Probabilities of Evolutionary Trees under a Rate-Varying Model of Speciation

Mike Steel

Biomathematics Research Centre
University of Canterbury,
Private Bag 4800
Christchurch, New Zealand

No. 167 December, 1998

Abstract

We analyse a simple model of biological speciation where the speciation rate for each species s depends on the time to the last speciation event involving s. We present recursions for calculating the probabilities of evolutionary trees under this model.

Keywords: Speciation, tree shapes, pure birth process.

Probabilities of Evolutionary Trees under a Rate-Varying Model of Speciation

December 14, 1998

Mike Steel[1]
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand

Abstract

We analyse a simple model of biological speciation where the speciation rate for each species s depends on the time to the last speciation event involving s. We present recursions for calculating the probabilities of evolutionary trees under this model.

Keywords: Speciation, tree shapes, pure birth process.

1 Introduction

Simple random models of speciation have been proposed in biology, and there is considerable interest in testing and refining such models by comparing their predictions with published phylogenetic trees ([1], [2], [3], [4], [5], [7], [9]). Such models make predictions about the shape of the (discrete) evolutionary tree connecting the extant species. In the simplest such model, at any time each existing species has the same probability of giving rise to a new species, and all lineages are treated independently. Here we consider a simple modification of this model, in which the rate of speciation events on a given lineage is a function of the time back to the last speciation event on that lineage.

More precisely, we suppose that at time $t = 0$ there is just one species, labelled s_0, subject to a 2-state Markov process on state space $\{1, 2\}$. Under this process, s_0 is initially in state 1, and state 2 corresponds to a "speciation event", that is, the replacement of the original species by two species (either two new species, or the original species plus one new one, and we will not distinguish here between these two possibilities). Let $s(t)$ denote the rate of change from state 1 to state 2 at time t, we call this the "speciation rate". Once a speciation event occurs (say at time Λ) the two species are again assumed to be independently subject to the same Markov process, with time reset to 0 (that is, with intensity function $s(t - \Lambda)$). Continuing in this way, we obtain a probability distribution on the trees of descent of species starting from s_0 up to some fixed time t which we can assume (by rescaling s if necessary) lies in the range $[0, 1]$.

The biological motivation for this model is that a recently evolved species, or the species that it has split off from, are often colonizing new regions or niches, and so may be more likely to give rise to further new species (in a given short time period) than a species that has existed for a very long

[1]Supported by the New Zealand Marsden Fund (M1019)

time without giving rise to any new species (thus we are thinking of s being a monotone decreasing function). It would also be interesting and useful to build extinctions into such a model, however we do not pursue this here.

Kubo and Iwasa [7] consider a rate-varying model of speciation, however in their case, the speciation rate is a function of (absolute) time, rather than the lineage-specific time back to the last speciation event. Our model has more similarity to that discussed by Heard [5] who used computer similation rather than analytical techniques in his analysis. We are interested in the probability distribution that this model induces on the tree that describes the species descendent from s_0. Since we are only interested in the "shape" of these speciation trees, we will mostly deal with trees in which the vertices are unlabelled.

2 Formulae

Definitions

- For $n \geq 1$, let $UB(n)$ denote the (finite) set of unlabelled binary trees consisting of n leaves (vertices of degree 1) together with an additional leaf, the *root leaf*, as in Fig. 1a (where the root leaf is the top-most vertex), and whose remaining internal vertices are all of degree 3.
- We say a vertex v (resp. a subtree) is a *descendant* of another vertex w, if v lies on the path between w (resp. the subtree) and s_0.
- For $n \geq 2$, let $URB(n)$ denote the (finite) set of unlabelled edge-rooted trees obtained from $UB(n)$ by deleting from each tree the root leaf and its incident edge. If $\tau \in UB(n)$ we will let τ^* denote the associated tree in $URB(n)$ (as in Fig. 1b).
- For the model described above, the speciation tree at time $t \in [0,1]$, $T(t)$, is the unlabelled tree of descent of the species that have evolved up to time t from the root leaf s_0.
- For a tree $\tau \in UB(n)$ with root leaf v and adjacent vertex v' let τ_1 and τ_2 denote the two subtrees of τ whose two vertex sets (i) intersect precisely on v' and (ii) cover all vertices of τ except v, as in Fig. 1c.
- For $0 \leq t \leq 1$ and $\tau \in UB(n)$, consider the following (absolute and conditional) probabilities

$$f(\tau,t) := \mathbb{P}[T(t) = \tau];\ p(\tau) := \mathbb{P}[T(1) = \tau | T(1) \text{ has } n \text{ leaves}],$$

- Let $\Lambda(s_0)$ denote the time until speciation of s_0, and set

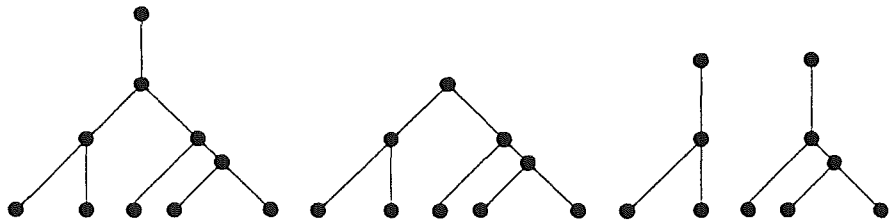
$$S(x) := \mathbb{P}[\Lambda(s_0) \geq x];\ \sigma(x) := s(x)S(x).$$

Since the speciation of s_0 is a time-dependent Poisson process we have, from [8]

$$\mathbb{P}[\Lambda(s_0) \geq x] = exp[-\int_0^x s(\lambda)d\lambda].$$

Thus, $\sigma(x) = lim_{\delta \to 0+} \frac{\mathbb{P}[\Lambda(s_0) \in (x, x+\delta)]}{\delta}$ and so, by the assumptions that define the model, we have the following fundamental recursion:

$$f(\tau,t) = 2^{\delta(\tau)} \int_0^t f(\tau_1, t-x) f(\tau_2, t-x) \sigma(x) dx \tag{1}$$



(a) (b) (c)

Figure 1: Unlabelled trees: (a) $\tau \in UB(5)$; (b) $\tau^* \in URB(5)$; and (c) τ_1, τ_2

where

$$\delta(\tau) = \begin{cases} 1 & \text{if } \tau_1 \neq \tau_2 \\ 0 & \text{otherwise,} \end{cases}$$

Let $N(t)$ denote the total number of species existing at time $t \in [0,1]$, and let

$$\nu(k,t) := \mathbb{P}[N(t) = k].$$

For $\tau \in UB(n)$ we wish to calculate the conditional probability:

$$p(\tau) = \mathbb{P}[T(1) = \tau | N(1) = n] = \frac{f(\tau,1)}{\nu(n,1)}. \tag{2}$$

The number $\nu(n,1)$ appearing in Equation (2) is given by:

$$\nu(n,1) = \sum_{\tau \in UB(n)} f(\tau,1)$$

however, the number of terms in this summation grows exponentially with n. Thus, we also give a simple recursion for computing the functions $\nu(1,t), \ldots, \nu(k,t)$ and thereby the number $\nu(n,1)$, as follows:

$$\nu(1,t) = S(t)$$

$$\nu(k,t) = \sum_{i=1}^{k-1} \int_0^t \nu(i, t-x)\nu(k-i, t-x)\sigma(x)dx.$$

We may also wish to compute the probability of the induced edge-rooted tree. Thus, given $\tau \in UB(n)$ and its associated tree $\tau^* \in URB(n)$ let:

$$p(\tau^*) := \lim_{\epsilon \to 0+} \mathbb{P}[T(1) = \tau | N(1) = n; \Lambda(s_0) < \epsilon].$$

The motivation for considering $p(\tau^*)$ is that one is frequently interested in the distribution on edge-rooted trees, and we can simplify matters by supposing that the first speciation event happened at time 0. We have the recursion:

$$p(\tau^*) = 2^{\delta(\tau)} p(\tau_1) p(\tau_2). \tag{3}$$

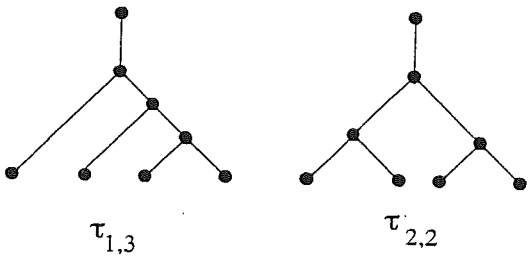


Figure 2: The two unlabelled binary trees on four leaves

Note that if we wish to compare the probability ratios of two trees then we can dispense with the function ν altogether, since $\frac{p(\tau)}{p(\tau')} = \frac{f(\tau,1)}{f(\tau',1)}$. For $i \in \{1,2,3\}$, let $\{\tau_i\} = UB(i)$. We have:

$$f(\tau_1, t) = S(t)$$

$$f(\tau_2, t) = \int_0^t S(t-x)^2 \sigma(x) dx$$

$$f(\tau_3, t) = 2\int_0^t S(t-x) f(\tau_2, t-x)\sigma(x) dx.$$

For the (only) two trees $\tau_{1,3}$ and $\tau_{2,2}$ in UB(4), as shown in Fig. 2, we have:

$$f(\tau_{1,3}, t) = 2\int_0^t S(t-x) f(\tau_3, t-x)\sigma(x) dx$$

and

$$f(\tau_{2,2}, t) = \int_0^t f(\tau_2, t-x)^2 \sigma(x) dx.$$

Thus we can obtain an explicit expression for the ratio of the probabilities of $\tau_{2,2}$ and $\tau_{1,3}$, and even simpler formulae for corresponding rooted trees, by a further application of Equations (1) and (3). This is summarized in Theorem 1.

Theorem 1

- $\frac{p(\tau_{2,2})}{p(\tau_{1,3})} = \frac{\int_0^1 \{\int_0^{1-x} S(1-x-s)^2\sigma(s)ds\}^2\sigma(x)dx}{4\int_0^1 S(1-x)\sigma(x)\{\int_0^{1-x} S(1-x-s)\sigma(s)\{\int_0^{1-x-s} S(1-x-s-r)^2\sigma(r)dr\}ds\}dx}$
- $\frac{p(\tau^*_{2,2})}{p(\tau^*_{1,3})} = \frac{\{\int_0^1 S(1-x)^2\sigma(x)dx\}^2}{4S(1)\int_0^1 S(1-x)\sigma(x)\{\int_0^{1-x} S(1-x-r)^2\sigma(r)dr\}dx}$

3 Two classes of models

1. The simplest model has $s(x) = s > 0$, a constant. For this (Yule) model the associated probability distribution on trees is described in [3] (see also [1] and [2]). In this case, $\sigma(x) = se^{-sx}$ and $N(t)$ models a pure birth process, so $\nu(k,t) = e^{-st}(1-e^{-st})^{k-1}$. Under this model, $p(\tau_{2,2}) = p(\tau^*_{2,2}) = 1/3$, and, more generally, $p(\tau) = p(\tau^*) = 2^{u(\tau)}\prod_{i>2}(i-1)^{-d_i(\tau)}$, where $d_i(\tau)$ denotes the number of internal vertices of τ which have exactly i descendant leaves, and $u(\tau)$ is the number of *unbalanced* internal vertices of τ - that is, internal vertices for which the two descendant subtrees are not identical.

2. A second class of models are those which satisfy the condition:

$$s(x) = 0 \text{ for } x > \epsilon,$$

which we will call "explosive radiation" models. In these model, unless a species has undergone a speciation event within the last ϵ time interval, it will never do so. Thus, in this model, speciation events would tend to be clustered close together. We now analyse this model, and show that, provided epsilon is sufficiently small, then this model is precisely that induced by a uniform distribution on leaf-labelled trees. This distribution on trees also arises under a conditioned critical Galton-Watson process -see [1] and the references therein.

We now describe this uniform model. For $\tau \in UB(n)$, let $L(\tau)$ be the set of distinct trees that can be obtained by assigning the (species) labels $\{1, \ldots, n\}$ bijectively to the n non-root leaves of τ. Let $LB(n) := \cup_{\tau \in UB(n)} L(\tau)$. Under the *uniform model* a tree is selected uniformly from $LB(n)$, and then it is viewed as an unlabelled tree $\tau \in UB(n)$. Thus $p_{unif}(\tau) = \frac{|L(\tau)|}{|LB(n)|}$. Fortunately, the numerator and denominator of this ratio can both be evaluated exactly, and so we get an explicit formula for $p_{unif}(\tau)$ as follows. We have $|L(\tau)| = n!2^{-b(\tau)}$ where $b(\tau)$ is the number of *balanced* internal vertices of τ–that is, internal vertices for which the two descendant subtrees are identical (and so $b(\tau)+u(\tau) = n-1$). Now, $|LB(n)| = (2n-3)!! = (2n-3)\times(2n-5)\times\ldots\times 3\times 1 = \frac{(2n-2)!}{(n-1)!2^{n-1}}$, and therefore, under the uniform model,

$$p_{unif}(\tau) = \frac{|L(\tau)|}{|LB(n)|} = n \begin{pmatrix} 2n-2 \\ n-1 \end{pmatrix}^{-1} 2^{u(\tau)}. \tag{4}$$

Theorem 2 *Under an explosive radiation model, with $\epsilon < 1/n$, the probability distribution on trees is precisely that induced by the uniform model. That is,*

$$p(\tau) = p_{unif}(\tau), \forall \tau \in UB(n).$$

Proof We use induction on n to establish the following:

$$CLAIM: \text{ if } \tau \in UB(n), \text{ then } f(\tau, t) = c(n)2^{u(\tau)} \text{ for } t > n\epsilon,$$

where $c(n) = e^{-n\int_0^\epsilon s(\lambda)d\lambda}(1 - e^{-\int_0^\epsilon s(\lambda)d\lambda})^{n-1}$.

The claim clearly holds for $n = 1$, since in this case, if $t > \epsilon$,

$$f(\tau, t) = S(t) = e^{-\int_0^t s(\lambda)d\lambda} = e^{-\int_0^\epsilon s(\lambda)d\lambda}.$$

Now suppose the result holds for $n = k \geq 1$, and let $\tau \in UB(k+1)$. Then, from Equation (1) and the fact that $s(x)$ is zero for $x > \epsilon$, we have:

$$f(\tau, t) = 2^{\delta(\tau)} \int_0^\epsilon f(\tau_1, t-x) f(\tau_2, t-x)\sigma(x)dx.$$

For $i = 1, 2$, let k_i denote the number of leaves of τ_i (thus, $k_1 + k_2 = k + 1$). If $t > (k+1)\epsilon$, and $x < \epsilon$, we have $t - x > k\epsilon \geq k_i\epsilon$ (since $k_1, k_2 \leq k$). Thus we may apply the induction hypothesis to $f(\tau_1, t-x)$ and $f(\tau_2, t-x)$ over the range of integration and deduce that:

$$f(\tau, t) = 2^{\delta(\tau)} \int_0^\epsilon c(k_1)2^{u(\tau_1)}c(k_2)2^{u(\tau_2)}\sigma(x)dx = 2^{u(\tau)}c(k_1)c(k_2)\int_0^\epsilon \sigma(x)dx = 2^{u(\tau)}c(k+1)$$

by the definition of the function c. By Equation (2), $p(\tau) = f(\tau, 1)/\nu(n, 1)$ and therefore, since $\epsilon < 1/n$, we can apply the above claim to deduce that $p(\tau) = c^*(n)2^{u(\tau)}$ for a function c^* that depends only on n and perhaps the function s. However, it is easy to show that c^* does not depend on s at all, and that it must equal $c'(n) := n \begin{pmatrix} 2n-2 \\ n-1 \end{pmatrix}^{-1}$, since we have, from Equation (4),

$$c'(n) \sum_{\tau \in UB(n)} 2^{u(\tau)} = \sum_{\tau \in UB(n)} p_{unif}(\tau) = 1 = \sum_{\tau \in UB(n)} p(\tau) = c^*(n) \sum_{\tau \in UB(n)} 2^{u(\tau)},$$

and thus $c'(n) = c^*(n) = 1/\sum_{\tau \in UB(n)} 2^{u(\tau)}$. □

3.1 Acknowledgements

The author thanks Andy McKenzie and Charles Semple for helpful comments on an earlier version of this manuscript.

References

[1] Aldous, D. (1996), Probability distributions on cladograms, in: D. Aldous, R. Permantle, eds. *Random Discrete Structures* (IMA Volume on Mathematics and its Applications Vol. 76 Springer) pp. 1-18.

[2] Brown, J.K.M. (1994), Probabilities of evolutionary trees, *Syst. Biol.* **43(1)**, 78-91.

[3] Harding, E.F. (1971), The probabilities of rooted tree-shapes generated by random bifurcation, *Adv. Appl. Prob.* **3**, 44-77 .

[4] Heard, S.B. (1992), Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, *Evolution*, **46(6)**, 1818-1826.

[5] Heard, S.B. (1996), Patterns in phylogenetic tree balance with variable and evolving speciation rates, *Evolution*, **50(6)**, 2141-2148.

[6] Hendy, M.D. and Penny, D. (1982), Branch and bound algorithms to determine minimal evolutionary trees, *Math. Biosci.* **59**, 277-290.

[7] Kubo, T. and Iwasa, Y. (1995), Inferring the rates of branching and extinction from molecular phylogenies, *Evolution* **49(4)**, 694-704.

[8] Medhi, J. (1982), *Stochastic Processes* (John Wiley and Sons Ltd).

[9] Mooers, A. O. and Heard, S.B. (1997), Inferring evolutionary process from phylogenetic tree shape, *Quart. Rev. Biol.* **72(1)**, 31-54.

[10] Steel, M.A. (1988), Distribution of the symmetric difference metric on phylogenetic trees, *SIAM J. Discr. Math.* **1(4)**, 541-551.

[11] Steel, M. and Penny, D. (1993), Distribution of tree comparison metrics - some new results, *Syst. Biol.* **42(2)**, 126-141.

[12] Thompson, E. A. (1975), *Human evolutionary trees* (Cambridge Univ. Press, Cambridge, England).