



# Quantifying the accuracy of ancestral state prediction in a phylogenetic tree under maximum parsimony

Lina Herbst<sup>1</sup> · Heyang Li<sup>2</sup> · Mike Steel<sup>3</sup>

Received: 26 April 2018 / Revised: 21 January 2019 / Published online: 13 February 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

In phylogenetic studies, biologists often wish to estimate the ancestral discrete character state at an interior vertex  $v$  of an evolutionary tree  $T$  from the states that are observed at the leaves of the tree. A simple and fast estimation method—maximum parsimony—takes the ancestral state at  $v$  to be any state that minimises the number of state changes in  $T$  required to explain its evolution on  $T$ . In this paper, we investigate the reconstruction accuracy of this estimation method further, under a simple symmetric model of state change, and obtain a number of new results, both for 2-state characters, and  $r$ -state characters ( $r > 2$ ). Our results rely on establishing new identities and inequalities, based on a coupling argument that involves a simpler ‘coin toss’ approach to ancestral state reconstruction.

**Keywords** Phylogenetic tree · Markov process · Maximum parsimony · Coupling

**Mathematics Subject Classification** 05C05 · 92D15

## 1 Introduction

Phylogenetic trees play a central role in evolutionary biology and in other related areas of classification (e.g. language evolution, stemmatology, ecology, epidemiology and medicine). Typically, these trees represent a set of sampled ‘taxa’ (e.g. species, genera, populations, individuals) as the leaves of the tree, with the vertices and edges of the tree providing a historical description of how these taxa evolved from a common ancestor (Felsenstein 2004). Biologists often use discrete characteristics of the species at the leaves of a tree to try to infer (or predict) an ancestral state deep within the tree. For

---

✉ Mike Steel  
mike.steel@canterbury.ac.nz

<sup>1</sup> Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany

<sup>2</sup> School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

<sup>3</sup> Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

example, in epidemiology, HIV sequences from sampled individuals have been used to estimate an ancestral form of the virus (e.g. for vaccine development) (Gaschen 2002); in another study, ancestral state reconstruction played a key role in investigating the evolution of complex traits involved in animal vision, which varies across different species (Plachetzki et al. 2010).

Assuming that the characteristic in question has also evolved with the species, various methods have been devised to infer the ancestral state of that characteristic inside the tree and, in particular, at the last common ancestor of the species under study (i.e. the root of the tree). A method that can predict this root state allows any other ancestral vertex in the tree to also be studied, since one can re-root that tree on that vertex. Thus, in this paper, we will assume that the root vertex is the one we wish to estimate an ancestral state for.

A variety of methods have been proposed for ancestral state reconstruction from the states at the leaves of a tree. An early method that is still used for certain types of data (e.g. morphological characters on a known tree) is maximum parsimony. This method minimizes the number of state changes required to fit the discrete data observed at the leaves to the rest of the vertices of the tree (including the root vertex). Other methods, now widely used, have also been developed within the maximum likelihood and Bayesian framework including Yang et al. (1995) and Koshi and Goldstein (1996) and leading to more efficient and refined techniques (see for example Pupko et al. 2000; Huelsenbeck and Bollback 2001).

Our reason for focussing on maximum parsimony in this paper is twofold: firstly, it is a classical method that has been used extensively over many decades, is fast, and is reliant only on the tree and the states at the leaves, and not the transition rates and branch lengths for the particular character under study (which may not be closely connected with parameters estimated from DNA sequence data). Recent examples of studies that used parsimony (and other methods) to estimate ancestral states on a tree include Göpel and Wirkner (2018), Sauquet et al. (2017), Hsiang et al. (2015) and Duchemin et al. (2017). The second reason is that we are able to derive new and exact mathematical results in this paper for ancestral state estimation using maximum parsimony for which comparable results for likelihood or Bayesian methods have yet to be formally established.

We stress, however, that our focus on maximum parsimony should not be interpreted as trying to suggest that it has some properties superior to other existing methods; indeed, it is well known that maximum parsimony is a problematic method for a different phylogenetic task—inferring evolutionary trees from a sequence of characters (such as DNA sequence sites) that follow some common stochastic model—since the method in that setting is known to be statistically inconsistent (see e.g. Felsenstein 2004).

The structure of this paper is as follows. First, we present some definitions concerning phylogenetic trees and a simple  $r$ -state Markovian model of character change on the tree, together with methods for predicting ancestral states, particularly maximum parsimony (MP). In Sect. 2, we concentrate on the 2-state model. We describe an exact relationship between the reconstruction accuracies of MP on any binary tree  $T$ , and the accuracy on two trees derived from  $T$  by deleting one and two leaves respectively. We show how this allows inequalities to be established easily by induction.

Next, in Sect. 3, we describe a simpler ancestral prediction method that is easier to analyse mathematically and yet is close enough to MP that it allows for inequality results for MP to be established. In particular, in Sect. 4, we show that the reconstruction accuracy for this simple method is always a lower bound to MP under the 2-state model, thereby improving on existing known lower bounds. In Sect. 5, we investigate the reconstruction accuracy for MP further in the more delicate setting when the number of states is greater than 2 and obtain some new inequality results. In Sect. 6, we present a novel combinatorial result that provides a sufficient condition for MP to infer the state at the root of a tree correctly, assuming only that the state changes in the tree are sufficiently well-spaced. In the final section, we present a conjecture for future work.

## 1.1 Definitions

In this paper, we consider rooted binary phylogenetic trees, which are trees in which every edge is directed away from a root vertex  $\rho$  that has in-degree 0 and out-degree 1 or 2, and in which every non-root vertex has in-degree 1 and out-degree 0 or 2. The vertices of out-degree 0 are the *leaves* of the tree. In the case where  $\rho$  has out-degree 2, we use  $T$  to denote the tree, but if  $\rho$  has out-degree 1, we will indicate this by writing  $\dot{T}$  instead of  $T$  and we will refer to the edge incident with this root as the *stem edge*. We will let  $X$  denote the set of leaves of  $T$ , and  $n = |X|$  the number of leaves of  $T$ .

Suppose that the root vertex  $\rho$  has an associated state  $F(\rho)$  that lies in some finite state space  $\mathcal{A}$  of size  $r \geq 2$ , and that the root state evolves along the edges of the tree to the leaves according to a Markov process in which each edge  $e$  has an associated probability  $p_e$  of a change of state (called a *substitution*) between the endpoints of  $e$ . We refer to  $p_e$  as the *substitution probability* for edge  $e$ . In this paper, we will assume that the underlying Markov process is the simple symmetric model on  $r$  states, often referred to as the Neyman  $r$ -state model, denoted  $N_r$ , which includes the (earlier) Jukes–Cantor model (Jukes and Cantor 1969) in the special case when  $r = 4$ . In this model, when a state change occurs on an edge  $e = (u, v)$ , each one of the  $r - 1$  states that are different from the state at  $u$  is assigned uniformly at random to the vertex  $v$ . In this way, each vertex  $v$  of the tree is assigned a random state, which we will denote as  $F(v)$ . We will denote the values of  $F$  on the leaves of  $T$  by the function  $f: X \rightarrow \mathcal{A}$ . This function  $f = F|_X$  (the restriction of  $F$  to the leaves of  $T$ ) is called a *character* in phylogenetics. Each such character has a well-defined probability under this stochastic model, and these probabilities sum to 1 over all the  $r^n$  possible choices for  $f$ .

Given  $f$ , consider the set  $\text{FS}(f, T)$  of possible states that can be assigned to the root vertex of  $T$  so as to minimise the total number of state changes required on the edges of  $T$  to generate  $f$  at the leaves. The set  $\text{FS}(f, T)$  can be found in linear time (in  $n$  and in  $r$ ) by the first pass of the ‘Fitch algorithm’ (Fitch 1971; Hartigan 1973). More precisely, to find  $\text{FS}(f, T)$ , we assign a subset  $\text{FS}(v)$  of  $\mathcal{A}$  to each vertex  $v$  of  $T$  in recursive fashion, starting from the leaves of  $T$  and working towards the root vertex  $\rho$  (we call  $\text{FS}(v)$  the *Fitch set* assigned to  $v$ ). First, each leaf  $x$  is assigned the singleton set  $\{f(x)\}$  as its Fitch set. Then for each vertex  $v$  for which its two children  $v_1$  and  $v_2$  have been assigned Fitch sets  $\text{FS}(v_1)$  and  $\text{FS}(v_2)$ , respectively, the Fitch set  $\text{FS}(v)$  is determined as follows:

$$FS(v) = \begin{cases} FS(v_1) \cap FS(v_2), & \text{if } FS(v_1) \cap FS(v_2) \neq \emptyset; \\ FS(v_1) \cup FS(v_2), & \text{if } FS(v_1) \cap FS(v_2) = \emptyset. \end{cases}$$

In this way, each vertex is eventually assigned a non-empty subset of  $\mathcal{A}$  as its Fitch set, and  $FS(f, T)$  is the Fitch set  $FS(\rho)$  that is assigned to the root vertex  $\rho$ .

When  $FS(f, T)$  consists of a single state, then the method of *maximum parsimony* uses this state as the estimate of the unknown ancestral state  $\alpha$  at the root. When  $FS(f, T)$  has more than one state, we will select one of the states in this set uniformly at random as an estimate of the root state (Fischer and Thatte 2009; Li et al. 2008; Zhang et al. 2010). We will let  $MP(f, T)$  be the state selected uniformly at random from  $FS(f, T)$ .

In this paper, we investigate the probability that this procedure correctly identifies the true root state  $\alpha$  [note that by the symmetry in the model there is nothing special about the choice of the root state  $F(\rho)$ ]. We call this probability the *reconstruction accuracy* for maximum parsimony, denoted  $RA_{MP}(T)$ . It is defined formally by:

$$RA_{MP}(T) := \mathbb{P}(MP(f, T) = F(\rho)).$$

Equivalently,  $RA_{MP}(T) = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \mathbb{P}(MP(f, T) = \alpha | F(\rho) = \alpha)$ . However, it is more useful in this paper to express  $RA_{MP}(T)$  as a weighted sum of probabilities involving Fitch sets as follows:

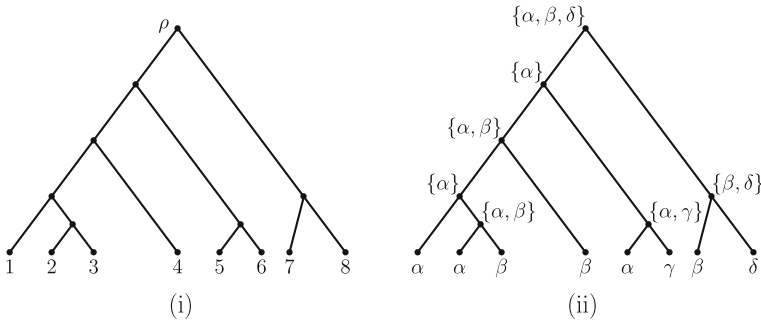
$$RA_{MP}(T) = \sum_{\substack{\mathcal{R}: \mathcal{R} \subseteq \mathcal{A} \\ \text{and } \alpha \in \mathcal{R}}} \frac{1}{|\mathcal{R}|} \cdot \mathbb{P}(FS(f, T) = \mathcal{R} | F(\rho) = \alpha). \tag{1}$$

This expression holds because MP selects one of the states in the Fitch set uniformly at random as the estimate of the root state, and so the probability that the actual root state (i.e.  $\alpha$ ) is correctly selected by MP is 0 if  $\alpha$  is not in the Fitch set for the root, otherwise it is 1 divided by the size of the Fitch set of the root (Fig. 1).

Because it is normally assumed that state changes occur according to an underlying continuous-time Markov process, one has:

$$p_e \leq (r - 1)/r.$$

We usually will assume that this inequality is strict, since  $p_e = (r - 1)/r$  would correspond to an infinite rate of change (or an infinite temporal length) on the edge  $e$  for a continuous-time Markov process. Given the substitution probability  $p_e$  for an edge  $e$ , we can formally associate a ‘length’ for this edge as the quantity  $\ell_e = -\frac{r-1}{r} \ln(1 - \frac{r}{r-1} p_e)$ . This ‘length’ corresponds to the expected number of state changes under a continuous-time Markov-process realization of the substitution process (see e.g. Felsenstein 2004; Steel 2016). Notice that we can write  $p_e = \frac{r-1}{r} \left(1 - \exp\left(-\frac{r}{r-1} \ell_e\right)\right)$ . If we let  $p(v)$  be the probability that vertex  $v$  is in a different state from the root  $\rho$  then  $p(v) = \frac{r-1}{r} \left(1 - \exp\left(-\frac{r}{r-1} L\right)\right)$ , where  $L$  is the sum of the  $\ell$ -lengths of the edges on the path from  $\rho$  to  $v$ .



**Fig. 1** **i** A rooted binary tree on leaf set  $X = \{1, 2, 3, \dots, 8\}$ . If we consider the character  $f: X \rightarrow \mathcal{A} = \{\alpha, \beta, \gamma, \delta\}$  defined by  $f(1) = f(2) = f(5) = \alpha, f(3) = f(4) = f(7) = \beta, f(6) = \gamma, f(8) = \delta$ , then the associated Fitch sets at the interior vertices are shown in **(ii)**. Notice that the root Fitch set  $FS(f, T)$  consists of three equally most-parsimonious root states, namely  $\{\alpha, \beta, \delta\}$  and so  $MP(f, T)$  would be one of these states chosen with equal probability ( $\frac{1}{3}$ ). An interesting feature of this example is that if the state of the leaf labelled 1 is changed from  $\alpha$  to  $\delta$ , then although  $\delta$  was initially one of the most parsimonious states for the root, it ceases to be so (instead,  $\beta$  becomes the unique most parsimonious root state)

A special condition that is sometimes further imposed on these edge lengths is that the edge lengths satisfy an *ultrametric condition* (called a ‘molecular clock’ in biology), which states that the sum of the lengths of the edges from the root to each leaf is the same. Under that assumption, the probability  $p(x)$  that leaf  $x$  is in a different state from the root takes the same value for all values of  $x$ . In this paper, our main results do not require this ultrametric assumption; however, we also point out how these results lead to particular conclusions in the ultrametric case.

Note that  $RA_{MP}(T)$  depends on  $T$ , the assignment of state-change probabilities (the  $p_e$  values) for the edges of  $T$ , and  $r$  (the size of the state space  $\mathcal{A}$ ). The aim of this paper is to provide new relationships (equations and inequalities) for reconstruction accuracy, extending earlier work by others (Herbst and Fischer 2018; Li et al. 2008; Zhang et al. 2010; Fischer and Thatte 2009). Note that two other methods for estimating the ancestral root state are *majority rule* (MR), which estimates the root state by the most frequently occurring state at the leaves (ties are broken uniformly at random), and *maximum likelihood estimation* (MLE), which estimates the root state by the state(s) that maximise the probability of generating the given character observed at the leaves. MR does not even require knowledge of the tree for estimating the root state, whereas MLE requires knowing not only the tree but also the edges lengths. Comparisons of these three methods were studied by Gascuel and Steel (2010, 2014).

Note that if the edge lengths particular to a single character under study in MLE are not known, and are therefore treated as ‘nuisance parameters’ to be estimated (in addition to the root state) then the resulting MLE estimate for the root state for that character can be shown to be precisely the MP estimate under the  $N_r$  model (Tuffley and Steel 1997, Theorem 6). However, if a collection of characters is used to estimate common branch lengths for the tree (under the assumption that the characters have all evolved under the same branch lengths, as with certain models of DNA sequence evolution Felsenstein 2004) then the MLE estimate of the ancestral root state is no longer directly given by MP.

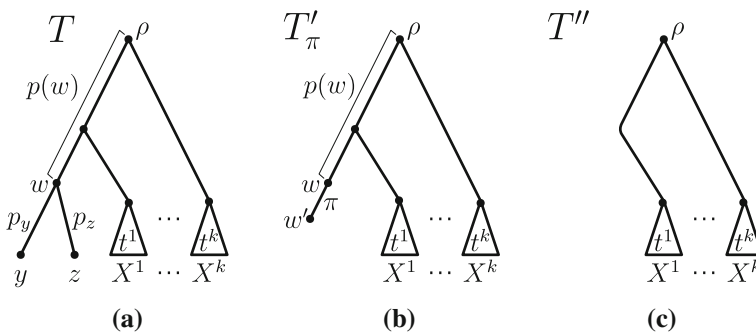
We end this section by collating some notation used throughout this paper.

- $T$  (resp.  $\dot{T}$ )—a rooted binary tree, with a root of out-degree 2 (resp. out-degree 1),
- $p_e$  (resp.  $p_\rho$ )—the substitution probability on edge  $e$ , (resp. the stem edge of  $\dot{T}$ ) under the  $N_r$  model,
- $p(x)$  [resp.  $p(w)$ ]—the probability that leaf  $x$  (resp. vertex  $w$ ) is in a different state from the root under the  $N_r$  model,
- $p_{\max}$ —the maximal value of  $p(x)$  over all leaves,
- $RA_{MP}(T)$ —the root-state reconstruction accuracy of maximum parsimony on  $T$  (with its  $p_e$  values) for a character generated under the  $N_r$  model.

## 2 A fundamental identity for reconstruction accuracy in the case where $r = 2$

For Theorem 1 (below) we consider a rooted binary phylogenetic tree  $T$  with a leaf set  $X$  of size at least 3, together with two associated trees  $T'_\pi$  and  $T''$  as indicated in Fig. 2, which are determined by selecting a pair of leaves  $y, z$  that are adjacent to a common vertex of  $T$  [such a pair of leaves, called a ‘cherry’, always exists in any binary tree with 3 or more leaves (Steel 2016)]. The rooted binary phylogenetic tree  $T'_\pi$  is obtained from  $T$  by deleting the leaves  $y$  and  $z$ ; in addition, we lengthen the edge leading to  $w$  slightly by putting an extra edge from  $w$  to a new leaf  $w'$  with substitution probability  $\pi$ . In order to keep  $T'_\pi$  binary, the vertex  $w$  is suppressed. An additional tree  $T''$  is obtained from  $T'_\pi$  by deleting the edge leading to  $w$  and edge  $(w, w')$ . Again, we suppress the resulting vertex of degree 2 in order to keep the tree binary.

We now state the main result of this section. Given  $T, T'_\pi$  and  $T''$  as described we have the following fundamental equation for MP as ancestral state reconstruction



**Fig. 2** **a** A rooted binary phylogenetic tree  $T$  with leaf set  $X$  where  $p(w)$  is the probability that  $w$  is in a different state from the root  $\rho$ , and  $p_y$  and  $p_z$  are the probabilities that leaves  $y$  and  $z$  are in a different state from  $w$ . The pendant subtrees adjacent to the path from  $w$  up to  $\rho$  are denoted  $t^1, \dots, t^k$  with leaf sets  $X^1, \dots, X^k$ , respectively, **b** a rooted binary phylogenetic tree  $T'_\pi$  derived from  $T$  by deleting leaves  $y$  and  $z$  and attaching a new leaf  $w'$  to  $w$  (which is then suppressed). The value  $\pi$  is the probability of a change of state from  $w$  to the new leaf  $w'$ , **c** the rooted binary tree  $T''$  obtained from  $T$  by deleting the leaves  $y$  and  $z$ , their incident edges and the other edge incident with  $w$ , then suppressing the resulting vertex of degree 2

method under the  $N_2$  model. This result is particular to the  $N_2$  model (i.e. it does not hold for  $N_r$  when  $r > 2$ ). The reason is that when  $r = 2$  and two leaves in a cherry have different states then we can essentially prune these two leaves from the tree without affecting the Fitch set recursion as it works towards the root. However for  $r > 2$  this no longer holds.

**Theorem 1** *Let  $T$  be a rooted binary phylogenetic tree with a leaf set  $X$  of size at least 3. For the reconstruction accuracy of maximum parsimony under the  $N_2$  model we then have:*

$$RA_{MP}(T) = \theta \cdot RA_{MP}(T'_\pi) + (1 - \theta) \cdot RA_{MP}(T''),$$

where  $\theta$  is the probability that the leaves  $y$  and  $z$  are in the same state, and  $\pi = p_y p_z / \theta \leq \min\{p_y, p_z\}$  (where  $p_y$  and  $p_z$  are the substitution probabilities for edges  $(w, y)$  and  $(w, z)$ , respectively).

**Proof** Let  $T$  be a rooted binary phylogenetic tree with root  $\rho$ . By the symmetry in the model, we assume, without loss of generality, that the root is in state  $\alpha$ . Let  $\mathcal{F}$  denote the event that  $MP(f, T) = \alpha$  (recall that in the case of two equally-most-parsimonious states, one is selected uniformly at random). Let  $\mathcal{E}_1$  be the event that leaf  $y$  and leaf  $z$  are in the same state (i.e.  $f(y) = f(z)$ ), and let  $\mathcal{E}_2$  be the complementary event (i.e.  $f(y) \neq f(z)$ ). Thus  $\theta = \mathbb{P}(\mathcal{E}_1)$  and  $1 - \theta = \mathbb{P}(\mathcal{E}_2)$ . By the law of total probability we have:

$$\begin{aligned} RA_{MP}(T) &= \mathbb{P}(\mathcal{F}) = \mathbb{P}(\mathcal{F}|\mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{F}|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2) \\ &= \mathbb{P}(\mathcal{F}|\mathcal{E}_1)\theta + \mathbb{P}(\mathcal{F}|\mathcal{E}_2)(1 - \theta). \end{aligned}$$

We use this to establish Theorem 1 by establishing the following two claims:

- Claim (i):  $RA_{MP}(T'_\pi)$  equals  $\mathbb{P}(\mathcal{F}|\mathcal{E}_1)$ ;
- Claim (ii):  $RA_{MP}(T'')$  equals  $\mathbb{P}(\mathcal{F}|\mathcal{E}_2)$ .

To establish Claim (i), we show that by an appropriate choice of  $\pi$ , the probability that the leaves  $y$  and  $z$  are in state  $\alpha$ , conditional on the event  $\mathcal{E}_1$ , is exactly equal to the probability that  $w'$  is in state  $\alpha$ ; that is:

$$\mathbb{P}(f(y) = f(z) = \alpha | \mathcal{E}_1) = \mathbb{P}(F(w') = \alpha). \tag{2}$$

A similar equality will then hold for  $\beta$  (i.e.  $\mathbb{P}(f(y) = f(z) = \beta | \mathcal{E}_1) = \mathbb{P}(F(w') = \beta)$ , since both probabilities sum up to 1). These two identities then ensure that  $RA_{MP}(T'_\pi)$  equals  $\mathbb{P}(\mathcal{F}|\mathcal{E}_1)$ , which is Claim (i). Thus for Claim (i), it suffices to establish Eq. (2) for a suitable choice of  $\pi$ .

Recall that  $1 - p(w)$  is the probability that  $w$  is in state  $\alpha$ , since the root is assumed to be in state  $\alpha$ . Then, the probability that  $y$  and  $z$  are in state  $\alpha$  is

$$\mathbb{P}(f(y) = f(z) = \alpha) = (1 - p(w))(1 - p_y)(1 - p_z) + p(w)p_y p_z,$$

where  $p_y$  and  $p_z$  are the probabilities of change on edge  $(w, y)$  and on edge  $(w, z)$ . Similarly, the probability that  $y$  and  $z$  are both in state  $\beta$  is:

$$\mathbb{P}(f(y) = f(z) = \beta) = p(w)(1 - p_y)(1 - p_z) + (1 - p(w))p_y p_z.$$

Adding these together, the probability of  $\mathcal{E}_1$  is given by:

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1) &= (1 - p(w))(1 - p_y)(1 - p_z) + p(w)p_y p_z \\ &\quad + p(w)(1 - p_y)(1 - p_z) + (1 - p(w))p_y p_z, \end{aligned}$$

and so

$$\mathbb{P}(\mathcal{E}_1) = (1 - p_y)(1 - p_z) + p_y p_z, \tag{3}$$

which is independent of  $p(w)$ .

Now,

$$\mathbb{P}(f(y) = f(z) = \alpha | \mathcal{E}_1) = \frac{(1 - p(w))(1 - p_y)(1 - p_z) + p(w)p_y p_z}{\theta}, \tag{4}$$

[recall that  $\theta = \mathbb{P}(\mathcal{E}_1)$ ]. We can write (4) as

$$\mathbb{P}(f(y) = f(z) = \alpha | \mathcal{E}_1) = (1 - p(w))U + p(w)V, \tag{5}$$

where  $U = \frac{(1-p_y)(1-p_z)}{\theta}$  and  $V = \frac{p_y p_z}{\theta}$  (note that  $U + V = 1$ ). Now, with substitution probability  $\pi$  on edge  $(w, w')$ , the probability that  $w'$  is in state  $\alpha$  is

$$\mathbb{P}(F(w') = \alpha) = (1 - p(w))(1 - \pi) + p(w)\pi. \tag{6}$$

Comparing (6) with (5), we see that if we take  $\pi = V = \frac{p_y p_z}{\theta}$ , then Eq. (2) [and hence Claim (i)] holds.

Notice also that with this choice,  $\pi$  is less or equal to  $p_y$  and to  $p_z$ . For example,  $\pi \leq p_y$  is equivalent to:

$$\begin{aligned} \pi &= \frac{p_y p_z}{(1 - p_y)(1 - p_z) + p_y p_z} \leq p_y \\ \Leftrightarrow p_y p_z &\leq p_y((1 - p_y)(1 - p_z) + p_y p_z) \\ \Leftrightarrow p_z &\leq 1 - p_z, \end{aligned}$$

which holds, since  $p_z \leq \frac{1}{2}$ .

To show that  $RA_{MP}(T'') = \mathbb{P}(\mathcal{F} | \mathcal{E}_2)$ , first notice that the probability of event  $\mathcal{E}_2$  does not depend on the state at  $w$  [i.e.  $\mathbb{P}(\mathcal{E}_2 | F(w) = \alpha) = \mathbb{P}(\mathcal{E}_2 | F(w) = \beta)$ ], because:

$$\mathbb{P}(\mathcal{E}_2) = 1 - \mathbb{P}(\mathcal{E}_1) = (1 - p_y)p_z + p_y(1 - p_z),$$



by Eq. (3). Moreover, notice that when the leaves  $y$  and  $z$  take the states  $\alpha, \beta$  (or  $\beta, \alpha$ ) then the Fitch set for  $w$  is  $\{\alpha, \beta\}$ , so the state that is chosen as the ancestral state for  $\rho$  is completely determined by the subtree  $T''$ .

Together with the argument above, this gives  $RA_{MP}(T'') = \mathbb{P}(\mathcal{F}|\mathcal{E}_2)$ , as required.  $\square$

Theorem 1 leads to the following corollary, which extends earlier results by Fischer and Thatte (2009) and by Zhang et al. (2010) in which the ultrametric constraint on the edge lengths was imposed (here this assumption is lifted).

**Corollary 1** *Let  $T$  be a rooted binary phylogenetic tree with leaf set  $X$ . Under the  $N_2$  model:*

$$RA_{MP}(T) \geq 1 - p_{\max},$$

where  $p_{\max} = \max\{p(x) : x \in X\}$ , and  $p(x)$  is the probability that leaf  $x$  has a different state from the root.

**Proof** We use induction on the number of leaves  $n$ . For  $n = 1$ ,  $p_{\max} = p_x$  and thus the reconstruction accuracy is given by  $RA_{MP}(T) = 1 - p_{\max}$ . For  $n = 2$ , and a tree with leaves  $x, y$ :

$$\begin{aligned} RA_{MP}(T) &= (1 - p_x)(1 - p_y) + \frac{1}{2}(p_x(1 - p_y) + (1 - p_x)p_y) \\ &= 1 - p_x - p_y + p_x p_y + \frac{1}{2}(p_x - p_x p_y + p_y - p_x p_y) \\ &= 1 - \frac{1}{2}p_x - \frac{1}{2}p_y \geq 1 - p_{\max}. \end{aligned}$$

This completes the base case of the induction.

Now, assume that the claim holds for all rooted binary phylogenetic trees with less than  $n$  leaves, where  $n \geq 3$ , and consider a tree with  $n$  leaves represented as shown in Fig. 2. Let  $p' := \max\{p(x) : x \in (X \setminus \{y, z\}) \cup \{w\}\}$  and let  $p'' := \max\{p(x) : x \in X \setminus \{y, z\}\}$ . Thus,  $p', p'' \leq p_{\max}$  (the inequality for  $p''$  is clear; for  $p'$  we use  $\pi \leq p_y, p_z$  from the last part of Theorem 1). Now, from Theorem 1, we have:

$$RA_{MP}(T) = \theta \cdot RA_{MP}(T'_\pi) + (1 - \theta) \cdot RA_{MP}(T''),$$

where  $RA_{MP}(T'_\pi) \geq 1 - p'$  and  $RA_{MP}(T'') \geq 1 - p''$  by the induction hypothesis. Thus:

$$\begin{aligned} RA_{MP}(T) &\geq \theta(1 - p') + (1 - \theta)(1 - p'') \\ &\geq \theta(1 - p_{\max}) + (1 - \theta)(1 - p_{\max}) \\ &\quad \text{since } p \geq p' \text{ and } p \geq p'' \\ &= (\theta + 1 - \theta)(1 - p_{\max}) = 1 - p_{\max}, \end{aligned}$$

which completes the proof.  $\square$

### 3 A 'coin-toss' reconstruction method ( $\varphi$ )

We now consider a method for estimating the ancestral state that is similar to the Fitch algorithm for MP, but which uses coin tosses to simplify the process. Note that the motivation for introducing this method is for purely formal reasons: It allows us to prove results concerning MP by a coupling argument that relates MP to this simpler method that is easier to analyse mathematically. In particular, we are not advocating this method as one to use on real data.

The coin-toss method works as follows: given a rooted binary phylogenetic tree  $T$  and a character  $f$  at the leaves of  $T$ , the method proceeds from the leaves to the root, just like the Fitch algorithm described earlier. However, rather than assigning sets of states to each vertex, the coin toss method assigns a single state to each vertex.

More precisely, the coin-toss method starts (similarly to the Fitch algorithm) by assigning each leaf the state given by the character  $f$ . For a vertex  $v$  for which both direct descendants have been assigned states, if both these states are the same, then this state is also assigned to  $v$ . On the other hand, if the direct descendants have different states, then a fair coin is tossed to decide which of the two states to assign to  $v$ . This procedure is continued upwards along the tree until the root is assigned a state. We let  $\varphi$  denote this coin-toss method for ancestral state reconstruction, and denote the state selected by this method as  $\varphi(T, f)$ . Let  $RA_\varphi(T)$  denote its reconstruction accuracy [i.e. the probability that it predicts the true root state in the  $r$ -state model, which equals  $\mathbb{P}(\varphi(T, f) = F(\rho))$ ].

**Theorem 2** *Let  $T$  be a rooted binary phylogenetic tree with leaf set  $X$ . For  $x \in X$ , let  $d(x)$  denote the number of edges between the root  $\rho$  of  $T$  and leaf  $x$ . For the  $N_r$  model (for any  $r \geq 2$ ) we have:*

- (i)  $RA_\varphi(T) = 1 - \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} p(x)$ ;
- (ii)  $RA_\varphi(T) \geq 1 - p_{\max}$ , and,
- (iii) *in the ultrametric setting,  $RA_\varphi(T) = 1 - p_{\max}$ ,*

where  $p_{\max} = \max\{p(x) : x \in X\}$ , and  $p(x)$  is the probability that leaf  $x$  has a different state from the root.

**Proof** *Part (i)* Let  $T$  be a rooted binary phylogenetic tree with root  $\rho$  and leaf set  $X$ . Start at the root of  $T$  and apply the following 'reverse' process: toss a fair coin and, depending on the outcome, select one of the two children of  $\rho$  with equal probability. We keep going away from the root in this way until a leaf is reached. The root state is then estimated as the state at that leaf. Note that the reverse procedure (which proceeds from the root to the leaves) is stochastically identical in its estimated root state as the original coin-toss procedure  $\varphi$ . Therefore, we have:

$$RA_\varphi(T) = \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} (1 - p(x)) = 1 - \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} p(x), \quad (7)$$

as claimed. This establishes Part (i).

For Part (ii), with  $p_{\max} = \max\{p(x) : x \in X\}$ , we have:

$$\begin{aligned}
 RA_\varphi(T) &= 1 - \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} p(x) && \text{by (7)} \\
 &\geq 1 - p_{\max} \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} && \text{because } p_{\max} \geq p(x) \text{ for all } x, \\
 &= 1 - p_{\max}, && \text{because } \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} = 1,
 \end{aligned}$$

which gives Part (ii).

For Part (iii), we again observe that the reverse procedure for  $\varphi$  is stochastically identical in its estimated root state to the coin toss procedure  $\varphi$ . Thus the reconstruction accuracy of  $\varphi$  is just the probability that the leaf that is sampled has the same state as the root, and this is clearly just  $1 - p_{\max}$  in case of an ultrametric tree and gives us Part (iii) of the theorem.  $\square$

Note that the reverse description of  $\varphi$  should not be confused with the following even simpler estimation method: select a leaf  $x$  uniformly at random and estimate the ancestral root state by the state at  $x$ . This method is stochastically equivalent to  $\varphi$  only when  $T$  is a complete balanced binary tree with  $n = 2^k$  leaves. In general, however, different leaves will have different probabilities of being chosen by the ‘reverse’ description of  $\varphi$ , depending on the shape of the tree.

### 3.1 Trees with a stem edge

Shortly, we will need to consider the reconstruction accuracy of a rooted binary tree  $\dot{T}$  that has a root  $\rho$  of out-degree 1, and so we pause to describe how this is related to the reconstruction accuracy of the tree  $T$  adjacent to  $\rho$ . Consider the stem edge leading from this degree-1 root  $\rho$  to its child  $\rho'$  and let  $T$  be the tree obtained by removing this edge. We can extend the definition of  $RA_{MP}$  and  $RA_\varphi$  to  $\dot{T}$  by simply assigning the predicted root state for  $\rho'$  (for  $T$ ) to the root  $\rho$  of  $\dot{T}$ . The following lemma describes a linear identity between the reconstruction accuracy of  $\dot{T}$  and  $T$  for MP and the coin-toss method  $\varphi$ .

**Lemma 1** *Under the  $N_r$  model, suppose that the substitution probability for the stem edge  $(\rho, \rho')$  of  $\dot{T}$  is  $p_\rho$ . If  $M$  denotes either the method MP or  $\varphi$ , we then have:*

$$RA_M(\dot{T}) = \left(1 - \frac{r}{r-1} p_\rho\right) RA_M(T) + \frac{p_\rho}{r-1}.$$

**Proof** By considering the two possible cases (no substitution on the stem edge, and a substitution to one of the  $r - 1$  non-root states), the law of total probability gives:

$$RA_M(\dot{T}) = (1 - p_\rho) \cdot RA_M(T) + p_\rho \cdot \mathbb{P}(M(f, T) = \alpha | F(\rho') = \beta). \tag{8}$$

for any state  $\beta \neq \alpha$  (the choice does not matter because of the symmetry in the model).  
Now:

$$\sum_{\gamma \in \mathcal{A}} \mathbb{P}(M(f, T) = \gamma | F(\rho') = \beta) = 1. \tag{9}$$

The term on the left of this last equation can also be written as:

$$\mathbb{P}(M(f, T) = \beta | F(\rho') = \beta) + \sum_{\gamma \neq \beta} \mathbb{P}(M(f, T) = \gamma | F(\rho') = \beta).$$

Moreover, the  $r - 1$  probabilities in the summation term on the right of this last equation are all equal (again by the symmetries in the model). In particular, each of these  $r - 1$  probabilities is  $\mathbb{P}(M(f, T) = \alpha | F(\rho') = \beta)$ . Combining this observation with Eq. (9) gives:

$$1 = RA_M(T) + (r - 1)\mathbb{P}(M(f, T) = \alpha | F(\rho') = \beta),$$

which rearranges to become:

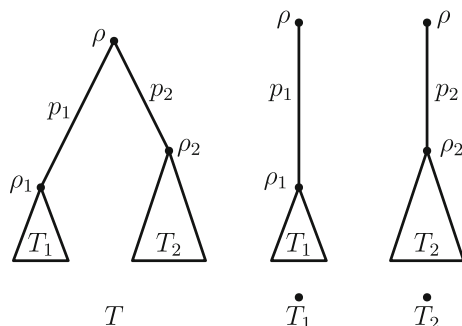
$$\mathbb{P}(M(f, T) = \alpha | F(\rho') = \beta) = \frac{1 - RA_M(T)}{r - 1}.$$

Finally, substituting this expression into Eq. (8) gives the expression in the lemma.  $\square$

### 3.2 Recursive equations for $RA_\varphi$

We now consider a rooted binary phylogenetic tree  $T$  with a root  $\rho$  of out-degree 2, along with its two maximal pendant subtrees  $T_1$  and  $T_2$  with roots  $\rho_1$  and  $\rho_2$ , respectively. Let  $\hat{T}_1$  be the tree obtained from  $T$  by deleting  $T_2$  and its incident edge  $e$  and associated  $p_e$  value (thus  $\hat{T}_1$  is  $T_1$  with the additional stem edge joining  $\rho_1$  to  $\rho$ ). Define  $\hat{T}_2$  similarly, as indicated in Fig. 3, and let  $p_i$  be the substitution probability for the edge  $(\rho, \rho_i)$ .

**Fig. 3** Left: The tree  $T$  with its two maximal subtrees  $T_1$  and  $T_2$ . Right: The trees  $\hat{T}_1$  and  $\hat{T}_2$  obtained by attaching a stem edge to  $T_1$  and  $T_2$ , with the same substitution probability  $p_i$  as in  $T$



**Theorem 3** *Let  $T$  be a rooted binary phylogenetic tree with leaf set  $X$ . Under the  $N_r$  model, the following identity for  $RA_\varphi(T)$  holds:*

$$RA_\varphi(T) = \frac{1}{2}(RA_\varphi(\dot{T}_1) + RA_\varphi(\dot{T}_2)).$$

**Proof** Let  $X_1$  and  $X_2$  be the leaf sets of the trees  $\dot{T}_1$  and  $\dot{T}_2$ . For the reconstruction accuracy of the coin-toss method under the  $N_r$  model, we then have:

$$\begin{aligned} RA_\varphi(T) &= 1 - \sum_{x \in X} \left(\frac{1}{2}\right)^{d(x)} p(x) && \text{by (7)} \\ &= 1 - \sum_{x \in X_1} \left(\frac{1}{2}\right)^{d(x)} p(x) - \sum_{x \in X_2} \left(\frac{1}{2}\right)^{d(x)} p(x) \\ &= \frac{1}{2} - \sum_{x \in X_1} \left(\frac{1}{2}\right)^{d(x)} p(x) + \frac{1}{2} - \sum_{x \in X_2} \left(\frac{1}{2}\right)^{d(x)} p(x) \\ &= \frac{1}{2} \left(1 - \sum_{x \in X_1} \left(\frac{1}{2}\right)^{d(x)-1} p(x)\right) + \frac{1}{2} \left(1 - \sum_{x \in X_2} \left(\frac{1}{2}\right)^{d(x)-1} p(x)\right) \\ &= \frac{1}{2} RA_\varphi(\dot{T}_1) + \frac{1}{2} RA_\varphi(\dot{T}_2) && \text{by (7)} \\ &= \frac{1}{2} (RA_\varphi(\dot{T}_1) + RA_\varphi(\dot{T}_2)), \end{aligned}$$

which completes the proof. □

### 4 The relationship between the two ancestral reconstruction methods

The aim of this section is to establish the following result.

**Theorem 4** *Let  $T$  be a rooted binary phylogenetic tree with leaf set  $X$ . Under the  $N_2$  model, the reconstruction accuracy of MP is at least equal to the reconstruction accuracy of the coin-toss method; that is:*

$$RA_{MP}(T) \geq RA_\varphi(T).$$

In order to establish this result, we first derive an analogue of the fundamental equation for MP (Theorem 1) for the coin-toss method, as given in Lemma 2. For this equation, we consider  $T'_{0.5}$  as depicted in Fig. 2, which is obtained from  $T$  as in Fig. 2 in the following way: again, we delete the leaves  $y$  and  $z$ . We then make the edge leading to  $w$  infinitely long by putting an extra edge from  $w$  to a new leaf  $w_{0.5}$  with the substitution probability  $\pi = \frac{1}{2}$  on this edge. Setting  $\pi = \frac{1}{2}$  simply means that both states are equally likely. Again, in order to keep the tree binary, vertex  $w$  is suppressed.

Under the  $N_2$  model, we have the following fundamental equation for the coin-toss method given  $T, T'_\pi$  and  $T'_{0.5}$  as described in Fig. 2 (note that  $T'_{0.5}$  is just  $T'_\pi$  with  $\pi = 0.5$ ).

**Lemma 2** *Let  $T$  be a rooted binary phylogenetic tree with leaf set  $X$ . Then, for the reconstruction accuracy of the coin-toss method under the  $N_2$  model, we have:*

$$RA_\varphi(T) = \theta \cdot RA_\varphi(T'_\pi) + (1 - \theta) \cdot RA_\varphi(T'_{0.5}),$$

where  $\theta$  and  $\pi$  are as defined as in Theorem 1.

**Proof** Let  $T$  be a rooted binary phylogenetic tree with root  $\rho$ , and assume without loss of generality that the root is in state  $\alpha$ . We define  $\mathcal{F}_\varphi$  to be the event that  $\alpha$  is the state chosen for  $\rho$  by the coin-toss method, and, as before, let  $\mathcal{E}_1$  be the probability that leaves  $y$  and  $z$  have the same state. By the law of total probability:

$$\begin{aligned} RA_\varphi(T) &= \mathbb{P}(\mathcal{F}_\varphi) = \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_2)\mathbb{P}(\mathcal{E}_2) \\ &= \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_1)\theta + \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_2)(1 - \theta). \end{aligned}$$

In order to prove Lemma 2, it remains to show  $RA_\varphi(T'_\pi) = \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_1)$  and  $RA_\varphi(T'_{0.5}) = \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_2)$  respectively. Now,  $RA_\varphi(T'_\pi) = \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_1)$  since, conditional on  $\mathcal{E}_1$ , the state chosen by  $\varphi$  at  $w$  in  $T$  has the same probability distribution as the state chosen by  $\varphi$  at  $w'$  in  $T'_\pi$ , and the remainder of application of  $\varphi$  to  $T$  and  $T'_\pi$  is identical.

We have  $RA_\varphi(T'_{0.5}) = \mathbb{P}(\mathcal{F}_\varphi|\mathcal{E}_2)$ , because by having the substitution probability  $\pi = \frac{1}{2}$  for the edge leading to  $w_{0.5}$  both states  $\alpha$  and  $\beta$  are equally probable. So the probability of choosing  $\alpha$  for  $w$  is  $\frac{1}{2}$ . Moreover, on  $T$  the probability of choosing  $\alpha$  for  $w$  from the states at the leaves  $y$  and  $z$  conditional on event  $\mathcal{E}_2$  (i.e.  $y$  and  $z$  are in different states), is  $\frac{1}{2}$  as well. □

**Proof of Theorem 4** The proof is by induction on the number of leaves. For  $n = 2$  and a tree with leaves  $x, y$ , we have:

$$\begin{aligned} RA_{MP}(T) &= (1 - p_x)(1 - p_y) + \frac{1}{2}(p_x(1 - p_y) + (1 - p_x)p_y) \\ &= \frac{1}{2}((1 - p_x) + (1 - p_y)). \end{aligned}$$

By Theorem 3, the reconstruction accuracy of the coin-toss method is exactly the average of the reconstruction accuracy of both subtrees. Therefore,  $RA_\varphi(T) = \frac{1}{2}((1 - p_x) + (1 - p_y))$ , which is equal to  $RA_{MP}(T)$ , and establishes the base case of the induction.

Now assume that the induction hypothesis holds for all rooted binary phylogenetic trees with fewer than  $n$  leaves, where  $n \geq 3$ . By Theorem 1, we have:

$$RA_{MP}(T) = \theta \cdot RA_{MP}(T'_\pi) + (1 - \theta) \cdot RA_{MP}(T''),$$

with  $T'_\pi$  and  $T''$  as in Fig. 2, and  $\theta$  as described above. Additionally, by Lemma 2 we have that

$$RA_\varphi(T) = \theta \cdot RA_\varphi(T'_\pi) + (1 - \theta) \cdot RA_\varphi(T'_{0.5}),$$

with  $T'_\pi$  and  $T'_{0.5}$  as in Fig. 2. By the induction hypothesis,  $RA_{MP}(T'_\pi) \geq RA_\varphi(T'_\pi)$  and  $RA_{MP}(T'') \geq RA_\varphi(T'')$  both hold, so in order to complete the proof, it remains to show that  $RA_\varphi(T'') \geq RA_\varphi(T'_{0.5})$ . The intuition behind this inequality is that when the leaf at the end of a pendant edge is completely random (i.e. no more likely to match the root state than not match it) then pruning this edge cannot reduce the reconstruction accuracy of  $\varphi$ . Note that  $T''$  has one leaf fewer than  $T'_{0.5}$ . In the following we consider both trees as shown in Fig. 2. As before, all vertices of degree 2 are suppressed to keep the tree binary. In order to calculate  $RA_\varphi(T'')$  and  $RA_\varphi(T'_{0.5})$ , consider the subtrees  $t^1, \dots, t^k$  of  $T$  and their corresponding leaf sets  $X^1, \dots, X^k$ , that are adjacent to the path from  $w$  up to  $\rho$ . These leaf sets partition the leaf set of  $T''$ , and if we add the additional set  $\{w_{0.5}\}$ , then this collection of  $k + 1$  sets partitions the leaves of  $T'_\pi$  and  $T'_{0.5}$ . By Theorem 2(i) we have:

$$RA_\varphi(T'') = 1 - \sum_{x \in X^1 \cup \dots \cup X^k} \left(\frac{1}{2}\right)^{d''(x)} p(x),$$

where  $d''(x)$  is the number of edges between the root and a leaf  $x$  in  $T''$ , and:

$$RA_\varphi(T'_{0.5}) = 1 - \sum_{x \in X^1 \cup \dots \cup X^k \cup \{w_{0.5}\}} \left(\frac{1}{2}\right)^{d'(x)} p(x),$$

where  $d'(x)$  is the number of edges between the root and leaf  $x$  in  $T'_{0.5}$ . Moreover, note that for  $i = 2, \dots, k$ :

$$\sum_{x \in X^i} \left(\frac{1}{2}\right)^{d''(x)} p(x) = \sum_{x \in X^i} \left(\frac{1}{2}\right)^{d'(x)} p(x).$$

Thus,  $RA_\varphi(T'') - RA_\varphi(T'_{0.5})$  becomes:

$$\begin{aligned} & RA_\varphi(T'') - RA_\varphi(T'_{0.5}) \\ &= 1 - \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d''(x)} p(x) - \left(1 - \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d'(x)} p(x) - \left(\frac{1}{2}\right)^k p(w_{0.5})\right) \\ &= \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d'(x)} p(x) - \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d''(x)} p(x) + \left(\frac{1}{2}\right)^k p(w_{0.5}). \end{aligned}$$

We have  $\pi = \frac{1}{2}$ , which gives us  $p(w_{0.5}) = p(w) + \pi - 2p(w)\frac{1}{2} = p(w) + \frac{1}{2} - p(w) = \frac{1}{2}$ . Again, note that the vertex  $w$  is suppressed in  $T'_{0.5}$  in order to keep the tree binary,

and thus  $k$  edges separate the root and the leaf  $w_{0.5}$ . Similarly, the vertex leading to subtree  $t^1$  is suppressed in  $T''$  to keep the tree binary. This gives us that  $k - 1$  edges separate the root of  $T''$  and the root of  $t^1$ , whereas  $k$  edges separate the root of  $T'_{0.5}$  and the root of the subtree  $t^1$ . Let  $d_1(x)$  denote the number of edges between the root of subtree  $t^1$  and leaf  $x$  in  $T$ , then we have

$$d'(x) = k + d_1(x) \text{ and } d''(x) = k - 1 + d_1(x).$$

If we now rearrange the above expression for  $RA_\varphi(T'') - RA_\varphi(T'_{0.5})$ , noting that  $p(w_{0.5}) = \frac{1}{2}$  we obtain:

$$\begin{aligned} RA_\varphi(T'') - RA_\varphi(T'_{0.5}) &= \sum_{x \in X^1} \left(\frac{1}{2}\right)^{k+d_1(x)} p(x) - \sum_{x \in X^1} \left(\frac{1}{2}\right)^{k-1+d_1(x)} p(x) + \left(\frac{1}{2}\right)^{k+1}, \\ &= \left(\frac{1}{2}\right)^k \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d_1(x)} p(x) - \left(\frac{1}{2}\right)^{k-1} \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d_1(x)} p(x) + \left(\frac{1}{2}\right)^{k+1} \\ &= -\left(\frac{1}{2}\right)^k \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d_1(x)} p(x) + \left(\frac{1}{2}\right)^{k+1} \\ &= \left(\frac{1}{2}\right)^{k+1} \left(1 - 2 \sum_{x \in X^1} \left(\frac{1}{2}\right)^{d_1(x)} p(x)\right) \\ &\geq 0, \end{aligned}$$

since  $\sum_{x \in X^1} \left(\frac{1}{2}\right)^{d_1(x)} = 1$  and  $0 \leq p(x) \leq \frac{1}{2}$  in the  $N_2$  model. Therefore, we have  $RA_\varphi(T'') \geq RA_\varphi(T'_{0.5})$ , which, together with the induction hypothesis, gives  $RA_{MP}(T'') \geq RA_\varphi(T'') \geq RA_\varphi(T'_{0.5})$  and thus completes the proof.  $\square$

Note that combining the statement of Theorem 4 with Theorem 2 gives us an alternative proof of Corollary 1, since  $RA_{MP}(T) \geq RA_\varphi(T) \geq 1 - p_{\max}$  (i.e. under the  $N_2$  model, the Fitch algorithm using all terminal taxa is at least as accurate for ancestral state reconstruction as selecting the state of a taxon  $x$  that maximises  $p(x)$ ).

### 5 Further results for the $r$ -state setting

In this section, we will indicate the set of states in  $\mathcal{A}$  by writing  $\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$ , and, unless stated otherwise, we assume the root is in state  $\alpha_1$ . For a set  $\mathcal{R} \subseteq \mathcal{A}$ ,  $\alpha_1 \in \mathcal{R}$ ,  $|\mathcal{R}| = k$ , let

$$P_k(T) := \mathbb{P}(\text{FS}(f, T) = \mathcal{R} | F(\rho) = \alpha_1).$$

Similarly, for a set  $\mathcal{R} \subseteq \mathcal{A}$ ,  $\alpha_1 \notin \mathcal{R}$ ,  $|\mathcal{R}| = k$ , let

$$Q_k(T) := \mathbb{P}(\text{FS}(f, T) = \mathcal{R} | F(\rho) = \alpha_1).$$



By the symmetry in the model, the values  $P_k(T)$  and  $Q_k(T)$  are independent of the choice of  $\mathcal{R}$ , subject to the constraints imposed on  $\mathcal{R}$  in their definition.

**Lemma 3** *For any rooted binary phylogenetic tree  $T$  under the  $N_r$  model, the reconstruction accuracy of MP is given by:*

$$RA_{MP}(T) = \frac{1}{r} \left( 1 + \sum_{k=1}^{r-1} \binom{r-1}{k} (P_k(T) - Q_k(T)) \right).$$

**Proof** Let  $T$  be a rooted binary phylogenetic tree and let  $P_k(T)$  and  $Q_k(T)$  be as defined above (so we assume the root to be in state  $\alpha_1$ ). For the reconstruction accuracy of MP under the  $N_r$  model, Eq. (1) and the law of total probability gives:

$$\begin{aligned} RA_{MP}(T) &= \sum_{\substack{\mathcal{R}: \mathcal{R} \subseteq \mathcal{A} \\ \text{and } \alpha_1 \in \mathcal{R}}} \frac{1}{|\mathcal{R}|} \cdot \mathbb{P}(\text{FS}(f, T) = \mathcal{R} | F(\rho) = \alpha_1) \\ &= \sum_{k=1}^r \frac{1}{k} P_k(T) \binom{r-1}{k-1} = \sum_{k=1}^{r-1} \frac{1}{k} P_k(T) \binom{r-1}{k-1} + \frac{1}{r} P_r(T) \\ &= \sum_{k=1}^{r-1} \frac{1}{k} P_k(T) \binom{r-1}{k-1} + \frac{1}{r} \left( 1 - \sum_{k=1}^{r-1} Q_k(T) \binom{r-1}{k} - \sum_{k=1}^{r-1} P_k(T) \binom{r-1}{k-1} \right). \end{aligned}$$

Rearranging this last expression gives:

$$\begin{aligned} RA_{MP}(T) &= \frac{1}{r} + \sum_{k=1}^{r-1} \left( \frac{1}{k} - \frac{1}{r} \right) P_k(T) \binom{r-1}{k-1} - \frac{1}{r} \sum_{k=1}^{r-1} Q_k(T) \binom{r-1}{k} \\ &= \frac{1}{r} + \frac{1}{r} \sum_{k=1}^{r-1} \frac{r-k}{k} P_k(T) \binom{r-1}{k-1} - Q_k(T) \binom{r-1}{k} \\ &= \frac{1}{r} + \frac{1}{r} \sum_{k=1}^{r-1} \binom{r-1}{k} (P_k(T) - Q_k(T)) \\ &= \frac{1}{r} \left( 1 + \sum_{k=1}^{r-1} \binom{r-1}{k} (P_k(T) - Q_k(T)) \right). \end{aligned}$$

□

For the following lemma we consider  $\dot{T}$  obtained from  $T$  by adding an additional stem edge  $(\rho, \rho')$  and substitution probability  $p_\rho$  on this edge. Let

$$P_{\alpha_1, \dots, \alpha_k}(\dot{T}) := \mathbb{P}(\text{FS}(f, \dot{T}) = \{\alpha_1, \dots, \alpha_k\} | F(\rho) = \alpha_1) \text{ and}$$

$$P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T}) := \mathbb{P}(\text{FS}(f, \dot{T}) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho) = \alpha_1)$$

**Lemma 4** Assume that  $\rho$  is in state  $\alpha_1$ . Under the  $N_r$  model and  $1 \leq k \leq r - 1$ , we have:

$$P_{\alpha_1, \dots, \alpha_k}(\dot{T}) = \left(1 - \frac{r - k}{r - 1} p_\rho\right) P_k(T) + \frac{r - k}{r - 1} p_\rho Q_k(T), \text{ and}$$

$$P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T}) = \left(1 - \frac{k}{r - 1} p_\rho\right) Q_k(T) + \frac{k}{r - 1} p_\rho P_k(T),$$

where  $P_k(T)$  and  $Q_k(T)$  are as defined above.

**Proof** For  $1 \leq k \leq r - 1$ , we can write  $P_{\alpha_1, \dots, \alpha_k}(\dot{T})$  as follows:

$$(1 - p_\rho) \mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1) + \frac{p_\rho}{r - 1} S, \tag{10}$$

where

$$S = \sum_{i=2}^r \mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1).$$

We can now split  $S$  into two sums depending on the range of  $k$ . Thus we have  $S = S_1 + S_2$ , where:

$$S_1 = \sum_{i=2}^k \mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1), \text{ and}$$

$$S_2 = \sum_{i=k+1}^r \mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1).$$

Notice also that, by the symmetry of the model, each of the  $k - 1$  terms in  $S_1$  is equal to

$$\mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1),$$

which is  $P_k(T)$ . Thus  $S_1 = (k - 1)P_k(T)$ .

Similarly, each of the  $r - k$  terms in  $S_2$  is equal to

$$\mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1),$$

which is just  $Q_k(T)$ , and thus  $S_2 = (r - k)Q_k(T)$ . Thus, from the expression for  $P_{\alpha_1, \dots, \alpha_k}(\dot{T})$  given by (10), we have:

$$P_{\alpha_1, \dots, \alpha_k}(\dot{T}) = (1 - p_\rho)P_k(T) + \frac{p_\rho}{r - 1}((k - 1)P_k(T) + (r - k)Q_k(T)).$$

Rearranging the term on the right gives the expression for  $P_{\alpha_1, \dots, \alpha_k}(\dot{T})$  in Lemma 4.

The second part of Lemma 4 follows by an analogous argument. For  $1 \leq k \leq r - 1$ , we can write  $P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T})$  as follows:

$$(1 - p_\rho)\mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1) + \frac{P_\rho}{r - 1} S', \tag{11}$$

where

$$S' = \sum_{i=2}^r \mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1).$$

Write  $S' = S'_1 + S'_2$  where:

$$S'_1 = \sum_{i=2}^k \mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1), \text{ and}$$

$$S'_2 = \sum_{i=k+1}^r \mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_i, F(\rho) = \alpha_1).$$

Notice also that, by the symmetry of the model, each of the  $k$  terms in  $S'_1$  is equal to

$$\mathbb{P}(\text{FS}(f, T) = \{\alpha_1, \dots, \alpha_k\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1),$$

which is  $P_k(T)$ . Thus  $S'_1 = kP_k(T)$ .

Similarly, each of the  $r - k - 1$  terms in  $S'_2$  is equal to

$$\mathbb{P}(\text{FS}(f, T) = \{\alpha_2, \dots, \alpha_{k+1}\} | F(\rho') = \alpha_1, F(\rho) = \alpha_1),$$

which is just  $Q_k(T)$ , and thus  $S'_2 = (r - k - 1)Q_k(T)$ . Thus, from the expression for  $P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T})$  given by (11) we have:

$$P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T}) = (1 - p_\rho)Q_k(T) + \frac{P_\rho}{r - 1}(kP_k(T) + (r - k - 1)Q_k(T)).$$

Rearranging the term on the right gives the expression for  $P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T})$  in Lemma 4. □

By the proof of Lemma 4, we have the following corollary.

**Corollary 2** *Let  $\dot{T}$  be a rooted binary phylogenetic tree with stem edge  $(\rho, \rho')$ . Consider the  $N_r$  model with state space  $\mathcal{A} = \{\alpha_1, \dots, \alpha_r\}$ , assume the root  $\rho$  is in state  $\alpha_1$ , and let  $p_\rho$  be the substitution probability on the stem edge. Then, for  $1 \leq k \leq r - 1$  we have:*

- (i)  $P_{\alpha_1, \dots, \alpha_k}(\dot{T}) = P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T}) + \left(1 - \frac{r}{r-1} p_\rho\right) (P_k(T) - Q_k(T))$
- (ii) If  $P_k(T) \geq Q_k(T)$ , then  $P_{\alpha_1, \dots, \alpha_k}(\dot{T}) \geq P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T})$ .

Notice also, that if the substitution probability on every edge is strictly less than  $\frac{r-1}{r}$  (as required by an underlying continuous-time Markov realisation of the process), then the following strict inequality result holds: if  $P_k(T) > Q_k(T)$ , then  $P_{\alpha_1, \dots, \alpha_k}(\dot{T}) > P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T})$ .

In Theorem 5 we consider a rooted binary tree  $T$  as depicted in Fig. 3.

**Theorem 5** *Let  $T$  be a rooted binary phylogenetic tree under the  $N_r$  model. For  $1 \leq k \leq r - 1$  we have:  $P_k(T) \geq Q_k(T)$ .*

**Proof** Since the root is assumed to be in state  $\alpha_1$  and by the definition of  $P_k(T)$  and  $Q_k(T)$  we have that

$$P_k(T) = P_{\alpha_1, \dots, \alpha_k}(T) \text{ and } Q_k(T) = P_{\alpha_2, \dots, \alpha_{k+1}}(T).$$

The proof is by induction on the number of leaves  $n$ . The inequality holds trivially for  $n = 1$ ; for  $n = 2$ , let  $p_x, p_y$  denote the substitution probabilities on the two edges of the tree. We then have:

$$\begin{aligned} P_{\alpha_1}(T) &= (1 - p_x)(1 - p_y); & P_{\alpha_2}(T) &= \frac{p_x}{r-1} \frac{p_y}{r-1}, \\ P_{\alpha_1\alpha_2}(T) &= (1 - p_x) \frac{p_y}{r-1} + \frac{p_x}{r-1} (1 - p_y); & P_{\alpha_2\alpha_3}(T) &= 2 \frac{p_x}{r-1} \frac{p_y}{r-1}. \end{aligned}$$

Moreover, we have:

$$\begin{aligned} &P_{\alpha_1}(T) - P_{\alpha_2}(T) \\ &= (1 - p_x)(1 - p_y) - \frac{p_x}{r-1} \frac{p_y}{r-1} = 1 - p_x - p_y + p_x p_y - \frac{p_x}{r-1} \frac{p_y}{r-1} \\ &= \left(1 - \frac{r}{r-1} p_x\right) \left(1 - \frac{r}{r-1} p_y\right) + \frac{p_x}{r-1} \left(1 - \frac{r}{r-1} p_y\right) \\ &\quad + \frac{p_y}{r-1} \left(1 - \frac{r}{r-1} p_x\right), \end{aligned}$$

and

$$\begin{aligned} P_{\alpha_1\alpha_2}(T) - P_{\alpha_2\alpha_3}(T) &= (1 - p_x) \frac{p_y}{r-1} + \frac{p_x}{r-1} (1 - p_y) - 2 \frac{p_x}{r-1} \frac{p_y}{r-1} \\ &= \frac{p_x}{r-1} \left(1 - \frac{r p_y}{r-1}\right) + \frac{p_y}{r-1} \left(1 - \frac{r p_x}{r-1}\right), \end{aligned}$$

which are both non-negative, since  $p_x, p_y \leq \frac{r-1}{r}$ . This gives the base case of the induction. We now assume that the induction hypothesis holds for all trees with fewer

than  $n$  leaves and show that it also holds for a tree  $T$  with  $n$  leaves. Consider the decomposition of  $T$  into its two maximal pending subtrees  $T_1$  and  $T_2$  and the associated trees  $\dot{T}_1$  and  $\dot{T}_2$  with a stem edge (as in Fig. 3). By the induction hypothesis,  $P_{\alpha_1, \dots, \alpha_k}(T_i) \geq P_{\alpha_2, \dots, \alpha_{k+1}}(T_i)$  holds for  $i \in \{1, 2\}$ . By combining this with Corollary 2 (ii), we obtain:

$$P_{\alpha_1, \dots, \alpha_k}(\dot{T}_i) \geq P_{\alpha_2, \dots, \alpha_{k+1}}(\dot{T}_i) \tag{12}$$

for  $i \in \{1, 2\}$ . Moreover,  $P_k(T)$  and  $Q_k(T)$  are given as follows. Let  $\omega_\alpha := \{\alpha_1, \dots, \alpha_k\}$  and  $\omega_\beta := \{\alpha_2, \dots, \alpha_{k+1}\}$ , and in the following equations,  $\omega_1$  and  $\omega_2$  vary over all the nonempty subsets of  $\mathcal{A}$  that satisfy the stated constraints under the summation signs of the following two equations:

$$P_k(T) = \underbrace{P_{\alpha_1, \dots, \alpha_k}(T)}_{:=\omega_\alpha} = \sum_{\omega_1 \cap \omega_2 = \omega_\alpha} P_{\omega_1}(\dot{T}_1) P_{\omega_2}(\dot{T}_2) + \sum_{\substack{\omega_1 \cap \omega_2 = \emptyset, \\ \omega_1 \cup \omega_2 = \omega_\alpha}} P_{\omega_1}(\dot{T}_1) P_{\omega_2}(\dot{T}_2) \tag{13}$$

and

$$Q_k(T) = \underbrace{P_{\alpha_2, \dots, \alpha_{k+1}}(T)}_{:=\omega_\beta} = \sum_{\omega_1 \cap \omega_2 = \omega_\beta} P_{\omega_1}(\dot{T}_1) P_{\omega_2}(\dot{T}_2) + \sum_{\substack{\omega_1 \cap \omega_2 = \emptyset, \\ \omega_1 \cup \omega_2 = \omega_\beta}} P_{\omega_1}(\dot{T}_1) P_{\omega_2}(\dot{T}_2). \tag{14}$$

To show that  $P_k(T) \geq Q_k(T)$ , our strategy is to show that the first term (summation) the right-hand side of Eq. (13) is greater or equal to the first term (summation) on the right-hand side Eq. (14). We then show that same inequality also holds for the second summation term.

For any set  $\omega_1^\alpha$  and  $\omega_2^\alpha$  there exist corresponding sets  $\omega_1^\beta$  and  $\omega_2^\beta$ . The corresponding set (for  $i \in \{1, 2\}$ ) is:

$$\omega_i^\beta = \begin{cases} \omega_i^\alpha \setminus \{\alpha_1\} \cup \{\alpha_{k+1}\} & \text{if } \alpha_1 \in \omega_i^\alpha \text{ and } \alpha_{k+1} \notin \omega_i^\alpha \\ \omega_i^\alpha & \text{otherwise.} \end{cases} \tag{15}$$

For the first half of this argument, take any two sets  $\omega_1^\alpha$  and  $\omega_2^\alpha$  for which  $\omega_1^\alpha \cap \omega_2^\alpha = \omega_\alpha$ . Note that  $\alpha_1$  is contained in  $\omega_1^\alpha$  and  $\omega_2^\alpha$ . Then, the corresponding sets  $\omega_1^\beta$  and  $\omega_2^\beta$  [from (15)] satisfy  $|\omega_1^\alpha| = |\omega_1^\beta|$  and  $|\omega_2^\alpha| = |\omega_2^\beta|$  and  $\omega_1^\beta \cap \omega_2^\beta = \omega_\beta$ . Here, we consider two cases.

**Case (i)**  $\alpha_1 \notin \omega_1^\beta$  and  $\alpha_1 \notin \omega_2^\beta$ .

By Eq. (12), we have  $P_{\omega_1^\alpha}(\dot{T}_1) \geq P_{\omega_1^\beta}(\dot{T}_1)$  and  $P_{\omega_2^\alpha}(\dot{T}_2) \geq P_{\omega_2^\beta}(\dot{T}_2)$ . Thus,  $P_{\omega_1^\alpha}(\dot{T}_1) P_{\omega_2^\alpha}(\dot{T}_2) \geq P_{\omega_1^\beta}(\dot{T}_1) P_{\omega_2^\beta}(\dot{T}_2)$ , which completes the first case.

**Case (ii)**  $\alpha_1$  is contained in  $\omega_1^\beta$  or in  $\omega_2^\beta$  (not both).

Without loss of generality, we have  $\alpha_1 \in \omega_1^\beta$  and  $\alpha_1 \notin \omega_2^\beta$ . We know that  $P_{\omega_1^\alpha}(\dot{T}_1) = P_{\omega_1^\beta}(\dot{T}_1)$  and by Eq. (12), we have  $P_{\omega_2^\alpha}(\dot{T}_2) \geq P_{\omega_2^\beta}(\dot{T}_2)$ . Thus,  $P_{\omega_1^\alpha}(\dot{T}_1)P_{\omega_2^\alpha}(\dot{T}_2) \geq P_{\omega_1^\beta}(\dot{T}_1)P_{\omega_2^\beta}(\dot{T}_2)$  holds.

This completes the first half of the argument.

We now compare the last terms on the right-hand side of the Eqs. (13) and (14) for  $P_k(T)$  and  $Q_k(T)$ . Take any two sets  $\omega_1^\alpha$  and  $\omega_2^\alpha$  for which  $\omega_1^\alpha \cap \omega_2^\alpha = \emptyset$  and  $\omega_1^\alpha \cup \omega_2^\alpha = \omega_\alpha$ . Without loss of generality, we have  $\alpha_1 \in \omega_1^\alpha$  and  $\alpha_1 \notin \omega_2^\alpha$ . Then, the corresponding sets  $\omega_1^\beta$  and  $\omega_2^\beta$  [from Eq. (15)] satisfy  $|\omega_1^\alpha| = |\omega_1^\beta|$  and  $|\omega_2^\alpha| = |\omega_2^\beta|$  such that  $\omega_1^\beta \cap \omega_2^\beta = \emptyset$  and  $\omega_1^\beta \cup \omega_2^\beta = \omega_\beta$ . Since  $\alpha_1 \in \omega_1^\alpha$  and  $\alpha_1 \notin \omega_2^\alpha$ , we have  $P_{\omega_2^\alpha}(\dot{T}_2) = P_{\omega_2^\beta}(\dot{T}_2)$  and, by Eq. (12), we have  $P_{\omega_1^\alpha}(\dot{T}_1) \geq P_{\omega_1^\beta}(\dot{T}_1)$ . Thus,  $P_{\omega_1^\alpha}(\dot{T}_1)P_{\omega_2^\alpha}(\dot{T}_2) \geq P_{\omega_1^\beta}(\dot{T}_1)P_{\omega_2^\beta}(\dot{T}_2)$  holds.

Therefore,  $P_k(T)$  is greater than or equal to  $Q_k(T)$  for tree  $T$  by induction from  $\dot{T}_1$  and  $\dot{T}_2$ . □

Combining Lemma 3 with Theorem 5 gives the following corollary, which states that the reconstruction accuracy of MP under the  $N_r$  model is greater or equal to  $\frac{1}{r}$ . In addition, note that if we assume the probabilities of change to be strictly less than  $\frac{r-1}{r}$ , we can then show that  $P_k(T) > Q_k(T)$  by induction on  $n$  similar to the proof of Theorem 5. This gives us  $RA_{MP}(T) > \frac{1}{r}$ .

**Corollary 3** *For any rooted binary phylogenetic tree  $T$  and the  $N_r$  model, we have:*

$$RA_{MP}(T) \geq \frac{1}{r}.$$

*Moreover, this inequality is strict under a continuous-time  $N_r$  model where  $p_e < \frac{r-1}{r}$ .*

## 6 A combinatorial sufficient condition for accurate ancestral state reconstruction

In this penultimate section, we present a new combinatorial property of ancestral state reconstruction using parsimony. More precisely, we provide a sufficient condition for MP to recover the ancestral state at the root vertex  $\rho$  correctly from the observed states at the leaves. Note that this does not make any model assumptions (as in the previous section) as to how the character  $f$  is generated—it simply requires the state changes to be spread sufficiently thinly in the tree as one moves way from the root. This result complements a related (but quite different) result from Steel and Penny (2005) (Theorem 9.4.5).

Let  $n_i$  ( $i = 1, 2, \dots$ ) be the number of edges descended from  $\rho$  and separated from  $\rho$  by  $i - 1$  other edges on which a substitution occurs. Thus  $n_1$  counts the number (0, 1, 2) of edges out of  $\rho$  on which substitutions occur. Note that  $n_i$  is not just a function of the tree and the character at the leaves; it depends on the actual evolution of this

character on the tree. We refer to  $n_i$  as the *substitution spectrum* of the character on the tree relative to the root vertex  $\rho$ .

The following theorem can be regarded as a type of combinatorial local ‘safety radius’ for MP to infer the ancestral state at a given vertex correctly (even though the states at other vertices may not be correctly reconstructed).

**Theorem 6** *Consider any binary tree  $T$  on any number of leaves, and any character (involving any number of states) that has evolved on this tree with a substitution spectrum relative to root vertex  $\rho$  that satisfies the inequality:*

$$\sum_{k \geq 1} n_k \left( \frac{1}{\sqrt{2}} \right)^k < \frac{1}{2}. \tag{16}$$

*The set of most parsimonious state at vertex  $\rho$  estimated from the states at the leaves descending from  $\rho$  consists precisely of the true ancestral state at  $\rho$  (i.e.  $\text{FS}(\rho) = \{F(\rho)\}$ ).*

**Proof** First observe that is sufficient to establish this result for a complete balanced binary tree  $T_h$  of arbitrary height  $h$ , with  $\rho$  being the root of  $T_h$ . We use induction on the height  $h$  of the tree. Let  $\alpha$  denote the state  $F(\rho)$  present at the root of  $T_h$ . For  $h \leq 2$  we have  $n_k = 0$  for all  $k > 2$ . Inequality (16) ensures that  $n_1 = n_2 = 0$ , in which case all leaves are in state  $\alpha$  and so the Fitch set  $\text{FS}(v)$  for  $\rho$  is the set  $\{\alpha\}$ . This establishes the result for  $h \leq 2$ .

For the induction step, suppose that the result holds for  $T_{h-2}$  and  $T_{h-1}$  and consider the tree  $T_h$  together with a character evolved on  $T_h$  for which Inequality (16) applies for vertex  $\rho$ . As before, this inequality ensures that none of the six edges at distance 1 or 2 descending from  $\rho$  have a substitution on them.

If  $T^1$  and  $T^2$  are the two maximal subtrees of  $T_h$ , then (i) each of these trees is of the type  $T_{h-1}$ , and (ii) the following identity holds for all  $k$ :

$$n_k = n_{k-1}^1 + n_{k-1}^2, \tag{17}$$

where  $n_i^1$  (resp.  $n_i^2$ ) is substitution spectrum for the character’s evolution on  $T^1$  and  $T^2$  [note that we are using the fact that no substitution occurs on either of the two edges outgoing from  $\rho$ , by Inequality (16)].

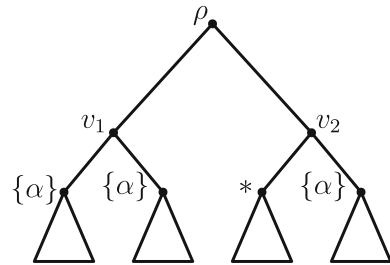
Thus if we let

$$p_h(\mathbf{n}, \theta) := \sum_{k \geq 1} n_k \theta^k,$$

where  $\mathbf{n} = [n_k]$ , then Eq. (17) allows us to write:

$$p_h(\mathbf{n}, \theta) = \theta \cdot \left[ p_{h-1}(\mathbf{n}^1, \theta) + p_{h-1}(\mathbf{n}^2, \theta) \right].$$

**Fig. 4** If the Fitch sets at the roots of three of the four subtrees at distance 2 from  $\rho$  in  $T_h$  ( $h \geq 3$ ) consist of the singleton set  $\{\alpha\}$ , then  $\text{FS}(v) = \{\alpha\}$  as well, regardless of the Fitch set  $*$  at the root of the fourth subtree



We can extend this argument one level further to obtain the following:

$$p_h(\mathbf{n}, \theta) = \theta^2 \cdot \left[ p_{h-2}(\mathbf{n}^{11}, \theta) + p_{h-2}(\mathbf{n}^{12}, \theta) + p_{h-2}(\mathbf{n}^{21}, \theta) + p_{h-2}(\mathbf{n}^{22}, \theta) \right], \tag{18}$$

where  $\mathbf{n}^{ij}$  refers the substitution spectra on the four subtrees of type  $T_{h-2}$  that are two edges descending from the vertex  $\rho$  in  $T$ . Note that in writing Eq. (18) we are again using the fact that Inequality (16) precludes any substitutions in the six edges descended from  $\rho$  and at distance at most 2 from it.

Now put  $\theta = \frac{1}{\sqrt{2}}$  in Eq. (18) and let  $x_{ij} := p_{h-2}(\mathbf{n}^{ij}, \frac{1}{\sqrt{2}})$ . We then obtain:

$$p_h(\mathbf{n}, \frac{1}{\sqrt{2}}) = \frac{1}{2}(x_{11} + x_{12} + x_{21} + x_{22}). \tag{19}$$

Since we are assuming that  $p_h(\mathbf{n}, \frac{1}{\sqrt{2}}) < \frac{1}{2}$  [by Inequality (16)], it follows from Eq. (19) that at least three of the four terms  $x_{ij}$  are strictly less than  $\frac{1}{2}$ , since if two of them were greater or equal to  $\frac{1}{2}$  then  $\frac{1}{2}(x_{11} + x_{12} + x_{21} + x_{22}) \geq \frac{1}{2}$ . By the induction hypothesis, three (or four) of the corresponding vertices (two edges descending from  $\rho$ ) have an FS value of  $\{\alpha\}$ , as shown in Fig. 4.

We now invoke a simple combinatorial observation: if a vertex  $v$  in a binary tree has the property that at least three vertices that are two edges descended from  $v$  have their Fitch set FS equal to  $\{\alpha\}$ , then  $\text{FS}(v) = \{\alpha\}$ . This establishes the induction step, and thereby the theorem. □

**Remark** An interesting question is the following: What is the smallest value of  $\theta$  for which there is a constant  $t$  so that the condition  $p_h(\mathbf{n}, \theta) < t$  implies that  $\text{FS}(v) = \{\alpha\}$  for all values of  $h$  and substitution spectra  $\mathbf{n}$ ? We have shown that the value  $\theta = \frac{1}{\sqrt{2}} \approx 0.7071$  (or any larger value) suffices, and it is known (from Theorem 2 of Steel and Charleston 1995) that  $\theta$  cannot be smaller than the reciprocal of the golden ratio (i.e.  $2/(1 + \sqrt{5}) \approx 0.6180$ ).



### 7 Concluding comments

Theorem 4 demonstrated that  $RA_{MP}(T) \geq RA_{\varphi}(T)$  when  $r = 2$ . An interesting question is whether or not this holds more generally. This leads us to pose the following conjecture:

**Conjecture 1** Let  $T$  be a rooted binary phylogenetic tree. Under the  $N_r$  model, the reconstruction accuracy of MP is at least equal to the reconstruction accuracy of the coin-toss method:

$$RA_{MP}(T) \geq RA_{\varphi}(T).$$

This conjecture holds for  $n = 2$  and all values of  $r \geq 2$ , as it is an exact equality in that case.

By using Theorem 3, and induction on the number of leaves, it can be shown that Conjecture 1 is equivalent to the following statement:

$$RA_{MP}(T) \geq \frac{1}{2}(RA_{MP}(\dot{T}_1) + RA_{MP}(\dot{T}_2)), \tag{20}$$

where  $\dot{T}_1$  and  $\dot{T}_2$  are the two pending subtrees of  $T$  as in Fig. 3.

Inequality (20) holds when  $r = 2$  since, as stated, it is equivalent to the above conjecture, and this holds when  $r = 2$  by Theorem 4. In the Appendix we give a direct alternative argument to justify Inequality (20) in the case  $r = 2$ .

**Acknowledgements** Lina Herbst thanks the University of Greifswald for the Landesgraduiertenförderung studentship and the German Academic Exchange Service (DAAD) for the DAAD-Doktorandenstipendium. Mike Steel thanks the New Zealand Marsden Fund (UOC-1709). We also thank Mareike Fischer for several helpful comments, Santiago Catalano for references to some recent biological studies, and the two anonymous reviewers for numerous helpful comments on an earlier version of this manuscript.

### 8 Appendix: Direct proof of inequality (20) when $r = 2$

**Proof** For the  $N_2$  model,  $RA_{MP}(T) = P_{\alpha}(T) + \frac{1}{2}P_{\alpha\beta}(T)$ . Thus,

$$RA_{MP}(T) = P_{\alpha}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\alpha\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + \frac{1}{2}(P_{\alpha\beta}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\beta}(\dot{T}_2) + P_{\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2)). \tag{21}$$

Moreover:

$$\begin{aligned} \frac{1}{2}(RA_{MP}(\dot{T}_1) + RA_{MP}(\dot{T}_2)) &= \frac{1}{2}(P_{\alpha}(\dot{T}_1) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_1) + P_{\alpha}(\dot{T}_2) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_2)) \\ &= \frac{1}{2}(P_{\alpha}(\dot{T}_1) + P_{\alpha}(\dot{T}_2)) + \frac{1}{4}(P_{\alpha\beta}(\dot{T}_1) + P_{\alpha\beta}(\dot{T}_2)) \\ &= \frac{1}{2}(P_{\alpha}(\dot{T}_1)(P_{\alpha}(\dot{T}_2) + P_{\beta}(\dot{T}_2) + P_{\alpha\beta}(\dot{T}_2)) + P_{\alpha}(\dot{T}_2)(P_{\alpha}(\dot{T}_1) + P_{\beta}(\dot{T}_1))) \end{aligned}$$

$$\begin{aligned}
 &+ P_{\alpha\beta}(\dot{T}_1)) \\
 &+ \frac{1}{4}(P_{\alpha\beta}(\dot{T}_1) + P_{\alpha\beta}(\dot{T}_2))(\text{by the law of total probability}) \\
 = &\frac{1}{2}(2P_{\alpha}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\beta}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2) \\
 &+ P_{\alpha\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2)) \\
 &+ \frac{1}{4}(P_{\alpha\beta}(\dot{T}_1) + P_{\alpha\beta}(\dot{T}_2)). \tag{22}
 \end{aligned}$$

In order to show that  $RA_{MP}(T) \geq \frac{1}{2}(RA_{MP}(\dot{T}_1) + RA_{MP}(\dot{T}_2))$ , we establish the following inequality:

$$RA_{MP}(T) - \frac{1}{2}(RA_{MP}(\dot{T}_1) + RA_{MP}(\dot{T}_2)) \geq 0.$$

By (21) and (22) we have:

$$\begin{aligned}
 &RA_{MP}(T) - \frac{1}{2}(RA_{MP}(\dot{T}_1) + RA_{MP}(\dot{T}_2)) \\
 = &P_{\alpha}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\alpha\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2) \\
 &+ \frac{1}{2}(P_{\alpha\beta}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\beta}(\dot{T}_2) + P_{\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2)) \\
 &- \frac{1}{2}(2P_{\alpha}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\beta}(\dot{T}_2) + P_{\alpha}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + P_{\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2) \\
 &+ P_{\alpha\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2)) \\
 &- \frac{1}{4}(P_{\alpha\beta}(\dot{T}_1) + P_{\alpha\beta}(\dot{T}_2)) \\
 = &\frac{1}{2}P_{\alpha}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_1)P_{\alpha}(\dot{T}_2) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_1)P_{\alpha\beta}(\dot{T}_2) \\
 &- \frac{1}{4}(P_{\alpha\beta}(\dot{T}_1) + P_{\alpha\beta}(\dot{T}_2)) \\
 = &\frac{1}{2}P_{\alpha\beta}(\dot{T}_1)(P_{\alpha}(\dot{T}_2) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_2) - \frac{1}{2}) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_2)(P_{\alpha}(\dot{T}_1) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_1) - \frac{1}{2}) \\
 = &\frac{1}{2}P_{\alpha\beta}(\dot{T}_1)(RA_{MP}(\dot{T}_2) - \frac{1}{2}) + \frac{1}{2}P_{\alpha\beta}(\dot{T}_2)(RA_{MP}(\dot{T}_1) - \frac{1}{2}).
 \end{aligned}$$

This last expression is non-negative because the reconstruction accuracy under the  $N_2$  model is greater or equal to  $\frac{1}{2}$  by Corollary 1, and (by Lemma 1),  $RA_{MP}(\dot{T}) \geq \frac{1}{2}$  if and only if  $RA_{MP}(T) \geq \frac{1}{2}$ . □

## References

- Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Bérard S, Chauve C, Scornavacca C, Daubin V, Tannier E (2017) DeCoSTAR: reconstructing in ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biol Evol* 9:1312–1319
- Felsenstein J (2004) *Inferring phylogenies*. Sinauer Press, Sunderland
- Fischer M, Thatte B (2009) Maximum parsimony on subsets of taxa. *J Theor Biol* 260:290–293
- Fitch WM (1971) Toward defining the course of evolution: minimal change for a specific tree topology. *Syst Zool* 20:406–416
- Gaschen B (2002) Diversity considerations in HIV-1 vaccine selection. *Science* 296:2354–2360
- Gascuel O, Steel M (2010) Inferring ancestral sequences in taxon-rich phylogenies. *Math Biosci* 227:125–135
- Gascuel O, Steel M (2014) Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Syst Biol* 63:421–435
- Göpel T, Wirkner CS (2018) Morphological description, character conceptualization and the reconstruction of ancestral states exemplified by the evolution of arthropod hearts. *PLoS ONE* 13:e0201702
- Hartigan JA (1973) Minimum mutation fits to a given tree. *Biometrics* 29:53–65
- Herbst L, Fischer M (2018) On the accuracy of ancestral sequence reconstruction for ultrametric trees with parsimony. *Bull Math Biol* 80:864–879
- Hsiang AY, Field D, Webster TH, Behlke A, Davis MB, Racicot RA, Gauthier JA (2015) The origin of snakes: revealing the ecology, behavior, and evolutionary history of early snakes using genomics, phenomics, and the fossil record. *BMC Evol Biol* 15:87
- Huelsbeck J, Bollback JP (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol* 50:351–366
- Jukes TH, Cantor C (1969) Evolution of protein molecules. In: Munro HN (ed) *Mammalian protein metabolism*. Academic Press, New York, pp 21–132
- Koshi JM, Goldstein RA (1996) Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol* 42:313–320
- Li G, Steel M, Zhang L (2008) More taxa are not necessarily better for the reconstruction of ancestral character states. *Syst Biol* 57:647–653
- Plachetzki DC, Fong CR, Oakley TH (2010) The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. *Proc R Soc Lond B Biol Sci* 277:1963–1969
- Pupko T, Pe I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* 17:890–896
- Sauquet H, von Balthazar M, Schönenberger J (2017) The ancestral flower of angiosperms and its early diversification. *Nat Commun* 8:16047
- Steel M (2016) *Phylogeny: discrete and random processes in evolution*. SIAM, Philadelphia
- Steel MA, Charleston M (1995) Five surprising properties of parsimoniously colored trees. *Bull Math Biol* 57:367–375
- Steel M, Penny D (2005) Maximum parsimony and the phylogenetic information in multi-state characters. In: Albert VA (ed) *Parsimony, phylogeny and genomics*. Oxford University Press, Oxford, pp 163–178
- Tuffley C, Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull Math Biol* 59:581–607
- Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650
- Zhang L, Shen J, Yang J, Li G (2010) Analyzing the Fitch method for reconstructing ancestral states on ultrametric phylogenetic trees. *Bull Math Biol* 72:1760–1782