# The Role of Fossils in Biogeographic Inference

# Annotated Bibliography

Reilly Hayes & Jaemin Lee

## MODELS

**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, v. 59, p. 2299-2311. doi: 10.1111/j.0014-3820.2005.tb00940.x.**

**Ree, R.H., and Smith, S.A., 2008, Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis: Systematic Biology, v. 57, p. 4-14. doi: 10.1080/10635150701883881.**

These two papers outline the maximum likelihood-based “dispersal-extinction-cladogenesis” (DEC) model of biogeographic inference. The 2008 method improves upon 2005 iteration, and remains popular despite the publication of an arguable improvement (see Matzke, 2013, 2014). The DEC model has proven especially popular among those who wish to incorporate fossils in their analyses.

The basic parameters of the DEC model include (1) dispersal rates between any number of analytical areas, (2) extinction rates within each area. From these a rate matrix is constructed to describe instantaneous transition rates between geographic ranges. The major distinction between the 2005 and 2008 iterations of DEC concerns the transition probabilities calculated from the rate matrix: the earlier version relies upon a sluggish simulation-based approach to derive these probabilities, while 2008 method calculates them exactly via exponentiation of the rate matrix. When coupled with prior probabilities for cladogenic range-inheritance scenarios, transition probabilities calculated from the rate matrix supply all the information necessary to calculate the likelihood of the observed geographic ranges of the tips of a phylogeny.

Calculating these likelihoods follows a procedure almost identical to that employed in ancestral character state reconstruction; the major caveat is that daughters do not inherit the identical ancestral state after cladogenesis, but rather a state following a range evolution scenarios permissible under the model. Permissible scenarios assume that one daughter lineage will inherit a single area while its sister lineage inherits (1) the remainder of the ancestral range, or (2) the entirety of the ancestral range. The model also accommodates easy implementation of biogeographic constraints. If, for example, it is known that dispersal between two areas would have been impossible before a certain time *t*, the dispersal rate between those areas can simply be set to zero before *t*, and range inheritance likelihoods calculated where branches cross *t* in addition to at internal nodes.

Simulated data in these papers indicate DEC tends to underestimate rates of dispersal and extinction, but performs well so long as these events are rarer than speciation and the overall rate of range evolution (i.e. sum of dispersal and extinction rates) is not too great. The papers also provide empirical examples: Ree et al. (2005) reconstruct a widespread ancestral range for *Cercis* (Fabaceae) across the northern hemisphere, while Ree and Smith (2008) recover Kaua'i as the best-supported area of origin for *Psychotria* (Rubiaceae) among the Hawaiian islands.

**Goldberg, E.E., Lancaster, L.T., and Ree, R.H., 2011, Phylogenetic inference of reciprocal effects between geographic range evolution and diversification: Systematic Biology, v. 60, p. 451-465. doi: 10.1093/sysbio/syr046.**

This paper describes the geographic state speciation and extinction and extinction ("GeoSSE") model, a likelihood-based method to jointly infer range evolution, speciation, and extinction on a phylogeny. This approach formulates spatial regions and rates of anagenetic range evolution as in DEC (Ree et al., 2005; Ree and Smith, 2008), but adds speciation rates to this basic framework. The two-area model presented here involves six parameters: two rates of dispersal, two of extinction, and two of speciation. It then performs likelihood calculations following a modified version of the binary state speciation and extinction ("BiSSE") model of Maddison et al. (2007; for that paper, see https://doi.org/10.1080/10635150701607033). The key distinction between BiSSE and GeoSSE is that while the former involves two states (i.e. character states 0 and 1), the latter has three (i.e. geographic ranges A, B, and AB). Although these authors do not explicitly discuss integration of fossils, no obvious barrier in the model restricts their use as noncontemporaneous tips alongside extant taxa.

Simulations indicate persistent correlations among the six parameters of GeoSSE, but it is possible to distinguish between parameters so long as phylogenies are sufficiently large. The authors describe this threshold as roughly "one or two hundred tip species". Moreover, GeoSSE consistently recovers the correct signals on large phylogenies in simulations of geographically-dependent speciation (speciation limited to only occur within or between regions) and source-sink and sink-sink scenarios (immigration rates, rather than speciation rates, maintain diversity in one or more regions). An empirical analysis of the California shrubs

*Arctostaphylos* (Ericaceae) and *Ceanothus* (Rhamnaceae) reveals asymmetry of rates among GeoSSE parameters, extrapolated by the authors to the broader California flora. These data establish GeoSSE as a useful tool for biogeographic inference, despite its limitation to taxon-rich contexts.

**Matzke, N.J., 2013, Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model testing [Ph.D. thesis]: Berkeley, University of California, 240 p.**

Nicholas Matzke's Ph.D. thesis introduces the R package 'BioGeoBEARS'. The package implements several biogeographic models—DIVA, DEC, and BayArea—in a common likelihood-based framework, and offers tweaked versions of these models accommodating the process of founder-event speciation (dubbed "+J" variants; see Matzke, 2014 below). The package also introduces procedures for incorporating fossil data in these models; we focus here on this second point.

BioGeoBEARS addresses explicitly the 'imperfect detection' problem of fossil data: the presence of fossils in a region can substantially influence the outcome of an analysis, yet the absence of fossils carries little information. To estimate the likelihood of true absence, BioGeoBEARS parameterizes sampling effort by way of taphonomic control groups (i.e. sets of taxa that are equally detectable in the fossil record as a taxon of interest). Low fractional abundance of a taxon of interest in a region, coupled with high abundance of taphonomic controls, increases the likelihood of true absence for that taxon. After sampling effort is jointly estimated alongside other parameters of biogeographic models implemented in BioGeoBEARS, the likelihood of any number of hypothetical taxon ranges may be calculated following this principle.

The procedure outlined above appeared as Chapter 4 of Matzke's Ph.D. thesis. Although this chapter has not yet been published, the model is already implemented in BioGeoBEARS. Interested readers should nevertheless consider emailing Matzke for details on implementation.

**Matzke, N.J., 2014, Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades: Systematic Biology, v. 63, p. 951-970. doi:10.1093/sysbio/syu056.**

The DEC model outlined by Ree and Smith (2008) only models cladogenic range evolution by sympatry or vicariance (i.e. one of the daughter lineages inherits a range that was present in the ancestral range). In this paper, Matzke describes a model variant accounting explicitly for founder-event speciation ("DEC+J"), in which the range of a daughter may be completely disparate from that of its ancestor. DEC+J achieves this by introducing an additional parameter weighting the probabilities of all such "jump dispersal" events possible at cladogenesis.

Matzke employs a model selection framework to compare the performance of DEC and DEC+J on simulated and empirical data. Likelihood-ratio tests and comparisons of alkaline information criterion weight consistently—and often dramatically—favor DEC+J over DEC in both these simulations and empirical data sets. This is because the explicit modeling of founder-event speciation allows DEC+J to reconstruct simpler biogeographic histories than DEC, with ancestral ranges at interior nodes both more narrowly constrained and more confidently estimated. For those interested, Matzke also provides qualitative comparisons of DEC+J to several alternative parsimony- and likelihood-based biogeographic methods.

**Silvestro, B., Zizka, A., Bacon, C.D., Cascales-Miñana, B., Salamin, N., and Antonelli, A., 2016, Fossil biogeography: a new model to infer dispersal, extinction, and sampling from paleontological data: Phil. Trans. R. Soc. B 371, 20150225. doi: 10.1098/rstb.2015.0225.**

This paper introduces the dispersal-extinction-sampling (DES) model for biogeographic inference in deep time. The DES model differs in three ways from all others presented here: (1) it does not rely upon phylogenetic information, but rather a record of fossil occurrences; (2) it explicitly models sampling biases that control the spatial distribution of fossils; and (3) it does not assume that anagenetic range evolution, commonly parameterized with a one constant dispersal and local extinction rate, is a time-homogeneous process.

DES relies upon a *Q* matrix identical to that employed by DES, and likewise models range evolution as a Markov process with an exponentially-distributed waiting time. However, DES differs in that coded absences are viewed as equivocal, not definitive. It introduces area-specific preservation rates—treated as homogeneous Poisson processes—as parameters underpinning the probability that a taxon might be present in a region in which it is unobserved. By combining the *Q* matrix with these preservation rates, the DES model calculates the likelihood of data identically to DES, with several exceptions: (1) range evolution is assumed to occur independently across lineages, so the joint likelihood of any number of lineages is simply the product of their respective likelihoods; (2) likelihoods are calculated in sequential time bins away (i.e. more ancient) from the present, rather than on a phylogeny; and (3) ancestral range probabilities are not treated as uniform, but rather set according to the preservation rates operative at a given time within each region in the analysis.

The agnosticism of DES to phylogenetic information makes it a poor choice for ancestral range reconstruction, but among competing models, it alone can produce satisfactory estimates of dispersal and extinction rates heterogeneous in time and space in the absence of well-developed phylogenetic data. The authors illustrate a potential application by contrasting global climate records with vascular plant dispersal and extinction rates between Cenozoic North America and Eurasia.

## APPLICATIONS

**Nesbitt, S.J, Smith, N.D., Irmis, R.B., Turner, A.H., Downs, A., and Norell, M.A., 2009, A complete skeleton of a Late Triassic saurischian and the early evolution of dinosaurs: Science, v. 326, p. 1530-1532. doi: 10.1126/science.1180350.**

The description of the basal dinosaur *Tawa hallae* occupies the bulk of this paper, but it is otherwise notable for presenting the first likelihood-based biogeographic analysis (DEC; Rees & Smith, 2008) of an entirely-extinct clade. Details of this analysis are limited to the Supplement.

Given the nature of fossil data, substantial uncertainties surround taxon ages and phylogenetic branch lengths in this analysis. The authors calibrated the tips by using the midpoint of the observed temporal range of each taxon as a point estimate for its age, and the remainder of the tree following three strategies: (1) all were branches set to an equal length of 1.0; (2) branch lengths were set according to taxon age point estimates and phylogenetic constraint, with zero-length branches set to an arbitrary minimum of 0.1; and (3) branch lengths were set according to taxon age point estimates and phylogenetic constraint, as before, but with zero-length branches spaced equally between internal calibration points. The authors did not prefer the results of analyses using any one of these calibration strategies, but instead drew their conclusion—rejection of a singular radiation of endemic North American theropods—given consistencies observed across all three.

**Lowen, M.A., Irmis, R.B., Sertich, J.J.W., Currie, P.J., and Sampson, S.D., 2013, Tyrant dinosaur evolution tracks the rise and fall of Late Cretaceous oceans: PLoS ONE, v. 8, e79420. doi:10.1371/journal.pone.0079420.**

Like Nesbitt et al. (2009), this paper describes a new dinosaur and employs it in a DEC-based biogeographical analysis. Because these authors estimated taxon ages, calibrated branch lengths, and imposed biogeographic constraints following the same procedures as Nesbitt et al. (2009), the lack of a complementary parsimony-based analysis represents the only major methodological deviation from that study.

**Wood, H.M., Matzke, N.J., Gillespie, R.G., and Griswold, C.E., 2013, Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders: Systematic Biology, v. 62, p. 264-284. doi: 10.1093/sysbio/sys092.**

This paper presents the first biogeographic analysis to include fossils as noncontemporaneous tips alongside extant taxa. An intriguing biogeographical puzzle motivated the study: although extant palpimanoid spiders occur only in the southern hemisphere, the clade features a rich Eurasian fossil record. The authors performed five independent divergence time estimates on a total-evidence phylogeny of the spider infraorder Araneomorphae, each with a unique implementations of fossil information (e.g. in-group fossils discarded, used strictly for node calibrations, or used as noncontemporaneous tips, with or without additional node calibrations provided by out-group fossils). Because they regarded vicariance as the mechanism likely underpinning the divergent distribution of fossil and extant palpimaniods, they performed DEC- and DIVA-based biogeographic analyses at two scales: (1) regional, a scale incompatible with large-scale vicariance because two daughters cannot both inherit multi-area ranges under the assumptions of DEC (but note that the authors here report summed probabilities of inter-hemisphere dispersals as probabilities of vicariance); and (2) hemispheric, compatible with vicariance because only two areas are available to inherit.

Divergence time estimates are older and less sensitive to priors in those analyses treating fossil taxa as noncontemporaneous tips rather than node calibrations. Moreover, all biogeographic analyses support a vicariance scenario consistent with these ancient divergence estimates: palpimanoids are reconstructed as widespread preceding the separation of Laurasia and Gondwana—as well as the divergence of fossil and extant taxa—and spatially restricted thereafter.

**Meseguer, A.S., Lobo, J.M., David, R.R., Beerling, D.J., and Sanmartin, I., 2015, Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of *Hypericum* (Hyperiaceae): Systematic Biology, v. 64, p. 215-232. doi: 10.1093/sysbio/sysu088.**

This study elaborates upon the DEC model of Ree and Smith (2008) with the integration of ecological niche modeling. A three-step process preceded biogeographic analysis. Phylogenetic and divergence-time analyses were first performed under a variety of clock and tree model priors in BEAST with previously-published chloroplast data. Unlike Nesbitt et al. (2009), Lowen et al. (2013), Wood et al. (2013), and Fu et al. (2018), fossils were not included here as tips, but rather assigned to internal nodes on the phylogeny following topological inference. This node-based approach is perhaps more tenable in the context of *Hypericum* than other study systems, as the genus features an extensive record of fossils characterized by easily-recognizable synapomorphies. P. The authors next developed their ecological niche model, based upon climate records and *Hypericum* occurrences throughout the Cenozoic, to describe evolving climatic tolerances of the genus that may have restricted potential dispersal routes.

These data permitted this study's unique biogeographical analysis, featuring four major departures from a standard DEC model: (1) ranges permissible at ancestral nodes were restricted to those compatible with fossils assigned to those nodes; (2) a time-stratified paleogeographic model restricted dispersal potential through time; (3) climatic restrictions inferred by the ecological niche model further restricted dispersal potential; and (4) a Geographic State Speciation and Extinction Model (GeoSSE; Goldberg et al. 2011) estimated heterogeneity of dispersal and extinction rates among regions. The analysis reconstructed an African ancestor for Hypericaceae, and a widespread high-latitude distribution for the ancestor to crown *Hypericum*. Moreover, the ecological niche model suggested spatial stability of climatic optima for *Hypericum* across the Cenozoic.

**Fu, C.-N., Mo, Z.-Q., Yang, J.-B., Ge, X.-J., Li, D.-Z., Xiang, Q.-Y., and Gao, L.-M., 2019, Plastid phylogenomics and biogeographic analysis support a trans-Tethyan origin and rapid early radiation of Cornales in the Mid-Cretaceous: Molecular Phylogenetics and Evolution, v. 140, 106601. doi: 10.1016/j.ympev.2019.106601.**

To our knowledge, this paper represents the most recent biogeographic study incorporating fossil data. The paper presents phylogenomic and biogeographic analyses of Cornales, an early-diverging asterid order with a rich fossil record. We focus here on the latter analysis.

This study includes two sets of biogeographic analyses—one in the parsimony-based DIVA, and the other in DEC—each with three differing treatments of fossil taxa. Analyses either (1) only included extant tip taxa on the phylogeny, and only considered ranges of extant taxa; (2) only included extant tips on the phylogeny, but considered ranges of both extant and fossil taxa; or (3) included both extant and fossil taxa as tips on the phylogeny, with the branch lengths subtending fossils on the time-calibrated tree set to the approximate age of each fossil, and considered ranges of fossil and extant taxa. Analyses were also performed under differing constraints of maximum ancestral range (3, 4, or 6 continental-scale areas), yielding consistent results across all nodes except crown Cornales.

The consideration of fossil taxa caused most discrepancies among these analyses. Analyses considering fossil and extant taxa generally yielded broader ancestral ranges than those considering extant taxa alone, with clades spread across two to four areas in the former analysis restricted to one or two in the latter. Notably, the results of analyses considering fossil ranges accorded regardless of whether those fossils were also included as tips on the phylogeny. The authors preferred the results of the fossil-based analyses, emphasizing the unique information only fossils provide. It is noteworthy, however, that both this study and Wood et al. 2013 (above) recovered widespread ancestral ranges when considering fossil information. This echoes the "ancestral area paradox" discussed by Ree et al. (2005). DEC frequently infers widespread ancestors because the presence of extant taxa at the tips of basal long branches depresses estimated extinction rates; survival of these taxa is simply unlikely under high extinction rates. Widespread ancestral ranges thus become more likely than narrow ones, as these narrow ranges necessitate extinction to explain the appearance of novel ranges among daughter taxa under the cladogenesis scenarios permissible under DEC.