# HYBRIDS IN REAL TIME

MIHAELA BARONI, CHARLES SEMPLE, MIKE STEEL

Abstract. We describe some new and recent results that allow for the analysis and representation of reticulate evolution by non-tree networks. In particular, we (1) present a simple result to show that, despite the presence of reticulation, there is always a well-defined underlying tree which corresponds to those parts of life that do not have a history of reticulation, (2) describe and apply new theory for determining the smallest number of hybridization events required to explain conflicting gene trees, and (3) present a new algorithm to determine whether an arbitrary rooted network can be realized by contemporaneous reticulation events. We illustrate these results with examples.

Address: Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

Email: mbaroni@ugal.ro; c.semple@math.canterbury.ac.nz; m.steel@math.canterbury.ac.nz

Corresponding author: Mike Steel

*Date*: 7 July 2005.

2000 *Mathematics Subject Classification.* 05C05, 92D15.

*Key words and phrases.* Directed acyclic graph, reticulate evolution, hybrid species, subtree prune and regraft.

We thank the New Zealand Marsden Fund (UOC-MIS-005) and the Allan Wilson Centre for Molecular Ecology and Evolution for supporting this research.

## Introduction

Evolutionary relationships are generally represented by non-reticulating trees, and for certain groups of taxa (e.g. mammals) this model seems well suited. However, for other groups (for example, plants, some fish, and bacteria), processes of reticulate evolution such as the formation of hybrid species, horizontal gene transfer, and other mechanisms (for example, endosymbiosis) suggest that evolutionary history would be better described by a network that is more complex than a tree, with some species arising from the genetic contribution of two (rather than one) ancestral lineages.

Although processes of reticulate evolution have long been recognized in biology, techniques for representing and analyzing reticulate evolution have tended to be fairly ad-hoc. For example, one might first build a tree and then heuristically add some additional edges if these improve the fit of the data (as in Legendre and Makarenkov, 2002). In the last few years there has been much new theoretical work by computer scientists and mathematicians (e.g., Baroni, 2004; Baroni et al., 2004; Gusfield, 2004; Gusfield et al., 2004; Holland et al., 2004; Huson et al., 2004; Huson et al., 2005; Moret et al., 2004; Song and Hein, 2004) with the aim of providing more rigorous approaches to the representation and analysis of reticulate evolution.

In the the third and fourth sections, we provide a brief overview of some of our recent work, and show how it can be applied to set lower bounds on the degree of reticulation required to explain two conflicting phylogenetic trees. We illustrate the application of these results on two trees that describe the evolution of alpine Ranunculi in New Zealand. In the fifth section, we present a fast algorithm that determines whether or not a hybrid phylogeny can be realized by hybridization events between species that existed at the same time—an obvious biological requirement, though one that is often overlooked in a formal mathematical representation. The last section contains some concluding remarks.

## Hybrid Phylogenies

In this section we introduce some terminology that is useful for describing and studying hybrid evolution. Informally, a 'hybrid phylogeny' is simply a rooted network in which each arc (directed edge) leads from an ancestral taxon to its immediate descendants. However, unlike a rooted phylogenetic tree, we allow for some (ancestral or extant) taxa to have two (or more) incoming arcs. In other words, we regard those taxa as being hybrids, consisting of a genetic composition from both (or all) of the incoming arcs. In this section, we formalize these notions in order to obtain precise results. Furthermore, we describe a tree that underlies any hybrid phylogeny, and provide some background and motivation for the rest of the paper. Throughout, the notation and terminology mostly follows Baroni (2004) and Baroni *et al.* (2004).

First we recall some graph-theoretic terminology. *Directed graphs* (also known as *digraphs*) are used in evolutionary biology to represent the evolutionary history of extant species. Usually, this representation takes the form of a rooted phylogenetic tree. However, in this paper we are mostly interested in representations called (rooted) *hybrid phylogenies.* A directed graph consists of a collection of nodes, and a collection of directed edges called *arcs* with each arc joining two nodes. Nodes typically represent species, individuals, or DNA sequences, while arcs represent relationships of ancestry. Thus if $u$ is the "parent" of $v$, then we denote this relationship with the arc $(u, v)$. The first node indicates where the arc is coming from and the second node indicates where the arc is going to, thus $(u, v) \neq (v, u)$.

The *degree* of a node $v$ is the number of arcs incident with $v$. In directed graphs, we often distinguish between arcs coming out of a node and those coming into a node. In particular, the *outdegree* of $v$ is the number of arcs whose first component is $v$ and is denoted $d^+(v)$. The *indegree* of $v$ is the number of arcs whose second component is $v$ and is denoted $d^-(v)$. In rooted phylogenies and hybrid phylogenies, the outdegree of a node $v$ is the number of "children" of $v$, while the indegree of $v$ is the number of "parents" of $v$.

A *directed path* in a digraph is an alternating sequence

$$v_0, a_1, v_1, a_2, v_2, \ldots, v_{k-1}, a_k, v_k$$

of nodes and arcs in which $a_i$ is an arc from $v_{i-1}$ to $v_i$ for all $i$, and no node or arc appears more than once. Essentially, a path describes one way in which we can get from one node to another following the direction of the arcs. A *directed cycle* of a digraph is directed path in which the first and last nodes are equal. A digraph is *acyclic* if it has no directed cycles.

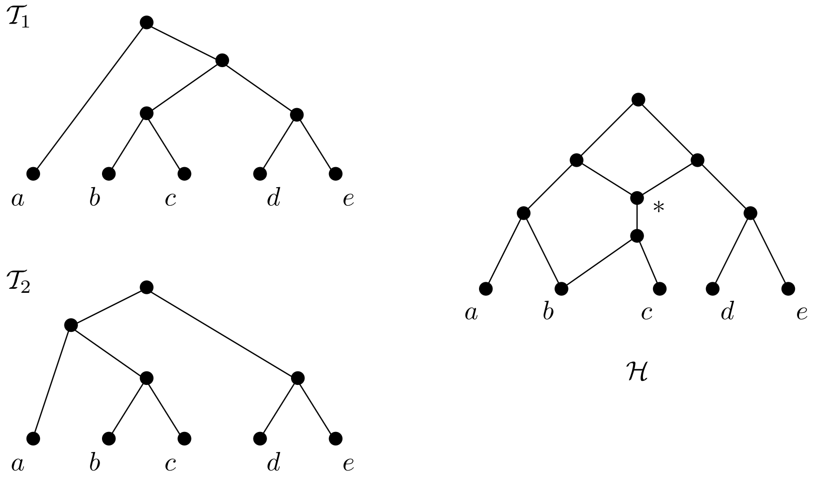


Figure 1. A hybrid phylogeny $\mathcal{H}$, and two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ displayed by $\mathcal{H}$.

An acyclic digraph $D$ with no underlying parallel edges (that is, no pair of arcs joining the same two nodes) is *rooted* if there is a distinguished node $\rho$, called the *root*, with the properties that $d^-(\rho) = 0$ and there is a directed path from $\rho$ to every node of $D$.

If $D$ is a rooted digraph, then a *rooted subtree* of $D$ is any rooted tree that is obtained from $D$ by deleting nodes (and any arcs incident with these nodes) and arcs.

We now formally describe rooted phylogenetic trees and hybrid phylogenies. Throughout these definitions, and indeed throughout this paper, $X$ will always denote a set of extant species. A *rooted phylogenetic tree* $\mathcal{T}$ on $X$ is a rooted tree with no nodes that have both indegree one and outdegree one, whose leaf set is $X$, and whose root has outdegree at least two. In addition, $\mathcal{T}$ is *binary* or *fully-resolved* if all interior nodes have outdegree two. We sometimes refer to $X$ as the *label* set of $\mathcal{T}$ and denote it has $\mathcal{L}(\mathcal{T})$. Indeed, for a collection $\mathcal{P}$ of rooted phylogenetic trees, we denote the union of the label sets of the trees in $\mathcal{P}$ by $\mathcal{L}(\mathcal{P})$. Two rooted binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ are shown in Fig. 1.

A *hybrid phylogeny* $\mathcal{H}$ on $X$ is a rooted acyclic digraph in which

(i) $X$ is the set of nodes of outdegree zero,
(ii) the root has outdegree at least two, and

(iii) for all nodes $v$ with $d^+(v) = 1$, we have $d^-(v) \geq 2$.

Nodes of indegree at least two (called *hybridization nodes*) represent hybridization events. These correspond to an exchange of genetic information between hypothetical ancestors by processes such as horizontal gene transfer, gene fusion etc. To illustrate, a hybrid phylogeny $\mathcal{H}$ on $X = \{a, b, c, d, e\}$ is shown in Fig 1, where the root is the topmost node. The node $*$ as well as the node labelled $b$ are hybridization nodes. Here and in all other figures, it is implicit that arcs are directed downwards. Observe that a rooted phylogenetic tree on $X$ is a particular type of hybrid phylogeny (one that contains no hybridization nodes).

Let $\mathcal{T}$ be a rooted phylogenetic tree on $X$ and let $\mathcal{H}$ be a hybrid phylogeny on $X'$, where $X \subseteq X'$. Then $\mathcal{H}$ *displays* $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{H}$ by deleting nodes and edges, and by replacing nodes of indegree one and outdegree one and their incident edges with a single edge (that is, *suppressing nodes of indegree one and outdegree one*). Extending this to a collection $\mathcal{P}$ of rooted phylogenetic trees, we say that $\mathcal{H}$ displays $\mathcal{P}$ if $\mathcal{H}$ displays every tree in $\mathcal{P}$. For example, in Fig 1, the hybrid phylogeny $\mathcal{H}$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$. Biologically speaking, saying that $\mathcal{H}$ displays $\mathcal{T}$ means that a gene tree with the topology described by $\mathcal{T}$ could arise from an evolutionary history depicted by $\mathcal{H}$ without requiring the action of other processes such as lineage sorting.

The concept of display can be generalized to allow refinement of non-binary trees, however, we do not require this in this paper.

**An underlying tree for a hybrid phylogeny.** Processes of reticulate evolution such as the evolution of hybrid species seem to call into question the very existence of any meaningful concept of a tree of life. However, we now describe a simple mathematical result that formalizes how there is always an underlying tree corresponding to those parts of life that do not have a history of reticulation. This result is similar in spirit (though different in detail) to results by Bafna and Bansal (2004), Gusfield (2004), and Huson et al. (2005).

Let $\mathcal{H} = (V, E)$ be a hybrid phylogeny on $X$ with root node $\rho$. Let $V_C$ be the set of nodes of $\mathcal{H}$ that lie on at least one undirected cycle (that is, a cycle that arises by ignoring the orientation of the arcs). Let $V_T = (V - V_C) \cup \{\rho\} \cup X$. For a node $v$ of $V$, let $c(v)$ denote the set of species $x$ in $X$ for which there is a directed path from $v$ to $x$

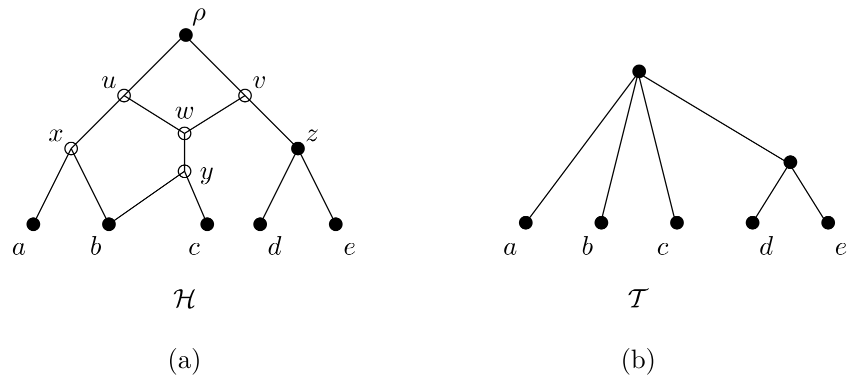


FIGURE 2. (a) A hybrid phylogeny $\mathcal{H}$ and (b) the rooted phylogenetic tree associated with $\mathcal{H}$ as described in Proposition 1.

(i.e. $c(v)$ is the extant species for which $v$ is an ancestor, often referred to as a cluster or a clade). To illustrate these concepts, consider the hybrid phylogeny $\mathcal{H}$ shown in Fig. 2(a). Here the nodes in $V_T$ are solid. Furthermore, $c(u) = \{a, b, c\}$ and $c(z) = \{d, e\}$.

A *hierarchy* $\mathcal{C}$ on $X$ is a collection of subsets of $X$, containing $X$ and all singleton subsets of $X$, and satisfying the property

$$A, B \in \mathcal{C} \Rightarrow A \cap B \in \{\emptyset, A, B\}.$$

Observe that the sets in $\mathcal{C}$ are *nested*—if they have one or more species in common, then one set is a subset of the other. It is a classical result in phylogenetics that a hierarchy on $X$ is exactly the set of clusters of a rooted phylogenetic $X$-tree. Given a hybrid phylogeny $\mathcal{H}$, the following result describes a tree that underlies $\mathcal{H}$. Informally speaking, it is the tree obtained by 'collapsing' portions of $\mathcal{H}$ where hybridization has occurred. This has the potential to give rise to trees that are poorly resolved in places.

**Proposition 1.** *Let $\mathcal{H}$ be a hybrid phylogeny on $X$ with node set $V$. Then the collection $\mathcal{C} = \{c(v) : v \in V_T\}$ is a hierarchy on $X$, in which case there is a rooted phylogenetic $X$-tree whose set of clusters is $\mathcal{C}$.*

*Proof.* The proof is by contradiction. Suppose that $\{c(v) : v \in V_T\}$ is not a hierarchy. By definition, there exist nodes $v_1, v_2 \in V_T$ and elements $a, b, x \in X$ such that $x \in c(v_1) \cap c(v_2)$, $a \in c(v_1) - c(v_2)$, and $b \in c(v_2) - c(v_1)$. Since $c(v_1)$ is not a subset of $c(v_2)$, there is no directed path in $\mathcal{H}$ from $v_2$ to $v_1$. Similarly, there is no directed path from $v_1$ to $v_2$. Since $x \in c(v_1) \cap c(v_2)$ there is a directed path $P_1$ from $v_1$ to $x$ and a directed path $P_2$ from $v_2$ to $x$. Let $v$ be the first node that is shared by both $P_1$ and $P_2$. Note that such a node exists since $x$ is a node shared by $P_1$ and $P_2$. Since there is no directed path from $v_1$ to $v_2$ or $v_2$ to $v_1$, we know that $v \neq v_1$ and $v \neq v_2$. Similarly, there exist directed paths $Q_i$ from $\rho$ to $v_i$ (for $i = 1, 2$) and we can let $w$ be the last node that is shared by $Q_1$ and $Q_2$. Again such a node exists since $\rho$ is shared by both $Q_1$ and $Q_2$. Now if we ignore the direction of the four paths $P_1$, $P_2$, $Q_1$, and $Q_2$ then the path from $w$ to $v_1$ (given by $Q_1$) and $w$ to $v_2$ (given by $Q_2$) and from $v_1$ to $v$ (given by $P_1$) and from $v_2$ to $v$ (given by $P_2$) constitutes an undirected cycle in $\mathcal{H}$, contradicting the assumption that $v_1, v_2 \in V_T$. □

For the hybrid phylogeny $\mathcal{H}$ shown in Fig. 2(a), the above construction yields the rooted phylogenetic tree $\mathcal{T}$ shown in Fig. 2(b). Here $\mathcal{C}$ in the statement of Proposition 1 is

$$\big\{\{a, b, c, d, e\}, \{d, e\}, \{a\}, \{b\}, \{c\}, \{d\}, \{e\}\big\}.$$

**Real-time hybrids.** Maddison (1997) (see also Moret et al., 2004) pointed out an important biological requirement of hybrid phylogenies. Namely, although a hybrid phylogeny might display two trees, there may be no process of hybridization between contemporaneous taxa (either past or present) that can realize this hybrid phylogeny. Nevertheless, by allowing for additional (unsampled, or perhaps extinct) taxa one can resolve this issue without introducing any additional hybridizations. Essentially the role of such an additional taxa is to 'carry' a gene (or combination of genes) from the past into some time when it can be inserted into the new hybrid species. Whether these taxa really are (or were) present is another question, but if we are concerned with just placing lower bounds on the degree of hybridization then we can (conservatively) allow them.

To illustrate this point, consider Fig. 3. Both hybrid phylogenies $\mathcal{H}$ and $\mathcal{H}'$ display $\mathcal{T}_1$ and $\mathcal{T}_2$ using two hybridization nodes. However, while $\mathcal{H}$ has a 'real-time' realization (see Fig 4)—a concept that will be formalized in the fifth section, $\mathcal{H}'$ has no such realization. To see the latter, observe that the "parents" of the hybrid species $b$ must coexist in time and the "parents" of the hybrid species $c$ must also coexist in time. Yet, by considering the ancestor-descendant relationships of these parents, this is not possible. Nevertheless, by allowing another

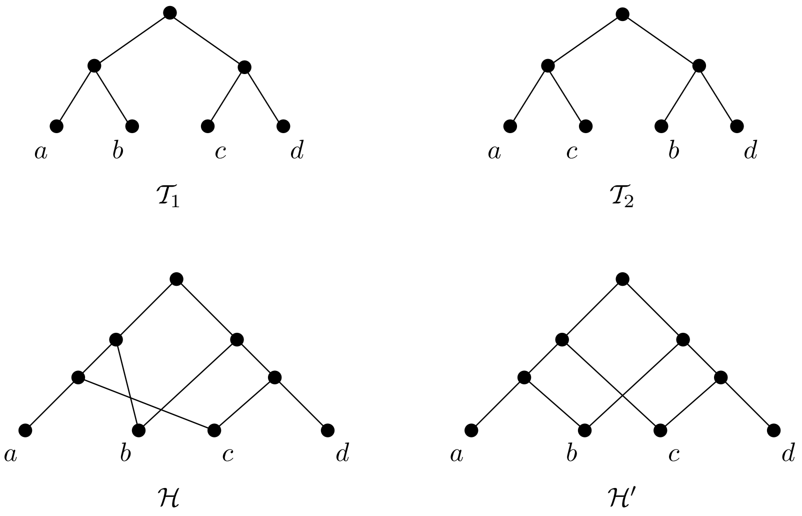


FIGURE 3. Two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ and two hybrid phylogenies $\mathcal{H}$ and $\mathcal{H}'$ that display $\mathcal{T}_1$ and $\mathcal{T}_2$.

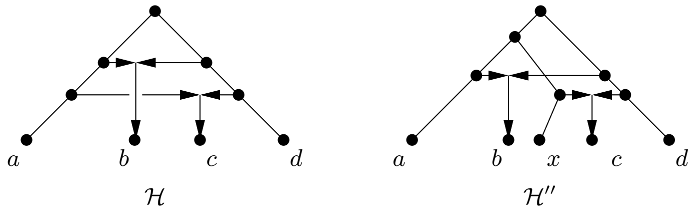


FIGURE 4. Two hybrid phylogenies that explain the real-time evolutionary histories of $\mathcal{T}_1$ and $\mathcal{T}_2$ in Fig. 3.

species $x$ that may be either extinct or not yet sampled, one can provide such a realization to $\mathcal{H}'$. This realization is shown as $\mathcal{H}''$ in Fig. 4.

In the fifth section we present an algorithm for determining whether a given hybrid phylogeny has a 'real-time' realization, or whether additional taxa (as in $\mathcal{H}''$ in Fig. 4) might be required.

**Finding the minimal degree of hybridization.** A topical question is: what is the smallest number or reticulation events required to explain a set of gene trees? This number sets a lower bound on the degree of reticulation that has occurred in the evolution of the species under consideration. If this initial set of data is a collection of rooted phylogenetic trees, this problem can be interpreted within the framework of hybrid phylogenies as follows.

For a hybrid phylogeny $\mathcal{H}$ with node set $V$ and root $\rho$, set

$$h(\mathcal{H}) = \sum_{v \in V; v \neq \rho} (d^-(v) - 1).$$

Note that, as $d^-(v)$ is the number of parents of $v$ and every node has exactly one parent if there is no hybridization, $d^-(v) - 1$ is the number of "extra parents" that $v$ has. Observe that $h(\mathcal{H}) \geq 0$, and $h(\mathcal{H}) = 0$ precisely if $\mathcal{H}$ is a rooted phylogenetic tree. Extending this definition, the *hybrid number* of a collection $\mathcal{P}$ of rooted phylogenetic trees is

$$h(\mathcal{P}) = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybrid phylogeny that displays } \mathcal{P}\}.$$

The value $h(\mathcal{P})$ represents the smallest number of hybridization events that are required to explain $\mathcal{P}$. Bordewich and Semple (2005) showed that computing this number is NP-hard even in the simplest case that $\mathcal{P}$ consists of just two rooted binary phylogenetic trees on the same leaf sets. However, despite this negative result, there are some attractive and useful positive results that have recently been described for computing and bounding $h(\mathcal{P})$. We describe these in the next section.

## The Minimum Number of Hybrid Events Required for Two Trees

We begin this section with some further graph-theoretic notation. Let $\mathcal{T}$ be a rooted binary phylogenetic $X$-tree and let $A$ be a subset of $X$. We denote the minimal rooted subtree of $\mathcal{T}$ that connects the elements in $A$ by $\mathcal{T}(A)$. Furthermore, we use $\mathcal{T}|A$ to denote the rooted subtree that is obtained from $\mathcal{T}(A)$ by suppressing all nodes of indegree one and outdegree one.

Now let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic $X$-trees. We will write $h(\mathcal{T}, \mathcal{T}')$ to denote $h(\mathcal{P})$ for $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$.

The first result we describe shows how one can simplify the calculation of $h(\mathcal{T}, \mathcal{T}')$ when one or more clusters are shared by both $\mathcal{T}$ and $\mathcal{T}'$. More precisely, suppose that $A \subset X$ is a cluster of both $\mathcal{T}$ and $\mathcal{T}'$ (that is, there is a node of each tree that has $A$ as its set of descendants in $X$). Let $\mathcal{T}|A$ and $\mathcal{T}'|A$ denote the subtree of $\mathcal{T}$ and $\mathcal{T}'$ (respectively) that have leaf set $A$, and let $\mathcal{T}_a$ and $\mathcal{T}'_a$ be the rooted trees obtained from $\mathcal{T}$ and $\mathcal{T}'$ (respectively) by replacing the subtree having leaf set $A$ with a new leaf $a$.

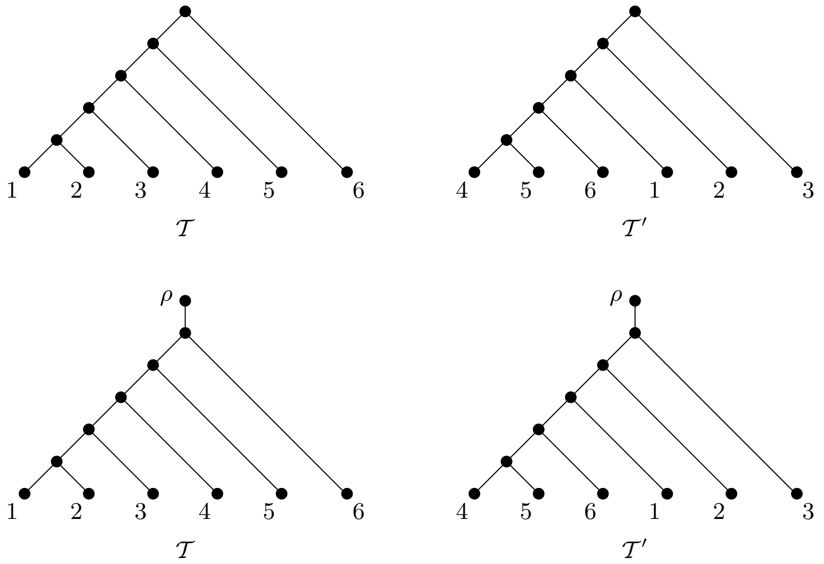


Figure 5. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ without (above) and with (below) their root labelled $\rho$.

**Theorem 1.** *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic $X$-trees. Suppose that $A \subset X$ is a cluster of both $\mathcal{T}$ and $\mathcal{T}'$. Then*

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a).$$

The proof of Theorem 1 is given in the Appendix. This result is typical of other relationships that can be established by exploiting a description of $h(\mathcal{T}, \mathcal{T}')$ in terms of what has recently been called a "good-agreement-forest" for the pair $\mathcal{T}$ and $\mathcal{T}'$ (see Baroni et al., 2005). ("Good" is an overused term, so in this paper we will refer to such agreement forests as "acyclic".) We describe this connection now, and provide an application in the next section to show how these results can be used in practice.

To make the interpretation work, we regard the root of both $\mathcal{T}$ and $\mathcal{T}'$ as a node $\rho$ that is adjoined to the original root by a new edge. Furthermore, we view $\rho$ as part of the label sets of both $\mathcal{T}$ and $\mathcal{T}'$; that is, we view the label sets of $\mathcal{T}$ and $\mathcal{T}'$ as $X \cup \{\rho\}$. For example, consider the two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ shown in the top part of Fig. 5. For the purposes of the interpretation, we view $\mathcal{T}$ and $\mathcal{T}'$ as shown in the bottom part of Fig. 5.

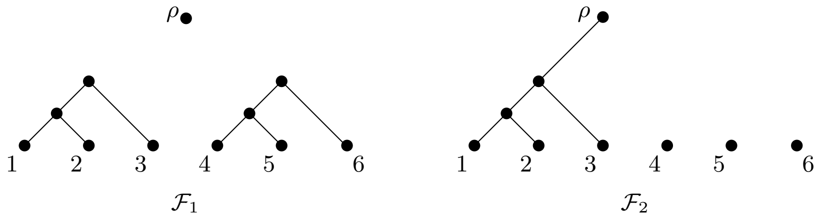


FIGURE 6. Two agreement forests for the two rooted binary phylogenetic trees shown in Fig. 5.

An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ with $k+1$ components is a collection $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is a rooted tree whose label set $\mathcal{L}_\rho$ includes $\rho$ and $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k$ are rooted binary phylogenetic trees with label sets $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$ such that the following properties are satisfied:

(i) The label sets $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$ partition $X \cup \{\rho\}$.
(ii) For all $i \in \{\rho, 1, 2, \dots, k\}$, $\mathcal{T}_i$ is the same as (isomorphic to) $\mathcal{T}|\mathcal{L}_i$ and $\mathcal{T}'|\mathcal{L}_i$.
(iii) The trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$ are node disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

More informally, $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ if, up to suppressing degree-two nodes, $\mathcal{F}$ can be obtained from each of $\mathcal{T}$ and $\mathcal{T}'$ by deleting $|\mathcal{F}|-1$ edges. As an example, the two forests $\mathcal{F}_1$ and $\mathcal{F}_2$ shown in Fig. 6 are both agreement forests for the two trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Fig. 5.

It has recently been shown (Bordewich and Semple, 2004) that for any two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on the same leaf set the smallest value of $k$ of any agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ equals the *rooted subtree prune and regraft distance* between $\mathcal{T}$ and $\mathcal{T}'$. Denoted $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$, this distance is the minimum number of rooted subtree prune and regraft operations required to transform $\mathcal{T}$ into $\mathcal{T}'$. It is tempting to conjecture that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ and $h(\mathcal{T}, \mathcal{T}')$ are identical, and indeed the former takes the value 1 if and only if the latter does. However, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ is only a lower bound for $h(\mathcal{T}, \mathcal{T}')$, and one can construct pairs of trees $\mathcal{T}$ and $\mathcal{T}'$ on $n$ species such that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 2$ yet $h(\mathcal{T}, \mathcal{T}') > \frac{n}{2} - 1$ (Baroni et al., 2005).

An agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is a *maximum-agreement forest* if, amongst all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, it has the smallest number

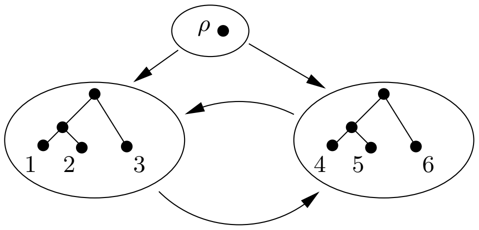


FIGURE 7. The graph $G_{\mathcal{F}_1}$.

of components. To continue the previous example, it is straightforward to check that the forest $\mathcal{F}_1$ in Fig. 6 is a maximum-agreement forest for the two trees $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 5. Thus the rooted subtree prune and regraft distance between these two trees is 2. For the interpretation of $h(\mathcal{T}, \mathcal{T}')$ in terms of agreement forest, we need one further definition.

Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $G_{\mathcal{F}}$ be the directed graph whose nodes represent the trees in $\mathcal{F}$ and for which $(\mathcal{T}_i, \mathcal{T}_j)$ is a directed edge from the node representing $\mathcal{T}_i$ to the node representing $\mathcal{T}_j$ precisely if $i \neq j$ and either

(I) the root of the subtree $\mathcal{T}(\mathcal{L}_i)$ in $\mathcal{T}$ is an ancestor of the root of the subtree $\mathcal{T}(\mathcal{L}_j)$ in $\mathcal{T}$, or

(II) the root of the subtree $\mathcal{T}'(\mathcal{L}_i)$ in $\mathcal{T}'$ is an ancestor of the root of the subtree $\mathcal{T}'(\mathcal{L}_j)$ in $\mathcal{T}'$.

Since $\mathcal{F}$ is an agreement forest, the roots of the subtrees $\mathcal{T}(\mathcal{L}_i)$ and $\mathcal{T}(\mathcal{L}_j)$, and the roots of the subtrees $\mathcal{T}'(\mathcal{L}_i)$ and $\mathcal{T}'(\mathcal{L}_j)$ are not the same. We call $\mathcal{F}$ a *acyclic-agreement forest* if $G_{\mathcal{F}}$ is acyclic; that is, if $G_{\mathcal{F}}$ has no directed cycles. Furthermore, if over all acyclic-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, $\mathcal{F}$ contains the smallest number of components, then $\mathcal{F}$ is a *maximum-acyclic-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$, in which case we denote this value of $k$ by $m_g(\mathcal{T}, \mathcal{T}')$. Observe that $m_g(\mathcal{T}, \mathcal{T}') = 0$ if and only if, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical. The forest $\mathcal{F}_2$ in Fig. 6 is a acyclic-agreement forest for the two trees $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 5. Indeed, this forest is a maximum-acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. To see that $\mathcal{F}_1$ is not a acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, observe that $G_{\mathcal{F}_1}$ contains a directed cycle (see Fig. 7, where the nodes are drawn as large circles enclosing the trees they represent).

The interpretation of the hybrid number of two rooted binary phylogenetic trees on the same label sets in terms of agreement forests is stated in following theorem which is established by Baroni et al. (2005).

**Theorem 2.** *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic $X$-trees. Then*

$$h(\mathcal{T}, \mathcal{T}') = m_g(\mathcal{T}, \mathcal{T}').$$

For example, it follows from Theorem 2 that the value of $h(\mathcal{T}, \mathcal{T}')$ for the two trees in Fig. 5 is 3.

We mentioned previously that computing $h(\mathcal{T}, \mathcal{T}')$ is NP-hard. The reason for this is that finding a maximum-acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is NP-hard. Currently, the best known method for finding such a forest is trial and error. However, if one has an acyclic-agreement forest $\mathcal{F}$ (not necessarily maximum) for $\mathcal{T}$ and $\mathcal{T}'$, then there is a simple algorithm using $\mathcal{F}$ for constructing a hybrid phylogeny that displays both $\mathcal{T}$ and $\mathcal{T}'$. This algorithm is provided by the inductive proof of Theorem 2 in Baroni et al. (2005) and is given below.

There is a simple, fast, and well-known way of deciding whether or not a directed graph $D$ is acyclic. Find a node, $v_1$ say, that has indegree zero. If there is no such node, then $D$ contains a directed cycle. Now delete $v_1$ (and all arcs incident with $v_1$) from $D$, and find a node, $v_2$ say, that has degree zero. Again, if there is no such node, $D$ contains a directed cycle. Deleting $v_2$ and continuing in this way, we eventually find that $D$ is not acyclic or obtain an ordering of the nodes, $v_1, v_2, \ldots, v_n$ say of $D$, so that for all $i \in \{1, 2, \ldots, n\}$, the node $v_i$ has indegree zero in the digraph obtained from $D$ by deleting the nodes $v_1, v_2, \ldots, v_{i-1}$ and all edges incident with these nodes. This ordering implies that $D$ is acyclic (see Lemma 1). Consequently, we will call such an ordering an *acyclic ordering* of $D$. We remark here that this process is formally incorporated in the algorithm given in the fifth section.

The algorithm for constructing a hybrid phylogeny from an acyclic-agreement forest $\mathcal{F}$ is as follows. Note that, in any acyclic ordering of $G_{\mathcal{F}}$, the node $\mathcal{T}_\rho$ always appears first.

**Algorithm:** HybridPhylogeny($\mathcal{F}$)
**Input:** An acyclic-agreement forest $\mathcal{F}$ for two rooted binary phylogenetic $X$-trees $\mathcal{T}$ and $\mathcal{T}'$ with $k+1$ components.
**Output:** A hybrid phylogeny $\mathcal{H}$ that displays both $\mathcal{T}$ and $\mathcal{T}'$ in which the number of hybridization nodes is $k$.

**1.** Find an acyclic ordering, $\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ say, of $G_{\mathcal{F}}$.
**2.** Set $\mathcal{H}_0 = \mathcal{T}_\rho$ and set $i = 1$.
**3.** Attach $\mathcal{T}_i$ to $\mathcal{H}_{i-1}$ via two new edges. Each of these edges join the root of $\mathcal{T}_i$ to some (not necessarily distinct) edge of $\mathcal{H}_{i-1}$. These edges are added so that the resulting hybrid phylogeny displays $\mathcal{T}|\mathcal{L}(\{\mathcal{T}_\rho, \mathcal{T}_1, \ldots, \mathcal{T}_i\})$ and $\mathcal{T}'|\mathcal{L}(\{\mathcal{T}_\rho, \mathcal{T}_1, \ldots, \mathcal{T}_i\})$.
Set $\mathcal{H}_i$ to be the resulting hybrid phylogeny, and return $\mathcal{H}_i$ if $i = k$.
**4.** Increment $i$ by $1$ and go to Step 3.

## Application

In this section, we apply the theory of the last section to two phylogenetic trees on 46 sequences of alpine Rununculi of New Zealand, reported by Lockhart et al. (2001). The first tree was constructed from nuclear ITS sequences, while the second was constructed from chloroplast ($J_{SA}$) sequences (for details see Lockhart et al., 2001). The two trees showed considerable agreement, however there was also a fair degree of incompatibility. One possible explanation for this incompatibility is the occurrence of hybrid evolution, whereby the nuclear ITS sequence has a different history to the chloroplast ($J_{SA}$) sequences. Of course, there may be other sources of phylogenetic error (sampling effects such as noise, model mis-specification, lineage sorting) that could cause the two trees to conflict, even in the absence of any hybrid evolution. Nevertheless, we can still ask the following question: Assuming the two trees correctly describe the history of the two genes, and their incongruence is due to hybrid evolution, what is the smallest number of hybrid events required to explain this? The study is complicated slightly by the fact that neither tree is binary. In this case, we took a conservative approach and allowed non-binary subtrees to be resolved in any way that helped minimize the required number of hybridization events. Also, for the sake of illustration in this paper, we will restrict attention to a subgroup ("Group I") of the sequences consisting of 20 sequences. This group is a candidate for reticulate evolution, since the $F_1$ progeny of hybrid origin are known to be fertile (Fisher, 1965). The two trees for these 20 sequences are shown in Fig. 8, with $\mathcal{T}_1$ the nuclear, and $\mathcal{T}_2$ the chloroplast tree.

For $\mathcal{T}_1$ and $\mathcal{T}_2$, one can identify five clusters (denoted $l_1$ to $l_5$ in Fig. 8) shared by these two trees; this allows us to apply Theorem 1. In this way we reduce the problem from comparing two 20-taxon trees to one of comparing two 5-taxon trees (each having leaf set $l_1, \ldots, l_5$),

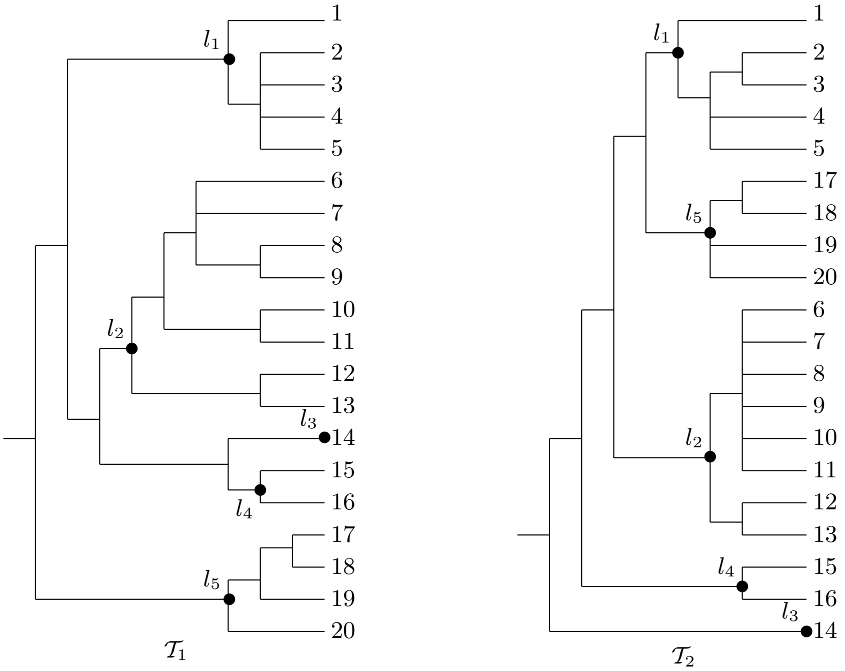


FIGURE 8. The tree $\mathcal{T}_1$ for nuclear ITS sequences and $\mathcal{T}_2$ for chloroplast $J_{SA}$ sequences from Lockhart et al. (2001) restricted to Group I.

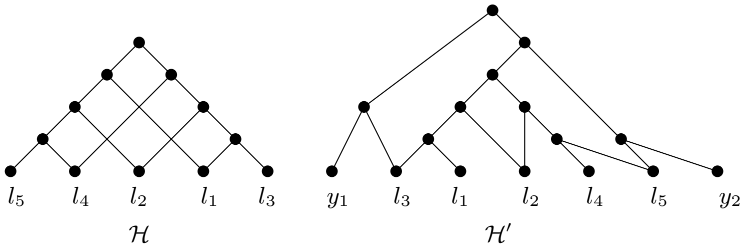


FIGURE 9. Two hybrid phylogenies that display $\mathcal{T}_1$ and $\mathcal{T}_2$, and requiring three hybridization events (the fewest possible for these two trees).

together with the trees on the shared clusters (in fact these latter trees do not contribute to the $h$ score, since all these pairs of cluster subtrees are compatible). Now using Theorem 2, one can show using a detailed case analysis that $h(\mathcal{T}_1, \mathcal{T}_2) = 3$. Fig. 9 shows one hybrid phylogeny ($\mathcal{H}$) that displays the five clusters shared by $\mathcal{T}_1$ and $\mathcal{T}_2$ with three hybrid events. Note that this is not the only such phylogeny. Similarly, for the full set of 46 sequences it can be shown (by hand) that the $h$ value lies between 7 and 12 (Baroni, 2004). Thus, assuming the trees are

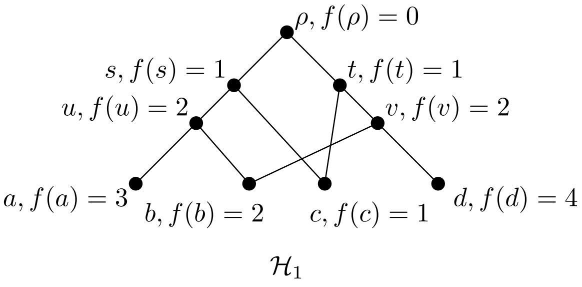


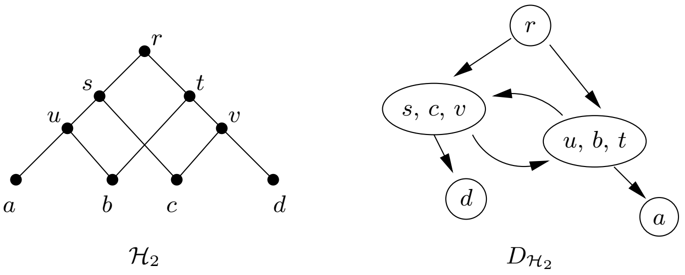
FIGURE 10. A temporal labelling of a hybrid.

correct we require at least 3 hybrid events to describe the evolution of the Group I sequences, and at least 7 hybrid events to describe the evolution of the entire group of 46 sequences. We should stress that this analysis is to illustrate the techniques, rather than to formally show that there has been this degree of hybrid evolution in the taxa described—as we mentioned there are other reasons why trees may disagree, and these need to be considered (these other processes often leave different statistical signatures from hybridization, see Holder et al., 2001; Huson et al. 2005).

Using an argument similar to that used to show that $\mathcal{H}'$ in Fig. 3 has no real-time realization (in the sense described in ), it is easily checked that the hybrid phylogeny $\mathcal{H}$ shown in Fig. 9 also has no real-time realization. However the hybrid phylogeny $\mathcal{H}'$ in Fig. 9 allows for a 'real-time' hybrid evolution scenario, with just two extra taxa $y_1$ and $y_2$. Although the analysis of deciding a real time realization could be resolved for this small-scale example by an ad-hoc case analysis, it is clear that such a task could be complicated for a large and complex hybrid phylogeny. In the next section, we present an algorithm to determine whether an arbitrary hybrid phylogeny can be realized by hybrid evolution between contemporaneous ancestral taxa.

## An Algorithm for 'Real-Time' Hybrids

The concept of a 'real-time hybrid' has been briefly and informally mentioned already; now we formalize this notion, and provide an algorithm to determine whether an arbitrary hybrid phylogeny can be realized in this way. Some of the more technical parts of this section have been moved to an Appendix to assist readability.



(a) (b)

FIGURE 11. (a) A hybrid phylogeny $\mathcal{H}_2$ with no temporal representation and (b) its associated digraph $D_{\mathcal{H}_2}$.

Let $\mathcal{H}$ be a hybrid phylogeny with node set $V$ and arc set $A$. We say that $\mathcal{H}$ has a *temporal representation* if there exists a map $f : V \to \mathbb{N} = \{0, 1, 2, \ldots, \}$ with the following properties:

(i) If $v$ is a node of $\mathcal{H}$ with $d^-(v) = 1$, then $f(u) < f(v)$ for the (only one) immediate ancestor $u$ of $v$.
(ii) If $v$ is a node of $\mathcal{H}$ with $d^-(v) \geq 2$, then $f(u) = f(v)$ for all immediate ancestors $u$ of $v$.

Such a map is a called a *temporal labelling* of $\mathcal{H}$. To illustrate, a temporal labelling of a hybrid phylogeny is shown in Fig. 10, where, for each node, the first element is the node and the second element is the element of $\mathbb{N}$ assigned under the temporal representation $f$. All rooted phylogenetic trees have a temporal representation. However, not all hybrid phylogenies have such a representation. For example, the hybrid phylogeny shown in Fig. 11(a), which has the same shape as $\mathcal{H}'$ shown in Fig. 3, has no temporal representation.

The main result of this section (Theorem 3) is to characterize exactly when an arbitrary hybrid phylogeny has a temporal representation. To this end, we next describe a particular digraph $D_{\mathcal{H}}$ associated with a fixed hybrid phylogeny $\mathcal{H}$ with node set $V$. This graph is not designed to depict the evolutionary relationships, instead it summarizes properties of $\mathcal{H}$. The node set for this new graph will be denoted $[V]$ and will consist of nodes $[v]$ which represent either a single node in $\mathcal{H}$, or a subset of nodes in $\mathcal{H}$ that must have been contemporaneous (because they are nodes involved in the same hybridization event, as parental species or as the child species). In particular, let $V$ and $A$ be the node

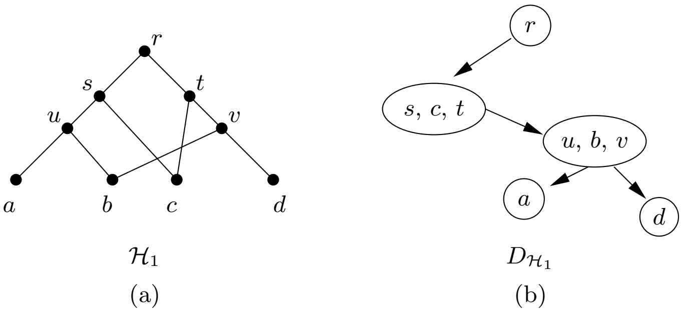


FIGURE 12. (a) A hybrid phylogeny $\mathcal{H}_1$ and (b) its associated digraph $D_{\mathcal{H}_1}$.

and arc sets of $\mathcal{H}$, respectively. Let

$$A_T = \{(u, v) \in A : d^-(v) = 1\}$$

and

$$A_H = \{(u, v) \in A : d^-(v) \geq 2\}.$$

Any arc in $A_T$ is called a *tree arc* and any arc in $A_H$ is called a *hybridization arc*. Note that the sets $A_T$ and $A_H$ partition $A$. Ignoring the direction of the arcs of $\mathcal{H}$, an equivalence relation on $V$ is now defined by setting

$$[v] = \{v\} \cup \{u \in V : \text{there is a path of hybridization arcs from } u \text{ to } v \text{ in } \mathcal{H}\}.$$

Observe that if $v$ is not incident with a hybridization arc, then $[v] = \{v\}$. Set

$$[V] = \{[v] : v \in V\}.$$

We describe our associated digraph $D_{\mathcal{H}}$ as follows. The node set of $D_{\mathcal{H}}$ is $[V]$, and $[u]$ and $[v]$ are joined by an arc $([u], [v])$ if there exists $a \in [u]$ and $b \in [v]$ such that $(a, b)$ is a tree arc in $A$. It is easily seen that $D_{\mathcal{H}}$ is connected. To illustrate, consider Figs 11 and 12. Figure 11(b) shows the digraph $D_{\mathcal{H}_2}$, where $\mathcal{H}_2$ is shown in Fig. 11(a) with

$$[V] = \big\{\{r\}, \{s, c, v\}, \{u, b, t\}, \{a\}, \{d\}\big\}.$$

Furthermore, for the hybrid phylogeny $\mathcal{H}_1$ shown in Fig. 12(a), the digraph $D_{\mathcal{H}_1}$ is shown in Fig. 12(b). To provide some intuition for $D_{\mathcal{H}}$, we note that the arcs in $D_{\mathcal{H}}$ represent the direction of time. Thus a directed cycle means that a descendant species is older than its ancestors, which is not possible.

Let $\mathcal{H}$ be a hybrid phylogeny and suppose that $f : V \to \mathbb{N}$ is a temporal labelling of $\mathcal{H}$. Let $\overline{f}$ be the map from $[V]$ to $\mathbb{N}$ that is defined by setting $\overline{f}([v]) = f(v)$ for all $v \in V$. To see that this map is well-defined, first observe that if $[u] = [v]$, then there is an (undirected) path from $u$ to $v$ consisting of hybridization arcs. Since the end nodes of any arc on this path are assigned the same natural number under $f$, it follows that all nodes in this path are assigned the same natural number under $f$. Hence, for all $w, w' \in [v]$, we have $f(w) = f(w')$. Thus $\overline{f}$ is well-defined. Moreover, as $f$ is a temporal labelling of $\mathcal{H}$, there is no $u$ and $v$ in the same equivalence class such that $(u, v)$ is a tree arc.

The following result provides a concise characterization for when a hybrid phylogeny has a temporal representation; its proof is given in the Appendix.

**Theorem 3.** *A hybrid phylogeny $\mathcal{H}$ has a temporal representation if and only if $D_{\mathcal{H}}$ is acyclic.*

Theorem 3 is the basis for a polynomial-time algorithm (TempRep) for determining whether or not a hybrid phylogeny has a temporal representation and, if so, providing a temporal labelling.

**Algorithm:** TempRep($\mathcal{H}$)
**Input:** A hybrid phylogeny $\mathcal{H}$ with node set $V$.
**Output:** A temporal labelling $f$ of $\mathcal{H}$ or the statement *$\mathcal{H}$ has no temporal representation.*

```
1. Construct $D_{\mathcal{H}}$.
2. Set $i = 0$ and $D_0 = D_{\mathcal{H}}$.
3. Choose $S_i$ to be any non-empty set of nodes of $D_i$ that have indegree
   zero. If there are no such nodes, then halt and return $\mathcal{H}$ has no temporal
   representation.
4. Set $D_{i+1}$ to the digraph resulting from $D_i$ by deleting the nodes $S_i$ and
   all arcs incident with these nodes. If $D_{i+1}$ is the empty graph, then go
   to Step 5. Otherwise, increment $i$ by $1$ and go to Step 3.
5. Define $f : V \rightarrow \mathbb{N}$ by setting $f(v) = i$ for all $v \in V$, where $[v] \in S_i$.
6. Return the map $f$.
```

The correctness of this algorithm is guaranteed by the following result, whose proof is given in the Appendix.

**Theorem 4.** *Let $\mathcal{H}$ be a hybrid, and suppose that* TempRep *is applied to $\mathcal{H}$.*

(i) *If $\mathcal{H}$ has a temporal representation, then* TempRep *returns a temporal labelling of $\mathcal{H}$.*
(ii) *If $\mathcal{H}$ has no temporal representation, then* TempRep *returns the statement $\mathcal{H}$* has no temporal representation.

*Moreover, the running time of* TempRep *is quadratic in the size of the node set of $\mathcal{H}$.*

For example, if one takes the hybrid phylogeny $\mathcal{H}_1$ in Fig. 12(a) and apply the algorithm TempRep, we can reconstruct the temporal representation shown in Fig. 10. Note that there is some choice as to the assignment of numbers for the leaves $a$ and $d$. Such choices will generally arise for any hybrid phylogeny that has a temporal representation. Observe that it is the relative ordering of the nodes and not the actual values assigned by a temporal labelling that is important. We can make this idea more precise as follows.

Let $\mathcal{H}$ be a hybrid phylogeny with node set $V$ that has a temporal representation, and let $f_1$ and $f_2$ be two temporal temporal labellings of $\mathcal{H}$. We say that $f_1$ and $f_2$ are *ordering isomorphic* if, for all $u, v \in V$, the following hold:

(i) $f_1(u) < f_1(v)$ if and only if $f_2(u) < f_2(v)$;
(ii) $f_1(u) = f_1(v)$ if and only if $f_2(u) = f_2(v)$.

Using the results in this section (and the Appendix) one can construct an algorithm that lists, up to ordering isomorphism, all temporal labellings of $\mathcal{H}$ so that each such labelling is outputted in polynomial time. An outline of this algorithm is given in the Appendix. It is important to note that, as this list may be exponential in the size of $V$, the algorithm itself is not guaranteed to run in polynomial time.

We end this section by noting that, although a hybrid phylogeny may have a temporal labelling, this does not mean that unsampled lineages could not have been involved in the event.

## Concluding Remarks

The reconstruction and analysis of hybrid phylogenies gives rise to many challenging mathematical and computational problems. We have described some results that can help set lower bounds on the extent of hybridization required to explain the conflict between two phylogenetic trees. This is currently an active area of research in bioinformatics (see e.g., Huydn et al., 2005; Huson et al., 2005). Ultimately statistical questions will also need to be addressed – for example, how can one use differing bootstrap (or Bayesian posterior probability) support values for different trees to quantify and distinguish genuine reticulate evolution from other phenomena (eg. lineage sorting) that can give rise to conflicting phylogenies? In the classical phylogenetic analysis on trees, a combinatorial analysis often lays the foundation for later statistical approaches (for example, Peter Buneman's work in the early 1970s concerning the four-point condition provided a basis for now widely-used distance-based approaches in phylogenetics such as neighbor-joining with model-corrected distances). Combinatorial insights into hybrid phylogenies are likely also to help in developing statistically-based approaches to the study of reticulate evolution.

We have also explored the issue of determining whether a hybrid phylogeny has a real-time realization, and provided a simple characterization (and algorithm) for this task. This algorithm runs in polynomial-time; and a naive implementation would allow a running time that is quadratic in the number of nodes, though it is possible that a more clever implementation could improve this.

Lastly, in general, a hybrid phylogeny on $X$ that displays a collection of rooted binary phylogenetic $X$-trees need not be unique. Deciding whether there exists such a hybrid phylogeny is an interesting question and one that may have an attractive combinatorial solution.

## Acknowledgments

We thank Peter Lockhart for encouragement to develop some of the theory in this paper, and for helpful subsequent comments. We also thank the referees for their helpful and constructive comments. A special thank you to Mark Holder (referee) for his detailed comments.

## References

Bafna, V., and V. Bansal. 2004. The number of recombination events in a sample history: conflict graph and lower bounds. IEEE/ACM Trans. Comput. Biol. Bioinf. 1(2):78–90.

Bang-Jensen, J., and G. Gutin. 2001. Digraphs: Theory, Algorithms and Applications. Springer-Verlag, London.

Baroni, M. 2004. Hybrid phylogenies: a graph-based approach to represent reticulate evolution. PhD thesis. University of Canterbury, Christchurch, New Zealand.

Baroni, M., C. Semple, and M. A. Steel. 2004. A framework for representing reticulate evolution. Ann. Combin. 8(4):391–408.

Baroni, M., S. Grünewald, V. Moulton, C. Semple. 2005. Bounding the number of hybridisation events for a consistent evolutionary history. Journal of Mathematical Biology 51:171–182.

Bordewich, M., and C. Semple. 2004. On the computational complexity of the rooted subtree prune and regraft distance. Ann. Combin. 8(4):409–423.

Bordewich, M., and C. Semple. 2005. Computing the minimum number of hybridisation events for a consistent evolutionary history. Submitted manuscript (Departmental report UCDMS2004/21, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand).

Fisher, F.J.F. 1965. The alpine Ranunculi of New Zealand. DSIR publishing, New Zealand.

Gusfield, D. 2004. A fundamental decomposition theory for phylogenetic networks and incompatible characters. Technical Report UC-Davis CSE-2004-32.

Gusfield, D., S. Eddhu, C. Langley. 2004. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J. Bioinf. Comput. Biol. 2(1):173–213.

Holder, M. T., J. A. Anderson and A. K. Holloway. 2001. Difficulties in detecting hybridization. Syst. Biol. 50(6):978–982.

Holland, B., K. Huber, V. Moulton and P. J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. 21:1459–1461.

Huson, D. H., T. Dezulian, T. Kloepper, and M. A. Steel. 2004. Phylogenetic super-networks from partial trees. IEEE Trans. Comput. Biol. Bioinf. 1(4):151–158.

Huson, D. H., T. Kloepper, P. J. Lockhart, M. A. Steel. 2005. Reconstruction of reticulate networks from gene trees. Pages 233–249 *in* Proceedings of RECOMB 2005 (S. Miyano et al. eds.). LNBI 3500, Springer-Verlag, Berlin.

Huydn, T.N.D., J. Jansson, N.B. Nguyen, and W.-K.Sung. 2005. Constructing a smallest refining galled phylogenetic network. Pages 265–280 *in* Proceedings of RECOMB 2005 (S. Miyano et al. eds.). LNBI 3500, Springer-Verlag, Berlin.

Legendre, P., and V. Makarenkov. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. 51(2):199–216.

Lockhart, P. J., P. A. McLenachan, D. Havell, D. Glenny, D. Huson, and U. Jensen. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. Ann. Missouri Bot. Gard. 88(3):458–477.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46(3):523–536.

Moret, B. M. E., L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun and R. Timme. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Trans. Comput. Biol. Bioinf. 1(1):1–11.

Song, Y., and J. Hein. 2004. On the minimum number of recombination events in the evolutionary history of DNA sequences. J. Math. Biol. 48:160–186.

## Appendix

**Proof of Theorem 1.** It is clear that the inequality holds if $A = X$. Therefore we may assume that $A \neq X$. We first show that

$$h(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a). \tag{1}$$

Let $\mathcal{F}_A$ be a maximum-acyclic-agreement forest for $\mathcal{T}|A$ and $\mathcal{T}'|A$, and let $\mathcal{F}_a$ be a maximum-acyclic-agreement forest for $\mathcal{T}|a$ and $\mathcal{T}'|a$. Let $\mathcal{T}_{i,a}$ be the unique tree in $\mathcal{F}_a$ with a node labelled $a$, and let $\mathcal{T}_{\rho,A}$ be the unique tree in $\mathcal{F}_A$ with a node labelled $\rho$. Let $\mathcal{T}_{A,a}$ be the tree obtained by adjoining $\mathcal{T}_{\rho,A}$ to $\mathcal{T}_{i,a}$ via an edge joining $\rho$ and $a$, removing the labels $\rho$ and $a$, and then suppressing any degree-two nodes. Because of the acyclic conditions on $\mathcal{F}_A$ and $\mathcal{F}_a$, we have that

$$\mathcal{F} = \big(\mathcal{F}_A \cup \mathcal{F}_a - \{\mathcal{T}_{i,a}, \mathcal{T}_{\rho,A}\}\big) \cup \{\mathcal{T}_{A,a}\}$$

is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_a| - 1$. It now follows by Theorem 2 that

$$\begin{aligned} h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a) &= |\mathcal{F}_A| - 1 + |\mathcal{F}_a| - 1 \\ &= |\mathcal{F}| - 1 \\ &\geq h(\mathcal{T}, \mathcal{T}'). \end{aligned}$$

This establishes (1).

We next show that

$$h(\mathcal{T}, \mathcal{T}') \geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a). \tag{2}$$

Let $\mathcal{F}$ be a maximum-acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. There are two cases to consider:

(i) there exists $\mathcal{T}_i \in \mathcal{F}$ such that $\mathcal{L}(\mathcal{T}_i) \cap A \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_i) \cap \big((X - A) \cup \{\rho\}\big) \neq \emptyset$, and

(ii) for all $\mathcal{T}_i \in \mathcal{F}$, either $\mathcal{L}(\mathcal{T}_i) \subseteq A$ or $\mathcal{L}(\mathcal{T}_i) \subseteq \big((X - A) \cup \{\rho\}\big)$.

**Case (i).** Assume that $\mathcal{T}_i$ is a such a tree in $\mathcal{F}$. Then the minimal subtree of $\mathcal{T}$ (and $\mathcal{T}'$) that contains the label set of $\mathcal{T}_i$ includes the root of $\mathcal{T}|A$ (and $\mathcal{T}'|A$). Since $\mathcal{F}$ is an agreement forest, this implies that $\mathcal{T}_i$ is the unique tree in $\mathcal{F}$ with the properties described in (i).

Let $x \in \mathcal{L}(\mathcal{T}_i) \cap A$, and let $\mathcal{T}_{i,a}$ be the tree obtained from $\mathcal{T}_i|\big((X - A) \cup \{\rho\} \cup \{x\}\big)$ by relabelling $x$ as $a$. Furthermore, let $\mathcal{T}_{i,A}$ be the tree obtained from $\mathcal{T}_i|A$ by adding $\rho$ at the end of a pendant edge adjoined to the root of $\mathcal{T}_i|A$. Then, as $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$,

$$\mathcal{F}_A = \{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq A\} \cup \{\mathcal{T}_{i,A}\}$$

is an acyclic-agreement forest for $T|A$ and $T'|A$, and

$$\mathcal{F}_a = \big\{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \big((X - A) \cup \{\rho\}\big)\big\} \cup \{\mathcal{T}_{i,a}\}$$

is an acyclic-agreement forest for $\mathcal{T}_a$ and $\mathcal{T}'_a$. Since $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_a| - 1$, we have that

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') &= |\mathcal{F}| - 1 \\ &= \big(|\mathcal{F}_A| + |\mathcal{F}_a| - 1\big) - 1 \\ &\geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a). \end{aligned}$$

This establishes (2) for (i).

**Case (ii).** Since $\mathcal{G}_{\mathcal{F}}$ does not contain any directed cycles, it follows that the sub-digraph of $\mathcal{G}_{\mathcal{F}}$ induced by the set $\{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\}$ does not contain any directed cycles. This means that this sub-digraph has a node, $\mathcal{T}_0$ say, of indegree zero. Let $\mathcal{T}_{0,\rho}$ be the tree obtained from $\mathcal{T}_0$ by adding $\rho$ at the end of a pendant edge adjoined to the root of $\mathcal{T}_0$. Since $\mathcal{F}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, it is easily seen that

$$\mathcal{F}_A = \big(\{\mathcal{T}_i \in \mathcal{F} : \mathcal{L}(\mathcal{T}_i) \subseteq A\} - \{\mathcal{T}_0\}\big) \cup \{\mathcal{T}_{0,\rho}\}$$

is an acyclic-agreement forest for $\mathcal{T}|A$ and $\mathcal{T}'|A$, and

$$\mathcal{F}_a = \big\{\mathcal{T}_j \in \mathcal{F} : \mathcal{L}(\mathcal{T}_j) \subseteq \big((X - A) \cup \{\rho\}\big)\big\} \cup \{a\}$$

is an acyclic-agreement forest for $\mathcal{T}_a$ and $\mathcal{T}'_a$, where $a$ is used denote the tree consisting of a single node labelled $a$. Thus $|\mathcal{F}| = |\mathcal{F}_A| + |\mathcal{F}_a| - 1$, and so, by Theorem 2,

$$\begin{aligned} h(\mathcal{T}, \mathcal{T}') &= |\mathcal{F}| - 1 \\ &= \big(|\mathcal{F}_A| + |\mathcal{F}_a| - 1\big) - 1 \\ &\geq h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a). \end{aligned}$$

This establishes (2) for (ii). Combining (1) and (2) completes the proof of the theorem.

**Proof of Theorem 3.** Let $D$ be a digraph with node set $V$ and arc set $A$, and suppose that $D$ is acyclic. In an earlier section, we described the concept of an acyclic ordering of $D$. It is easily seen that this is equivalent to there being a map $g : V \rightarrow \mathbb{N}$ such that, for all $(u, v) \in A$, we have $g(u) < g(v)$. Such a map $g$ will prove useful in proving Theorem 3.

The following lemma is well-known and easily proved (for example, see Bang-Jensen and Guitin, 2001).

**Lemma 1.** *A digraph is acyclic if and only if it has an acyclic ordering.*

**Proposition 2.** *Let $\mathcal{H}$ be a hybrid phylogeny with node set $V$ and suppose that $f : V \rightarrow \mathbb{N}$ is a temporal labelling of $\mathcal{H}$. Then $\overline{f}$ induces an acyclic ordering of $[V]$. In particular, $D_{\mathcal{H}}$ is acyclic.*

*Proof.* Let $f : V \rightarrow \mathbb{N}$ be a temporal labelling of $\mathcal{H}$, and consider $D_{\mathcal{H}}$. Let $([u],[v])$ be an arc of $D_{\mathcal{H}}$. To prove the proposition it suffices to show by Lemma 1 that $\overline{f}([u]) < \overline{f}([v])$. Now, by definition, there exists elements $a \in [u]$ and $b \in [v]$ such that $(a, b)$ is a tree arc of $\mathcal{H}$. Since $f$ is a temporal labelling of $\mathcal{H}$, we have that $f(a) < f(b)$, which in turn implies that $\overline{f}([u]) < \overline{f}([v])$ as required. □

**Proposition 3.** *Let $\mathcal{H}$ be a hybrid phylogeny with node set $V$, and suppose that $D_{\mathcal{H}}$ is acyclic. Let $g$ be an acyclic ordering of $[V]$. Let $f$ be the map from $V$ into $\mathbb{N}$ defined by setting $f(v) = g([v])$. Then $f$ is a temporal labelling of $\mathcal{H}$.*

*Proof.* Let $(u, v)$ be an arc of $\mathcal{H}$. First assume that $(u, v)$ is a tree arc. Then $u$ and $v$ are in different equivalence classes; otherwise, $D_{\mathcal{H}}$ contains a loop contradicting the fact that $D_{\mathcal{H}}$ is acyclic. Furthermore, there is an arc from $[u]$ to $[v]$ in $D_{\mathcal{H}}$. It now follows that $f(u) < f(v)$.

Now assume that $(u, v)$ is a hybridization arc of $\mathcal{H}$. Then $[u] = [v]$, and so $f(u) = f(v)$. Hence, by definition, $f$ is a temporal labelling of $\mathcal{H}$. □

Combining Propositions 2 and 3, we obtain Theorem 3.

**Proof of Theorem 4.** To see that TempRep does indeed work, we begin with the following well-known and easily proved lemma.

**Lemma 2.** *Let $D$ be a digraph that contains no directed cycle. Then there exists a node of $D$ whose indegree is zero.*

To prove part (i) of Theorem 4, suppose that $\mathcal{H}$ has a temporal representation. Then, by Theorem 3, $D_{\mathcal{H}}$ has no directed cycles. By Lemma 2, this implies that every subdigraph obtained from $D_{\mathcal{H}}$ by deleting nodes (and their incident arcs) contains at least one node of indegree zero. It now follows that TempRep applied to $\mathcal{H}$ returns a map $f : V \rightarrow \mathbb{N}$. To see that $f$ is a temporal labelling of $\mathcal{H}$, define $g : [V] \rightarrow \mathbb{N}$ by setting $g([v]) = S_i$, where $[v] \in S_i$. Because of the way in which $S_0, S_1, S_2, \ldots$ are constructed, $g$ is an acyclic ordering of the nodes of $D_{\mathcal{H}}$. Therefore, by Proposition 3, the map $f$ is a temporal labelling of $\mathcal{H}$.

For the proof of part (ii) of Theorem 4 suppose that $\mathcal{H}$ has no temporal representation. Then, by Theorem 3, $D_{\mathcal{H}}$ contains a directed cycle. Let $\{[v_1], [v_2], \ldots, [v_k]\}$ be the nodes in this cycle, where we may assume that $([v_j], [v_{j+1}])$ for all $j$ and $([v_k], [v_1])$ are arcs of this cycle. It is now easily seen that beginning with $D_{\mathcal{H}}$, and selecting and removing only nodes with indegree zero none of the nodes in this cycle can ever be removed. Thus at some iteration $i$ of TempRep when applied to $\mathcal{H}$, no node of $D_i$ has indegree zero, in which case TempRep halts and returns $\mathcal{H}$ *has no temporal representation.* This completes the proof of (ii).

We leave the straightforward check that the running time of TempRep applied to $\mathcal{H}$ is quadratic in the size of the node set of $\mathcal{H}$ to the reader.

**Outline of an algorithm to output all temporal labellings of a hybrid phylogeny, up to order isomorphism.** By Proposition 2, each temporal labelling of $\mathcal{H}$ induces an acyclic ordering of the node set $[V]$ of $D_{\mathcal{H}}$. Conversely, by Proposition 3, each acyclic ordering of $[V]$ induces a temporal labelling of $\mathcal{H}$. It follows that if $\mathcal{H}$ has a temporal representation, then all temporal labellings of $\mathcal{H}$ can be found by finding all acyclic orderings of $[V]$. Using the first part of the proof of Theorem 3, it is easily checked that all such orderings can be obtained by considering all possible ways of reducing $D_{\mathcal{H}}$ to the empty graph by sequentially selecting and then deleting subsets of nodes of indegree zero. Since it is only the relative ordering of the nodes of $\mathcal{H}$ that are of interest, it follows that it is only the order in which these subsets are chosen that is important. Each possible way of reducing $D_{\mathcal{H}}$ to the empty graph gives rise to a unique sequence of chosen subsets of nodes of $D_{\mathcal{H}}$. In TempRep, this corresponds to all possible choices for the sequence $S_0, S_1, S_2, \ldots$. Furthermore, each such sequence induces, up to ordering isomorphism, a unique temporal labelling of $\mathcal{H}$. Hence to list, up to ordering isomorphism, all temporal labellings of $\mathcal{H}$ one simply needs to systematically find all possible choices for selecting $S_0, S_1, S_2, \ldots$ in TempRep.

□