IB Topics in Phylogenetics
Jun Ying Lim and Isaac Marck

# Historical Biogeography Models

## Parsimony-based models

**Ronquist, F. (1997) Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic Biology*, 46, 195 - 203.**

Proposes DIVA, a parsimony-based approach that accommodated vicariance and dispersal in a and does not assume hierarchical vicariant scenarios a prior. Dispersal-vicariance analysis ( thus DIVA) is a method for optimizing ancestral areas on a cladogram by minimizing the number of historical events (area vicariance, dispersal, and extinction) required to explain the geographic distribution of terminal taxa. DIVA has since come to be widely applied in studies of historical biogeography.

**Nylander , J.A.A, Olsson, U., Alström, P. & Sanmartin, I (2008)** Accounting for Phylogenetic Uncertainty in Biogeography: A Bayesian Approach to Dispersal-Vicariance Analysis of the Thrushes (Aves: Turdus). *Systematic Biology*, 57, 257 - 268.

A simple extension that accounts for phylogenetic uncertainty in DIVA parsimony optimizations by running DIVA on a random sample from a Bayesian posterior distribution of trees. The marginal probabilities of ancestral ranges for each node are then calculated as the frequency of an ancestral range recovered from the posterior distribution of trees.

**Yu, Y., Harris, A.J. & He, X (2010)** S-DIVA (Statistical Dispersal-Vicariance Analysis): A tool for inferring biogeographic histories. *Molecular Phylogenetics and Evolution*, 56, 848- 850.

A program that implements methods of dealing with phylogenetic uncertainty in DIVA ancestral area reconstructions (including Bayes-DIVA, see Nylander *et al.* 2008). This program is fairly popular due to its relatively useful graphical user interface.

## Probabilistic models

**Pagel, M. (1999)** The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48, 612 - 622.

In character models such as the one described in this paper, character state transitions are typically assumed to occur according to a Markov process that can be modeled as parameters in a Q matrix, with the rates of transition from state i to state j along a branch of length *t* a function of the transition rates in an instantaneous rate matrix. In this paper, Pagel describes how this approach can be used to reconstruct ancestral character states at the nodes in a phylogeny using maximum likelihood or bayesian methods, and he argues based on simulations for a “local” approach to inferring ancestral states estimated without conditioning on other nodes in the tree based on finding the state at that node that maximized the overall likelihood of the tree. This paper is relevant because it is was the conceptual and practical basis for models of biogeographic history developed later on.

IB Topics in Phylogenetics
Jun Ying Lim and Isaac Marck

**Nepokroeff, M.K., Sytsma, J., Wagner, W.L. & Zimmer, E.A. (2003)** Reconstructing ancestral patterns of colonization and dispersal in the Hawaiian understory tree genus *Psychotria* (Rubiaceae): a comparison of parsimony and likelihood approaches. *Systematic Biology*, 820 - 838.

To our knowledge, this is the first paper to implement a maximum-likelihood CTMC modelling approach on biogeographic ranges. Nepokroeff *et al.* adapted the methods for testing hypotheses about transition rates (dispersal and exctinction) and reconstructing ancestral states described in Pagel 1999 on a system of Hawaiian plants with the convenient pattern of single species endemism on four islands (analagous to the four bases of dna). They hack the popular software PAUP* and harness nucleotide substitution model implementations in the program to test hypothesis about historical biogeography. Specifically they compare likelihood of models with equal island transition rates to ones in which colonization of newer islands are favored. They found strong evidence for colonization from older islands to younger islands, and they reconstruct a Kauai origin of all of the major subclades of Psychotria.

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

Ree *et al.* develops a novel parametric model of historical biogeography that models lineage geographic range histories as anagenetic events (along the branches) and cladogenetic events (at the nodes). Unlike the model by Nepokroeff et al. (2003), Ree *et al.* allow lineages to occupy more than one areas (discrete geographic units), and is thus concerned with the combination of areas that a lineage occupies (the geographic range), with transition rates representing dispersal from other areas occupied by a lineage, or extirpation in particular areas. To estimate transition probabilities along branches, Ree *et al.* employed a Monte Carlo approach with waiting times drawn from an exponential distribution with mean 1/(rate of dispersal + rate of extirpation). At the nodes, Ree *et al.* consider three main flavours of cladogenetic range inheritance scenarios, 1) sympatry (parent lineage occupies only one area and both daughter lineages inherit that area, e.g., A → A + A, B → B + B), 2) allopatry/vicariance (parental range is divided between daughter lineage, e.g., AB → A + B) and 3) peripheral isolate / parapatry (parent lineage occupies more than one area, and there has been speciation in one of those areas, leading to one daughter in that area, and another inheriting the whole parent range, e.g., AB → A + AB or B + AB). Note that Ree *et al.* 2005 does not explicit name these scenarios, but they have been referred to as such in the literature (e.g., Matzke 2014). They correspond to scenarios 1, 2 and 3 respectively in Figure 3. To calculate the likelihood of any set of observed species ranges on a tree, they use a flat prior on possible ancestral ranges and on the possible range inheritance scenarios for each of these possible ancestral ranges. He compares the performance of DEC under different scenarios of dispersal and extirpation rates to ancestral ranges inferred under DIVA using both simulated and empirical data. Most differences can be explained by the way events are treated in DEC and DIVA, but they suffer from interesting side-effects as they do not deal with extinction of whole lineages. For example, long branches are better explained by extremely low extirpation rates, giving rise to a “paradox of widespread ancestors”.

**Ree and Smith (2008)** Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*, 57, 4 - 14.

In this paper, Ree & Smith treat transitions between geographic ranges using a rate matrix, allowing transition probabilities to be calculated analytically and computed much more quickly. This represents a refinement of the Monte Carlo approach of Ree *et al.* (2005) which was computationally

more demanding. They also describe and implement constraints such as time-stratified transition rates (i.e., time slices of the phylogeny and downpassing probabilities). Interestingly, they find that while the DEC model consistently underestimated dispersal and extirpation rates on simulated data, the DEC model was able to accurately reconstruct ancestral ranges, although accuracy declines with increasing rates of range dispersal and extinction. They also fit their model to a time-calibrated phylogeny of Hawaiian *Psychotria*, under various constraints (max range area of two, west-to-east dispersal, incorporating geologic ages of the Hawaiian islands). Model comparisons favour a Kauai origin of the clade and the model where range sizes were constrained suggesting that ancestral ranges were small (or widespread ancestors did not persist for long).

**Sanmartin, I., van der Mark, P. & Ronquist, F. (2008)** Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *Journal of Biogeography*, 428 - 449.

**Sanmartin, I., Anderson, C.L., Alarcon, M., Ronquist, F. & Aldasoro, J.J. (2010)** Bayesian island biogeography in a continental setting: the Rand Flora case. *Biology Letters*, 6, 703 - 706.

The above two papers use a novel Bayesian approach to biogeographic modelling. They use a typical character model (e.g., Nepokroeff *et al.* 2003), but fit it to multiple lineages simultaneously. To accommodate for lineage-specific differences, they allow for separate molecular clocks for each lineage, and lineage-specific mutation rate modifiers and dispersal rate modifiers. The benefit of this approach is that it gains statistical power and biogeographic signal from multiple lineages. However, one drawback is that dispersal rates are averaged across all lineages and so interesting lineage-specific biogeographic patterns may not be recovered.

**Goldman, E.E., Lancaster, L.T. & Ree, R.H. (2011)** Phylogenetic inference of reciprocal effects between geographic range evolution and diversification

In this paper, Goldman and colleagues extend the BiSSE model (Maddison *et al.* 2007) to include geographic states. Unlike the BiSSE model (and analagous to the DEC model), lineages can occupy more than one area with transitions between states described by rates of dispersal, extirpation and speciation. Like all other BiSSE type models, the probabilities of extinct lineages that do not survive to the present are explicitly accounted for. However, the degree to which reconstructed phylogenies contain enough information for trait-based diversification models to recover extinction rate parameter values accurately remains an open issue.

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

To account for rare, long-distance colonization events, Matzke extends the DEC model by allowing for a jump-speciation (J) cladogenetic scenario which permits one daughter lineage at any splitting event to not have to share any part of ancestral range of the parent lineage. The added philosophy is that genetic isolation through this mechanism is effectively instantaneous, and should not be modelled as anagenetic range expansions. Matzke also implements some changes to how cladogenetic scenarios are weighted, by allowing different classes of event (“sympatry”, “sympatric-subset” and "vicariance") to either be fixed (i.e, like in DEC), free (i.e., fully estimated from the data) or deterministic functions of each other. He then compares model performances between the DEC and DEC+J models. To ensure fairness,

he allows different event classes to have a weight of (3-j)/3, where j = [0,3] (weights reduce to that in DEC when j = 0; whereas only jump speciation is allowed when j = 3). By fitting DEC+J using DEC-derived parameter estimates of dispersal (d) and extirpation (e), they find similar log-likelihoods for both models. However, when they fit the DEC+J model with free parameter values to a variety of island clades, DEC+J appears to have substantially higher parameter estimates (log-likelihood difference of ~ 10 in many cases). Ancestors are often estimated with narrower ranges, and with more confident range estimates. Clades also tend to have simpler biogeographic histories (i.e., less reticulation). Using statistical methods to gauge model performance, Matzke argues that DEC+J is more realistic, and that the two models should be compared in a statistical framework. He also discusses the role of parametric biogeographic models in the context of other widely used approaches (e.g., parsimony, island models etc).

**Landis, M.J., Matzke, N.J., Moore, B.R. & Huelsenbeck, J.P. (2013)** Bayesian analysis of biogeography when the number of areas is large. *Systematic Biology*, 62, 789 - 804.

Parametric models of biogeography become unfeasible as the number of areas grows (number of possible ranges = 2^number of areas - 1). Matrix exponentiation of large matrices (> 10 areas) becomes intractable. To circumvent this limitation, Landis *et al.* integrate over possible biogeographic histories using MCMC. They do so by sampling biogeographic histories (using stochastic character mapping) **for each** area under a given model (i.e., one parameter to describe area gain, and one parameter to describe area loss). Likelihood calculations for any given biogeographic history is thus straightforward as the simulated events and waiting times are known. Landis *et al.* do not incorporate cladogenetic events in his model. He describes two types of models from which biogeographic histories can be sampled, a distance-independent model (rate of area gain and loss equal across for all areas) and a distance-dependent model where rates of area gain for any particular area are dependent upon its weighted connectivity to other occupied areas, and an additional distance exponent (beta). They also show how the performance of the two models can be objectively compared using Bayes factors under both simulated and empirical data (Rhododendrons in South East Asia).