Lab 13:
Probabilistic Models of Geographic Range Evolution

By Will Freyman
Modified by Ixchel González Ramírez

1 Before you begin

Today's lab will be completed in R. Please install R: `https://www.r-project.org/`

We'll be using the excellent R package **BioGeoBears** [Matzke, 2013] which was developed by Nick Matkze, who was a GSI of this course! His website (and the source of the example script we'll use) is here: `http://phylo.wikidot.com/biogeobears`

2 Introduction to probabilistic biogeography models

Probabilistic modeling of geographic range evolution allows us to use the standard set of statistical model choice procedures (e.g. AIC, BIC, likelihood ratios, etc.) to test different biogeographical scenarios, and to make statistically sound inference of ancestral geographic ranges over a phylogeny. For example, we can test whether the observed biogeographic distribution of a clade is best explained with a model that allows for vicariance and long-distance dispersal versus a model that allows for only vicariance.

There are many parsimony based approaches for studying biogeography that will be covered in lecture. The first likelihood approaches to biogeography were presented in the ground breaking Ree and Smith [2008], which introduced the Dispersal-Extinction-Cladogenesis (DEC) model. We'll discuss these probabilistic models and their many extensions in lecture (also see Figure 1 for a summary), so here we'll focus on applying them to the the phylogeny of the Hawaiian members of the plant genus *Psychotria* published in Ree and Smith [2008].

3 Using BioGeoBears with the DEC and DEC+J models

We will use BioGeoBears to implement the 2-parameter Dispersal-Extinction-Cladogenesis (DEC) model from Ree and Smith [2008] as well as the DEC+J model that includes a long distance dispersal (jump or founder event speciation) parameter and was described in Matzke [2014]. See Figure 1 for details on the different models. We'll use the Likelihood Ratio Test (LRT) and Akaike information criterion (AIC) to see which model fits the data better.

3.1 Install and setup BioGeoBears

Start up R and be sure to set your working directory so that you can find the output file we will be generating (using the function `setwd()`). Install the BioGeoBears package and useful dependencies

```
#install useful packages
install.packages("rexpokit")
install.packages("cladoRcpp")
#install BioGeoBEARS from github repository
library(devtools)
devtools::install_github(repo="nmatzke/BioGeoBEARS")
```

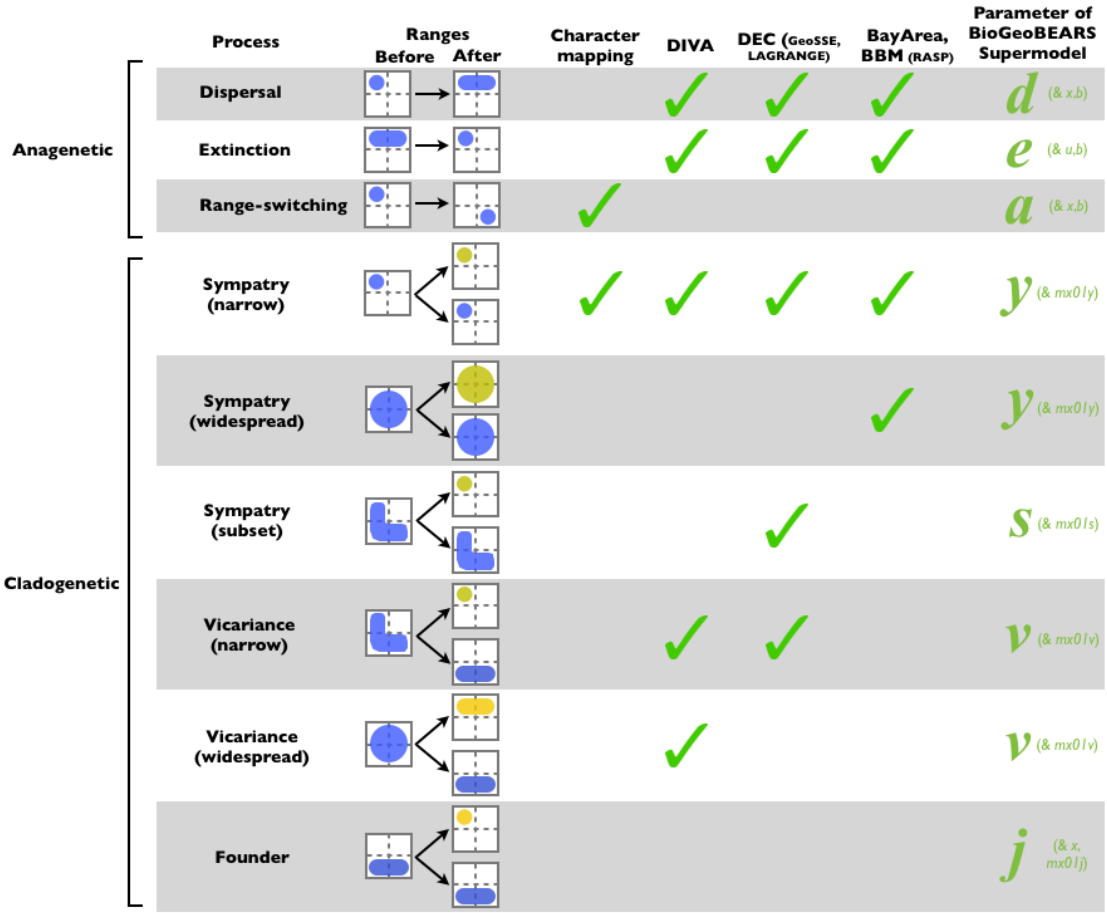

Figure 1: The processes assumed by different historical biogeography methods. Each of these processes is controlled by the specified parameter(s) in the BioGeoBEARS supermodel, allowing them to be turned on or off, or estimated from the data. Note that whether or not the data support a particular free parameter is an empirical question that should be tested with model choice procedures. Note also that this graphic deals only with the range-changing processes assumed by the different methods. BioGeoBEARS does not attempt to replicate e.g. the parsimony aspect of DIVA, just the processes allowed by DIVA (the BioGeoBEARS "DIVA" model could be called "DIVALIKE" to emphasize that it is a likelihood implementation of the processes assumed by DIVA). Text and image from `http://phylo.wikidot.com/biogeobears#BioGeoBEARS_supermodel_graphic`.

Now lets load the package and some dependencies:

```
#loading packages
library(optimx)
library(FD)
library(snow)
library(parallel)
library(BioGeoBEARS)
```

And finally, let's get the directory of the example data that was installed with BioGeoBears:

```
extdata_dir = np(system.file("extdata", package="BioGeoBEARS"))
```

3.2 Load the phylogeny

This is the phylogeny of the Hawaiian members of the plant genus *Psychotria* from Ree and Smith [2008].

```
tree_file_name = np(paste(addslash(extdata_dir), "Psychotria_5.2.newick", sep=""))
tr = read.tree(tree_file_name)
```

Let's plot the tree:

```
plot(tr)
title("Example Psychotria phylogeny from Ree & Smith (2008)")
axisPhylo()
```

3.3 Load the geography data

Now we need to load data on the geographic range of each extant species in our phylogey.

```
geo_file_name = np(paste(addslash(extdata_dir), "Psychotria_geog.data", sep=""))
tipranges = getranges_from_LagrangePHYLIP(lgdata_fn=geo_file_name)
```

Let's take a look at the geographic range data

```
tipranges
```

Question 1:
Geographic range is being modeled here as a discrete character state, where the character state represents the combination of areas that a lineage inhabits at any given point in time. Models of biogeographic range evolution are essentially no different than the other models of discrete character evolution we have looked at over the course of the semester (with the exception of including cladogenetic events).

1. DNA substitution models have 4 discrete states. How many states will our biogeographic model have? Remember, these models allow for a lineage to be in more than one area at a time.

2. Does it make sense to allow the character state `0 0 0 0`? What would this represent?

3. To calculate likelihoods of discrete character evolution models we have to exponentiate the transition matrix, which can be very computationally demanding for large matrices. If we were performing inference on a dataset with 10 areas instead of 4, how many states would our model have? How large would the anagenetic transition rate matrix be?

4. To model cladogenetic change we need an additional transition probability matrix that represents probabilities for all combinations of the state before cladogenesis and after on each of the two daughter lineages. Given 10 areas, how large would the cladogenetic transition probability matrix be?

3.4 Setup the DEC model

First we will run the analysis using the DEC model. The DEC is the default model in BioGeoBears, so it is relatively straightforward to setup.

```
BioGeoBEARS_run_object = define_BioGeoBEARS_run()
```

Give BioGeoBEARS the location of the example input files:

```
BioGeoBEARS_run_object$trfn = tree_file_name
BioGeoBEARS_run_object$geogfn = geo_file_name
```

And let's configure our analysis. If you are running this on your own dataset be sure to adjust the maximum range size (that is the maximum number of areas a lineage can inhabit at an given point – usually this is just the total number of areas in your dataset).

```
BioGeoBEARS_run_object$max_range_size = 4
BioGeoBEARS_run_object$min_branchlength = 0.000001
BioGeoBEARS_run_object$include_null_range = TRUE
BioGeoBEARS_run_object$num_cores_to_use = 1
BioGeoBEARS_run_object$force_sparse = FALSE
BioGeoBEARS_run_object = readfiles_BioGeoBEARS_run(BioGeoBEARS_run_object)
BioGeoBEARS_run_object$return_condlikes_table = TRUE
BioGeoBEARS_run_object$calc_TTL_loglike_from_condlikes_table = TRUE
BioGeoBEARS_run_object$calc_ancprobs = TRUE
```

Now we are ready to run the analysis:

```
results_DEC = bears_optim_run(BioGeoBEARS_run_object)
```

Question 2:
Take a look at the `results_DEC` object. What is the maximum likelihood estimate of the rate of anagenetic "dispersal" (range expansion)? And the rate of anagenetic "extinction" (range contraction)?

3.5 Setup the DEC+J model

Now we will run the analysis using the DEC+J model. This include the "jump" parameter for long distance dispersal / founder speciation events.

```
BioGeoBEARS_run_object = define_BioGeoBEARS_run()
```

Give BioGeoBEARS the location of the example input files:

```
BioGeoBEARS_run_object$trfn = tree_file_name
BioGeoBEARS_run_object$geogfn = geo_file_name
```

Just like before, let's configure the analysis. If you are running this on your own dataset be sure to adjust the maximum range size. These settings are all the same as they were for the DEC model above.

```
BioGeoBEARS_run_object$max_range_size = 4
BioGeoBEARS_run_object$min_branchlength = 0.000001
BioGeoBEARS_run_object$include_null_range = TRUE
BioGeoBEARS_run_object$num_cores_to_use = 1
BioGeoBEARS_run_object$force_sparse = FALSE
BioGeoBEARS_run_object = readfiles_BioGeoBEARS_run(BioGeoBEARS_run_object)
BioGeoBEARS_run_object$return_condlikes_table = TRUE
BioGeoBEARS_run_object$calc_TTL_loglike_from_condlikes_table = TRUE
BioGeoBEARS_run_object$calc_ancprobs = TRUE
```

And now finally, set up DEC+J model. We'll use the maximum likelihood parameter value estimates from the DEC analysis we already ran to get good starting points for the hill-climbing heuristic.

```
dstart = results_DEC$outputs@params_table["d","est"]
estart = results_DEC$outputs@params_table["e","est"]
```

Set the starting values in our analysis object:

```
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["d","init"] = dstart
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["d","est"] = dstart
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["e","init"] = estart
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["e","est"] = estart
```

The DEC is a 2-parameter model nested within the 3-parameter DEC+J, so we need to add J as a new free parameter to estimate. We also need to assign it an initial value.

```
jstart = 0.0001
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["j","type"] = "free"
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["j","init"] = jstart
BioGeoBEARS_run_object$BioGeoBEARS_model_object@params_table["j","est"] = jstart
```

Now we are ready to run the analysis:

```
results_DECJ = bears_optim_run(BioGeoBEARS_run_object)
```

3.6 Plotting ancestral states

Let's plot the results of the two models and look at the ancestral state reconstructions.

```
pdffn = "Psychotria_DEC_vs_DEC+J.pdf"
pdf(pdffn, width=6, height=6)
```

Plot the DEC ancestral states. These plots will be written to a PDF file.

```
analysis_titletxt = "DEC on Psychotria"
results_object = results_DEC
scriptdir = np(system.file("extdata/a_scripts", package="BioGeoBEARS"))
res2 = plot_BioGeoBEARS_results(results_object, analysis_titletxt, addl_params=list("j"), plotwhat="text",
          label.offset=0.45, tipcex=0.7, statecex=0.7, splitcex=0.6, titlecex=0.8, plotsplits=TRUE,
          cornercoords_loc=scriptdir, include_null_range=TRUE, tr=tr, tipranges=tipranges)
plot_BioGeoBEARS_results(results_object, analysis_titletxt, addl_params=list("j"), plotwhat="pie",
          label.offset=0.45, tipcex=0.7, statecex=0.7, splitcex=0.6, titlecex=0.8, plotsplits=TRUE,
          cornercoords_loc=scriptdir, include_null_range=TRUE, tr=tr, tipranges=tipranges)
```

Plot the DEC+J ancestral states to the PDF file.

```
analysis_titletxt ="DEC+J on Psychotria"
results_object = results_DECJ
scriptdir = np(system.file("extdata/a_scripts", package="BioGeoBEARS"))
res1 = plot_BioGeoBEARS_results(results_object, analysis_titletxt, addl_params=list("j"), plotwhat="text",
          label.offset=0.45, tipcex=0.7, statecex=0.7, splitcex=0.6, titlecex=0.8, plotsplits=TRUE,
          cornercoords_loc=scriptdir, include_null_range=TRUE, tr=tr, tipranges=tipranges)
plot_BioGeoBEARS_results(results_object, analysis_titletxt, addl_params=list("j"), plotwhat="pie",
          label.offset=0.45, tipcex=0.7, statecex=0.7, splitcex=0.6, titlecex=0.8, plotsplits=TRUE,
          cornercoords_loc=scriptdir, include_null_range=TRUE, tr=tr, tipranges=tipranges)
```

And finally let's check out our PDF file. We'll see each model's ancestral state reconstructions plotted twice: 1) with the the maximum likelihood estimate of the ancestral ranges, and 2) with pie charts at each node that show the probability of each ancestral state.

```
dev.off()
cmdstr = paste("open ", pdffn, sep="")
system(cmdstr)
```

Question 3:
Compare the estimated ancestral ranges of the lineages leading to `P_hexandra_Oahu` all the way back to the root of the tree. Explain the results in context of biogeographic hypothesis testing. Which hypothesis makes more sense to you given Hawaiian island geography?

3.7 Model testing

Let's use the Likelihood Ratio Test (LRT) and Akaike information criterion (AIC) to see which model fits the data better.

First get the log likelihoods:

```
LnL_1 = get_LnL_from_BioGeoBEARS_results_object(results_DECJ)
LnL_2 = get_LnL_from_BioGeoBEARS_results_object(results_DEC)
```

The AIC is calculated with the log likelihoods and the number of parameters in each model:

```
numparams1 = 3
numparams2 = 2
```

Calculate AIC:

```
#calculate stats
stats = AICstats_2models(LnL_1, LnL_2, numparams1, numparams2)
#AIC values
stats$AIC1
stats$AIC2
# pvalue of likelihood ratio test
stats$pval
```

Question 4:

1. Which model does the AIC support?
2. These models incorporate cladogenetic evolutionary events, where evolutionary change occurs at speciation events. However, in our reconstructed phylogenies we usually only consider the speciation events that led to the extant taxa. How might unobserved speciation events (lineages that went extinct) affect our inferences?

4 More complex models and Bayesian approaches

BioGeoBears can test many other models besides DEC and DEC+J, enabling other biogeographic scenarios to be tested. For example, the effect of distance on dispersal events between

areas can be modeled using the x parameter, which allows us to model a lower probability of dispersal to distant regions. BioGeoBears can also use time stratified models, in which different areas are available to be inhabited at different points in time over the phylogeny. Bruce Baldwin and Will Freyman (a brilliant former GSI of this class) used this approach to reconstruct the colonization of the Hawaiian islands from California of the Silversword alliance (an adaptive radiation in the Asteraceae) given the different emergence times of the Hawaiian islands.

Another previous GSI of our class, Michael Landis, has pioneered Bayesian approaches to biogeographic modeling, both in his software BayArea [Landis et al., 2013] and RevBayes [Höhna et al., 2014]. Nearly all the models described above have been implemented in RevBayes. Furthermore, Michael has extended biogeographic models to use paleogeographic dates to help calibrate speciation times, allowing for simultaneous inference of ancestral geographic ranges and divergence times [Landis, 2015].

Again I highly recommend you to visit the BioGeoBears and RevBayes pages for exploring more biogeographic models.
`http://phylo.wikidot.com/biogeobears#script`
`https://revbayes.github.io/tutorials/`

Please email me the following:

1. Your PDF of the ancestral state reconstructions.
2. The answers to questions 1-4.

References

Sebastian Höhna, Tracy A Heath, Bastien Boussau, Michael J Landis, Fredrik Ronquist, and John P Huelsenbeck. Probabilistic graphical model representation in phylogenetics. *Systematic biology*, 63(5):753–771, 2014.

Michael J Landis. Biogeographic dating of speciation times using paleogeographically informed processes. *bioRxiv*, page 028738, 2015.

Michael J Landis, Nicholas J Matzke, Brian R Moore, and John P Huelsenbeck. Bayesian analysis of biogeography when the number of areas is large. *Systematic biology*, page syt040, 2013.

Nicholas J Matzke. Biogeobears: biogeography with bayesian (and likelihood) evolutionary analysis in r scripts. *R package, version 0.2*, 1, 2013.

Nicholas J Matzke. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology*, 63(6):951–970, 2014.

Richard H Ree and Stephen A Smith. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*, 57(1):4–14, 2008.