Supplementary Document on More Taxa are not Necessarily Better for the Reconstruction of Ancestral Character States

Guoliang Li, Mike Steel, Louxin Zhang

1 The Jukes-Cantor model

Given the phylogenetic tree for a group of species, we assume that the character evolves by a Markov process, starting with a state at the root and proceeding to the leaves node by node. The probability that a node x receives a state t_x depends only on its parent node p and the conditions along the branch from p to x. The evolutionary model specifies the probability that state c evolves to state d on a branch from p to x as a conditional probability $\Pr[s_x = d|s_p = c]$. Here, we consider a simple Jukes-Cantor model. In this symmetric model, there are only two states, say 0 and 1, and the probability of a substitution change of any sort on a branch is the same.

2 Reconstruction Accuracy

Assume the character evolves in a phylogeny with the root A according to a probabilistic evolutionary model. The evolutionary model specifies a prior probability for each state at A. When we say D is a state configuration at the leaves, we mean that it contains a state for each leaf node in the phylogenetic tree. For a state c and a state configuration D at the leaves, we let $P(D|c)$ to denote the probability that c at the root evolves into the states given by D at the leaves in the phylogeny. The reconstruction accuracy of a method M is

$$P_{accuracy} = \sum_{c,D} prior(c)P(D|c)I(c, D, M),$$

where $I(c, D, M) = 1$ if the method M reconstructs c correctly from D at the root and 0 otherwise.

Here we consider a symmetric evolutionary mode with two states 0 and 1. In this case, the reconstruction accuracy is independent of the prior distribution of the states

and hence

$$P_{accuracy} = \sum_D P(D|0)I(0, D, M) = \sum_D P(D|1)I(1, D, M).$$

2.1 Computing Accuracy for Fitch's Method

For reconstructing character evolution, parsimony methods assign to each internal node those states that allow for the fewest number of substitutions throughout the tree. In Particular, the Fitch's method assigns a set of states to each node by node downward through the tree, starting with the leaves and using the sets previously computed for the node's two children (Fitch, 1971). For each leaf node, the observed state forms the state set. The following rules are used to combine the state sets S_B and S_C for the two children of a node A to yield the state set S_A:

$$S_A = \begin{cases} S_B \cup S_C & if \;\; S_B \cap S_C = \phi, \\ S_B \cap S_C & if \;\; S_B \cap S_C \neq \phi. \end{cases}$$

The state set at the root contains all the possible states that will be assigned to it. We say that the method **unambiguously** reconstructs a state at the root if the state set constructed at the root contains only that state and **ambiguously** reconstructs a state at the root if the state set is $\{0, 1\}$.

For the Fitch's method, $P_{accuracy}$ can be calculated through a dynamic programming approach (Maddison, 1992). Note that there are three different state sets $\{0\}$, $\{1\}$ and $\{0, 1\}$ with two states 0 and 1. For a state set t, and a state c, we use $P_N[t|c]$ to denote the probability that the state set t is assigned to the node N by the Fitch's method given the true state c at N. Under the symmetrical model, the unambiguous reconstruction accuracy of the Fitch method is

$$UA_F = P_A[\{0\}|0] = P_A[\{1\}|1];$$

the reconstruction accuracy of the Fitch method is

$$RA_F = P_A[\{1\}|1] + \frac{1}{2}P_A[\{0, 1\}|1].$$

At a leaf x with observed state s, we have

$$P_x[\{s\}|s] = 1, \;\; P_x[\{t\}|s] = P_x(\{0, 1\}|s] = 0$$

where $t \neq s$. Let N be an internal node with the children L and R. Then, for $c, d = 0, 1$,

$$\begin{aligned} & P_N[\{d\}|c] \\ = \;& \textstyle\sum_{x,y=0,1} \Pr[s_L = x|s_N = c] \Pr[s_R = y|s_N = c] P_L[\{d\}|x] P_R[\{d\}|y] \\ & + \textstyle\sum_{x,y=0,1} \Pr[s_L = x|s_N = c] \Pr[s_R = y|s_N = c] \{P_L[\{d\}|x] P_R[\{0, 1\}|y] + P_L[\{0, 1\}|x] P_R[\{d\}|y]\} \end{aligned}$$

and

$$P_N[\{0,1\}|c] = 1 - P_N[\{0\}|c] - P_N[\{1\}|c].$$

The above recurrence relations give immediately a dynamic programming approach for computing the reconstruction accuracy of the Fitch's method, which is used in our analysis in the rest of this paper. Such a method first appeared in Maddison (1995).

2.2 Computing Accuracy for the ML Method

For each state configuration D, the conditional probabilities $P_A(D|\ c)$ varies with state c at the root A. One likelihood reconstruction method takes state c with the largest conditional probability $P_A(D|\ c)$ as the best estimate for the site at the root A (Koshi and Goldstein, 1996, Schluter et al., 1997). This is called the marginal (or local) ML method (Felsenstein, 2004).

For the MML method, the indicator function can be defined as follows:

$$I(c, D, \text{MML}) = \begin{cases} 1 & P_A(D|\ c) = \max_y P_A(D|\ y) \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the accuracy of the likelihood method at the root A is the sum of the conditional probabilities $P_A(D|\ c]$ over all D such that the marginal method reconstructs c correctly from D:

$$RA_{MML} = \sum_D \Pr[D|\ 0] I_L(0, D, \text{MML}). \tag{1}$$

under the symmetric Jukes-Cantor model.

3 Accuracy Analysis on Complete Phylogenies

Let T_n be the complete binary tree with root A and 2^n leaves. Set

$$P_n = \Pr[t_{A_n} = \{x\}|A_n = x]$$

$$Q_n = \Pr[t_{A_n} = \{x\}|A_n = y], x \neq y.$$

Then

$$\Pr[t_{A_n} = \{0,1\}|A_n = x] = 1 - P_n - Q_n.$$

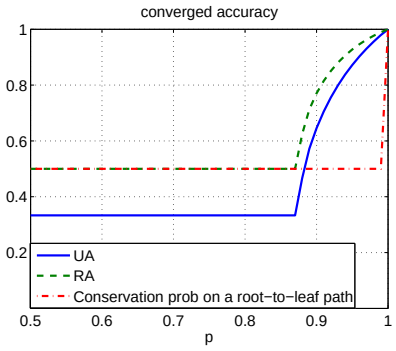

Figure 1: The limit of the reconstruction accuracy on complete phylogenetic trees. UA denotes the unambiguous reconstruction accuracy; RA denotes the reconstruction accuracy.

Let B and C be the children of A. Since the subtree rooted at each child of A_n is T_{n-1}, we have

$$\begin{aligned}
& P_n \\
= \; & \textstyle\sum_{x=0,1}\sum_{y=0,1} \Pr[s_B = x|s_A = 1]\Pr[s_C = y|s_A = 1]\Pr[t_B = \{1\}|s_B = x]\Pr[t_C = \{1\}|s_C = y] \\
& + \textstyle\sum_{x=0,1}\sum_{y=0,1} \Pr[s_B = x|s_A = 1]\Pr[s_C = y|s_A = 1]\Pr[t_B = \{1\}|s_B = x]\Pr[t_C = \{0,1\}|s_C = y] \\
& + \textstyle\sum_{x=0,1}\sum_{y=0,1} \Pr[s_B = x|s_A = 1]\Pr[s_C = y|s_A = 1]\Pr[t_B = \{0,1\}|s_B = x]\Pr[t_C = \{1\}|s_C = y] \\
= \; & (p^2P_{n-1}^2 + q^2Q_{n-1}^2 + 2pqP_{n-1}Q_{n-1}) \\
& +2(1 - P_{n-1} - Q_{n-1})[p^2P_{n-1} + pqP_{n-1} + pqQ_{n-1} + q^2Q_{n-1}] \\
= \; & (pP_{n-1} + qQ_{n-1})^2 + 2(1 - P_{n-1} - Q_{n-1})(pP_{n-1} + qQ_{n-1}) \\
= \; & (pP_{n-1} + qQ_{n-1})[2 - (1+q)P_{n-1} - (1+p)Q_{n-1}].
\end{aligned}$$

Similarly, we also have

$$Q_n = (pQ_{n-1} + qP_{n-1})[2 - (1+p)P_{n-1} - (1+q)Q_{n-1}]$$

The numerical computation based on these two recurrence relations shows that, when n goes to infinity, the unambiguous reconstruction accuracy $UA_n(p) = P_n$ converges to $1/3$ for $p \in [1/2, 7/8]$ and to some $\lambda(p)$ for $p \in (7/8, 1)$, where

$$\lambda(p) = \frac{1}{2(2p-1)}\Big[\,(4p-3) + \frac{1}{(2p-1)}\sqrt{(8p-7)(4p-3)}\;\Big]$$

(Steel, 1989). The curve of $\lambda(p)$ is drawn in Figure 1. The reconstruction accuracy $RA_n(p)$ converges to $1/2$ when $p \in [1/2, 7/8]$ and

$$\frac{1}{2} + \frac{1}{2(2p-1)^2}\sqrt{(8p-7)(4p-3)}.$$

4 Accuracy Analysis on Comb-shaped Phylogenies

Let I_n be the comb-shaped phylogeny with n nodes and root A_n as shown in Figure 2. In this phylogeny, one child of A is a leaf and the other child is the root of the imbalanced phylogeny I_{n-1}. Let the leaf child of A_n be u and let the other child of A be B. Again, we set

$$P_n = \Pr[t_A = \{1\}|s_A = 1], \;\; Q_n = \Pr[t_A = \{0\}|s_A = 1]$$

We have

$$\begin{array}{rl} & P_n \\ = & \sum_{x=0,1} \Pr[s_u = 1|s_A = 1] \Pr[s_B = x|s_A = 1] \Pr[t_B \neq \{0\}|s_B = x] \\ = & p(1 - pQ_{n-1} - qP_{n-1}). \end{array}$$

Similarly, we also have

$$Q_n = q(1 - pP_{n-1} - qQ_{n-1}).$$

Simplifying these recurrence relations, we have

$$P_{n+1} + (1-p)P_n + p(1 - 3p + 2p^2)P_{n-1} - p(2p^2 - 3p + 2) = 0$$

and so

$$P_n = a_1\lambda_1^n + a_2\lambda_2^n + \frac{p(2 - 3p + 2p^2)}{2 - 3p^2 + 2p^3}$$

where λ's are the two roots of the characteristic equation

$$\lambda^2 + (1-p)\lambda + p(1 - 3p + 2p^2) = 0.$$

Since $|\lambda_1|, |\lambda_2| < 1$ when $0 < p < 1$, P_n converges to $\frac{p(2-3p+2p^2)}{2-3p^2+2p^3}$ when n goes to infinity.

Similarly, we can show that $\Pr[t_A = \{0, 1\}|s_A = 1] = 1 - P_n - Q_n$ converges to $\frac{(1-p)(1-p+2p^2)}{2-3p^2+2p^3}$ when n goes to infinity. Hence, the reconstruction accuracy converges $RA_n(p)$ to

$$RA(p) = \frac{1}{2} + \frac{2p - 1}{2(2 - 3p^2 + 2p^3)} = \frac{1 + 2p - 3p^2 + 2p^3}{2(2 - 2p^2 + 2p^3)}.$$

Since $2 - 3p^2 + 2p^3 > 1$ for any $p \in [1/2, 1)$,

$$RA(p) < \frac{1}{2} + \frac{2p - 1}{2} = p.$$

Hence, the reconstruction accuracy is smaller than p in the limit case unless $p = 1$ (see Figure 2).

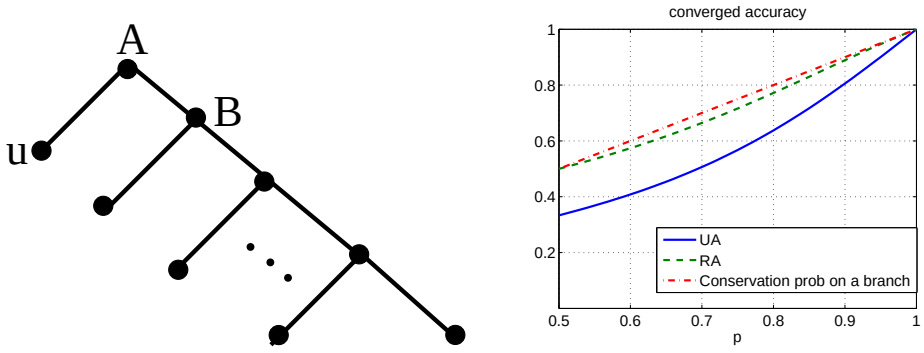

Figure 2: The limit of the reconstruction accuracy on comb-shaped trees.

5 Accuracy Analysis on Imbalanced Phylogenies

In this section, we consider imbalanced phylogenies in which the root has a leaf as its child. Let T be such a phylogeny and let its root A have leaf child Y and non-leaf child Z as illustrated in Figure 3. We assume that the conservation probability on the branches leading to Y and Z be p_1 and p_2 respectively.

5.1 Parsimony Method

Setting $P_Z[\{a\} \mid a] = \beta$ and $P_Z[\{a\} \mid b] = \gamma$ for different $a, b \in \{0, 1\}$, we have

$$\begin{aligned} & P_A[\{0\} \mid 0] \\ = \; & Pr[s_Y = 0 \mid s_A = 0] \\ & \times(\Pr[s_Z = 0 \mid s_A = 0]P_Z[t_Z \neq \{1\} \mid 0] + \Pr[s_Z = 1 \mid s_A = 0]P_Z[t_Z \neq \{1\} \mid 1]) \\ = \; & p_1(p_2(1-\gamma) + q_2(1-\beta)) \\ = \; & p_1(1 - p_2\gamma - q_2\beta), \end{aligned}$$

$$\begin{aligned} & P_A[\{1\} \mid 0] \\ = \; & Pr[s_Y = 1 \mid s_A = 0] \\ & \times(\Pr[s_Z = 0 \mid s_A = 0]P_Z[t_Z \neq \{0\} \mid 0] + \Pr[s_Z = 1 \mid s_A = 0]P_Z[t_Z \neq \{0\} \mid 1]) \\ = \; & q_1(p_2(1-\beta) + q_2(1-\gamma)) \\ = \; & q_1(1 - q_2\gamma - p_2\beta), \end{aligned}$$

and

$$\begin{aligned} & P_A[\{0,1\} \mid 0] \\ = \; & 1 - P_A[\{0\} \mid 0] - P_A[\{1\} \mid 0] \\ = \; & (p_1p_2 + q_1q_2)\gamma + (p_1q_2 + q_1p_2)\beta. \end{aligned}$$

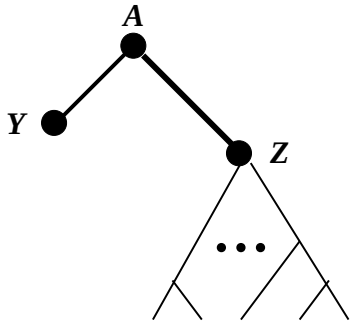

Figure 3: The imbalanced phylogeny

The reconstruction accuracy of the Fitch's method on T is

$$\begin{aligned} & AA_P \\ = \; & P_A[\{0\} \mid 0] + \tfrac{1}{2} P_A[\{0,1\} \mid 0] \\ = \; & p_1 - \tfrac{1}{2}(p_1 p_2 - q_1 q_2)\gamma - \tfrac{1}{2}(p_1 q_2 = q_1 p_2)\beta \\ = \; & p_1 + \tfrac{1}{2}(1 - p_1 - p_2)\gamma + \tfrac{1}{2}(p_2 - p_1)\beta \end{aligned}$$

If $p_1 \geq p_2 > \frac{1}{2}$, $AA_P < p_1$ since $\beta, \gamma > 0$.

5.2 The Accuracy of the Marginal ML Method

For $a \in \{0,1\}$, we use aD_Z to denote the state configuration of the leaves in which the leaf Y has state a and the state configuration of the leaves below Z is D_Z. Then, the probability that 0 at the root A evolves into the states specified by $0D_Z$ is

$$P_A(0D_Z \mid 0) = p_1[p_2 P_Z(D_Z \mid 0) + q_2 P_Z(D_Z \mid 1)],$$

and the probability that 1 at the root evolves into the states specified by $0D_Z$ is

$$P_A(0D_Z \mid 1) = q_1[p_2 P_Z(D_Z \mid 1) + q_2 P_Z(D_Z \mid 0)].$$

Assume $p_1 \geq p_2 > 0.5$. We have

$$P_A(0D_Z \mid 0) - P_A(0D_Z \mid 1) = (p_1 + p_2 - 1)P_Z(D_Z \mid 0) + (p_1 - p_2)P_Z(D_Z \mid 1) > 0.$$

This implies that the marginal ML method select 0 as the ancestor state at A when the state at Y is 0. As a result, the accuracy of the marginal ML method is

$$AA_{MML} = \sum_{D_Z} P_A(0D_Z \mid 0) = p.$$