The Problem with Clade-specific Sampling Fractions

Jeremy M. Beaulieu

We get a lot of requests to include clade-specific sampling capabilities, like those available in BAMM and MEDUSA. For a time, HiSSE actually had such capabilities, but we removed them because it leads to what we believe is an incorrect likelihood behavior. In an SSE model, the sampling fraction, f_i sets the initial conditions for both $D_{N_i}(0)$ – the probability that a lineage in state i at time t would evolve into the extant clade N as observed – and $1 - f_i$ sets the initial conditions for $E_i(0)$ – the probability that a lineage in state i at time t would go completely extinct by the present. The important variable to pay attention to is $E_i(t)$, which does not depend on the tree structure, only on time. Thus, unlike with $D_{N_i}(t)$, at a node (termed N here) the probability for the left and right descendants (L and R, respectively) are not combined. Instead, because $E_i(t)$ at any node is anchored by the character states, $E_i(t)$ for the L and R descendants converge to the exact probabilities, and so it is arbitrary which set is carried down the subtending branch. However, as Moore et al. (2016) pointed out with BAMM, when the diversification parameters vary between the L and R descendants, due to a shift in diversification along one branch, $E_i(t)$ does not converge at the node.

The same problem occurs when the sampling fraction is allowed to vary across the tree because it actually transforms the speciation and extinction rates to account for the missing diversity. Stadler (2013) showed that the relationship between sampling fraction and speciation/extinction rates is as follows:

$$\lambda_i^* = \lambda_i / f_i; \mu_i^* = \mu_i - \lambda_i(1 - 1/f_i)$$

In other words, rescaling a global set of speciation and extinction rate parameters in different parts of the tree due to different sampling fractions effectively creates "rate shifts" in the tree. This also means that wherever in the tree any two clades that differ in their sampling fractions coalesce it is not possible to simply arbitrarily choose one set of extinction probabilities to carry down in the likelihood calculation.

We can demonstrate this behavior using a slightly involved but straightforward example. Here we will use a three-taxon tree to show that even after combining probabilities at a nested node, under ideal conditions, $E_i(t)$ for a L and R descedant branches should converge at a node. We will assume the tree has a total height of 20 time units, and we will also assume that the character states of the unsampled species are unknown completely, which means f_i simply reflects the number of extant species sampled. We will use the following function to conduct our branch calculations:

```
GetProbs <- function(yini, times) {
    times = times
    prob.subtree.cal.full <- lsoda(yini, times, func = "maddison_DE_bisse",
        padded.pars, initfunc = "initmod_bisse", dllname = "hisse",
        rtol = 1e-08, atol = 1e-08)
    probs.out <- prob.subtree.cal.full[-1, -1]
    return(probs.out)
}
```

In the case of a global sampling fraction of 50%, we can calculate the probability of the left lineage, L, which is a single branch of length 20, the extinction probabilities would be:

```
times = c(0, 20)
yini <- c(E0 = 1 - 0.5, E1 = 1 - 0.5, D0 = 0.5, D1 = 0)
```

```r
left.branch.probs <- GetProbs(yini, times)
left.branch.probs[1:2]
```

```
##        E0        E1
## 0.3453536 0.1793777
```

The calculation of the extinction probabilities for the right clade, which contains two taxa that coalesce at 10 time units, is a bit more involved. We first have to combine the probabilities of the two tip branches at node R:

```r
times = c(0, 10)
yini <- c(E0 = 1 - 0.5, E1 = 1 - 0.5, D0 = 0.5, D1 = 0)
right.subA.probs <- GetProbs(yini, times)
right.subB.probs <- GetProbs(yini, times)

nodeR <- c(right.subA.probs[3:4] * right.subB.probs[3:4] * c(0.1,
    0.2))
# Arbritarily using the extinction probabilities from side
# A:
phi_A <- right.subA.probs[1:2]
```

But, from here we simply use these probabilities as the initial conditions when calculating probabilities for the subtending branch representing the remaining 10 time units down to the root:

```r
times = c(10, 20)
yini <- c(E0 = phi_A[1], E1 = phi_A[2], D0 = nodeR[1], D1 = nodeR[2])
right.branch.probs <- GetProbs(yini, times)
right.branch.probs[1:2]
```

```
##     E0.E0     E1.E1
## 0.3453536 0.1793776
```

As expected, the probabilities of the left and right lineages converge to the same extinction probabilities (note, there may be slight precision issues due to our reliance on ode integration):

```r
round(left.branch.probs[1:2], 4) == round(right.branch.probs[1:2],
    4)
```

```
##   E0   E1
## TRUE TRUE
```

Let's assume that the right clade is actually sampled at 25%, and the left clade is still sampled at 50%:

```r
times = c(0, 10)
yini <- c(E0 = 1 - 0.25, E1 = 1 - 0.25, D0 = 0.25, D1 = 0)
right.subA.probs <- GetProbs(yini, times)
right.subB.probs <- GetProbs(yini, times)
nodeR <- c(right.subA.probs[3:4] * right.subB.probs[3:4] * c(0.1,
    0.2))
# Again, arbritarily using the extinction probabilities
# from side A:
phi_A <- right.subA.probs[1:2]

times = c(10, 20)
yini <- c(E0 = phi_A[1], E1 = phi_A[2], D0 = nodeR[1], D1 = nodeR[2])
right.branch.probs <- GetProbs(yini, times)
right.branch.probs[1:2]
```

```
##     E0.E0     E1.E1
## 0.4846120 0.2302621
```

```
left.branch.probs[1:2]
```

```
##        E0        E1
## 0.3453536 0.1793777
```

```
round(left.branch.probs[1:2], 4) == round(right.branch.probs[1:2],
    4)
```

```
##    E0    E1
## FALSE FALSE
```

Now, these extinction probabilities no longer converge at the root, and the likelihood is incorrect. An important question worth further examination is if sampling is uneven, is the final answer worse by forcing the assumption that sampling is even or is it better to allow clade-specific sampling but with mathematical errors? Future work will address this question. But until then, for mathemetical reasons we feel it is best to avoid using clade-specific sampling fractions, which is why we have removed this capability within all HiSSE functions. We recommend for now using either a global sampling fraction as shown above, or a global fraction for each of the states included in a given model.

References

Moore, B.R., S. Hohna, M.R. May, B. Rannala, and J.P. Huelsenbeck. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. Proceedings of the National Academy of Sciences, USA, 113:9569-9574.

Stadler, T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Systematic Biology, 68:321-329.