A Polynomial-Time Algorithm for the Phylogeny Problem when the Number of Character States is Fixed

Richa Agarwala
Iowa State University

David Fernández-Baca
Iowa State University, fernande@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/cs_techreports

Part of the Theory and Algorithms Commons

Recommended Citation
http://lib.dr.iastate.edu/cs_techreports/13

This Article is brought to you for free and open access by the Computer Science at Iowa State University Digital Repository. It has been accepted for inclusion in Computer Science Technical Reports by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A Polynomial-Time Algorithm for the Phylogeny Problem when the Number of Character States is Fixed

Richa Agarwala* and David Fernández-Baca†

Department of Computer Science, Iowa State University, Ames, IA 50011

March 15, 1993

Abstract

We present a polynomial-time algorithm for determining whether a set of species, described by the characters they exhibit, has a phylogenetic tree, assuming the maximum number of possible states for a character is fixed. This solves an open problem posed by Kannan and Warnow. Our result should be contrasted with the proof by Steel and Bodlaender, Fellows, and Warnow that the phylogeny problem is NP-complete in general.

1 Introduction

A fundamental problem in biology is that of inferring the evolutionary history of a set of species, each of which is specified by the set of traits or characters that it exhibits [6, 7]. In mathematical terms, the problem can be expressed as follows. Let \( \mathcal{C} = \{1, \ldots, m\} \) be the character set, and for every \( c \in \mathcal{C} \), let \( \mathcal{A}_c = \{1, \ldots, r_c\} \) be the set of allowable states for character \( c \). We write \( r \) to denote \( \max_{c \in \mathcal{C}} r_c \). A species \( s \) is a vector \( (s_1, \ldots, s_m) \) such that \( s \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_m \); \( s_c \) is referred to as the state of character \( c \) for \( s \). The perfect phylogeny problem is to determine whether a given set of \( n \) distinct species \( \mathcal{S} \) has a tree \( T \) with the following properties:

\begin{align*}
\text{(C1)} & \quad \mathcal{S} \subseteq V(T) \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_m, \\
\text{(C2)} & \quad \text{Every leaf in } T \text{ is in } \mathcal{S}. \\
\text{(C3)} & \quad \text{For every } c \in \mathcal{C} \text{ and every } j \in \mathcal{A}_c, \text{ the set } U \text{ of all } u \in V(T) \text{ such that } u_c = j \text{ induce a subtree of } T.
\end{align*}

*Supported in part by a College of Liberal Arts and Sciences Research Assistantship, Iowa State University.

†Supported in part by the National Science Foundation under grants CCR-890626 and CCR-9211262. This author gratefully acknowledges the support of DIMACS, at Rutgers University, where much of this work was carried out. E-mail address: fernande@cs.iastate.edu.
Figure 1: Forced and unforced states.

The tree $T$, if it exists, is called a phylogeny for $S$. We should note that instances of the phylogeny problem are often expressed in matrix form, by giving the set of species $S$ as an $n \times m$ matrix $M$ whose rows are the species in $S$.

The phylogeny problem was shown to be NP-complete by Bodlaender et al. [2] and, independently, by Steel [12]. This fact suggests at least two lines of attack: one is to restrict $m$, the number of characters; the other is to restrict $r$. Pursuing the first approach, McMorris, Warnow, and Wimer have shown that, for every fixed $m$, the phylogeny problem is solvable in polynomial time [11]. In this paper, we pursue the second approach. When $r = 2$, the binary character case, an elegant characterization of the set of “yes” instances [5] leads to efficient algorithms [1, 8]. Dress and Steel [4] devised a $O(nm^2)$ algorithm for $r \leq 3$. Kannan and Warnow [10] gave a polynomial-time algorithm for $r \leq 4$ and conjectured the existence polynomial-time algorithms for any fixed $r$. Here we prove this conjecture by giving a $O(2^r(nm^3 + m^4))$ algorithm for the perfect phylogeny problem.

2 Preliminaries

We now introduce some definitions and prove certain preliminary results.

**Definition 1** Suppose $T$ is a phylogeny for $S$ and let $p$ be some vertex in $T$. We shall say that the state of $p$ on character $c$ is forced if $p$ lies on the path between vertices $a$ and $b$ in $S$ such that $a_c = b_c$. (Observe that if this is the case, in order to satisfy condition (C3) we must have $p_c = a_c = b_c$.)

If the state of a character of a node is unforced, several assignments may be possible. In Figure 1, for example, the state of the fourth character of the internal node is unforced and we could assign it a value of 1, 2 or 3.

**Lemma 1** A set of species $S$ has a phylogeny if and only if every subset of $S$ has one.
Proof The “if” part is trivial. For the “only if” part, let \( S' \) be any subset of \( S \) and let \( T \) be a phylogeny of \( S \). Clearly, \( T \) satisfies (C1) and (C3) for \( S' \), but, possibly, not (C2). To obtain a phylogeny for \( S' \), repeatedly delete from \( T \) any leaf that is not in \( S' \), until this operation is no longer possible. Since each deletion preserves properties (C1) and (C3) for \( S' \), the final tree will also satisfy (C2) for \( S' \). □

Definition 2 Two vectors \( a \) and \( b \) are said to be compatible on a set of characters \( A \) if and only if \( a_c = b_c \) for every \( c \in A \).

Definition 3 Suppose \( G \subseteq S \) and let \( G' = S - G \). \( \mathcal{D}(G) \), the set of distinguishing characters of \( G \), is the set of all \( c \in C \) such that for every \( a \in G \) and every \( b \in G' \), \( a_c \neq b_c \). \( \mathcal{M}(G) \), the set of common characters, is \( C - \mathcal{D}(G) \).

Obviously, \( \mathcal{D}(G) = \mathcal{D}(G') \) and \( \mathcal{M}(G) = \mathcal{M}(G') \).

Definition 4 A pair \((G, G')\) where \( G \subseteq S \) and \( G' = S - G \) is called a split if, for every character, the number of common character states between \( G \) and \( G' \) is at most one. A split \((G, G')\) is a \( c \)-split if \( \mathcal{D}(G) \neq \emptyset \). If \((G, G')\) is a split \( c \)-split, \( G \) and \( G' \) are called clusters \( c \)-clusters).

Note that, whereas the number of splits is at most \( 2^{n-1} - 1 \), the total number of \( c \)-splits is at most \((2^{n-1} - 1) \cdot m \). Observe that we can determine whether a partition \((G, G')\) of \( S \) is a split in \( O(nm) \) time. Note also that if \( G \) is a cluster that is not a \( c \)-cluster, then \( \mathcal{D}(G) = \emptyset \).

Definition 5 Let \((G, G')\) be a split. We say that \((G, G')\) is of type I if there exists an \( s \in G \) such that for all \( c \in \mathcal{M}(G) \), \( s_c \) equals the unique common state between \( G \) and \( G' \) on character \( c \) and \( |G - \{s\}|, |G'| \geq 1 \). If \((G, G')\) is of type I, we refer to \( s \) as a connecting species. If \((G, G')\) is not of type I, we say that it is of type II.

Lemma 2 If all \( c \)-splits are of type II, then, in every phylogeny \( T \) of \( S \), every internal node of \( T \) is not in \( S \).

Proof Suppose some phylogeny \( T \) of \( S \) has an internal node \( v \) in \( S \). Let \( T' \) be any connected component of \( T - v \), let \( G' = S \cap V(T') \) and \( G = S - G' \). Clearly, \((G, G')\) is a \( c \)-split with \( |G'| \geq 1 \) and \( |S - G'| \geq 2 \). One can also readily verify that \((G, G')\) is of type I, with \( v \) as a connecting species. □

Note that we can check whether or not a split \((G_1, G_2)\) is of type I in \( O(nm) \) time, since we simply need to check whether \( G_1 \) or \( G_2 \) contain a suitable species \( s \).

3 Subphylogenies

Definition 6 A subphylogeny \( T_G \) for a cluster \( G \) is a phylogeny for \( G \) containing a node \( x \) such that for every \( c \in \mathcal{M}(G) \), \( x_c \) equals the (unique) common state for character \( c \) between \( G \) and \( S - G \) and for every \( c \in \mathcal{D}(G) \), \( x_c \) is the state of some species in \( G \) on character \( c \). Node \( x \) is referred to as the connection of \( T_G \).
The next result implies that, in searching for a phylogeny for $\mathcal{S}$, we can restrict our attention to phylogenies constructed entirely from subphylogenies.

**Lemma 3** $\mathcal{S}$ has a phylogeny if and only if there exists a split $(G_1, G_2)$ such that both $G_1$ and $G_2$ have subphylogenies.

**Proof** For the “if” part, let $(G_1, G_2)$ be a split satisfying the requirements of the lemma and let $T_1$ and $T_2$ be subphylogenies for $G_1$ and $G_2$, respectively. Let $x^1$ and $x^2$ be the connections of $T_1$ and $T_2$. We can obtain a phylogeny for $\mathcal{S}$ by taking $T_1$ and $T_2$ and connecting them as follows. If $D(G_1) = \emptyset$, identify $x^1$ and $x^2$. Otherwise, add an edge $(x^1, x^2)$. It is not hard to check that conditions (C1)–(C3) hold.

For the “only if” part, let $T$ be a phylogeny for $\mathcal{S}$ and let $(u, v)$ be any edge in $T$. Without loss of generality, assume that every node in $T$ that is not in $\mathcal{S}$ has degree at least three. Let $T_1$ and $T_2$ be the subtrees of $T - (u, v)$ containing $u$ and $v$, respectively, and let $G_1 = \mathcal{S} \cap V(T_1)$ and $G_2 = G - G_1$. $T_1$ and $T_2$ are obviously phylogenies for $G_1$ and $G_2$. We can construct a subphylogeny for $G_1$ from $T_1$ as follows. For each $c \in D(G)$, let $B_c$ be the set of all $b \in V(T_1)$ such that $b_c = u_c$. Since $T$ satisfies (C3), the states of these nodes on character $c$ are unforced. Let $d$ be any node in $V(T_1) - B_c$ that is adjacent to a node in $B_c$. Obviously, such a node must exist. Now, set $b_c = d_c$ for all $b \in B_c$. Note that, since $T_1$ satisfies (C3), for every $c \in M(G_1)$, $u_c$ equals the unique common state between $G_1$ and $G_2$. The resulting modification of $T_1$ is therefore a subphylogeny for $G_1$ with connection $u$. An analogous construction can be used to obtain a subphylogeny for $G_2$. □

**Definition 7** A cluster $G$ is said to be compatible with a vector $s$ if for every $c \in M(G)$, $s_c$ equals the unique common state for character $c$ between $G$ and $\mathcal{S} - G$.

The following result demonstrates that a subphylogeny for a cluster can always be assembled from subphylogenies for $c$-clusters.

**Lemma 4** Let $G$ be a cluster. Then, $G$ has a subphylogeny if and only if there exist pairwise disjoint $c$-clusters $G_1, \ldots, G_k$ and a vector $x$ such that (i) for every $c \in M(G)$, $x_c$ equals the (unique) common state for character $c$ between $G$ and $\mathcal{S} - G$, (ii) $\bigcup_{i=1}^k G_i = G$, and (iii) each $G_i$ is compatible with $x$ and has a subphylogeny.

**Proof** For the “if” part, let $T_1, \ldots, T_k$ be the subphylogenies for $G_1, \ldots, G_k$ with roots $x^1, \ldots, x^k$. Clearly, the tree $T$ consisting of a node for $x$ and the trees $T_1, \ldots, T_k$ connected to $x$ by edges $(x^i, x), \ldots, (x^k, x)$ is a subphylogeny for $G$.

For the “only if” part, let $T$ be a subphylogeny for $G$ with connection $x$. Without loss of generality, assume that all nodes in $T$ are distinct. Let $x^1, \ldots, x^k$ be the neighbors of $x$ in $T$ and for $1 \leq i \leq k$, let $T_i$ be the subtree of $T - x$ containing $x^i$ and let $G_i = V(T_i) \cap \mathcal{S}$. For each $c \in M(G)$, $x_c$ equals the unique common state between $G_i$ and $\mathcal{S} - G_i$. This is because either this state is shared with some species in $G_j$, for some $j \neq i$, or it is shared with some species in $\mathcal{S} - G$. In either case, due to condition (C3), the value of $x^i_c$ must equal the common state and, hence, $G_i$ is compatible with $x$. Also, as done in the proof of Lemma 4, we can insure that for every character $c$, the state of
any \( v \in V(T_i) \) on character \( c \) will be that of some species in \( G_i \) on \( c \), by altering unforced states, if needed. Thus, \( T_i \) can be transformed into a subphylogeny for \( G_i \). All that is left is to verify that each \( G_i \) is indeed a c-split; i.e., that \( \mathcal{D}(G_i) \neq \emptyset \). Suppose \( \mathcal{D}(G_i) = \emptyset \). Then we must have had \( \mathbf{a}^i = \mathbf{a} \) in \( T \) (contradicting our earlier assumption that all nodes are distinct), since for every character \( c \), there is one common character state between \( G \) and \( S - G \) and condition (C3) must be satisfied in \( T \). □

To find a phylogeny for \( S \), we shall rely on certain properties of subphylogenies which allow them to be combined into larger subphylogenies. These properties are discussed next.

**Lemma 5** Let \( G, G_1, G_2 \) be clusters such that \( G = G_1 \cup G_2 \) and \( G_1 \cap G_2 = \emptyset \). If \( G_1 \) and \( G_2 \) have subphylogenies, then there exists a subphylogeny \( T \) for \( G \).

**Proof** Let \( T_1 \) and \( T_2 \) be subphylogenies for \( G_1 \) and \( G_2 \) respectively. Let \( \mathbf{a}^1 \) and \( \mathbf{a}^2 \) be the connections of \( T_1 \) and \( T_2 \). Construct \( T \) by adding a node \( \mathbf{a} \) and the edges \((\mathbf{a}, \mathbf{a}^1)\) and \((\mathbf{a}, \mathbf{a}^2)\), \( \mathbf{a} \) will be the connection of \( T \). The character states of \( \mathbf{a} \) are defined as follows. For each \( c \in \mathcal{M}(G) \), set \( x_c \) equal to the common state between \( G \) and \( S - G \). For \( c \in \mathcal{D}(G) \), set \( x_c = x_c^1 \). Since \( \mathbf{a} \) has the required states on characters in \( \mathcal{M}(G) \), it suffices to prove that \( T \) is a phylogeny for \( G \). For this we must show that \( \mathbf{a}^1 \) and \( \mathbf{a}^2 \) are compatible with \( \mathbf{a} \) on \( \mathcal{M}(G_1) \) and \( \mathcal{M}(G_2) \) respectively. We have two cases to consider:

\( c \in \mathcal{M}(G) \): Then \( x_c \) equals the unique common state for character \( c \) between \( G \) and \( S - G \) and \( c \) must be in \( \mathcal{M}(G_1) \cup \mathcal{M}(G_2) \). There are two possibilities:

- \( c \in \mathcal{M}(G_1) \cap \mathcal{D}(G_2) \) or \( c \in \mathcal{M}(G_2) \cap \mathcal{D}(G_1) \). Consider the first of these two cases — the other can be dealt with analogously. Then there is no species in \( G_2 \) that has a common character state with a species in \( S - G_2 \). Now since \( G = G_1 \cup G_2 \) and \( G \) has a common character state with \( S - G \), the common state must be with respect to a species in \( G_1 \) and hence \( x_c^1 = x_c \).

- \( c \in \mathcal{M}(G_1) \cap \mathcal{M}(G_2) \). Then we claim that \( x_c^1 = x_c^2 = x_c \). First, note that we must have \( x_c^1 = x_c \) or \( x_c^2 = x_c \) because the species in \( G \) sharing the common character state with a species in \( S - G \) belongs to either \( G_1 \) or \( G_2 \).

We now show that \( x_c^1 = x_c^2 \). Assume the contrary. Since \( c \in \mathcal{M}(G_1) \), there exists a species in \( G_1 \) that shares the character state \( x_c^1 \) with a species \( \mathbf{a} \) in \( S - G_1 \). Similarly since \( c \in \mathcal{M}(G_2) \), there exists a species in \( G_2 \) that shares the character state \( x_c^2 \) with a species \( \mathbf{b} \) in \( S - G_2 \). Then \( \mathbf{b} \notin G_1 \), for if \( \mathbf{b} \in G_1 \), \( G_1 \) and \( S - G_1 \) would share two character states on character \( c \), contradicting the assumption that \( G_1 \) is a cluster. Similarly, \( \mathbf{a} \notin G_2 \). Hence, \( \mathbf{a}, \mathbf{b} \in S - G \) and there are two common character states between \( G \) and \( S - G \) for character \( c \). This contradicts the assumption that \( G \) is a cluster.

\( c \in \mathcal{D}(G) \): There are three subcases:

- \( c \in \mathcal{D}(G_1) \cap \mathcal{D}(G_2) \). Then, setting \( x_c = x_c^1 \) will not affect the compatibility of \( \mathbf{a} \) with \( \mathbf{a}^2 \).

- \( c \in \mathcal{M}(G_1) \cap \mathcal{M}(G_2) \). Then \( x_c = x_c^1 = x_c^2 \) as desired.
\[(c \in \mathcal{M}(G_1) \cap \mathcal{D}(G_2)) \text{ or } (c \in \mathcal{D}(G_1) \cap \mathcal{M}(G_2)).\] Impossible, since \(c \in \mathcal{D}(G)\).

Hence, \(x^1\) and \(x^2\) are compatible with \(x\) on \(\mathcal{M}(G_1)\) and \(\mathcal{M}(G_2)\) respectively. Note that if \(\mathcal{D}(G_1) = \emptyset\), then, rather than adding an edge \((x, x^1)\), we can simply identify nodes \(x^1\) and \(x\). A similar situation arises when \(\mathcal{D}(G_2) = \emptyset\). \(\square\)

**Lemma 6** Let \(G, G_1, G_2\) be clusters such that \(G = G_1 \cup G_2\) and \(G_1 \cap G_2 = \emptyset\). Suppose that \(G_1\) has a subphylogeny \(T_1\) and that there exists a subphylogeny \(T\) for \(G\) with \(T_1\) as a subtree at the connection \(x\) of \(T\). Then if \(G_2\) is not a c-cluster, the value of \(x_c\) on every \(c \in \mathcal{D}(G)\) is forced.

**Proof** We first show that \(\mathcal{M}(G_1) \cup \mathcal{M}(G) = \mathcal{C}\). Suppose there exists some \(c \notin \mathcal{M}(G_1) \cup \mathcal{M}(G)\). We claim that there can be no common state between \(G_2\) and \(S - G_2\) on \(c\). Suppose the common state is with a species in \(G_1\). Then \(c \in \mathcal{M}(G_1)\), a contradiction. We also arrive at a contradiction if we assume that the common state is with a species in \(S - G\). Thus, since there is no common state on \(c\), we must have \(c \in \mathcal{D}(G_2)\), contradicting the assumption that \(G_2\) is not a c-cluster.

Hence, for every \(c \in \mathcal{D}(G)\), we must have that \(c \in \mathcal{M}(G_1)\) and thus \(G_1\) and \(G_2\) share a common state on \(c\). Therefore, \(x_c\) must equal this common state. \(\square\)

### 4 The Algorithm

We now describe the algorithm \textsc{Phylogeny}, which constructs a phylogeny for \(S\), if it has one. The algorithm first tries to find if one of the \(O(2^{-1} \cdot m)\) c-splits is of type I. If there is a type I c-split \((G_1, G_2)\) where \(s\) is a connecting species, the algorithm recursively attempts to construct phylogenies \(T_1\) and \(T_2\) for \(G_1 \cup \{s\}\) and \(G_2 \cup \{s\}\). If one or both of the latter sets has no phylogeny, by Lemma 1, neither does \(S\). If both of the sets have phylogenies, then, a phylogeny for \(S\) is obtained by identifying the nodes for \(s\) in \(T_1\) and \(T_2\).

If there is no type I c-split, by Lemma 2, none of the species appears as an internal node in any phylogeny for \(S\). \textsc{Phylogeny} deals with this situation as a special case, by invoking a procedure called \textsc{Phase-II}, whose description shall occupy most of the remainder of this section. Before proceeding, however, we now summarize the steps of \textsc{Phylogeny}.

**Algorithm** \textsc{Phylogeny}(\(S\))

```
begin
  if \(|S| = 1\) then
    return the tree \(T\) consisting of the single species \(a \in S\)
  if there exists a type I c-split \((G_1, G_2)\) then
    Let \(s\) be the connecting species
    Call \textsc{Phylogeny}(\(G_1 \cup \{s\}\)) and \textsc{Phylogeny}(\(G_2 \cup \{s\}\))
    if both calls succeed then
      Combine the resulting trees into a phylogeny for \(S\)
    else return FAILURE
```
else Call Phase-II($\mathcal{S}$)
end

**Implementing Phase-II**

The heart of Phase-II is a procedure Subphylogeny that determines whether a cluster $G$ has a subphylogeny and, if so, constructs one. It assumes that for every c-cluster $G \subseteq G$, a subphylogeny has been constructed, if it exists.

**Algorithm Subphylogeny($G$)**

**Step 1.** If $|G| = 1$, then return the tree $T_G$ where $V(T_G) = G$, whose connection is the single node in $T_G$.

**Step 2.** Consider each c-cluster $G_1$ such that $G_1 \subseteq G$ and $G_1$ has a subphylogeny. Let $G_2 = G - G_1$. For each $G_1$, apply the following steps to find if there exists a subphylogeny for $G$ whose connection $x$ has a subphylogeny for $G_1$ as one of its subtrees.

**Step 2a.** If $G_2$ is a c-cluster having a subphylogeny, construct a subphylogeny $T_G$ for $G$ by applying the construction in the proof of Lemma 5. Return $T_G$.

**Step 2b.** If $G_2$ is not a c-cluster, compute the states of the connection $x$ as explained in Lemma 6. Let $R = \{H : H \subseteq G_2$ is a c-cluster that is compatible with $x$ and has a subphylogeny$\}$. Initialize $T_G$ to consist of $x$ with a subphylogeny for $G_1$ as its subtree and repeat the following step while both $R$ and $G_2$ are not empty.

**Add-Tree.** Choose any $H \in R$ and set $R = R - H$. If $H \subseteq G_2$, set $G_2 \leftarrow G_2 - H$ and make $T_H$ a subtree of $x$ in $T_G$, where $T_H$ is a subphylogeny for $H$.

Now, if $G_2 = \emptyset$, return $T_G$.

If all possible $G_1$’s have been exhausted, return FAILURE.

**Lemma 7** Let $G$ be a cluster and suppose that for every c-cluster $G'$ such that $G' \subseteq G$, we have determined whether $G'$ has a subphylogeny and, if so, one has been constructed. Then, if $G$ has a subphylogeny, Subphylogeny($G$) constructs it. Otherwise, the procedure returns FAILURE.

**Proof** When $|G| = 1$, a node for the single species $s \in G$ is indeed a subphylogeny for $G$. Hence, the tree returned in Step 1 is correct.

Suppose $|G| > 1$ and that $G$ has a subphylogeny $T_G$ with connection $x$. Then, there must exist a c-cluster $A \subseteq G$ having a subphylogeny $T_A$ such that $T_A$ is a subtree of $x$ in $T_G$. At some point during the execution of Step 2, $G_1$ will equal one such $A$. If $G_2 = G - G_1$ is a c-cluster having a subphylogeny, then, by Lemma 5, Step 2a returns a subphylogeny for $G$. If $G_2$ is not a c-cluster, then, by Lemma 6, the states of the connection $x$ are completely determined and, by Lemma 4, there exists a set of pairwise disjoint c-clusters.
$H_1, \ldots, H_k$ having subphylogenies such that (i) $G_1 \cup \bigcup_{i=1}^k H_i = G$ and (ii) $H_1, \ldots, H_k$ are compatible with $x$. By assumption, subphylogenies $T_1, \ldots, T_k$ for $H_1, \ldots, H_k$ have already been constructed. Furthermore, all the $H_j$'s will be contained initially in the set $R$ of Step 2b.

It is clear that at all times during the execution of Step 2b $R$ will contain every c-cluster $B$ such that $B \subseteq G_2$, $B$ has a subphylogeny, and $B$ is compatible with $x$. We shall show that in addition $R$ contains a set $\mathcal{N}$ of pairwise disjoint c-clusters such that $\bigcup \{ N : N \in \mathcal{N} \} = G_2$. As stated above, the initial $R$ has the desired properties. We now prove that $\text{Add-Tree}$ preserves the invariant. During each execution of $\text{Add-Tree}$, some $H$ is removed from $R$. If $H \not\in G_2$, then clearly it is not in $\mathcal{N}$, and the invariant is maintained. Otherwise, $H$ is subtracted from $G_2$. Let $R^0, G_2^0$ and $R^1, G_2^1$ be the sets $R$ and $G_2$ immediately before and immediately after an application of $\text{Add-Tree}$. Then, $R^1 = R^0 - \{ H \}$ and $G_2^1 = G_2^0 - H$. Let the subset $\mathcal{N}$ corresponding to $R^0$ be $\mathcal{N}^0$; we shall show that $R^1$ will have a corresponding subset $\mathcal{N}^1$. Now if $H \in \mathcal{N}^0$, $R^1$ clearly contains a subset $\mathcal{N}^1$ with the desired properties, namely $\mathcal{N}^1 = \mathcal{N}^0 - H$.

If $H \not\in \mathcal{N}^0$, then, since $\bigcup \{ N : N \in \mathcal{N}^0 \} = G_2^0$, there exists $\mathcal{G} \subseteq \mathcal{N}^0$ such that for each $J \in \mathcal{G}$, $J \cap H \neq \emptyset$ and $\bigcup \{ J : J \in \mathcal{G} \} \supseteq H$. We claim that for every $J \in \mathcal{G}$ there exist disjoint c-clusters compatible with $x$ whose union is $J - H$, such that each c-cluster has a subphylogeny.

Since $J$ is compatible with $x$, there exists a phylogeny for $J \cup \{ x \}$. Therefore, by Lemma 1, there exists a phylogeny $T'$ for $(J \cup \{ x \}) - H$. Now, let $x^1, \ldots, x^l$ be the neighbors of $x$ in $T'$ and for $1 \leq i \leq l$, let $T_i$ be the subtree of $T' - x$ containing $x^i$ and let $T_i = V(T_i) \cap \mathcal{S}$. Then, a subphylogeny can be constructed for each $T_i$ in much the same way as was done in the proofs of Lemmas 3 and 4. Observe too that each $T_i$ must be in $R^0$, since it is compatible with $x$. Let $\mathcal{N}^1$ consist of all these c-clusters for all $J \in \mathcal{G}$. Then $\mathcal{N}^1 = (\mathcal{N}^0 - \mathcal{G}) \cup \mathcal{N}^0$ has the required properties.

Thus, assuming we have chosen the right $G_1$, a subphylogeny for $G$ will be found. If no choice of $G_1$ yields a subphylogeny for $G$, then $G$ has no subphylogeny and Step 2 correctly returns FAILURE. $\square$

Subphylogeny is invoked by the following algorithm.

Algorithm Phase-II

Step 1. Enumerate all c-clusters of size at most $n - 1$ by non-decreasing order of cardinality. For each c-cluster $G$, determine whether $G$ has a subphylogeny $T_G$. If so, record $T_G$ and its connection.

Step 2. For each $s \in \mathcal{S}$, determine if $G = \mathcal{S} - \{ s \}$ has a subphylogeny. If so, let $T_G$ be the tree and $x$ be its connection. Return the tree obtained by adding a node $s$ and the edge $(s, x)$ to $T_G$.

Theorem 8 If $\mathcal{S}$ has a phylogeny, then Phase-II finds one.

Proof Suppose there exists a phylogenetic tree $T$ for $\mathcal{S}$. Let $s$ be any species that is a leaf in $T$. Then $\mathcal{S} - \{ s \}$ is a c-cluster and has a subphylogeny. By Lemma 7, Phase-II computes subphylogenies for all c-clusters $G$ such that $|G| \leq n - 1$ and, in particular, it
finds a subphylogeny for $S = \{s\}$. Hence, at some point, Step 2 of Phase-II obtains a phylogeny for $S$. □

**Analysis**

We first establish the correctness of our algorithm.

**Theorem 9** **Phylogeny** correctly determines whether or not $S$ has a phylogeny and, if so, constructs one.

**Proof**  Follows from Theorem 8, and the fact that we are dealing with type I c-splits properly. □

Next, we consider the running time of **Phylogeny**. First, we analyze the running time of a call to **Subphylogeny**$(G)$. This procedure will consider each of the $O(2^r m)$ c-clusters $G_1$ such that $|G_1| < |G|$. For each such c-cluster, it verifies that $G_1 \subset G$, which can be done in $O(n)$ time. With a particular $G_1$, the algorithm goes through $O(2^r m)$ c-clusters, checking in $O(n + m)$ time whether they are subsets of $G_2$ that are compatible with $x$. The total time of **Subphylogeny** is therefore $O(2^{2r}(nm^2 + m^3))$.

In Phase-II, we generate all c-clusters, which takes $O(2^r nm^2)$ time. Since **Subphylogeny** is applied to each c-cluster, Phase-II takes $O(2^{3r}(nm^3 + m^4))$ time. **Phylogeny** spends $O(2^{r}nm^2)$ time generating c-clusters and testing each of these to find out whether it is of type I. It is clear that, in the worst case, the running time of **Phylogeny** is dominated by the time required to deal with the case where all c-clusters are of type II. Hence, the running time of **Phylogeny** is $O(2^{3r}(nm^3 + m^4))$.

**5 Remarks**

Our algorithm constructs a phylogeny by working from the bottom up, following what is essentially a dynamic programming approach. One can use memoization (a technique described in some detail in pp. 312–314 of [3]) to obtain an equivalent top-down recursive algorithm with the same running time. Such a procedure has been proposed to us by E.L. Lawler (personal communication).

Algorithm **Phylogeny** can be modified to work correctly and within the same time bounds even if instances with type I c-splits are not treated separately. However, in practice, identifying such splits and exploiting their presence may tend to reduce the running time of the algorithm when the phylogeny has one or more elements of $S$ as internal nodes.

Kannan and Warnow (personal communication) have discovered a clever way to reduce the running time of our algorithm by a factor of $2^r$. Their technique speeds up Step 2 of **Subphylogeny** by providing a way to determine in $O(nm)$ time whether there exists a subphylogeny for $G$ having a subphylogeny for a given c-cluster $G_1$ as a subtree. Even with this improvement, the algorithms presented in [10] and [4] are faster than ours for the cases where $r \leq 4$ and $r \leq 3$, respectively. It is an open problem whether our algorithm can be improved to match those bounds on those special cases.
Acknowledgements

We thank Martin Farach, Sampath Kannan, and Tandy Warnow for their encouragement and Mike Steel for bringing his results to our attention. Special thanks are due to Eugene Lawler, who carefully read our manuscript and suggested changes in the terminology.

References


