Analyzing diversification with `diversitree`: GeoSSE addendum

Emma E. Goldberg

February 9, 2011

Contents

1 Geographic distributions and diversification: GeoSSE

The GeoSSE model (Geographic State Speciation and Extinction) combines features of the constant-rates birth-death model with a three-state Markov model. It differs from BiSSE in parameterizing the model to represent diversification and range shifts among two regions, which includes allowing widely-distributed species whose ranges may change in conjunction with a speciation event. See Goldberg et al. (in press) for a full explanation of the model assumptions.

1.1 Parameters and a tree

Simulating trees under GeoSSE is not yet part of `diversitree`, but it can be done with a separate program, SimTreeSDD. A simulated tree is included and shown in Fig. 1.

```
> library(diversitree)
> data("geosse")
```

```
> statecols <- c(AB = "purple", A = "blue", B = "red")
> plot(geosse.phy, tip.color = statecols[geosse.phy$tip.state +
+     1], cex = 0.5)
```

A crude starting point for parameter estimation can be obtained with:

```
> p <- starting.point.geosse(geosse.phy)
> p
```

```
      sA       sB      sAB       xA       xB       dA       dB
2.018992 2.018992 2.018992 1.009496 1.009496 1.009496 1.009496
```

The parameters are: speciation within region A (`sA`), speciation within region B (`sB`), between-region speciation (`sAB`), extinction from region A (`xA`), extinction from region B (`xB`), dispersal from A to B (range expansion, `dA`), and dispersal from B to A (`dB`).

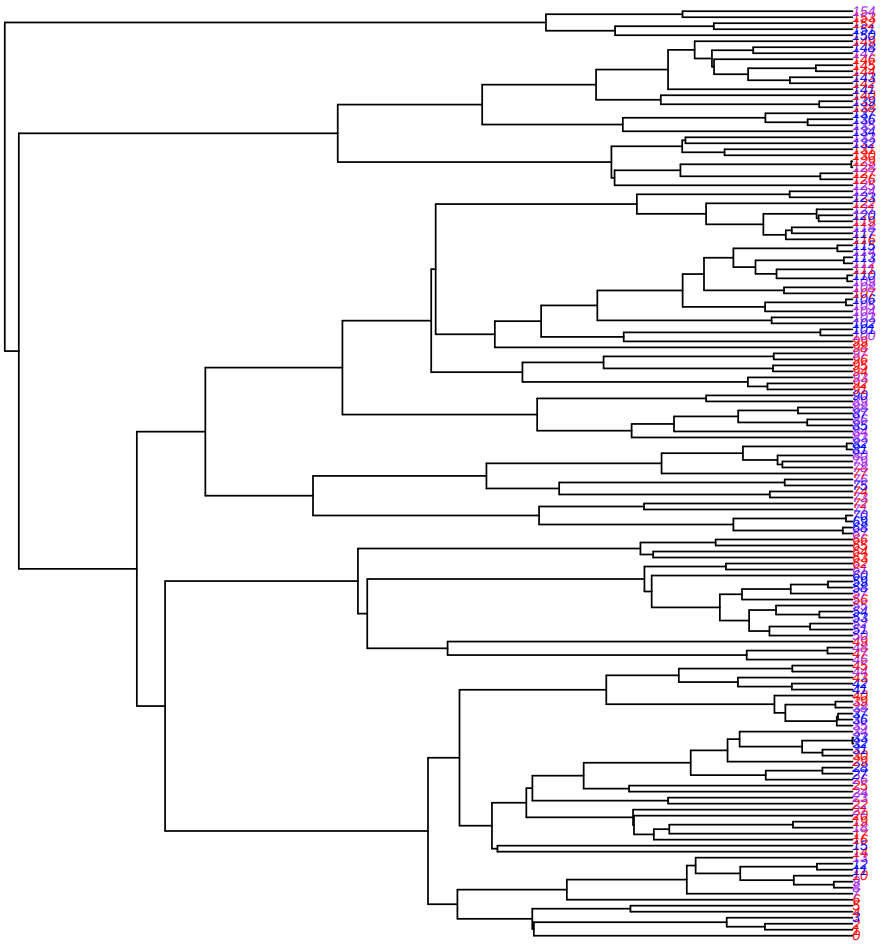

Figure 1: A GeoSSE tree simulated with `params = c(1.5, 0.5, 1.0, 0.7, 0.7, 2.5, 0.5)`. The tip state colors are purple for species present in both regions (AB), blue for species only in region A, and red for species only in region B.

1.2 Model construction and constraining

Constructing and constraining likelihood functions works as for the other models. Here we will consider the full model, a model without between-region speciation, and a model without regional dependence of speciation or extinction rates.

```
> lik1 <- make.geosse(geosse.phy, geosse.phy$tip.state)
> lik2 <- constrain(lik1, sAB ~ 0)
> lik3 <- constrain(lik1, sA ~ sB, xA ~ xB)
```

1.3 Maximum likelihood

ML parameter estimation and model comparisons:

```
> ml1 <- find.mle(lik1, p)
> p <- coef(ml1)
> ml2 <- find.mle(lik2, p[argnames(lik2)])
> ml3 <- find.mle(lik3, p[argnames(lik3)])

> round(rbind(full = coef(ml1), no.sAB = coef(ml2, TRUE), eq.div = coef(ml3,
+     TRUE)), 3)

          sA    sB   sAB    xA    xB    dA    dB
full   1.258 0.432 0.545 0.519 0.412 2.083 0.423
no.sAB 1.643 0.489 0.000 1.115 0.611 2.129 0.705
eq.div 0.802 0.802 0.605 0.237 0.237 1.546 0.629

> anova(ml1, no.sAB = ml2, eq.div = ml3)

       Df   lnLik    AIC   ChiSq Pr(>|Chi|)
full    7 -295.03 604.06
no.sAB  6 -296.00 603.99  1.9394   0.163731
eq.div  5 -300.52 611.04 10.9823   0.004123 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On this tree, we reject the model of equal speciation and extinction in the two regions, concluding that there are regional differences in diversification. Including the between-region mode of speciation does not, however, significantly improve the fit.

1.4 Markov chain Monte Carlo

We will only consider the 6-parameter model here. Use the ML rate estimates as a starting point. Place a broad exponential prior on each parameter.

```
> p <- coef(ml2)
> prior <- make.prior.exponential(1/2)
```

Use a pilot run to obtain reasonable step sizes:

```
> set.seed(1)
> tmp <- mcmc(lik2, p, nsteps = 100, prior = prior, lower = 0,
+     w = rep(1, 6))
> w <- diff(sapply(tmp[2:7], quantile, c(0.025, 0.975)))
```

Now the real analysis, which will take awhile to run:

```
> mcmc2 <- mcmc(lik2, p, nsteps = 10000, prior = prior, lower = 0,
+     w = w)
```

Marginal posterior distributions are shown in Fig. 2. We can also compare rate estimates by looking at posterior probabilities of their differences.

```
> mcmc2diff <- with(mcmc2, data.frame(s.diff = sA - sB, x.diff = xA -
+     xB, d.diff = dA - dB, div.A = sA - xA, div.B = sB - xB))
> colMeans(mcmc2diff > 0)

s.diff x.diff d.diff  div.A  div.B
1.0000 0.8219 0.9306 0.9413 0.3307
```

We correctly and confidently recover regional biases in speciation and dispersal, and positive net diversification in region A. With less confidence, we recover regional bias in extinction and negative net diversification for region B (i.e., extinction exceeds speciation in region B).

1.5 Additional options

GeoSSE likelihood functions can be built with randomly-incomplete sampling (FitzJohn et al. 2009). There is not currently support for unresolved clades.

```
> p <- coef(ml1)
> lik1(p)

[1] -295.0275

> lik4 <- make.geosse(geosse.phy, geosse.phy$tip.state, sampling.f = c(0.9,
+     0.6, 0.4))
> lik4(p)

[1] -307.7467
```

When using the likelihood function, one can condition on survival of the clade (not done by default):

```
> lik4(p, condition.surv = TRUE)

[1] -308.0192
```

External information about the geographic distribution of the common ancestor of the clade can be enforced by fixing the root state. For example, if you are absolutely positive that the MRCA was found only in region B:

```
> lik4(p, root.p = c(0, 0, 1), root = ROOT.GIVEN)

[1] -310.7291
```

Use this procedure with caution, and only in the face of truly external data, e.g., fossil or geologic information.

```
> col1 <- c("red", "orange", "blue", "purple", "black", "gray")
> col2 <- col1[c(1, 3, 5)]
> mcmc2diff <- with(mcmc2, data.frame(s.diff = sA - sB, x.diff = xA -
+     xB, d.diff = dA - dB))
> par(mfrow = c(2, 1), mar = c(3, 4, 0, 1))
> profiles.plot(mcmc2[2:7], col.line = col1, xlab = "", ylab = "")
> legend("topright", argnames(lik2), col = col1, lty = 1)
> profiles.plot(mcmc2diff, col.line = col2, xlab = "", ylab = "")
> legend("topright", colnames(mcmc2diff), col = col2, lty = 1)
> title(xlab = "rate", ylab = "posterior probability density",
+     outer = T, line = -1)
```

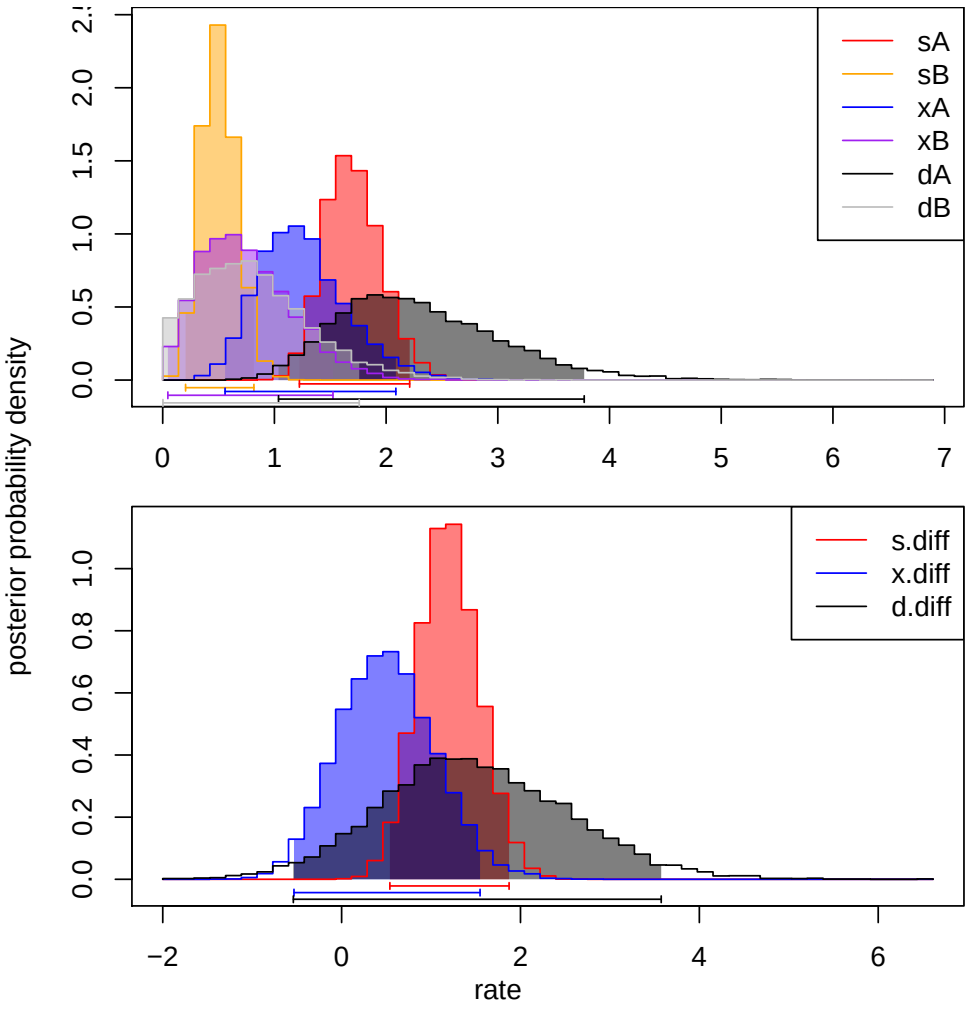

Figure 2: Posterior probability distributions for the six-rate GeoSSE model, for the tree shown in Fig. 1. Uncertainty is largest for dispersal and smallest for speciation. Regional differences in speciation and, to some extent, in dispersal are recovered.

2 Multi-clade analysis

Some applications of GeoSSE have combined multiple clades into a single analysis (Anacker et al. 2011; Goldberg et al. in press). This has the advantage of providing a larger dataset and hence presumably more power, but it is important to keep in mind the assumptions that go into such an analysis. First, it treats all clades as evolving according to the same model, with the same values for the rate parameters. This can be tested by first fitting the clades individually, or it may be inherent to the hypothesis at hand. Second, it assumes that the clades are independent of each other, since their likelihoods are simply multiplied together to form a joint likelihood function. In some cases, this may be a defensible approximation, for example, when the clades being considered are only very distantly related. If you have decided that you can convince yourself and your reviewers that a multi-clade analysis is appropriate, here is one way to do it.

Assemble the data as lists of trees and character state vectors. Each list element is one clade. Here we will use two trees from the chaparral study, one each from the posterior sets for *Ceanothus* and *Arctostaphylos*. The trees `geosse.phy.cea` and `geosse.phy.arc` and the tip states `geosse.chars.cea` and `geosse.chars.arc` (A = chaparral, B = forest) were read in with the call to `data("geosse")` above.

```
> chaparral.trees <- list(geosse.phy.cea, geosse.phy.arc)
> chaparral.states <- list(geosse.chars.cea, geosse.chars.arc)
> chaparral.samp <- list(c(0.913, 0.941, 0.875), c(0.674, 0.533,
+     0.75))
```

Create an individual likelihood function for each clade.

```
> lnL.each <- list()
> for (i in seq_along(chaparral.trees)) lnL.each[[i]] <- make.geosse(chaparral.trees[[i]],
+     chaparral.states[[i]], sampling.f = chaparral.samp[[i]])
```

Now define the joint, multi-clade likelihood function. (I actually used a somewhat messier procedure, but Rich suggested this `combine()` function.)

```
> combine <- function(liks) {
+     if (length(unique(lapply(liks, class))) != 1)
+         stop("All functions must have the same class")
+     ret <- function(pars, ...) {
+         ans <- lapply(liks, function(f) f(pars, ...))
+         sum(unlist(ans))
+     }
+     class(ret) <- c("combined", class(liks[[1]]))
+     ret
+ }
> lnL.multi <- constrain(combine(lnL.each), sAB ~ 0)
```

Now `lnL.multi()` can be used like any other likelihood function, for example

```
> chaparral.params <- c(0.19, 0.08, 0.29, 0.48, 1.29, 0.87)
> lnL.multi(chaparral.params)

[1] -390.7837
```

or in `find.mle()` or `mcmc()`.

References

Anacker, B. L., J. B. Whittall, E. E. Goldberg, and S. P. Harrison, 2011. Origins and consequences of serpentine endemism in the California flora. Evolution 65:365–376.

FitzJohn, R. G., W. P. Maddison, and S. P. Otto, 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology 58:595–611.

Goldberg, E. E., L. T. Lancaster, and R. H. Ree, in press. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Systematic Biology .

SimTreeSDD, 2010. Simulating phylogenetic trees under state-dependent diversification. URL http://tigger.uic.edu/~eeg/code/code.html.