Journal of Theoretical Biology 360 (2014) 315–318

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Letter to Editor

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

ARTICLE INFO

Keywords:
Yule pure-birth tree
Markov process
Ancestral state
Information loss

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.

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

http://dx.doi.org/10.1016/j.jtbi.2014.07.029
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

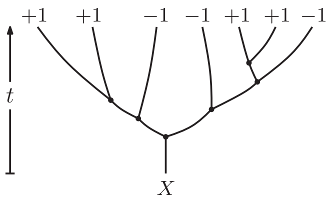

Fig. 1. A tree generated under a Yule process for time t with seven leaves. For the states at the leaves shown, majority rule will assign state $+1$ to the unknown root state $X=\pm 1$, while maximum parsimony will assign state -1 to X (since $X=-\ 1$ requires just two state changes in the tree, while $X=+1$ 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.

Lemma 3.1. *Provided $\lambda > 4m$, then for $r = \lambda - 4m > 0$:*

$$d_{TV}(S_{t,+}, S_{t,-}) \geq \frac{1-4m/\lambda}{1-4me^{-rt}/\lambda},$$

and the expression on the right is a monotone decreasing function of t, from 1 *(at* t=0*) to* $1-4m/\lambda$ *(as* $t \to \infty$*).*

Proof. We first recall that the total variation distance between any two random variables X, Y on the same probability space Ω is defined by

$$d_{TV}(X,Y) := \frac{1}{2} \sum_{\omega \in \Omega} \left| \mathbb{P}[X=\omega] - \mathbb{P}[Y=\omega] \right|. \tag{9}$$

A dual definition, which will be used in the proof of the lemma, is given by:

$$d_{TV}(X,Y) := \inf\{\mathbb{P}[X' \neq Y'] : X' \sim X, Y' \sim Y\}, \tag{10}$$

where the infimum is taken over all random variables on Ω^2 with X' having the same distribution as X, and with Y' having the same distribution as Y (such (X', Y') is called a coupling of X and Y).

Let (X, Y) be a coupling of $S_{t,+}$ and $S_{t,-}$. We will use (10) to place a lower bound on the total variation distance by providing a lower bound on $\mathbb{P}[X \neq Y]$. This is a standard application of the second moment method (see e.g. Levin et al., 2010, Proposition 7.8). Using Paley–Zygmund's second moment inequality, one has

$$\mathbb{P}[X \neq Y] \geq \frac{(\mathbb{E}[|X-Y|])^2}{\mathbb{E}[(X-Y)^2]}.$$

By Jensen's inequality one has

$$(\mathbb{E}[|X-Y|])^2 \geq (\mathbb{E}[X] - \mathbb{E}[Y])^2 = 4\mathbb{E}[S_{t,+}]^2.$$

Moreover,

$$\mathbb{E}[(X-Y)^2] \leq 2\mathbb{E}[X^2] + 2\mathbb{E}[Y^2] = 4 \times \frac{1}{2}(\mathbb{E}[S_{t,+}^2] + \mathbb{E}[S_{t,-}^2]) = 4\mathbb{E}[S_{t,+}^2].$$

Thus we have proved that

$$\mathbb{P}(X \neq Y) \geq \frac{\mathbb{E}[S_{t,+}]^2}{\mathbb{E}[S_{t,+}^2]},$$

and Lemma 3.1 now follows from Proposition 2.1 (noting that S_t in that proposition is $S_{t,+}$). This completes the proof of the lemma. □

4. Majority reconstruction

In order to complete the proof, we will establish the following lemma.

Lemma 4.1. *For all $t \geq 0$, the probability M_t that majority rule reconstructs the root state correctly is given by*

$$M_t = \frac{1}{2} + \frac{1}{2} d_{TV}(S_{t,+}, S_{t,-}).$$

Proof. Let σ denote the root value. Then, by rewriting (9), we see that

$$D_t := d_{TV}(S_{t,+}, S_{t,-}) = \frac{1}{2} \sum_s \left| \mathbb{P}[S_t = s \middle| \sigma = +1] - \mathbb{P}[S_t = s \middle| \sigma = -1] \right|. \tag{11}$$

Moreover, the probability of reconstruction by majority rule is given by

$$M_t = \sum_{s>0} \mathbb{P}[S_t = s]\mathbb{P}[\sigma = +1 | S_t = s] + \sum_{s<0} \mathbb{P}[S_t = s]\mathbb{P}[\sigma = -1 | S_t = s] + \frac{1}{2}\mathbb{P}[S_t = 0]. \tag{12}$$

Since $\mathbb{P}[\sigma = +1 | S_t = s] + \mathbb{P}[\sigma = -1 | S_t = s] = 1$, we can rewrite $\mathbb{P}[\sigma = +1 | S_t = s]$ as $0.5 + 0.5(\mathbb{P}[\sigma = +1 | S_t = s] - \mathbb{P}[\sigma = -1 | S_t = s])$ and similarly for the other terms. We thus obtain the following from (12):

$$M_t = \frac{1}{2} + \frac{1}{2}\sum_s \mathbb{P}[S_t = s](\mathbb{P}[\sigma = +1 \middle| S_t = s] - \mathbb{P}[\sigma = -1 \middle| S_t = s]) \operatorname{sgn}(s) \tag{13}$$

$$M_t = \frac{1}{2} + \frac{1}{4}\sum_s (\mathbb{P}[S_t = s \middle| \sigma = +1] - \mathbb{P}[S_t = s \middle| \sigma = -1]) \operatorname{sgn}(s). \tag{14}$$

Comparing (14) and (11), we see that in order to prove the lemma, it suffices to show that if $s > 0$ then $\mathbb{P}[S_t = s | \sigma = +1] > \mathbb{P}[S_t = s | \sigma = -1]$, while if $s < 0$ then $\mathbb{P}[S_t = s | \sigma = +1] < \mathbb{P}[S_t = s | \sigma = -1]$.

The proof of this follows from the reflection principle. Consider the Markov chain (N_t, S_t) where N_t is the population size. Let T be the first stopping time where S_T=0 ($T = \infty$ if it does not happen). Then for $s > 0$, we have (where σ is the root state)

$\mathbb{P}[S_t = s | T > t, \sigma = +1] > 0, \quad \mathbb{P}[S_t = s | T > t, \sigma = -1] = 0,$ and
$\mathbb{P}[S_t = s | T \leq t, \sigma = +1] = \mathbb{P}[S_t = s | T \leq t, \sigma = -1].$

From this, it follows that

$$\mathbb{P}[S_t = s | \sigma = +1] > \mathbb{P}[S_t = s | \sigma = -1],$$

as needed. The symmetric argument applies when $s < 0$. □

Recall that when $\lambda/m < 4$ then $\lim_{t\to\infty} M_t = \frac{1}{2}$ (from Gascuel and Steel, 2014 or Li, 2011). We can now state our main result which describes what happens when $\lambda/m > 4$, and whose proof is immediate from Lemmas 3.1 and 4.1.

Theorem 4.2. *Let M_t denote the probability that majority rule correctly infers the root state for a Yule tree grown at speciation rate λ for time t, and with a character evolved on this tree under a 2-state symmetric process with transition rate m, where $\lambda/m > 4$. Then for all $t \geq 0$*

$$M_t \geq \frac{1}{2} + \frac{1}{2}\left(\frac{1-4m/\lambda}{1-4me^{-rt}/\lambda}\right),$$

where the term on the right is monotone decreasing from 1 *(at* t=0*) to* $1-4m/\lambda$ *(as* $t \to \infty$*). In particular, for all finite* $t \geq 0$:

$$M_t > \frac{1}{2} + \frac{1}{2}\left(1 - \frac{4m}{\lambda}\right).$$

Acknowledgments

The authors thank the Simons Institute at UC Berkeley, where this work was carried out. E.M. would like to acknowledge the support of NSF (Grant nos. DMS 1106999 and CCF 1320105) and ONR (DOD ONR Grant N000141110140). M.S. would like to thank the NZ Marsden Fund and the Allan Wilson Centre for funding support. We also thank the two anonymous reviewers for helpful comments.

References

Bartoszek, K., Sagitov, S., 2013. Phylogenetic Confidence Intervals for the Optimal Trait Value. arxiv:1207.6488v3 [q-bio.PE].

Evans, W., Kenyon, C., Peres, Y., Schulman, L., 2000. Broadcasting on trees and the Ising model. Ann. Appl. Probab. 10 (2), 410–433.

Gascuel, O., Steel, M., 2010. Inferring ancestral sequences in taxon-rich phylogenies. Math. Biosci. 227, 125–135.

Gascuel, O., Steel, M., 2014. Predicting the ancestral character changes in a tree is typically easier than predicting the root state. Syst. Biol. 63 (3), 421–435.

Levin, D.A., Peres, Y., Wilmer, E.L., 2010. Markov Chains and Mixing Times. American Mathematical Society 2009.

Li, H., 2011. Ancestral Reconstruction: Comparing Majority Rule with Parsimony. Available from ⟨www.math.canterbury.ac.nz/bio/downloads/files/mathHonoursProjectpaperTLi.pdf⟩.

Liberles, D., 2007. Ancestral Sequence Reconstruction. Oxford University Press, New York.

Mossel, E., Peres, Y., 2003. Information flow on trees. Ann. Probab. 13 (3), 817–844.

Royer-Carenzi, M., Pontarotti, P., Didier, G., 2013. Choosing the best ancestral state reconstruction method. Math. Biosci. 242, 95–109.

Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J. Theor. Biol. 297, 33–40.

Yule, G.U., 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. Roy. Soc. Lond. Ser. B, 213: 21–87.

Elchanan Mossel
Department of Statistics, UC Berkeley, Berkeley, CA, USA

Mike Steel*
Biomathmatics Research centre, University of Canterbury,
Christchurch, New Zealand
E-mail address: mike.steel@canterbury.ac.nz

Received 10 April 2014

Available online 7 August 2014

* Corresponding author. Present address: School of Mathematics and Statistics, University of Canterbury, Christchurch 8140, New Zealand. Fax: +64 33642587.