

Efficient Estimation of the Accuracy of the Maximum Likelihood Method for Ancestral State Reconstruction

Bin Ma¹, Louxin Zhang²

¹ David R. Cheriton School of Computer Science, University of Waterloo

² Department of Mathematics, National University of Singapore

Corresponding author:

LX Zhang, Email: matzlx@nus.edu.sg; Phone: +65-6516-6579; Fax: +65-6779-5452

Abstract

The marginal maximum likelihood method is a widely-used method for ancestral state reconstruction. Given an evolution model (a phylogeny tree and the edge mutation rates) and the extant states (states on leaves), the method computes efficiently the most likely ancestral state on the root. However, when the extant states are randomly generated by using the evolutionary model, it is unknown how to efficiently calculate the expected reconstruction accuracy of the marginal maximum likelihood method. In this paper, a fully polynomial time approximation scheme (FPTAS) is presented for the calculation.

Keywords Ancestral state reconstruction, reconstruction accuracy, maximum likelihood method, polynomial time approximation

1 Introduction

Ancestral state reconstruction is an important approach to investigating the evolution of molecular or morphological features of living organisms (Jermann et al. 1995; Gaucher et al. 2003; Thornton et al. 2003; Blanchette et al. 2004). Sophisticated methods have been developed for reconstruction of ancestral state for DNA or protein sequences (Yang, Kumar and Nei 1995; Koshi and Goldstein 1996), multistate discrete data (Schultz et al. 1996; Mooers and Schluter 1999; Pagel 1999) and continuous data (Martins 1999). These different methods have also been assessed analytically (Lucena and Haussler 2005; Maddison 1995, Evens et al. 2000) and empirically (Schultz et al. 1996; Zhang and Nei 1997; Salisbury and Kim 2001; Blanchette et al. 2004; Mooers 2004; Williams et al. 2006; Hall 2006). The reader is referred to the books of Felsenstein (2004) and Liberles (2007) for further information on ancestral state reconstruction.

We are interested in the problem of reconstructing ancestral state from extant states of a character, with the assumption that the true phylogeny that relates the extant taxa to the target ancestor, together with an evolutionary model of the character, is given. In evolutionary biology, the states could represent particular traits or morphological features of organisms, In ancestral sequence reconstruction, states are nucleotides. Most reconstruction methods assume that the character under consideration evolves by a Markov process, starting with a state, which will be reconstructed, at the root and proceeding to the leaves. Hence, an evolutionary model of a character specifies the length of each branch or, equivalently, the probability that a state c evolves to a state d on a branch from node u to node v as the conditional probability $\Pr[s_v = d | s_u = c]$ in the given phylogeny. The model also often specifies the priori probability of each state at the root.

Two popular reconstruction methods are the Fitch (1971) and marginal (or local) maximum likelihood (ML) methods. Fitch method is the first used for ancestral protein reconstruction. It is efficient, but is based on an unrealistic simplification of the evolutionary process (see Farris (1983) for a contrary view). As a result, it has less reconstruction accuracy. In contrast, the marginal (or local) maximum likelihood uses an explicit evolutionary model of characters to improve reconstruction accuracy. The marginal ML is proved to be the ‘most accurate’ method once an evolutionary model is known (Berger 1985, p.159 and Steel and Szekely 1999, Theorem 4). By saying ‘most accurate’, we mean that it has the highest expected probability of returning the correct root state. Recall that for estimating the root state, the marginal ML method simply selects the state that has the highest likelihood given the observed states at all the leaves under the model (with the branch lengths specified)(see Section ?? for details), and any ties are broken uniformly (Koshi and Goldstein 1996; Schluter et al 1997).

Although the ML method itself is efficient when the extant states are given, it is extremely hard to estimate the expected reconstruction accuracy of the ML method on *all* possible configurations of the extant states. Studying how to calculate the accuracy is motivated as follows. Due to resource constraint, it is often impossible to sequence all the extant genomes that are evolved from the target ancestral genome for the reconstruction purpose¹. Accordingly, it is important to select the best subset of extant genomes for ancestral genome reconstruction. Without knowing the extant states, the best strategy is to select a subset of extant genomes such that the expected reconstruction accuracy on all possible state configurations of the selected genomes is maximized. This appeared to be a difficult problem (Li, Ma, and Zhang 2007). A related and seemingly easier problem is to compute the expected accuracy for a given subset of genomes (or equivalently, a given phylogeny). An efficient algorithm on this problem will facilitate the comparison between different selections of genomes in an ancestral genome reconstruction project. In this paper, we present a fully polynomial time approximation scheme (FPTAS) for this problem.

This paper is divided into four sections. In Section ??, we briefly introduce the Fitch and marginal ML methods and formally define the accuracy of a reconstruction method. In Section ??, we first give a formula to calculate the reconstruction accuracy RA_{ML} of the marginal ML method, which is not seen in phylogeny literature. We then present an FPTAS for estimating $RA_{ML}(T)$ given a rooted phylogeny T together with an evolutionary model. More specifically, for any small $\epsilon > 0$, our method outputs an estimate that falls into the interval $[(1 - \epsilon)RA_{ML}(T), RA_{ML}(T)]$ in time polynomially proportional to the number of leaves of the input phylogeny, the logarithm of the minimum substitution rate on each branch, and $\frac{1}{\epsilon}$. In Section ??, we conclude this work by giving several remarks on the problems related to estimating the reconstruction accuracy of the ML method.

2 Ancestral State Reconstruction Problem

Let C denote the set of states of a character and T a phylogeny. A model \mathcal{M} of the evolution of the character specifies a prior distribution $\{p_{prior}(c) \mid c \in C\}$ of all possible states at the root r of T . In ancestral state reconstruction, the prior distribution is often estimated from the state distribution at leaves. For each branch $e = (u, v)$ from u to v of T , \mathcal{M} also specifies the probability $\Pr_e[d|c] = \Pr_e[s_v = d \mid s_u = c]$ that the state c evolves into d along the branch for any $c, d \in C$.

The problem of reconstructing ancestral state is to estimate the ancestral state of a character at the root r of a phylogeny T from its states at the leaves (or equivalently the extant species) given T and an evolutionary model.

2.1 Reconstruction methods

Different methods have been developed for ancestral state reconstruction. Two widely used methods are the parsimony and maximum likelihood (ML) method. The **Fitch** parsimony method assigns one or more equally good states to each non-leaf node (called internal node) to minimize the total number of state changes along all the branches of the given rooted phylogeny. The states assigned to an internal node are obtained from the states assigned to the children of the node in a recursive

¹In fact, more genomes do not necessarily give higher accuracy for reconstruction when the parsimony method is used (Li, Steel, and Zhang 2008).

manner. The reconstruction depends on the character states at the extant species and the topology of the given phylogeny.

The **marginal (or local) ML** method reconstructs the ancestral state based on the likelihood of a state given the states at the leaves. When we say D is a state configuration for the leaves of a rooted phylogeny T , we mean that it associates a state with each leaf of T . In literature, a state configuration for the leaves is also called a leaf labeling. Let the root of T be r . For a state c and a state configuration D for the leaves of T , the likelihood of a given D is defined as

$$\Pr[s_r = c | D] = \Pr[\text{The root state is } c | \text{Observing } D \text{ at leaves}].$$

The marginal ML method chooses as the ancestral state at r the one that has the maximum likelihood. A tie is broken by choosing the first state that reaches the maximum. This tie-breaking strategy is used throughout the paper. The reconstructed root state depends on the states at the extant species, the topology of the given phylogeny, and the given evolutionary model.

2.2 The reconstruction accuracy of a method

For a state $c \in C$ and a state configuration D for the leaves of the given phylogeny T , we use $\Pr[D|s_r = c]$ to denote the probability that c as the root state evolves into the state configuration D of the leaves under the given evolutionary model. The probability $\Pr[D|s_r = c]$ can be recursively calculated as follows.

Assume that the root r of T has two children u_1 and u_2 . Then,

$$\Pr[D|s_r = c] = \sum_{a_1, a_2 \in C} (\Pr_{ru_1}[a_1|c] \Pr[D(u_1)|s_{u_1} = a_1]) (\Pr_{ru_2}[a_2|c] \Pr[D(u_2)|s_{u_2} = a_2]),$$

where the summation is over all the states and $D(u_i)$ denotes the restriction of D on the leaves of the subtree rooted at u_i for each i .

The accuracy of a method M for reconstructing the ancestral state at the root of T is

$$\text{RA}_M(T) = \sum_{c, D} p_{\text{prior}}(c) \Pr[D|s_r = c] \chi_M(c, D), \quad (1)$$

where $\chi_M(c, D)$ is an ‘indicator’ function to show whether the method can correctly reconstruct c from D . Here, $\chi_M(c, D)$ indicates whether c is the output state from the method or not. For marginal ML method,

$$\chi(c, D) = \begin{cases} \frac{1}{t} & \text{if } \Pr[s_r = c|D] \text{ is the maximum likelihood,} \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where t is the number of states that have the maximum likelihood given D .

If the evolutionary model M is a Jukes-Cantor model, that is,

$$\Pr_e[c|c] = \Pr_e[d|d]$$

and

$$\Pr_e[d|c] = \frac{1}{|C| - 1} [1 - \Pr_e[c|c]]$$

for any $c, d \in C$ and any edge e , the reconstruction accuracy is independent of the prior distribution of the states at the root, where $|C|$ denotes the number of states in C . In this case,

$$\text{RA}_M(T) = \sum_D \Pr[D|s_r = c] \chi_M(c, D) \quad (3)$$

for any state c .

3 FPTAS for computing accuracy for the marginal ML method

3.1 A formula for the ML method

Let T be a phylogeny having root r . For a character state a and a state configuration D of the leaves of T , by the following Bayes' formula,

$$\Pr[s_r = a|D] = \frac{\Pr[D|s_r = a] \times p_{\text{prior}}(a)}{\Pr[D]}$$

and the definition of the indicator function given in (??), the accuracy of the marginal ML method for reconstructing the ancestral state at r becomes

$$\begin{aligned} & \text{RA}_{ML}(T) \\ &= \sum_{a \in C} p_{\text{prior}}(a) \times \left\{ \sum_D \Pr[D|s_r = a] \chi \left(\Pr[s_r = a|D] = \max_{c \in C} \Pr[s_r = c|D] \right) \right\} \\ &= \sum_D \Pr[D] \times \left\{ \sum_{a \in C} \Pr[s_r = a|D] \chi \left(\Pr[s_r = a|D] = \max_{c \in C} \Pr[s_r = c|D] \right) \right\} \\ &= \sum_D \Pr[D] \max_{a \in C} \Pr[s_r = a|D] \\ &= \sum_D \Pr[D] \max_{a \in C} \frac{\Pr[D|s_r = a] p_{\text{prior}}(a)}{\Pr[D]} \\ &= \sum_D \max_{a \in C} \{ p_{\text{prior}}(a) \Pr[D|s_r = a] \}. \end{aligned} \quad (4)$$

Let $a_{\max}(D)$ denote the state $a \in C$ such that $p_{\text{prior}}(a) \Pr[D|s_r = a]$ is maximized. Then the error rate of the marginal ML method, $\text{Error}_{ML}(T) = 1 - \text{RA}_{ML}(T)$, can be similarly written as

$$\text{Error}_{ML}(T) = \sum_D \sum_{a \neq a_{\max}(D)} p_{\text{prior}}(a) \Pr[D|s_r = a]. \quad (5)$$

Obviously, for two-state character, we have that

$$\text{Error}_{ML}(T) = \sum_D \min_{a \in C} p_{\text{prior}}(a) \Pr[D|s_r = a]$$

3.2 A fully polynomial-time approximation scheme

In this section, we present an FPTAS for computing the accuracy of reconstructing the root state for the marginal ML method. For simplicity, we assume that (i) there are only two character states 0 and 1, (ii) the character evolves in the Jukes-Cantor model, and (iii) all the branches of T have the same length, i.e., $\Pr_e[a|a] = p$ for any edge e and $a = 0, 1$, where p is a fixed number between $\frac{1}{2}$ and 1. We further set $\Pr_e[1 - a|a] = 1 - p = q$ for any edge e of T . However, our method can be easily extended to the general model without these limitations.

Let T have n leaves. Since there are only two states 0 and 1, each state configuration of the leaves corresponds uniquely to a binary string of length n . Using $\{0, 1\}^n$ to denote all the binary strings of length n , we have

$$\text{RA}_{ML}(T) = \sum_{D \in \{0,1\}^n} \max_{c=0,1} \{p_{\text{prior}}(c) \Pr[D|s_r = c]\}$$

by Equation (??)

Given a state configuration D for the leaves of T and a state a , the conditional probability $\Pr[D|s_r = a]$ can be calculated by the dynamic programming method in polynomial time (Pupko, Peer and Shamir 2000). But, this fact is not adequate for approximating $\text{RA}_{ML}(T)$ in polynomial time since the former formula contains as many as 2^n such probabilities.

To approximate $\text{RA}_{ML}(T)$, we shall use the following bucket idea: We divide the range of $\text{RA}_{ML}(T)$ into small intervals and estimate how many state configurations D having

$$(\Pr[D|s_r = 0], \Pr[D|s_r = 1]) \in X \times Y$$

for every pair of intervals X and Y .

We first establish a lower bound for the value of each conditional probability that we will approximate.

Proposition 3.1 *Let u be an internal node and T_u be a subtree rooted at u . Then, for any state configuration D_u of the leaves of T_u ,*

$$\Pr[D_u|s_u = a] \geq q^{2(l_u-1)}, \quad a = 0, 1 \tag{6}$$

where l_u is the number of leaves in T_u .

Proof. We prove it by induction. Assume that u has children v and w and that D_v and D_w are the restrictions of D_u in T_v and T_w , respectively. For any $a = 0, 1$, let $b = 1 - a$. Then,

$$\begin{aligned} & \Pr[D_u|s_u = a] \\ &= (p \Pr[D_v|s_v = a] + q \Pr[D_v|s_v = b]) \times (p \Pr[D_w|s_w = a] + q \Pr[D_w|s_w = b]) \\ &\geq q^2 \Pr[D_v|s_v = b] \Pr[D_w|s_w = b] \\ &\geq q^{2+2(l_v-1)+2(l_w-1)} \\ &= q^{2(l_u-1)}. \end{aligned}$$

This concludes the proof.

Theorem 3.1 For any phylogeny T in which there are n leaves and a binary character mutates with probability $q < 1/2$ on each branch, the accuracy of the marginal ML method for reconstructing the root state of the character can be computed with approximation ratio $1 - \epsilon$ in $O(nN^4)$ time, where $N = \lceil 2(n-1)^2 \ln q / \ln(1-\epsilon) \rceil$. Note that $\ln q / \ln(1-\epsilon) \leq (1/\epsilon) \ln(1/q)$.

Proof. For any small $\epsilon > 0$, we define

$$\delta = (1 - \epsilon)^{1/(n-1)}$$

and divide the interval $[q^{2(n-1)}, 1]$ into N subintervals $[x_i, x_{i+1}]$ ($0 \leq i \leq N-1$) such that $x_0 = q^{2(n-1)}$ and $x_i/x_{i+1} = \delta$ for $0 \leq i \leq N-1$. It is easy to see that

$$x_i = q^{2(n-1)}/\delta^i, \quad 0 \leq i \leq N, \quad (7)$$

and $x_{N-1} < 1 \leq x_N$.

Let u be an internal node. We introduce N^2 variables B_{uij} ($0 \leq i, j \leq N-1$) for u . For each state configuration D_u of the leaves of T_u , we compute estimates $E_{u1}^{(D)}$ and $E_{u0}^{(D)}$ of $\Pr[D_u|s_u = 1]$ and $\Pr[D_u|s_u = 0]$, respectively. B_{uij} is set to the number of all possible state configurations D' of the leaves of T_u such that $E_{u1}^{(D')} \in [x_i, x_{i+1}]$ and $E_{u0}^{(D')} \in [x_j, x_{j+1}]$. The estimates $E_{u1}^{(D)}$ and $E_{u0}^{(D)}$ are recursively computed as follows. For simplicity, we just drop the superscript (D) in the rest of discussion.

Let u have children v and w . Initially, we set $B_{uij} = 0$ for every i and j . We update B_{uij} by considering the following cases.

Case 1. Both v and w are leaves.

Recalling $b = 1 - a$, we have, for each $a = 0, 1$,

$$\Pr[aa|s_u = a] = p^2, \Pr[ab|s_u = a] = pq, \Pr[ba|s_u = a] = pq, \Pr[bb|s_u = a] = q^2,$$

where aa, ab, ba, aa are all the possible state configurations of leaves v and w in T_u .

Let

$$\begin{aligned} k &= \lfloor 2(n-1) \ln q / \ln \delta - 2 \ln p / \ln \delta \rfloor, \\ k' &= \lfloor 2(n-1) \ln q / \ln \delta - \ln(pq) / \ln \delta \rfloor, \\ k'' &= \lfloor 2(n-2) \ln q / \ln \delta \rfloor. \end{aligned}$$

Then

$$\begin{aligned} x_k &\leq p^2 < x_{k+1}, \\ x_{k'} &\leq pq < x_{k'+1}, \end{aligned}$$

and

$$x_{k''} \leq q^2 < x_{k''+1}.$$

Therefore, we have that

$$B_{ukk''} = 1, B_{uk'k'} = 2, B_{uk''k} = 1,$$

and other B_{uij} are equal to 0.

Case 2. u has a leaf child, say v , and an internal child, say w .

For any state configuration D_u for the leaves of T_u , we have that

$$\begin{aligned} & \Pr[D_u | s_u = a] \\ = & \prod_{x=v,w} (p \Pr[D_x | s_x = a] + q \Pr[D_x | s_x = b]) \end{aligned} \quad (8)$$

$$= \begin{cases} p (p \Pr[D_w | s_w = a] + q \Pr[D_w | s_w = b]), & \text{if } D_v = a, \\ q (p \Pr[D_w | s_w = a] + q \Pr[D_w | s_w = b]), & \text{if } D_v \neq a, \end{cases} \quad (9)$$

where D_v is the state assigned to the leaf v in the state configuration D_u , and D_w is the restriction of D_u on the leaves of T_w .

For every pair of i and j such that $B_{wij} \neq 0$, we do the following calculations:

(i) Calculate E_{u1} and E_{u0} as

$$\begin{aligned} E_{u1} &= p(px_i + qx_j), \\ E_{u0} &= q(qx_i + px_j). \end{aligned}$$

Setting

$$\begin{aligned} k &= \lfloor 2(n-1) \ln q / \ln \delta - E_{u1} / \ln \delta \rfloor, \\ k' &= \lfloor 2(n-1) \ln q / \ln \delta - E_{u0} / \ln \delta \rfloor, \end{aligned}$$

we have $E_{u1} \in [x_k, x_{k+1}]$ and $E_{u0} \in [x_{k'}, x_{k'+1}]$. Thus, we update $B_{ukk'}$ as

$$B_{ukk'} = B_{ukk'} + B_{wij}.$$

(ii) Similarly, calculate E'_{u1} and E'_{u0} as

$$\begin{aligned} E'_{u1} &= q(px_i + qx_j), \\ E'_{u0} &= p(qx_i + px_j). \end{aligned}$$

We then set

$$\begin{aligned} m &= \lfloor 2(n-1) \ln q / \ln \delta - E'_{u1} / \ln \delta \rfloor, \\ m' &= \lfloor 2(n-1) \ln q / \ln \delta - E'_{u0} / \ln \delta \rfloor, \end{aligned}$$

and update $B_{umm'}$ as

$$B_{umm'} = B_{umm'} + B_{wij}.$$

Case 3. Both v and w are internal nodes.

For any two pairs (i, j) and (i', j') such that $B_{vij} \neq 0$ and $B_{wi'j'} \neq 0$, by Equation (??), we first calculate E_{u1} and E_{u0} as

$$\begin{aligned} E_{u1} &= (px_i + qx_j)(px_{i'} + qx_{j'}), \\ E_{u0} &= (qx_i + px_j)(qx_{i'} + px_{j'}). \end{aligned}$$

According to E_{u0}, E_{u1} , we update B_{uij} as follows:

$$B_{ukk'} = B_{ukk'} + B_{vij} B_{wi'j'},$$

where

$$k = \lfloor 2(n-1) \ln q / \ln \delta - E_{u1} / \ln \delta \rfloor,$$

$$k' = \lfloor 2(n-1) \ln q / \ln \delta - E_{u0} / \ln \delta \rfloor,$$

There are $n-1$ internal nodes, for each internal node u , we need to update N^2 variables. In the worst case (which happens in the case 3), we use $O(N^2 \times N^2)$ operations to update the N^2 variables associated with each internal node. Overall, the whole updating procedure takes about $O(nN^4)$ operations to compute the values of the N^2 variables B_{rij} associated with the root r . After B_{rij} are computed, we output

$$\text{Acc} = \sum_{0 \leq i, j \leq N-1} B_{rij} \max\{p_{\text{prior}}(1)x_i, p_{\text{prior}}(0)x_j\}$$

as the estimate of $\text{RA}_{ML}(T)$. By Theorem ?? (to be proved later),

$$\begin{aligned} \text{Acc} &= \sum B_{rij} \max\{p_{\text{prior}}(1)x_i, p_{\text{prior}}(0)x_j\} \\ &\leq \sum_D \max_{c=0,1} \{p_{\text{prior}}(c) \Pr[D|s_r = c]\} \\ &= \text{RA}_{ML}(T). \end{aligned}$$

On the other hand, by Theorem ?? and the definition of δ ,

$$\begin{aligned} &\text{RA}_{ML}(T) \\ &= \sum_D \max_{c=0,1} \{p_{\text{prior}}(c) \Pr[D|s_r = c]\} \\ &\leq \sum B_{rij} \max\{p_{\text{prior}}(1) \frac{x_i}{\delta^{n-1}}, p_{\text{prior}}(0) \frac{x_j}{\delta^{n-1}}\} \\ &\leq \frac{1}{\delta^{n-1}} \sum B_{rij} \max\{p_{\text{prior}}(1)x_i, p_{\text{prior}}(0)x_j\} \\ &= \frac{1}{1-\epsilon} \text{Acc}. \end{aligned}$$

This shows that Acc is a $(1-\epsilon)$ -approximation of the reconstruction accuracy and finishes the proof.

Remarks: 1. The statement in above theorem still holds when the mutational probabilities on all branches are not equal after we replace q with the minimum mutational probability on a branch.

2. The theorem can be generalized to the general case in which there are multiple states. Assume there are m states. In each internal nodes, we need to introduce N^m instead of N^2 variables. The values of these variables can be updated using N^{2m} operations in the same way as in the binary case. The complexity of the algorithm then becomes $O(nN^{2m})$, where N is defined in the above theorem. The theorem also holds for any general evolutionary model. To prove it, we just need to modify Formulas (??) and (??) as well as Theorem ??.

Theorem 3.2 *Let n_u denote the number of internal nodes in the subtree T_u rooted at u . For $a = 0, 1$ and any state configuration D_u of the leaves in T_u , if the estimates E_{ua} of $\Pr[D_u|s_u = a]$ falls in the subinterval $[x_i, x_{i+1}]$ for some i , then,*

$$x_i \leq \Pr[D_u|u = a] \leq x_{i+1} / \delta^{n_u}.$$

Proof. By the computation of B_{uij} in the case 1 in the proof of Theorem ??, the statement is true if u has two leaf children, where $n_u = 1$ and $E_{ua} = \Pr[D_u|s_u = a]$.

In general, we assume u has two internal children v and w and the statement is true for both v and w . For any state configuration D_u for leaves in T_u , D_v and D_w are used to denote the restrictions of D_u in T_v and T_w , respectively. Moreover, for $y = v, w$ and $a = 0, 1$, we use E_{ya} to denote the estimate of $\Pr[D_y|s_y = a]$ and assume $E_{ya} \in [x_{i_{ya}}, x_{i_{ya+1}}]$ for some i_{ya} . By assumption,

$$x_{i_{ya}} \leq \Pr[D_y|s_y = a] \leq \frac{x_{i_{ya}}}{\delta^{n_y}}, \quad (10)$$

for $y = v, w$ and $a = 0, 1$.

By the computation in the case 3 of the proof of Theorem ??, for $a = 0, 1$, the estimate E_{ua} of $\Pr[D_u|s_u = a]$ is calculated as

$$E_{ua} = (px_{i_{va}} + qx_{i_{vb}})(px_{i_{wa}} + qx_{i_{wb}}),$$

where $b = 1 - a$. Let $x_m \leq E_{ua} < x_{m+1}$. Then, by (??), we have that

$$\begin{aligned} & \Pr[D_u|s_u = a] \\ &= (p \Pr[D_v|s_v = a] + q \Pr[D_v|s_v = b]) \times (p \Pr[D_w|s_w = a] + q \Pr[D_w|s_w = b]) \\ &\geq (px_{i_{va}} + qx_{i_{vb}})(px_{i_{wa}} + qx_{i_{wb}}) \\ &= E_{ua} \\ &\geq x_m, \end{aligned}$$

and

$$\begin{aligned} & \Pr[D_u|s_u = a] \\ &= (p \Pr[D_v|s_v = a] + q \Pr[D_v|s_v = b]) (p \Pr[D_w|s_w = a] + q \Pr[D_w|s_w = b]) \\ &\leq \left(p \frac{x_{i_{va}}}{\delta^{n_v}} + q \frac{x_{i_{vb}}}{\delta^{n_v}} \right) \left(p \frac{x_{i_{wa}}}{\delta^{n_w}} + q \frac{x_{i_{wb}}}{\delta^{n_w}} \right) \\ &\leq E_{ua} / \delta^{n_v+n_w} \\ &\leq x_{m+1} / \delta^{n_v+n_w} \\ &= x_m / \delta^{n_u}, \end{aligned}$$

since $x_{m+1} = x_m / \delta$ and $n_u = n_v + n_w + 1$. This proves the statement for the case that both v and w are two internal nodes. Similarly, we can prove the statement for the case that u has just one leaf child. Therefore, the theorem is proved.

Note that Theorem ?? is very strong. For $\delta = (1 - \epsilon)^{1/(n-1)}$ as defined in the proof of Theorem ??, Theorem ?? says that each term in Equations (??) and (??) can be approximated within ratio $1 - \epsilon$. In Theorem ??, we used this property and Equation (??) to prove the approximation ratio for computing the accuracy. It is easy to see that the same proof works for computing the error rate by replacing (??) with (??). Thus, we have the following corollary.

Corollary 3.1 *For any phylogeny T in which there are n leaves and a binary character mutates with probability q along each branch, the reconstruction error rate of the marginal ML method for reconstructing the ancestral state at the root of T can be computed with approximation ratio $1 - \epsilon$ in $O(nN^4)$ time, where $N = \lceil 2(n-1)^2 \ln q / \ln(1 - \epsilon) \rceil$.*

4 Concluding remarks

Although finding a maximum likelihood tree is NP-hard (Chor and Tuller, 2006), the complexity of computing the accuracy of the marginal ML method for reconstruction of ancestral states is unknown. Here, we have presented an FPTAS for estimating the accuracy of the marginal ML method for reconstruction of ancestral states. Because the accuracy of a reconstruction method is non-zero in the case that there are a constant number of states, it is possible to use random sampling to estimate the accuracy of the ML method. However, our algorithm is a deterministic algorithm. Moreover, the random sampling cannot be used to efficiently estimate the error rate of the ML method as we do in Corollary ???. This is because when the error rate is very small, it takes too many trials in order to obtain one state configuration sample that gives an error.

One limitation of our algorithm is its high complexity especially for multiple-state characters. The complexity is too high to be implemented for practical use. As such, this investigation is just a first step toward exposing the complexity of the problem. One way to improve the running time is to get a lower bound on $\Pr[D_u|s_u = a]$ better than one established in Proposition ???. Obviously, this improvement is not enough. How to improve it further will be one of our future research focuses.

Another variant of the ML method is the joint ML method. To estimate the ancestral state at the root of the input phylogeny, the joint ML method finds the combination of states for all the internal nodes that maximizes the likelihood. It is unknown how to approximate efficiently the accuracy of the joint ML method for reconstruction of ancestral states.

Finally, reconstruction of ancestral genomes that evolved into the extant genomes is a grand challenging task and has only been tackled in recent years. It needs an integrated, multi-disciplinary approach for making progress (Rocchi, Archidiacono and Stanyon 2006). Since resource constraint often prohibits researchers from sequencing all the extant genomes and the accuracy of a reconstruction method is not necessarily a monotone function of the size of taxon sampling (Li, Steel and Zhang 2008), identifying a subset of taxa for the best reconstruction of an ancestral genome becomes extremely important. With these motivations, Li, Ma and Zhang (2007) proposed to study the following problems:

Taxon Selection Problem

INSTANCE: A phylogeny T over n extant species, an evolutionary model of a character, and a reconstruction method M .

OBJECTIVE: Find a subset of the extant species from which the character state at the root of T is reconstructed with the highest probability when M is applied.

K -taxon Selection Problem

INSTANCE: A phylogeny T over n extant species, an evolutionary model of a character, a reconstruction method M and an integer k .

OBJECTIVE: Find k extant species from which the character state at the root of T is reconstructed with the highest probability, over all the subsets of k extant species, when M is used.

Obviously, if the K -taxon selection problem is polynomial-time solvable, the taxon selection problem is also solvable in polynomial time by calling repeatedly the polynomial-time algorithm for the former for each k . Unfortunately, it still remains open whether these two problems are NP-hard or not for any of the known methods including the Fitch method.

In addition, with $\Pr[X|s_r = c]$ denoting the conditional probability that the Fitch method outputs a subset X of state at the root r of a given phylogeny given that the true state of r is c , the

reconstruction accuracy RA_{Fitch} of the Fitch method is equal to

$$RA_{Fitch}(T) = \sum_{c \in S} p_{\text{prior}}(c) \left(\sum_{X \subseteq S} \frac{1}{|X|} \Pr[X|_{s_r} = c] \right),$$

where S is the state set and $|X|$ the number of states contained in X . Moreover, $\Pr[X|_{s_r} = c]$ can be computed by using the dynamic programming technique in polynomial time (Maddison, 1995). Combining these two facts gives an PTAS for the above taxon selection problems (Mossel, personal communication, 2007).

The situation becomes complicated in the case of the marginal ML method. By the information processing lemma, the reconstruction accuracy of the marginal ML method is a monotonic function of the size of taxon sampling (Mossel, personal communication, 2007). It means that reconstruction the root state from all the leaf states is more accurate than from the states of any subset of leaves. Accordingly, the taxon selection problem is trivial for the marginal ML method. However, the k -taxon selection problem is much harder. In this paper, we show how to estimate the reconstruction accuracy efficiently. It is interesting to generalize our algorithm to solve the k -taxon selection problem. This is definitely one interesting problem for future research.

Acknowledgment

Louxin Zhang is supported by NUS ARF grant R146-000-109-112. He would like to thank E. Mossel for sharing his bucket idea for approximating the taxon selection problems for the Fitch method and for suggesting this author work on the ML methods. Bin Ma is supported by an NSERC grant, a start up grant at University of Waterloo, and China national high-tech R&D program 2008AA02Z313. Some of Bin Ma's work was done when he was an associate professor at the University of Western Ontario.

References

- [1] Blanchette M, Green ED, Miller W, Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* 14:2412-2423.
- [2] Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed., Springer Series in Statistics, Springer-Verlag.
- [3] Chor B, Tuller T (2006) Finding a maximum likelihood tree is hard. *J. Assoc Comput. Mach.* 53:722-744.
- [4] Evens W, Kenyon C, Peres Y, Schulman, LJ (2000) Broadcasting on trees and the ising model. *Annals of Applied Probab.* 10:410-433.
- [5] Farris JS (1983) The logical basis of phylogenetic analysis. In *Advances in Cladistics* (ed. V.A. Funk and D.R. Brooks), New York Botanical Garden, Bronx, USA.
- [6] Felsenstein J (2004) *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts.

- [7] Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst Zoology* 20:406-416.
- [8] Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285-288.
- [9] Hall BG (2006) Simple and accurate estimation of ancestral protein sequences. *Proc. Natl. Acad. Sci. USA* 103:5431-5436.
- [10] Jermann TM, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57-59.
- [11] Koshi, JM, Golstein RA (1992) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42:313-321.
- [12] Li GL, Ma J, Zhang LX (2007) Selecting Genomes for Reconstruction of Ancestral Genomes. In: Durand D, Tesler G (ed) *Proc. RECOMB-CG 2007*. LNCS, vol. 4751, pp 110-121, Springer.
- [13] Li GL, Steel M, Zhang LX (2008) More taxa are not necessarily better for the reconstruction of ancestral sequence by parsimony, *Syst. Biol.* 57:647-653.
- [14] Liberles DA (2007) *Ancestral sequence reconstruction*, Oxford University Press, Oxford.
- [15] Lucena B, Haussler D (2005) Counterexample to a claim about the reconstruction of ancestral character states. *Syst Biol.* 54:693-695.
- [16] Maddison WP (1995) Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Syst. Biol.* 44:474-481.
- [17] Martins EP (1999) Estimating of ancestral states of continuous characters: A computer simulation study. *Syst. Biol.* 48:642-650.
- [18] Mooers A Ø (2004) Effects of tree shape on the accuracy of maximum likelihood-based ancestor reconstruction. *Syst. Biol.* 53:809-814.
- [19] Mooers, A Ø, Schluter D (1999) Reconstructing ancestor states with maximum likelihood: Support for one-and two-rate models. *Syst. Biol.* 48:623-633.
- [20] Pagel M (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* 48:612-622.
- [21] Pupko T, Peer I, Shamir R (2000) A fast algorithm for joint reconstruction of amino acid sequences. *Mol. Biol. Evol.* 17:890-896.
- [22] Rocchi M, Archidiacono, Stanyon R (2006) Ancestral genomes reconstruction: an integrated, multi-disciplinary approach is needed. *Genome Res.* 16:1441-1444.
- [23] Salisbury BA, Kim J (2001) Ancestral state estimation and taxon sampling density. *Syst. Biol.* 50:557-564.
- [24] Schultz TR, Cocroft RB, Churchill GA (1996) The reconstruction of ancestral character states, *Evolution* 50:504-511.

- [25] Steel, MA, and Székely LA (1999) Inverting random functions. *Annals Combin.* 3:103-113.
- [26] Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301:1714-1717.
- [27] Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol.* 2:e69.
- [28] Yang ZH, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences, *Genetics* 141:1641-1650.
- [29] Zhang J, Nei M (1997) Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods. *J. Mol. Evol.* 44(S1):139-146.