map that assigns positive real values to the interior vertices v of $\mathcal{T}'$ as follows. If a vertex v in $\mathring{V}(\mathcal{T}')$ is on the path from the root of $\mathcal{T}'$ to the vertex labeled $f_{36p^2+72p^2+2p-1}$, then assign v the value as shown in Figure 7. For the interior vertices that are part of the subtrees B_r and C_i and their respective parents, their assigned values depend upon whether or not the element r is in a tuple in M and whether or not the tuple corresponding to C_i is in M. In particular, we make the following assignments, where R in the superscript of the temporal labels assigned to the subtrees B_r and their parent vertices equates to W, X, Y, or Z depending upon whether r is an element of W, X, Y, or Z, respectively:

(i) If neither r_i nor r_j is in a tuple in M, then the interior vertices of B_r and its parent are assigned values as shown in Figure 7.

(ii) If r_i is in a tuple in M and so r_j is not in a tuple in M, then the interior vertices of B_r and its parent are assigned values as shown in Figure 8(a).

(iii) If r_j is in a tuple in M and so r_i is not in a tuple in M, then the interior vertices of B_r and its parent are assigned values as shown in Figure 8(b).

(iv) If M does not contain the tuple corresponding to C_i, then the interior vertices of C_i and its parent are assigned values as shown in Figure 7.

(v) Lastly, if M contains the tuple corresponding to C_i, then the interior vertices of C_i and its parent are assigned values as shown in Figure 9.

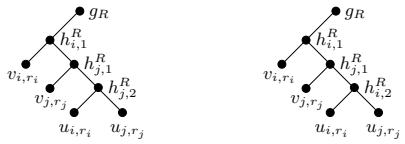

Figure 8: The temporal labeling of the subtree B_r if (a) r_i is contained in a tuple of M and (b) r_j is contained in a tuple of M.

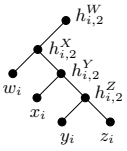

Figure 9: The temporal labeling of the subtree C_i for when M contains the 4-tuple corresponding to C_i.

It is easily checked that t' is a temporal labeling of $\mathcal{T}'$.

We next specify a set of edge separations for $\mathcal{T}$ and $\mathcal{T}'$. For $\mathcal{T}$, we make the following edge separations:

(I) For each i, separate the edge attaching A_i to the rest of $\mathcal{T}$.

(II) For each i in which A_i corresponds to a tuple in M, separate each of the pendant edges attaching w_i, x_i, y_i, and z_i, and then separate, in turn, the edges attaching the subtrees containing u_{i,z_i} and v_{i,z_i}, u_{i,y_i} and v_{i,y_i}, and u_{i,x_i} and v_{i,x_i}. In this case, A_i is broken into 8 components.

(III) For each i in which A_i corresponds to a tuple not in M, separate each of the pendant edges attaching $u_{i,w_i}, v_{i,w_i}, u_{i,x_i}, v_{i,x_i}, u_{i,y_i}, v_{i,y_i}, u_{i,z_i}, v_{i,z_i}$. In this case, A_i is broken into 9 components.

Altogether, this process breaks $\mathcal{T}$ into $1 + 8k + 9(2p - k) = 18p - k + 1$ components. For $\mathcal{T}'$, we make the following edge separations:

(I′) For each r and i, separate the edge attaching B_r and C_i to the rest of $\mathcal{T}'$.

(II′) For each $r \in W \cup X \cup Y \cup Z$, if neither r_i nor r_j is in a tuple in M, then separate, in turn, the pendant edges attaching u_{j,r_j}, u_{i,r_i}, and v_{j,r_j}, so that B_r is broken into 4 components. If r_i is in a tuple in M, then separate the pendant edges attaching u_{j,r_j} and v_{j,r_j}, so that B_r is broken into 3 components. If r_j is in a tuple in M, then separate the pendant edges attaching u_{i,r_i} and v_{i,r_i}, so that B_r is broken into 3 components.

(III′) For each i, if the tuple corresponding to C_i is not in M, then it remains as 1 component. If the tuple corresponding to C_i is in M, then separate the pendant edges attaching z_i, y_i, and x_i, so that C_i is broken into 4 components.

Collectively, this process breaks $\mathcal{T}'$ into $1+4(4p-4k)+3\cdot 4k+(2p-k)+4k = 18p-k+1$. Thus the two processes break $\mathcal{T}$ and $\mathcal{T}'$ into the same number of components. Furthermore, a routine check shows that the two sets of components are identical. It now follows that the resulting set of components is a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$.

For the converse, let $\mathcal{F}$ be a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of size at most $18p+1$. Note that the temporal labelings assigned to the interior vertices of $\mathcal{T}$ and $\mathcal{T}'$ play no role in this part of the proof. Let

$$S = \{s_1, s_2, \ldots, s_{36p^2}, t_1, t_2, \ldots, t_{72p^2}, q_1, q_2, \ldots, q_{2p}\}.$$

We first show that if $\mathcal{T}_j \in \mathcal{F}$ and, for some i, the intersection $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}(A_i)$ is non-empty, then $\mathcal{L}(\mathcal{T}_j) \subseteq \mathcal{L}(A_i)$. Furthermore, we also show that if $\mathcal{T}_j \in \mathcal{F}$ and, for some r, the intersection $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}(B_r)$ is non-empty, then $\mathcal{L}(\mathcal{T}_j) \subseteq \mathcal{L}(B_r)$.

Suppose that $\mathcal{T}_j \in \mathcal{F}$ and $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}(A_i)$ is non-empty. Let $\ell \in \mathcal{L}(\mathcal{T}_j)$ such that $\ell \notin \mathcal{L}(A_i)$. Assume $\ell \in \mathcal{L}(A_{i'})$ for some $i' \neq i$. Then, as $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, at least $18p$ members of $\{s_1, s_2, \ldots, s_{36p^2-18p}\}$ appear as planted singletons in $\mathcal{F}$. By property (iii) in the definition of an agreement forest, no label set of a component of $\mathcal{F}$ contains $\mathcal{L}(A_i)$, and so $\mathcal{F}$ has at least $18p+2$ components; a contradiction. Now assume that $\ell \notin \mathcal{L}(A_{i'})$. Thus $\ell \in S$. If $\ell \in S - \{s_1, s_2, \ldots, s_{36p^2}\}$, then either each of the $18p$ elements in $\{s_{36p^2-18p+1}, \ldots, s_{36p^2}\}$ if $i = 2p$ or each of the $18p$ elements in $\{s_{36p^2-18p}, s_{36p^2-18p+2}, \ldots, s_{36p^2}\}$ if $i \neq 2p$ appear as planted singletons in

$\mathcal{F}$. Since no label set of a component in $\mathcal{F}$ contains $\mathcal{L}(A_i)$, this implies that $\mathcal{F}$ contains at least $18p+2$ components; a contradiction. Thus we may assume that $\ell \in \{s_1, s_2, \ldots, s_{36p^2}\}$, in which case, it is easily checked that at least $18p-1$ members of $\{s_{36p^2-18p+1}, \ldots, s_{36p^2}\}$ appear as planted singletons in $\mathcal{F}$. However, $\mathcal{L}(A_i)$ is not contained in the label set of a single component and, by property (iii) in the definition of an agreement forest, the label set of no component other than $\mathcal{T}_j$ has a non-empty intersection with $\mathcal{L}(A_i)$ and a non-empty intersection with $S - \{s_1, s_2, \ldots, s_{36p^2}\}$. Thus $\mathcal{F}$ contains at least $18p+2$ components. This last contradiction now implies that $\mathcal{L}(\mathcal{T}_j) \subseteq \mathcal{L}(A_i)$. With this in hand, it is now easily seen that if $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}(B_r)$ is non-empty for some r, then $\mathcal{L}(\mathcal{T}_j) \subseteq \mathcal{L}(B_r)$.

Now suppose that $\mathcal{F}$ is a temporal-agreement forest of size $1+8k+9(2p-k) = 18p-k+1$. By the above argument, for each i and r, there is a subset of components of $\mathcal{F}$ for which the union of the label sets equates to $\mathcal{L}(A_i)$ and $\mathcal{L}(B_r)$, respectively. It is now easily seen that, for each i, the number of components of $\mathcal{F}$ in such a subset for which the union of the label sets equates to $\mathcal{L}(A_i)$ is at least 8. Moreover, it is precisely 8 if the partition of $\mathcal{L}(A_i)$ induced by these label sets is

$$\big\{\{w_i\}, \{x_i\}, \{y_i\}, \{z_i\}, \{u_{i,w_i}, v_{i,w_i}\}, \{u_{i,x_i}, v_{i,x_i}\}, \{u_{i,y_i}, v_{i,y_i}\}, \{u_{i,z_i}, v_{i,z_i}\}\big\}.$$

As $\mathcal{F}$ has size $1 + 8k + 9(2p-k)$, it follows that at least k of the A_is are 'partitioned' in this way. Let A_i and A_j be two such subtrees in $\mathcal{T}$, and let (w_i, x_i, y_i, z_i) and (w_j, x_j, y_j, z_j) be the two 4-tuples associated with A_i and A_j. Assume that the first coordinates agree, that is $w_i = w_j$. The label sets of two components of $\mathcal{F}$ are $\{u_{i,w_i}, v_{i,w_i}\}$ and $\{u_{j,w_j}, v_{j,w_j}\}$. However, as $w_i = w_j$, this implies that the third condition of an agreement forest is violated in the corresponding subtree B_r; a contradiction. Hence the first coordinate does not agree and, similarly, the other coordinates do not agree. We deduce that Q has a 4-dimensional matching of size at least k, thereby completing the proof of the lemma. □

Despite the modifications and additions in the above construction, the relationship between the size of a 4-dimensional matching in Q and the size of a temporal-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ as described in Lemma 4.1 is precisely the same as that for the key lemma in [4] for showing that the non-temporal version of Minimum Temporal Hybridization is APX-hard, where the type of agreement forest involved is more general. As a

consequence, the proofs of the remaining results in this section are essentially identical to the analogous results in [4], and are thus omitted.

Theorem 4.2. *The optimization problem* Maximum-Temporal-Agreement Forest *is APX-hard. In particular, unless* P=NP*, there is no polynomial-time approximation scheme for* Maximum-Temporal-Agreement Forest.

Chlebík and Chlebíková [7] showed that, unless P=NP, there is no polynomial-time approximation algorithm for Max-4DM-2 with an approximation ratio better than $\frac{48}{47}$. Using this result, one can establish the next corollary.

Corollary 4.3. *Unless* P=NP*, there is no polynomial-time approximation algorithm for* Maximum-Temporal-Agreement Forest *with an approximation ratio better than* $\frac{2113}{2112}$.

It immediately follows from Theorem 3.3 that there is no polynomial-time approximation algorithm with ratio c for Maximum-Temporal-Agreement Forest if and only if there is no polynomial-time approximation algorithm with ratio c for Minimum Temporal Hybridization. Combining Theorem 4.2 and Corollary 4.3, we have the following result.

Corollary 4.4. *The optimization problem* Minimum Temporal Hybridization *is APX-hard. In particular, unless* P=NP*, there is no polynomial-time approximation algorithm for* Minimum Temporal Hybridization *with an approximation ratio better than* $\frac{2113}{2112}$.

Acknowledgements. S.L. thanks the New Zealand Marsden Fund and the German Academic Exchange Service (DAAD) for financial support. C.S. was supported by the New Zealand Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution.

References

[1] B. Albrecht, C. Scornavacca, A. Cenci, D. H. Huson, Fast computation of minimum hybridization networks, Bioinformatics 28 (2012) 191–197.

[2] M. Baroni, S. Grünewald, V. Moulton, C. Semple, Bounding the number of hybridization events for a consistent evolutionary history, J. Math. Biol. 51 (2005) 171–182.

[3] M. Baroni, C. Semple, and M. Steel. Hybrids in real time, Sys. Biol. 44 (2006) 46–56.

[4] M. Bordewich, C. Semple, Computing the minimum number of hybridization events for a consistent evolutionary history, Discrete Appl. Math. 155 (2007) 914–928.

[5] G. Cardona, F. Rossello, G. Valiente, Comparison of tree-child phylogenetic networks, IEEE Trans. Comput. Biol. Bioinf. 6 (2009) 552–569.

[6] Z. Z. Chen, L. Wang, HybridNET: a tool for constructing hybridization networks, Bioinformatics 26 (2010) 2912–2913.

[7] M. Chlebík, J. Chlebíková. Inapproximability results for bounded variants of optimization problems, in: A. Lingas, B.J. Nilsson (Eds.), Fundamentals of Computation Theory, 14th International Symposium (FCT), Lecture Notes in Computer Science, vol. 2751, Springer-Verlag, 2003, pp. 27–38.

[8] J. Collins, S. Linz, and C. Semple, Quantifying hybridization in realistic time, J. Comp. Biol. 18 (2011) 1305–1318 .

[9] L. van Iersel, S. Kelk, When two trees go to war, J. Theo. Biol. 269 (2011) 245–255.

[10] L. van Iersel, S. Kelk, Constructing the simplest possible phylogenetic network from triplets, Algorithmica 60 (2011) 207–235.

[11] L. van Iersel, S. Kelk, R. Rupp, D. H. Huson, Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters, Bioinformatics 26 (2010) i124–i131.

[12] S. Linz, C. Semple, T. Stadler, Analyzing and reconstructing reticulation networks under timing constraints, J. Math. Biol. 61 (2010) 715–735.

[13] W. Maddison, Gene trees in species trees, Sys. Biol. 46 (1997) 523–536.

[14] B. M. E. Moret, L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme, Phylogenetic networks: modeling, reconstructibility, and accuracy, IEEE Trans. Comput. Biol Bioinf. 1 (2004) 13–23.

[15] C. H. Papadimitriou, Computational Complexity, Addison Wesley, 1993.

[16] C. H. Papadimitriou, M. Yannakakis, Optimization, approximation, and complexity classes, J. Comput. System Sci. 43 (1991)425–440.

[17] C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.