Integrative Biology 200B *University of California, Berkeley*
"Systematics" *Spring 2011*
by Nick Matzke

Lab 15: Maximum Likelihood Estimation of Biogeographic History on Phylogenies using DIVA and Lagrange

---

Setup

The main goal for today's lab is just to get these programs working. This can be a bit of a challenge – but part of the skill set you need in order to do evolutionary analyses is the ability to figure out how to get scientific software to function.

Summary of lab

If you have done DIVA and Python-Lagrange in IB200a, you can skip those parts and focus on getting the C++ version of Lagrange to work (and, help others get DIVA/Lagrange to work, work through the example files with them, etc.).

If you can't get some version of Lagrange to work (or if e.g. you are doing the long slow download of the Enthought Python Distribution for Python Lagrange), work with someone who does have it working.

The lab files (and some, but not all executables) are downloadable most simply here:
http://ib.berkeley.edu/courses/ib200b/labs/lab15/lab15.zip

What to turn in

Read the questions in the lab and answer them for yourself, but all I want you to turn in is:

- an email saying what you did/didn't get to work. Probably everyone should be able to get DIVA to work. Lagrange will be harder. If you end up looking on someone else's computer when they run Lagrange, say that.
- Ignore the other instructions to email answers to questions
- I just want to get a sense of how feasible people in the class found these programs to be.

Background on why academic software is sometimes hard to use

Scientists are not trained software developers, they usually have no staff who can work on usability and graphical user interfaces and the like, and they often don't have the ability/time to keep software up to date with continually changing operating systems at the like. For all of that, the software often does amazing things – just don't expect it to be as easy as Microsoft Word and their thousands of software engineers, user trials, and billions of dollars.

The main point of saying the above is to encourage patience and tolerance when you are trying to use academic software.

Some useful guidelines on attempting to install academic software (derived from long, hard experience)

- The easiest option is always to find an "executable" or "binary" version of a program that is appropriate for your operating system (OS). These have been compiled for your OS ("precompiled binaries") by someone else. Typically all you have to do with these is download them, put them in a desired directory, and run them from the command line.
- If you don't have an executable, you can try to compile the software yourself from the source code. For well-established packages that are used by hundreds of thousands of people, this often works (you have to have the right compilers installed on your machine). For specialist software like phylogenetics software, in my experience it's a 50-50 thing.
- If you are stuck trying to compile something from source, sometimes you will discover that you have to compile another package, which itself requires another package, etc. This can get *extremely* frustrating. Even after putting a lot of time in, and make an attempt to install all dependencies, in my experience it's still only maybe a 75% chance that some bit of software will successfully compile and work.
- Some software packages are available via things like *fink* and *port*, which attempt to find and download all dependencies for you. This can be useful for common software libraries which might be dependencies for your specialist program.
- When you get error messages, the first thing to do is **google the error message** and see what people say about it.
- It is possible to waste *days* messing around with getting software to work, so an important rule is: KNOW WHEN TO GIVE UP. Give it an hour or two, and if you're not making progress, GIVE UP. Try emailing colleagues that use the software, and/or the authors of the software.
- Sometimes the best option is just to have someone else run your data on their computer, which has the ability to run the software. This is one reason collaborations happen.
- Once you have a compiled executable, ALWAYS START WITH THE RUNNING THE EXAMPLE FILES. This will show you how the formatting etc.

works, and will also double-check whether or not the program works, before you invest a lot of time in reformatting your data.

Setup for DIVA – Dispersal-Vicariance Analysis

DIVA was written by Fredrik Ronquist (coauthor with John Huelsenbeck of MrBayes). For a long time only a Windows executable was available. Recently, Jonathan Nylander compiled a Mac version.

The Windows version used to be here:
http://www.ebc.uu.se/systzoo/research/diva/diva.html

…but that website seems to have disappeared. Therefore, I am putting both versions, along with the original manual, etc., on our website here:

http://ib.berkeley.edu/courses/ib200b/labs/lab15/

DOWNLOAD EVERYTHING INTO THE SAME DIRECTORY, e.g.:
something/something/lab15/

This is mostly simply done by downloading the zipfile:
http://ib.berkeley.edu/courses/ib200b/labs/lab15/lab15.zip

Lagrange: general background

Lagrange was written by Stephen Smith and Rick Ree.

- The original version was written in Python and will work on any system (after you have successfully installed the Enthought Python Distribution)

- The new version, currently in Beta form, was written in C++ and is much faster (probably 100s of times faster). However, compiled binaries only currently exist for Mac 10.4 and 10.6 (the 10.6 might work on 10.5, we'll see). If you are a computer jock you might try compiling the source code for Windows, but it took me a whole weekend of plinking to successfully compile it for 10.4, so be advised.
- The C++ version is totally rewritten from the Python version. As far as I know, they use different input files etc.

Lagrange C++ BETA version

Basically, if you have a Mac, you should see if one of these compiled binaries will work:

http://code.google.com/p/lagrange/downloads/list

If they don't, or if you have a Windows PC, try getting the Python version to work.

Lagrange Python Version

a. First, download and install the Enthought Python Distribution (free academic download, has all Lagrange requirements: http://www.enthought.com/products/edudownload.php

   (However it is very large, ~1 GB; you might wait to download until you are plugged into a wired connection, it may take a long time over AirBears)

b. (see the end of lab for additional help in getting the install recognized by the Mac Terminal or Windows Command Line)

c. Lagrange Python version: http://code.google.com/p/lagrange/

d. Website that generates input files for Python Lagrange: http://www.reelab.net/lagrange/configurator/index

---

Biogeography: Background

Today we are going to be looking at programs that compare trees between two associated groups of objects to deduce their common history. This could be a comparison of host & parasite, organism & gene or area & organism trees. The different relationships can be analogized like this:

Host	Organism	Area
Parasite	Gene	Organism
Host switch	Horizontal transfer	Dispersal
Cospeciation	Orthology	Vicariance
Parasite speciation on one host	Gene duplication or allelic divergence	Sympatric speciation (kind of)
Parasite extinction	Gene loss or fixation	Extinction

In all three cases comparisons can be made between the two trees to see how often dispersal or vicariance (or their analogous events) best explains the situation.

There are several different approaches to this, for example parsimony reconstruction of character states on a cladogram, where the "character states" are simply geographic

location. Additional methods involve converting the phylogenies of several groups into matrices, relabeling the species as their areas, then building a supertree of the areas, which will allegedly represent the common geographical signal between the clades input into the supertree (MRP: Matrix Representation of Parsimony).

Today we will just deal with two of the most popular methods, which were explicitly designed for biogeographical purposes.

DIVA

DIVA is a program by Fredrick Ronquist, which is (was) freely available on the web at http://www.ebc.uu.se/systzoo/research/diva/diva.html (Windows version; see above for Mac version). Furthermore, it does not just maximize the number of cospeciation events, but instead has a cost matrix that describes the cost of all possible events. Should these assumptions be different for a host-parasite as opposed to an area-organism reconstruction? *DIVA* does not require a cladogram for the relationships among different areas. It only requires a tree describing the relationship between the different taxa and a description of which areas those taxa are associated with.

(Note: although DIVA does technically allow extinction events, when it is running on defaults it does not use them, since any reduction in range can be explained at no cost as a vicariance event from a more widespread ancestor.)

Open the file fruits_geog.txt in a text editor. This file describes the relationships and distributions of several species of domestic fruit. The matrix is a description of where the taxa are found. A 0 represents absence from that area and a 1 indicates presence. Following that is a tree describing the relationship of the taxa. Can you read the tree in this format to understand the relationships?

1. Open *DIVA*. (move to the DIVA directory, type ./diva on Mac Terminal)

2. Type these commands:

```
output fruits_geog_output.txt;
echo status;
proc fruits_geog.txt;
optimize;
quit;
```

It will quickly optimize the data to fit the tree, although it would take much longer if you had more taxa or more areas.

To understand the output file (open it) you must recognize two things. First what is the tree that describes the relationship among the taxa. Each line of output gives the name of only two taxa to describe a node so you must know the tree to understand what other taxa

are descended from that node. The second thing you need to know is the order that the areas were listed in, as *DIVA* refers to them only with letters. Thus **A** refers to the first area listed, South America, B to Africa, etc.

This should be enough for you to interpret the output. For example the second line of output:

Node 10 (anc. of terminals orange-kiwi):E

means that the common ancestor of oranges, bananas, papayas and kiwis lived in Asia.

Try it again with the bird data file bird_tree_diva_in_file.txt. Commands:

```
output tmp_diva_out_file;
echo status;
proc bird_tree_diva_in_file.txt;
optimize maxareas=2;
quit;
```

Try this again, increasing maxareas to 3, then 4, etc. "maxareas" is the maximum number of areas any species is allowed to inhabit. Stop when it starts getting slow.

Answer these questions on email:

- At what level of maxareas does it begin to get slow?
- How would one chose what to set for "maxareas"? (i.e., think of your own study groups, and what a reasonable choice might be)
- Look at:

  Donoghue, M. J. and Moore, B. R., 2003. Toward an Integrative Historical Biogeography. *Integrative and Comparative Biology*. 43 (2), 261-270. http://dx.doi.org/10.1093/icb/43.2.261

  What is their main critique of DIVA-like approaches, and what do you think of their point?

Lagrange: Background

For many years DIVA (Ronquist, 1996, 1997) and a few even older pattern-based methods have been the standard methods in historical biogeography. However, DIVA was published in 1997, and is considered obsolete by many. Furthermore, even the author

of DIVA, Fredrick Ronquist (coauthor of MrBayes), says his biggest regret in his academic career is that people still use DIVA.

Nevertheless, DIVA was the best thing going for many years, and was used in some interesting large-scale analyses of plants and animals (Donoghue and Smith, 2004; Sanmartin and Ronquist, 2004). It was useful in that it was an "event-based" method, instead of a "pattern-based" method (Ronquist, 1996), i.e. it explicitly hypothesized a history of events, and then sought the history that minimized the number of dispersal and extinction events. Scientists could run the program and then conclude that X number of dispersal events occurred between Island A and B, Y number between B and C, and do this for each clade of interest.

This whole approach was criticized in:

Donoghue, M. J. and Moore, B. R., 2003. Toward an Integrative Historical Biogeography. *Integrative and Comparative Biology*. 43 (2), 261-270.

…which argued that biogeographical histories and patterns were not very useful without an explicit time component. E.g., the same pattern could be produced by different events and different times, and the available methods would not point this out. Congruence, typically taken as strong evidence of common history, could in biogeography very easily be due to "pseudocongruence." In addition, time estimates for biogeographic events were often either much too early or too late for the geological/climatic events that had been hypothesized to be behind inferred vicariance events (de Queiroz, 2005; Bush *et al.*, 2006).

From 2005-present, Rick Ree, Stephen Smith, Brian Moore, and others have developed a maximum-likelihood method for inference in historical biogeography:

Ree, R. H., Moore, B. R., Webb, C. O. and Donoghue, M. J., 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*. 59 (11), 2299-2311.

Ree, R. H. and Smith, S. A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst Biol*. 57 (1), 4-14.

Moore, B. R., Smith, S. A., Ree, R. H. and Donoghue, M. J., 2009. Incorporating Fossil Data in Biogeographic Inference: A Likelihood Approach. *Evolution*. In press.

Lagrange

The currently available program "Lagrange" (**L**ikelihood **A**nalysis of **G**eographic **R**ange **E**volution) is from Ree & Smith (2008) (the 2005 version was very complex and much slower). The figures below are from this paper.

The Lagrange program takes as input:

1. an ultrametric phylogeny (nodes are dated)
2. locations of the tips
3. a list of possible ranges (area 1, area 2, area 1+2, etc.)
4. area adjacency matrix (which areas are connected such that they could share the same species)
5. dispersal matrix (relative probability of dispersal between regions; note that adjacent areas will not have a higher rate of dispersal unless you specify this explicitly here)

Unlike DIVA, which calculates the number of dispersal and extinction events and tries to minimize them, Lagrange works down the tree to calculate the relative likelihood of each possible ancestral range at each node, given a particular probability of dispersal and extinction. Here is the rate matrix:

$$Q = \left[\begin{array}{c|cccccccc} & \emptyset & 1 & 2 & 3 & 12 & 13 & 23 & 123 \\ \hline \emptyset & — & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & E_1 & — & 0 & 0 & D_{12} & D_{13} & 0 & 0 \\ 2 & E_2 & 0 & — & 0 & D_{21} & 0 & D_{23} & 0 \\ 3 & E_3 & 0 & 0 & — & 0 & D_{31} & D_{32} & 0 \\ 12 & 0 & E_2 & E_1 & 0 & — & 0 & 0 & D_{13} + D_{23} \\ 13 & 0 & E_3 & 0 & E_1 & 0 & — & 0 & D_{12} + D_{32} \\ 23 & 0 & 0 & E_3 & E_2 & 0 & 0 & — & D_{21} + D_{31} \\ 123 & 0 & 0 & 0 & 0 & E_3 & E_2 & E_1 & — \end{array}\right]. \quad (1)$$

E1-E3 are instantaneous extinction rates (all the same in our example), the Ds are the instantaneous dispersal rates. This rate matrix is exponentiated to give the probability of change as a function of time (branch length):

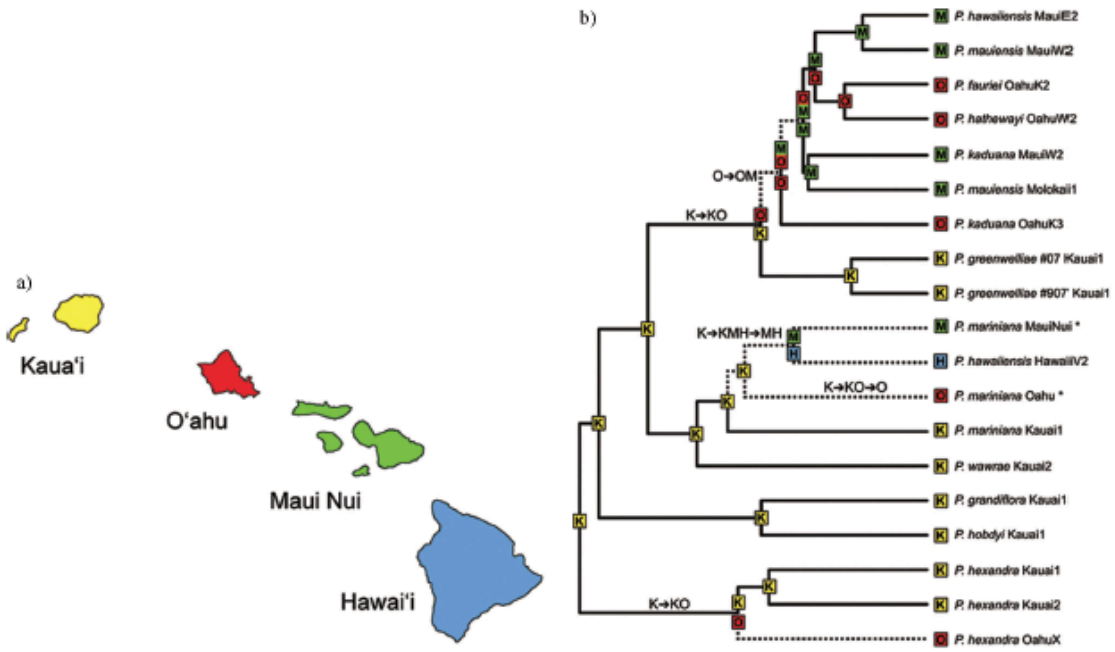
$\mathbf{P}(t) = e^{-Qt}$

Thus, for an ancestral node, the likelihood of it being in Area 1 can be calculated given the ranges of the two daughter nodes, and their branch lengths (distance in time) to the ancestral node.

Using the above, the algorithm can calculate the likelihood for a whole history on a phylogeny, and then vary the extinction and dispersal parameters, calculate again, etc., optimizing for the ML estimates of dispersal and extinction rates. The output consists of the history resulting in the maximum likelihood, its log likelihood, and the estimated rates.

Via the input files, the user can prohibit certain histories (i.e., if an island doesn't exist at a certain point in time) or events (i.e., disallow certain dispersals), and then compare the likelihoods with a less constrained model.

Here is Ree & Smith's inference for their example dataset, *Psychotria*, with an unconstrained model (no blockage of certain dispersals, range can be any combination of islands):



Here is the log likelihood of each possible ancestral range for the root of *Psychotria*, for the unconstrained (M0) and more constrained models:

TABLE 1. Inferences about the ancestral area and range evolution parameters of Hawaiian *Psychotria* under DEC models. The unconstrained model (M0) allows geographic ranges to include any combination of islands in the archipelago and permits direct dispersal between any pair of islands. M1 and M2 restrict ranges to include a maximum of two adjacent islands. M2 further limits dispersal to be eastward between adjacent islands. The stratified model permits dispersal to islands only after their time of geological origin, thus with a root age of 5.1 Ma, the only ancestral area possible is Kaua'i.

Model	Area	$-\ln(L)$	Dispersal	Extinction
M0	Kaua'i	35.758	0.040	0.0358
	O'ahu	40.700	0.041	0.024
	Maui Nui	44.378	0.054	0.076
	Hawai'i	45.323	0.058	0.085
M1	Kaua'i	34.636	0.093	0.017
	O'ahu	38.877	0.112	0.052
	Maui Nui	48.683	0.207	0.164
	Hawai'i	55.396	0.377	0.280
M2	Kaua'i	32.434	0.132	0.009
	O'ahu	106.018	0.174	0.103
	Maui Nui	107.701	0.216	0.101
	Hawai'i	118.930	0.173	0.066
Stratified	Kaua'i	40.777	0.075	0.082

Running it – Python version (skip to C++ version, if you are running that)

We will be doing well if people can run the default *Psychotria* dataset from the Ree/Smith paper.

Prerequisites:

Lagrange runs in Python. Python is a free language/scripting environment widely used in bioinformatics (BioPython), website management, etc. Programs written in Python are then platform-independent, because Python can be installed on any operating system.

1. Mac OS X systems come with Python pre-installed (just type "python" at the prompt in Terminal.app). However, Lagrange needs some math/science libraries (scipy, numpy) in addition to the basic Python, plus it is best to use the most up-to-date version (Lagrange needs Python 2.4, many Macs have Python 2.3 or older, type "python –V" to check).

   So, the simplest thing to do is download everything you could ever need in the free Enthought Python Distribution. This is a single executable file, you can get it here free as an educational user:

http://www.enthought.com/products/getepd.php

It is a ~300 MB download, so don't everyone try and download it at once during class on Airbears! (it will download over wireless elsewhere, it just takes awhile)

Once it is downloaded, double click and follow instructions to install.*

2. Download Lagrange from here and unzip it:

   http://code.google.com/p/Lagrange/

   Place the "Lagrange" directory wherever you will be saving/editing your data files (but keep those files outside of the Lagrange directory).

Generating the inputs
(note, this is changing slightly day by day as Ree improves it)

1. Rick Ree has set up a "Lagrange configurator." Go to: http://www.reelab.net/Lagrange
2. Click on the "phylogenetic tree" link. Input the default tree.
3. Click on "species ranges". Use the example matrix.
4. Now that you have ranges and the tree, click through all the other options and figure out what each of them is for.
5. When you are done, click on "Save/Download" and download psychotria_demo.Lagrange.py (which will be the default dataset unless you changed something). Save the file to your data directory.
6. Open that file up in a text editor and look at the text between "### begin data" and "### end data". You should be able to see how the various inputs from the web form are now represented in text.
7. Go to your Terminal/Command Line window, and navigate to your data directory. Your data directory should have the /Lagrange directory in it.
8. Type "python psychotria_demo.Lagrange.py" and sit back and watch the results
9. Open up the output file "psychotria_demo.results.txt" in a text editor. You will see the "Global ML at root node", the estimated dispersal and extinction rates, and the estimated ranges for each ancestral node on the phylogeny.

Questions:

- Email me the results file, and give a brief interpretation of the results.

Lagrange, C++ version

Once you have the executable, running the example file should be simple. From the appropriate directory:

./lagrange_cpp test.lg

…or…

./lagrange test.lg

(Depending on the name of your executable)

Examine the output files and figure out what they are telling you.

Footnote

* Hopefully unnecessary detail follows on installing Python:

Once EPD is installed, type "python –V" at a Terminal/Command Line prompt. You should see something like "Python 2.5.2 |EPD Py25 4.1.30101|", NOT "Python 2.3". (If you installed EPD but don't see this, you may have to add something to your PYTHONPATH variable in the /Users/nick/.bash_profile text file, like this:

export PYTHONPATH=<insert EPD directory here>${PYTHONPATH}

…and then restart Terminal. For various Windows machines you do something similar: http://www.chem.gla.ac.uk/~louis/software/faq/q1.html

---

References

Bush, M. B., Gosling, W. D. and Colinvaux, P. A., 2006. Climate change in the lowlands of the Amazon Basin in: Flenley, J. R. and Bush, M. B. (Eds.), *Tropical Rainforest Responses to Climatic Change*, USA and UK: Springer, jointly published with Praxis Publishing, UK, pp. 55–79.

de Queiroz, A., 2005. The resurrection of oceanic dispersal in historical biogeography. *Trends Ecol Evol*. 20 (2), 68-73.

Donoghue, M. J. and Moore, B. R., 2003. Toward an Integrative Historical Biogeography. *Integrative and Comparative Biology*. 43 (2), 261-270.

Donoghue, M. J. and Smith, S. A., 2004. Patterns in the assembly of temperate forests around the Northern Hemisphere. *Philosophical Transactions of the Royal Society of London B Biological Sciences*. 359 (1450), 1633-1644.

Moore, B. R., Smith, S. A., Ree, R. H. and Donoghue, M. J., 2009. Incorporating Fossil Data in Biogeographic Inference: A Likelihood Approach. *Evolution*. In press

Ree, R. H., Moore, B. R., Webb, C. O. and Donoghue, M. J., 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*. 59 (11), 2299-2311.

Ree, R. H. and Smith, S. A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst Biol*. 57 (1), 4-14.

Ronquist, F., 1996. DIVA version 1.1. Computer program and manual, accessed online. URL: http://www.ebc.uu.se/systzoo/research/diva/diva.html.

Ronquist, F., 1997. Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. *Syst Biol*. 46 (1), 195-203.

Sanmartin, I. and Ronquist, F., 2004. Southern hemisphere biogeography inferred by event-based models: plant versus animal patterns. *Syst Biol*. 53 (2), 216-243.