Maximum Likelihood Estimation of Biogeographic History on Phylogenies using Lagrange

by Nick Matzke

This is an optional, but should be interesting/useful for those interested in biogeographical questions.

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, runs only on PCs, 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 2005-2008, 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.

The currently available program "Lagrange" (Likelihood analysis of geographic range evolution) 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):

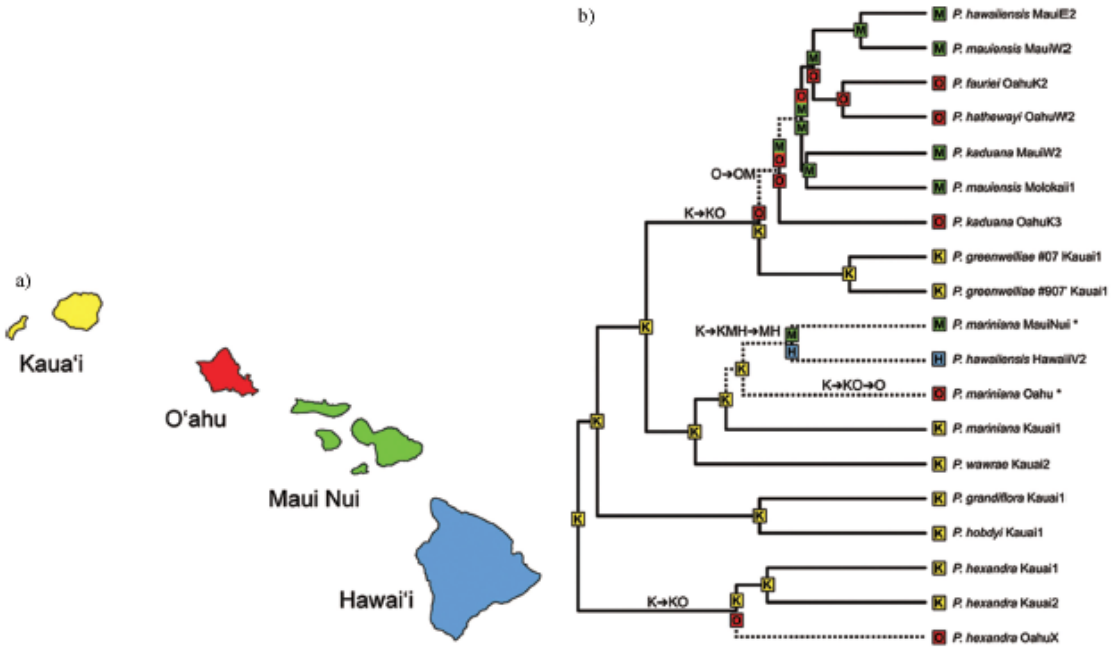
$\mathbf{P}(t) = e^{-Qt}$

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

This is a very new program, so we will be doing well if people can run the default *Psychotria* dataset from the Ree/Smith paper.

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.