

# Analyzing the Fitch Method for Reconstructing Ancestral States on Ultrametric Phylogenetic Trees

Louxin Zhang<sup>a,\*</sup>, Jian Shen<sup>b,†</sup>, Jialiang Yang<sup>c</sup>, Guoliang Li<sup>d</sup>

<sup>a</sup>*Department of Mathematics, National University of Singapore (NUS), Singapore 117543, Singapore*

<sup>b</sup>*Department of Mathematics, Texas State University, San Marcos, TX 78666, USA*

<sup>c</sup>*MPI-CAS Institute of Computational Biology, CAS at Shanghai, Shanghai, China*

<sup>d</sup>*Department of Computer Science, NUS, Singapore, Singapore*

Received: 31 July 2009 / Accepted: 7 January 2010  
© Society for Mathematical Biology 2010

**Abstract** The accuracy of the Fitch method for reconstructing ancestral states on ultrametric phylogenetic trees is studied. Two recurrence relations for computing the accuracy are given here. Using these relations, we analyze the convergence of the accuracy of the Fitch method for reconstructing the root state on a complete binary tree of  $2^n$  leaves as  $n$  goes to infinity, present a closed-form formula for the accuracy on ultrametric comb trees, and provide a lower bound on the accuracy on arbitrary ultrametric phylogenetic trees.

**Keywords** Ancestral state reconstruction · Reconstruction accuracy · Fitch method · Phylogenetic trees

## 1. Introduction

Ancestral sequence reconstruction incorporates sequences from modern living things into evolutionary models to estimate the corresponding sequence of an ancestor that lived millions of years ago. This approach to understanding proteins was first suggested by Pauling and Zuckerkandl (1963) in their seminal work in 1963. With the rapid accumulation of biomolecular sequence data and advances in computational biology, it has become an important approach to studying the origin and evolution of genes, proteins and even whole genomes (see, for example, Liberles, 2007 and Thornton, 2004). The reader is referred to the survey paper by Pachter (2007) for the mathematical issues of ancestral sequence reconstruction.

---

\*Corresponding author.

E-mail addresses: [matzlx@nus.edu.sg](mailto:matzlx@nus.edu.sg) (Louxin Zhang), [js48@txstate.edu](mailto:js48@txstate.edu) (Jian Shen), [yangjialiang@picb.ac.cn](mailto:yangjialiang@picb.ac.cn) (Jialiang Yang), [lgl@gis.a-star.edu.sg](mailto:lgl@gis.a-star.edu.sg) (Guoliang Li).

†Work of L. Zhang was supported by ARF(R146-000-109-112).

‡Work of J. Shen was partially supported by NSF (CNS 0835834) and Texas Higher Education Coordinating Board (ARP 003615-0039-200).

The evolutionary history of a set of extant taxa is modeled as a rooted binary tree with leaves each labeled by a taxon, called a phylogenetic tree, in which any internal node denotes the common ancestor of the taxa associated with the leaves below it. The state of a character changes during evolutionary course and, therefore, the extant taxa may have character states different from their common ancestor in a phylogenetic tree.

The goal of ancestral state reconstruction is to estimate the ancestral character state at the root from the states of the extant taxa in a phylogenetic tree. The Fitch method (Fitch, 1971) is the first phylogenetic technique used for inferring the ancestral states of a character when the phylogenetic tree that evolutionarily relates the ancestor to the extant species is known (Baba et al., 1984). Here, character states represent particular traits or morphological features. In ancestral DNA sequence reconstruction, the characters are simply sequence sites and the states are four nucleotides. As a parsimony method, it estimates the ancestral states by minimizing the total number of hypothetical changes on all branches that are used to explain the evolution of the character states. It is efficient and accurate for sequences that are reasonably similar to each other. However, the accuracy of the Fitch method for reconstructing ancestral states has yet to be well studied (Li et al., 2008; Maddison, 1995; Salisbury and Kim, 2001; Zhang and Nei, 1997).

In this work, we study three mathematical problems arising from the analysis of the Fitch method for ancestral sequence reconstruction. To solve these problems, we first prove two useful recurrence relations for calculating the reconstruction accuracy of the Fitch method (in Theorem 3.1). These relations are derived from formulas in earlier studies (Maddison, 1995) and (Steel, 1989).

The first problem is to analyze the convergence of the accuracy of the Fitch method for reconstructing the root state in a complete phylogenetic tree in the equal-length branch and two-state Jukes–Cantor model (see Section 2 for details). Let  $p$  denote the conservation probability on each branch. Steel (1989) showed that, when the Fitch method is applied, the accuracy of reconstructing the root state from all leaf states in the complete binary tree of  $2^n$  leaves converges as  $n$  goes to infinity to  $\frac{1}{2}$  if  $\frac{1}{8} \leq p \leq \frac{7}{8}$  and  $\frac{1}{2} + \frac{1}{2} \frac{\sqrt{(8p-7)(4p-3)}}{(2p-1)^2}$  if  $\frac{7}{8} \leq p \leq 1$ . This result was proved under the assumption that certain limits exists. However, the existence of these limits is not trivial as shown in this paper. We fill the gap left in Steel (1989) by proving their existence. Additionally, we show that the reconstruction accuracy diverges when  $p \leq \frac{1}{8}$ .

Complete phylogenetic trees in which all branches have equal length are special ultrametric trees. In an ultrametric tree, each branch has its own branch length  $l$ , with conservation rate  $p(l) = \frac{1}{2}(1 + e^{-\lambda l})$  in a symmetry two-state Jukes–Cantor model, but requiring that the sum of branch lengths is constant in each path from the root to a leaf. Hennigian comb trees are trees in which each internal node has at least one leaf child. Hennigian comb trees are also called caterpillar trees. We give a closed-form formula for the reconstruction accuracy of the Fitch method on ultrametric Hennigian comb trees.

Lastly, we study the reconstruction accuracy of the Fitch method on arbitrary ultrametric trees. Contradicting the intuitive, the reconstruction accuracy of the Fitch method is not a monotonic function of the number of taxa selected for reconstruction of the root state (Li et al., 2008). Therefore, in Li et al. (2008), it is asked whether the accuracy  $RA_F$  of the Fitch method for reconstructing the root state from all leaf states is always larger than or equal to the conservation probability along a root-to-leaf path or not on an ultrametric tree. Recently, this problem is positively answered by Fischer and Thatté (2009). In this paper, we present a stronger lower bound on  $RA_F$  for arbitrary ultrametric trees.

Our lower bound implies that  $RA_F$  is not less than the accuracy of reconstructing the root state from any three leaves in an arbitrary ultrametric tree with three or more leaves.

## 2. The Fitch method and its reconstruction accuracy

Let  $C$  be a character with multiple states. Given a phylogenetic tree  $T$  in which each leaf has a state for character  $C$ , the Fitch method estimates the root state from the leaf states in two steps. It first computes a subset  $S_u$  of states for each node  $u$  of  $T$  as follows:

1. If  $u$  is a leaf,  $S_u$  contains only the state of  $u$ ;
2. If  $u$  is an internal node having children  $v$  and  $w$ ,  $S_u$  is equal to  $S_v \cup S_w$  if  $S_v$  and  $S_w$  are disjoint and  $S_v \cap S_w$  otherwise.

After the subset  $S_r$  for the root  $r$  of  $T$  is computed, the method selects a state as the root state from  $S_r$  randomly. In other words, a state is selected as the root state with probability  $\frac{1}{|S_r|}$ , where  $|S_r|$  denotes the number of states contained in  $S_r$ .

Assume the state mutation process along each branch of the given phylogenetic tree  $T$  is modeled as a stochastic process in which a state is replaced by another with some probability. The Fitch method reconstructs correctly a root state  $s$  from a set  $D$  of leaf states only if  $s$  evolves into the leaf states in  $D$ . Hence, the accuracy of the Fitch method for reconstructing the state of the root in the tree  $T$ , denoted by  $RA_F(T)$ , is defined to be the expected probability that it outputs the true state on the given set  $D$  of leaf states. Let  $\Pr_r[D|s]$  denote the probability that the root state  $s$  evolves into the leaf states in  $D$ . Then

$$RA_F(T) = \sum_{s,D} p_r(s) \Pr_r[D|s] \Pr[s \text{ is selected by the Fitch method from } D], \quad (1)$$

where  $p_r(s)$  is the prior probability of  $s$  being the root state.

## 3. Recurrence relations for analyzing the reconstruction accuracy

In the rest of this paper, we assume that a character has only two states 0 and 1 and the root of a given phylogenetic tree takes these two states with equal prior probability. By definition, the Fitch method selects 1 with probability 1 if  $\{1\}$  is the state subset  $S_r(D)$  computed from  $D$  at the root (see the last section for the calculation of  $S_r(D)$ ); it selects 1 with probability  $\frac{1}{2}$  if  $S_r(D) = \{0, 1\}$ . By symmetry, Eq. (1) becomes

$$RA_F(T) = \sum_D \Pr_r[D|1] \left( \Pr[S_r(D) = \{1\}] + \frac{1}{2} \Pr[S_r(D) = \{0, 1\}] \right). \quad (2)$$

Let

$$\Pr_X[S|s] = \sum_{D'} \Pr_X[D'|s] \Pr[S_X(D') = S]$$

for a node  $X$  of a phylogenetic tree, a state  $s \in \{0, 1\}$ ,  $S \subseteq \{0, 1\}$ , where the sum is over all the combinations  $D'$  of the states of the leaves below  $X$ . Clearly,  $\Pr_X[S|s]$  is the probability that the Fitch method outputs state subset  $S$  at  $X$  in its first step given the true state of  $X$  is  $s$ . By symmetry,

$$\begin{aligned}\Pr_X[\{1\}|1] &= \Pr_X[\{0\}|0], \\ \Pr_X[\{0\}|1] &= \Pr_X[\{1\}|0], \\ \Pr_X[\{0, 1\}|1] &= \Pr_X[\{0, 1\}|0].\end{aligned}$$

For a node  $X$  and a state  $s \in \{0, 1\}$ , we further set

$$\alpha_X = \Pr_X[\{s\}|s], \quad \beta_X = \Pr_X[\{1-s\}|s].$$

Then

$$\Pr_X[\{0, 1\}|s] = 1 - \alpha_X - \beta_X.$$

Then Eq. (2) becomes

$$\begin{aligned}\text{RA}_F(T) &= \Pr_r[\{1\}|1] + \frac{1}{2}\Pr_r[\{0, 1\}|1] \\ &= \frac{1}{2} + \frac{1}{2}(\Pr_r[\{1\}|1] - \Pr_r[\{0\}|1]) \\ &= \frac{1}{2} + \frac{1}{2}(\alpha_r - \beta_r).\end{aligned}\tag{3}$$

Let  $Z$  be an internal node having  $X$  and  $Y$  as its children. Furthermore, we let the conservation probability on branches  $ZX$  and  $ZY$  be  $p_X$  and  $p_Y$ , respectively, i.e.,  $p_v = \Pr[Z \text{ and } v \text{ have the same state}]$  for  $v = X, Y$ . The subset  $S_Z$  computed at  $Z$  is  $\{1\}$  if and only if one of  $S_X$  and  $S_Y$  is  $\{1\}$  and the other is  $\{1\}$  or  $\{0, 1\}$ . Hence,

$$\begin{aligned}\alpha_Z &= (p_X\alpha_X + q_X\beta_X)(p_Y\alpha_Y + q_Y\beta_Y) \\ &\quad + (p_X\alpha_X + q_X\beta_X)(1 - \alpha_Y - \beta_Y) \\ &\quad + (1 - \alpha_X - \beta_X)(p_Y\alpha_Y + q_Y\beta_Y),\end{aligned}\tag{4}$$

where  $q_X = 1 - p_X$  and  $q_Y = 1 - p_Y$ . Similarly,

$$\begin{aligned}\beta_Z &= (q_X\alpha_X + p_X\beta_X)(q_Y\alpha_Y + p_Y\beta_Y) \\ &\quad + (q_X\alpha_X + p_X\beta_X)(1 - \alpha_Y - \beta_Y) \\ &\quad + (1 - \alpha_X - \beta_X)(q_Y\alpha_Y + p_Y\beta_Y).\end{aligned}\tag{5}$$

These two recurrence relations presented in Maddison (1995) lead to an efficient dynamic programming method for calculating  $\alpha_r$  and  $\beta_r$ . But these two relations are not simple enough for the theoretical study of the reconstruction accuracy. In the rest of this section, we shall establish two recurrence relations for the purpose of the theoretical analysis.

Let

$$C_Z = 1 - \alpha_Z - \beta_Z$$

and

$$D_Z = \alpha_Z - \beta_Z.$$

If  $Z$  is a leaf, we have that

$$C_Z = 0, \quad D_Z = 1. \quad (6)$$

Otherwise, we have the following recurrence relations.

**Theorem 3.1.** *Let  $Z$  be an internal node with children  $X$  and  $Y$ . Then*

$$C_Z = \frac{1}{2} \times [1 - C_X - C_Y + 3C_X C_Y - (2p_X - 1)(2p_Y - 1)D_X D_Y], \quad (7)$$

and

$$D_Z = \frac{1}{2}(2p_X - 1)(1 + C_Y)D_X + \frac{1}{2}(2p_Y - 1)(1 + C_X)D_Y. \quad (8)$$

*Proof:* These two recurrence equations can be verified by using Eqs. (4) and (5). The details can be found in the Appendix A.  $\square$

Although  $C$ 's and  $D$ 's are considered in earlier works such as Steel (1989) and Steel and Székely (2007), the above powerful recurrence equations are different from those presented on page 127 in Steel (1989) and Steel and Székely (2007). In Steel (1989), Steel presented a recurrence formula for  $C$ 's and  $D$ 's involving not only  $C$ 's and  $D$ 's but also  $\alpha$ 's and  $\beta$ 's. The same formula for  $D$  appears also in Steel and Székely (2007).

As the first application of above equations, we obtain the following fact. This result can be found in Steel (1989). Here, we give a simple argument.

**Corollary 3.1.** *For any phylogenetic tree  $T$  with root  $r$  in which the conservation probability on each branch is at least  $\frac{1}{2}$ ,  $\Pr[\{0, 1\}|s] = C_r \leq \frac{1}{2}$  for  $s = 0, 1$ .*

*Proof:* We prove the inequality by induction on  $m$ , the number of nodes of  $T$ . For  $m = 1$ , it follows from Eq. (6). Suppose  $C_r \leq \frac{1}{2}$  for any tree with less than  $m$  nodes. Now, consider a phylogenetic tree  $T$  of  $m$  nodes. Let the root  $r$  of  $T$  have children  $X$  and  $Y$ . Then by induction,  $0 \leq C_X, C_Y \leq \frac{1}{2}$ . Since  $p_X, p_Y \geq \frac{1}{2}$ , by Eq. (7),

$$\begin{aligned} C_r &= \frac{1}{2} \left[ \frac{2}{3} + 3 \left( C_X - \frac{1}{3} \right) \left( C_Y - \frac{1}{3} \right) - (2p_X - 1)(2p_Y - 1)D_X D_Y \right] \\ &\leq \frac{1}{2} \left[ \frac{2}{3} + 3 \left| C_X - \frac{1}{3} \right| \times \left| C_Y - \frac{1}{3} \right| \right] \end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{2} \left[ \frac{2}{3} + 3 \times \left( \frac{1}{3} \right)^2 \right] \\ &= \frac{1}{2}. \end{aligned}$$

Hence, this completes the proof.  $\square$

#### 4. Accuracy on complete binary trees

In this section, we shall focus on the reconstruction accuracy of the Fitch method on the complete binary trees. Let  $T_n$  be the complete binary tree of  $2^n$  leaves in which the conservation probability is  $p$  on each branch. Let  $r$  denote the root of  $T_n$  and  $C_n(p) = C_r$  and  $D_n(p) = D_r$  in  $T_n$ . Since the subtree rooted at each child of the root in  $T_n$  is the complete binary tree of  $2^{n-1}$  leaves, Eqs. (7) and (8) imply that, for  $n \geq 1$ ,

$$\begin{aligned} 2C_n(p) &= 1 - 2C_{n-1}(p) + 3C_{n-1}^2(p) - (2p - 1)^2 D_{n-1}^2(p), \\ D_n(p) &= (2p - 1)(1 + C_{n-1}(p))D_{n-1}(p), \end{aligned} \quad (9)$$

where  $0 \leq p \leq 1$ .

**Lemma 4.1.** For any  $n \geq 1$  and  $0 \leq p \leq 1$ ,

$$C_n(p) = C_n(1 - p), \quad |D_n(p)| = |D_n(1 - p)|.$$

*Proof:* We prove by induction on  $n$ . For  $n = 1$ , by Eq. (6),  $C_1(p) = 2p(1 - p)$ ,  $D_1(p) = 2p - 1$ , and hence the facts hold.

Assume that the facts are true for  $n - 1$ , i.e.  $C_{n-1}(p) = C_{n-1}(1 - p)$  and  $|D_{n-1}(p)| = |D_{n-1}(1 - p)|$ . Then

$$\begin{aligned} 2C_n(p) &= 1 - 2C_{n-1}(p) + 3C_{n-1}^2(p) - (2p - 1)^2 D_{n-1}^2(p) \\ &= 1 - 2C_{n-1}(1 - p) + 3C_{n-1}^2(1 - p) - (2(1 - p) - 1)^2 D_{n-1}^2(1 - p) \\ &= 2C_n(1 - p) \end{aligned}$$

and

$$\begin{aligned} |D_n(p)| &= |2p - 1| \cdot (1 + C_{n-1}(p)) \cdot |D_{n-1}(p)| \\ &= |2(1 - p) - 1| \cdot (1 + C_{n-1}(1 - p)) \cdot |D_{n-1}(1 - p)| \\ &= |D_n(1 - p)|. \end{aligned}$$

This completes the proof.  $\square$

By Lemma 4.1, we have

$$\lim_{n \rightarrow \infty} C_n(p) = \lim_{n \rightarrow \infty} C_n(1 - p)$$

and

$$\lim_{n \rightarrow \infty} |D_n(p)| = \lim_{n \rightarrow \infty} |D_n(1-p)|,$$

if they exist. Therefore, it suffices to assume that  $\frac{1}{2} \leq p \leq 1$ . For simplicity, we put  $C_i = C_i(p)$  and  $D_i = D_i(p)$  for all  $i$  and  $p$ . The two equalities in (9) become

$$2C_n = 1 - 2C_{n-1} + 3C_{n-1}^2 - (2p-1)^2 D_{n-1}^2, \quad (10)$$

$$D_n = (2p-1)(1 + C_{n-1})D_{n-1}. \quad (11)$$

**Lemma 4.2.** *For any  $n \geq 1$ ,*

$$0 \leq C_n \leq \frac{1}{2}, \quad 0 \leq D_n \leq 1.$$

*Proof:* By the assumption  $\frac{1}{2} \leq p \leq 1$ , the first inequality follows from Corollary 3.1. The second one is trivial since  $D_n$  is some probability.  $\square$

**Lemma 4.3.** *Let  $n \geq 1$ . If  $C_{n-1} \leq \frac{1}{3}$ , then  $C_n \leq \frac{1}{3}$ .*

*Proof:* We rewrite Eq. (10) as

$$2\left(\frac{1}{3} - C_n\right) + 3\left(\frac{1}{3} - C_{n-1}\right)^2 - (2p-1)^2 D_{n-1}^2 = 0. \quad (12)$$

This implies that

$$0 \leq 2\left(\frac{1}{3} - C_{n-1}\right) \leq (2p-1)^2 D_{n-2}^2,$$

and

$$4\left(\frac{1}{3} - C_{n-1}\right)^2 \leq (2p-1)^4 D_{n-2}^4.$$

By Lemma 4.2, we have that

$$\begin{aligned} 2C_n &= \frac{2}{3} + 3\left(\frac{1}{3} - C_{n-1}\right)^2 - (2p-1)^2 D_{n-1}^2 \\ &\leq \frac{2}{3} + \frac{3}{4}(2p-1)^4 D_{n-2}^4 - (2p-1)^2 [(2p-1)(1 + C_{n-2})D_{n-2}]^2 \\ &= \frac{2}{3} + (2p-1)^4 \left(\frac{3}{4} D_{n-2}^2 - (1 + C_{n-2})^2\right) D_{n-2}^2 \\ &\leq \frac{2}{3} + (2p-1)^4 \left(\frac{3}{4} - (1+0)^2\right) D_{n-2}^2 \\ &\leq \frac{2}{3}. \end{aligned}$$

This completes the proof.  $\square$

**Lemma 4.4.** *Let  $n \geq 1$ . If  $C_{n-1} \geq \frac{1}{3}$ , then  $C_n \leq C_{n-1}$ .*

*Proof:*

$$\begin{aligned} 2C_n &= 1 - 2C_{n-1} + 3C_{n-1}^2 - (2p-1)^2 D_{n-1}^2 \\ &= 2C_{n-1} + (1 - C_{n-1})(1 - 3C_{n-1}) - (2p-1)^2 D_{n-1}^2 \\ &\leq 2C_{n-1}. \end{aligned} \quad \square$$

**Theorem 4.1.** *Suppose  $\frac{1}{8} \leq p < \frac{7}{8}$ . Then*

$$\lim_{n \rightarrow \infty} C_n = \frac{1}{3}, \quad \lim_{n \rightarrow \infty} D_n = 0.$$

*Proof:* The proof is divided into two cases.

Case 1:  $C_n \geq \frac{1}{3}$  for all  $n$ . By Lemma 4.4,  $C_n$  is a decreasing positive sequence, and thus  $\lim_{n \rightarrow \infty} C_n$  exists and its value is at least  $\frac{1}{3}$ . The equation  $2C_n = 1 - 2C_{n-1} + 3C_{n-1}^2 - (2p-1)^2 D_{n-1}^2$  implies that  $\lim_{n \rightarrow \infty} D_n$  exists. Taking limit on all terms in Eq. (11) implies that  $\lim_{n \rightarrow \infty} D_n = 0$  since  $\lim_{n \rightarrow \infty} C_n \geq \frac{1}{3}$ . Again, taking limit on all terms in Eq. (10) gives that

$$2 \lim_{n \rightarrow \infty} C_n = 1 - 2 \lim_{n \rightarrow \infty} C_n + 3 \left( \lim_{n \rightarrow \infty} C_n \right)^2 - 0;$$

that is,  $\lim_{n \rightarrow \infty} C_n = \frac{1}{3}$  or 1. Since  $C_n$  is decreasing and  $C_1 = 2p(1-p) < \frac{1}{2}$ ,  $\lim_{n \rightarrow \infty} C_n \neq 1$ . Thus,  $\lim_{n \rightarrow \infty} C_n = \frac{1}{3}$ .

Case 2:  $C_N < \frac{1}{3}$  for some  $N$ . By Lemma 4.3,  $C_n \leq \frac{1}{3}$  for all  $n \geq N$ . Equation (11) implies that

$$D_n = (2p-1)(1+C_{n-1})D_{n-1} \leq \left( \frac{4}{3}(2p-1) \right)^{n-N} D_{N-1}$$

for any  $n \geq N$ . Since  $\frac{1}{2} \leq p < \frac{7}{8}$ ,  $\frac{4}{3}(2p-1) < 1$  and hence  $\lim_{n \rightarrow \infty} D_n = 0$ .

By Eq. (12),

$$2 \left( \frac{1}{3} - C_n \right) = (2p-1)^2 D_{n-1}^2 - 3 \left( C_{n-1} - \frac{1}{3} \right)^2$$

and hence

$$2 \left( \frac{1}{3} - C_n \right) \leq (2p-1)^2 D_{n-1}^2$$

for all  $n \geq N$ . Since

$$0 \leq 2 \left( \frac{1}{3} - C_n \right)$$

and

$$\lim_{n \rightarrow \infty} (2p - 1)^2 D_{n-1}^2 = (2p - 1)^2 \left( \lim_{n \rightarrow \infty} D_{n-1} \right)^2 = 0,$$

by the Sandwich theorem

$$\lim_{n \rightarrow \infty} 2 \left( \frac{1}{3} - C_n \right) = 0$$

and thus  $\lim_{n \rightarrow \infty} C_n = 1/3$ . This completes the proof.  $\square$

To prove the convergence of  $C_n$  and  $D_n$  for  $p \geq \frac{7}{8}$ , we set

$$c_n = \frac{2(1-p)}{2p-1} - C_n$$

and

$$d_n = D_n^2.$$

Then Eq. (12) implies that

$$\begin{aligned} & 2 \left( \frac{2(1-p)}{2p-1} - c_n \right) \\ &= \frac{2}{3} + 3 \left( \frac{1}{3} - \frac{2(1-p)}{2p-1} + c_{n-1} \right)^2 - (2p-1)^2 d_{n-1} \\ &= \frac{2}{3} + 3 \left( \frac{8p-7}{3(2p-1)} + c_{n-1} \right)^2 - (2p-1)^2 d_{n-1} \\ &= \frac{2}{3} + \frac{(8p-7)^2}{3(2p-1)^2} + \frac{2(8p-7)}{2p-1} c_{n-1} + 3c_{n-1}^2 - (2p-1)^2 d_{n-1}, \end{aligned}$$

or equivalently

$$2c_n = (2p-1)^2 d_{n-1} - \frac{2(8p-7)}{2p-1} c_{n-1} - 3c_{n-1}^2 - \frac{(8p-7)(4p-3)}{(2p-1)^2}. \quad (13)$$

Equation (11) implies that

$$d_n = (2p-1)^2 \left( \frac{1}{2p-1} - c_n \right)^2 d_{n-1} = [1 - (2p-1)c_n]^2 d_{n-1}. \quad (14)$$

**Lemma 4.5.** For any  $k \geq 2$  and  $p \geq \frac{7}{8}$ ,

- (1)  $c_k \geq 0$ .
- (2)  $d_{k+1} \leq d_k$ .
- (3)  $c_k \leq \frac{5(1-p)}{4(2p-1)}$ .

The proof of this lemma involves sophisticated inequalities, and hence is put in Appendix B.

**Theorem 4.2.** *Suppose  $\frac{7}{8} \leq p \leq 1$ . Then*

$$\lim_{n \rightarrow \infty} C_n = \frac{2(1-p)}{2p-1}$$

and

$$\lim_{n \rightarrow \infty} D_n^2 = \frac{(8p-7)(4p-3)}{(2p-1)^4}.$$

*Proof:* Since  $c_n \geq 0$  for all  $n$ , Eq. (13) implies

$$d_n \geq \frac{(8p-7)(4p-3)}{(2p-1)^4}$$

for all  $n$ . Since  $d_n = D_n^2$  is a decreasing sequence,  $\lim_{n \rightarrow \infty} d_n$  exists and is at least  $\frac{(8p-7)(4p-3)}{(2p-1)^4}$ , which is larger than 0 for  $p > \frac{7}{8}$ . Since  $0 \leq c_n \leq 1$ ,

$$0 \leq 1 - (2p-1)c_n \leq 1.$$

For  $p > \frac{7}{8}$ , Eq. (14) implies that

$$\lim_{n \rightarrow \infty} 1 - (2p-1)c_n = 1$$

and so

$$\lim_{n \rightarrow \infty} c_n = 0.$$

Hence,  $\lim_{n \rightarrow \infty} C_n = \frac{2(1-p)}{2p-1}$ .

For  $p = \frac{7}{8}$ , Eqs. (13) and (14) become

$$2c_n + 3c_{n-1}^2 = \frac{9}{16}d_{n-1}$$

and

$$d_n = \left(1 - \frac{3}{4}c_{n-1}\right)^2 d_{n-1}.$$

As a decreasing sequence,  $d_n$  has a non-negative limit. If  $\lim_{n \rightarrow \infty} d_n = 0$ , by the Sandwich theorem,  $\lim_{n \rightarrow \infty} c_n = 0$  from the fact that  $0 \leq 2c_n \leq \frac{9}{16}d_{n-1}$ . Therefore,

$$\lim_{n \rightarrow \infty} C_n = \frac{2(1-p)}{2p-1}$$

and

$$\lim_{n \rightarrow \infty} D_n^2 = \frac{(8p-7)(4p-3)}{(2p-1)^4}.$$

If  $\lim_{n \rightarrow \infty} d_n > 0$ , then,

$$d_n = d_{n-1} \left( 1 - \frac{3}{4} c_{n-1} \right)^2$$

implies that  $\lim_{n \rightarrow \infty} c_n = 0$  and hence  $\lim_{n \rightarrow \infty} d_n = 0$ , a contradiction.  $\square$

**Theorem 4.3.** Let  $T_n$  be the complete binary tree of  $2^n$  leaves in which the conservation probability is  $p$  along each branch. In the two-state Jukes–Cantor model,

- (a) (Steel, 1989) the accuracy of the Fitch method for reconstructing the root state in  $T_n$  converges as  $n$  goes to infinity to  $\frac{1}{2} + \frac{1}{2(2p-1)^2} \sqrt{(8p-7)(4p-3)}$  if  $p \in [\frac{7}{8}, 1]$  and  $\frac{1}{2}$  if  $p \in [\frac{1}{8}, \frac{7}{8}]$ ;
- (b) it diverges as  $n$  goes to infinity if  $p \in (0, \frac{1}{8})$ .

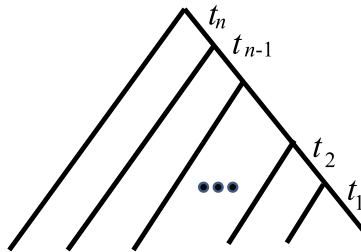
*Proof:* By Eq. (3) and the definition of  $D_n$ ,  $\text{RA}_F(T_n) = \frac{1}{2} + \frac{1}{2} D_n$ . Hence, part (a) follows from Theorems 4.1 and 4.2.

When  $0 < p < \frac{1}{8}$ ,  $D_n > 0$  for even integers  $n$  and  $D_n < 0$  for odd integers  $n$ . By Lemma 4.1 and Theorem 4.2,  $|D_n|$  converges to a positive number. Hence,  $D_n$  and  $\text{RA}_F(T_n)$  diverge. This proves part (b).  $\square$

## 5. The reconstruction accuracy on ultrametric comb trees

In the phylogenetic study, a rooted tree in which each internal node has at least one leaf child is called a ‘Hennigian comb tree’, or a caterpillar tree. In this section, we present a closed-form formula for the reconstruction accuracy of the Fitch method on ultrametric Hennigian comb trees.

Let  $H_n$  ( $n > 0$ ) denote the ultrametric Hennigian comb tree with  $n + 1$  levels as shown in Fig. 1, in which each branch between two internal nodes has positive length  $t_i$ . Under



**Fig. 1** The Hennigian comb tree  $H_n$  with  $n + 1$  leaves. Its height is  $h = \sum_{i=1}^n t_i$ .

the two-state Jukes–Cantor model, the conservation probability  $p(t)$  along a branch of length  $t$  is

$$p(t) = \frac{1}{2}(1 + e^{-\lambda t}),$$

where  $\lambda > 0$ , which is 2 times the substitution rate. Equivalently,

$$2p(t) - 1 = e^{-\lambda t}. \quad (15)$$

On the tree  $H_n$ , by (6), (7), (8), and (15),

$$C_n = \frac{1}{2}[1 - C_{n-1} - e^{-\lambda(t_n + \sum_1^n t_i)} D_{n-1}], \quad (16)$$

$$D_n = \frac{1}{2}e^{-\lambda t_n} D_{n-1} + \frac{1}{2}e^{-\lambda(\sum_1^n t_i)}(1 + C_{n-1}) \quad (17)$$

for  $n \geq 1$ , where we simply define  $C_0 = 0$  and  $D_0 = 1$ . We remark that  $H_1$  has two leaves and that  $C_0$  and  $D_0$  are not the corresponding values for the degenerated comb tree in which there is a root and a leaf. It is easy to see that

$$D_1 = e^{-\lambda t_1}. \quad (18)$$

For  $n \geq 2$ , we have that

$$\begin{aligned} D_n &= \frac{1}{2}e^{-\lambda t_n} D_{n-1} + \frac{1}{2}e^{-\lambda(\sum_1^n t_i)}(1 + C_{n-1}) \\ &= \frac{1}{4}e^{-\lambda(t_n + t_{n-1})} D_{n-2} + \frac{1}{4}e^{-\lambda(\sum_1^n t_i)}(1 + C_{n-2}) + \frac{1}{2}e^{-\lambda(\sum_1^n t_i)}(1 + C_{n-1}) \\ &= \frac{1}{4}e^{-\lambda(t_n + t_{n-1})} D_{n-2} + \frac{1}{2}e^{-\lambda(\sum_1^n t_i)}\left(\frac{3}{2} + C_{n-1} + \frac{1}{2}C_{n-2}\right) \\ &= \frac{1}{4}e^{-\lambda(t_n + t_{n-1})} D_{n-2} + \frac{1}{2}e^{-\lambda(\sum_1^n t_i)}\left(2 - \frac{1}{2}e^{-\lambda(t_{n-1} + \sum_1^{n-1} t_i)} D_{n-2}\right) \\ &= e^{-\lambda(\sum_1^n t_i)} + \frac{1}{4}e^{-\lambda(t_n + t_{n-1})}(1 - e^{-2\lambda(\sum_1^{n-1} t_i)})D_{n-2}, \end{aligned} \quad (19)$$

where the second last equality is obtained from Eq. (16). Using this recurrence formula iteratively, we obtain

$$D_n = \begin{cases} e^{-\lambda(\sum_1^n t_i)}[1 + \sum_{j=1}^{n/2} \frac{1}{4^j} \prod_{k=1}^j (1 - e^{-2\lambda(\sum_1^{n-2k+1} t_i)})], & n \text{ is even,} \\ e^{-\lambda(\sum_1^n t_i)}[1 + \sum_{j=1}^{(n-1)/2} \frac{1}{4^j} \prod_{k=1}^j (1 - e^{-2\lambda(\sum_1^{n-2k+1} t_i)})], & n \text{ is odd.} \end{cases} \quad (20)$$

Additionally, we can also establish the following bounds on  $C_n$  and  $D_n$ .

**Theorem 5.1.** *Assume  $\lambda, t_i > 0$ . The following facts hold.*

(i)  $C_n$  and  $D_n$  can be bounded as

$$\left| \frac{1}{3} - C_n \right| \leq \frac{4}{3} \frac{c^{n+4}(1-c^n)}{1-c} + \frac{1}{3 \times 2^{n-1}}, \quad (21)$$

where  $c = \max\{\frac{1}{2}, e^{-\lambda t_i} \mid 1 \leq i \leq n\} < 1$ , and

$$\left( \frac{4}{3} - \frac{1}{3 \times 2^{n-1}} \right) \left( 1 - \frac{1}{3} e^{-2\lambda t_1} \right) e^{-\lambda(\sum_1^n t_i)} \leq D_n \leq \left( \frac{4}{3} - \frac{1}{3 \times 2^n} \right) e^{-\lambda(\sum_1^n t_i)} \quad (22)$$

on the Hennigian comb tree  $H_n$  for  $n \geq 2$ .

(ii) Let  $h = \sum_{i=1}^n t_i$ . The reconstruction accuracy of the Fitch method on  $H_n$  goes to  $\frac{1}{2}$  when the height  $h$  of  $H_n$  goes to infinity.

*Proof:* (i) We prove the inequalities by induction. By Eq. (19),

$$D_2 = e^{-\lambda(t_1+t_2)} \left( \frac{5}{4} - \frac{1}{4} e^{-2\lambda t_1} \right) \leq \frac{5}{4} e^{-\lambda(t_1+t_2)},$$

$$D_3 = e^{-\lambda(t_1+t_2+t_3)} \left( \frac{5}{4} - \frac{1}{4} e^{-2\lambda(t_1+t_2)} \right) \leq \frac{5}{4} e^{-\lambda(t_1+t_2+t_3)}.$$

Thus, the inequality (22) is true for  $n = 2, 3$  because  $\frac{4}{3} - \frac{1}{3 \times 2^2} = \frac{5}{4}$ . Assume it is true for  $n \leq k$ . Setting  $n = k - 1$ , we have that

$$\left( \frac{4}{3} - \frac{1}{3 \times 2^{k-2}} \right) \left( 1 - \frac{1}{3} e^{-2\lambda t_1} \right) e^{-\lambda(\sum_1^{k-1} t_i)} \leq D_{k-1} \leq \left( \frac{4}{3} - \frac{1}{3 \times 2^{k-1}} \right) e^{-\lambda(\sum_1^{k-1} t_i)}.$$

By Eq. (19), we obtain that

$$\begin{aligned} D_{k+1} &\leq e^{-\lambda(\sum_1^{k+1} t_i)} + \frac{1}{4} e^{-\lambda(t_{k+1}+t_k)} D_{k-1} \\ &\leq e^{-\lambda(\sum_1^{k+1} t_i)} + \frac{1}{4} e^{-\lambda(t_{k+1}+t_k)} \left[ e^{-\lambda(\sum_1^{k-1} t_i)} \left( \frac{4}{3} - \frac{1}{3 \times 2^{k-1}} \right) \right] \\ &= e^{-\lambda(\sum_1^{k+1} t_i)} \left( \frac{4}{3} - \frac{1}{3 \times 2^{k+1}} \right). \end{aligned}$$

On the other hand, by Eq. (19) and the fact that  $t_i > 0$  for each  $i$ , we obtain that

$$\begin{aligned} D_{k+1} &\geq e^{-\lambda(\sum_1^{k+1} t_i)} + \frac{1}{4} e^{-\lambda(t_{k+1}+t_k)} (1 - e^{-2\lambda t_1}) D_{k-1} \\ &\geq e^{-\lambda(\sum_1^{k+1} t_i)} + \frac{1}{4} e^{-\lambda(t_{k+1}+t_k)} (1 - e^{-2\lambda t_1}) \end{aligned}$$

$$\begin{aligned}
& \times \left( \frac{4}{3} - \frac{1}{3 \times 2^{k-2}} \right) \left( 1 - \frac{1}{3} e^{-2\lambda t_1} \right) e^{-\lambda(\sum_1^{k-1} t_i)} \\
& \geq \left( \frac{4}{3} - \frac{1}{3 \times 2^k} \right) \left( 1 - \frac{1}{3} e^{-2\lambda t_1} \right) e^{-\lambda(\sum_1^{k+1} t_i)}.
\end{aligned}$$

This proves the inequality (22).

Now we estimate  $C_n$  as follows. By Eq. (16),

$$\begin{aligned}
\frac{1}{3} - C_n &= \frac{1}{3} - \frac{1}{2} \left[ 1 - C_{n-1} - e^{-\lambda(t_n + \sum_1^n t_i)} D_{n-1} \right] \\
&= \frac{1}{2} e^{-\lambda(t_n + \sum_1^n t_i)} D_{n-1} - \frac{1}{2} \left( \frac{1}{3} - C_{n-1} \right).
\end{aligned}$$

Hence, we have

$$\left| \frac{1}{3} - C_n \right| \leq \frac{1}{2} e^{-\lambda(t_n + \sum_1^n t_i)} D_{n-1} + \frac{1}{2} \left| \frac{1}{3} - C_{n-1} \right|.$$

Applying this inequality iteratively, we

$$\begin{aligned}
\left| \frac{1}{3} - C_n \right| &\leq \sum_{k=1}^n \frac{1}{2^k} e^{-\lambda(t_{n-k+1} + \sum_1^{n-k+1} t_i)} D_{n-k} + \frac{1}{2^{n-1}} \left| \frac{1}{3} - C_0 \right| \\
&\leq \frac{4}{3} \sum_{k=1}^n \frac{1}{2^k} e^{-2\lambda \sum_1^{n-k+1} t_i} + \frac{1}{2^{n-1}} \left| \frac{1}{3} - 0 \right|,
\end{aligned}$$

where the second inequality follows from (22). By the definition of  $c$ , we have that  $\frac{1}{2} \leq c$  and  $e^{-\lambda t_i} \leq c$ . Hence,

$$\left| \frac{1}{3} - C_n \right| \leq \frac{4}{3} \sum_{k=0}^{n-1} c^{2n+2-k} + \frac{1}{3 \times 2^{n-1}} = \frac{4}{3} \frac{c^{n+4}(1-c^n)}{1-c} + \frac{1}{3 \times 2^{n-1}}.$$

Hence, we proves inequality (21).

(ii) Since  $\lambda > 0$ ,  $\lim_{h \rightarrow \infty} e^{-2\lambda h} = 0$ . Moreover,  $n$  goes to infinity when  $h$  goes to infinity. Hence, inequality (22) implies that  $\lim_{h \rightarrow \infty} D_n = 0$  and the reconstruction accuracy of the Fitch method goes to  $\frac{1}{2}$  by Eq. (3).  $\square$

When a set of species originated in a burst of speciation events, their phylogenetic tree has a star-like topology. It has been known that a star-like topology allows more accurate reconstruction than others. A star-like topology can approximately be considered as one obtained from an ultrametric comb phylogenetic tree by contracting the branches between internal nodes. By Theorem 21(i),

$$D_n \sim \frac{4}{3} e^{-\lambda h}$$

by letting  $t_i$  go to zero for each  $i \geq 2$ . Hence, we conjecture that the accuracy of reconstructing the root state on a star-like phylogenetic tree is about  $\frac{1}{2}(1 + \frac{4}{3}e^{-\lambda h})$  when the Fitch method is used if the height  $h$  of the tree is large.

## 6. The reconstruction accuracy on arbitrary ultrametric trees

We now consider the accuracy of reconstructing the root state in arbitrary ultrametric phylogenetic trees. In an ultrametric phylogenetic tree  $T$ , a branch  $xy$  has a length  $t_{xy}$ , but all the leaves have the same distance from the root. For an internal node  $u$  of  $T$ , the distance between it and any of its leaf descendants is defined as its depth, denoted by  $d(u)$ .

**Lemma 6.1.** *Let  $T$  be an ultrametric phylogenetic tree and  $u$  an internal node. Under the 2-state Jukes–Cantor model, for any path  $P(x, y)$  from an internal node  $x$  to its leaf descendant  $y$ ,*

$$\prod_{uv \in P(x,y)} (2p_{uv} - 1) = e^{-\lambda d(x)}. \quad (23)$$

*Proof:* It follows from that  $2p_{uv} - 1 = e^{-\lambda t_{uv}}$  for each branch  $uv$  and that  $d(x) = \sum_{uv \in P(x,y)} t_{uv}$ .  $\square$

Let  $T$  be an ultrametric tree that has three or more leaves. For any internal node  $w$  with children  $w_1$  and  $w_2$ , by Eq. (8) and Lemma 6.1, we have that

$$D_w \geq \frac{1}{2}(2p_{ww_1} - 1)D_{w_1} + \frac{1}{2}(2p_{ww_2} - 1)D_{w_2} \quad (24)$$

because  $C_{w_1}, C_{w_2} \geq 0$ . By induction, we can easily show the following fact from Eq. (24).

**Lemma 6.2.**

$$D_w \geq \prod_{(u,v) \in P(w,l)} (2p_{uv} - 1) = e^{-\lambda d(w)},$$

where  $l$  is a leaf below  $w$ .

By (3), the above lemma implies that the accuracy of reconstructing the root state from all the leaf states is not less than from a single leaf. Such a fact was established by Fischer and Thatte (2009). It can be strengthened as follows.

**Theorem 6.1.** *Let  $T$  be an ultrametric tree having three or more leaves and let  $x$  be a child of its root  $r$ . If  $x$  has two children, then*

$$D_r \geq e^{-\lambda d(r)} \left[ 1 + \frac{1}{4}(1 - e^{-2\lambda d(x)}) \right]. \quad (25)$$

*Proof:* By Lemma 6.2,  $D_y \geq e^{-\lambda d(y)}$ , where  $y$  is the other child of  $r$ , different from  $x$ . Since  $C_y \geq 0$ , by Eq. (8), we have that

$$D_r = \frac{1}{2}(2p_{rx} - 1)(1 + C_y)D_x + \frac{1}{2}(2p_{ry} - 1)(1 + C_x)D_y$$

$$\geq \frac{1}{2}(2p_{rx} - 1)D_x + \frac{1}{2}e^{-\lambda d(r)}(1 + C_x). \quad (26)$$

Let  $u$  and  $v$  be the children of  $x$ . By Lemma 6.2,  $D_u \geq e^{-\lambda d(u)}$  and  $D_v \geq e^{-\lambda d(v)}$ . Let

$$D_u = e^{-\lambda d(u)}(1 + \Delta(u)), \quad D_v = e^{-\lambda d(v)}(1 + \Delta(v)),$$

where  $\Delta(u), \Delta(v) \geq 0$ . We then have

$$\begin{aligned} D_x &= \frac{1}{2}(2p_{xu} - 1)(1 + C_v)D_u + \frac{1}{2}(2p_{xv} - 1)(1 + C_u)D_v \\ &= e^{-\lambda d(x)} \left\{ \frac{1}{2}[1 + C_v + \Delta(u) + C_v\Delta(u)] + \frac{1}{2}[1 + C_u + \Delta(v) + C_u\Delta(v)] \right\} \\ &\geq e^{-\lambda d(x)} \left\{ 1 + \frac{1}{2}[C_u + C_v + \Delta(u) + \Delta(v)] \right\}. \end{aligned} \quad (27)$$

Combining inequalities (26) and (27) gives that

$$D_r \geq e^{-\lambda d(r)} \left\{ 1 + \frac{1}{2}C_x + \frac{1}{4}[C_u + C_v + \Delta(u) + \Delta(v)] \right\}.$$

By Eq. (7),

$$\begin{aligned} C_x &= \frac{1}{2}[1 - C_u - C_v + 3C_uC_v - (2p_{xu} - 1)(2p_{xv} - 1)D_uD_v] \\ &\geq \frac{1}{2}[1 - C_u - C_v - (2p_{xu} - 1)(2p_{xv} - 1)D_uD_v] \\ &= \frac{1}{2}[1 - C_u - C_v - e^{-2\lambda d(x)}[1 + \Delta(u)][1 + \Delta(v)]]. \end{aligned}$$

We further have that

$$D_r \geq \frac{1}{4}e^{-\lambda d(r)} \{5 + \Delta(u) + \Delta(v) - e^{-2\lambda d(x)}[1 + \Delta(u)][1 + \Delta(v)]\}.$$

Since  $d(x) > d(v)$ ,

$$[1 + \Delta(v)]e^{-2\lambda d(x)} \leq [1 + \Delta(v)]e^{-\lambda d(v)} = D_v \leq 1.$$

Therefore, we obtain that

$$\begin{aligned} &\Delta(u) + \Delta(v) - e^{-2\lambda d(x)}[1 + \Delta(u)][1 + \Delta(v)] \\ &\geq \Delta(u) - e^{-2\lambda d(x)}[1 + \Delta(u)] \\ &\geq -e^{-2\lambda d(x)} \end{aligned}$$

and

$$D_r \geq \frac{1}{4}e^{-\lambda d(r)}(5 - e^{-2\lambda d(x)}) = e^{-\lambda d(r)} \left[ 1 + \frac{1}{4}(1 - e^{-2\lambda d(x)}) \right]. \quad \square$$

It is known that there exists an ultrametric tree in which the root state can be reconstructed more accurately from the states of a subset of four leaves than from all the nine leaf states (Li et al., 2009). Let  $l_1, l_2$ , and  $l_3$  be three leaves in  $T$ . Assume that the least common ancestor (lca)  $y$  of  $l_2$  and  $l_3$  is below the lca of  $l_1, l_2$ , and  $l_3$ . Let  $y$  be below  $x$  and have depth  $d(y)$ .

**Proposition 6.1.** *The accuracy of reconstructing the state at the root  $r$  from the states of  $l_1, l_2$ , and  $l_3$  is*

$$\frac{1}{2} + \frac{1}{2}e^{-\lambda d(r)} \left[ 1 + \frac{1}{4}(1 - e^{-2\lambda d(y)}) \right].$$

*Proof.* There are two cases as shown in Fig. 2.

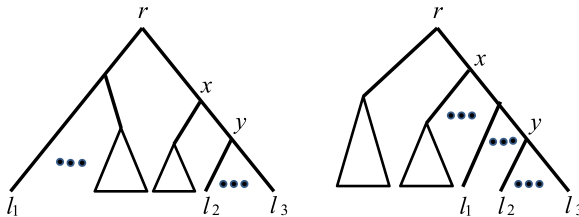
Case 1. The lca of  $l_1$  and  $t$  is the root. In this case, the fact follows from (19) because  $D_0 = 1$  by definition.

Case 2. The lca of  $l_1$  and  $y$  is below  $x$ . In this case, we let the lca of  $l_1$  and  $y$  be  $z$ . Then by the case 1 discussed above, the accuracy of reconstructing the state at  $z$  is

$$A_z = \frac{1}{2} + \frac{1}{2}e^{-\lambda d(z)} \left[ 1 + \frac{1}{4}(1 - e^{-2\lambda d(y)}) \right].$$

Since our model is symmetric two-state model and that the conservation probability on the path from  $r$  to  $z$  is  $p_{rz} = \frac{1}{2}(1 + e^{-\lambda(d(r)-d(z))})$ , the accuracy of reconstructing the state at the root  $r$  from  $l_i$ 's is

$$\begin{aligned} & p_{rz}A_z + (1 - p_{rz})(1 - A_z) \\ &= 1 - p_{rz} + (2p_{rz} - 1)A_z \\ &= 1 - p_{rz} + \frac{1}{2}e^{-\lambda(d(r)-d(z))} + \frac{1}{2}e^{-\lambda d(r)} \left[ 1 + \frac{1}{4}(1 - e^{-2\lambda d(y)}) \right] \end{aligned}$$



**Fig. 2** In the left tree, the least common ancestor (lca)  $y$  of  $l_2$  and  $l_3$  is below  $x$ , but the lca of  $y$  and  $l_1$  is the root of the tree. In the right tree, both the lca's are below  $x$ .

$$= \frac{1}{2} + e^{-\lambda d(r)} \left[ 1 + \frac{1}{4} (1 - e^{-2\lambda d(y)}) \right].$$

This proves the result.  $\square$

Under the assumption that  $l_2$  and  $l_3$  are below  $x$ , the above proposition suggests that the accuracy of reconstructing the root state from  $l_1, l_2$ , and  $l_3$  is not larger than  $\frac{1}{2} + e^{-\lambda d(r)} [1 + \frac{1}{4} (1 - e^{-2\lambda d(x)})]$ . Thus, by Theorem 6.1, the reconstruction of the root state from all the leaf states is at least as accurate as from the states of any three leaves in an ultrametric tree.

## 7. Conclusion

In this paper, we present two recurrence relations for studying the accuracy of the Fitch method for reconstructing the root state on phylogenetic trees. Using these two relations, we have addressed three theoretical problems. First, we prove that the accuracy of the Fitch method on a complete phylogenetic tree with equal branch length converges when the number of taxa goes to infinite if and only if the conservation probability on each branch is larger than  $\frac{1}{8}$ . This fills a technical gap left in the study in Steel (1989) (also see Steel and Charleston, 1995). Second, we give a closed-form formula for computing the accuracy of the Fitch method on an ultrametric Hennigian comb tree. Lastly, we present a lower bound on the accuracy of the Fitch method on an arbitrary phylogenetic tree, improving the bound given in Fischer and Thatte (2009).

The maximum likelihood (ML) method is another important method for ancestral state reconstruction. It assigns a state at the root based on the likelihood of the observed states of taxa given the state at the root. The convergence of the reconstruction accuracy of the ML method on infinite trees has extensively been studied in information theory and statistical physics in the past three decades (see Bleher et al., 1995; Evans et al., 2000; Mossel, 1998 for example). An interesting and important question for future research is how to generalize the convergence result on complete phylogenetic trees in Section 4 to arbitrary large trees.

## Acknowledgement

L.X. Zhang thanks Mike Steel for helpful comments on the first draft of this manuscript and Yun Cui for discussions.

## Appendix A: Proof of Theorem 3.1

We first have that

$$\begin{aligned} & (p_X \alpha_X + q_X \beta_X)(1 - \alpha_Y - \beta_Y) - (q_X \alpha_X + p_X \beta_X)(1 - \alpha_Y - \beta_Y) \\ &= (2p_X - 1)C_Y D_X, \end{aligned} \tag{A.1}$$

$$\begin{aligned}
 & (1 - \alpha_X - \beta_X)(p_Y\alpha_Y + q_Y\beta_Y) - (1 - \alpha_X - \beta_X)(q_Y\alpha_Y + p_Y\beta_Y) \\
 & = (2p_Y - 1)C_X D_Y,
 \end{aligned} \tag{A.2}$$

and

$$\begin{aligned}
 & (p_X\alpha_X + q_X\beta_X)(p_Y\alpha_Y + q_Y\beta_Y) - (q_X\alpha_X + p_X\beta_X)(q_Y\alpha_Y + p_Y\beta_Y) \\
 & = (p_X + p_Y - 1)(\alpha_X\alpha_Y - \beta_X\beta_Y) + (p_Y - p_X)(\beta_X\alpha_Y - \alpha_X\beta_Y).
 \end{aligned} \tag{A.3}$$

Since

$$\alpha_X\alpha_Y - \beta_X\beta_Y = (\alpha_X - \beta_X)\alpha_Y + \beta_X(\alpha_Y - \beta_Y)$$

and

$$\beta_X\alpha_Y - \alpha_X\beta_Y = \alpha_X(\alpha_Y - \beta_Y) - (\alpha_X - \beta_X)\alpha_Y,$$

combining the equalities (A.1)–(A.3) given above leads to

$$D_Z = (2p_X - 1)(1 - \beta_Y)D_X + (2p_Y - 1)(1 - \alpha_X)D_Y.$$

By symmetry,

$$D_Z = (2p_X - 1)(1 - \alpha_Y)D_X + (2p_Y - 1)(1 - \beta_X)D_Y.$$

Therefore,

$$\begin{aligned}
 D_Z &= \frac{1}{2}(2p_X - 1)(2 - \alpha_Y - \beta_Y)D_X + \frac{1}{2}(2p_Y - 1)(2 - \alpha_X - \beta_X)D_Y \\
 &= \frac{1}{2}(2p_X - 1)(1 + C_Y)D_X + \frac{1}{2}(2p_Y - 1)(1 + C_X)D_Y.
 \end{aligned} \tag{A.4}$$

Moreover, we also have that

$$\begin{aligned}
 & \alpha_Z + \beta_Z \\
 & = (p_X p_Y + q_X q_Y)(\alpha_X \alpha_Y + \beta_X \beta_Y) + (q_X p_Y + p_X q_Y)(\beta_X \alpha_Y + \alpha_X \beta_Y) \\
 & \quad + (\alpha_X + \beta_X)(1 - \alpha_Y - \beta_Y) + (1 - \alpha_X - \beta_X)(\alpha_Y + \beta_Y).
 \end{aligned}$$

Since

$$\alpha_X\alpha_Y + \beta_X\beta_Y = \frac{1}{2}((1 - C_X)(1 - C_Y) + D_X D_Y)$$

and

$$\beta_X\alpha_Y + \alpha_X\beta_Y = \frac{1}{2}((1 - C_X)(1 - C_Y) - D_X D_Y),$$

we obtain that

$$1 - C_Z = \frac{1}{2}[1 + C_X + C_Y - 3C_X C_Y + (2p_X - 1)(2p_Y - 1)D_X D_Y],$$

or equivalently

$$C_Z = \frac{1}{2} [1 - C_X - C_Y + 3C_X C_Y - (2p_X - 1)(2p_Y - 1)D_X D_Y]. \quad (\text{A.5})$$

## Appendix B: Proof of Lemma 4.5

We prove it by induction on  $k$ . The three inequalities clearly hold for  $k = 2, 3$ . Assume they hold for  $k \leq n - 1$ . We now prove they hold for  $k = n$ .

(1) By induction,  $0 \leq c_{n-2}, c_{n-1} \leq \frac{5(1-p)}{4(2p-1)}$ . Hence,

$$\begin{aligned} & [1 - (2p - 1)c_{n-2}]^2 - \frac{8p - 7}{2p - 1} - \frac{3}{2}c_{n-1} \\ &= \frac{6(1-p)}{2p-1} - 2(2p-1)c_{n-2} - \frac{3}{2}c_{n-1} + (2p-1)^2 c_{n-2}^2 \\ &\geq \frac{6(1-p)}{2p-1} - \frac{8p-1}{2} \times \frac{5(1-p)}{4(2p-1)} + 0 \\ &= \frac{1-p}{2p-1} \times \frac{53-40p}{8} \\ &\geq 0. \end{aligned} \quad (\text{B.1})$$

Setting  $\Delta = \frac{(8p-7)(4p-3)}{(2p-1)^2}$ , we have

$$\begin{aligned} 2c_n &= (2p-1)^2 d_{n-1} - \frac{2(8p-7)}{2p-1} c_{n-1} - 3c_{n-1}^2 - \Delta \\ &= (2p-1)^2 d_{n-1} - 2c_{n-1} \left( \frac{8p-7}{2p-1} + \frac{3}{2}c_{n-1} \right) - \Delta. \end{aligned}$$

By using recurrence Eqs. (13) and (14), we obtain that

$$\begin{aligned} 2c_n &= (2p-1)^2 (1 - (2p-1)c_{n-2})^2 d_{n-2} - \left[ \frac{8p-7}{2p-1} + \frac{3}{2}c_{n-1} \right] \\ &\quad \times \left[ (2p-1)^2 d_{n-2} - \frac{2(8p-7)}{2p-1} c_{n-2} - 3c_{n-2}^2 - \Delta \right] - \Delta \\ &= (2p-1)^2 \left[ (1 - (2p-1)c_{n-2})^2 - \frac{8p-7}{2p-1} - \frac{3}{2}c_{n-1} \right] d_{n-2} \\ &\quad + \left[ \frac{8p-7}{2p-1} + \frac{3}{2}c_{n-1} \right] \left[ \frac{2(8p-7)}{2p-1} c_{n-2} + 3c_{n-2}^2 + \Delta \right] - \Delta. \end{aligned}$$

Since  $c_{n-1} \geq 0$ , Eq. (13) implies that

$$(2p-1)^2 d_{n-2} \geq \frac{2(8p-7)}{2p-1} c_{n-2} + 3c_{n-2}^2 + \Delta.$$

This inequality and inequality (B.1) implies that

$$\begin{aligned}
 2c_n &\geq \left[ (1 - (2p - 1)c_{n-2})^2 - \frac{8p - 7}{2p - 1} - \frac{3}{2}c_{n-1} \right] \\
 &\quad \times \left[ \frac{2(8p - 7)}{2p - 1}c_{n-2} + 3c_{n-2}^2 + \Delta \right] \\
 &\quad + \left[ \frac{8p - 7}{2p - 1} + \frac{3}{2}c_{n-1} \right] \left[ \frac{2(8p - 7)}{2p - 1}c_{n-2} + 3c_{n-2}^2 + \Delta \right] - \Delta \\
 &= \frac{8(8p - 7)(1 - p)}{2p - 1}c_{n-2} + [3 + (8p - 7)(4p - 7)]c_{n-2}^2 \\
 &\quad + 4(2p - 1)(4p - 5)c_{n-2}^3 + 3(2p - 1)^2c_{n-2}^4.
 \end{aligned}$$

By assumption,  $c_{n-2} \leq \frac{5(1-p)}{4(2p-1)}$  and  $4p - 5 < -1$ . Replacing  $c_{n-2}^3$  with  $\frac{5(1-p)}{4(2p-1)}c_{n-2}^2$  in the right-hand side of the last inequality, we have that

$$\begin{aligned}
 2c_n &\geq \frac{8(8p - 7)(1 - p)}{2p - 1}c_{n-2} + [3 + (8p - 7)(4p - 7)]c_{n-2}^2 \\
 &\quad + 5(1 - p)(4p - 5)c_{n-2}^2 + 3(2p - 1)^2c_{n-2}^4 \\
 &= \frac{8(8p - 7)(1 - p)}{2p - 1}c_{n-2} + 3(1 - p)(9 - 4p)c_{n-2}^2 \\
 &\quad + 3(2p - 1)^2c_{n-2}^4 \\
 &\geq 0.
 \end{aligned}$$

(2) We have proved that  $c_n \geq 0$ . Therefore,  $d_{n+1} = [1 - (2p - 1)c_n]^2 d_n \leq d_n$ .

(3) Since  $d_k$  decreases for  $k \leq n$ ,

$$d_n \leq d_2 = D_2^2 < (2p - 1)^2. \quad (\text{B.2})$$

Let  $q = 1 - p$ . Note that  $p \geq \frac{7}{8}$  and  $q \leq \frac{1}{8}$ . Therefore, we have that

$$\frac{1}{1 - 2q} \leq \frac{4}{3}$$

and

$$16q(1 - 5q) \leq 16 \times \frac{1}{10} \times \left( 1 - 5 \times \frac{1}{10} \right) = \frac{4}{5}.$$

Recalling that  $c_{n-1} \geq 0$ , by inequality (B.2), we have that

$$\begin{aligned}
 c_n &= \frac{1}{2} \left[ (2p - 1)^2 d_{n-1} - \frac{2(8p - 7)}{2p - 1}c_{n-1} - 3c_{n-1}^2 - \frac{(8p - 7)(4p - 3)}{(2p - 1)^2} \right] \\
 &\leq \frac{1}{2} \left[ (2p - 1)^2 d_{n-1} - \frac{(8p - 7)(4p - 3)}{(2p - 1)^2} \right]
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2(2p-1)^2} [(2p-1)^4 d_{n-1} - (8p-7)(4p-3)] \\
&\leq \frac{1}{2(2p-1)^2} [(2p-1)^6 - (8p-7)(4p-3)] \\
&= \frac{1}{2(2p-1)^2} [(1-2q)^6 - (1-8q)(1-4q)] \\
&= \frac{q}{(2p-1)^2} [2q(7-40q+60q^2-48q^3+16q^4)] \\
&\leq \frac{q}{(2p-1)^2} [2q(7-40q+60q^2+16q^4)] \\
&\leq \frac{q}{(2p-1)^2} \left[ 2q \left( 7-40q + \frac{60}{64} + \frac{1}{256} \right) \right] \\
&\leq \frac{q}{(2p-1)^2} [2q(8-40q)] \\
&= \frac{q}{2p-1} \frac{16q(1-5q)}{1-2q} \\
&\leq \frac{4q}{5(2p-1)} \frac{1}{1-2q}.
\end{aligned}$$

Since  $q \leq \frac{1}{8}$  and  $\frac{1}{1-2q} \leq \frac{4}{3}$ ,  $c_n \leq \frac{16q}{15(2p-1)} \leq \frac{5q}{4(2p-1)}$ . This concludes the proof.

## References

- Baba, M.L., Goodman, M., Berger-Cohn, J., Demaille, J.G., Matsuda, G., 1984. The early adaptive evolution of calmodulin. *Mol. Biol. Evol.* 1, 442–455.
- Bleher, P.M., Ruiz, J., Zagrebnoy, V.A., 1995. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Stat. Phys.* 79, 473–482.
- Evans, W., Kenyon, C., Peres, Y., 2000. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* 10, 410–433.
- Fischer, M., Thatte, B.D., 2009. Maximum parsimony on subsets of taxa. *J. Theor. Biol.* 260, 290–293.
- Fitch, W.M., 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20, 406–416.
- Li, G.L., Steel, M., Zhang, L.X., 2008. More taxa are not necessarily better for the reconstruction of ancestral character states. *Syst. Biol.* 57, 647–653.
- Li, G.L., Ma, J., Zhang, L.X., 2009. Greedy selection of species for ancestral state reconstruction on phylogenies: Elimination is better than insertion. *PLOS One*, in press.
- Liberles, D.A. (Ed.), 2007. *Ancestral Sequence Reconstruction*. Oxford University Press, New York.
- Maddison, W.P., 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Syst. Biol.* 44, 474–481.
- Mossel, E., 1998. Recursive reconstruction on periodic trees. *Random Struct. Algorithms* 16, 252–260.
- Pauling, L., Zuckerkandl, E., 1963. Chemical paleogenetics: molecular restoration studies of extinct forms of lives. *Acta Chem. Scand.* 17, S9–S16.
- Pachter, L., 2007. An introduction to reconstructing ancestral genomes. In: *Proceedings of Symposia in Applied Mathematics*, AMS Short Course Subseries, vol. 64, pp. 1–20.
- Salisbury, B.A., Kim, J., 2001. Ancestral state estimation and taxon sampling density. *Syst. Biol.* 50, 557–564.
- Steel, M., 1989. Distribution in bicoloured evolutionary trees. PhD thesis, Massey University, New Zealand.

- Steel, M., Charleston, M., 1995. Five surprising properties of parsimoniously colored tree. *Bull. Math. Biol.* 57, 367–375.
- Steel, M.A., Székely, L.A., 2007. Teasing apart two trees. *Comb. Probab. Comput.* 16, 903–922.
- Thornton, J.W., 2004. Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nat. Rev. Genet.* 5, 366–375.
- Zhang, J., Nei, M., 1997. Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods. *J. Mol. Evol.* 44, 139–146.