Letter to Editor

Majority rule has transition ratio 4 on Yule trees under a 2-state symmetric model

ARTICLE INFO

ABSTRACT

Inferring the ancestral state at the root of a phylogenetic tree from states observed at the leaves is a problem arising in evolutionary biology. The simplest technique – majority rule – estimates the root state by the most frequently occurring state at the leaves. Alternative methods – such as maximum parsimony - explicitly take the tree structure into account. Since either method can outperform the other on particular trees, it is useful to consider the accuracy of the methods on trees generated under some evolutionary null model, such as a Yule pure-birth model. In this short note, we answer a recently posed question concerning the performance of majority rule on Yule trees under a symmetric 2-state Markovian substitution model of character state change. We show that majority rule is accurate precisely when the ratio of the birth (speciation) rate of the Yule process to the substitution rate exceeds the value 4. By contrast, maximum parsimony has been shown to be accurate only when this ratio is at least 6. Our proof relies on a second moment calculation, coupling, and a novel application of a reflection principle.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Given a binary tree, T, suppose that a state from some set S is assigned uniformly at random to the root of T. This state then evolves down the tree to the tips of the tree according to a continuous-time Markovian process on S, acting along the edges of the tree. Given the states at the tips of a tree, *ancestral state reconstruction* aims to estimate the state that was present at the root of the tree. This problem is particularly relevant to certain questions in evolutionary biology (Liberles, 2007; Royer-Carenzi et al., 2013).

The performance of any ancestral state reconstruction methods depends on the underlying tree (its topology and branch lengths); accordingly, to compare the performance of different ancestral state reconstruction methods, it is helpful to sample trees from some underlying null distribution. In evolutionary biology, a natural and widely used null process is the Yule pure-birth model (Stadler and Steel, 2012; Yule, 1925), starting with a single lineage at time 0 and grown for time t with birth rate λ, and this is the model we study here. Moreover, for the rest of this paper, we will consider the simple continuous-time Markov process, on the state space $S = \{+1, -1\}$ with an instantaneous substitution rate m between the two states. Notice that there are two random processes at play here – the generation of the tree and the substitution process that then applies along the edges of this tree.

A straightforward information-theoretic argument shows that any method for estimating the root state at a Yule tree cannot achieve an accuracy that is strictly bounded above 1/2 as t grows, when $\lambda < 4m$, even when the tree and its branch lengths are given (Gascuel and Steel, 2014). If just the tree topology is known (but not necessarily its branch lengths) then a natural and often used ancestral state reconstruction approach is to assign a root state that minimizes the number of state changes in the tree required to explain the states at the leaves. This method is known as *maximum*

parsimony and it was shown in Gascuel and Steel (2010) and Li (2011) that when $\lambda/m < 6$, this method does no better than guessing the root state, as $t \to \infty$ (for $\lambda/m > 6$, the probability of correct reconstruction (as $t \to \infty$) is strictly greater than 1/2 and converges to 1 (as $\lambda/m \to \infty$). The difference between these two ancestral state reconstruction methods is illustrated in Fig. 1.

There is an even simpler way to estimate the root state from the leaf states, which does not even require us to know the tree topology. This is to simply count the number of leaves in each state and use a majority state as the estimate (ties are broken randomly). For this *majority rule* method, the question of determining the ratio of λ/m at which root state estimation retains some accuracy as $t \to \infty$ was posed in Gascuel and Steel (2014). In this note, we show that this transition occurs for majority rule at $\lambda/m = 4$, which is therefore the smallest possible ratio. In particular, there is a range $(4 < \lambda/m < 6)$ within which simple majority rule will outperform a recursive method that explicitly uses the tree topology information (maximum parsimony), despite the fact that for some trees, maximum parsimony can have higher accuracy than majority rule (Gascuel and Steel, 2014). Our findings are consistent with simulations that have suggested that majority rule tends to have higher overall accuracy for ancestral state reconstruction on Yule trees than maximum parsimony (Gascuel and Steel, 2014), and complement a recent study of ancestral state reconstruction on Yule trees for continuous characters evolving under an Ornstein–Uhlenbeck process (Bartoszek and Sagitov, 2013).

It is interesting to compare our results to results on census reconstruction from Mossel and Peres (2003). Theorem 1.4 in Mossel and Peres (2003) implies that when $\lambda > 4m$, then the reconstruction problem is *census solvable*. This means that there is a linear estimator $\sum a_v \sigma_v$ of the root in terms of the leaves σ_v which is correlated with the root of the tree. The coefficients of this linear estimator depend on the topology and edge lengths of
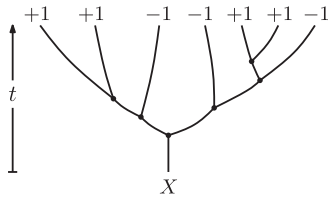
Fig. 1. A tree generated under a Yule process for time t with seven leaves. For the states at the leaves shown, majority rule will assign state $+1$ to the unknown root state $X=\pm 1$, while maximum parsimony will assign state -1 to X (since $X=-1$ requires just two state changes in the tree, while $X=+1$ requires at least three).

the tree. In contrast, we are interested in the simpler estimator which is simply given by the majority of the leaf values and show that it results in correlated reconstruction for $\lambda > 4m$. Interestingly, our proof shows that for the Yule tree, the majority reconstruction estimator maximizes the reconstruction probability among all reconstruction methods which are functions of the number of $+1$ and -1 leaves only.

We note further that the threshold $\lambda > 4m$ is the threshold for reconstruction of *spherically symmetric trees* if the number of leaves at distance t is $\Theta(e^{\lambda t})$ and that, in this case, majority reconstruction achieves the threshold. See Evans et al. (2000) for more details and the definition of spherically symmetric trees.

1.1. Preliminaries

First recall that under the symmetric 2-state process, if the initial state is $+1$, the state $\sigma_t \in S$ after time t is the random variable with distribution:

$$\sigma_t = \begin{cases} +1 & \text{with probability } \frac{1}{2}(1+e^{-2mt}); \\ -1 & \text{with probability } \frac{1}{2}(1-e^{-2mt}). \end{cases}$$

Notice that

$$\mathbb{E}[\sigma_t] = e^{-2mt}. \tag{1}$$

Let L_t be the set of leaves at time t. It is well known that $N_t := |L_t|$ has a geometric distribution with parameter $p = e^{-\lambda t}$ (and so $N_t e^{-\lambda t}$ converges in distribution to an exponential distribution with mean 1). In particular, we have

$$\mathbb{E}[N_t] = e^{\lambda t}. \tag{2}$$

Let

$$S_t = \sum_{i \in L_t} \sigma_t(i),$$

where $\sigma_t(i)$ is the state at leaf i on the resulting Yule tree, conditional on the root of the tree being in state $+1$. We first compute the first moment of S_t. Eq. (1) gives $\mathbb{E}[S_t] = \mathbb{E}[\mathbb{E}[S_t|N_t]] = \mathbb{E}[N_t \cdot e^{-2mt}]$ from which Eq. (2) gives

$$\mathbb{E}[S_t] = e^{(\lambda - 2m)t}. \tag{3}$$

2. Second moment calculation

Calculating the second moment of S_t requires a little more care. First, observe that we may write

$$S_t^2 = \sum_{i \in L_t} \sigma_t(i)^2 + \sum_{(i,j) \in \tilde{L}_t} \sigma_t(i)\sigma_t(j),$$

where $\tilde{L}_t = \{(i,j) \in L_t \times L_t : i \neq j\}$. Consequently, since $\sum_{i \in L_t} \sigma_t(i)^2 = N_t$, we have

$$\mathbb{E}[S_t^2] = e^{\lambda t} + F(t), \tag{4}$$

where $F(t) = \mathbb{E}[\tilde{S}_t]$ for $\tilde{S}_t = \sum_{(i,j) \in \tilde{L}_t} \sigma_t(i)\sigma_t(j)$.

Now, suppose that, for the Yule tree grown for time t, two leaves i and j have a most-recent common ancestor at time $t - t'$. Then conditional on this,

$$\mathbb{E}[\sigma_t(i)\sigma_t(j)] = e^{-4mt'}, \tag{5}$$

where expectation is with respect to the substitution process alone.

The function $F(t)$ satisfies $F(0) = 0$, and, by the nature of the Yule pure-birth process, and Eq. (5), we have

$$F(t+\delta) = (1 + 2\lambda\delta + O(\delta^2)) \cdot (e^{-4m\delta}F(t)) \\ + (\lambda\delta + O(\delta^2))(1 - O(\delta))\mathbb{E}[N_t] \tag{6}$$

Here the first of the two summands

$$(1 + 2\lambda\delta + O(\delta^2)) \cdot (e^{-4m\delta}F(t)),$$

is the total contribution to $F(t+\delta)$ coming from all pairs of different leaves at time t. The main contribution is $e^{-4m\delta}F(t)$ from all pairs at time t but we have to include the additional contribution when one of the two leaves in a pair splits into two lineages given by the $2\lambda\delta e^{-4m\delta}F(t)$ term; the probability that two neighboring leaves split is $O(\delta^2)$. The second summand

$$(\lambda\delta + O(\delta^2))(1 - O(\delta))\mathbb{E}[N_t]$$

is the contribution made by all pairs of children of the same leaf that splits in the δ period. More precisely, conditional on N_t, exactly one leaf will split into two leaves (call them l and l') in the interval $(t, t+\delta)$ with probability $\lambda\delta N_t + O(\delta^2)$ (the probability of more than one leaf splitting in this interval is $O(\delta^2)$). Moreover, the length of the two new branches ending in l and l' is $O(\delta)$, and so $\sigma_{t+\delta}(l)\sigma_{t+\delta}(l')$ equals $+1$ with probability $1 - O(\delta)$. Taking expectation gives the second summand (i.e. $(\lambda\delta + O(\delta^2))(1 - O(\delta))\mathbb{E}[N_t]$).

Now, $e^{-4m\delta} = 1 - 4m\delta + O(\delta^2)$, so if we apply this, along with Eq. (2) in Eq. (6), and collect together all terms of quadratic or higher order in δ, we obtain

$$F(t+\delta) = (1 - (4m - 2\lambda)\delta)F(t) + \lambda\delta e^{\lambda t} + O(\delta^2).$$

Rearranging this, and letting $\delta \to 0$, we obtain the following linear differential equation for $F(t)$:

$$\frac{dF}{dt} + 2(2m - \lambda)F = \lambda e^{\lambda t}. \tag{7}$$

Solving for F is standard (using the integrating factor $I(s) = e^{(4m - 2\lambda)s}$ and the initial condition $F(0) = 0$ gives $F(t) = e^{(2\lambda - 4m)t} \int_0^t \lambda e^{-(\lambda - 4m)s} ds$) and so

$$F(t) = e^{(2\lambda - 4m)t} \times \frac{\lambda}{\lambda - 4m}(1 - e^{-(\lambda - 4m)t}). \tag{8}$$

This and Eq. (3) leads to the following result:

Proposition 2.1. $\mathbb{E}[S_t^2] = e^{\lambda t} + F(t)$, where $F(t)$ is given by (8). In particular, when $\lambda > 4m$, then for all $t \geq 0$:

$$\frac{\mathbb{E}[S_t^2]}{\mathbb{E}[S_t]^2} = e^{-rt} + \frac{1}{(1 - 4m/\lambda)}(1 - e^{-rt}),$$

where $r = \lambda - 4m > 0$.

We note that exactly the same proof can be applied to $S_{t,+}$ (resp. $S_{t,-}$) which is S_t conditioned on the root being in state $+1$ (resp. -1). We will use this to establish our desired result.

3. A lower bound on the total variation distance of $S_{t,-}$ and $S_{t,+}$

Out next goal is to show the following.