Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Mihaela Baroni · Stefan Grünewald · Vincent Moulton · Charles Semple

Bounding the Number of Hybridisation Events for a Consistent Evolutionary History

Received: date / Revised version: date –

Abstract. Evolutionary processes such as hybridisation, lateral gene transfer, and recombination are all key factors in shaping the structure of genes and genomes. However, since such processes are not always best represented by trees, there is now considerable interest in using more general networks instead. For example, in recent studies it has been shown that networks can be used to provide lower bounds on the number of recombination events and also for the number of lateral gene transfers that took place in the evolutionary history of a set of molecular sequences. In this paper we describe the theoretical performance of some related bounds that result when merging pairs of trees into networks.

1. Introduction

Although phylogenetic trees have proven an invaluable tool in studying evolution [5], it is now well-accepted that in certain circumstances they may not provide an appropriate representation of the evolutionary history of organisms (see e.g. [12]). For example, hybridisation, lateral gene transfer, and recombination all constitute important evolutionary processes that may not be best represented using trees. As a result there has been increasing interest in how to represent such processes using phylogenetic networks (cf. e.g. [4,6,11,16]). In this setting, we expect that the following concept introduced in [2] will prove useful (see also [7,9,11,16,17] where related concepts are presented).

A *hybrid phylogeny* or *hybrid* $\mathcal{H}$ (*on a finite set* X) is an ordered pair $(D;\phi)$ consisting of a rooted acyclic digraph $D = (V, A)$ and a bijective map ϕ from X into the set of vertices of V with out-degree zero such that

Mihaela Baroni, Stefan Grünewald, Charles Semple: Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, e-mail: mihaela.baroni@ugal.ro, s.grunewald@math.canterbury.ac.nz, c.semple@math.canterbury.ac.nz

Vincent Moulton: School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK, e-mail: vincent.moulton@cmp.uea.ac.uk

Key words: phylogenetic tree – phylogentic network – hybrid network – recombination – lateral gene transfer – hybridisation

Mathematics Subject Classification (2000): 92D15, 05C05, 05C20

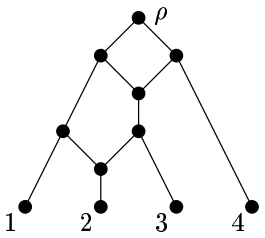

Fig. 1. A hybrid $\mathcal{H}$ with label set $X = \{1, 2, 3, 4\}$ and root ρ. This hybrid has hybridisation number 2. Since we will always draw hybrids with arcs directed downwards away from the root, in this and all subsequent figures we omit the arrow heads of the arcs.

the root has out-degree at least two and, for all vertices $v \in V - \phi(X)$ with $d^-(v) = 1$, we have $d^+(v) \geq 2$ (cf. Fig. 1). As usual, $d^-(v)$ and $d^+(v)$ denote the in-degree and the out-degree of v, respectively. If $|X| = 1$, then the digraph consisting of a single-root vertex v together with the mapping from X into $\{v\}$ is also defined to be a hybrid phylogeny on X. The set X represents a collection of organisms and is called the *label set* of $\mathcal{H}$. Vertices of in-degree at least two are called *hybridisation* vertices. These vertices represent an exchange of genetic information between organisms that we will generically call *hybridisation events*. For a hybrid $\mathcal{H}$ on X with root ρ, the *hybridisation number* of $\mathcal{H}$, denoted $h(\mathcal{H})$, is the value

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

One of the main problems when studying the evolution of certain organisms is to estimate the number of hybridisation events that took place in their past, and much work has been done in this direction for recombination events (see e.g. [7,10,15,17]), and also some for lateral gene transfer events [9,11]. Indeed, it is well-known that bounds on the number of recombination events can be obtained by merging trees into networks [14,16]. In this paper, our main goal is to better understand related bounds for hybridisation numbers that can be computed using pairs of trees.

To describe our main result, we first require some more definitions. Given a set X as above, a *rooted binary phylogenetic X-tree* $\mathcal{T}$ is a rooted tree whose root has degree two and all other interior vertices have degree three, and whose leaf set is X (cf. [13]). If $|X| = 1$, then the rooted tree consisting of a single-root vertex labelled by the element in X is a rooted binary phylogenetic X-tree. The set X is called the *label set* of $\mathcal{T}$ and it is sometimes denoted by $\mathcal{L}(\mathcal{T})$. For technical reasons, given a rooted tree $\mathcal{T}$, it is useful to define the *planted tree* $\mathcal{P}(\mathcal{T})$, which is the tree obtained by adding an additional edge to $\mathcal{T}$, called the *root edge*, that contains the root of $\mathcal{T}$ together with an extra vertex (see Fig. 3 for an example).

Now, let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $\mathcal{H}$ be a hybrid such that the label set of $\mathcal{H}$ contains X. Then $\mathcal{H}$ *displays* $\mathcal{T}$ if $\mathcal{T}$

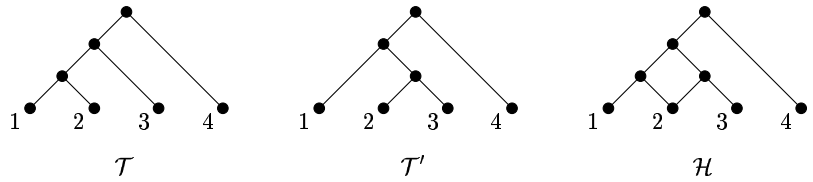

Fig. 2. A hybrid phylogeny $\mathcal{H}$ that displays the trees $\mathcal{T}$ and $\mathcal{T}'$.

is a refinement of a rooted subtree of $\mathcal{H}$. Moreover, for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, let

$$h(\mathcal{T}, \mathcal{T}') = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybrid on } X \text{ that displays } \mathcal{T}, \mathcal{T}'\}.$$

For example, in Figure 2, a pair of trees $\mathcal{T}, \mathcal{T}'$ is presented with $h(\mathcal{T}, \mathcal{T}') = 1$, and a hybrid phylogeny $\mathcal{H}$ that displays both of these trees having hybridisation number 1. The number $h(\mathcal{T}, \mathcal{T}')$ should be regarded as an estimate of the minimum number of hybridisation events required to preserve all of the ancestral relationships described by $\mathcal{T}$ and $\mathcal{T}'$. Note that if $\mathcal{T}$ is isomorphic to $\mathcal{T}'$, then $h(\mathcal{T}, \mathcal{T}') = 0$ as $\mathcal{T}$ itself is a hybrid that displays both $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, if $\mathcal{T}$ is not isomorphic to $\mathcal{T}'$, then $h(\mathcal{T}, \mathcal{T}') \geq 1$.

Pairs of trees $\mathcal{T}, \mathcal{T}'$ for which $h(\mathcal{T}, \mathcal{T}') = 1$ are strongly related. In particular, it is not difficult to show that $h(\mathcal{T}, \mathcal{T}') = 1$ if and only if the trees $\mathcal{T}, \mathcal{T}'$ differ by a single (non-trivial) rSPR operation, a tree rearrangement operation whose definition we now recall.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $e = \{u, v\}$ be an edge of $\mathcal{T}$, where u is the vertex that is in the path from the root of $\mathcal{T}$ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic tree obtained from $\mathcal{P}(\mathcal{T})$ by deleting e and then adjoining a new edge f between v and the component C_u that contains u as follows. Create a new vertex u' which subdivides an edge in C_u, and adjoin f between u' and v. Then suppress the degree-two vertex u and delete the root edge. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted subtree prune and regraft* (rSPR) operation. For example, in Fig. 2, the tree $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by performing a single rSPR operation (namely, the operation that cuts off the edge incident with the leaf labelled by 2 and reattaches it to the edge incident with the leaf that is labelled by 3). If $\mathcal{T}$ and $\mathcal{T}'$ are two arbitrary rooted binary phylogenetic X-trees, then it is always possible to transform $\mathcal{T}$ into $\mathcal{T}'$ by a sequence of rooted subtree prune and regraft operations. Moreover, the minimum number of rooted subtree prune and regraft operations required to transform a tree $\mathcal{T}$ into another tree $\mathcal{T}'$ is a metric on the set of binary phylogenetic X-trees (see [1]). This number is called the rSPR *distance* and is denoted by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$.

The rSPR distance has been used to provide lower bounds for the number of recombination events that took place in the evolutionary history of a set of DNA sequences [14,16]. However, as pointed out in [16, Section 3.4] the rSPR distance can underestimate the number of events. In the terminology that we introduce above, this translates to the fact that the rSPR distance

between two trees $\mathcal{T}$ and $\mathcal{T}'$ can underestimate $h(\mathcal{T},\mathcal{T}')$. In this paper, we will prove the following theorem that provides a more precise relationship between $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$ and $h(\mathcal{T},\mathcal{T}')$.

Theorem 1. *Let $|X| = n \geq 2$, and let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') \leq h(\mathcal{T},\mathcal{T}') \leq n-2. \tag{1}$$

Moreover, the bounds on $h(\mathcal{T},\mathcal{T}')$ in (1) are sharp and, for all $n \geq 4$, there are trees $\mathcal{T}$ and $\mathcal{T}'$ with

$$h(\mathcal{T},\mathcal{T}') - d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = n - 2\lfloor\sqrt{n}\rfloor - c$$

where $c = 0$ if n is a square, $c = 1$ if $1 \leq n - \lfloor\sqrt{n}\rfloor^2 < \sqrt{n}$, and $c = 2$ otherwise.

As can be seen from Theorem 1, $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$ can greatly underestimate $h(\mathcal{T},\mathcal{T}')$. Thus in future it will be of importance to find better lower bounds for $h(\mathcal{T},\mathcal{T}')$.

The rest of the paper is organised as follows. In Section 2 we introduce our main tool, maximum agreement forests, and prove a key theorem (Theorem 2) that characterises the hybridisation number in terms of this concept. Using this theorem, in Sections 3 and 4 we prove two results (Theorem 3 and Corollary 1) from which Theorem 1 immediately follows.

2. Maximum Agreement Forests

In establishing Theorem 1, we also characterise $h(\mathcal{T},\mathcal{T}')$ for a pair $\mathcal{T},\mathcal{T}'$ of rooted binary phylogenetic X-trees in terms of a particular type of "agreement forest". Agreement forests are an invaluable tool for analysing and understanding tree rearrangment operations such as rooted subtree prune and regraft as shown in [1,3]. They were introduced in a slightly different form in [8].

To state this characterization, we first need some additional definitions. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let X' be a non-empty subset of X. We denote the rooted minimal subtree of $\mathcal{T}$ that connects the vertices in X' by $\mathcal{T}(X')$. Furthermore, the *restriction* of $\mathcal{T}$ to X', denoted $\mathcal{T}|X'$, is obtained from $\mathcal{T}(X')$ by suppressing any degree-two vertices apart from the root. Observe that, as $\mathcal{T}$ is a binary phylogenetic tree, $\mathcal{T}|X'$ is also a binary phylogenetic tree.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. We define an *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ as follows. For the purposes of the definition, we label the root of both $\mathcal{P}(\mathcal{T})$ and $\mathcal{P}(\mathcal{T}')$ by ρ (see Fig. 3 for an example). Furthermore, in addition to the elements in X, we also view ρ as an element of the label set of both $\mathcal{P}(\mathcal{T})$ and $\mathcal{P}(\mathcal{T}')$. An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\{\mathcal{T}_\rho,\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is either the tree containing ρ as an isolated vertex, or a planted rooted tree whose label set includes ρ, and $\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_k$ are rooted binary phylogenetic trees such that the following hold:

(i) The label sets $\mathcal{L}(\mathcal{T}_\rho), \mathcal{L}(\mathcal{T}_1), \mathcal{L}(\mathcal{T}_2), \ldots, \mathcal{L}(\mathcal{T}_k)$ partition $X \cup \{\rho\}$.
(ii) For all $i \in \{\rho, 1, 2, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}(\mathcal{T}_i) \cong \mathcal{T}'|\mathcal{L}(\mathcal{T}_i)$.
(iii) The trees in $\{\mathcal{T}(\mathcal{L}(\mathcal{T}_i)) : i \in \{\rho, 1, 2, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}(\mathcal{T}_i)) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are vertex disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

A *maximum agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is an agreement forest in which the number of components over all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$ is minimised. If $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ is a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, we denote this value for k by $m(\mathcal{T}, \mathcal{T}')$. Observe that if, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical, then $m(\mathcal{T}, \mathcal{T}') = 0$. Moreover, it is shown in [3] that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$. To illustrate these definitions, Fig. 4 shows a maximum agreement forest $\mathcal{F}_1$ for the two rooted binary phylogenetic trees shown in Fig. 3.

We now introduce a particular type of agreement forest that will play an essential role in this paper. Again, let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. 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 vertex set is $\mathcal{F}$ and for which $(\mathcal{T}_i, \mathcal{T}_j)$ is an arc precisely if $i \neq j$ and either

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

Note that, as $\mathcal{F}$ is an agreement forest, the roots of $\mathcal{T}(\mathcal{L}(\mathcal{T}_i))$ and $\mathcal{T}(\mathcal{L}(\mathcal{T}_j))$, and the roots of $\mathcal{T}'(\mathcal{L}(\mathcal{T}_i))$ and $\mathcal{T}'(\mathcal{L}(\mathcal{T}_j))$ are not the same. We call $\mathcal{F}$ a *good* agreement forest if $G_\mathcal{F}$ does not contain a directed cycle. A *maximum good agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a good agreement forest in which the number of components over all good agreement forests $\mathcal{T}$ and $\mathcal{T}'$ is minimised. This number minus one is denoted by $m_g(\mathcal{T}, \mathcal{T}')$. An example of a maximum good agreement forest $\mathcal{F}_2$ for the two trees shown in Fig. 3 is shown in Fig. 4. Observe that, in this figure, $\mathcal{F}_1$ is not a good agreement forest for these two trees. Since every good agreement forest is an agreement forest, $m(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}')$. As shown by the example illustrated in Fig.s 3 and 4, this inequality may be strict.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree with vertex set V. For all $v \in V$, let $c(v)$ denote the subset of X that consists of those elements x for which there is a directed path in $\mathcal{T}$ from v to x.

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

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

Proof. First we prove that $h(\mathcal{T}, \mathcal{T}') \geq m_g(\mathcal{T}, \mathcal{T}')$. For convenience in this part of the proof, we will rewrite $\mathcal{T}$ and $\mathcal{T}'$ as $\mathcal{T}_1$ and $\mathcal{T}_2$. Let $\mathcal{H}$ be a hybrid that displays both $\mathcal{T}_1$ and $\mathcal{T}_2$, and has the property that $h(\mathcal{H})$ is minimised. Because of this minimality, it is easily seen that, for each vertex of $\mathcal{H}$, the number of incoming arcs is at most two. Let $\mathcal{F}$ be the forest obtained from

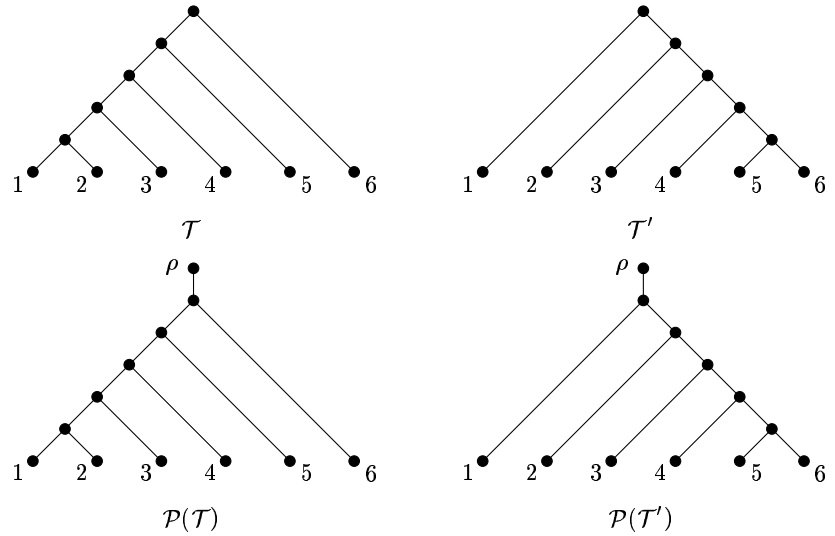

Fig. 3. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ without (above) and with (below) their roots labelled.

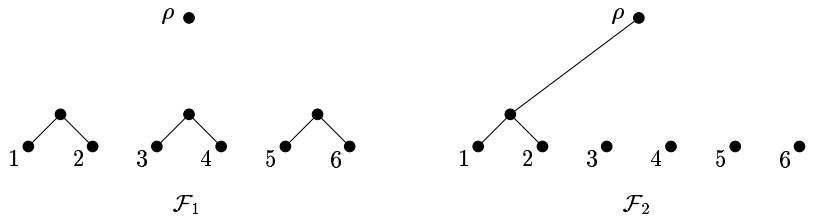

Fig. 4. A maximum agreement forest $\mathcal{F}_1$ for $\mathcal{T}$ and $\mathcal{T}'$, and a maximum good agreement forest $\mathcal{F}_2$ for $\mathcal{T}$ and $\mathcal{T}'$.

$\mathcal{H}$ by deleting, for each hybridisation vertex u of $\mathcal{H}$, the two incoming arcs and then suppressing any resulting degree-two vertex. We show by induction on $h(\mathcal{H})$ that $\mathcal{F}$ is a good agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$ with $h(\mathcal{H})+1$ components, thus showing that $h(\mathcal{T}_1, \mathcal{T}_2) \geq m_g(\mathcal{T}_1, \mathcal{T}_2)$.

If $h(\mathcal{H}) = 0$, then $\mathcal{T}_1$ and $\mathcal{T}_2$ are, up to isomorphism, identical and so the result clearly holds. Now let $h(\mathcal{H}) = n \geq 1$ and assume that the result holds for all pairs of rooted binary phylogenetic X-trees in which $h(\mathcal{H})$ is at most $n-1$. Let v be a hybrid vertex of $\mathcal{H}$ in which $\mathcal{H}|c(v)$ is a rooted binary phylogenetic tree $\mathcal{T}_v$ on $c(v)$. It is easily seen that there must exist such a vertex. Observe that $\mathcal{T}_v$ is a rooted binary phylogenetic subtree of $\mathcal{T}_1$ and $\mathcal{T}_2$, and so one of the arcs coming in to v, e_1 say, is used by $\mathcal{H}$ to display $\mathcal{T}_1$ and the other arc coming in to v, e_2 say, is used by $\mathcal{H}$ to display $\mathcal{T}_2$. We next define a hybrid $\mathcal{H}'$ on X, and two rooted binary phylogenetic X-trees $\mathcal{T}_1'$ and $\mathcal{T}_2'$.

Viewing the root ρ of $\mathcal{H}$ as a vertex at the end of a pendant edge adjoined to the orginal root, $\mathcal{H}'$ is obtained from $\mathcal{H}$ by deleting e_1 and e_2, and then

adjoining the root of $\mathcal{T}_v$ to ρ. Similarly, for each i, viewing the root ρ_i of $\mathcal{T}_i$ as vertex at the end of a pendant edge adjoined to the original root, $\mathcal{T}_i'$ is obtained from $\mathcal{T}_i$ by pruning the rooted subtree T_v and then adjoining the root of this subtree to ρ with a new edge. Evidently, $h(\mathcal{H}') = h(\mathcal{H}) - 1 = n-1$. Moreover, as $\mathcal{H}$ displays both $\mathcal{T}_1$ and $\mathcal{T}_2$, it is also clear by construction that $\mathcal{H}'$ displays both $\mathcal{T}_1'$ and $\mathcal{T}_2'$. Therefore, by the induction assumption, the forest $\mathcal{F}'$ obtained from $\mathcal{H}'$ by deleting, for each hybrid vertex u of $\mathcal{H}'$, the two incoming arcs and then suppressing any resulting degree-two vertex is a good agreement forest for $\mathcal{T}_1'$ and $\mathcal{T}_2'$ with n components.

Consider $\mathcal{F}'$ and, in particular, the component $\mathcal{T}_\rho$ of $\mathcal{F}'$ that contains ρ (the root label of $\mathcal{T}_1'$ and $\mathcal{T}_2'$). By the maximality of $\mathcal{F}'$, the label set of $\mathcal{T}_\rho$ contains $c(v)$. This is easily seen by observing that at most one tree in $\mathcal{F}'$ has the property that its label set contains elements of $c(v)$ and $X - c(v)$, in which case this tree is $\mathcal{T}_\rho$. Now let $\mathcal{F}$ be the forest obtained from $\mathcal{F}'$ by deleting the edge joining ρ to the root of $\mathcal{T}_v$. Since $\mathcal{F}'$ is a good agreement forest for $\mathcal{T}_1'$ and $\mathcal{T}_2'$, $\mathcal{F}$ is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$. Furthermore, as $\mathcal{T}_1$ and $\mathcal{T}_2$ both contain $\mathcal{T}_v$ as a rooted subtree, it follows that $\mathcal{F}$ is also a good agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_2$. Since $\mathcal{F}$ has $n+1$ components, we deduce that $h(\mathcal{T}_1, \mathcal{T}_2) \geq m_g(\mathcal{T}_1, \mathcal{T}_2)$.

We now show that $h(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}')$. The proof is by induction on $m_g(\mathcal{T}, \mathcal{T}')$. If $m_g(\mathcal{T}, \mathcal{T}') = 0$, then, up to isomorphism, $\mathcal{T}$ and $\mathcal{T}'$ are identical and so $h(\mathcal{T}, \mathcal{T}') \leq m_g(\mathcal{T}, \mathcal{T}')$. Now let $m_g(\mathcal{T}, \mathcal{T}') = k$ and assume the result holds for all pairs of rooted binary phylogenetic X-trees in which the minimum number of components over all good agreement forests is at most k.

Let $\mathcal{F} = \{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Since $\mathcal{F}$ is good, $G_\mathcal{F}$ has no directed cycles and so there is a vertex of $G_\mathcal{F}$ whose out-degree is zero. Without loss of generality we may assume that this vertex is $\mathcal{T}_k$. Note that the vertex $\mathcal{T}_\rho$ in $G_\mathcal{F}$ is the unique vertex that has in-degree zero. Since $\mathcal{T}_k$ has out-degree zero in $G_\mathcal{F}$, it follows that $\mathcal{T}_k$ is a rooted subtree of both $\mathcal{T}$ and $\mathcal{T}'$. Let $X_k = X - \mathcal{L}(\mathcal{T}_k)$ and let $\mathcal{F}_k = \mathcal{F} - \{\mathcal{T}_k\}$. Then it is easily checked that $\mathcal{F}_k$ is a good agreement forest for $\mathcal{T}|X_k$ and $\mathcal{T}'|X_k$. Since $|\mathcal{F}_k| < |\mathcal{F}|$, it now follows by the induction assumption that there is a hybrid $\mathcal{H}_k$ on X that displays both $\mathcal{T}|X_k$ and $\mathcal{T}'|X_k$, and has the property that $h(\mathcal{H}_k) \leq k-1$.

Now let $\mathcal{H}$ be the hybrid on X that is obtained from $\mathcal{H}_k$ as follows. Since $\mathcal{H}_k$ displays $\mathcal{T}|X_k$ and since $\mathcal{T}_k$ is a rooted subtree of $\mathcal{T}$, it is easily seen that there is a hybrid that can be obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ with a new edge e that connects the root of $\mathcal{T}_k$ and a new vertex that subdivides an edge of $\mathcal{H}_k$. Similarly, there is a hybrid that can be obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ using a new edge e' that displays $\mathcal{T}'|X_k$. Let $\mathcal{H}$ be the hybrid obtained from $\mathcal{H}_k$ by adjoining $\mathcal{T}_k$ using exactly the edges e and e'. Evidently, $\mathcal{H}$ displays both $\mathcal{T}$ and $\mathcal{T}'$. Furthermore, as $\mathcal{T}_k$ is a rooted binary phylogenetic tree and the vertex of $\mathcal{H}$ corresponding to the root of $\mathcal{T}_k$ has in-degree two, $h(\mathcal{H}) \leq k$. Hence $h(\mathcal{T}, \mathcal{T}') \leq k = m_g(\mathcal{T}, \mathcal{T}')$. This completes the proof of the theorem. □

3. Lower and Upper Bounds for $h(\mathcal{T}, \mathcal{T}')$

In this section, we establish lower and upper bounds on the value of $h(\mathcal{T}, \mathcal{T}')$ for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$. In particular, we prove the following result which is a restatement of the first part of Theorem 1.

Theorem 3. *Let $|X| = n$, and let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then, for all $n \geq 2$,*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq h(\mathcal{T}, \mathcal{T}') \leq n - 2. \tag{2}$$

Moreover, the bounds on $h(\mathcal{T}, \mathcal{T}')$ in (2) are sharp.

The proof of Theorem 3 is an immediate consequence of Theorem 4 and Propositions 1, 2, and 3. The first of these results is established in [3]. Since a good agreement forest for two rooted binary phylogenetic trees is also an agreement forest for the same pair of trees, Theorem 4 in combination with Theorem 2 establishes the lower bound in Theorem 3.

Theorem 4. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}').$$

The next result establishes the upper bound in Theorem 3.

Proposition 1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $|X| = n$, and suppose that $n \geq 2$. Then*

$$h(\mathcal{T}, \mathcal{T}') \leq n - 2.$$

Proof. To prove the proposition, it suffices by Theorem 2 to construct a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with $n - 1$ components. Let $X = \{x_1, x_2, \ldots, x_n\}$. Let $\mathcal{T}_\rho$ be the restriction of $\mathcal{T}$ to the set $\{\rho, x_{n-1}, x_n\}$ and, for all $i \in \{1, 2, \ldots, n-2\}$, let $\mathcal{T}_i$ be the rooted phylogenetic tree consisting of a single vertex labelled by x_i. Since $\mathcal{T}_\rho \cong \mathcal{T}'|\{\rho, x_{n-1}, x_n\}$, it follows that

$$\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{n-2}\}$$

is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Moreover, this forest has exactly $n - 1$ components. □

The next two propositions show that the bounds in Theorem 3 are sharp.

Proposition 2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$ if and only if $h(\mathcal{T}, \mathcal{T}') = 1$.*

Proof. The sufficient part is established in [2]. The necessary part is straightforward and omitted. □

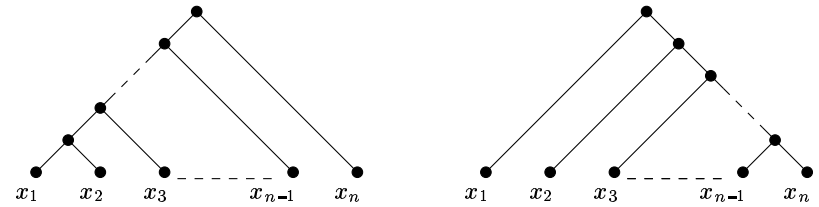

Fig. 5. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on n leaves with $h(\mathcal{T}, \mathcal{T}') = n - 2$.

For the next proposition, we make use of the following lemma.

Lemma 1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $\mathcal{T}_\rho$ be the rooted tree in a maximum good agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ whose label set contains ρ. Then $\mathcal{L}(\mathcal{T}_\rho) \cap X$ is non-empty.*

Proof. Suppose that $\mathcal{L}(\mathcal{T}_\rho) \cap X$ is empty. Since $\mathcal{F}$ is a good agreement forest, there is a vertex, $\mathcal{T}_1$ say, of $G_\mathcal{F} \backslash \mathcal{T}_\rho$ whose in-degree is zero. It now follows that the forest obtained from $\mathcal{F}$ by adding an arc directed from $\mathcal{T}_\rho$ to the root of $\mathcal{T}_1$ is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with one less component. This contradicts the maximality of $\mathcal{F}$. □

Proposition 3. *For all $n \geq 2$, there exist two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ with $|X| = n$ such that $h(\mathcal{T}, \mathcal{T}') = n - 2$.*

Proof. Let $X = \{x_1, x_2, \ldots, x_n\}$, and let $\mathcal{T}$ and $\mathcal{T}'$ be the two rooted binary phylogenetic trees shown in Fig. 5. To establish the proposition, it suffices to show that $m_g(\mathcal{T}, \mathcal{T}') = n - 2$.

Suppose that $\mathcal{F}$ is a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{T}_\rho$ denote the tree in $\mathcal{F}$ that has ρ as a vertex label. We begin by first making the following elementary, but important observations about $\mathcal{F}$:

(i) No tree in $\mathcal{F}$ has a label set that contains at least three elements of X; for otherwise, such a tree is not a restriction of either $\mathcal{T}$ or $\mathcal{T}'$.

(ii) At most one tree in $\mathcal{F}$ has the property that its label set contains two elements of X. To see this, suppose that there are two such trees $\mathcal{T}_i$ and $\mathcal{T}_j$ in $\mathcal{F}$. Since $\mathcal{T}_i$ and $\mathcal{T}_j$ are vertex disjoint subtrees of both $\mathcal{T}$ and $\mathcal{T}'$, it is easily seen that neither $\mathcal{L}(\mathcal{T}_i)$ nor $\mathcal{L}(\mathcal{T}_j)$ contains ρ. This implies that $\mathcal{L}(\mathcal{T}_i) = \{x_i, x_j\}$ and $\mathcal{L}(\mathcal{T}_j) = \{x_k, x_l\}$ with $i < j < k < l$. But then, by considering the vertices of $\mathcal{T}$ and $\mathcal{T}'$ corresponding to the roots of $\mathcal{T}_i$ and $\mathcal{T}_j$, we deduce that $G_\mathcal{F}$ contains a directed cycle. Thus $\mathcal{F}$ does indeed satisfy (ii).

(iii) By Lemma 1, the label set of $\mathcal{T}_\rho$ contains at least one element of X.

Combining (i)—(iii), we deduce that in $\mathcal{F}$ either the label set of $\mathcal{T}_\rho$ contains two elements of X and the label sets of the remaining trees in $\mathcal{F}$ are singletons, or the label set of $\mathcal{T}_\rho$ contains exactly one element of X and there

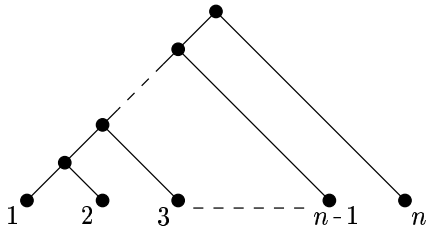

Fig. 6. The rooted caterpillar tree on $(1, 2, \ldots, n)$.

is one other tree whose label set is of size two, while the label sets of the remaining trees in $\mathcal{F}$ are singletons. In both cases, $\mathcal{F}$ has precisely $n-1$ components and so $m_g(\mathcal{T}, \mathcal{T}') = n-2$. □

4. Bounds on the Difference Between $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ and $h(\mathcal{T}, \mathcal{T}')$

In this section, we establish the second part of Theorem 1. In particular, we will show that, for all X with $|X| \geq 4$, there exist rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ such that the difference between $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ and $h(\mathcal{T}, \mathcal{T}')$ is large relative to the size of X.

For all $n \geq 4$, let $\mathcal{T}$ be the rooted binary phylogenetic tree with label set $\{1, 2, \ldots, n\}$ as shown in Fig. 6. This tree is an example of a rooted caterpillar tree. For our purposes the ordering of the labels are important, thus we will call $\mathcal{T}$ the rooted caterpillar on $(1, 2, \ldots, n)$. For a permutation π of $\{1, 2, \ldots, n\}$, let $\mathcal{T}_\pi$ be the rooted caterpillar tree on $(\pi(1), \pi(2), \ldots, \pi(n))$.

Now let l be a natural number with $2 \leq l \leq \lceil \frac{n}{2} \rceil$, and let a and b be the unique non-negative integers such that $n = al + b$ and $b \leq l - 1$. It is straightforward to deduce that

$$n = (l-b)\left\lfloor \frac{n}{l} \right\rfloor + b\left(\left\lfloor \frac{n}{l} \right\rfloor + 1\right).$$

Set $g = \lfloor \frac{n}{l} \rfloor$ and, for $1 \leq j \leq l$, set

$$I_j = \begin{cases} ((j-1)g+1, \ldots, jg), & \text{if } 1 \leq j \leq l-b; \\ ((j-1)(g+1)-(l-b), \ldots, jg+j-(l-b)), & \text{if } l-b+1 \leq j \leq l. \end{cases}$$

For all j, we say that I_j is an *interval*. Furthermore, define π_l to be the permutation of $\{1, 2, \ldots, n\}$ that cuts $(1, 2, \ldots, n)$ into pieces corresponding to the intervals I_j and then rearranges the pieces in the opposite order. In other words, π_l is obtained from $(I_l, I_{l-1}, \ldots, I_1)$ by removing the brackets around each interval. As an illustration, the rooted binary phylogenetic tree shown in Fig. 7 is the rooted caterpillar tree $\mathcal{T}_{\pi_3}$ for $n = 8$.

Theorem 5. *Let $\mathcal{T}$ be the rooted caterpillar tree on $(1, 2, \ldots, n)$. Then, for all $n \geq 4$, we have $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_{\pi_l}) = l$ and $h(\mathcal{T}, \mathcal{T}_{\pi_l}) = n - \lceil \frac{n}{l} \rceil$.*

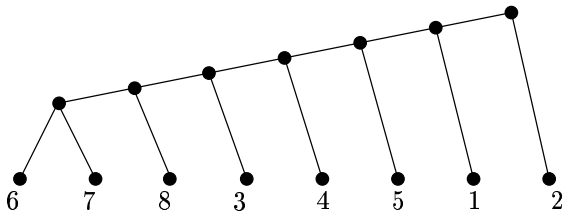

Fig. 7. The rooted binary phylogenetic tree $\mathcal{T}_{\pi_3}$ for $n = 8$.

Proof. Since, for $1 \leq j \leq l$, the restrictions of $\mathcal{T}$ to I_j, together with an isolated vertex labelled ρ form an agreement forest for $\mathcal{T}$ and $\mathcal{T}_{\pi_l}$, we have $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_{\pi_l}) \leq l$. Furthermore, $h(\mathcal{T}, \mathcal{T}_{\pi_l}) \leq n - \lceil \frac{n}{l} \rceil$ since the restriction of the planted tree $\mathcal{P}(\mathcal{T})$ to $I_l \cup \{\rho\}$ together with each of the remaining elements in $\{1, 2, \ldots, n\}$ labelling isolated vertices is a good agreement forest for $\mathcal{T}$ and $\mathcal{T}_{\pi_l}$.

We now establish the reverse inequalities. Firstly, assume that $\mathcal{F}$ is a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}_{\pi_l}$ with at most l trees. If the label set of the tree $\mathcal{T}_\rho$ of $\mathcal{F}$ containing ρ also contains an element $x \in I_r$ for some $r \in \{1, 2, ..., l\}$, then all elements not contained in I_r must label an isolated vertex in $\mathcal{F}$. Since I_r contains at most $\lceil \frac{n}{l} \rceil$ elements, $\mathcal{F}$ has at least $n - \lceil \frac{n}{l} \rceil + 1$ components. But $2 \leq l \leq \lceil \frac{n}{2} \rceil$ and $n \geq 4$, and so $n - \lceil \frac{n}{l} \rceil + 1 \geq l + 1$; a contradiction. Thus the tree in $\mathcal{F}$ whose label set contains ρ consists of an isolated vertex.

Since $\mathcal{F}$ contains at most l trees, there must be a tree $\mathcal{T}_1$ say in $\mathcal{F}$ whose label set contains an element $x_i \in I_i$ and an element $x_j \in I_j$ for different intervals I_i and I_j. Without loss of generality, we may assume that $i < j$. It follows by construction that, for all $x \in I_m$ and for all $i \leq m \leq j$ with $x \notin \{x_i, x_j\}$, the tree in $\mathcal{F}$ whose label set contains x consists of an isolated vertex. But then the forest $\mathcal{F}'$ obtained from $\mathcal{F}$ by replacing those isolated vertices and $\mathcal{T}_1$ with the restrictions of $\mathcal{T}$ to I_m for $i \leq m \leq j$ is an agreement forest. Moreover, this forest contains strictly less trees than $\mathcal{F}$ provided $|I_i|, |I_j| \geq 2$; a contradiction to the maximality of $\mathcal{F}$. In the remaining case, $|I_i| = 1$, it is easily seen that there is another distinct pair of intervals $I_{i'}$ and $I_{j'}$ in which there is a tree in $\mathcal{F}$ whose label set contains an element of each. Applying the same argument, we again deduce a contradiction. It follows that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_{\pi_l}) \leq l$ and so $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_{\pi_l}) = l$

To show that $h(\mathcal{T}, \mathcal{T}_{\pi_l}) \geq n - \lceil \frac{n}{l} \rceil$, assume that $\mathcal{F}_g$ is a maximum good agreement forest for $\mathcal{T}$ and $\mathcal{T}_{\pi_l}$ with at most $n - \lceil \frac{n}{l} \rceil$ components. If there is a tree in $\mathcal{F}_g$ whose label set contains elements x_i and x_j such that x_i and x_j are in distinct intervals, then all remaining elements of X must label an isolated vertex in $\mathcal{F}_g$; otherwise $\mathcal{F}_g$ is not good. This implies that $\mathcal{F}_g$ contains at least $n - 1$ trees; a contradiction. Furthermore, as $\mathcal{F}_g$ is a good agreement forest, there is no distinct pair $\mathcal{T}_i$ and $\mathcal{T}_j$ of trees in $\mathcal{F}_g$ such that $|\mathcal{L}(\mathcal{T}_i) \cap I_i| \geq 2$ and $|\mathcal{L}(\mathcal{T}_j) \cap I_j| \geq 2$ for some distinct intervals I_i and I_j. Hence, except for one interval, I_k say, all elements in $X - I_k$ label isolated

vertices in $\mathcal{F}_g$. Since I_k contains at most $\lceil \frac{n}{l} \rceil$ elements, $\mathcal{F}_g$ contains at least $n - \lceil \frac{n}{l} \rceil + 1$ components. This contradiction establishes the inequality and thus the theorem. □

Corollary 1. *For every natural number $n \geq 4$, there are trees $\mathcal{T}$, $\mathcal{T}_1$, $\mathcal{T}_2$ with n leaves such that*

$$\frac{h(\mathcal{T}, \mathcal{T}_1)}{d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_1)} = \frac{1}{2} \left\lfloor \frac{n}{2} \right\rfloor$$

and

$$h(\mathcal{T}, \mathcal{T}_2) - d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}_2) = n - 2\lfloor \sqrt{n} \rfloor - c$$

where $c = 0$ if n is a square, $c = 1$ if $1 \leq n - \lfloor \sqrt{n} \rfloor^2 < \sqrt{n}$ and $c = 2$ else.

Proof. The result follows from Theorem 5 with $\mathcal{T}_1 = \mathcal{T}_{\pi_2}$ and $\mathcal{T}_2 = \mathcal{T}_{\pi_{\lfloor \sqrt{n} \rfloor}}$. □

We conclude by remarking that, though the maximal difference between h and d_{rSPR} can be large for large n, they are equal for $n \leq 5$.

Acknowledgements. All authors thank the Swedish Foundation for International Cooperation in Research and Education (STINT). The second author, and the first and fourth authors also thank the New Zealand Institue of Mathematics and its Applications and the New Zealand Marsden Fund, respectively. The second and the third author thank The Linnaeus Centre for Bioinformatics, Uppsala University, where they undertook the main part of this work.

References

1. B. L. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. **5** (2001) 1-13.
2. M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, Ann. Comb. **8** (2004) 391-408.
3. M. Bordewich and C. Semple, On the computational complexity of the rooted subtree prune and regraft distance, Ann. Comb. **8** (2004) 409-423.
4. D. Bryant and V. Moulton, NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks, Mol. Biol. Evol. **21**, no. 2 (2004) 255-265.
5. J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, 2003.
6. D. Gusfield, S. Eddhu, C. Langley, Optimal, efficient reconstruction of phylogenetic networks with constrained recombination, Journal of Bioinformatics and Computational Biology **2**, no. 1 (2004) 173-213.
7. J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, Math. Biosci. **98** (1990) 185-200.
8. J. Hein, T. Jiang, L. Wang, K. Zhang, On the complexity of comparing evolutionary trees, Discrete Appl. Math. **71** (1996) 153-169.
9. W. Maddison, Gene trees in species trees, Syst. Biol. **46** (1997) 523-536.
10. S. Myers, R.Griffiths, Bounds on the minimum number of recombination events in a sample history, Genetics **163** (2003) 375-394.
11. L. Nakhleh, T. Warnow, and C. Randal Linder, Reconstructing reticulate evolution in species - theory and practice, in: Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 2004, 337-346.

12. D. Posada and K. Crandall, Intraspecific gene geneologies: trees grafting into networks, Trends Ecol. Evol. **16**, (2001) 37-45.
13. C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2003.
14. Y. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotyde blocks: finding the minimum number of recombination events, in: Algorithms in Bioinformatics (WABI), G. Benson and R. Page, Eds., Lecture Notes in Bioinformatics, vol. 2812, 2003, pp. 287-302.
15. Y. Song and J. Hein, On the minimum number of recombination events in the evolutionary history of DNA sequences, J. Math. Biol. **48** (2004) 160-186.
16. Y. Song and J. Hein, Constructing minimal ancestral recombination graphs, J. Comp. Biol., to appear.
17. L. Wang, K. Zhang, L. Zhang, Perfect phylogenetic networks with recombination, J. Comp. Biol. **8** (2001) 69-78.