Chapter 11: Fitting birth-death models

pdf version

Introduction: Diversity hotspots

The number of species on the Earth remains highly uncertain, but ranges somewhere above 10 million. As far as we know, all of those species are descended from a single common ancestor that lived some 4.5 billion years ago. All species on Earth, then, formed by the process of speciation, the process by which one species splits into two (or more) descendants. Imbalance in diversity across the tree of life tells us that speciation is much more common in some lineages than others. Moreover, numerous studies have argued that certain habitats are "hotbeds" of speciation. For example, the high Andes ecosystem called the Páramo - a peculiar landscape of alien-looking plants and spectacled bears - harbors the highest speciation rates on the planet.

Figure 11.1 Páramo ecosystem, Chingaza Natural National Park, Colombia. This photo is mine.

In this chapter we will explore how we can fit birth-death models to data and, in the process, learn about speciation and extinction rates. Birth-death models can be applied to clade ages and diversities or to patterns of branching times in phylogenetic trees. We will explore both maximum likelihood and Bayesian methods to do this.

Key questions:

1. How can we calculate speciation and extinction rates using clade ages and diversities?
2. How can we fit a birth-death model to the pattern of branching times on a phylogenetic tree?

Clade age and diversity

If we know the age of a clade and its current diversity, then we can calculate the net diversification rate for that clade. The simplest way follows Magallon and Sanderson (2001), who give an equation for the estimate of net diversification rate as:

(eq. 11.1) r^=(log(n))/t_stem

where tstem is the stem group age and r= - . Alternatively, one can use tcrown, the crown group age:

(eq. 11.2) r^=(log(n)-log(2))/t_crown

The two equations differ because at the crown group age one is considering the clade's diversification starting with two lineages rather than one (Figure 11.2). Furthermore, these two equations give ML estimates for r only if there is no extinction. If there has been extinction in the history of the clade, then these estimates will be biased. Under a scenario with extinction, one can define = / and use the following method-of-moments estimators from Rohatgi (1976, following the notation of Magallon and Sanderson 2001):

(eq. 11.3) r̂ =(log[n(1-)+])/t_stem

for stem age, and

(eq. 11.4) r̂ =(log[n(1- ^2)/2+2 +((1-) √(n(n ^2-8 +2n +n))/2]-log(2))/t_crown

for crown age. (Note that eq. 11.3 and 11.4 reduce to 11.1 and 11.2, respectively, when = 0). Magallon and Sanderson (2001), following Strathmann and Slatkin (1983), describe how to use eq. 10.13 and 10.15 to calculate confidence intervals for the number of species at a given time.

For example, consider these data, which summarize the ages and diversities of a number of plant lineages in the Páramo (from Madriñán et al. 2013). For each lineage, I have calculated the pure-birth estimate of speciation rate (from equation 11.2, since these are crown ages), and net diversification rates under three scenarios for extinction (=0.1, =0.5, and =0.9).

Lineage n Age r̂ _pb r̂ _([=0.1]) r̂ _([=0.5]) r̂ _([=0.9]) Aragoa 17 0.42 5.10 5.08 4.53 2.15 Arcytophyllum 14 10.96 0.18 0.18 0.16 0.07 Berberis 32 3.8 0.73 0.73 0.66 0.36 Calceolaria 65 2.5 1.39 1.39 1.28 0.78 Draba 55 3.05 1.09 1.08 1.00 0.59 Espeletiinae 120 4.04 1.01 1.01 0.94 0.62 Festuca 36 4.28 0.68 0.67 0.61 0.34 Jamesonia + Eriosorus 32 7.6 0.36 0.36 0.33 0.18 Lupinus 66 1.47 2.38 2.37 2.19 1.34 Lysipomia 27 8.96 0.29 0.29 0.26 0.14 Oreobolus 5 3.01 0.30 0.30 0.26 0.09 Puya 46 0.8 3.92 3.91 3.58 2.07 Valeriana 53 14.58 0.22 0.22 0.21 0.12

Table 11.1. Estimates of net diversification rates for Páramo lineages using equations 11.2 and 11.4.

We can also estimate birth and death rates for clade ages and diversities using ML or Bayesian approaches. We already know the full probability distribution for birth-death models starting from any standing diversity N(0)=a (see equations 10.13 and 10.15). We can use these equations to calculate the likelihood of any particular combination of N and t (either tstem or tcrown) given particular values of and . We can then find parameter values that maximize that likelihood. Of course, with data from only a single clade, we cannot estimate parameters reliably; in fact, we are trying to estimate two parameters from a single data point, which is a futile endeavor. (It is common, in this case, to assume some level of extinction and calculate net diversification rates based on that). One can also assume that a set of clades have the same speciation and extinction rates and fit them simultaneously, estimating ML parameter values. This is the approach taken by Magallon and Sanderson in their 2001 paper on diversification rates across angiosperms.

When we apply this approach to the Paramo data, shown above, we obtain ML estimates of rˆ=0.27 and ˆ=0. If we were forced to estimate an overall average rate of speciation for all of these clades, this might be a reasonable estimate. However, the table above also suggests that, perhaps, some of these clades might be diversifiying faster than others. We will return to the issue of rate variation across clades in the next chapter.

Another approach is to use a Bayesian approach to calculate posterior distributions for birth and death rates based on clade ages and diversities. This approach has not, to my knowledge, been implemented in any software package, although the method is straightforward (for similar approaches, see xxx B Moore). To do this, we will modify the basic algorithm for Bayesian MCMC (see Chapter 2) as follows:

`Sample a set of starting parameter values, r and , from their prior distributions. For this`
`Given the current parameter values, select new proposed parameter values using the proposal`
`Q(pˆ' p)~U(p-w_p/2,p+w_p/2).`

We can either choose both parameter values simultaneously, or one at a time (the latter is typically more effective). Calculate three ratios: The prior odds ratio. This is the ratio of the probability of drawing the parameter values p and p' from the prior. Since we have exponential priors for both parameters, we can always calculate this ratio as: a_1=(_p eˆ(- _p p'))/(_p eˆ(- _p p))=eˆ(_p (p-pˆ‘)). The proposal density ratio. This is the ratio of probability of proposals going from p to p' and the reverse. We have already declared a symmetrical proposal density, so that = and . The likelihood ratio. This is the ratio of probabilities of the data given the two different parameter values. We can calculate these probabilities from equations 11.3 or 4 (depending on if the data are stem ages of crown ages). Then:

`Find the product of the prior odds, proposal density ratio, and the likelihood ratio. In thi`
`a=a_1 a_3`

`Draw a random number x from a uniform distribution between 0 and 1. If x < a, accept the pro`
`Repeat steps 2-5 a large number of times.`

When we apply this technique to the Páromo (from Madriñán et al. 2013), we obtain posterior distributions for both r (mean = 0.497, 95% CI = 0.08-1.77) and (mean = 0.36, 95% CI = 0.02-0.84; Figure xxx).

Thus, we can estimate diversification rates from data on clade ages and diversities. If we have a whole set of such clades, we can (in principal) estimate both speciation and extinction rates, so long as we are willing to assume that all of the clades share equal diversificiation rates. However, as we will see in the next section, this assumption is almost always dubious!

Tree Balance

We can use the concept of tree balance to evaluate the fit of constant-rate birth-death models to phylogenetic trees. Tree balance is based on comparing the number of species descended from each pair of sister lineages in the tree. For single nodes, we already know that the distribution of sister taxa species richness is uniform over all possible divisions of Nn species into two clades of size Na and Nb. We can use this fact to derive a simple test of whether the distribution of species between two sister clades is unusual compared to the expectation under a birth-death model. This test can be used, for example, to test whether the diversity of exceptional clades, like passerine birds, is higher than one would expect when compared to their sister clade. This approach traces back to Raup and colleagues, who applied stochastic birth-death models to paleontology in a series of influential papers in the 1970s (e.g. Raup et al. 1973; Raup and Gould 1974). Slowinsky and Guyer (1993) developed a test based on calculating a p-value for a division at least as extreme as seen in a particular comparisons of sister clades. We consider Nn total species divided into two sister clades of sizes Na and Nb, where Na < Nb and Na + Nb = Nn. Then:

(eq. 11.5) P=(2N_a)/(N_n-1)

(if Na = Nb, then P = 1).

For example, we can assess diversification in the Andean representatives of the legume genus Lupinus (Hughes and Eastwood 2006). In particular, this genus includes a young radiation of 81 Andean species, spanning a wide range of growth forms. The likely sister clade to this spectacular Andean radiation is a clade of Lupinus species in Mexico that includes 46 species (Drummond et al. 2013). We can then calculate a P-value testing the null hypothesis that both of these clades have the same diversification rate:

(eq. 11.6) P=(2N_a)/(N_n-1)=(2 46)/(91-1)=1.0

We cannot reject the null hypothesis. Indeed, later work suggests that the actual increase in diversification rate for Lupinus occurred deeper in the phylogenetic tree, in the ancestor of a more broadly ranging New World clade (Drummond et al. 2013, Hughes et al. 2015).

Often, we are interested in testing whether a particular trait - say, dispersal into the Paramo - is responsible for the increase in species richness that we see in some clades. In that case, a single comparison of sister clades may be unsatisfying, as sister clades almost always differ in many characters, beyond just the trait of interest. Even if the clade with our putative "key innovation" is more diverse, we still might not be confidence in inferring a correlation from a single observation. To address this problem, many studies have used natural replicates across the tree of life, comparing the species richnesses of many pairs of sister clades that differ in a given trait of interest. Following Slowinsky and Guyer's logic above, we could calculate a p-value for each clade, and then combine those p-values

into an overall test. In this case, one clade (with diversity N1) has the trait of interest and the other does not (N0), and our formula is half of equation 11.5 since we will consider this a one-tailed test:

(eq. 11.7) P=N_0/(N_n-1)

Slowinsky and Guyer (1993) recommended combining these p-values using Fisher's combined probability test, so that:

`P_total=-2 lnP_i`

Follow-up work showed that this test, though, can be very sensitive to outliers - that is, clades with extreme differences in diversity - and can, in some cases with two characters, show that both characters significantly increase diversity (Vamosi and Vamosi 2005)! Fortunately, there are a number of improved methods that can be used that are similar in spirit to the original Slowinsky and Guyer test but more statistically robust (xxx add references).

Finally, we can assess the overall balance of an entire phylogenetic tree using tree balance statistics. There are a relatively large number of such statistics, and different indices capture different aspects of diversification. Since the test statistics are based on descriptions of patterns in trees rather than particular processes, the relationship between imbalance and evolutionary processes can be difficult to untangle! But all tree balance indices allow one to reject the null hypothesis that the tree was generated under a birth-death model. Actually, the expected patterns of tree balance are absolutely identical under a broader class of models called "Equal-Rates Markov" (ERM) models. ERM models specify that diversification rates (both speciation and extinction) are equal across all lineages for any particular point in time. However, those rates may or may not change through time. If they don't change through time, then we have a constant rate birth-death model, as described above - so birth-death models are ERM models. But ERM models also include, for example, models where birth rates slow through time, or extinction rates increase through time, and so on. All of these models predict exactly the same pattern of tree balance.

Typical steps for using tree balance indices to test the null hypothesis that the tree was generated under an ERM model are as follows:

1. Calculate tree balance using a tree balance statistic.
2. Simulate pure birth trees to general a null distribution of the test statistic. We are considering the set of ERM models as our null, but since pure-birth is simple and still ERM we can use it to get the correct null distribution.
3. Compare the actual test statistic to the null distribution. If the actual test statistic is in the tails of the null distribution, then your data deviates from an ERM model.

Step 2 is unnecessary in cases where we know null distributions for tree balance statistics analytically (e.g. xxx). There are also some examples in the literature of considering null distributions other than ERM. For example, Mooers and Heard consider two other null models, PDA and EPT, which consider different

statistical distributions of tree shapes (but both of these are difficult to tie to any particular evolutionary process).

Typically, phylogenetic trees are more imbalanced than expected under the ERM model. In fact, this is one of the most robust generalizations that one can make about macroevolutionary patterns in phylogenetic trees. This deviation means that diversification rates vary among lineages in the tree of life. We will discuss how to quantify and describe this variation in later chapters. These tests are all similar in that they use multiple non-nested comparisons of species richness in sister clades to calculate a test statistic, which is then compared to a null distribution, usually based on a constant-rates birth-death process (reviewed in Vamosi and Vamosi 2005, Paradis 2012).

XXX example

Fitting birth-death models to branching times

Another approach that uses more of the information in a phylogenetic tree involves fitting birth-death models to the distribution of branching times in a phylogenetic tree. This approach traces all the way back to Yule (1924), who first applied stochastic process models to the growth of phylogenetic trees. More recently, Raup et al. (1973 and various follow-ups) spurred modern approaches to quantitative macroevolution by demonstrating how variable clades grown under simple birth-death models can be.

Most modern approaches to fitting birth-death models to phylogenetic trees use the intervals between speciation events on a tree - the "waiting times" between successive speciation - to estimate the parameters of birth-death models. Figure xxx shows these waiting times. Frequently, information about the pattern of species accumulation in a phylogenetic tree is summarized by a lineage-through-time (LTT) plot, which is a plot of the number of lineages in a tree against time. Typically, the y-axis of LTT plots is log-transformed, so that the expected pattern under a constant-rate pure-birth model is a straight line. Note also that LTT plots ignore the relative order of speciation events - that is, the two trees shown in Figure xxx (below) have the same LTT plot. Stadler (2013) calls models justifying such an approach "species-exchangable" models - we can change the identity of species at any time point without changing the expected behavior of the model. Because of this, approaches to understanding birth-death models based on branching times are complementary to approaches based on tree topology.

For phylogenetic trees with only extant (living) species, such plots are always strictly increasing - even when the total number of species in the clade has gone up and down through time. But even though we often have no information about extinct species in a clade, we can still (in theory) infer the presence of extinction from an LTT plot. This is because in our birth-death model we assume that each lineage has a constant probability of extinction per unit time. Lineages that

have been around for the longest, then, have the highest cumulative probability of extinction - and, likewise, young lineages have had little time to go extinct. This leads to an excess of young lineages, which is seen as a steep upturn in the LTT plot towards the present day. This upturn, called the "pull of the recent," allows our statistical methods to detect the signature of extinction in the branch lengths of a tree.

In order to use ML and Bayesian methods for estimating the parameters of birth-death models from comparative data, we need to write down the likelihoods of the waiting times between speciation events in a tree. There is a little bit of variation in notation in the literature, so I will follow Stadler (xxx) to maintain consistency. We will assume that the clade begins at time 0 with a single species. Speciation and extinction events occur at various times, and the process ends at time t0 when the clade has n extant species. Extinction will result in species that do not extend all the way to t0. For now, we will assume that we only have data on extant species. We will refer to the phylogenetic tree that shows branching times leading to the extant species as the reconstructed tree (Nee xxx). For a reconstructed tree with n species, there are n-1 speciation times, which we will denote as t1, t2, …, tn. Note that in this notation, t1 < t2 < … < tn-1, that is, our speciation times are constantly increasing (this is an important notational difference between Stadler (2013) and Nee (xxx). For now, we will assume complete sampling; that is, all n species alive at the present day are in the tree.

We can now write down the likelihood of observing the set of speciation times t1, t2, …, tn given our total age, t0, the extant diversity of the clade, n, and our birth-death model parameters and . There are a number of ways to condition this likelihood (see Stadler 2012 for a review). We will follow most of the R packages and condition the process as starting at some time t1 in the past with two lineages (since we rarely have information on the stem age of our clade) and both surviving to the present day (e.g. Stadler 2012 equation 5). We then have:

Where p_0 (t_i) and p_1 (t_i) are the probabilities of observing 0 and 1 species, respectively, after sampling a birth-death tree of age t, and can be calculated as:

p_0 (t)=1-(-)/(- e^(-(-)t))

p_1 (t)=((-) ^2 e^((-(-)t))/(- e^(-(-)t))^2

<< THIS SHOULD MATCH 10.14 but I don't think it does - why not? >>

Given equation xxx for the likelihood, we can estimate birth and death rates using both ML and Bayesian approaches. For the ML estimate, we maximize equation xxx over and . For a pure-birth model, that is when = 0, the maximum likelihood estimate of can be calculated analytically as:

= (n-2)/s

where s is the sum of branch lengths in the tree,

This is the Kendal-Moran estimator of the speciation rate (Baldwin and Sanderson 1998).

For a birth-death model, we can use numerical methods to maximize the likelihood over and . EXAMPLE

We can also estimate birth and death rates using a Bayesian MCMC. We can use exactly the method spelled out above for clade ages and diversities, but substitute equation xxx for the likelihood, thus using the waiting times derived from a phylogenetic tree to estimate model parameters. EXAMPLE.

It is important to think about sampling when fitting birth-death models to phylogenetic trees. If any species are missing from your phylogenetic tree, they will lead to biased parameter estimates. This is because missing species are disproportionally likely to connect to the tree on short, rather than long, branches. If we randomly sample lineages from a tree, we will end up badly underestimating both speciation and extinction rates (and wrongly inferring slowdowns; see chapter 12).

Fortunately, the mathematics for incomplete sampling of reconstructed phylogenetic trees has also been worked out. We can substitute in equations that include , the proportion of sampled species, for equations xxx and xxx above:

p_0 (t)=1- (-)/(+((1-)-)e ^(-(-)t))

p_1 (t)=((-) ^2 e^(-(-)t))/(+((1-)-)e (-(-)t))^2

EXAMPLE

Chapter Summary

In this chapter, I described how to estimate parameters from birth-death models using data on species diversity and ages, and how to use patterns of tree balance to test hypotheses about changing birth and death rates. I also described how to calculate the likelihood for birth-death models on trees, which leads directly to both ML and Bayesian methods for estimating birth and death rates. Next, we will explore elaborations on birth-death models, and discuss models that go beyond constant-rates birth-death models to analyze the diversity of life on Earth.

References