Linz *BMC Evolutionary Biology* 2010, **10**:334
http://www.biomedcentral.com/1471-2148/10/334

BMC Evolutionary Biology

CORRESPONDENCE Open Access

On Hill et al's conjecture for calculating the subtree prune and regraft distance between phylogenies

Simone Linz

Abstract

Background: Recently, Hill et al. [1] implemented a new software package–called SPRIT–which aims at calculating the minimum number of horizontal gene transfer events that is needed to simultaneously explain the evolution of two rooted binary phylogenetic trees on the same set of taxa. To this end, SPRIT computes the closely related so-called rooted subtree prune and regraft distance between two phylogenies. However, calculating this distance is an NP-hard problem and exact algorithms are often only applicable to small- or medium-sized problem instances. Trying to overcome this problem, Hill et al. propose a divide-and-conquer approach to speed up their algorithm and conjecture that this approach can be used to compute the rooted subtree prune and regraft distance exactly.

Results: In this note, we present a counterexample to Hill et al's conjecture and subsequently show that a modified version of their conjecture holds.

Conclusion: While Hill et al's conjecture may result in an overestimate of the rooted subtree prune and regraft distance, a slightly more restricted version of their approach gives the desired outcome and can be applied to speed up the exact calculation of this distance between two phylogenies.

Background

In recent years, one of the main research foci in the development of theoretical frameworks that aim at approaching questions in evolutionary biology turns from the reconstruction of phylogenetic trees towards the reconstruction of phylogenetic networks. This has partly been triggered by the exponentially growing amount of available sequence data arising from whole genome sequencing projects and a successive detection of genes whose sequences are chimeras of distinct ancestral gene sequences, and hence, are likely to be the result of reticulation (e.g. horizontal gene transfer or hybridization). Although evolutionary biologists are now mostly acknowledging the existence of species arising from reticulation within certain groups of organisms, the extent to which such events have influenced the evolutionary history for a set of present-day species remains controversially discussed until today. To shed light on this question, Hill et al. [1] recently published a study that is centered around the identification and quantification of horizontal gene transfer. The authors have implemented a new software package–called SPRIT–consisting of a heuristic as well as an exact algorithm, applied it to several data sets of variable size, and compared their results and running times with those obtained from other algorithms that have previously been developed to analyze reticulate evolution.

Algorithmically, SPRIT draws on ideas that are borrowed from work that has been done in the context of the graph-theoretic operation of rooted subtree prune and regraft (rSPR) which is a popular tool to quantify the dissimilarity between two trees. Loosely speaking, an rSPR operation cuts (prunes) a subtree and reattaches (regrafts) it to another part of the tree. A lower bound on the number of reticulation events that is needed to simultaneously explain two phylogenies is the minimum number of rSPR operations that transform one phylogeny into the other [2,3]. This minimum number, which is computed by SPRIT, is referred to as the rSPR distance. However, since the task of calculating this distance is an NP-hard optimization problem, the

Correspondence: simone_linz@yahoo.de
Department of Computer Science, Technical University of Catalonia, Barcelona, Spain

application of exact algorithms is often restricted to medium-sized data sets.

In trying to overcome this obstacle, thus to speed up SPRIT, Hill et al. propose a divide-and-conquer-type reduction that breaks the problem into several smaller and more tractable subproblems before calculating the rSPR distance for each subproblem separately. Briefly, the authors conjecture that the sum of rSPR distances over all smaller subproblems is equal to the rSPR distance of the original unreduced trees. In this note, we give a counterexample to their conjecture. Nevertheless, we subsequently show that a slightly more restricted version of their conjecture holds and can be used to exactly calculate the rSPR distance between two phylogenies by breaking the problem into smaller subproblems.

The remainder of this paper is organized as follows. The next section contains some mathematical preliminaries that are needed to formally state Hill at al's conjecture. This conjecture is then given in the subsequent section which also contains the aforementioned counterexample. We then show that a modified version of the conjecture holds in the following section. We end this note with a brief conclusion.

Preliminaries

In this section, we give some preliminary definitions that are used throughout this paper. Unless otherwise stated, the notation and terminology follows [4].

Phylogenetic Trees

A *rooted binary phylogenetic X-tree* $\mathcal{T}$ is a rooted tree whose root has degree two while all other interior vertices have degree three and whose leaf set is X. The set X is the *label set* of $\mathcal{T}$ and is frequently denoted by $\mathcal{L}(\mathcal{T})$. Furthermore, let X' be a subset of X. The *minimal rooted subtree* of $\mathcal{T}$ that connects all the leaves in X' is denoted by $\mathcal{T}(X')$ while the *restriction of* $\mathcal{T}$ *to* X', denoted by $\mathcal{T}|X'$, is the rooted binary phylogenetic X'-tree obtained from $\mathcal{T}(X')$ by contracting all degree-two vertices apart from the root.

Rooted Subtree Prune and Regraft

Let $\mathcal{T}$ be a rooted binary phylogenetic X-trees. For the purposes of the upcoming definition, we view the root of $\mathcal{T}$ as a vertex ρ adjoined to the original root by a pendant edge. Now, let $e = \{u, v\}$ be any edge of $\mathcal{T}$ that is not incident with ρ such that u is the vertex on the path from ρ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and reattaching the resulting subtree with root v via a new edge, say f, as follows. Subdivide an edge of the component that contains ρ with a new vertex u', join u' and v with f, and contract u. Then $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted subtree prune and regraft (rSPR) operation*. The rSPR *distance* between two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is the minimum number of rSPR operations that transform $\mathcal{T}$ into $\mathcal{T}'$. We denote this distance by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$.

Agreement Forests

Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Again, to make the following work, regard the roots of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ adjoined to the original root by a pendant edge. An *agreement forest* $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ is a partition of $X \cup \{\rho\}$ such that $\rho \in \mathcal{L}_\rho$ and the following properties are satisfied:

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

Throughout the remainder of this note, we will interchangeably refer to $\{\mathcal{T} \mid \mathcal{L}_\rho, \mathcal{T} \mid \mathcal{L}_1, \mathcal{T} \mid \mathcal{L}_2, ..., \mathcal{T} \mid \mathcal{L}_k\}$ and $\{\mathcal{L}_\rho, \mathcal{L}_1, ..., \mathcal{L}_k\}$ as an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. A *maximum-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ with the smallest number of elements over all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$. Note that a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is not necessarily unique.

Bordewich and Semple [5] established the following characterization which directly relates the rSPR distance to the number of elements in a maximum-agreement forest and is crucial to many algorithms that exactly compute the rSPR distance between two rooted binary phylogenetic trees.

Theorem 1. *Let* $\mathcal{T}$ *and* $\mathcal{T}'$ *be two rooted binary phylogenetic X-trees, and let* $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_k\}$ *be a maximum-agreement forest for* $\mathcal{T}$ *and* $\mathcal{T}'$. *Then*

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

Clusters

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree, and let A be a subset of X with $|A| \geq 2$. We say that A is a *cluster* of $\mathcal{T}$ if there is a vertex v in $\mathcal{T}$ whose set of descendants is precisely A. We denote this cluster by $\mathcal{C}_\mathcal{T}(v)$.

We next consider several different types of clusters that will play an important role in the remainder of this paper. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and let A be a cluster that is common to $\mathcal{T}$ and $\mathcal{T}'$; that is there exists a vertex v in $\mathcal{T}$ and a vertex v' in $\mathcal{T}'$ such that $\mathcal{C}_\mathcal{T}(v) = \mathcal{C}_{\mathcal{T}'}(v')$. Furthermore, let u (resp. u') be the parent vertex of v (resp. v') in $\mathcal{T}$ (resp.

$w \neq v$), and let w (resp. w') be the child vertex of u (resp. u') with $w \neq v$ (resp. $w' \neq v'$). If no proper subset of A is a common cluster of $\mathcal{T}$ and $\mathcal{T}'$, we refer to A as a *minimal cluster*. Moreover, A is a *solvable cluster* if A is minimal and $\mathcal{C}_{\mathcal{T}}(u) = \mathcal{C}_{\mathcal{T}'}(u')$. Lastly, we say that A is a *subtree-like cluster* if A is a solvable cluster and $\mathcal{T}|\mathcal{C}_{\mathcal{T}}(w) \cong \mathcal{T}'|\mathcal{C}_{\mathcal{T}'}(w')$. Roughly speaking, the condition $\mathcal{T}|\mathcal{C}_{\mathcal{T}}(w) \cong \mathcal{T}'|\mathcal{C}_{\mathcal{T}'}(w')$ is satisfied if the subtree with root w in $\mathcal{T}$ is identical to the subtree with root w' in $\mathcal{T}'$. We refer to $\mathcal{T}|\mathcal{C}_{\mathcal{T}}(w)$ as the *common subtree associated with A* and note that it can exclusively consist of an isolated vertex. For example, $A = \{1, 2, ..., 6\}$ is a solvable cluster of the two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ that are shown in Figure 1 since $\mathcal{C}_{\mathcal{T}}(u) = \mathcal{C}_{\mathcal{T}'}(u') = \{1, 2, ..., 12\}$. However, as $\mathcal{T}|(7,8,\ldots,12) \not\cong \mathcal{T}'|(7,8,\ldots,12)$, it follows that A is not a subtree-like cluster of $\mathcal{T}$ and $\mathcal{T}'$.

Now, let $\Theta \in \{$*minimal, solvable, subtree-like*$\}$. We next describe algorithmically how to obtain a sequence of tree pairs–which is important to mathematically state Hill et al's conjecture–by decomposing two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ into smaller subtrees. As previously, view the roots of $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ adjoined to the original root by a pendant edge, and regard ρ as part of the label set; that is $\mathcal{L}(\mathcal{T}) = X \cup \{\rho\}$. Setting i to be 1, let A_i be a common Θ cluster of $\mathcal{T}$ and $\mathcal{T}'$ with $|\mathcal{L}(\mathcal{T})| - |A_i| > 1$. Let $\mathcal{T}_i$ denote the rooted binary phylogenetic tree $\mathcal{T}|A_i$ (viewing the root of $\mathcal{T}_i$ as a vertex ρ_i adjoined to the original root by a pendant edge) and reset $\mathcal{T}$ to be the tree obtained from $\mathcal{T}$ by replacing $\mathcal{T}(A_i)$ with a new vertex a_i. Analogously, let $\mathcal{T}'_i$ denote the rooted binary phylogenetic tree $\mathcal{T}'|A_i$ (viewing the root of $\mathcal{T}'_i$ as a vertex ρ_i adjoined to the original root by a pendant edge) and reset $\mathcal{T}'$ to be the tree obtained from $\mathcal{T}'$ by replacing $\mathcal{T}'(A_i)$ with a new vertex a_i. If $\mathcal{T}$ and $\mathcal{T}'$ contain a Θ cluster A_{i+1} with $|\mathcal{L}(\mathcal{T})| - |A_{i+1}| > 1$, stop or increment i by 1 and repeat this process; otherwise, stop. Eventually, we obtain a sequence

$$(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$$

of pairs of rooted binary phylogenetic trees, where $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$ denote the two trees after the replacement of $\mathcal{T}(A_t)$ and $\mathcal{T}'(A_t)$ with a vertex a_t. We call this sequence a *cluster sequence* of $\mathcal{T}$ and $\mathcal{T}'$ with respect to a specific cluster type Θ. An example of a cluster sequence with respect to Θ = *solvable* for the two rooted binary phylogenetic trees depicted in Figure 1 is shown in Figure 2.

Hill et al's Conjecture and a Counterexample

We begin this section by formally stating Hill et al's conjecture which was introduced in [1].

Conjecture 2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$ with respect to Θ = solvable. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = \sum_{i=1}^{t} d_{\mathrm{rSPR}}(\mathcal{T}_i, \mathcal{T}'_i) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho). \quad (1)$$

Next, we detail a counterexample to the above conjecture which is based on the two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ that are shown in Figure 1. A maximum-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ contains 5 elements and is shown in the top of Figure 3. By Theorem 1, this implies that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 4$. Now, consider the cluster sequence with respect to Θ = *solvable* for $\mathcal{T}$ and $\mathcal{T}'$ that contains three tree pairs and is depicted in Figure 2. The first tree pair $(\mathcal{T}_1, \mathcal{T}'_1)$ consists

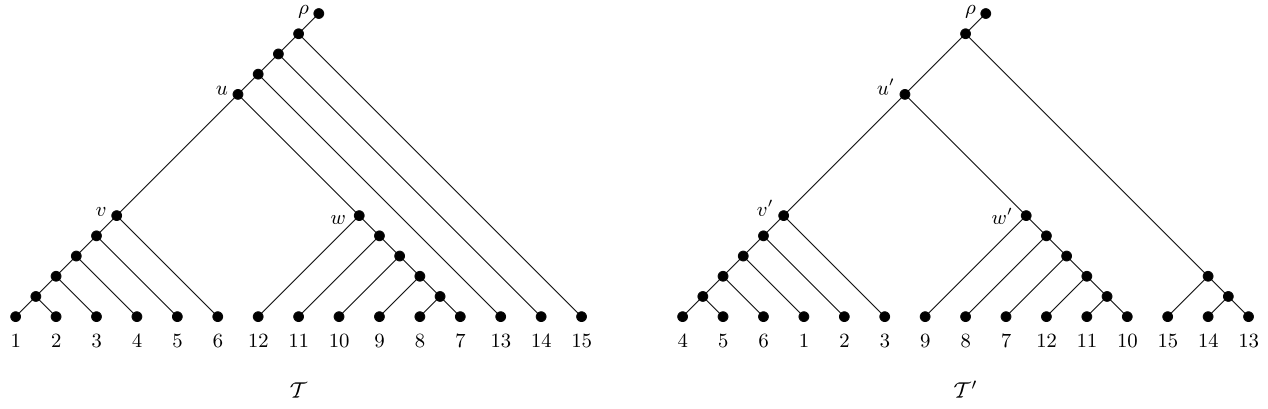

Figure 1 Two rooted binary phylogenetic *X*-trees $\mathcal{T}$ and $\mathcal{T}'$. Note that $\mathcal{T}$ and $\mathcal{T}'$ have an additional vertex ρ adjoined to the original root by a pendant edge.

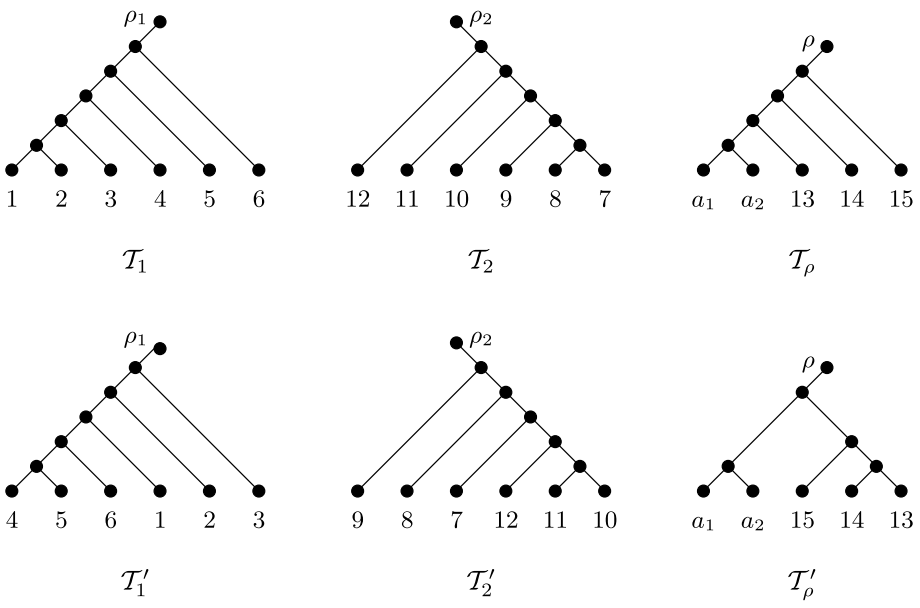

Figure 2 A cluster sequence with respect to Θ = *solvable* for the two rooted binary phylogenetic *X*-trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Figure 1. Details on how the tree pairs have been obtained are given in the text.

of the restricted subtrees of $\mathcal{T}$ and $\mathcal{T}'$ whose leaf set is the solvable cluster A_1 = {1, 2, ..., 6} of $\mathcal{T}$ and $\mathcal{T}'$; thus $\mathcal{T}_1 = \mathcal{T}\,|\,(A_1 \cup \{\rho_1\})$ and $\mathcal{T}_1' = \mathcal{T}'\,|\,(A_1 \cup \{\rho_1\})$. Similarly, the second tree pair $(\mathcal{T}_2, \mathcal{T}_2')$ consists of the restricted subtrees of the two trees that have been obtained from $\mathcal{T}$ and $\mathcal{T}'$ by replacing $\mathcal{T}(A_1)$ and $\mathcal{T}'(A_1)$, respectively, with a single leaf a_1 whose leaf set is the solvable cluster A_2 = {7, 8, ..., 12}. Lastly, the third tree pair $(\mathcal{T}_\rho, \mathcal{T}_\rho')$ can be regarded as being obtained from $\mathcal{T}$ and $\mathcal{T}'$ by replacing $\mathcal{T}(A_1)$ and $\mathcal{T}'(A_1)$ with

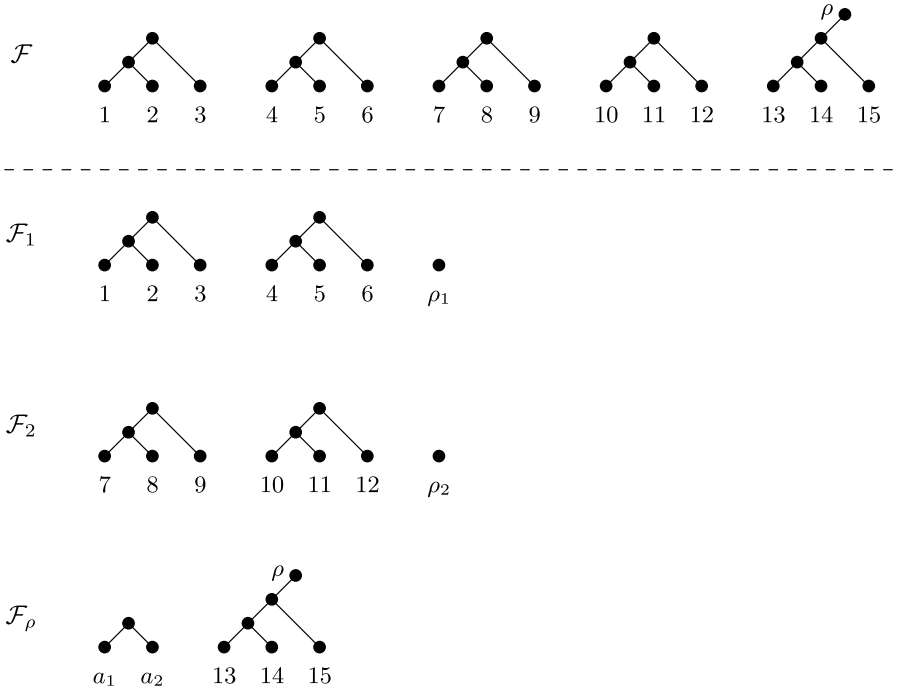

Figure 3 Maximum-agreement forests. Top: A maximum-agreement forest $\mathcal{F}$ for $\mathcal{T}$ and $\mathcal{T}'$ depicted in Figure 1. Bottom: A maximum-agreement forest $\mathcal{F}_i$ for each tree pair $\mathcal{T}_i$ and $\mathcal{T}_i'$ shown in Figure 2.

a leaf a_1 and replacing $\mathcal{T}(A_2)$ and $\mathcal{T}'(A_2)$ with a leaf a_2. For each tree pair $(\mathcal{T}_i, \mathcal{T}'_i)$ of the cluster sequence shown in Figure 2, a maximum-agreement forest $\mathcal{F}_i$ with $i \in \{1, 2, \rho\}$ is depicted in the bottom part of Figure 3. Note that each forest $\mathcal{F}_i$ is the unique maximum-agreement forest for $\mathcal{T}_i$ and $\mathcal{T}'_i$ Now, by Equation 1, we have

$$d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}'_1) + d_{\mathrm{rSPR}}(\mathcal{T}_2, \mathcal{T}'_2) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho) = 2 + 2 + 1 = 5$$

which is strictly greater than $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$; thus showing that Conjecture 2 does not hold.

Using Subtree-Like Clusters to Prove Hill et al's Conjecture

In this section, we show that Conjecture 2 holds, if we consider a subtree-like cluster instead of a solvable cluster in each iteration of computing a cluster sequence for two rooted binary phylogenetic trees. We first prove the result for a cluster sequence of size two and then see that this result generalizes to cluster sequences of greater size.

Lemma 3. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $(\mathcal{T}_1, \mathcal{T}'_1)$, $(\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$ with respect to Θ = subtree-like. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}'_1) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho).$$

Proof. Let A_1 be the subtree-like cluster $\mathcal{L}(\mathcal{T}_1) - \{\rho_1\}$ of $\mathcal{T}$ and $\mathcal{T}'$. We start by making an observation that is crucial for what follows. By the definition of a subtree-like cluster, there exists a common subtree, say $\mathcal{S}$, that is associated with A_1 in $\mathcal{T}$ and $\mathcal{T}'$. Clearly, $\mathcal{S}$ is also a common subtree of $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$. Furthermore, as $\mathcal{T}_\rho$ has been obtained from $\mathcal{T}$ by replacing $\mathcal{T}(A_1)$ with a single vertex a_1 and as $\mathcal{T}'_\rho$ has been obtained from $\mathcal{T}'$ by replacing $\mathcal{T}'(A_1)$ with a single vertex a_1, it is easily checked that $\mathcal{T}\,|\,(\mathcal{L}(\mathcal{S}) \cup \{a_1\})$ is a common subtree of $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$.

We now show that

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \leq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}'_1) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho). \quad (2)$$

Let $\mathcal{F}_1$ be a maximum-agreement forest for $\mathcal{T}_1$ and $\mathcal{T}'_1$, and let $\mathcal{F}_\rho$ be a maximum-agreement forest for $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$. By the observation prior to this paragraph, it follows from Proposition 3.2 of [5] that $\mathcal{L}(\mathcal{S}) \cup \{a_1\}$ is a subset of an element, say $\mathcal{L}_{a_1}$, in $\mathcal{F}_\rho$. Furthermore, let $\mathcal{L}_{\rho_1}$ be the label set of $\mathcal{F}_1$ with $\rho_1 \in \mathcal{L}_{\rho_1}$. As $\mathcal{F}_1$ is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}'_1$ and as $\mathcal{F}_\rho$ is such a forest for $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$, it follows that

$$\mathcal{F} = (\mathcal{F}_1 \cup \mathcal{F}_\rho - \{\mathcal{L}_{\rho_1}, \mathcal{L}_{a_1}\}) \cup \{(\mathcal{L}_{\rho_1} - \{\rho_1\}) \cup (\mathcal{L}_{a_1} - \{a_1\})\}$$

is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. As $\mathcal{L}_{a_1}$ - $\{a_1\}$ always contains an element, note that $(\mathcal{L}_{\rho_1} - \{\rho_1\}) \cup (\mathcal{L}_{a_1} - \{a_1\})$ is never the empty set. Thus $|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_\rho| - 1$ and, by Theorem 1, we have

$$d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho) = |\mathcal{F}_1| - 1 + |\mathcal{F}_\rho| - 1 = |\mathcal{F}| - 1 \geq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}').$$

This establishes Equation 2.

We now turn to the second part of this proof and show that

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') \geq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}'_1) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho). \quad (3)$$

Let $\mathcal{F}$ be a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. The remainder of this part splits into two cases. First, assume that there exists an element in $\mathcal{F}$, say $\mathcal{L}_m$, such that $\mathcal{L}_m \cap A_1 \neq \emptyset$ and $\mathcal{L}_m \cap (X - A_1) \cup \{\rho\} \neq \emptyset$. Note that $\mathcal{L}_m$ is the only label set with the described properties, as otherwise, $\mathcal{F}$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $\mathcal{L}_{m'} = (\mathcal{L}_m \cap A_1) \cup \{\rho_1\}$, and let $\mathcal{L}_{m''} = (\mathcal{L}_m \cap ((X - A_1) \cup \{\rho\})) \cup \{a_1\}$. Since $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$,

$$\mathcal{F}_1 = \{\mathcal{L} \in \mathcal{F} : \mathcal{L} \subseteq A_1\} \cup \{\mathcal{L}_{m'}\}$$

is such a forest for $\mathcal{T}_1$ and $\mathcal{T}'_1$ and

$$\mathcal{F}_\rho = \{\mathcal{L} \in \mathcal{F} : \mathcal{L} \subseteq ((X - A_1) \cup \{\rho\})\} \cup \{\mathcal{L}_{m''}\}$$

is an agreement forest for $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$. Second, assume that no such element $\mathcal{L}_m$ exists. Hence, every element $\mathcal{L}$ in $\mathcal{F}$ is either a subset of A_1 or a subset of $(X - A_1) \cup \{\rho\}$. Furthermore, as A_1 is a subtree-like cluster of $\mathcal{T}$ and $\mathcal{T}'$ whose associated common subtree is $\mathcal{S}$, it again follows from Proposition 3.2 of [5], that $\mathcal{L}(\mathcal{S})$ is a subset of an element, say $\mathcal{L}_\mathcal{S}$, in $\mathcal{F}$. Now, as $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, it follows that

$$\mathcal{F}_1 = \{\mathcal{L} \in \mathcal{F} : \mathcal{L} \subseteq A_1\} \cup \{\{\rho_1\}\}$$

is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}'_1$ and

$$\mathcal{F}_\rho = (\{\mathcal{L} \in \mathcal{F} : \mathcal{L} \subseteq ((X - A_1) \cup \{\rho\})\} - \{\mathcal{L}_\mathcal{S}\}) \cup \{\mathcal{L}_\mathcal{S} \cup \{a_1\}\}$$

is such a forest for $\mathcal{T}_\rho$ and $\mathcal{T}'_\rho$. Regardless of whether or not $\mathcal{L}_m$ exists, we have $|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_\rho| - 1$, and therefore,

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = |\mathcal{F}| - 1 = |\mathcal{F}_1| + |\mathcal{F}_\rho| - 2 \geq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}'_1) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho).$$

This establishes Equation 3, and combining Equations 2 and 3 completes the proof of this lemma.■

The next theorem directly follows from repeated applications of Lemma 3.

Theorem 4. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Let $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a cluster sequence for $\mathcal{T}$ and $\mathcal{T}'$ with respect to* Θ = *subtree-like. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = \sum_{i=1}^{t} d_{\mathrm{rSPR}}(\mathcal{T}_i, \mathcal{T}'_i) + d_{\mathrm{rSPR}}(\mathcal{T}_\rho, \mathcal{T}'_\rho).$$

Conclusion

In this paper, we have shown that Hill et al's conjecture [1] and the underlying divide-and-conquer approach cannot be used to calculate the rSPR distance between two phylogenies exactly. To provide some intuition why this conjecture fails, consider the following. Let $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ be a cluster sequence with respect to Θ = *solvable* for two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$. Calculating a maximum-agreement forest for each tree pair $(\mathcal{T}_i, \mathcal{T}'_i)$, taking their union, and, for each $i \in$; {1, 2, ..., *t*}, joining the element containing a_i with the element containing ρ_i can potentially result in a set, say $\mathcal{G}$, which contains an element that is a subset of $\{a_1, a_2, \ldots, a_t, \rho_1, \rho_2, \ldots, \rho_t\}$. In the case of our counterexample,

$$\mathcal{G} = \{\{1,2,3\}, \{4,5,6\}, \{7,8,9\}, \{10,11,12\}, \{13,14,15, \rho\}, \{a_1, a_2, \rho_1, \rho_2\}\}$$

contains one such element. Trivially, this element is not part of any agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ while $\mathcal{G}$ - $\{\{a_1, a_2, \rho_1, \rho_2\}\}$ is precisely a maximum-agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Consequently, a divide-and-conquer approach that exactly calculates $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$ needs to take into account the number of elements in $\mathcal{G}$ that are subsets of $\{a_1, a_2, \ldots, a_t, \rho_1, \rho_2, \ldots, \rho_t\}$; otherwise, the result may be an overestimate of the exact solution. Alternatively, one can approach the problem by finding a strategy which guarantees that no element in $\mathcal{G}$ is a subset of $\{a_1, a_2, \ldots, a_t, \rho_1, \rho_2, \ldots, \rho_t\}$. This is the underlying idea of Theorem 4 which uses a slightly more restricted version of Hill et al's conjecture and finally gives the desired outcome. Hence, decomposing $\mathcal{T}$ and $\mathcal{T}'$ into a cluster sequence with respect to Θ = *subtree-like* can be used to speed up the exact calculation of $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$.

However, for practical problem instances, it may be unlikely to find many subtree-like clusters. For example, the two phylogenies shown in Figure 1 do not have any common subtree-like cluster. This is due to the restricted definition of such a cluster which requires that a vertex whose set of descendants is a common cluster of two rooted binary phylogenetic *X*-trees $\mathcal{T}$ and $\mathcal{T}'$ has the same parent vertex than a common subtree of $\mathcal{T}$ and $\mathcal{T}'$. To lessen this problem, an alternative approach–that has recently been published by Linz and Semple [6]–can be applied. This paper describes a more general divide-and-conquer approach that exactly computes the rSPR distance between $\mathcal{T}$ and $\mathcal{T}'$ for when a cluster sequence $(\mathcal{T}_1, \mathcal{T}'_1), \ldots, (\mathcal{T}_t, \mathcal{T}'_t), (\mathcal{T}_\rho, \mathcal{T}'_\rho)$ with respect to Θ = *minimal* for $\mathcal{T}$ and $\mathcal{T}'$ is given. Loosely speaking, the authors calculate a so-called minimum-weight partition $\mathcal{G}$ of $X \cup \{\rho\} \cup \{a_1, a_2, \ldots, a_t, \rho_1, \rho_2, \ldots, \rho_t\}$ such that $\mathcal{G}$ contains an agreement forest (not necessarily a maximum-agreement forest) for each tree pair $(\mathcal{T}_i, \mathcal{T}'_i)$. To compute $\mathcal{G}$, it has been shown that applying a 'bottom-up' approach which locally works on subtrees of each tree pair $(\mathcal{T}_i, \mathcal{T}'_i)$ guarantees that the number of elements in $\mathcal{G}$ that are subsets of $\{a_1, a_2, \ldots, a_t, \rho_1, \rho_2, \ldots, \rho_t\}$ is maximized while $|\mathcal{G}|$ is minimized.

Acknowledgements

I thank Maria Luisa Bonet, Mareike Fischer, and Charles Semple for useful discussions and comments on an earlier version of this paper. Financial support from MEC (TIN2007-68005-C04-03) is gratefully acknowledged.

Response

By Helgi B Schiöth
E-Mail: helgis@bmc.uu.se
Address: Department of Neuroscience, Functional Pharmacology, Uppsala University, BMC, Box 593, 751 24, Uppsala, Sweden

"We have found that the manuscript by Linz is correct and to the point. We have therefore updated the SPRIT software and published the new version online.

The new version supports both the old incorrect conjecture as well as the new correct one to allow for comparisons to be made."

Received: 21 June 2010 Accepted: 29 October 2010
Published: 29 October 2010

References

1. Hill T, Nordström KJV, Thollesson M, Säfström TM, Vernersson AKE, Fredriksson R, Schiöth HB: **SPRIT: Identifying horizontal gene transfer in rooted phylogenetic trees.** *BMC Evol Biol* 2010, **10**:42.
2. Hein J, Jing T, Wang L, Zhang K: **On the complexity of comparing evolutionary trees.** *Discrete Appl Math* 1996, **71**:153-169.
3. Baroni M, Grünewald S, Moulton V, Semple C: **Bounding the number of hybridization events for a consistent evolutionary history.** *J Math Biol* 2005, **51**:171-182.
4. Semple C, Steel M: *Phylogenetics* Oxford University Press; 2003.
5. Bordewich M, Semple C: **On the computational complexity of the rooted subtree prune and regraft distance.** *Ann Comb* 2004, **8**: 409-423.
6. Linz S, Semple C: **A cluster reduction for computing the subtree distance between phylogenies.** *Ann Comb*, in press.

doi:10.1186/1471-2148-10-334
Cite this article as: Linz: On Hill et al's conjecture for calculating the subtree prune and regraft distance between phylogenies. *BMC Evolutionary Biology* 2010 **10**:334.