Pure birth model

Dan Rabosky

June 10, 2019

1 Derivation of the model

Consider a single lineage that exists at time t with probability $P(t)$. In the CTMC framework, we can think about this lineage as being in a state S_1. If the lineage undergoes a speciation event, we consider it to have transitioned to a new state S_2, which consists of two lineages. E.g., any state S_k is a cohort of k lineages, any of which can undergo a speciation event. For our purposes, though, we will only worry about the transition from S_1 to S_2.

Let's go back to our single lineage in state S_1. Allow that the lineage can undergo speciation on some time interval Δt with some probability $\lambda \Delta t$.

It must then be true that, on the same time interval Δt, the lineage will *not undergo speciation* with probability $1 - \lambda \Delta t$.

We can then write down an equation for the *new probability* that the lineage exists in its current state some amount of time Δt later, as

$$P(t + \Delta t) = (1 - \Delta t \lambda) P(t) \tag{1}$$

So, the probability of the lineage in its current state S_1 some time Δt later is simply the current probability of existence $P(t)$ multiplied by the probability that nothing happens on the focal interval. We can rearrange this equation to:

$$P(t + \Delta t) = P(t) - \Delta t \lambda P(t) \tag{2}$$

and then

$$P(t + \Delta t) - P(t) = -\Delta t \lambda P(t) \tag{3}$$

Now, we can maybe see where this is going. Remember from calculus that the definition of a derivative is:

$$\frac{dX}{dt} = \lim_{\Delta t \to 0} \frac{X(t + \Delta t) - X(t)}{\Delta t} \tag{4}$$

We want to make a differential equation for the change in probability as a function of time, dP/dt, and we can do this by dividing both sides by Δt and taking limits as $\Delta t \to 0$:

$$\frac{P(t + \Delta t) - P(t)}{\Delta t} = \frac{-\Delta t \lambda P(t)}{\Delta t} \tag{5}$$

$$\lim_{\Delta t \to 0} \frac{P(t + \Delta t) - P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{-\Delta t \lambda P(t)}{\Delta t} \tag{6}$$

So this gives us

$$\frac{dP}{dt} = -\lambda P(t) \tag{7}$$

This is a simple differential equation that can be solved to yield an equation for the probability that a lineage *will not speciate* after some time t.

Representing $P(t)$ as P, we can make some simple rearrangements:

$$\frac{dP}{P} = -\lambda dt \tag{8}$$

We can solve this by integrating both sides:

$$\int \frac{dP}{P} = \int -\lambda dt \tag{9}$$

$$\ln(P) = -\lambda t + c \tag{10}$$

where c is the constant of integration. Exponentiating both sides:

$$P = e^{-\lambda t + c} = e^{-\lambda t} e^{c} \tag{11}$$

Rewriting with $P(t)$:

$$P(t) = e^{-\lambda t} e^{c} \tag{12}$$

To deal with the constant of integration, we note that we have the initial condition $P(0) = 1$, or

$$P(t) = 1 = e^{0} e^{c} \tag{13}$$

or

$$1 = e^{c} \tag{14}$$

So $c = 0$, and the probability density of a given waiting time t_i, under a pure-birth process, is:

$$P(t_i) = e^{-\lambda t_i} \tag{15}$$

2 Likelihood of a phylogenetic tree

To construct the likelihood of a phylogenetic tree under the pure-birth model, we need to compute the probability density of two components that make up our tree:

- the waiting times between each speciation event, using the probability model derived above
- the speciation events themselves

For a fully-resolved (nonzero branch length) phylogeny of N species, we have a total of $2N - 2$ waiting times, each of which represents the transition between state S_1 and S_2. We also have a total of $N - 1$ speciation events represented in our tree. The density (likelihood) of the waiting times is the product of the probability densities of each individual waiting time t_i (e.g., individual branch lengths), or

$$\prod e^{-\lambda t_i} = e^{-\lambda \sum t_i} = e^{-\lambda T} \tag{16}$$

where T is the sum of all branch lengths in the tree. And the speciation events have a probability that is proportional to their value, such that their contribution to the likelihood is λ^{N-1}. Putting these together, the likelihood of the data given the speciation rate λ is

$$L(D) = \lambda^{N-1} e^{-\lambda T} \tag{17}$$

Now, we typically *condition* this expression on the occurrence of the basal (root) node: if this speciation event hadn't happened, we wouldn't be looking at a tree. So we simply divide the equation through by λ, giving us:

$$L(D) = \lambda^{N-2} e^{-\lambda T} \tag{18}$$

Finally, it is usually easier to work with the natural logarithm of the likelihood, so we'll take the log of both sides:

$$logL(D) = (N - 2)\log(\lambda) - \lambda T \tag{19}$$

And that's it: this simple expression gives us the likelihood of a given phylogenetic tree under the pure-birth-model.

3 Finding the ML speciation rate under the model

The likelihood given above is pretty simple, suggesting that we might be able to find (analytically) the value of λ that maximizes the likelihood. To do so, we'll take the derivative of the expression for the log-likelihood. We'll then set the resulting equation equal to zero, which will give us all values of λ for which the slope of the likelihood is zero (e.g., a local maximum or minimum of the function). Recalling that $d/d\lambda \log \lambda = 1/\lambda$, we have:

$$\frac{d \log L}{d\lambda} = \frac{N-2}{\lambda} - T \tag{20}$$

Setting the left side equal to zero and rearranging terms gives us:

$$\hat{\lambda} = \frac{N-2}{T} \tag{21}$$

where $\hat{\lambda}$ is the value of λ that maximizes the log-likelihood of the data. Take a minute to consider how simple this expression is, and what it means. If you think about it, this is actually quite intuitive. How do you estimate the rate at which things happen? The simplest and most natural thing to do is to divide the *number of things that happened* by the *amount of time available for those things to happen*. Here, the number of things that happened is $N-2$ ($N-2$ speciation events), and the total time available for speciation events to happen is the sum of branch lengths, or T. So, a bit of math has given us exactly the answer that we should intuitively expect!