NOTE ON THE HYBRIDIZATION NUMBER AND SUBTREE DISTANCE IN PHYLOGENETICS

PETER J. HUMPHRIES AND CHARLES SEMPLE

ABSTRACT. For two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$, the rooted subtree prune and regraft distance between $\mathcal{T}$ and $\mathcal{T}'$ has often been used as a replacement for the hybridization number of $\mathcal{T}$ and $\mathcal{T}'$. However, Baroni *et al.* [1] constructed particular instances that showed both the difference and the ratio between this number and distance can be arbitrarily large. In this note, we show that the difference and ratio values obtained in [1] are the best possible, thus answering a problem posed in [2].

1. INTRODUCTION

Reticulation processes are now widely recognized in the evolution history of certain groups of species (e.g. [3, 4]). These processes include hybridization, horizontal gene transfer, and recombination, and result in species being a composite of DNA regions derived from different ancestors.

A fundamental problem for biologists studying the evolution of present-day species whose past includes reticulation is the following: given a collection of rooted phylogenetic (evolutionary) trees that correctly represents the tree-like evolution of different parts of the species genomes, find the smallest number of reticulation events that explains the evolution of the species under consideration (e.g. [5, 6]). This smallest number sets a lower bound on the number of such events and provides an indication of the extent to which reticulation has influenced the evolutionary history of the present-day species. Computationally speaking, this problem is NP-hard even when the initial collection consists of just two rooted binary phylogenetic trees [7]. Partly because of this, much of the interest in the problem has been on this particular instance.

Historically, a graph-theoretic operation, 'rooted subtree prune and regraft' (rSPR), has been used as one of the main tools for analyzing and modeling reticulation (e.g. [5, 8, 9]). Informally, this operation prunes a subtree of a rooted tree and then reattaches this subtree to another part of the tree. A single rSPR operation models a single reticulation event. Indeed, it is easy to observe that if two rooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ are inconsistent but this inconsistency can be explained by a single reticulation event, then $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by a single rSPR operation. Extending this further, it is tempting to conjecture that the minimum number of reticulation events to explain the inconsistency of $\mathcal{T}$ and $\mathcal{T}'$ equates to the rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$, that is the minimum number of rSPR operations to transform $\mathcal{T}$ into $\mathcal{T}'$. While the rSPR distance is a lower bound

Date: August 8, 2008.
Key words and phrases. subtree prune and regraft, hybridization number, agreement forests.
We thank the New Zealand Marsden Fund for supporting this work.

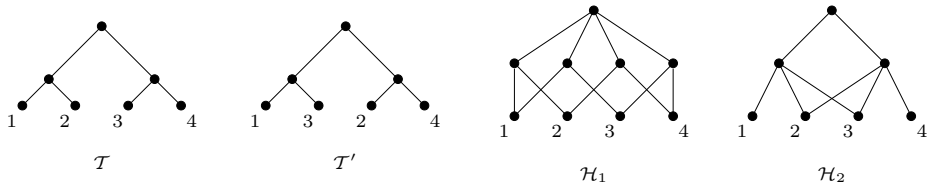

FIGURE 1. Two hybridization networks $\mathcal{H}_1$ and $\mathcal{H}_2$ that display the two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$.

and has been used as such for the minimum number of reticulation events [9, 10], it can underestimate this number of events [10]. This underestimation motivated Baroni *et al.* [1] to investigate the possible differences and ratios between these two minimum numbers. In particular, they show that these values can be arbitrarily large relative to the size of the leaf sets. In this note, we show the values obtained in [1] are the best possible, thus answering a question posed in [2].

The next section contains some preliminaries and a formal statement of the main result of the paper, while Section 3 contains the proof of this result.

2. Main Theorem

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. For completeness, if $|X| = 1$, then the tree consisting of the isolated vertex in X is a rooted binary phylogenetic X-tree. The set X is called the *label set* of $\mathcal{T}$ and is often denoted by $\mathcal{L}(\mathcal{T})$. For example, two rooted binary phylogenetic trees are shown on the left-hand-side of Fig. 1.

Hybridization networks. A *hybridization network* $\mathcal{H}$ (on X) is a rooted acyclic digraph with root ρ and the following properties:

(i) X is the set of vertices of out-degree zero;
(ii) the out-degree of ρ is at least two; and
(iii) each vertex with out-degree one has in-degree at least two.

Again for completeness, if $|X| = 1$, then the digraph consisting of the isolated vertex in X is a hybridization network on X. Rooted phylogenetic trees are special examples of hybridization networks. The set X represents a set of taxa (e.g. species) and vertices of in-degree at least two represent an exchange of genetic information between their parents. Generically, we refer to such vertices as *hybridization vertices*. The *hybridization number* of $\mathcal{H}$ is

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

where $d^-(v)$ denotes the in-degree of v. Since every vertex apart from the root has at least one parent, $(d^-(v) - 1)$ is the number of additional parents of v. The number $h(\mathcal{H})$ quantifies the number of hybridization events of $\mathcal{H}$. In Fig. 1, $\mathcal{H}_1$ and $\mathcal{H}_2$ are hybridization networks with $h(\mathcal{H}_1) = 4$ and $h(\mathcal{H}_2) = 2$. Note, in contrast to the networks $\mathcal{H}_1$ and $\mathcal{H}_2$ in this figure, hybridization vertices may be 'internal'.

A hybridization network $\mathcal{H}$ *displays* a rooted binary phylogenetic tree $\mathcal{T}$ if $\mathcal{T}$ can be obtained from a rooted subtree of $\mathcal{H}$ by contracting degree-two vertices. The

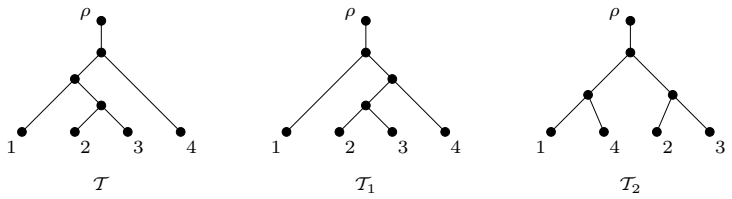

FIGURE 2. Each of $\mathcal{T}_1$ and $\mathcal{T}_2$ can be obtained from $\mathcal{T}$ by a single rSPR operation.

hybridization number of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is

$$h(\mathcal{T},\mathcal{T}') = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybridization network that displays } \mathcal{T} \text{ and } \mathcal{T}'\}.$$

In Fig. 1, each of $\mathcal{H}_1$ and $\mathcal{H}_2$ display $\mathcal{T}$ and $\mathcal{T}'$. Moreover, it is easily seen that $\mathcal{H}_2$ minimizes the hybridization number of $\mathcal{T}$ and $\mathcal{T}'$, in other words, $h(\mathcal{T},\mathcal{T}') = 2$.

Rooted subtree prune and regraft operation. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. For the purposes of the definitions in this subsection, we view the roots of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ at the end of a pendant edge adjoined to the original root. For example, see the phylogenetic trees shown in Fig. 2. Also, we regard ρ as part of the label sets of $\mathcal{T}$ and $\mathcal{T}'$, and so $\mathcal{L}(\mathcal{T}) = \mathcal{L}(\mathcal{T}') = X \cup \{\rho\}$.

Let $e = \{u, v\}$ be any edge of $\mathcal{T}$ that is not incident with ρ, where u is the vertex on the path from ρ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by deleting e and reattaching the resulting rooted subtree via a new edge, f say, as follows. Subdivide an edge of the component that contains ρ with a new vertex u', adjoin f between u' and v, and then contract u. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted subtree prune and regraft* (rSPR) operation. The rSPR *distance* between two arbitrary rooted binary phylogenetic X-trees, denoted by $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$, is the minimum number of rSPR operations that is needed to transform $\mathcal{T}$ into $\mathcal{T}'$. This distance is well-defined as it is well-known that one can always transform $\mathcal{T}$ to $\mathcal{T}'$ via a sequence of rSPR operations. Referring to Fig. 2, each of $\mathcal{T}_1$ and $\mathcal{T}_2$ can be obtained from $\mathcal{T}$ by a single rSPR operation.

The main result of this paper is the following theorem.

Theorem 2.1. *Let $n \geq 4$, and let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees with $|X| = n$. Then*

(i) $\frac{h(\mathcal{T},\mathcal{T}')}{d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')} \leq \frac{1}{2}\left\lfloor \frac{n}{2} \right\rfloor$, *and*
(ii) $h(\mathcal{T},\mathcal{T}') - d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') \leq n - \lceil 2\sqrt{n} \rceil$.

Moreover, the inequalities in (i) and (ii) are sharp for all n.

Baroni *et al.* [1] constructed explicit examples to show that, for all $n \geq 4$, there exists pairs of rooted binary phylogenetic X-trees for which the ratio and difference values given in (i) and (ii) can be obtained. Thus to prove Theorem 2.1 it suffices to show that they are the best possible.

Note that, in [1], the right-hand-side of (ii) is expressed as follows:

$$n - 2\lfloor \sqrt{n} \rfloor - c,$$

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. To see that $n - \lceil 2\sqrt{n} \rceil$ is equal to this expression, first observe, for all $n \geq 4$, that

$\lceil 2\sqrt{n} \rceil = 2\lfloor \sqrt{n} \rfloor + c$ for some $c \in \{0, 1, 2\}$. Now n is a square if and only if $c = 0$, so assume n is not a square. The two expressions coincide if we can now show that $\lceil 2\sqrt{n} \rceil = 2\lfloor \sqrt{n} \rfloor + 1$ if and only if $n - \lfloor \sqrt{n} \rfloor^2 < \sqrt{n}$. Now

$$\begin{aligned} \lceil 2\sqrt{n} \rceil = 2\lfloor \sqrt{n} \rfloor + 1 &\Leftrightarrow \sqrt{n} - \lfloor \sqrt{n} \rfloor \leq \frac{1}{2} \\ &\Leftrightarrow \sqrt{n} \leq \frac{1}{2} + \lfloor \sqrt{n} \rfloor \\ &\Leftrightarrow n \leq \frac{1}{4} + \lfloor \sqrt{n} \rfloor + \lfloor \sqrt{n} \rfloor^2 \\ &\Leftrightarrow n - \lfloor \sqrt{n} \rfloor^2 \leq \frac{1}{4} + \lfloor \sqrt{n} \rfloor \end{aligned}$$

But $n - \lfloor \sqrt{n} \rfloor^2$ is a positive integer, and so $n - \lfloor \sqrt{n} \rfloor^2 \leq \lfloor \sqrt{n} \rfloor < \sqrt{n}$. The rest of this section contains additional preliminaries to be used in the proof of Theorem 2.1.

Agreement forests. We noted earlier that computing the hybridization number of two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is NP-hard. Perhaps not surprisingly, computing the rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$ is also NP-hard [11]. Nevertheless, for each of $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ and $h(\mathcal{T}, \mathcal{T}')$, there is an attractive and useful graph-theoretic characterization in terms of 'agreement forests'. These characterizations have been crucially used a number of times by various authors—in particular, they were used by [1] to show the ratio and difference values in Theorem 2.1 can be obtained—and will again be used here.

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $X' \subseteq X$. We denote the minimum rooted subtree of $\mathcal{T}$ that connects the vertices labelled with elements in X' by $\mathcal{T}(X')$. Furthermore, we use $\mathcal{T}|X'$ to denote the rooted binary phylogenetic X'-tree obtained from $\mathcal{T}(X')$ by contracting any non-root vertices of degree two.

Now let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. As in the definitions associated with the rSPR operation, we again view the root ρ of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex at the end of a pendant edge adjoined to the original root, and as part of the label sets of $\mathcal{T}$ and $\mathcal{T}'$. An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a partition $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \ldots, \mathcal{L}_k\}$ of the label set $X \cup \{\rho\}$ of $\mathcal{T}$ and $\mathcal{T}'$, where $\rho \in \mathcal{L}_\rho$, and

(i) the trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, \ldots, k\}\}$ are vertex disjoint subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively, and
(ii) for all $i \in \{\rho, 1, \ldots, k\}$, we have $\mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$.

If, amongst all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, $\mathcal{F}$ contains the smallest number of parts, then $\mathcal{F}$ is a *maximum-agreement* forest for $\mathcal{T}$ and $\mathcal{T}'$, in which case we denote this value of k by $m(\mathcal{T}, \mathcal{T}')$. To illustrate, $\mathcal{F}_1 = \{\{\rho\}, \{1, 2, 3\}, \{4, 5, 6\}\}$ and $\mathcal{F}_2 = \{\{\rho, 1, 2, 3\}, \{4\}, \{5\}, \{6\}\}$ are agreement forests for $\mathcal{T}$ and $\mathcal{T}'$ in Fig 3. Indeed, it is easily checked that $\mathcal{F}_1$ is a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Agreement forests can be used to characterize the rSPR distance. To characterize the hybridization number in terms of agreement forests we need an additional condition.

Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \ldots, \mathcal{L}_k\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $G_\mathcal{F}$ be the directed graph that has vertex set $\mathcal{F}$ and a directed edge $(\mathcal{L}_i, \mathcal{L}_j)$ if $i \neq j$ and either

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

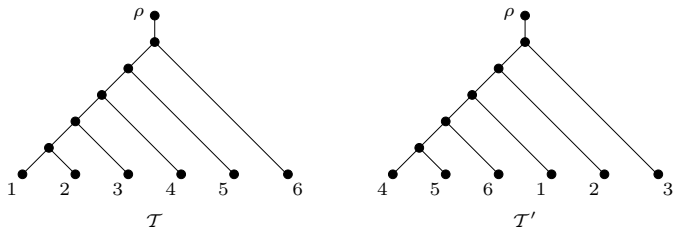

FIGURE 3. Two rooted binary phylogenetic trees.

We say $\mathcal{F}$ is an *acyclic-agreement* forest for $\mathcal{T}$ and $\mathcal{T}'$ if $G_{\mathcal{F}}$ contains no directed cycles. If $\mathcal{F}$ has the smallest number of parts over all acyclic-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, 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 $m_a(\mathcal{T}, \mathcal{T}')$. Extending the last example, $\mathcal{F}_1$ is not an acyclic-agreement forest, while $\mathcal{F}_2$ is such a forest. It is straightforward to show that $\mathcal{F}_2$ is a maximum-acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ in Fig. 3. Parts (i) and (ii) of the following theorem are established in [11] and [1], respectively.

Theorem 2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then*

(i) $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$, *and*
(ii) $h(\mathcal{T}, \mathcal{T}') = m_a(\mathcal{T}, \mathcal{T}')$.

3. Proof of Theorem 2.1

Despite its simplicity, the next lemma is the key to proving Theorem 2.1.

Lemma 3.1. *Let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \ldots, \mathcal{L}_k\}$ be an agreement forest for two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, where $k \geq 2$. Then there is an acyclic-agreement forest $\mathcal{F}_a$ for $\mathcal{T}$ and $\mathcal{T}'$ such that $|\mathcal{F}_a| \leq n - \lceil \frac{n}{k} \rceil + 1$, where $n = |X|$.*

Proof. It suffices to construct an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ of the specified size. Let $|\mathcal{L}_\rho| = r + 1$. Without loss of generality, we may assume that $\mathcal{L}_k$ is a maximum-sized set in $\mathcal{F} - \mathcal{L}_\rho$, and so $|\mathcal{L}_k| \geq \frac{n-r}{k}$.

If $r = 0$ or $r = 1$, then $\mathcal{T}|(\mathcal{L}_\rho \cup \mathcal{L}_k) \cong \mathcal{T}'|(\mathcal{L}_\rho \cup \mathcal{L}_k)$ and so $\mathcal{F}_a = \{\mathcal{L}_\rho \cup \mathcal{L}_k\} \cup \{\{y\} : y \in X - (\mathcal{L}_\rho \cup \mathcal{L}_k)\}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Here $|\mathcal{F}_a| = n - (|\mathcal{L}_\rho| - 1 + |\mathcal{L}_k|) + 1$. In the case $r = 0$, this gives

$$|\mathcal{F}_a| \leq n - \left\lceil \frac{n}{k} \right\rceil + 1,$$

while if $r = 1$, we obtain

$$|\mathcal{F}_a| \leq n - \left\lceil \frac{n-1}{k} \right\rceil \leq n - \left\lceil \frac{n-k}{k} \right\rceil \leq n - \left\lceil \frac{n}{k} \right\rceil + 1.$$

On the other hand, if $r \geq 2$, then $\mathcal{F}_a = \{\mathcal{L}_\rho, \mathcal{L}_k\} \cup \{\{y\} : y \in X - (\mathcal{L}_\rho \cup \mathcal{L}_k)\}$ is an acyclic-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Now

$$|\mathcal{F}_a| \leq n - \left\lceil \frac{n-r}{k} \right\rceil - r + 2 = n - \left\lceil \frac{n-r}{k} + r - 1 \right\rceil + 1 \leq n - \left\lceil \frac{n}{k} \right\rceil + 1$$

since $-\frac{r}{k} + r - 1 \geq 0$ as $k, r \geq 2$. This completes the proof of the lemma. □

Lemma 3.2. *Let k and n be positive integers. For all $n \geq 4$, if $2 \leq k \leq n$, then*

$$\frac{n - \lceil \frac{n}{k} \rceil}{k} \leq \frac{1}{2} \left\lfloor \frac{n}{2} \right\rfloor.$$

Proof. A routine check shows that the lemma holds for $4 \leq n \leq 8$, and for all $n \geq 4$ with $k = 2$. So assume that $n \geq 9$ and $k \geq 3$. Then

$$\frac{n - \lceil \frac{n}{k} \rceil}{k} \leq \frac{n - \frac{n}{k}}{k} = \frac{n(k-1)}{k^2} \leq \frac{2n}{9},$$

as $\frac{k-1}{k^2}$ is strictly decreasing for $k \geq 3$. Since $\frac{2n}{9} \leq \frac{n-1}{4}$ if $n \geq 9$,

$$\frac{2n}{9} \leq \frac{n-1}{4} \leq \frac{1}{2} \left\lfloor \frac{n}{2} \right\rfloor$$

completing the proof of the lemma. □

Lemma 3.3. *Let k and n be positive integers. For all $n \geq 4$, if $2 \leq k \leq n$, then*

$$k + \left\lceil \frac{n}{k} \right\rceil \geq \lceil 2\sqrt{n} \rceil.$$

Proof. As a particular consequence of the classical inequality of arithmetic and geometric means (see, for example, [12]), we have, for all non-negative real numbers k and l,

$$k + l \geq 2\sqrt{kl}.$$

Setting $l = \frac{n}{k}$ and then taking the ceiling of both sides gives the desired result. □

Proof of Theorem 2.1. Suppose that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = k$. If $k = 1$, then $h(\mathcal{T}, \mathcal{T}') = 1$ and the theorem holds. So assume $k \geq 2$. Then, by Theorem 2.2 and Lemma 3.1, $h(\mathcal{T}, \mathcal{T}') \leq n - \lceil \frac{n}{k} \rceil$. Therefore, by Lemma 3.2,

$$\frac{h(\mathcal{T}, \mathcal{T}')}{d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')} \leq \frac{n - \lceil \frac{n}{k} \rceil}{k} \leq \frac{1}{2} \left\lfloor \frac{n}{2} \right\rfloor$$

and, by Lemma 3.3,

$$h(\mathcal{T}, \mathcal{T}') - d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq n - \left\lceil \frac{n}{k} \right\rceil - k = n - \left(\left\lceil \frac{n}{k} \right\rceil + k \right) \leq n - \left\lceil 2\sqrt{n} \right\rceil.$$

This establishes (i) and (ii), thus completing the proof of the theorem. □

References

[1] M. Baroni, S. Grünewald, V. Moulton and C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, *J. Math. Biol.* **51**, 171-182 (2005).

[2] C. Semple, Hybridization networks, In *Reconstructing Evolution: New Mathematical and Computational Advances* (Edited by O. Gascuel and M. Steel), pp. 277-314, Oxford University Press (2007).

[3] J. Mallet, Hybridization as an invasion of the genome, *Trends Ecol. Evol.* **20**, 229-237 (2005).

[4] K. McBreen and P.J. Lockhart, Reconstructing reticulate evolutionary histories of plants, *Trends Plant Sci.* **11**, 398-404 (2007).

[5] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, *Math. Biosci.* **98**, 185-200 (1990).

[6] S. Myers and R. Griffiths, Bounds on the minimum number of recombination events in a sample history, *Genetics* **163**, 375-394 (2003).

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

[8] W. Maddison, Gene trees in species trees, *Syst. Biol.* **46**, 523-536 (1997).

[9] Y.S. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotyde blocks: finding the minimum number of recombination events, In *Algorithms in Bioinformatics*, (Edited by G. Benson and R. Page), Lecture Notes in Bioinformatics **2812**, pp. 287-302, Springer, Berlin (2003).

[10] Y.S. Song and J. Hein, Constructing minimal ancestral recombination graphs, *J. Comput. Biol.* **12**, 147-169 (2005).

[11] M. Bordewich and C. Semple, On the computational complexity of the rooted subtree prune and regraft distance, *Ann. Comb.* **8**, 409-423 (2004).

[12] A.-L. Cauchy, *Cours d'analyse de l'École Royale Polytechnique*, premier partie, Analyse algébraique, Paris, (1821).

Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand.

E-mail address: p.humphries@math.canterbury.ac.nz, c.semple@math.canterbury.ac.nz