Birth-death-sampling

Graham Jones

13th December 2007

1 Introduction

This note is aimed at providing tree priors for Bayesian phylogenetic analysis. The model is a standard birth-death process which produces a binary tree, followed by random sampling of the tips. Assume that at some time t_1 in the past a speciation occurred, and that this forms the root of the tree that the analysis attempts to recover: it is the last common ancestor of the species in the sample, so both lineages starting at t_1 must leave at least one species in the sample. A birth-death process with speciation rate λ and extinction rate μ starting at this root produces some number n species at the present time. The present time is taken to be zero, and times are measured 'backwards' so that $t_1 > 0$. After the birth-death process reaches time zero, the n species are then sampled at a rate $\rho \leq 1$, which produces a sample of size s.

This is clearly an idealised account of how a sample of species is chosen for a real phylogenetic analysis. To carry it out strictly, the researcher would need to identify a monophyletic clade (how?) and a number $\rho <= 1$, and then accept or reject each species from the clade at random using ρ, and then analyse whatever list of species this produced.

2 Calculation

In [1], equation (3) gives an expression for the joint probability density for the node times after t_1, namely $t_2, ..., t_{s-1}$, given t_1, s, ρ, λ, and μ. Rewriting their formula, this is

$$f(t_2, ...t_{s-1}|t_1, s; \rho, \lambda, \mu) = (s-2)! \left(\frac{\rho\lambda + (\lambda - \rho\lambda - \mu)e^{(\mu-\lambda)t_1}}{\rho(1 - e^{(\mu-\lambda)t_1})} \right)^{s-2} \prod_{j=2}^{s-1} p_1(t_j) \tag{1}$$

where $p_i(x)$ is the probability that a lineage at time x leaves exactly i species in the sample, after the birth-death process and sampling. Only $p_1(x)$ appears in the above equation, but I use the same notation for all $i \geq 0$. The rest of their analysis sets $t_1 = 1.0$ and makes inferences about λ and μ relative to t_1. For more general phylogenetic analysis, a formula for the density of all nodes including the root is more useful, and what is missing is therefore a density for t_1 given s, ρ, λ, and μ. A calculation of the value for this follows, starting with a calculation of the $p_i(x)$ for $i \geq 1$. Let $q_i(x)$ be the probability that a lineage at time x leaves exactly i species at present, after the birth-death process, but before sampling. A standard result from the theory of the birth-death process (see [2], or one of many books on probability) gives, for $i \geq 1$,

$$q_i(x) = q_1(x)\alpha(x)^{i-1}$$

where

$$\alpha(x) = \frac{\lambda - \lambda e^{(\mu-\lambda)x}}{\lambda - \mu e^{(\mu-\lambda)x}}$$

and

$$q_1(x) = \frac{(\lambda - \mu)^2 e^{(\mu-\lambda)x}}{(\lambda - \mu e^{(\mu-\lambda)x})^2}.$$

If j species are produced by the birth-death process, then the probability that i 'survive' the sampling is $\binom{j}{i}\rho^i(1-\rho)^{j-i}$ so

$$p_i(x) = \sum_{j=i}^{\infty} q_j(x)\binom{j}{i}\rho^i(1-\rho)^{j-i}$$

Substituting for $q_j(x)$ and rearranging,

$$p_i(x) = q_1(x)\alpha(x)^{-1}\rho^i(1-\rho)^{-i}\sum_{j=i}^{\infty}\binom{j}{i}[(1-\rho)\alpha(x)]^j.$$

The sum can be recognised as an expansion of $\eta^i(1-\eta)^{-i-1}$, where $\eta = (1-\rho)\alpha(x)$, so

$$p_i(x) = q_1(x)\alpha(x)^{-1}\rho^i(1-\rho)^{-i}[(1-\rho)\alpha(x)]^i[1-(1-\rho)\alpha(x)]^{-i-1}$$

which simplifies to

$$p_i(x) = \frac{\rho^i q_1(x)\alpha(x)^{i-1}}{(1-(1-\rho)\alpha(x))^{i+1}}.$$

Substituting for $q_1(x)$ and $\alpha(x)$ and more rearranging gives

$$p_i(x) = p_1(x)\beta(x)^{i-1}$$

where

$$\beta(x) = \frac{\rho\lambda(1-e^{(\mu-\lambda)x})}{\rho\lambda + (\lambda - \rho\lambda - \mu)e^{(\mu-\lambda)x}}$$

and

$$p_1(x) = \frac{\rho(\lambda-\mu)^2 e^{(\mu-\lambda)x}}{(\rho\lambda + (\lambda - \rho\lambda - \mu)e^{(\mu-\lambda)x)})^2}.$$

Up to this point, the results are standard, though I have not been able to find a reference I can quote directly for the values of $p_i(x)$. Nee et al [3] do similar (and more general) calculations.

From here on, errors are more likely!

Now, the probability that the root node produces i species in the sample from one lineage and $(s-i)$ from the other is $p_i(t_1)p_{s-i}(t_1)$, which is $p_1(t_1)^2\beta(t_1)^{s-2}$ for all i, and since we require at least one species from each lineage, this needs to be summed from $i = 1$ to $s-1$, that is, multiplied by $(s-1)$ to give

$$P(s|t_1;\rho,\lambda,\mu) = (s-1)p_1(t_1)^2\beta(t_1)^{s-2} \tag{2}$$

This can be regarded as a likelihood function $L(t_1|s;\rho,\lambda,\mu)$ for t_1. I assume that before observing the sample (before knowing the number of species) the last common ancestor is equally likely to have been any time in the past. That is, an improper uniform prior for t_1 on $[0,\infty)$ is assumed. Then the posterior density for t_1 is proportional to the right hand side of equation 2. This can be multiplied by equation (1) to provide a density for all the node times including the root. Note that the term raised to the power of $(s-2)$ in equation (1) is $\lambda/\beta(t_1)$, so this density is

$$g(t_1, t_2, ...t_{s-1}|s; \rho, \lambda, \mu) = c(\rho, \lambda, \mu, s) p_1(t_1)^2 \prod_{j=2}^{s-1} p_1(t_j) \tag{3}$$

where $c(\rho, \lambda, \mu, s)$ is a normalisation constant. In Bayesian phylogenetic analysis, ρ, λ, and μ are generally parameters which will be estimated along with the tree, so $c(\rho, \lambda, \mu, s)$ is required (although the dependence on s is not needed) and this can be found as follows.

$$c(\rho, \lambda, \mu, s)^{-1} = \int_0^\infty p_1(t_1)^2 \left(\int \prod_{j=2}^{s-1} p_1(t_j) \right) dt_1$$

where the inner integral is taken with respect to $t_2, ..., t_{s-1}$ over the region $0 \leq t_2, ..., t_{s-1} \leq t_1$. It is easily shown that the derivative of $\beta(x)/\lambda$ is $p_1(x)$, so this inner integral is $\beta(x)^{s-2}\lambda^{-(s-2)}$ and

$$c(\rho, \lambda, \mu, s)^{-1} = \lambda^{-(s-2)} \int_0^\infty p_1(x)^2 \beta(x)^{s-2} dx$$

This one-dimensional integral could be evaluated numerically. In the case of a pure birth process where $\mu = 0$ and $\rho = 1$, it is

$$c(0, \lambda, 0, s)^{-1} = \lambda^{-(s-2)} \int_0^\infty e^{-2\lambda x}(1 - e^{-\lambda x})^{s-2} dx = \lambda^{-(s-1)}/(s(s-1))$$

so that in this case

$$g(t_1, t_2, ...t_{s-1}|s; \lambda) = s(s-1)\lambda^{s-1} e^{-\lambda(2t_1 + t_2 + t_3 + ...t_{s-1})}. \tag{4}$$

References

[1] Z. Yang and B. Rannala, *Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method*, Mol. Biol. Evol. 14(7):717-724. 1997

[2] Z. Yang and B. Rannala, *Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference*, J Mol Evol (1996) 43:304311

[3] Sean Nee, Robert M. May, Paul H. Harvey, *The Reconstructed Evolutionary Process*, Philosophical Transactions: Biological Sciences, Vol. 344, No. 1309 (May 28, 1994), pp. 305-311