Short Communication

Root location in random trees: A polarity property of all sampling consistent phylogenetic models except one

Mike Steel

Allan Wilson Centre for Molecular Ecology and Evolution, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

A B S T R A C T

Neutral macroevolutionary models, such as the Yule model, give rise to a probability distribution on the set of discrete rooted binary trees over a given leaf set. Such models can provide a signal as to the approximate location of the root when only the unrooted phylogenetic tree is known, and this signal becomes relatively more significant as the number of leaves grows. In this short note, we show that among models that treat all taxa equally, and are sampling consistent (i.e. the distribution on trees is not affected by taxa yet to be included), all such models, except one (the so-called PDA model), convey some information as to the location of the ancestral root in an unrooted tree.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Random neutral models of speciation (and extinction) have been a central tool for studying macroevolution, since the pioneering work of G.U. Yule in the 1920s (Yule, 1925). Such models typically provide a probability distribution on rooted binary trees for which the leaf set comprises some given subset of present-day taxa. The 'shape' of these trees has been investigated in various phylogenetic studies (see, for example, Blum and Francois, 2006) as it reflects properties of the underlying processes of speciation and extinction. Ignoring the branch lengths and considering just the topology of the trees provides not only a more tractable analysis, it also allows for a fortuitous robustness: several different processes (e.g. time- or density-dependent speciation and extinction rates) lead to the same probability distribution on discrete topologies, even though the processes are quite different when branch lengths are considered (Aldous, 1995).

In this short paper, we are concerned with the extent to which phylogenetic models for tree topology convey information as to where the tree is rooted. This is motivated in part by the fact that such models are used as priors in phylogenetic analysis (Jones, 2011; Velasco, 2008) and that sequence data analysed assuming the usual time-reversible Markov processes typically returns an unrooted tree (i.e. the location of the root is unknown). Many techniques attempt to estimate the root of the tree using the data and additional assumptions (e.g. a molecular clock or the inclusion of

an additional taxon that is known to be an 'outgroup'), or using properties of the tree that depend on branch length (e.g. 'midpoint rooting', where the tree is rooted in the middle of the longest path between any two leaves) (Boykin et al., 2010).

Here, we are interested in a much more basic question: what (if any) information the prior distribution on the topology alone itself might carry as to the location of the root of the tree. While it has been known that some models convey root-location information, we show here a stronger result – all models that satisfy two natural requirements (exchangeability and sampling consistency) have preferred root locations for a tree, except for one very special model. We begin by recalling some phylogenetic terminology.

2. Neutral phylogenetic models and their properties

Let $R(n)$ denote the finite set of rooted binary phylogenetic trees on the leaf set $[n] = \{1, 2, \ldots, n\}$ for $n \geqslant 2$; this set has size $(2n - 3)!! = 1 \times 3 \times 5 \times \cdots \times (2n - 3)$ (see, for example, Semple and Steel, 2003).

Given a tree $T \in R(n)$, let $T^{-\rho}$ denote the unrooted binary phylogenetic tree obtained by suppressing the root vertex ρ. This many-to-one association $T \mapsto T^{-\rho}$ is indicated in Fig. 1, where we have used $*$ to indicate the edge of $T^{-\rho}$ on which the root vertex ρ would be inserted in order to recover T. Notice that placing ρ at the midpoint of any edge of the tree on the right in Fig. 1 leads to a different rooted tree.

Thus, since an unrooted binary tree with n leaves has $2n - 3$ edges, if $B(n)$ denotes the set of unrooted binary phylogenetic trees on leaf set $[n]$ then $|R(n)| = (2n - 3)|B(n)|$.

E-mail address: mike.steel@canterbury.ac.nz
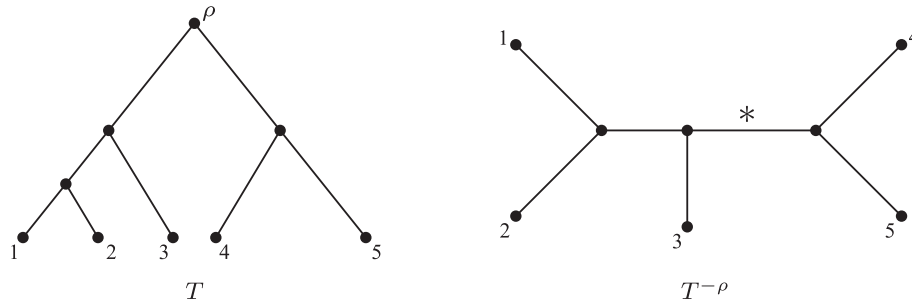
Fig. 1. One of the 105 trees in $R(5)$ and the associated unrooted tree $T^{-\rho}$ in $B(5)$ obtained from T by identifying the two edges incident with ρ by a single edge (indicated by $*$). Each of the seven edges of $T^{-\rho}$ corresponds to a different rooted tree.

Given a subset Y of $[n]$, and a (rooted or unrooted) binary phylogenetic tree T with leaf set $[n]$, let $T|Y$ denote the induced binary tree (root or unrooted, respectively) that connects the leaves in Y (for further details, see Semple and Steel, 2003).

Now suppose we have some random process for generating a rooted binary tree on leaf set $[n]$. We will denote the resulting randomly-generated tree as $\mathscr{T}_n$. Thus, $\mathscr{T}_n$ is an element of $R(n)$ while $\mathscr{T}_n^{-\rho}$ is an element of $B(n)$. To distinguish more clearly between rooted and unrooted trees, we will often write T_U to make it clear that we are referring to an unrooted tree.

A well-studied probability distribution on $R(n)$ is the Yule–Harding model (Harding, 1971; Yule, 1925), which can be described recursively as follows. Start with a tree with two (unlabelled) leaves, and repeatedly apply the following construction: For the tree constructed thus far, select a leaf uniformly at random and attach a new leaf to the edge incident with this leaf (by a new edge) and continue until the tree has n leaves. This produces a random rooted binary tree with n unlabelled leaves (sometimes referred to as a 'tree shape'). We then generate an element of $R(n)$ by assigning the elements of $[n]$ randomly to the leaves of this tree.

This probability distribution arises under quite general conditions, provided an exchangeability assumption is made. We describe this briefly here (for more detail, see Aldous, 1995). Consider any Markovian model of speciation and extinction in which the rates of these two events can either be constant or vary arbitrarily with time (and even depend on the past or on the number of lineages present). Then provided that each event (speciation or extinction) is equally likely to affect any one of the extant lineages at any given time, the resulting probability distribution on $R(n)$ is that described by the Yule–Harding model (moreover, this model also provides an equivalent distribution on tree topologies to that given by the Kingman coalescent model from population genetics, when, once again, branch lengths are ignored (Aldous, 1995; Zhu et al., 2011)).

A feature of the Yule–Harding model is that given just the associated unrooted tree $T^{-\rho}$, one can readily calculate the maximum likelihood (ML) estimate of the edge(s) of $T^{-\rho}$ on which the root node ρ was located (such ML edges are always incident with one of the (at most two) centroid vertices[1] of $T^{-\rho}$); moreover, the probability that an ML-edge contains the root node tends to a non-zero constant $(4 \log(4/3) - 1 \sim 0.15)$ as $n \to \infty$ (Steel and McKenzie, 2001). Indeed, if we consider the edges within three edges of this ML edge, the probability that at least one of them contains the root node is close to 0.9 (Steel and McKenzie, 2001). Thus, on a very large unrooted tree, T_U, we can isolate the likely location of the root to a relatively small proportion of edges (edges that are incident with, or near to the centroid vertex of T_U). Similar results concerning the

initial (root) vertex for a quite different model of tree growth were derived in 1970 (Haigh, 1970).

On the other hand, a model which provides no hint as to which edge in the associated unrooted tree might have contained the root is the *PDA model* (for 'proportional to distinguishable arrangements'), which is simply the uniform probability distribution on $R(n)$. This model is not directly described by a model of macroevolution involving speciation and/or extinction, in the same way that the Yule–Harding distribution is, although it is possible to derive the PDA distribution under somewhat contrived evolutionary assumptions. These include (i) conditioning on events such as a short window of opportunity for speciation and the survival of the tree to produce n leaves (Steel and McKenzie, 2001); (ii) the tree on n species sampled randomly from the leaves of a large tree generated by a conditioned critical branching process (Aldous, 1995); and (iii) a general conditional independence assumptions regarding parent and daughter branches in a tree (Pinelis, 2003).

Notice that both the PDA and Yule–Harding models comprise a *family* of probability distributions (i.e. they provide a probability distribution on $R(n)$ for each $n \geqslant 2$). Thus we will refer to any such sequence $(p_n : n \geqslant 2)$ of probability distributions on $(R(n) : n \geqslant 2)$ as a *phylogenetic model* $\mathscr{M}$ for $(R(n) : n \geqslant 2)$, and we will also write $\mathbb{P}(\mathscr{T}_n = T)$ for the probability $p_n(T)$ when $T \in R(n)$.

We now list three properties that any family of probability distributions on $(R(n) : n \geqslant 2)$ can posses. The first two (the exchangeability property (EP) and sampling consistency (SC)) are satisfied by several models (including the PDA and Yule–Harding model), while the third, root invariance (RI), does not hold for the Yule–Harding model, as we saw above, however it does hold for the PDA model.

EP [Exchangeability property] For each $n \geqslant 2$, if T and T' are trees in $R(n)$ and T' is obtained from T by permuting its leaf labels, then $p_n(T') = p_n(T)$.

SC [Sampling consistency] For each $n \geqslant 2$, and each tree T in $R(n)$ we have:

$$\mathbb{P}(\mathscr{T}_{n+1}|[n] = T) = p_n(T).$$

RI [Root invariance] For each $n \geqslant 2$, if T and T' are trees in $R(n)$ and they are equivalent up to the placement of their roots (i.e. $T^{-\rho} = T'^{-\rho}$) then $p_n(T') = p_n(T)$.

The exchangeability property (EP), from Aldous (1995), requires the probability of a particular phylogenetic tree to depend just on its shape and not on how its leaves are labelled (this property is called 'label-invariance' in Steel and Penny (1993)).

The sampling consistency (SC) property, also from Aldous (1995), states that the probability distribution on trees for a given set of taxa does not change if we add another taxon (namely, $n + 1$) and consider the induced marginal distribution on the original set of taxa. The condition seems reasonable if we wish the model to be

---

[1] A vertex v of an unrooted tree T_U is a *centroid vertex* if each of the subtrees of T_U obtained by deleting vertex v from T_U contain at most half the vertices of T_U (Kang and Ault, 1975).

'stable' in that sense that, in the absence of any data associated with the taxa, p_n should depend just on the taxa present, and not on taxa yet to be discovered (or not included in the set of taxa under study).[2] An equivalent statement of the (SC) condition is that the phylogenetic model describing the distribution of trees on *all* the taxa also correctly describes the distribution of trees when we restrict these trees to *any* sample of taxa specified in advance (i.e. selected without reference to the tree on all the taxa).

The root-invariance (RI) property states that the model does not prefer any particular rooting of a tree (i.e. any re-rooting of the tree would have equal probability).

Note that the three properties EP, SC and RI pertain to distributions on trees in which the taxa have yet to have any biological data associated with them (i.e. they are 'prior' to considering any particular data). If the taxa come with data that is used to construct a probability distribution on trees, then clearly EP will often not hold, and SC could also fail, since data provided by an additional taxon can influence the relative support for trees on an existing set of taxa. RI may or may not fail, depending on the assumptions of the model (e.g. whether or not it is time-reversible, or whether or not a molecular clock is imposed). We now consider pairwise combinations of these three properties.

[EP + SC] Several families of probability distributions on $(R(n) : n \geqslant 2)$ satisfy both EP and SC, including the 'β-splitting model', a one-parameter family that was described by Aldous (1995) and which includes, as special cases, the Yule–Harding model (when $\beta = 0$) and the PDA model (when $\beta = -3/2$). Other phylogenetic models satisfying SC have also been studied recently in Jones (2011). Models satisfying EP and SC have been studied (and characterised) recently by Haas et al. (2008) and McCullagh et al. (2008).

[EP + RI] If we consider the combination of EP and RI several possible distributions on $R(n)$ satisfy these two properties, since one may simply define *any* probability distribution on unrooted binary tree shapes, and extend this to labelled and rooted trees by imposing RI and EP.

[SC + RI] Finally, consider the combination of SC and RI. The PDA distribution satisfies these two properties, and, as noted already, it also satisfies EP.

The point of this short note is to show that, unlike the other two combinations of properties, apart from the PDA model, there is no other phylogenetic model that satisfies this last combination of SC and RI. This 'impossibility' result is of similar spirit to (but is quite unrelated to) the result of Velasco (2007) concerning phylogenetic models that are uniform on clades of all given sizes.

3. Results

We now state the main result of this short note, the proof of which is given in Appendix A.

Theorem 1. *A phylogenetic model $\mathcal{M} = (p_n : n \geqslant 2)$ for $(R(n) : n \geqslant 2)$ satisfies the two properties SC and RI if and only if $\mathcal{M}$ is the PDA model.*

The relevance of this theorem is that the PDA model does not describe the shape of most published phylogenetic trees derived from biological data very well, as the latter trees are typically more balanced than the PDA model predicts (Aldous, 1995; Aldous et al., 2011; Blum and Francois, 2006). Moreover, as noted already, the PDA model does not have a compelling biological motivation. Thus, the significance of Theorem 1 is that any 'biologically realistic' sampling consistent distribution on discrete rooted phylogenetic trees necessarily favours some root locations over others in the associated unrooted tree topology. And this holds without knowledge of the branch lengths or, indeed, of any data.

As a corollary of Theorem 1, the only value of β for which the β-splitting model satisfies sampling consistency and root invariance is $\beta = -3/2$, which corresponds to the PDA model. For other values of β, it may be of interest to determine how accurately one can estimate the exact (or approximate) location of the edge of an unrooted tree that contained the root node, when the rooted tree evolved under the β-splitting model.

Notice that Theorem 1 does not require EP to hold at any point, however it falls out as a second consequence of the theorem that any phylogenetic model that satisfies SC and RI must also satisfy EP (since this holds for the PDA model). The conditions EP and SC also apply to phylogenetic models for *unrooted* trees – simply replace $R(n)$ with $B(n)$ in their definition – and the reader may wonder whether Theorem 1 is merely a consequence of a result that states that any phylogenetic model on unrooted trees that satisfies SC is uniform. Such a result, if true, would indeed imply Theorem 1, but such a result does not hold, even if we append condition EP to SC; a simple counterexample is the probability distribution on unrooted trees induced by the Yule–Harding model.

Finally, notice that Theorem 1 is not merely a consequence of RI alone, as it does not hold if condition SC is dropped, or replaced by EP. This is because, as noted already, there exists phylogenetic models that satisfy EP + RI, and that are different from the PDA model.

3.1. Acknowledgments

Appendix A. Proof of Theorem 14

Since the PDA model clearly satisfies RI and SC, it remains to establish the 'only if' claim. To this end, we first establish the following: for any phylogenetic model $\mathcal{M} = (p_n : n \geqslant 2)$ for $(R(n) : n \geqslant 2)$ that satisfies SC, and any $T_U \in B(n)$, the following identity holds:

$$\mathbb{P}(\mathscr{T}_{n+1}^{-\rho}|[n] = T_U) = \mathbb{P}(\mathscr{T}_n^{-\rho} = T_U). \qquad (1)$$

To establish (1), first observe that for any tree T_{n+1} in $R(n + 1)$, one has $T_{n+1}^{-\rho}|[n] = (T_{n+1}|[n])^{-\rho}$. Thus,

$$\mathbb{P}(\mathscr{T}_{n+1}^{-\rho}|[n] = T_U) = \mathbb{P}((\mathscr{T}_{n+1}|[n])^{-\rho} = T_U)$$
$$= \sum_{T \in R(n) : T^{-\rho} = T_U} \mathbb{P}(\mathscr{T}_{n+1}|[n] = T),$$

and, by SC, we can express this as:

$$\sum_{T \in R(n) : T^{-\rho} = T_U} \mathbb{P}(\mathscr{T}_n = T) = \mathbb{P}(\mathscr{T}_n^{-\rho} = T_U),$$

which establishes the claimed identity (1).

Returning to the proof of Theorem 1, observe that, by RI, we can describe the probability distribution p_n as being equivalent to the one in which we first generate an unrooted phylogenetic tree

---

$T_U \in B(n)$ with some associated probability $q(T_U)$ and then select one of the edges of T_U uniformly at random to subdivide as the root vertex. We will refer to this second (uniform) process as the *root-edge selection* process. Thus, for any $T_U \in B(n)$, we have:

$$q(T_U) = \mathbb{P}(\mathscr{T}_n^{-\rho} = T_U), \qquad (2)$$

and, for any $T \in R(n)$, we have, from RI that:

$$\mathbb{P}(\mathscr{T}_n = T) = \frac{1}{(2n-3)} q(T^{-\rho}). \qquad (3)$$

Let $b(n) = |B(n)|$ (i.e. $b(n) = (2n-5)!! = |R(n-1)|$). We will show by induction on n that q is uniform on $B(n)$ for all $n \geqslant 2$ (i.e. $q(T_U) = \frac{1}{b(n)}$ for all $T_U \in B(n)$ and all $n \geqslant 2$). This induction hypothesis holds for $n = 2$, so supposing that it holds for $n \geqslant 2$, we will use this to show that q is uniform on $B(n+1)$. First observe that, for $T_U \in B(n)$, we can apply identity (1), since SC holds, to deduce that:

$$\mathbb{P}(\mathscr{T}_{n+1}^{-\rho}|[n] = T_U) = q(T_U) = \frac{1}{b(n)}, \qquad (4)$$

where the second equality holds by the induction hypothesis.

Now, for each edge e of T_U consider the tree $T_U^e \in B(n+1)$ obtained from T_U by attaching leaf $n+1$ to the midpoint of edge e with a new edge. Let $p(e) = \mathbb{P}(A_e|B_{T_U})$ where A_e, B_{T_U} denote the nested events defined by:

$$A_e := \text{`}\mathscr{T}_{n+1}^{-\rho} = T_U^e\text{'} \text{ and } B_{T_U} := \text{`}(\mathscr{T}_{n+1}^{-\rho})|[n] = T_U\text{'}.$$

Notice that Eqn. (4) says that B_{T_U} has uniform probability on $B(n)$, and we will now show that p (the conditional probability of A_e given B_{T_U}) also has uniform probability across the edge of T_U; this will allow us to establish the induction hypothesis.

To this end, recall that on T_U^e each edge has a uniform root-edge selection probability by RI. This implies (by SC) that the root-edge selection process on T_U will select e with the following probability:

$$p(e)\frac{3}{2n-1} + (1-p(e))\frac{1}{2n-1}. \qquad (5)$$

In this expression, the term $2n-1$ in the denominator is simply the number of edges of T_U^e, the numerator term 3 corresponds to the three edges of T_U^e consisting of the edge incident with $n+1$ and the two edges incident with that edge. Since the root-edge selection process on the edges of T_U is also uniform, and this tree has $2n-3$ edges, it follows from the expression in (5) that:

$$p(e)\frac{3}{2n-1} + (1-p(e))\frac{1}{2n-1} = \frac{1}{2n-3}.$$

Thus, $p(e) = \frac{1}{(2n-3)}$, and so p is the uniform distribution on the edges of T_U, as claimed.

Finally, observe that the uniformity of p now entails that q is uniform on $B(n+1)$ since for any tree $T'_U \in B(n+1)$, we can select $T_U \in B(n)$ and an edge e of T_U for which $T'_U = T_U^e$ and then:

$$q(T'_U) = q(T_U^e) = \mathbb{P}(\mathscr{T}_{n+1}^{-\rho} = T_U^e) = \mathbb{P}(A_e)$$
$$= \mathbb{P}(A_e \& B_{T_U}) = \mathbb{P}(A_e|B_{T_U})\mathbb{P}(B_{T_U}),$$

and thus:

$$q(T'_U) = \frac{1}{2n-3} \cdot \frac{1}{b(n)} = \frac{1}{b(n+1)}.$$

Thus, we have established the induction step required to show that q is uniform on $B(n)$ for all $n \geqslant 2$. It now follows from Eq. (3) that p_n is uniform on $R(n)$, for all $n \geqslant 2$, which completes the proof of Theorem 1. $\square$

References

Aldous, D., 1995. Probability distributions on cladograms. In: Aldous, D., Pemantle, R.E. (Eds.), Random Discrete Structures. IMA Volumes in Mathematics and its Applications, vol. 76. Springer, pp. 1–18.

Aldous, D., Krikun, M., Popovic, L., 2011. Five statistical questions abou the tree of life. Syst. Biol. 60, 318–328.

Blum, M., Francois, O., 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55, 685–691.

Boykin, L., Kubatko, L., Lowrey, T., 2010. Comparison of methods for rooting phylogenetic trees: a case study using Orcuttieae (Poaceae: Chloridoideae). Mol. Phylogenet. Evol. 54 (3), 687–700.

Haas, B., Miermont, G., Pitman, J., Winkel, M., 2008. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. Ann. Probab. 36 (5), 1790–1837.

Haigh, J., 1970. The recovery of the root of a tree. J. Appl. Prob. 7, 79–88.

Harding, E., 1971. The probabilities of rooted tree shapes generated by random bifurcation. Adv. Appl. Prob. 3, 44–77.

Jones, G., 2011. Tree models for macroevolution and phylogenetic analysis. Syst. Biol. 60 (6), 735–746.

Kang, A., Ault, D., 1975. Some properties of a centroid of a free tree. Inf. Proc. Lett. 4, 18–20.

McCullagh, P., Pitman, J., Winkel, M., 2008. Gibbs fragmentation trees. Bernoulli 14 (4), 988–1002.

Pinelis, I., 2003. Evolutionary models of phylogenetic trees. Proc. Roy. Soc. Lond. (B) 270 (1522), 1425–1431.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press.

Steel, M., McKenzie, A., 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170, 91–112.

Steel, M., Penny, D., 1993. Distributions of tree comparison metrics – some new results. Syst. Biol. 42, 126–141.

Velasco, J., 2007. Why non-uniform priors on clades are both unavoidable and unobjectionable. Mol. Phylogenet. Evol. 45, 748–749.

Velasco, J., 2008. The prior probabilities of phylogenetic trees. Biol. Philos. 23, 455–547.

Yule, G., 1925. A mathematical theory of evolution.Based on the conclusions of Dr. J.C. Willis, F.R.S. Philos. Trans. Roy. Soc. 213, 21–87.

Zhu, S., Degnan, J., Steel, M., 2011. Clades, clans and reciprocal monophyly under neutral evolutionary models. Theor. Popul. Biol. 79, 220–227.