Journal of Theoretical Biology 364 (2015) 275–283

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

On the distribution of interspecies correlation for Markov models of character evolution on Yule trees

Willem H. Mulder [a], Forrest W. Crawford [b,*]

[a] Department of Chemistry, The University of the West Indies, Mona Campus, Kingston 7, Jamaica
[b] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

HIGHLIGHTS

- We investigate Markov models of character evolution on Yule trees.
- We derive the marginal distribution of pairwise interspecies covariance.
- We derive the distribution of the number of segregating sites on a Yule tree.
- A new measure of phylogenetic information is proposed for trees with n tips.

ARTICLE INFO

Article history:
Received 7 June 2014
Received in revised form
15 August 2014
Accepted 9 September 2014
Available online 18 September 2014

Keywords:
Infinite sites
Mutation model
Phylogenetics
Phylogenetic informativeness
Yule process

ABSTRACT

Efforts to reconstruct phylogenetic trees and understand evolutionary processes depend fundamentally on stochastic models of speciation and mutation. The simplest continuous-time model for speciation in phylogenetic trees is the Yule process, in which new species are "born" from existing lineages at a constant rate. Recent work has illuminated some of the structural properties of Yule trees, but it remains mostly unknown how these properties affect sequence and trait patterns observed at the tips of the phylogenetic tree. Understanding the interplay between speciation and mutation under simple models of evolution is essential for deriving valid phylogenetic inference methods and gives insight into the optimal design of phylogenetic studies. In this work, we derive the probability distribution of interspecies covariance under Brownian motion and Ornstein–Uhlenbeck models of phenotypic change on a Yule tree. We compute the probability distribution of the number of mutations shared between two randomly chosen taxa in a Yule tree under discrete Markov mutation models. Our results suggest summary measures of phylogenetic information content, illuminate the correlation between site patterns in sequences or traits of related organisms, and provide heuristics for experimental design and reconstruction of phylogenetic trees.

1. Introduction

Simple stochastic models of speciation and trait evolution have proven useful for reconstruction of phylogenetic trees describing the ancestral relationship between sets of taxa. The simplest continuous-time model of speciation is the Yule process, in which each extant lineage gives birth at constant rate λ. A Yule tree is a phylogenetic tree in which the branching times of the tree are drawn from the Yule distribution. Despite the apparent simplicity of the Yule process, Yule trees have complex structural properties (Steel and McKenzie, 2002; Rosenberg, 2006; Gernhard et al., 2008; Steel and Mooers, 2010; Mulder, 2011; Crawford and Suchard, 2013). The Yule process is usually employed as a prior or null distribution on the space of phylogenetic trees within a broader scheme of phylogenetic reconstruction (Nee et al., 1994; Rannala and Yang, 1996; Nee, 2006). Researchers impose a model for the evolution of a character (trait, DNA, RNA, or amino acid sequence) on the branches of this phylogenetic tree. By jointly estimating the phylogenetic tree topology, branch lengths, and the parameters underlying the evolutionary model, researchers hope to understand the evolutionary history and process that gave rise to the observed data.

Research on the interaction of tree topology, branch lengths, and evolutionary processes generally falls into one of two categories. The first is the search for better measures of phylogenetic information for *prospective* experimental design. Most of these studies examine the probability of correctly reconstructing a simple tree or optimal design

* Correspondence to: Department of Biostatistics, Yale School of Public Health, PO Box 208034 New Haven, CT 06510, CT, USA. Tel.: (203) 785 6125.
E-mail addresses: willem.mulder@uwimona.edu.jm (W.H. Mulder), forrest.crawford@yale.edu (F.W. Crawford).
URL: http://crawford.research.yale.edu (F.W. Crawford).

http://dx.doi.org/10.1016/j.jtbi.2014.09.016
0022-5193/

of phylogenetic studies (Yang, 1998; Sullivan et al., 1999; Shpak and Churchill, 2000; Zwickl and Hillis, 2002; Susko et al., 2002). Several authors have attempted to determine whether it is better to add more taxa or additional characters to maximize the chance of reconstructing the correct tree (Graybeal, 1998; Zwickl and Hillis, 2002). Steel and Penny (2000) analyze basic models of evolution to understand the theoretical properties of stochastic models on phylogenetic trees. Fischer and Steel (2009) consider asymptotic sequence length bounds for correct reconstruction under maximum parsimony. Townsend (2007) introduces "phylogenetic informativeness", the probability of observing site patterns allowing correct reconstruction of a four-taxon tree. Susko (2011) and Susko and Roger (2012) find expressions for correct reconstruction probability for small internal edges on four-taxon trees. Real-world phylogenetic studies often involve large numbers of taxa, and it remains controversial whether properties of mutation models on four-taxon trees generalize to trees with larger numbers of taxa (see e.g. Townsend, 2007; Klopfstein et al., 2010; Townsend and Leuenberger, 2011).

The second class of approaches focuses on *retrospective* inferences about evolutionary parameters and the derivation of estimators and confidence intervals. Following the work of Stadler (2009), who describes sampling properties of birth–death trees and the distribution of the age of the most recent common ancestor (MRCA) of subsets of randomly chosen taxa, Bartoszek and Sagitov (2012) and Bartoszek (2013) find expressions for the expectation of the interspecies correlation under models of continuous trait evolution via diffusion and Ornstein–Uhlenbeck processes. Bartoszek and Sagitov (2012) derive asymptotic confidence intervals for ancestral trait values under these models. Crawford and Suchard (2013) give an estimator for the evolutionary variance under Brownian motion for an unobserved Yule tree.

In this paper we study the distribution of character values observed at the tips of a phylogenetic tree generated by the Yule process. We first state two theorems that describe the distribution of the time of shared ancestry between two randomly chosen taxa in a Yule tree of age τ with n taxa and speciation rate λ. Next we extend results presented by Bartoszek and Sagitov (2012) and Bartoszek (2013) to find the exact probability distribution and covariance between pairs of randomly chosen tip values under Brownian motion and Ornstein–Uhlenbeck evolution of a continuous trait. These results give insight into the finite-time, finite-n dynamics of interspecies correlation. Next we examine discrete character evolution on Yule trees under Poisson and reversible Poisson mutation models. We suggest a new measure of phylogenetic information and give a method for deciding whether it is better to add taxa or sites to a phylogenetic analysis.

2. Background

A Yule process $Y(t)$ is a continuous-time Markov chain on the positive integers in which a jump from state n to $n+1$ occurs with rate $n\lambda$. Define $P^Y_{mn}(t)=\Pr(Y(t)=n|Y(0)=m)$ to be the transition probability from state m to n in time t. The Yule process obeys the forward Kolmogorov equations:

$$\frac{dP^Y_{mn}(t)}{dt}=(n-1)\lambda P^Y_{m,n-1}(t)-n\lambda P^Y_{mn}(t) \tag{1}$$

for $m\geq 1$ and $n\geq m$. The transition probabilities are

$$P^Y_{mn}(t)=\binom{n-1}{m-1}e^{-\lambda mt}(1-e^{-\lambda t})^{n-m} \tag{2}$$

(Bailey, 1964). A Yule tree is a binary tree in which the number of extant lineages at time t is given by the Yule process $Y(t)$. If there are n extant lineages and a "birth" event occurs, one of the n lineages is chosen uniformly at random and split into two. In this paper, we assume that at the MRCA of all n taxa existed at time 0. We model $t=0$ as the time of the first split, so $Y(0)=2$, and both tree size (number of taxa) n and age τ are given. In what follows, we limit our attention to the $(n-1)!$ unlabelled, ranked, oriented trees that make up an n-forest, since our conclusions readily carry over to the $n!(n-1)!/2^{n-1}$ leaf-labelled, ranked Yule trees of phylogenetic interest (Gernhard et al., 2008; Mulder, 2011).

We now consider pairs of tips on a Yule tree whose MRCA is the kth birth event. We call these events "nodes" in the tree. The kth node is preceded chronologically by $k-1$ nodes, and this node emerges at time x since the first split. The kth node corresponds to the "crown age" of the sub-tree or clade below the node. Fig. 1 shows an example in which the kth birth event, preceded by $k-1$ such events, takes place at time x. In continuous time each n-tree pattern of this type, with tree age τ and with the kth node appearing at time x, has the same probabilistic weight, and hence these trees can be dealt with on equal footing using purely combinatorial arguments. The following result gives the probability of two randomly chosen tips in a phylogenetic tree having their MRCA at the kth node. It was first derived by Stadler (2009).

Theorem 1. *The probability of randomly choosing two tips in a tree of size n whose MRCA is the kth node is*

$$P(n,k)=\frac{2(n+1)}{(n-1)(k+1)(k+2)} \tag{3}$$

for $n\geq k+1$ (Stadler, 2009).

Appendix A gives a simple alternative proof of this fact using recurrence relations.

We now consider the time of shared ancestry of two randomly chosen taxa, the age of their MRCA. Theorem 1 provides the probability of choosing two tips whose MRCA is the kth node; here we seek the distribution of the age x of this node.

Lemma 1. *The probability density of time x of the kth node of a Yule tree of age* τ *and size n is*

$$p(x|k,n,\tau,\lambda)=\begin{cases}\delta(x), & k=1\\ \lambda(n-2)\binom{n-3}{k-2}e^{-(k-1)\lambda(\tau-x)}(1-e^{-\lambda x})^{k-2}(1-e^{-\lambda(\tau-x)})^{n-k-1}\\ \hline (1-e^{-\lambda\tau})^{n-2}, & k\geq 2\end{cases} \tag{4}$$

for $0\leq x\leq\tau$, *where* $\delta(x)$ *is the Dirac delta function.*

Appendix B provides a derivation.

Now we study the age of the MRCA of two randomly chosen taxa without conditioning on the MRCA being the kth node in the tree. Finding the marginal distribution of x by summing $P(n,k)$ over k with respect to (4), we arrive at

$$p(x|n,\tau,\lambda)=\sum_{k=1}^{n-1}P(n,k)\;p(x|k,n,\tau,\lambda). \tag{5}$$

where $P(n,k)$ is given by Theorem 1 and $p(x|k,n,\tau,\lambda)$ is given by Lemma 1. The following Theorem gives a closed-form expression for this probability.