Estimating rates of evolution for ecological data

Jeremy M. Beaulieu

Getting started

In today's demonstration, we are going to explore how to fit some regime-based Brownian motion and Ornstein-Uhlenbeck models to ecological data. We are going to rely on an empirical data set of Antirrhineae (snapdragons, toadflax, and their relatives) from a recent publication from my lab (Boyko et al. 2023). In this paper we wrote a pipeline that took a data set that consisted of life form information (i.e., annual and perennial) and then assembled BIOCLIM data. The raw data that we will be using should be found on the workshop website. We will focus on mean annual temperature (MAT) throughout, and you will need the following R packages:

```r
library(corHMM)
library(OUwie)
library(phytools)
```

Note the version of `OUwie` that works with this demonstration is the most up to date version on GitHub. You can download directly doing the following:

```r
library(devtools)
install_github("thej022214/OUwie")
```

Loading data and matching trait data with tip labels in a phylogeny

Be sure that you have R pointed to the proper working directory for where the raw tree and trait files are located. If that is the case, then let's load these raw data files into R:

```r
trait_raw <- read.csv("Antirrhineae-Gorospe_et_al-2020_bio_1_full.csv")
tree_raw <- read.tree("Antirrhineae-Gorospe_et_al-2020_cleaned.tre")
```

In the Boyko et al. (2023) publication, the first thing we did was run these files through the following function:

```r
organizeData <- function(phy, dat) {
    dat$species <- gsub(" ", "_", dat$species)
    dat <- dat[dat$species %in% phy$tip.label, ]
    dat <- dat[match(phy$tip.label, dat$species), ]
    dat <- dat[!dat$life_form == "no_life_form_on_database",
        ]
    plot_data <- data.frame(id = dat$species, life_form = as.factor(dat[,
        grep("life_form", colnames(dat))]), value = as.numeric(dat[,
        grep("mean", colnames(dat))]), se = as.numeric(dat[,
        grep("se", colnames(dat))]))
    plot_data <- plot_data[apply(plot_data, 1, function(x) !any(is.na(x))),
        ]
    pruned_phy <- keep.tip(phy, phy$tip.label[phy$tip.label %in%
        plot_data$id])
    return(list(dat = plot_data, phy = pruned_phy))
}
snapdragon <- organizeData(phy = tree_raw, dat = trait_raw)
```

What does this function do? Well often times when you download data the species names are Genus specific epithet. That was the case here. If you notice that within the raw tree file the tip labels have underscores instead of the spaces. Also, for the climate and life form data there are missing records for some species, and sometimes the tree is missing taxa for where we have trait data. So this function is *ONE* of the many ways to organize the data so that the trait and tree information exactly match.

We can take a look at the new output:

```r
dim(snapdragon$dat)
snapdragon$phy
```

Both should have 132 species, the trait file should have underscores in the species names, and the tree and trait sampling should match perfectly.

Let's also take a quick look how MAT is spread out across both taxa and time. For this we are going to use the `contMap()` function within the R package `phytools`:

```r
physig.dat <- snapdragon$dat$value
names(physig.dat) <- snapdragon$dat$id
contMap(tree=snapdragon$phy, x = physig.dat, fsize=.5)
```

Q: How much variation do we see across Antirrhineae in MAT?

Reconstructing a hypothesis

For the analyses within `OUwie`, we are testing whether a discrete character, in this case life form, can impact the evolution of a continuous character, in this case it is BIO1 (mean annual temperature). But, right now we only have the discrete character coded at the tips, and we need to determine how these "regimes" change across the tree. To do this, we can generate a map using an ancestral state reconstruction method in another package I've developed called `corHMM()`. Note that this package can do a lot! But, for the purposes of this demonstration we're going to do a very straightforward reconstruction of life form across the tree using a very simple model:

```r
anc.recon <- corHMM(snapdragon$phy, snapdragon$dat[, c(1, 2)],
    model = "ER", rate.cat = 1)
```

We can then plot the reconstruction to get a sense of the overall changes in life form across the tree using a built-in plotting function in `corHMM`:

```r
plotRECON(anc.recon$phy, likelihoods = anc.recon$states)
```

Q: What is the state for life form within Antirrhineae at the root of the tree? Is there a lot of certainty in this? What are the minimum number of changes in life form?

One thing to notice, however, is this:

```
anc.recon$phy$node.label
```

What you should see is that for the node labels, corHMM converted our life form data into a discrete character – into 1's and 2's. If you look at the corHMM() output it provides you with legend: annual was converted to a "1", and perennial was converted to a "2". This is fine, but we need to alter our input data set to deal with this:

```
life_form_coded <- numeric(length(snapdragon$dat$life_form))
life_form_coded[snapdragon$dat$life_form == "annual"] <- 1
life_form_coded[snapdragon$dat$life_form == "perennial"] <- 2
life_form_coded <- as.numeric(life_form_coded)
snapdragon$dat <- cbind(snapdragon$dat, life_form_coded)
```

That last line appended our the data file within the snapdragon object with our new column of discretized life form data.

Fitting single evolutionary rate models

First we will fit a simple, single rate Brownian motion model, BM1. Note that the input files. We are inputting the phylo object that contains labels for the internal nodes and specific columns from our cleaned snapdragon data set. There is a specific order: species names, regime scoring, and the continuous character:

```
bm1 <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "BM1", algorithm = "three.point")
bm1
```

Take note of the AIC, and note the parameter sigma.sq – this is the Brownian motion variance parameter. That is, how fast the trait evolves on the tree in units of variance per Myr. Also, an interesting thing to note in the output is that under any BM model, there are no "optima". Here the optima actually represents the ancestral state for the whole tree (under BM where we start is where we expect to be after a stretch of time *on average*).

Next we will fit an OU model with a single global optimum, OU1, applied to the whole tree:

```
ou1 <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OU1", algorithm = "three.point")
ou1
```

Take note again of the AIC, as well as the parameter sigma.sq. We also have a new parameter, alpha, which tells us about the strength of the "pull" towards the optima. Here the optima takes on a new meaning in the sense that the trait is being pulled towards this specific value.

Which model fits best? We need to find the one with the minimum AIC:

```
which.min(c(bm1$AIC, ou1$AIC))
```

Should suggest OU1 is the better fit. In other words, the lack of variation in MAT that we observed above is consistent with an Orstein-Uhlenbeck model of evolution. But, let's take a look at the model estimate.

We see a pretty low rate of evolution. The half-life is also a good way of understanding the OU model. The half-life is the expected time it takes for the influence of the ancestral trait value to decay by half, under the stabilizing pull toward the optimum. In other words, after one half-life, the expected deviation from the optimum is only 50% of what it was at the start. Bottom-line: larger alpha, lower half-life, ancestral influence lost quickly; smaller alpha, higher half-life, ancestral influence persists.

So, here, we see that the half-life is 3.82 Myr. The total length of the tree is 44 Myr, so the influence of the ancestor gets lost very quickly. In other words, After 3.82 time units, the expected deviation from the optimum is only half of what it was initially. After 2 × 3.82 = 7.64 units, only 25% of the initial deviation remains; after about 5 half-lives (~19 units), less than 5% remains. In practical terms, under this model trait values near the tips are determined almost entirely by the stochastic variation around the optimum, with very little direct influence from the root state.

Fitting multiregime models

Remember what we really want to know is whether life form can impact the evolution of MAT within snapdragons. So we need to fit more complex models to better address this question. There are several more complex models we can try. For example, an OU model that has an specific optimum for each of our two regimes – this is called OUM. We can also fit an OU model that allows for the rate of evolution to vary as well as the optima between our two regimes (OUMV). We can also look at whether just the rate differs by life form – BMS. Let's fit them and then compare their fit using AIC:

```
bms <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "BMS", algorithm = "three.point")
oum <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OUM", algorithm = "three.point")
oumv <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OUMV", algorithm = "three.point")
which.min(c(bm1$AIC, ou1$AIC, bms$AIC, oum$AIC, oumv$AIC))
```

Q: Which model is the best fit of the set we tried?

Let's take a look at the output for this model. One thing to notice is the "best" model is suggesting a different optimal MAT between annual vs. precipitation. But like, come on, 5.667037 vs. 5.659484? Well one thing to understand is that the climate variables are log transformed and in units of Kelvin. So to really understand what the model is telling us, we need back transform the optima estimates:

```
annual.mat <- exp(5.667037) - 273.15
perenn.mat <- exp(5.659484) - 273.15
annual.mat - perenn.mat
```

So that's a near 2 degree difference, on average, for annuals than perennials. In other words, the model suggests that annuals within Antirrhineae are able to deal with generally higher temperatures than perennials.

"Tip fog"

As ecologists, hopefully, maybe, some of you are like, *Jeremy, you are just using the species means. Nature is full of variation. Climate means? Come on. What about their entire range?* I couldn't agree more.

Recently, we found that errors in measuring evolutionary change, whether in speciation rates or body size evolution, can profoundly bias rate estimates and comparative analyses of rate shifts within phylogenies. While such errors are often assumed negligible, they encompass broader uncertainties, termed variously as "specific variances," "residual variation," or "measurement error". To avoid ambiguity, we propose the term "tip fog" to describe the variance between true species means and observed values at present. This concept applies broadly to both discrete and continuous traits. Our methods for estimating tip fog demonstrate its significance in comparative studies, urging its inclusion as a standard correction to prevent biased parameter estimates and spurious model selection. We have recently published a paper showing how tip fog can be estimated in various models of continuous and discrete trait evolution (Beaulieu & O'Meara, 2025).

Let's refit our set of models and take estimate this so-called "tip fog" and see how this impacts our model fits:

```r
bm1 <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "BM1", algorithm = "three.point", tip.fog = "estimate")
ou1 <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OU1", algorithm = "three.point", tip.fog = "estimate")
bms <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "BMS", algorithm = "three.point", tip.fog = "estimate")
oum <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OUM", algorithm = "three.point", tip.fog = "estimate")
oumv <- OUwie(anc.recon$phy, snapdragon$dat[, c("id", "life_form_coded",
    "value")], model = "OUMV", algorithm = "three.point", tip.fog = "estimate")
which.min(c(bm1$AIC, ou1$AIC, bms$AIC, oum$AIC, oumv$AIC))
```

Q: What is the best fit model now? Did the results change?

Also take note of the tip fog estimate relative to the rate of evolution. Tip fog is a direct estimate of variance from fog; variance from the process depends on the details of the tree and rate, but an easy way to think about it is the wiggling that happens along the branches. An upper estimate of that is just the height times the rate. We can see for this tree of height 44 MY, we get 6.568824e-05 variance units from tip fog and 43*6.0132e-06 from the evolutionary process. So in this case, 6.568824e-05 / (6.568824e-05 + 0.0002645808) = 20% of the variance comes from just the tip fog process. In practical terms, when tip fog is ignored, the model interprets this as evidence for life form-specific differences in the optima.

Extensions

Within the `corHMM` framework, we have also incorporated a way to model "tip fog" — uncertainty in the true state of a species. Such uncertainty can arise from data collection errors or genuine polymorphisms, and it can bias transition rate estimates and exaggerate biological patterns. This, in turn, may distort the reconstructed regime mapping across the tree. The package `corHMM` includes an option to estimate character misassignments and state transition rates simultaneously, helping to mitigate these effects. In this demonstration, I have shown the regime mapping under a simple single-rate transition model, but `corHMM` supports far more complex discrete evolution models that could influence regime mapping. Similarly, we have recently introduced `hOUwie` (Boyko et al. 2023), an extension of `OUwie` that estimates both regime mappings and BM/OU rate models simultaneously. This approach improves null-hypothesis testing and lets the continuous trait inform the regime mapping, and vice versa, in a single function call.

References

Beaulieu, J.M., and B.C. O'Meara. (2025). Navigating "tip fog": embracing uncertainty in tip measurements. Evolution, In press, qpaf067.

Beaulieu, J.M., B.C. O'Meara, and M.J. Donoghue. (2013). Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Systematic Biology 62: 725-737.

Beaulieu, J.M., D-J. Jhueng, and B.C. O'Meara. (2012). Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution. 66: 2369- 2383.

Boyko, J.D., E.R. Hagen, J.M. Beaulieu, and Vasconcelos, T. (2023). The evolutionary responses of life-history strategies to climatic variability in flowering plants. New Phytologist 240: 1587-1600.

Boyko, J.D., B.C. O'Meara, and J.M. Beaulieu. (2023). A novel method for jointly modeling the evolution of discrete and continuous traits. Evolution, 77:836-851.