Journal of Biogeography (*J. Biogeogr.*) (2009) **36**, 1211–1220

SPECIAL ISSUE

Prospects and challenges for parametric models in historical biogeographical inference

Richard H. Ree[1]* and Isabel Sanmartín[2]

[1]*Department of Botany, Field Museum of Natural History, Chicago, IL, USA and* [2]*Department of Biodiversity and Conservation, Real Jardín Botánico, CSIC, Madrid, Spain*

*Correspondence: Richard H. Ree, Department of Botany, Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, IL 60605, USA.
E-mail: rree@fieldmuseum.org

ABSTRACT

In historical biogeography, phylogenetic trees have long been used as tools for addressing a wide range of inference problems, from explaining common distribution patterns of species to reconstructing ancestral geographic ranges on branches of the tree of life. However, the potential utility of phylogenies for this purpose has yet to be fully realized, due in part to a lack of explicit conceptual links between processes underlying the evolution of geographic ranges and processes of phylogenetic tree growth. We suggest that statistical approaches that use parametric models to forge such links will stimulate integration and propel hypothesis-driven biogeographical inquiry in new directions. We highlight here two such approaches and describe how they represent early steps towards a more general framework for model-based historical biogeography that is based on likelihood as an optimality criterion, rather than having the traditional reliance on parsimony. The development of this framework will not be without significant challenges, particularly in balancing model complexity with statistical power, and these will be most apparent in studies of regions with many component areas and complex geological histories, such as the Mediterranean Basin.

Keywords

Ancestral range estimation, dispersal, Mediterranean Basin, parametric model, range expansion, stochastic process, vicariance.

INTRODUCTION

The use of phylogenetic trees in historical biogeography dates to the origins of cladistic theory, when Hennig's progression rule argued that the branching order of lineages could contain information on their geographic origins (Hennig, 1966). Since then, phylogenies have become integral to historical biogeographical inference, being applied to a range of problems, from reconstructing general area relationships and inferring the causes of common distribution patterns across species (cladistic biogeography), to reconstructing ancestral ranges and biogeographical events on branches of the tree of life (taxon biogeography). However, in spite of this progress, statistical methods for detecting significant patterns and trends in such reconstructions remain poorly developed, a deficiency that contrasts sharply with other areas of phylogenetic research, such as comparative analysis of phenotypic evolution. By and large, such methods have not been transferred to historical biogeography, limiting the power with which biogeographical hypotheses can be tested in a phylogenetic context.

That statistical inference has not taken hold in historical biogeography appears to stem from an ingrained adherence to parsimony (Occam's razor) as a means of inferring the past. This has been particularly evident in the development of cladistic biogeography, the study of detecting general area relationships from the phylogenies and distributions of species among areas of endemism, in which the principle of parsimony underlies the conviction that congruence in observed patterns necessarily results from a common cause (Parenti, 2006). Patterns, in this context, are the branching relationships, or topologies, of taxon–area cladograms (phylogenetic trees with the geographic ranges of species arrayed at their tips), and a common cause of particular interest is vicariance, the splitting apart of a contiguous ancestral area. The concordant splitting of ancestral species by vicariance is predicted to leave a phylogenetic signal of sister clades having common disjunct distributions. Dispersal, on the other hand, is viewed by cladistic biogeographers as idiosyncratic, generating unpredictable topological patterns, and is not considered to be a valid common cause.

doi:10.1111/j.1365-2699.2008.02068.x

Adherence to parsimony arguably led to a disproportionate emphasis on taxon–area cladogram topologies, at the expense of other relevant sources of information, especially those related to time. Donoghue & Moore (2003) held up clade age estimates as an example of data that, given advances in molecular clock methods, cladistic biogeographers today are imperiled to ignore, as they are needed to distinguish, for example, between true congruence (common patterns with a common cause) and pseudo-congruence (common patterns with different causes). Reflecting on the need for more integrative methods for historical biogeography, they noted how new sources of data can spur the development of new analysis methods, which in turn can make new kinds of inferences feasible; eventually, changes in which questions are perceived to be important and interesting can occur, and the overall direction of a field can shift. In this light, molecular clock estimates of clade ages are a new source of data in the sense of only recently becoming relatively common and trustworthy. These authors further speculated that parametric models integrating time, vicariance, dispersal and phylogenetic uncertainty could be developed, leading to new analysis methods and new biogeographical insights.

In this essay, we take a closer look at the potential use of parametric models in historical biogeography, with the view that their development may help to invigorate the field, by allowing a broader range of hypotheses to be tested using statistical methods than has been previously feasible under parsimony frameworks. We develop a rationale for such models, and discuss how they have been and could be applied to historical biogeographical inference.

WHY PARAMETRIC MODELS?

In abstract terms, parametric models define the likelihoods of alternative scenarios, given a set of probability distributions and their corresponding parameter values. In Markov models, likelihoods of a system being in alternative states at a given point in time are defined by probabilities of state change, conditional on its prior state. Markov models are commonly used for historical inference, wherein data observed in the present may be seen as the outcome of stochastic change through time. The likelihood of observed data is thus a function of parameters for the probability distributions underlying such change, as well as of any additional hypotheses about the past, for example concerning states at previous time points.

In terms of historical biogeography, this translates to extant species ranges (observed data) being the outcome of stochastic change through time, or, in other words, arising from processes of geographic range evolution. Examples of such processes are dispersal, extinction and speciation, which can be thought of as generating discrete events that cause ranges to evolve by expansion, contraction and splitting, respectively. A common Markov model for temporal sequences of discrete events draws successive waiting times randomly and independently from an exponential distribution. Models of geographic range evolution, then, could thus be parameterized by exponential rates for different event types, with the rates estimated from the observed data by maximum likelihood.

In such a model, the phylogeny itself is an important parameter, because its topology structures the hierarchical sequence of ancestor–descendant events of range evolution, and its branch lengths indicate relative amounts of expected change (e.g. proportional to time). This highlights an important difference between parametric models and parsimony: the latter cannot effectively account for stochastic events, such as the fact that dispersal is more likely to occur on longer branches than on shorter branches, and thus tends to underestimate the number of events on long branches. Time, then, is a primary reason why parametric models are better suited to historical biogeographical inference, and indeed to comparative evolutionary inference in general, in being a common axis along which processes of change and phylogenetic relationships can be integrated. This is true even in the absence of an absolute (i.e. geological) time-scale, but the calibration of phylogenetic branches to absolute time greatly expands the scope of possible inquiry, allowing sources of data based on geological information, such as fossils, plate tectonics, palaeoclimate reconstructions, sea levels, etc., to be integrated into models and hypotheses about the past distributions and movements of species.

Parametric models are widely used in comparative biology because they provide a hypothesis-testing framework based on likelihoods of alternative parameterized scenarios generating the observed data. Other variables that are not directly related to the question of interest may be factored out, either by fixing them to assumed values, estimating them by maximum likelihood, or, in Bayesian approaches, by integrating over their posterior densities. Objective methods for choosing among alternative models, such as likelihood ratio tests, the Akaike information criterion, and Bayes factors are becoming ubiquitous in the literature (see Sullivan & Joyce, 2005, for a review), and new approaches such as mixture models (Pagel & Meade, 2004), model averaging (Posada & Buckley, 2004) and reversible-jump Markov chain Monte Carlo methods (Pagel & Meade, 2006) allow ancestral inference without the need to apply a specific model. Such techniques run counter to a common criticism that individual models inevitably oversimplify the data (e.g. Ebach *et al.*, 2003; Brooks, 2005).

PARAMETRIC METHODS IN HISTORICAL BIOGEOGRAPHY

Previous approaches

A forerunner to model-based, statistical approaches in historical biogeography was the so-called 'event-based' approach (Ronquist, 2003), in which different kinds of events causing range evolution (dispersal, extinction and vicariance) are assigned fixed costs, which are then minimized according to a parsimony criterion. The most widely used of these is dispersal–vicariance analysis (DIVA; Ronquist, 1997). The assignment of costs is somewhat arbitrary, but the detection of

phylogenetically conserved patterns (i.e. similarity in geographic ranges inherited from ancestors to descendants) hinges on dispersal having a higher cost than vicariance, and extinction a lower cost than dispersal (Ronquist, 2003). The reason for this is that dispersal events result in descendants that occur outside the range of their immediate ancestors – a non-conservative pattern (Sanmartín *et al.*, 2007). By making explicit the relationship between biogeographical processes and expected distribution patterns on phylogenetic trees, event-based methods provided some unexpected insights. For example, they suggested that animal distributions are more likely to show the footprint of ancient vicariance events, whereas plant distributions are more influenced by recent dispersal events (Donoghue & Smith, 2004; Sanmartín & Ronquist, 2004). But despite such empirical advances, event-based methods remain constrained by their reliance on parsimony. Costs of biogeographical events cannot be estimated from the data, but must be defined in advance. Moreover, temporal and geological information is considered only heuristically in the interpretation of inferred events, not incorporated directly into inference algorithms (Sanmartín *et al.*, 2007). And hypothesis testing is problematic compared with likelihood models: event-based methods rely on non-parametric permutations of the data (e.g. Sanmartín *et al.*, 2007) that bear uncertain connections to the underlying processes of biogeographical evolution.

Recent years have witnessed increasing interest in parametric biogeography, such as the use of empirical Bayesian methods with dispersal–vicariance analysis to account for phylogenetic uncertainty (Nylander *et al.*, 2008), and the application of likelihood models of character evolution to reconstruct ancestral geographic ranges (e.g. Nepokroeff *et al.*, 2003; Olsson *et al.*, 2006; McGuire *et al.*, 2007; Pereira *et al.*, 2007). Two new inference methods have recently emerged from work by ourselves and colleagues that are based on distinct parametric models of biogeographical processes. We describe them below as representing initial forays into biogeographical model development that may set the stage for future extensions.

Range evolution by dispersal, extinction and cladogenesis

The first approach is based on a continuous-time model for geographic range evolution by dispersal, extinction and cladogenesis (the DEC model: Ree *et al.*, 2005; Ree & Smith, 2008), which may be thought of as a parametric, extended version of dispersal–vicariance analysis (Ronquist, 1997). As alluded to in the previous section, the DEC model treats dispersal and local extinction as stochastic processes that cause range expansion and contraction events, respectively, according to exponential rate parameters. Change is assumed to occur in the context of a predefined set of discrete geographic areas in which a species may be present or absent, independently of other species. The geographic range of a species (the subset of areas in which it is present) is thus analogous to a heritable character that evolves through time. For ancestor–descendant range evolution along a phylogenetic branch, the DEC model defines a matrix of transition rates between ranges based on the assumption that only a single dispersal or local extinction event can occur in an instant of time.

The model differs from those for character evolution in its treatment of cladogenesis events. Unlike character states, which are typically assumed to be inherited identically by daughter lineages following speciation, geographic ranges can potentially be inherited non-identically, as a consequence of spatial subdivision of the ancestral range. The DEC model enumerates the distinct scenarios by which this can occur, making three simplifying assumptions. The first is that speciation is dichotomous, forming two (and only two) daughter lineages. The second is that lineage divergence can occur within an area, or, for widespread ancestral ranges, between areas. The third assumption is that cladogenesis results in one of the two daughter species arising in, and inheriting a range of, a single area. For ancestors endemic to one area, range inheritance following divergence is therefore identical: the daughter species form within and inherit the ancestral area. For ancestors present in multiple areas, range inheritance is not identical, because divergence subdivides the ancestral range either between a single area and the others (a vicariant pattern), or within a single area, in which case one daughter species inherits the entire ancestral range. Range subdivision–inheritance scenarios, rather than ranges themselves, are thus the discrete states of interest at phylogenetic nodes representing cladogenesis events, and are directly analogous to ancestral character states in models of discrete character evolution.

The DEC model allows existing statistical methods that were developed for inferring character evolution on phylogenetic trees (e.g. Pagel, 1997) to be applied directly to geographic ranges, to estimate ancestral ranges and biogeographical parameters (dispersal and extinction rates) by maximum likelihood. At present, the primary implementation of the DEC model is in a cross-platform software package, Lagrange (Ree & Smith, 2007). A minimal analysis using Lagrange specifies a set of areas, a data matrix of species ranges coded as binary presence–absence values, and a phylogenetic tree with branch lengths proportional to expected change. By default, all possible ranges (area subsets) are allowed as valid states in the model. In addition, the rate of dispersal from one area to another and the rate of local extinction within an area are each assumed to be uniform across areas and constant across all branches of the phylogeny. However, these assumptions are unlikely to be suitable in practice if factors such as the spatial arrangement, size, etc. of areas are assumed to influence ranges and their evolution.

One might wish, for example, to remove some ranges from consideration because they contain spatial disjunctions, or are much larger than those of extant species. These can be removed globally from the model (reducing the dimensions of the transition matrix), or locally, limiting their exclusion to specific parts of the phylogeny. Globally reducing the set of allowed ranges becomes imperative as the number of areas increases, because without such constraints the size of the

transition matrix becomes computationally difficult. (With the current version of Lagrange on a conventional desktop computer, the maximum number of areas feasible for an unconstrained model is seven or eight.) Local range constraints might also exclude those that are inconsistent with fossil biogeographical evidence. In a similar vein, dispersal between areas can be constrained to reflect area-specific dispersal opportunity by introducing parameters that scale the overall rate. For example, one might want to allow dispersal only between spatially adjacent areas, by scaling the rate to zero for non-adjacent areas; the rate could also be scaled inversely to distance, as shown by hypothetical examples in Fig. 1. Scaling parameters could also be estimated by maximum likelihood. As with range constraints, dispersal constraints can be applied uniformly across the phylogeny, or selectively across lineages and/or through time. The latter is implemented in Lagrange by assigning separate scaling matrices to discrete time periods that stratify the phylogeny, allowing one to describe expectations about changes in dispersal opportunity through time as a result of plate movements, land bridges, mountain building, etc.

Given observed species ranges, a phylogeny, and a DEC model with or without constraints, a standard analysis using Lagrange first estimates optimal rates of dispersal and local extinction by maximum likelihood, integrating over all possible ancestral range subdivision–inheritance scenarios (ancestral states) at internal nodes on the phylogeny. These rates are then treated as fixed. Next, for each internal node, the likelihood of the data is iteratively recalculated for each range-inheritance scenario, without conditioning on assumptions about scenarios elsewhere in the tree. This procedure allows alternative scenarios at a single node to be ranked by their contributions to the overall likelihood.

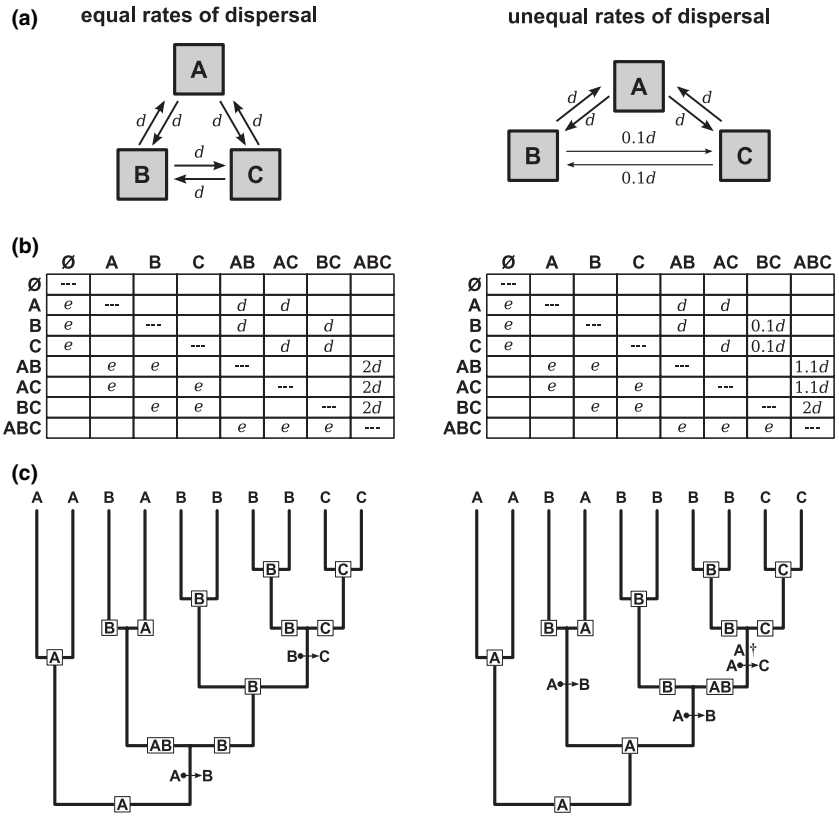

Figure 1 The effect of assuming equal and unequal (constrained) rates of dispersal on inferences of ancestral ranges and biogeographical events using DEC (dispersal, extinction and cladogenesis) models. The three component areas are labelled A, B and C; parameters for dispersal and local extinction rates are denoted by *d* and *e*, respectively. For each model, from top to bottom is shown (a) a schematic diagram of dispersal rates between areas, (b) the corresponding matrix of instantaneous transition rates between geographic ranges, and (c) a hypothetical phylogeny with observed species ranges, on which maximum likelihood ancestral ranges and implied dispersal and extinction events are mapped. The constrained model scales dispersal between B and C to one-tenth of the overall rate, as might be assumed based on relative distances. The transition matrices show rates between ranges separated by a single dispersal or extinction event. All other transitions have an instantaneous rate of zero, and elements along the diagonal are defined such that the sum of rates across a row is zero. For transitions involving dispersal, the rate is the sum of rates from areas in the starting range to the target area. At internal phylogenetic nodes, identical range inheritance is shown as a single area, whereas non-identical inheritance is shown by each daughter's range at the base of descendant branches, with the ancestral range being the union of these ranges. Dispersal events implied by ancestral ranges are shown by arrows between the source and destination areas; extinction is denoted by †. Assuming equal dispersal favours one dispersal event from A to B and one from B to C. Assuming a lower rate between B and C favours two dispersal events from A to B, one from A to C, and one local extinction event in A.

Bayesian island biogeography

The second new parametric method is a Bayesian approach to island biogeography (Sanmartín *et al.*, 2008). It also uses a model of range evolution in which lineages disperse stochastically, but, unlike the DEC model, geographic ranges are restricted to single islands (areas) and extinction is not considered (but see below). Islands here are thus analogous to nucleotides in a DNA substitution model. Lineage ranges undergo transitions determined by parameters for dispersal rate and island 'carrying capacity' (equilibrium frequencies of species diversity). The method is oriented towards drawing inferences about the movement and distribution of species in an island system from multiple clades of inhabitants. It treats all parameters, including the phylogenies themselves, as probability distributions that are estimated simultaneously from the data gathered for phylogeny reconstruction (typically molecular sequence alignments) and from the biogeographical data on species ranges. Analyses yield Bayesian posterior probabilities about the parameters of interest, such as island carrying capacities or directional trends in dispersal, while integrating over the distributions of all other parameters.

Similar analogies between DNA substitution models and island biogeography models have been used before, notably for plants in the Hawaiian archipelago by Nepokroeff *et al.* (2003). The main novelty of the Bayesian method is the use of Markov chain Monte Carlo for integrating over the uncertainty in tree topology and branch lengths while inferring general biogeographical trends across multiple groups with different dispersal capabilities and life-history traits. Another advantage is the flexibility of the model, in that it allows dispersal rates to vary among islands, groups of islands, etc., or to be scaled by geographic distance, area size, etc. Future work should allow asymmetric dispersal models and directional models. The Bayesian island model has been implemented in the program MrBayes v. 4.0 (source code available from http://www.mrbayes.net; a beta version for Mac available from I.S.).

A standard Bayesian island analysis involves multiple groups of species. For each group, data are needed for molecular phylogeny reconstruction (a DNA sequence alignment) and for island distribution. Each group is assigned a private model of DNA evolution, whereas the island character is shared across groups – in other words, molecular parameters and phylogenies are estimated independently, and biogeographical parameters are estimated globally. A composite phylogenetic–biogeographical Markov chain Monte Carlo run is constructed to sample stochastically the tree topology, branch lengths and the parameters of the biogeographical and molecular models, yielding estimates of their posterior distributions given the data. This enables extraction of the marginal probabilities of biogeographical parameters of interest, for example estimates of dispersal rate that do not condition on any particular phylogeny or set of branch lengths. To account for variation in rates of molecular evolution across groups, branch lengths (measured in units of the expected number of substitutions per site) are scaled according to group-specific molecular clocks, which then can be converted from relative to absolute times using fossil calibrations. Subsequent conversion to the expected number of dispersal events per unit time is done according to scaling parameters that account for differences in dispersal rates across groups.

Restricting geographic states to single islands (i.e. disallowing widespread states) may seem like a disadvantage if one is interested in detailed inferences of ancestral ranges of species inhabiting contiguous areas, but less so if inquiry is focused on the general dispersal patterns of multiple clades among isolated land masses. Since it is unrealistic to assume that widespread species persist for any significant duration in geological time before diverging in allopatry between isolated areas, dispersal (migration) events between such areas could be considered to be effectively equivalent to speciation events (even though they are modelled as occurring along phylogenetic branches, not at internal nodes). Extinction is not explicitly considered as a parameter in the Bayesian island model, but it is implicit in the reconstructed phylogeny, which includes only extant species (Sanmartín *et al.*, 2008). For the island model, dispersal is best seen as a process of historical biotic exchange: the net rate of successful migration and survival. It is thus applicable generally to any biogeographical setting in which areas are discrete and isolated from one another by barriers. For example, it could be used to estimate rates of dispersal in mountain systems, where peaks are isolated by intervening lowlands.

Time, particularly with respect to change in area relationships, can also be integrated into the Bayesian method by converting branch lengths from units of relative to absolute time by means of fossil calibration points. As with the DEC model, one could specify different transition rates across distinct time periods to reflect changing area configurations (connections), with the durations of those periods being fixed to values inferred from current palaeogeographic knowledge, or estimated from the data. Temporally dynamic island models of this kind could be constructed to study reticulate biogeographical scenarios in which dispersal barriers appear and disappear over time. They could also be helpful in designing conservation policies. For example, in a recent study of the South African Cape Flora, Forest *et al.* (2007) used the phylogenetic diversity (PD) index – a measure of the phylogenetic depth of an area's biota – to distinguish between an old, relict eastern flora and a younger, more diverse western flora. The PD index, however, gives no direct information on the extent to which each flora originated by *in situ* diversification or migration from other areas. Such insight could be provided by combining Bayesian island models with dated phylogenies.

Current challenges

The parametric methods described in the previous section were initially developed for different purposes (ancestral range estimation within a single clade vs. parameter estimation across multiple clades), drawing inspiration from existing, analogous methods for character-state reconstruction and phylogenetic model selection, respectively. Reflecting their

roots, they differ in assuming whether the phylogeny is known at the outset, and accordingly adopt different underlying models of geographic range evolution. In the Bayesian approach, the similarity of the island model to nucleotide substitution models facilitates the simultaneous analysis of geographic ranges and molecular sequences using Markov chain Monte Carlo methods. The DEC method, in emulating the analysis of character evolution on a known phylogeny, focuses attention on range inheritance scenarios at cladogenesis events and adopts a somewhat more complex view of states that allows for widespread taxa and flexible constraints on ranges and dispersal. These differences are not fundamental, however, and hybrid approaches are easily imagined: for example, DEC-type models could in principle be used in Bayesian frameworks that integrate over phylogenetic uncertainty. However, effort would perhaps be better spent extending these methods so that they overcome current limitations, rather than simply merging them. Below, we highlight what we perceive to be the most significant challenges in this endeavour and suggest some potentially fruitful paths to their solution.

Defining areas

A critical step in biogeographical model construction is in circumscribing the component areas, as these form the foundation on which model states and parameters rest. In contrast to the traditional emphasis on current areas of endemism, parametric biogeography requires that areas be defined with greater consideration towards specific hypotheses of interest, such as the movement of lineages through particular corridors. Care must be taken not to recognize more areas than necessary, out of concern towards computational feasibility (Ree & Smith, 2008), but also for theoretical reasons. All else being equal, more areas translates to less phylogenetic signal in the species ranges of any given clade, because fewer areas (states) are likely to be shared by descent. Similarly, events such as dispersal are less likely to involve the same areas, reducing the power to detect general trends. To compensate for this decline in information content, it becomes increasingly important to adjust dispersal parameters to reflect spatial relationships and other factors related to the connectivity of areas. To illustrate this point, consider that, for four areas arranged in a 2×2 grid, one might be satisfied with a model having a single rate of dispersal between any pair of areas; if, however, the grid dimensions were larger (e.g. 16×16 areas), one should be more inclined to scale downwards the dispersal rate between non-adjacent areas.

Defining ranges

For models allowing widespread ranges as states, defining a large number of component areas also means that care must be taken to ensure that widespread ranges make biological sense: for example, that they are not discontiguous or contain disjunctions across which dispersal is highly unlikely or impossible. In the context of the grid example above, this might mean constraining widespread ranges to include areas that share an edge. The maximum size and configurations of valid ranges in the model might also be empirically constrained based on those observed in extant species (Ree & Smith, 2008). The bottom line is that as the number of states increases relative to the amount of data, it becomes increasingly worthwhile to impose geographic structure on model parameters governing the transitions between those states.

Complex area histories

A related obstacle to inferring biogeographical events and patterns from phylogenetic trees is that the component areas themselves can have complex histories, involving tectonic activity, changes in climate, sea levels, etc., which may be known to varying degrees of certainty. Indeed, area identities may change through time as a result of geological evolution, increasing the difficulty of circumscribing transition matrices. It is also important to consider the historical interplay between the ecological niches of the species of interest and dispersal between areas. Needless to say, physical connections between areas are not in themselves sufficient for successful dispersal; for potential migrants, environmental conditions must be tolerable, and ecological opportunity must be waiting on the other side. For example, during Pleistocene glacial periods, Beringia served as a dispersal corridor between Eurasia and North America for frost-tolerant species, but remained a barrier for more warm-adapted organisms.

Range-dependent diversification

Perhaps the most significant limitation of both the DEC and the Bayesian island model is their failure to integrate parameters for lineage diversification with those for range evolution. This is illustrated by considering the role of dispersal in each case. In the DEC model, a dispersal event merely expands the range to include a new area, whereas in the island model it 'jumps' the lineage from one area to another. In both cases, dispersal events are assumed to occur along phylogenetic branches in proportion to their length. In neither case are dispersal events explicitly associated with phylogenetic nodes representing lineage divergence, as might be expected from a rare long-distance dispersal event establishing a small founder population across a wide barrier. More generally, the issue is whether lineage diversification, the stochastic process that generates waiting times between birth (speciation) and death (extinction) events, is dependent on geographic range evolution. In other words, rather than viewing phylogenetic branch lengths simply as indicators of expected ancestor–descendant change, should they be regarded as being partially determined by events such as dispersal? In cases of long-distance dispersal events, as in island systems, this makes intuitive sense: a founder population is unlikely to backcross with its source or receive new migrants, and coalescence of a new lineage will be rapid. Even if dispersal is viewed as simply range expansion, as conceived in the DEC model, it still seems biologically

reasonable to expect that larger ranges are more likely than smaller ranges to undergo divergence, owing to the greater potential for local adaptation, peripheral isolation, disruptions of habitat continuity inhibiting gene flow, and so on (Rosenzweig, 1995; Fine & Ree, 2006). As a result, dispersal events causing range expansion should lead to shorter waiting times between speciation events.

A model of range-dependent diversification incorporating the above concepts would represent a significant advance over current models, particularly if it allowed flexible specification of the relative effects of range expansion and long-distance dispersal on rates of speciation and extinction. In such a model, the expected waiting time between speciation events would be a function of such parameters, conditional on the ancestral range. Under this framework, the topology and branch lengths of a given phylogeny would be assumed to have been generated by the biogeographical model itself, rather than as an independent structure on which biogeographical events passively occurred. Likelihood calculations using this model would require integration over all possible ways that range-dependent diversification could have produced a given tree and observed range data. Such a model has not yet been implemented, but there exists a template on which its development could be based. As with the DEC and Bayesian island models, inspiration is found in phylogenetic methods for inferring character evolution, specifically in the binary state speciation–extinction model (BiSSE) introduced by Maddison *et al.* (2007). This model describes the simultaneous evolution of a discrete binary character with phylogenetic tree growth, in which character transition rates may differ and rates of speciation and extinction are state-dependent. The likelihood of observed states is calculated on a phylogeny by numerically integrating probabilities of state transitions and lineage diversification along its branches. By substituting parameters for character evolution with those for geographic range evolution, and including non-identical range inheritance scenarios, the BiSSE approach could be co-opted for use in historical biogeography.

Integration with phylogeography

As the time horizon of interest shrinks towards the present, and genetic data below the species level become increasingly relevant, biogeographical inquiry converges on phylogeography, where the aim is to identify and test hypotheses about geographic patterns of genetic variation and the historical factors that have shaped them. Advancements in model-based phylogeographic methods are on the rise, notably through the blending of demographic parameter estimation (e.g. Beerli, 2006) with explicit biogeographical hypothesis testing (Knowles & Maddison, 2002) in combination with ecological niche modelling (see Richards *et al.*, 2007; Yesson & Culham, 2006). These efforts have thus far yielded significant insights into the tempo and mode of species divergence in the Pleistocene (e.g. Knowles *et al.*, 2007). Studies combining niche modelling with palaeogeographic reconstructions are yielding inferences about the presence of land corridors of suitable habitat across regions that are now uninhabitable (e.g. Weaver *et al.*, 2006), and these can be used in turn to model the availability of dispersal routes in biogeographical inference. Also of note is a new maximum likelihood approach to phylogeography based on a random-walk migration model (Lemmon & Lemmon, 2008) that is aimed at estimating the locations of ancestors on gene trees of georeferenced individuals and testing hypotheses about per-generation dispersal distance, centres of origin, and directional trends in the development of species ranges. Such advances bode well for future integration with historical biogeography, such that improved knowledge about recent range dynamics will hopefully inform parametric models targetting deeper time-scales.

TOWARDS A GENERAL PARAMETRIC FRAMEWORK

We view current methods as first steps towards more general parametric frameworks that are both inclusive and flexible in allowing disparate sources of data and assumptions to be incorporated into studies of historical biogeography. Ideally, future frameworks would include models of geographic range evolution suited to a variety of spatial contexts (continental areas, islands, etc.) and time-scales. A general workflow would begin with hypotheses of interest, available data, and assumptions about historical conditions (constraints on ranges, dispersal, extinction, etc. resulting from biotic and abiotic factors). These inputs would guide the choice of model of range evolution, and of which parameters are to be estimated, fixed, or marginalized. Inferences drawn from the data using the model could then be used to refine the initial hypotheses and assumptions.

As with all inference problems, the key is identifying the simplest model that adequately captures the relevant features of the geographic system, and maximizes the power of the comparative data at hand to distinguish between biologically interesting scenarios. In the Mediterranean Basin, for example, geological evidence points to a complex mosaic of microplates, tectonic belts and island arcs characterized by a history of repeated fragmentation and reticulation of component areas, highlighted in a recent biogeographical study of aroid flowering plants (Mansion *et al.*, 2008). The authors identified a system of 14 geologically significant areas, with most of their study species restricted to one or two areas, but a few occurring in up to five (not including the cosmopolitan *Arisarum vulgare*). Dispersal–vicariance analysis (DIVA) initially yielded multiple thousands of equally parsimonious reconstructions of ancestral ranges, a result probably resulting in part from the lack of geographic structure in assumptions about dispersal (by default, DIVA assigns equal costs across all areas). From these results, the authors extracted the most empirically reasonable solutions, based on assumptions of spatially contiguous range expansion.

A parametric approach to the DIVA analysis of Mansion *et al.* (2008) would require careful consideration of the geographic structure and dispersal opportunities between the

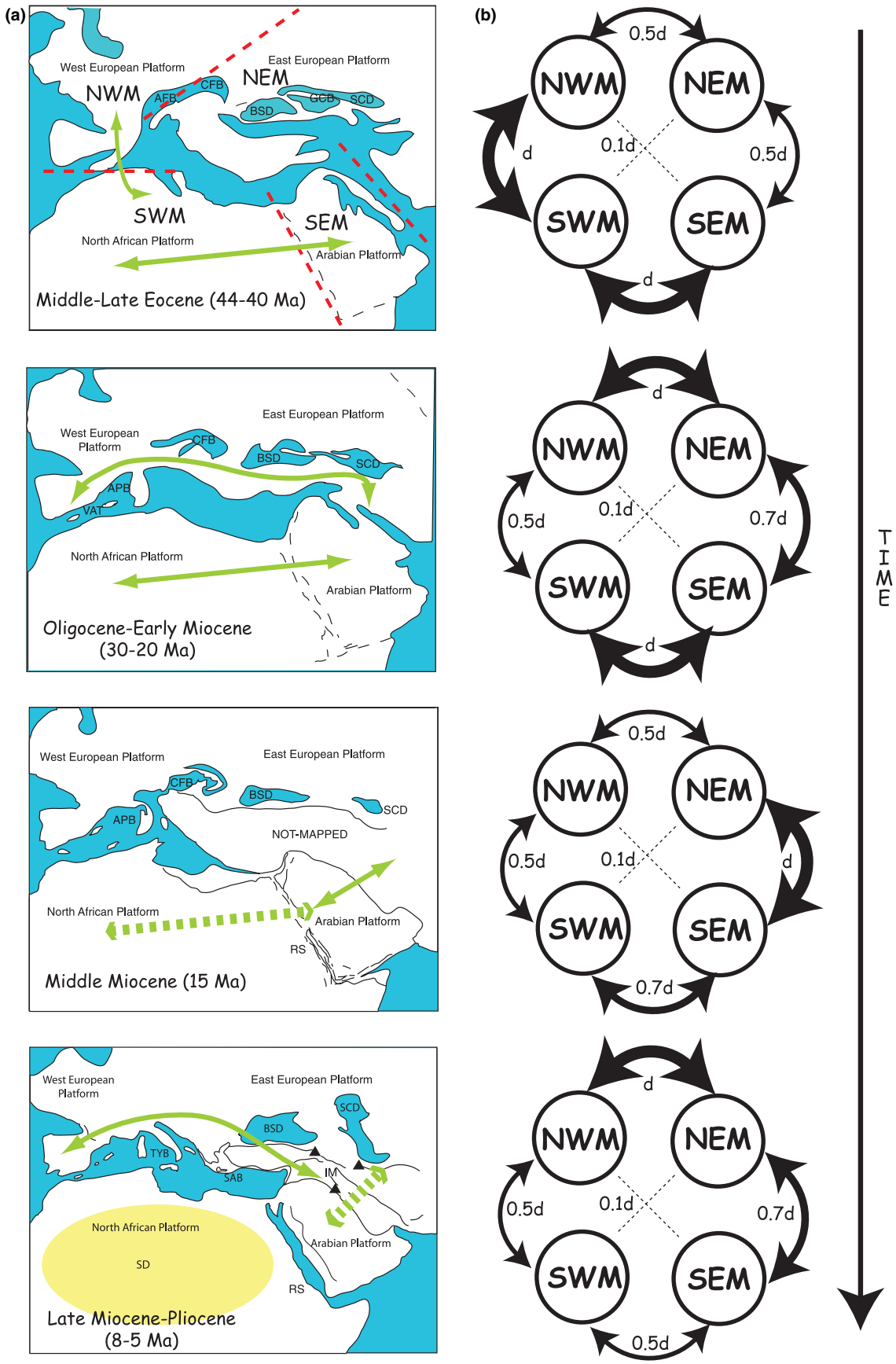

Figure 2 Simplified Mediterranean biogeography interpreted as a parametric model. (a) Palaeogeographic reconstructions of Mediterranean areas through the Cenozoic, reflecting the main collision and splitting events between four major plates. Maps are reinterpreted from Meulenkamp & Sissingh (2003). AFB, Alpine Foreland Basin; APB, Algero-Provenzal Basin; BSD, Black Sea Depression; CFB, Carpathian Foreland Basin; GCD, Great Caspian Depression; IM, Iranian Mountain chains; RS, Red Sea; SAB, South Aegean Basin; SCD, South Caspian Depression; SD, Saharan Desert; TYB, Tyrrenian Basin; VAT, Valencia Trough. Green lines indicate primary dispersal routes; red dashed lines show the limits between the four major areas. (b) Dispersal parameters for the four time periods depicted in (a). NWM, North Western Mediterranean (West European Platform); SWM, South Western Mediterranean (African Platform); NEM, North Eastern Mediterranean (East European Platform); SEM, South Eastern Mediterranean (Arabian Platform).

14 component areas through time in order to construct a useful and computationally tractable transition matrix. We regard this as a worthwhile and achievable goal, but in the interim suggest a simplified model (Fig. 2), reflecting the main collision and splitting events between four major plates (Meulenkamp & Sissingh, 2003). We delimit four temporal periods in the Cenozoic with corresponding parameter values that reflect changing dispersal opportunities through time. This simplified model could be used as a base for either DEC models that allow widespread species, or Bayesian island models for clades composed of single-area endemics.

CONCLUSIONS

Parametric methods depart rather substantially from the earlier tradition in cladistic biogeography of avoiding assumptions about process. Instead, biogeographical processes are explicitly modelled in probabilistic terms, and integrated into the inference framework. We see this as a strength rather than as a weakness, but we also recognize that, as with other applications of model-based inference in evolutionary biology, a significant challenge is how to balance the complexity and realism of models against computational feasibility and inferential power in an optimal way. In particular, attention must be paid to the potential accuracy of inferences as model complexity increases. However, despite these challenges, we remain optimistic. Donoghue & Moore (2003) argued convincingly for integrating time and other kinds of data into the inference of historical cause in current biogeographical patterns. Here, we make the case that parametric models of biogeographical evolution represent a means for making such integration happen that is more viable than traditional parsimony frameworks. Movement in this direction is as yet at an early stage, but if the development of model-based methods in other areas of phylogenetics is any indication, it will yield new opportunities for quantitative statistical analysis in the field of historical biogeography.

ACKNOWLEDGEMENTS

This paper stems from a contribution initially presented at the conference 'Origin and Evolution of Biota in Mediterranean Climate Zones: an Integrative Vision', held in Zürich on July 14–15, 2007, for which we thank Elena Conti for the invitation to participate. We also thank the participants of 'Developing an Integrative Algorithmic Method for Historical Biogeography', a working group supported by NESCent (US NSF #EF-0423641) and led by Joel Cracraft and Mark Siddall, for discussions and inspiration.

REFERENCES

Beerli, P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, **22,** 341–345.

Brooks, D.R. (2005) Historical biogeography in the age of complexity: expansion and integration. *Revista Mexicana de Biodiversidad*, **76,** 79–94.

Donoghue, M.J. & Moore, B.R. (2003) Toward an integrative historical biogeography. *Journal of Integrative and Comparative Biology*, **43,** 261–270.

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

Ebach, M.C., Humphries, C.J. & Williams, D.M. (2003) Phylogenetic biogeography deconstructed. *Journal of Biogeography*, **30,** 1285–1296.

Fine, P.V.A. & Ree, R.H. (2006) Evidence for a time-integrated species–area effect on the latitudinal gradient in tree diversity. *The American Naturalist*, **168,** 786–804.

Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Proches, S. & van der Bank, M. (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, **445,** 757–760.

Hennig, W. (1966) *Phylogenetic systematics*. University of Illinois Press, Urbana.

Knowles, L.L. & Maddison, W.P. (2002) Statistical phylogeography. *Molecular Ecology*, **11,** 2623–2635.

Knowles, L.L., Carstens, B.C. & Keat, M.L. (2007) Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology*, **17,** 940–946.

Lemmon, A.R. & Lemmon, E.M. (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic Biology*, **57,** 544–561.

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56,** 701–710.

Mansion, G., Rosenbaum, G., Schoenenberger, N., Bacchetta, G., Rosselló, J.A. & Conti, E. (2008) Phylogenetic analysis informed by geological history supports multiple, sequential invasions of the Mediterranean Basin by the angiosperm family Araceae. *Systematic Biology*, **57,** 269–285.

McGuire, J.A., Witt, C.C., Altshuler, D.L. & Remsen, J.V. (2007) Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. *Systematic Biology*, **56,** 837–856.

Meulenkamp, J.E. & Sissingh, W. (2003) Tertiary palaeogeography and tectonostratigraphic evolution of the Northern and Southern Peri-Tethys platforms and the intermediate domains of the African–Eurasian convergent plate boundary zone. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **196,** 209–228.

Nepokroeff, M., Sytsma, K.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*, **52,** 820–838.

Nylander, J.A.A., Olsson, U., Alström, P. & Sanmartín, 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.

Olsson, U., Alström, P., Gelang, M., Ericsson, P.G. & Sundberg, P. (2006) Phylogeography of Indonesian and Sino-Himalayan bush warblers (*Cettia*, Aves). *Molecular Phylogenetics and Evolution*, **41,** 556–561.

Pagel, M. (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, **26,** 331–348.

Pagel, M. & Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, **53,** 571–581.

Pagel, M. & Meade, A. (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, **167,** 808–825.

Parenti, L. (2006) Common cause and historical biogeography. *Biogeography in a changing world* (ed. by M.C. Ebach and R.S. Tangney), pp. 61–71. CRC Press, London.

Pereira, S.L., Johnson, K.P., Clayton, D.H. & Baker, A.J. (2007) Mitochondrial and nuclear DNA sequences support a Cretaceous origin of Columbiformes and a dispersal-driven radiation in the Paleogene. *Systematic Biology*, **56,** 656–672.

Posada, D. & Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, **53,** 793–808.

Ree, R.H. & Smith, S.A. (2007) *Lagrange (Software for likelihood analysis of geographic range evolution), version 2*. Distributed by the authors at http://lagrange.googlecode.com.

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

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.

Richards, C.L., Carstens, B.C. & Knowles, L.L. (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, **34,** 1833–1845.

Ronquist, F. (1997) Dispersal–vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic Biology*, **45,** 195–203.

Ronquist, F. (2003) Parsimony analysis of coevolving associations. *Tangled trees: phylogeny, cospeciation, and coevolution* (ed. by R.D.M. Page), pp. 22–64. University of Chicago Press, Chicago.

Rosenzweig, M.L. (1995) *Species diversity in space and time*. Cambridge University Press, Cambridge.

Sanmartín, I. & Ronquist, F. (2004) Southern Hemisphere biogeography inferred by event-based models: plant versus animal patterns. *Systematic Biology*, **53,** 216–243.

Sanmartín, I., Wanntorp, L. & Winkworth, R.C. (2007) West Wind Drift revisited: testing for directional dispersal in the Southern Hemisphere using event-based tree fitting. *Journal of Biogeography*, **34,** 398–416.

Sanmartín, 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*, **35,** 428–449.

Sullivan, J. & Joyce, P. (2005) Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **36,** 445–466.

Weaver, K.F., Anderson, T. & Guralnick, R. (2006) Combining phylogenetic and ecological niche modeling approaches to determine distribution and historical biogeography of Black Hills mountain snails (Oreohelicidae). *Diversity and Distributions*, **12,** 756–766.

Yesson, C. & Culham, A. (2006) Phyloclimatic modeling: combining phylogenetics and bioclimatic modeling. *Systematic Biology*, **55,** 785–802.

BIOSKETCHES

Richard Ree is a curator of flowering plants at the Field Museum of Natural History (Chicago, IL, USA). His research focuses on plant evolution, emphasizing phylogenetic methods, and on theoretical approaches to comparative inference.

Isabel Sanmartín is a senior researcher at the Real Jardín Botánico, CSIC (Madrid, Spain). Her main research interests include the study of large-scale historical biogeographical patterns for both plants and animals, and the development of new analytical methods of biogeographical inference with special reference to model-based approaches.

Editor: John Lambshead

This paper stems from a contribution initially presented at the conference *Origin and Evolution of Biota in Mediterranean Climate Zones: an Integrative Vision*, held in Zurich on 14–15 July 2007.