Refining Phylogenetic Trees Given Additional Data: An Algorithm Based on Parsimony

Taoyang Wu, Vincent Moulton, and Mike Steel

Abstract—Given a set X of taxa, a phylogenetic X-tree T that is only partially resolved, and a collection of characters on X, we consider the problem of finding a resolution (refinement) of T that minimizes the parsimony score of the given characters. Previous work has shown that this problem has a polynomial time solution provided certain strong constraints are imposed on the input. In this paper, we provide a new algorithm for this problem and show that it is fixed-parameter tractable under more general conditions.

Index Terms—Maximum parsimony, Fitch-Hartigan algorithm, optimal tree refinement, hitting set problem, fixed parameter tractability.

1 Introduction

One of the most basic methods for phylogenetic tree reconstruction is maximum parsimony (MP). The starting point for constructing the most parsimonious tree on a set X of n taxa is a collection $(\chi_1, \ldots, \chi_k)$ of r-state characters, i.e., functions χ_i from X into some discrete set S_i for which $|\chi_i(X)| \leq r$ for all i. In applications, the case $r = 2$ typically corresponds to presence-absence of a character state, while $r = 4$ frequently corresponds to the four nucleotides. The problem is to find a *phylogenetic X-tree T* (i.e., a tree $T = (V, E)$ with vertex set V, edge set E, and leaf set $X \subseteq V$) and extensions $\overline{\chi}_i : V \to S_i$ for each character χ_i, so that the total number of changes of states along the edges of T taken over all extensions is minimized over all possible trees and extensions. This problem, the *MP problem*, is **NP**-hard even for $r = 2$, [5], [8], and various methods have been proposed for its solution (see, e.g., [6]).

In [3], the following variant of the MP problem was introduced (precise definitions follow in Section 2). A phylogenetic X-tree T' *refines* T if T can be obtained by contracting some edges in T'. Now, suppose that T is a phylogenetic X-tree with maximum vertex degree d, and that, as above, $(\chi_1, \ldots, \chi_k)$ is a collection of r-state characters on X. The *optimal parsimony refinement (OPR) problem* consists of the following two parts: calculating the minimal parsimony score among all refinements of T (the **OPR**-score problem) and finding a refinement of T' that has such score (the **OPR**-tree problem).

- *T. Wu is with the Department of Computer Science and School of Mathematical Sciences, Queen Mary, University of London, London E1 4NS, UK. E-mail: Taoyang.Wu@dcs.qmul.ac.uk.*
- *V. Moulton is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK. E-mail: vincent.moulton@cmp.uea.ac.uk.*
- *M. Steel is with the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand. He is also with the Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: m.steel@math.canterbury.ac.nz.*

Manuscript received 19 Feb. 2008; revised 1 July 2008; accepted 5 Sept. 2008; published online 24 Sept. 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2008-02-0034.
Digital Object Identifier no. 10.1109/TCBB.2008.102.

A simple instance of the **OPR** problem is illustrated in Fig. 1, where we have just a single character (rather than a sequence of characters) and the state this character assigns to each leaf is denoted by its subscript. Notice that the parsimony score of this character for the tree in Fig. 1a is 4, while for the refinement of this tree shown in Fig. 1b, the parsimony score is 3.

The biological motivation for considering the **OPR** problem is given as follows: Often, in applications, one can be confident in the historical correctness of a partially resolved tree, and one would like to resolve the tree further by "expanding" vertices of degree at least 4. These higher degree vertices (called "polytomies") can arise from a variety of biological processes: for example, a rapid speciation event, or a divergence event deep in the past, or from other processes (such as lineage sorting) that cause conflicting phylogenetic signal. The partially resolved tree would typically be some consensus (summary) tree showing well-supported phylogenetic splits, and based on trees obtained from various data sets and/or methods (not necessarily based on MP). One wishes to use additional informative characters to help resolve these vertices. The use of the MP criterion for this purpose can be reasonable when these additional characters concern rare genomic changes where homoplasy (reverse or convergent evolution) is infrequent [16]. For example, this approach has been employed for analyzing short interspersed nuclear element (SINE) data, which describe the presence or absence of an insertion of a particular DNA sequence into a specific location of the genomes of the species under study [14]. Such markers have been used to refine the mammal tree, providing, for example, evidence for the sister relationship between whale and hippopotamus [13]. Note that the number k of characters provided by SINE (and related) data is typically quite small, which is relevant to the running time of our algorithm as we describe below.

The **OPR** problem is **NP**-hard even when $r = 2$ (taking T to be the star tree, the **OPR** problem reduces to the MP problem). However, in [3, Theorem 3], it was shown that the **OPR** problem can be solved in $O\left(nkr^{k(d+1)+1}df(d)\right)$ time, where $f(d) := (2d-3)!! := (2d-3)(2d-5)\ldots 3$ (this is the number of rooted binary phylogenetic trees with d leaves). In particular, the **OPR** problem is fixed-parameter tractable

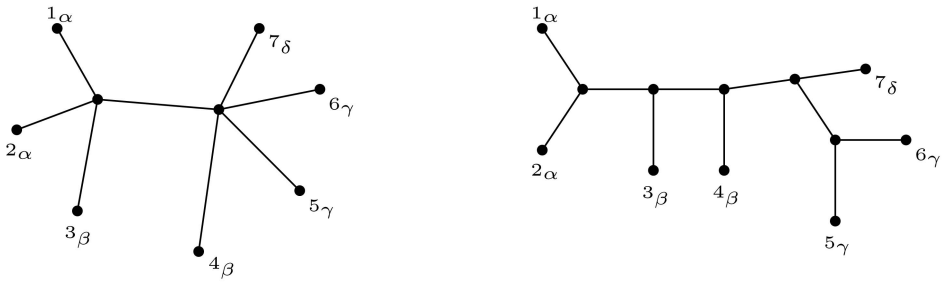


Fig. 1. (a) A partially resolved phylogenetic X-tree, for $X = \{1, 2, \ldots, 7\}$, and a single character on four states, where x_θ denotes that leaf x receives state θ for $\theta \in \{\alpha, \beta, \gamma, \delta\}$. (b) An OPR of the tree in (a) for this single character.

[12], in that it can be solved in $n \cdot h(k, r, d)$ time, where h is a function that grows exponentially in k, r, and d.

In this paper, we shall present an alternative algorithm to solve the **OPR** problem that is fixed-parameter tractable with respect to just k and r, that is, it can be used to solve the **OPR** problem in polynomial time in n when k and r are fixed, even for large d. More particularly, while [3, Theorem 3] requires all of k, d, and r to be bounded to obtain an algorithm that runs in polynomial time in n, our result requires that k and either d or r be bounded for polynomial time in n. Thus, our algorithm is suited to applications in which one wishes to refine a partially resolved phylogenetic tree using a small number of low-homoplasy characters (such as occurs with SINE data). Provided the characters are binary (for example, showing presence or absence of some features), or at least near binary (i.e., if r is small), or if the tree is already moderately resolved (so, d is small), the algorithm has a desirable running time.

Summarizing, our main result is the following.

Theorem 1.1. *Suppose that T is a phylogenetic X-tree on an n-set X with maximum vertex degree d and that $(\chi_1, \ldots, \chi_k)$ is a collection of r-state characters on X. Then, the* **OPR** *problem can be solved in*

$$O\Big(ndkrs\binom{m}{s}f(s)\Big)$$

time, where $m = r^k$ and $s = \min(d, r^k)$. In particular, the **OPR** *problem can be solved in $O\left(ndg(k, r)\right)$ time with g some function of k and r.*

The proof of Theorem 1.1 will be presented in Section 6. Note that this theorem provides an improved complexity bound to the bound $O\left(nkr^{k(d+1)+1}df(d)\right)$ given by [3, Theorem 3]: if $d < m = r^k$, then the ratio of two asymptotics is

$$\frac{nkr^{k(d+1)+1}df(d)}{nd^2kr\binom{r^k}{d}f(d)} = \frac{r^{k(d+1)}}{\binom{r^k}{d}d} \approx r^k(d-1)!,$$

for the other case $d > m = r^k$, the ratio is

$$\frac{nkr^{k(d+1)+1}df(d)}{ndkr^{k+1}f(r^k)} = \frac{r^{kd}f(d)}{f(k)} \geq r^{kd}.$$

Compared to the exact algorithm in [3], the main difference in our approach is to introduce a local version of the **OPR**-score problem (see Section 3 for details), which provides a smaller search space and which allows us to solve the **OPR** problem by working up through the vertices of a rooted phylogenetic tree from the leaves to the root, in a similar manner to the well-known Fitch-Hartigan algorithm [7], [10].

The remainder of this paper is organized as follows: In Section 2, we present some preliminaries. Then, in Section 3, we present the local version of the **OPR**-score problem. In Sections 4 and 5, we present an optimization problem for bipartite graphs, a generalized hitting set (**GHS**) problem, and show that the local **OPR**-score problem can be solved as a special instance of this problem. We then present an algorithm for solving the **OPR** problem in Section 6 and prove Theorem 1.1. We conclude in Section 7 with a discussion of the **OPR** problem for some special families of characters and also outline future directions.

2 Preliminaries

We refer readers unfamiliar with this area to [15] for further details concerning phylogenetic trees that we shall use freely in this paper.

To ease notation, we will consider a collection $(\chi_1, \ldots, \chi_k)$ of r-state characters to be a map χ from X to the set $\Gamma = \Gamma_{k,r}$ consisting of the set of words of length k over the alphabet $\{0, 1, \ldots, r-1\}$ (i.e., Γ is the sequence space of length k sequences over the alphabet $\{0, 1, \ldots, r-1\}$). We also refer to such maps χ as (k, r)-*characters*. Note that in solving either the MP or **OPR** problem the state space for each character could be much larger than r since we can always restrict to the states we see at the leaves, namely $\{0, 1, \ldots, r-1\}$. We also denote by d_H the Hamming distance on Γ, that is, the metric defined, for $\gamma, \gamma' \in \Gamma$, by

$$d_H(\gamma, \gamma') := \big|\{i : 0 \leq i \leq k-1 \text{ and the } i\text{th letter } \gamma_i \text{ and } \gamma'_i \text{ are different}\}\big|.$$

An *extension* of a character $\chi : X \to \Gamma$ to a phylogenetic X-tree T is a function $\bar{\chi}$ from the vertex set of T, denoted by $V(T)$, to Γ, such that $\bar{\chi}(v) = \chi(v)$ for any leaf v of T. Given $e = \{u, v\} \in E(T)$, we set

$$\Delta(e, \bar{\chi}) := d_H(\bar{\chi}(u), \bar{\chi}(v))$$

and define the *changing score* of $\bar{\chi}$, denoted by $\mathrm{ch}(\bar{\chi})$, to be

$$\mathrm{ch}(\bar{\chi}) := \sum_{e=\{u,v\}\in E(T)} d_H(\bar{\chi}(u), \bar{\chi}(v)) = \sum_{e=\{u,v\}\in E(T)} \Delta(e, \bar{\chi}).$$

Now, given a phylogenetic X-tree T and a character χ on X, its *parsimony score* is defined as

$$l(\chi, T) := \min\{\mathrm{ch}(\bar{\chi}) \,:\, \bar{\chi} \text{ is an extension of } \chi \text{ on } T\}.$$

An extension $\bar{\chi}$ that achieves this score, i.e., an extension $\bar{\chi}$ with $\mathrm{ch}(\bar{\chi}) = l(\chi, T)$, is called a *minimum extension* of χ to T. Given a character χ on X, the *maximal parsimony score* of χ is defined as

$$l(\chi) := \min\{l(\chi, T) \,:\, T \text{ is a phylogenetic } X\text{-tree}\}.$$

Note that $l(\chi)$ is determined only by $\chi(X)$, and hence, we will also denote it by $l(\chi(X))$ without mentioning X explicitly. Correspondingly, the phylogenetic X-tree T that minimizes $l(\chi, T)$ is called a *maximal parsimony tree* for χ. In this notation, the MP problem can be stated as follows:

Problem 1. MP.
Input: A (k, r)-character χ on a finite set X of size n.
Task: Calculate $l(\chi)$ and find a maximal parsimony tree T for χ.

As mentioned in Section 1, this problem is **NP**-hard (even for the $(k, 2)$-characters) [5], [8].

We conclude this section with a statement of the **OPR** problem using our new notation. If T and T' are phylogenetic X-trees such that T' refines T, we shall write $T' \geq T$. The *OPR-score* for a pair (χ, T) consisting of a character and a phylogenetic tree T is then

$$rl_T(\chi) := \min_{T'\geq T} l(\chi, T') = \min_{T'\geq T} \min_{\bar{\chi}} \mathrm{ch}(\bar{\chi}, T').$$

A tree T' achieving such score is called an *OPR-tree* of (χ, T).

Problem 2. OPR.
Input: A phylogenetic X-tree T and a (k, r)-character χ on X.
Tasks: Calculate $rl_T(\chi)$ and construct an OPR-tree for (χ, T).

As indicated in Section 1, to distinguish between the two tasks in the above problem, we shall call them the **OPR**-score problem and **OPR**-tree problem, respectively. Note that if T' is an **OPR**-tree of (χ, T) and $\bar{\chi}$ is a minimum extension of χ to T', then $rl_T(\chi) = \mathrm{ch}(\bar{\chi}, T')$, and we call $(\bar{\chi}, T')$ an **OPR**-pair.

3 The Local OPR-Score Problem

In this section, we shall introduce a local version of the **OPR**-score problem, which is defined on rooted phylogenetic X-trees, and show that this problem is well defined. Note that a *rooted phylogenetic X-tree* T is a phylogenetic X-tree with a distinguished leaf, called the *root*, which we usually denote by $\rho = \rho_T$. Note also that our concepts concerning **MP** and **OPR** extend naturally to rooted trees, and so, we will only point out key differences where necessary. We refer the reader to [15] for more details concerning such trees.

Now, given a rooted phylogenetic X-tree T and a vertex v of T, we define T_v to be the subtree of T lying below v together with an additional root vertex ρ_v connected by an additional edge to v, unless $v = \rho$ in which case we set $T_v = T$ and $\rho_v = \rho$. Given a character $\chi : X \to \Gamma$, we associate a *valid pair* to v,

$$\mathbb{S}_v = \mathbb{S}_{v,\chi} := (\mathrm{cost}(v), \Gamma_v)$$

consisting of a natural number $\mathrm{cost}(v)$ together with a collection of nonintersecting subsets Γ_v of Γ. In particular, given $\gamma \in \Gamma$, we define $\mathrm{cost}(v, \gamma)$ to be the **OPR**-score of (χ, T_v) with ρ_v labeled by γ, thus

$$\mathrm{cost}(v) := \min_{\gamma\in\Gamma} \mathrm{cost}(v, \gamma),$$

and we set

$$\Gamma_v := \{\Gamma_v^0, \ldots, \Gamma_v^{k-1}\},$$

where, for any nonnegative integer i,

$$\Gamma_v^i := \{\gamma \in \Gamma \,:\, \mathrm{cost}(v, \gamma) = \mathrm{cost}(v) + i\}.$$

Note that if v is a leaf of T, then we have

$$\mathrm{cost}(v, \gamma) = d_H(\gamma, \chi(v)) \tag{1}$$

for all $\gamma \in \Gamma$ (so that $\mathrm{cost}(v) = 0$ and $\Gamma_v^0 = \{\chi(v)\}$ as well). Note that given a rooted phylogenetic tree T and a character χ on X, the valid pair $\mathbb{S}_v$ for each vertex $v \in V(T)$ is unique.

The local **OPR**-score problem is now stated as follows:

Problem 3. Local **OPR**-score.
Input: A rooted phylogenetic X-tree T, a (k, r)-character χ on X, an internal vertex $v \in V(T)$, and the valid pair $\mathbb{S}_{v'}$ for each child v' of v.
Task: Determine the valid pair $\mathbb{S}_v = (\mathrm{cost}(v), \Gamma_v)$ for v.

In the remainder of this section, we shall show that this problem is well defined, that is, a valid pair $\mathbb{S}_v$ for v can always be determined from the input data.

To this end, given a subset A of X (i.e., a subset A of the leaf set of T), we let $\mathrm{lca}_T(A)$ denote the *most recent common ancestor* of A in T (that is, the vertex in T furthest below the root that lies above every element of A). Note that if T' is a (binary) refinement of T, then we can define a canonical mapping $\phi : V(T) \to V(T')$, by setting $\phi(v)$ to be the most recent common ancestor in T' of the subset of elements in X that lie below v in T. Clearly, the map ϕ is injective. In addition, it induces a mapping from the set of rooted subtrees T_v of T ($v \in V(T)$) to the set of rooted subtrees of T', for which the image of T_v is $T'_{\phi(v)}$ (which we will also denote by T'_v in case no confusion may arise). Now, we state a lemma concerning valid pairs.

Lemma 3.1. *Let T be a rooted phylogenetic X-tree and let $\chi : X \to \Gamma$ be a character, together with a valid pair $\mathbb{S}_v = (\mathrm{cost}(v), \Gamma_v)$ for each $v \in V(T)$. Then, there exists an* **OPR***-pair $(\bar{\chi}, T')$ such that*

$$\bar{\chi}(\phi(v)) \in \bigcup_{\Gamma'\in\Gamma_v} \Gamma'$$

for all $v \in V$.

Proof. For any **OPR**-pair $(\bar{\chi}, T')$ of (χ, T), set $v' := \phi(v)$ and

$$\mathrm{dis}(\bar{\chi}, T') := \left\{ v \in V(T) \,:\, \bar{\chi}(v') \notin \bigcup_{\Gamma' \in \Gamma_v} \Gamma' \right\}.$$

Note that, by definition, $\mathrm{dis}(\bar{\chi}, T')$ does not contain the root ρ_T or any leaf of T. Furthermore, it does not contain the unique child of ρ_T in T.

Using this notation, the lemma is equivalent to the following claim: There exists an **OPR**-pair $(\bar{\chi}, T')$ of (χ, T) such that $\mathrm{dis}(\bar{\chi}, T') = \emptyset$ **holds**. We shall prove this claim using induction on $h(T)$, the height of T (i.e., the length of a longest path from any leaf of T to ρ_T).

Clearly, this claim holds for the base case $h(T) = 1$ (that is, when T is a tree with one edge). So, assume that the claim holds for all phylogenetic X-trees T with $h(T) \leq q$ for some integer $q \geq 1$.

Suppose that T is a tree with height $q+1$ and that the claim fails for a pair (χ, T). Let $(\bar{\chi}, T')$ be an **OPR**-pair of (χ, T) that minimizes the cardinality of $\mathrm{dis}(\bar{\chi}, T')$. Furthermore, let $v \in \mathrm{dis}(\bar{\chi}, T')$ be a vertex of T such that $h(T_v)$ is minimal. Note that $1 < h(T_v) < q+1$ **holds** since $\mathrm{dis}(\bar{\chi}, T')$ does not contain any leaf of T, ρ_T, or the child of ρ_T.

Now, let T^* be the tree obtained from T' by pruning off T'_v, and let u' be the direct ancestor of v' in T'. Then,

$$\mathrm{ch}(\bar{\chi}, T') = \Delta(\{u', v'\}, \bar{\chi}) + \sum_{e \in E(T'_v)} \Delta(e, \bar{\chi}) + \sum_{e \in E(T^*)} \Delta(e, \bar{\chi}).$$

Moreover, by the induction assumption, there exists an **OPR**-pair $(\widetilde{\chi}, \widetilde{T}_v)$ of (χ, T_v) such that $\mathrm{dis}(\widetilde{\chi}, \widetilde{T}_v) = \emptyset$. Since $v \notin \bigcup_{\Gamma' \in \Gamma_v} \Gamma'$, it follows that

$$\mathrm{ch}(\widetilde{\chi}, \widetilde{T}_v) = \mathrm{cost}(v) \leq -k + \sum_{e \in E(T'_v)} \Delta(e, \bar{\chi}).$$

Now, consider the tree T'' that is obtained from T' by replacing T'_v with $\widetilde{T}_v$. Let v'' be the image of v under the canonical mapping ϕ between $V(T)$ and $V(T'')$. Furthermore, let χ'' be a map on T'' defined by

$$\chi''(v) = \begin{cases} \widetilde{\chi}(v), & \text{if } v \in \widetilde{T}_v, \\ \bar{\chi}(v), & \text{else.} \end{cases}$$

Then, we have

$$\begin{aligned} \mathrm{ch}(\chi'', T'') - \mathrm{ch}(\bar{\chi}, T') &= d_H(\widetilde{\chi}(v''), \bar{\chi}(u')) + \mathrm{ch}(\widetilde{\chi}, \widetilde{T}_v) \\ &\quad - \Delta(\{u', v'\}, \bar{\chi}) - \sum_{e \in E(T'_v)} \Delta(e, \bar{\chi}) \\ &\leq -k + d_H(\widetilde{\chi}(v''), \bar{\chi}(u')) - \Delta(\{u', v'\}, \bar{\chi}) \\ &\leq 0. \end{aligned}$$

Conversely, $\mathrm{ch}(\bar{\chi}, T') \geq \mathrm{ch}(\chi'', T'')$ holds since $(\bar{\chi}, T')$ is an **OPR**-pair. Hence, we obtain $\mathrm{ch}(\chi'', T'') = \mathrm{ch}(\bar{\chi}, T')$ so that (χ'', T'') is an **OPR**-pair of (χ, T) as well. But, this implies $|\mathrm{dis}(\chi'', T'')| < |\mathrm{dis}(\bar{\chi}, T')|$, a contradiction to the minimality of $\mathrm{dis}(\bar{\chi}, T')$. This completes the proof of the induction step and, hence, of the lemma. □

For an internal vertex v of T with children $\{v_1, \ldots, v_{p-1}\}$, we now describe how to obtain the valid pair $\mathbb{S}_v$ for v from the set of valid pairs $\mathbb{S}_{v_i}$, $1 \leq i \leq p-1$. First, define the set of $(p-1)$-tuples

$$A_v := \left\{ (a_1, a_2, \ldots, a_{p-1}) \,:\, a_i \in \bigcup_{\Gamma' \in \Gamma_{v_i}} \Gamma' \right\},$$

noting that A_v depends only on the valid pairs $\mathbb{S}_{v_i}, 1 \leq i \leq p-1$. Now, for $\underline{a} = (a_1, a_2, \ldots, a_{p-1}) \in A_v$, set $\kappa'(a_i) = t$ for the unique t in $\{0, \ldots, k-1\}$ such that $a_i \in \Gamma^t_{v_i}$ holds, and

$$\kappa(\underline{a}) := \sum_{i=1}^{p-1} \kappa'(a_i).$$

In addition, for $\gamma \in \Gamma$, let $\xi_{\underline{a},\gamma} : \{0, 1, \ldots, p-1\} \to \Gamma$ be the character

$$\xi_{\underline{a},\gamma}(i) := \begin{cases} \gamma, & \text{if } i = 0, \\ a_i, & \text{otherwise.} \end{cases}$$

We now show, for $\gamma \in \Gamma$, how we can compute $\mathrm{cost}(v, \gamma)$ using A_v and $\mathrm{cost}(v_i)$, $1 \leq i \leq p-1$. Since the valid pair $\mathbb{S}_v$ can be directly deduced from the set of scores $\{\mathrm{cost}(v, \gamma) \,:\, \gamma \in \Gamma\}$, it immediately follows that the local **OPR**-score problem is well defined (Fig. 2).

v	$\mathrm{cost}(v)$	Γ^0_v	Γ^1_v
u_1	3	{00}	{01,10}
u_2	2	{02,11}	{20,12,22}
u_3	5	{20,11}	{22,21}

Fig. 2. Assuming $\Gamma = \{00, 01, 02, 10, 11, 12, 20, 21, 22\}$ (i.e., $k = 3$ and $r = 2$) and considering an internal vertex $u \in V(T)$ with valid sets associated to its children $\{u_1, u_2, u_3\}$ as given in the table, then we have $A_u = \{(a_1, a_2, a_3) : a_1 \in \{00, 01, 10\},\ a_2 \in \{02, 11, 20, 12, 22\},\ a_3 = \{20, 11, 22, 21\}\}$. For the element $\underline{a} = (00, 20, 21)$ in A_u, we have $\kappa(\underline{a}) = \kappa'(a_1) + \kappa'(a_2) + \kappa'(a_3) = 0 + 1 + 1 = 2$.

Proposition 3.2. *Suppose that T is a rooted phylogenetic X-tree and $\chi : X \to \Gamma$ is a character. For an internal vertex $v \in V(T)$, let $\{v_1, v_2, \ldots, v_{p-1}\}$ denote the set of children of v. Then*

$$\mathrm{cost}(v, \gamma) = \min\{ l(\xi_{\underline{a},\gamma}) + \kappa(\underline{a}) \,:\, \underline{a} \in A_v \} + \sum_{i=1}^{p-1} \mathrm{cost}(v_i). \quad (2)$$

Proof. Given $\gamma \in \Gamma$, note that

$$\mathrm{cost}(v, \gamma) \leq l(\xi_{\underline{a},\gamma}) + \sum_{i=1}^{p-1} \mathrm{cost}(v_i, a_i) = l(\xi_{\underline{a},\gamma}) + \kappa(\underline{a}) + \sum_{i=1}^{p-1} \mathrm{cost}(v_i)$$

holds for any $\underline{a} = (a_1, a_2, \ldots, a_{p-1}) \in A_v$ by definition; therefore, to prove that (2) holds, it suffices to show that

$$\mathrm{cost}(v, \gamma) \geq l(\xi_{\underline{a},\gamma}) + \kappa(\underline{a}) + \sum_{i=1}^{p-1} \mathrm{cost}(v_i)$$

holds for some $\underline{a} \in A_v$.

Let $(\bar{\chi}, T'_v)$ be an **OPR**-pair for (χ, T_v) with root labeled by γ such that $\mathrm{ch}(\bar{\chi}, T'_v) = \mathrm{cost}(v, \gamma)$ and

$$\bar{\chi}(u') \in \bigcup_{\Gamma' \in \Gamma_u} \Gamma'$$

holds for each vertex u in T_v, which exists by Lemma 3.1.

Now, for each $1 \leq i \leq p-1$, set $a_i = \bar{\chi}(\phi(v_i))$. Then,

$$\underline{a} = (a_1, \ldots, a_{p-1}) \in A_v$$

and

$$\mathrm{cost}(v_i) + \kappa'(a_i) = \mathrm{cost}(v_i, a_i) \leq \mathrm{ch}\Big(\bar{\chi}|_{V(T'_{v_i})}, T'_{v_i}\Big),$$

where the root of T'_{v_i} is labeled by a_i. Moreover, for the tree T^* obtained by pruning the trees T'_{v_i} from T'_v (i.e., deleting all edges and vertices below v_i) for all $i \in \{1, 2, \ldots, p-1\}$, the inequality

$$l(\xi_{\underline{a},\gamma}) \leq \mathrm{ch}\Big(\bar{\chi}|_{V(T^*)}, T^*\Big)$$

clearly holds. Therefore, we have

$$\begin{aligned}
\mathrm{cost}(v,\gamma) &= \mathrm{ch}\big(\bar{\chi}, T'_v\big) \\
&= \mathrm{ch}\Big(\bar{\chi}|_{V(T^*)}, T^*\Big) + \sum_{i=1}^{p-1} \mathrm{ch}\Big(\bar{\chi}|_{V(T'_{v_i})}, T'_{v_i}\Big) \\
&\geq l(\xi_{\underline{a},\gamma}) + \sum_{i=1}^{p-1} (\mathrm{cost}(v_i) + \kappa'(a_i)) \\
&= l(\xi_{\underline{a},\gamma}) + \kappa(\underline{a}) + \sum_{i=1}^{p-1} \mathrm{cost}(v_i),
\end{aligned}$$

as required. □

4 A Generalized Hitting Set Problem

We now concentrate on finding an efficient algorithm to solve the local **OPR**-score problem. To do this, in this section, we introduce an optimization problem for bipartite graphs, called the **GHS** problem, and describe a simple way to solve it together with complexity bounds. In the next section, we show how the local **OPR**-score problem can be reduced to the **GHS** problem.

To this end, suppose that $G = (U \sqcup W, E)$ is a bipartite graph with $W := \Gamma$, $\theta : U \to \mathbb{N}$ as a function and that $\{U^0, U^1, \ldots, U^{p-1}\}$ is a partition of U for a positive integer p. For F any subset of E and U' any subset of U (W' any subset of W), we denote by U'_F (respectively, W'_F) the set of vertices in U' (respectively, W') that are contained in some edge of F. In addition, for $0 \leq i \leq p-1$, we set $u^i_F := \min\{\theta(u) : u \in U^i_F\}$ if $U^i_F \neq \emptyset$ and $u^i_F := \infty$ else.

Now, for any subset W' of W, we define

$$\omega(F) := l(W_F) + \sum_{j=0}^{p-1} u^j_F$$

and

$$\omega^*(G) := \min\{\omega(F) : F \subseteq E\},$$

and we say that F is a *minimal hitting set* if $|F| = p$ and $\omega(F) = \omega^*(G)$.

Problem 4. GHS.
Input: A bipartite graph $G = (U \sqcup W, E)$ with a map $\theta : U \to \mathbb{N}$, and a partition $\{U^0, \ldots, U^{p-1}\}$ of U.
Task: Calculate $\omega^*(G)$.

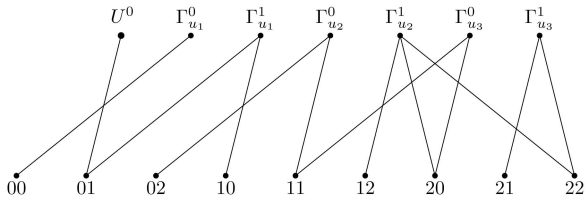


Fig. 3. The bipartite graph constructed for the data given in Fig. 2 with $\gamma = 01$. Here, $U^i = \{\Gamma^0_{u_i}, \Gamma^1_{u_i}\}$ for $i = 1, 2, 3$.

Note that in case $\Gamma = \Gamma_{1,r}$ and $\theta(u) = 0$ for all $u \in U$, the **GHS** problem is equivalent to the well-known Hitting Set problem (see [9] and [12] and the references therein).

Now, for $Y \subseteq W$, define E_Y to be the subset of E consisting of those edges in E that are incident to some vertex in Y. Then, the **GHS** problem can be solved by simply computing $\omega(E_Y)$ for each subset $Y \subseteq W$ with $|Y| \leq p$, and then setting

$$\omega^*(G) = \min\{\omega(E_Y) : Y \subseteq W \text{ and } |Y| \leq p\}.$$

Indeed, for $F \subseteq E$, we have $F \subseteq E_{W_F}$, i.e., $\omega(F) \geq \omega(E_{W_F})$. And, on the other hand, for each subset $Y \subseteq W$, if $\omega(E_Y) < \infty$, we can construct a subset F of E_Y, with $|F| = p$ and $\omega(F) \leq \omega(E_Y)$.

In view of these considerations, we obtain the following lemma.

Lemma 4.1. *For any instance of the* **GHS** *problem with $|W| = r^k$ and U consisting of p blocks, where the cardinality of each block in U is bounded by t, the* **GHS** *problem can be solved in*

$$\sum_{i=1}^{\min(p, r^k)} \binom{r^k}{i} (kri(2i-5)!! + pti)$$

time.

Proof. Let $Y \subseteq W$, and set $y := |Y|$. First, note that to compute $\omega(E_Y)$ we need to calculate $l(Y)$, i.e., we need to solve the **MP** problem for the canonical character on Y that associates each element in Y with itself. This can be done in $O(kry(2y-5)!!)$ time, by making an exhaustive search through all $(2y-5)!!$ unrooted binary trees with leaf set Y [15] and applying the Fitch-Hartigan algorithm [7], [10] to each tree, which takes $O(kry)$ time.

Now, to complete the proof, note that there at most ty edges in E that are incident with U^j and Y for each $0 \leq j \leq p-1$. Therefore, we can calculate $u^j_{E_Y}$ in ty time. In particular, $\omega(Y)$ can be computed in $kry(2y-5)!! + pty$ time. □

5 A Solution to the Local OPR-Score Problem

In this section, we show that the local **OPR**-score problem can be solved as a special instance of the **GHS** problem. Using Lemma 4.1, this will also give us a bound on the complexity for solving the local **OPR**-score problem.

Suppose that T is a rooted phylogenetic X-tree, v is an internal vertex in $V(T)$ with children $v_1, \ldots, v_{p-1}$, and that we are given the valid pairs $\mathbb{S}_{v_i}$ for $1 \leq i \leq p-1$. Given $\gamma \in \Gamma$, we now use this data to define a bipartite graph $G_{v,\gamma} = (U \sqcup W, E)$ (see Fig. 3 for an example), together with

Algorithm I
Input: A phylogenetic X-tree T, and a (k, r)-character $\chi : X \to \Gamma$.
Output: The **OPR**-score $rl_T(\chi)$ and an **OPR**-tree having this score.
1. Arbitrarily introduce a root vertex ρ and corresponding edge to T to form a rooted tree T^ρ.
2. Working up from the leaves to the root, assign to each vertex $v \in V(T^\rho)$ the valid pair $\mathbb{S}_v = (\mathrm{cost}(v), \Gamma_v)$ and the set $\{T'_{v,\gamma} : \gamma \in \bigcup_{\Gamma' \in \Gamma_v} \Gamma'\}$, where $T'_{v,\gamma}$ a refinement of T_v whose root is labelled by $\gamma \in \Gamma$ and that satisfies $\mathrm{ch}(T'_{v,\gamma}, \chi) = \mathrm{cost}(T_v, \gamma)$, as follows.
2a. For each leaf v, let $\mathrm{cost}(v) = 0$ and construct Γ_v using formula (1) (see Section III), and let $T'_{v,\chi(v)}$ be the rooted tree with one edge.
2b. For each interior vertex v with children $v_1, \cdots, v_{p-1}$ do:
2c. Solve the local **OPR**-score problem (i.e. compute $\mathbb{S}_v$ from the valid pairs of the children of v) by reducing it to an instance of the **GHS** problem, and find a corresponding optimal tree $T^*_{v,\gamma}$ as given by Lemma V.2.
2d. Construct $T'_{v,\gamma}$ from the tree $T^*_{v,\gamma}$ by replacing each of its leaves γ^i $(1 \leq i \leq p-1)$ with the tree T'_{v_i,γ^i}.
3. Output $\mathrm{cost}(\rho)$, and $T_{\rho,\gamma}$ for some $\gamma \in \Gamma^0_\rho$.

Fig. 4. An algorithm for solving the **OPR** problem.

function $\theta : U \to \mathbb{N}$ and a partition $\{U^0, \ldots, U^{p-1}\}$ of U as follows:

- $U^0 := \{\{\gamma\}\}$, $U^i := \{\Gamma^0_{v_i}, \ldots, \Gamma^{k-1}_{v_i}\}$, $1 \leq i \leq p-1$ and $U := \sqcup_{i=0}^{p-1} U^i$ (so that, in particular, U consists of subsets of Γ);
- $W := \Gamma$;
- E consists of those $\{u, w\}$, $u \in U$ and $w \in W$, with $w \in u$;
- $\theta : U \to \mathbb{N}$ is given by setting $\theta(u) = 0$ if $u \in U^0$ and $\theta(u) = t$ if $u = \Gamma^t_{v_i}$ for some $1 \leq i \leq p-1$.

As a consequence of the following result, it immediately follows by the discussion just preceding Proposition 3.2 that the local **OPR**-score problem can be solved as an instance of the **GHS** problem.

Proposition 5.1. *In the terminology just defined above, we have*

$$\mathrm{cost}(v, \gamma) = \omega^*(G_{v,\gamma}) + \sum_{i=1}^{p-1} \mathrm{cost}(v_i).$$

Proof. First, we show that there is a bijection ψ between elements of A_v and the following family of subsets of F:

$$\mathcal{F} := \big\{F \subseteq E : |F| = p \text{ and } u^i_F \neq \emptyset \text{ for all } 0 \leq i \leq p-1\big\}.$$

Let $u = \{\gamma\}$ be the unique vertex in U^0. Define $\psi : A_v \to \mathcal{F}$ by setting, for $\underline{a} = (a_1, \ldots, a_{p-1})$,

$$\psi(\underline{a}) := \{\{u, \gamma\}\} \sqcup \Big\{\Big\{\Gamma^t_{v_i}, a_i\Big\} : 1 \leq i \leq p-1 \text{ and } a_i \in \Gamma^t_{v_i} \text{ with } 0 \leq t \leq k-1\Big\}.$$

Note that, by definition, $\psi(A_v) \subseteq \mathcal{F}$. To see that ψ is indeed a bijection, it is straight-forward to check that it has the inverse ψ^{-1}, which is defined, for any $F \in \mathcal{F}$, by setting

$$\psi^{-1}(F) := \big(b_1, \ldots, b_{p-1}\big),$$

where b_i is the unique edge in F that is incident to some vertex contained in U^i, $1 \leq i \leq p-1$.

Moreover, note that

$$l(\xi_{\underline{a},\gamma}) + \kappa(\underline{a}) = \omega(\psi(\underline{a}))$$

and

$$l\big(\xi_{\psi^{-1}(F),\gamma}\big) + \kappa\big(\psi^{-1}(F)\big) = \omega(F).$$

Therefore, since ψ is a bijection, the proposition now follows by Proposition 3.2 and the definition of $\omega^*(G_{v,\gamma})$.□

Since the graph $G_{v,\gamma}$ can be constructed in $O(pkr^k)$ time (as it is specified by its edge set whose cardinality is bounded above by pr^k and the cardinality of U^i is bounded above by k for all $0 \leq i \leq p-1$), and since in solving the **GHS** problem for $G_{v,\gamma}$ we can obtain a resolved tree with leaves $\{\gamma^i\}_{i=1}^p$ in the proof of Lemma 4.1, where $\{U^i_F, \gamma^i\}$ $(1 \leq i \leq p)$ belongs to a minimal hitting set F, the following result immediately follows by Lemma 4.1.

Lemma 5.2. *By reducing to an instance of the* **GHS** *problem as described above, for a vertex with $p-1$ children, the local* **OPR***-score problem can be solved in*

$$pkr^k + \sum_{i=1}^{\min(p, r^k)} \binom{r^k}{i} (kri(2i-5)!! + pki)$$

time. Moreover, we can find an optimal tree $T^*_{v,\gamma}$ solving the corresponding local **OPR***-tree problem that is resolved, has root γ and leaves $\gamma^i \in \Gamma^0_{v_i} \sqcup \cdots \sqcup \Gamma^{k-1}_{v_i}$ for $1 \leq i \leq p-1$, and constructing such tree does not require additional time.*

6 An Algorithm for Solving the OPR Problem

We now present our algorithm for solving the **OPR** problem (Algorithm I depicted in Fig. 4). This algorithm can be regarded as a generalization of the well-known Fitch-Hartigan algorithm [7], [10] in that it works by arbitrarily introducing a root vertex to the (unresolved) phylogenetic tree of interest and then works up through the vertices of the resulting tree from the leaves to the root. At each stage, it uses the data previously computed for the children of a given vertex to solve the local **OPR**-score and local **OPR**-tree problems. Once it reaches the root, this gives the required solutions for the **OPR** problem.

The correctness of Algorithm I follows from the fact that if T^ρ is a rooted version of T, then $l(\chi, T) = l(\chi, T^\rho)$ and

$rl_T(\chi) = rl_{T^\rho}(\chi)$, and the fact that for the tree $T^*_{v,\gamma}$ constructed in Step 2d, we have

$$\mathrm{ch}\left(T^*_{v,\gamma}, \chi\right) = \mathrm{cost}(T_v, \gamma) - \sum_{i=1}^{p-1} \mathrm{cost}\left(T_{v_i}, \gamma^i\right),$$

from which it follows that $\mathrm{ch}(T'_{v,\gamma}, \chi) = \mathrm{cost}(v, \gamma)$.

The proof of Theorem 1.1. Recall that $m = r^k$ and $s := \min(d, m)$. Also, for an integer $i \geq 3$, set $h(i) := kri(2i-5)!!$.

Now, Steps 1 and 3 in Algorithm I can be computed in constant time, and Step 2a can be performed in nkm time. Therefore, since Step 2c is performed for each interior vertex of T, and $T'_{v,\gamma}$ can be constructed from $T^*_{v,\gamma}$ and T'_{v_i,γ^i} $(1 \leq i \leq p-1)$ in Step 2d in $p-1$ time, it follows by Lemma 5.2 that Algorithm I takes

$$n\left(dkm + km + dm + \sum_{i=1}^{\min(d,m)} \binom{m}{i}(h(i) + dki)\right)$$

time.

Now, since

$$h(i) + dki = kri(2i-5)!! + dki \leq 2dkr(2i-3)!!,$$

using the fact that $\binom{m}{i}(2i-3)!!$ is a monotone increasing function of $i \in \{1, \ldots, m-1\}$, $m \geq 3$, and so is maximized for $i \in \{1, \ldots, s\}$ by $\binom{m}{s}(2s-3)!!$, we conclude that

$$\begin{aligned}
&n\left(dkm + km + dm + \sum_{i=1}^{\min(d,m)} \binom{m}{i}(h(i) + dki)\right) \\
&\quad \leq n\left(3dkm + 2dkr \sum_{i=1}^{\min(d,m)} \binom{m}{i}(2i-3)!!\right) \\
&\quad \leq 5ndkr \sum_{i=1}^{\min(d,m)} \binom{m}{i}(2i-3)!! \\
&\quad \leq 5ndkrs\binom{m}{s}(2s-3)!!
\end{aligned}$$

from which Theorem 1.1 immediately follows. □

7 Discussion

In this paper, we have presented an algorithm for solving the **OPR** problem for a (k, r)-character on an n-set. However, for some special values of k and r, it is possible to obtain a more efficient and/or explicit algorithm. For example, in [11], an approach is proposed for solving the **OPR** problem for (1,2)-characters, and in [1] and [4], the **MP** problem for $(2, r)$-characters is investigated.

We also note that for $(1, r)$-characters the local **OPR**-score problem can be reduced to an instance of the **Hitting Set** problem in polynomial time (since $l(\chi) = |\chi(X)| - 1$, and for $F \subseteq E$, we clearly have $\omega(F) = |W_F| - 1$; see Section 4). In particular, it follows that the **OPR**-score problem for $(1, r)$-characters can be solved in $O(ndr2^t)$ time, where $t = \min\{d, r\}$. In addition, for (2,2)-characters, the local **OPR**-score problem can be solved in a more direct manner without reducing to the **GHS** problem. For example, it can be shown that the local **OPR**-score problem for a vertex v with $p-1$ children can be solved in $O(p)$ time, from which it follows that the **OPR**-score problem can be solved in $O(nd)$ time.

In this paper, we have mainly focused on the exact algorithms solving the **OPR** problem, but the approaches designed here might also be helpful to study PTAS algorithms, or fixed parameter algorithms for other variants of the **MP** problem (cf. [2]).

There are several additional interesting directions for possible future work. One is to consider trees with small degree, say $d = 4$ (since the case that $d = 3$ is trivial). Either a positive result (an efficient algorithm) or a negative result (a complexity conclusion) would provide further insight into the **OPR** problem. Another direction is to see whether the approach taken in this paper could lead to better randomized or approximation algorithms. Finally, it would also be interesting to consider the optimal refining problem with respect to other optimization criteria, such as maximum likelihood.

Acknowledgments

Vincent Moulton and Mike Steel would like to thank the Engineering and Physical Sciences Research Council (EPSRC, Grant EP/D068800/1) for its support. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, for hosting them in the context of its Phylogenetics Programme where part of the research presented in this paper was carried out. The authors would also like to thank the editor and three anonymous referees for their valuable suggestions.

References

[1] E. Althaus and R. Naujoks, "Computing Steiner Minimum Trees in Hamming Metrics," *Proc. 17th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA '06)*, pp. 172-181, 2006.

[2] G. Blelloch, K. Dhamdhere, E. Halperin, R. Ravi, R. Schwartz, and S. Sridhar, "Fixed Parameter Tractability of Binary Near-Perfect Phylogenetic Tree Reconstruction," M. Bugliesi, B. Preneel, V. Sassone, I. Wegener, eds., *Proc. 33rd Int'l Colloquium Automata, Languages and Programming (ICALP '06)*, Part I, pp. 667-678, 2006.

[3] M. Bonet, M. Steel, T. Warnow, and S. Yooseph, "Better Methods for Solving Parsimony and Compatibility," *J. Computational Biology*, vol. 5, no. 3, pp. 391-408, 1998.

[4] T. Bruen and D. Bryant, "A Subdivision Approach to Maximum Parsimony," *Annals of Combinatorics*, vol. 12, pp. 45-51, 2008.

[5] W.H.E Day, "Computationally Difficult Parsimony Problems in Phylogenetics Systematics," *J. Theoretical Biology*, vol. 103, pp. 429-438, 1983.

[6] J Felsenstein, *Inferring Phylogenies*. Sinauer Assoc., Inc., 2004.

[7] W. Fitch, "Toward Defining the Course of Evolution: Minimum Change for a Specified Tree Topology," *Systematic Zoology*, vol. 20, pp. 406-416, 1971.

[8] L.R. Foulds and R.L. Graham, "The Steiner Problem in Phylogeny Is NP-Complete," *Advances in Applied Math.*, vol. 3, pp. 43-49, 1982.

[9] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979.

[10] J.A. Hartigan, "Minimum Mutation Fits to a Given Tree," *Biometrics*, vol. 29, pp. 53-65, 1973.

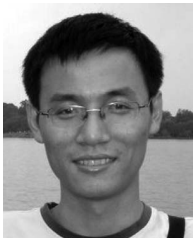
[11] D. Huson, M. Steel, and J. Whitfield, "Reducing Distortion in Phylogenetic Networks," *Proc. Sixth Workshop Algorithms in Bioinformatics (WABI '06)*, pp. 150-161, 2006.

[12] R. Niedermeier, *Invitation to Fixed-Parameter Algorithms*. Oxford Univ. Press, 2006.

[13] O. Nomura, Z.H. Lin, Muladno, Y. Wada, and H. Yasue, "A SINE Species from Hippopotamus and Its Distribution among Animal Species," *Mammalian Genome*, vol. 9, no. 7, pp. 550-555, 1998.

[14] A.M. Shedlock and N. Okada, "SINE Insertions: Powerful Tools for Molecular Systematics," *Bioessays*, vol. 22, pp. 148-160, 2000.

[15] C. Semple and M. Steel, *Phylogenetics.* Oxford Univ. Press, 2003.
[16] M. Steel and D. Penny, "Maximum Parsimony and the Phylogenetic Information in Multi-State Characters," *Parsimony, Phylogeny and Genomics,* V. Albert, ed., pp. 163-178, Oxford Univ. Press, 2005.



Taoyang Wu received the BE degree from Harbin Institute of Technology and the MSc degree from Peking University. He is a PhD candidate of computer sciences and mathematics in the Department of Computer Science and School of Mathematical Sciences, Queen Mary, University of London, and a visiting scientist at the CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai. His research interests include combinatorics, computational biology, and theoretical computer science. He is a student member of the London Mathematical Soceity.

Vincent Moulton received the PhD degree in mathematics from Duke University in 1994. He did postdoctoral research at the University of Bielefeld, University of Canterbury, and Massey University. He was a senior lecturer in discrete mathematics at Mid Sweden University from 1999 to 2002 and a professor in bioinformatics at Uppsala University from 2002 to 2004. In 2004, he moved to the University of East Anglia, where he is a professor in computational biology. His research interests are in phylogenetics, computational biology of RNA, metabolic modeling, algorithms in bioinformatics, and the study of discrete structures such as graphs and finite metric spaces.

Mike Steel received the MSc degree in mathematics from Canterbury University and the PhD degree in mathematical biology from Massey University in 1989. He had a brief career as a journalist. After various postdoctoral positions, starting with Andreas Dress in Germany, he was appointed to a tenured position at the University of Canterbury in 1994. He is currently a professor and the director of the Biomathematics Research Centre, University of Canterbury and is a principal investigator in the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University.