Trends in
Ecology & Evolution

Opinion

Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues

Hélène Morlon, [1,5,*,@] Stéphane Robin, [2,3] and Florian Hartig [4]

A lot of what we know about past speciation and extinction dynamics is based on statistically fitting birth–death processes to phylogenies of extant species. Despite their wide use, the reliability of these tools is regularly questioned. It was recently demonstrated that vast 'congruent' sets of alternative diversification histories cannot be distinguished (i.e., are not identifiable) using extant phylogenies alone, reanimating the debate about the limits of phylogenetic diversification analysis. Here, we summarize what we know about the identifiability of the birth–death process and how identifiability issues can be addressed. We conclude that extant phylogenies, when combined with appropriate prior hypotheses and regularization techniques, can still tell us a lot about past diversification dynamics.

Highlights

Reconstructing past speciation and extinction dynamics from extant phylogenies is one of the main approaches to study the build-up of biodiversity on geological time scales.

These reconstructions typically involve a prior hypothesis on the functional form of temporal variations in speciation and extinction rates.

Avoiding to formulate *a priori* hypotheses while still being able to separate speciation and extinction dynamics will require other information such as fossils, constraints on complexity, or statistical regularization techniques.

Molecular phylogenies and diversification dynamics

The diversity of life on Earth has arisen from a succession of **speciation** and **extinction** events (see Glossary). The rates at which ancestral species give rise to new daughter species (the speciation rate, λ) or go extinct (the extinction rate, μ) reflect underlying ecological and evolutionary processes, and shape species richness over geological timescales. Understanding how these rates have changed through time has long been of interest to evolutionary biologists [1–8]. While the first estimates of speciation and extinction rates were derived from the fossil record, researchers now also widely use dated phylogenies of present-day species (so-called **reconstructed (or extant) phylogenies**, hereafter referred to as 'phylogenies' for simplicity) to study past speciation and extinction dynamics [9–12].

Nee *et al.* [13] showed, using the **homogeneous birth–death (BD) process**, that despite extinct species being absent from a phylogeny, extinctions leave a distinctive signal in the timing of branching patterns, known as the 'pull of the present'. Under the assumption of homogeneous and constant speciation and extinction rates, it is therefore possible to estimate these rates from phylogenies. A wide range of more complex models grounded on the homogeneous BD process have now been developed, and are used to test hypotheses about past diversification dynamics [14–22].

Increasing flexibility of the models brings new issues, however, such as parameters that may not be **identifiable**. Here, we discuss the identifiability of speciation and extinction rates in a variety of homogeneous BD models, and clarify the theoretical limits that nonidentifiability imposes on phylogenetic diversification analysis. We conclude that although speciation and extinction histories are statistically unidentifiable if the underlying functions are completely unconstrained [23], this does not imply that phylogenies cannot reveal speciation and extinction dynamics [23,24]. We hold that in most practical scenarios, *a priori* hypotheses, biological knowledge, or statistical **regularization** can make the problem identifiable.

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, PSL Research University, Paris, France
[2]UMR MIA-Paris, AgroParisTech, INRA, Paris-Saclay University, 75005 Paris, France
[3]Centre d'Ecologie et des Sciences de la Conservation (CESCO), Muséum National d'Histoire Naturelle, CNRS, Sorbonne University, Paris, France
[4]Theoretical Ecology, University of Regensburg, Regensburg, Germany
[5]http://www.phyloeco.biologie.ens.fr/

*Correspondence:
helene.morlon@bio.ens.psl.eu
(H. Morlon).
[@]Twitter: @HMorlon@florianhartig

Identifiability of speciation and extinction rates

To clarify the issue of identifiability, it is useful to make a distinction between asymptotic (or fundamental) and practical unidentifiability. **Asymptotic unidentifiability** corresponds to the case when distinct parameter combinations cannot be told apart, even in the limit of an infinite number of observations; **practical unidentifiability** corresponds to the case when parameters cannot be told apart from the limited number of observations available in practice.

Asymptotic identifiability of the homogeneous BD model

Nee *et al.* [13] showed that the homogeneous constant rate BD model with complete sampling (i.e., all present-day species are represented in the phylogeny) is asymptotically identifiable. Incomplete sampling can be accounted for by assuming that each extant species is sampled with the same probability ρ ($\rho < 1$), but already in this simple extension of the model, if ρ is a parameter to be estimated, λ, μ, and ρ are not asymptotically identifiable [25]. To solve this identifiability problem, the fraction of present-day species represented in the phylogeny is often included as prior information on ρ, which renders λ and μ asymptotically identifiable.

Extending the work of Nee *et al.* [13], Stadler [16] showed that the 'episodic' birth–death model (also called birth–death shift, BDS), where diversification rates are piecewise constant (i.e., constant on successive time intervals, or epochs) is asymptotically identifiable. More recently, Legried and Terhorst [26] confirmed this result and showed that it holds even if the epochs are not fixed. However, the BDS model with mass extinction events, that is, including the possibility that sudden (simultaneous) extinction events can occur at the end of each epoch (equivalent to sampling each species with an epoch-specific probability ρ), is not identifiable [16].

In the case when $\lambda(t)$ [or $\mu(t)$] is a smooth function of time and not constrained to follow specific functional forms such as the exponential or any other biologically motivated function, Louca and Pennell [23] showed that there is an infinity of 'congruent' functions that yield the same **likelihood**, meaning that this process is not asymptotically identifiable (Box 1).

Practical identifiability of the homogeneous BD model

When applying BD models to real data, a further issue arises: the size of phylogenies is typically not huge. Finite data sizes impose limits to the identifiability of any given model, as the confidence in the parameter estimates decreases with decreasing sample sizes. This is well illustrated by estimates of the extinction rate and the **extinction fraction** ($\frac{\mu}{\lambda}$), which typically have wide confidence intervals even for asymptotically identifiable models (see, e.g., Table S9 in [16]), such that accurate estimates often require sample sizes that are not achieved in practice. Speciation rates, by contrast, can be estimated with good accuracy on phylogenies of moderate size for the constant-rate BD model [27], as well as for the BDS model if the number of epochs is kept small [16]. Similarly, in BD models with rates that are constrained to follow a specific and simple functional dependency (such as the exponential) to time [14,15] or the environment [28], parameters determining the time or environment dependency of the extinction rate have wide confidence intervals, while those associated with the speciation rate can be estimated with good accuracy [15,28]. In general, by the usual arguments about degrees of freedom, the functional complexity that can be supported by a typically sized phylogeny of a few hundred tips is probably in the order of a few parameters. Thus, practical identifiability alone dictates that we must put constraints on the flexibility of the models used to infer diversification dynamics.

Dealing with practical versus asymptotic identifiability issues

Asymptotic and practical identifiability issues are common in science, and a large set of ideas has emerged to address such problems. Practical identifiability issues are commonly understood as

Glossary

Asymptotic (or theoretical) unidentifiability: a situation when there are distinct combinations of the model parameters that cannot be told apart even in the limit of an infinite number of observations.

Bias–variance trade-off: a trade-off between systematic model error due to limited flexibility (bias) and uncertainty of the parameter estimates (variance).

Extinction: disappearance of a species, corresponding to the death of its last individual.

Extinction fraction: extinction rate divided by speciation rate.

Homogeneous birth–death (BD) process: the BD process where speciation and extinction rates are identical across lineages at any time. Rates may vary in time, but not across lineages.

Identifiability: when fitting statistical models, identifiability means that any two combinations of parameter values can be distinguished.

Likelihood: function of the parameters of a given model that measures the probability of the observations given the model and its parameter values.

Model misspecification: a situation when the distribution of data implied by the model (under best possible parameterization) differs from the distribution of data under the true generating process.

Net diversification rate: speciation rate minus extinction rate.

Practical unidentifiability: a situation when there are distinct combinations of the model parameters that cannot be told apart from the limited number of observations available in practice.

Reconstructed phylogeny: estimated phylogenetic tree for present-day species (missing lineages that have gone extinct and are thus unsampled).

Regularization: a set of statistical techniques that consist in adding a regularization term (or penalty) to the optimization function (typically the likelihood) to solve an ill-posed problem or avoid overfitting.

Speciation: a process by which two populations of the same ancestral species give rise to two distinct descendant species.

Box 1. Model congruency and pulled diversification rates

Louca and Pennell [23] consider the homogeneous (i.e., lineage independent) stochastic BD process of cladogenesis with rates of speciation (birth, λ) and extinction (death, μ) that can change arbitrarily over time t. They show that for any given differentiable (and therefore continuous) time-dependent speciation function $\lambda > 0$ and extinction function $\mu \geq 0$, there exists an infinite set of alternative functions $\lambda^* > 0$ and $\mu^* \geq 0$ such that the probability distribution of extant trees under the corresponding BD processes M and M* is identical. Consequently, M or M* yield identical likelihood values for any given empirical tree, which implies that $\lambda(t)$ and $\mu(t)$ are not uniquely identifiable unless further constraints are imposed on their functional form.

Louca and Pennell then reparameterize the problem to have only identifiable quantities, which they call the pulled rates. The pulled speciation rate is given by:

$$\lambda_p = \lambda(1 - \phi) \qquad \text{[I]}$$

where ϕ is a function of time that denotes the probability that a lineage alive at time t has no descendant in the phylogeny, and which analytical expression is given, for example, by Equation 2 in [15]. The pulled diversification rate is given by:

$$r_p = \lambda - \mu + \frac{1}{\lambda}\frac{d\lambda}{dt} \qquad \text{[II]}$$

Congruent models are found by solving Equation 2 in [23]:

$$\frac{d\lambda}{dt} = \lambda\left(r_p - \lambda + \mu\right) \qquad \text{[III]}$$

Given any μ^*, we can compute λ^* using the solution to this equation, provided in Louca and Pennell [23]'s supplementary information [Equations 39 and 40, $\eta_0 = \rho\lambda(0)$, $\mu_0 = \mu(0)$]:

$$\lambda\,(t) = \frac{\eta_0 e^{\Lambda(t)}}{\rho + \eta_0 \int_0^t e^{\Lambda(s)} ds} \qquad \text{[IV]}$$

with

$$\Lambda(t) = \int_0^t \left[r_p(s) + \mu\,(s)\right] ds \qquad \text{[V]}$$

Alternatively, given any λ^*, we can compute μ^* as:

$$\mu\ = \frac{1}{\lambda}\frac{d\lambda}{dt} + \lambda\ - r_p \qquad \text{[VI]}$$

manifestations of the **bias–variance trade-off**, which states that model complexity must be adjusted to the data size to minimize the total error (bias + variance) of the inference (Box 2). This can be achieved by a variety of statistical model selection or regularization techniques (Box 2). For example, the practical identifiability of the asymptotically identifiable BDS model (without mass extinctions) can be improved by introducing temporally autocorrelated rates drawn from a Bayesian prior, rendering parameter estimates with time divided in hundreds of epochs identifiable on relatively small phylogenies (200 tips) [29].

Addressing asymptotic identifiability issues, such as the nonidentifiability of the BD model with unconstrained λ and μ highlighted by Louca and Pennell [23], is a different problem, as the error of our inference does not decrease with increasing data size. Yet there are approaches for dealing with asymptotic identifiability as well, which we detail in the following sections.

Box 2. Reasons and approaches to select simple models

Deciding between alternative hypotheses through a preference for simplicity is ubiquitous in statistics and the sciences. Mathematically, this is expressed by viewing the evidence in favor of a respective hypothesis (or model, denoted by M) as a combination of:

$$\text{Evidence} = \text{Likelihood}(\text{M}) - \text{Penalty} \times \text{Complexity}(\text{M}) \qquad [\text{I}]$$

where the penalty term controls the ‘strength’ of the preference for simplicity.

In statistics, the traditional motivation to favor simplicity is based on the bias–variance trade-off, which posits that increasing model complexity reduces the systematic misfit (bias), but at the cost of increasing variance (uncertainty) of the parameter estimates. One can prove that, with limited data, inducing a bias toward simpler models decreases total estimation error (bias + variance), even if the true underlying model is more complex. The complexity penalty is selected to optimize the total error. This logic underlies most frequentist regularization and model selection approaches.

There is a second argument for constraining model complexity, which is independent of the data size and the bias–variance trade-off. This argument, known as the law of parsimony or Occam’s razor, relies on an *a priori* assumption that natural processes tend to be simple and smooth. The principle of parsimony is not a mathematically provable law, but it underlies centuries of thinking and experience from physics to machine learning, and from philosophy as well (see [61] for a discussion).

When implementing preferences for simplicity, it typically makes no difference if they originate from bias–variance or parsimony principles. The main difference is that in the former the penalty is chosen from the data, such that more complex models are preferred as the data size increases, whereas in the latter the penalty is chosen independent of the data, based on prior beliefs. How to best define complexity is a question of constant debate and development in statistics: we may, for example, decide that a model is simple if it is interpretable, if it involves less parameters, if it prevents fast variations, or yet other criteria. Various statistical regularization techniques implementing these criteria exist. For example, information-theoretical measures (e.g., the Akaike information criterion or Bayesian information criterion, [42,62]) add a direct penalty for the number of parameters, shrinkage estimators such as lasso or ridge or their corresponding Bayesian priors add a penalty on the deviation of model parameters from 0 [52], and statistical smoothers [63] penalize the roughness of the fitted model (as in generalized additive models, see [53,54]).

Reparametrization

A solution to asymptotic identifiability issues is to reparameterize the model with identifiable quantities. For example, in the BD model with incomplete sampling and free ρ (which needs to be considered when total diversity is unknown, which is the case of most microbial and insect groups), the **net diversification rate** $\lambda - \mu$ and $\lambda\rho$ are identifiable. The drawback of this approach, however, is that the reparameterized quantities are often scientifically less interesting. For example, Louca and Pennell [23] suggest estimating the pulled speciation and diversification rates λ_p and r_p instead of $\lambda(t)$ and $\mu(t)$ (Box 1), but these pulled rates are difficult to interpret biologically (see [30] for an attempt), which considerably limits their practical utility.

Independent data sources

Another approach to dealing with asymptotic identifiability issues is to add additional, independent data sources. Considerable progress has been made in recent years to use both phylogenetic and fossil data, which is achieved by adding fossil sampling processes to the BD process [31–38]. In the most elaborate versions of these ‘fossilized’ birth–death (FBD) models, two distinct sampling processes are considered: one with rate ψ for fossils with character (or molecular) data, which are included in the tree, and one for simple fossil occurrences without character data. The former process is asymptotically identifiable when λ, μ, and ψ are constant [34], unless samples are removed upon sampling [34,39]. The latter, however, is irrelevant in the case of modeling diversification dynamics, as extinctions and fossilizations are independent processes. As long as samples are not removed upon sampling, the process remains identifiable even if the sampling probability at present ρ is unknown (a case when the process is not identifiable from extant species alone), which illustrates that fossils can alleviate identifiability issues [34].

Despite these encouraging results, more work is needed to determine if and under which circumstances the FBD process is identifiable when λ, μ, and ψ vary as piecewise constant or continuous functions of time, to assemble empirical data sets on which to apply FBD models for diversification analyses (the FBD has so far mainly been applied to improve divergence times rather than diversification rate estimates, but see, e.g., [35,40]), to improve their computational efficiency (current implementations limit the applicability of the model to small data sets), as well as to assess whether the inclusion of fossils provides realistic estimates of extinction rates [41] (see Outstanding questions).

Constraints from *a priori* hypotheses

Identifiability issues are more likely to arise the more flexible our models are. Flexibility is put to the extreme by Louca and Pennell [23], who set the task to be able to identify any possible functional forms $\lambda(t)$ or $\mu(t)$ from extant phylogenies. A hypothesis-driven research framework limits this complexity by comparing only a small number of alternative *a priori* ideas about the underlying process [42]. Such *a priori* hypotheses will usually constrain the functional forms of λ and μ and thus render the corresponding BD models identifiable.

The foundational study of Nee *et al.* [43] followed such a hypothesis-driven philosophy. After demonstrating that their bird phylogeny was incompatible with a constant-rate diversification model and grounded in Simpson's evolutionary theory of adaptive radiations [44], they hypothesized that rates of cladogenesis might be affected by niche-filling processes. Finding that a diversity-dependent model indeed fitted their data better, they concluded that diversity-dependent cladogenesis was a more plausible scenario to explain the diversification of birds.

This hypothesis-driven approach has inspired more than 30 years of research in phylogenetic diversification analyses [10]. Exponential time dependencies have been used, for example, to mimic early burst patterns expected from adaptive radiation theory [44], or as an approximation to diversity-dependent cladogenesis [45] (see Box 3 for an illustration with the Madagascan vangas, Vangidae). In the context of environment-dependent models, functional hypotheses have often been derived from foundational theories of biodiversity, such as the metabolic theory of biodiversity [18] and MacArthur and Wilson's theory of island biogeography [20]. Phenomenological models, such as simple linear time or environmental dependencies, have also been used, but typically either as null models [45] or as the simplest way to model the effect of an explanatory environmental variable on evolutionary rates [18]. The primary goal of this research has been to fit, test, and compare diversification scenarios that were defined *a priori* to reflect different evolutionary hypotheses. Louca and Pennell's congruent models do not correspond *a priori* to any evolutionary hypotheses, and would never be considered in a hypothesis-driven model selection procedure in the first place [42] (Box 3).

A drawback of hypothesis-driven research is that the biological conclusions we draw are contingent on the *a priori* hypotheses we formulate. In particular, our hypotheses typically do not correspond completely to the process underlying the empirical data (the truth). Still, it is usually assumed that if a given hypothesis is statistically supported within a well-chosen set of alternatives, it is likely that this hypothesis is the closest to the truth. Whether this is the case for BD models, considering the existence of a large number of congruent models, remains an open question to be explored in more details (see 'The future of phylogenetic-based diversification research' section and Outstanding questions).

Constraints on complexity and statistical regularization techniques

Even in the absence of additional data or *a priori* hypotheses, there are certain philosophical, statistical, or information-theoretic principles that may allow us to prefer some congruent solutions over others.

For example, a widely accepted scientific method of deciding between alternative explanations is the principle of parsimony (or Occam's razor, Box 2). If we follow this traditional thinking in science, when several explanations with different degrees of complexity are asymptotically unidentifiable, we should prefer the simplest, which is most probably true, all other things being equal. A possible solution to the identifiability issue highlighted by Louca and Pennell [23] consists then in selecting the simplest diversification scenario in a congruence class. This preference for simplicity is distinct from the problem of optimizing complexity to avoid overfitting in the case of finite data, and applies to the case of infinite data as well. Quantifying and penalizing complexity can be challenging, but it is a classical problem that can be addressed with a variety of statistical regularization techniques (Box 2).

Penalizing complexity is just one example of a more general class of regularization techniques that add additional constraints to solve an ill-posed (e.g., asymptotically unidentifiable) problem [46]. Constraints can also come from prior biological knowledge, information theory or model selection principles, added in the statistical inference in the form of shrinkage estimators [47], or as priors in the case of Bayesian inference (Box 2). For example, as shown by May *et al.* [19], using Bayesian priors that represent the prior belief that on average 10% of species survive a mass extinction event in the BDS model with mass extinction events (an asymptotically unidentifiable model) allows distinguishing rate shifts from mass extinction events. This example provides a clear counterexample to the conclusion of Louca and Pennell that regularization cannot solve asymptotic identifiability issues (see S2.2 in [39]). Another well-known example in phylogenetics is the dating of divergence times: substitution rates and time are unidentifiable with only sequence data from extant species, but Bayesian priors on divergence times (e.g., informed by fossils) combined with relaxed clock models solve this issue (see, e.g., Figure 1 in [48]).

Box 3. Diversification of the Madagascan vangas

We illustrate hypothesis-driven research by performing an analysis of the diversification of the Madagascan vangas (Vangidae) using the logic that would be applied in the field [64], but simplified for illustrative purposes. We hypothesize that diversification followed an 'early burst' pattern [65], with fast speciation at the origin of the group and subsequent slowdown, rather than constant-rate diversification. The early burst pattern, related to the idea of adaptive radiations [44], is modeled by an exponential decay of the speciation rates through time, used as an approximation of diversity dependence. We also consider the hypothesis that a substantial number of extinction events occurred during the diversification of this group. Among the four corresponding models, the model with an exponentially declining speciation rate $\lambda(t) = \lambda_0 e^{\alpha t}$ (time t is measured from the present to the past), with speciation rate at present $\lambda_0 = 0.018$, rate of decline $\alpha = 0.1$, and no extinction $\mu(t) = 0$, noted M, is best supported by the data (see Table S1 in the supplemental information online). We conclude that the hypothesis of early burst diversification with negligible extinctions is the most likely of the four hypotheses we considered.

To better grasp the nature of congruent models, we explore models congruent to our best model M (see Text S1 in the supplemental information online). First, we choose the extinction function to be a constant $\mu_1\ (t) = \mu_0$ and compute $\lambda_1\ (t)$. Second, we choose the speciation function to be a constant $\lambda_2\ (t) = \lambda_0$ and compute $\mu_2\ (t)$. We find (see Text S1 in the supplemental information online; Figure I; here we take $\rho = 1$ as the tree of the Madagascan vangas is complete [64]):

$$\lambda_1\ (t) = \frac{\lambda_0 e^{\frac{-\lambda_0}{\alpha}} e^{(\alpha + \mu_0)t} e^{\frac{\lambda_0}{\alpha}\ e^{\alpha t}}}{1 + \lambda_0 e^{\frac{-\lambda_0}{\alpha}} \int_0^t e^{(\alpha + \mu_0)s} e^{\frac{\lambda_0}{\alpha}\ e^{\alpha s}} ds} \qquad [I]$$

and

$$\mu_2\ (t) = \lambda_0 - \alpha - \lambda_0 e^{\alpha t} \qquad [II]$$

The biological interpretation of these models and of their parameters is not obvious. The equation for μ_2 looks more interpretable at first, but it expresses the temporal change and the extinction rate at present through the same parameter α, which means that a positive extinction rate at present ($\alpha < 0$) will force extinction rates to decline over time. Here M_2* infers negative extinction rates, and is therefore not plausible (Figure I). M_1* infers a decline in speciation rate from the origin of the group to the present for extinction rates μ_0 ranging from at least 0.05 to 0.3, consistent with our previous results (Figure I). While rate estimates do vary substantially, the general temporal trend is preserved.

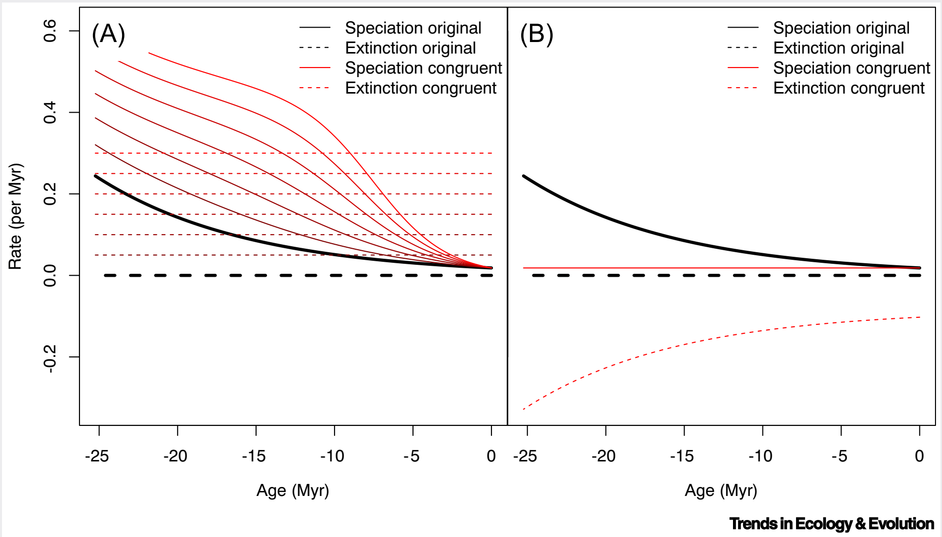
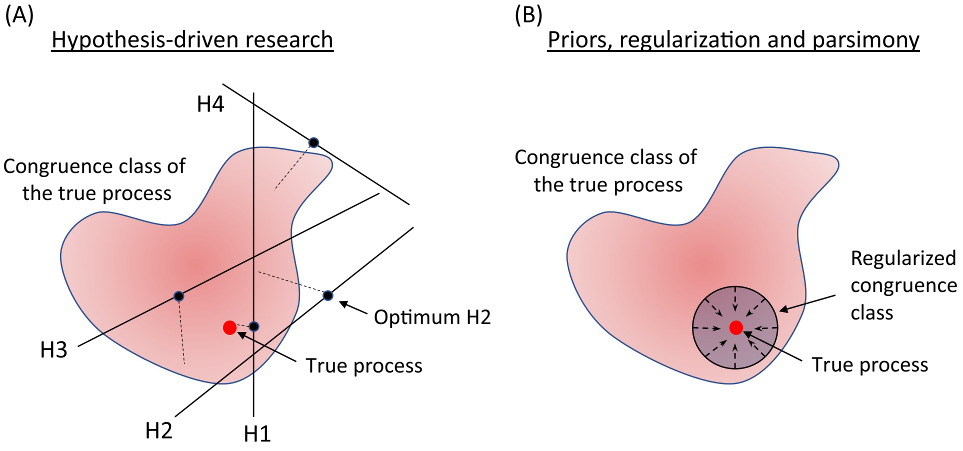


Figure I. Diversification of the Madagascan vangas as inferred from congruent models. The black curves represent the dynamics of speciation (solid line) and extinction (dashed line) corresponding to our best-fit model M (exponential decline in speciation rate, nonsignificant extinctions). The colored curves illustrate the rate dynamics of congruent models that were obtained by (A) fixing increasing values of a constant extinction rate (M_1*) and (B) fixing the speciation rate to λ_0 (M_2*). In the case of constant extinction (A), we can choose any value for μ_0 and find λ_1 (t) (so there is an infinity of congruent models), while in the case of constant speciation (B), λ_0 needs to be taken equal to the λ_0 of model M, as two congruent models necessarily have the same speciation rate at present if ρ is fixed [23] (so there is only one congruent model). Note that M_1* infers a speciation rate decline regardless of the assumed extinction rate and that M_2* infers biologically implausible negative extinction rates. See also Figures S1 and S2 in the supplemental information online. Abbreviation: Myr, million years.

The future of phylogenetic-based diversification research

The asymptotic nonidentifiability of the homogeneous BD process led Louca and Pennell [23] to conclude that phylogenetic-based diversification research should switch from a focus on speciation and extinction rates to a focus on the identifiable pulled rates. Yet, scientists interested in testing specific evolutionary hypotheses would have a hard time formulating their hypotheses in terms of these quantities, which do not correspond to a particular biological mechanism. Moreover, estimating these rates from limited-size phylogenies is still a challenging task (see Text S2 and S3 in the supplemental information online).

Instead of abandoning the goal of developing models with explicit hypotheses on speciation and extinction rates, we argue to put more efforts in using all available data (including fossil data), and testing how robust the inference from these models really is in practice, when using either a hypothesis-driven research approach or appropriate statistical regularization techniques (Figure 1). In this area, two key questions remain: how robust are biological conclusions in practice, when we use a hypothesis-driven research framework, given the existence of congruence classes? And can parsimony considerations or other regularizing techniques successfully shrink solutions in the congruence class toward the truth? The answer to these questions depends on the nature of congruence classes, for example, on whether congruence classes typically contain a wide range of disjunct models that all correspond to reasonable biological hypotheses, or that have similar parsimony/regularization properties, which remains to be explored by future research.


(A)
Hypothesis-driven research
H4
Congruence class of the true process
H3
Optimum H2
True process
H2
H1
(B)
Priors, regularization and parsimony
Congruence class of the true process
Regularized congruence class
True process

Trends in Ecology & Evolution

Figure 1. Conceptual figure illustrating how constraints imposed by prior hypotheses and regularization may help to approach the true process. Following Figure 3 in [23], the pink area represents the congruence class of the true process (red circle). (A) When considering a small number of biologically motivated hypotheses (H1–H4), the models will usually be identifiable, meaning that the optimum solution under a given hypothesis is unique (one black circle per hypothesis), and we will select the hypothesis that comes closest to the congruence class (here, H1, dashed lines convey the distance to the congruence class). This hypothesis, which is the one with highest likelihood, is traditionally assumed to be the closest to the true process. (B) Parsimony and regularization assumptions constrain the congruence class (grey circle). From the experience in other fields, we would expect the congruence class to be constrained toward the true process. These two expectations are likely to be met if biologically and statistically (i.e., with respect to parsimony and regularity properties) reasonable models within the congruence class cluster around the true process. Whether this assumption holds in reality is a question for future research.

We can think of several ways to explore these questions, such as (i) Studying the geometric properties of congruence classes mathematically, as Louca and Pennell have started to do but without definitive conclusions (see S.1.8 in [23]). This would help make the regularization choices most likely to render the models identifiable. (ii) Simulating phylogenies under general eco-evolutionary models [49–51] and checking whether the application of a hypothesis-driven framework (with well-chosen *a priori* hypotheses) selects the hypothesis that best captures a given simulated scenario; in comparison to the simulation analyses that are already usually performed to evaluate the power and type I error rates of newly developed methods, in which simulations correspond exactly to one of the fitted models, this requires using less idealized simulation models representing the eco-evolutionary processes that shape diversification dynamics. (iii) Pursuing current efforts to develop regularized models, as detailed in the following paragraph, and using eco-evolutionary simulations [as in (ii)] to check whether these models provide estimates of speciation and extinction rates that approach simulated rates.

Moreover, in real applications, practical identifiability is often as much a problem as asymptotic identifiability. Given that regularization can solve practical as well as asymptotic identifiability issues, developing suitable and biologically motivated regularization approaches that act directly on speciation and extinction rates seems more promising to us. Such approaches have already started to be developed (e.g., [19,29]), and including further general ideas from statistics and machine learning, for example, the fused lasso [52] or generalized additive models [53,54], could lead to further advances (Box 2).

The problems as well as their solutions discussed here are likely not limited to homogeneous BD models. In recent years, models with diversification rates that vary across lineages have been developed to understand why some groups of organisms are richer than others and to avoid biased inferences linked to **model misspecification** [15,55–59]. Unlike for the homogeneous BD model, for which all topologies are equally likely and therefore only branching times are informative, both branching times and topology are informative in the case of heterogeneous BD models. Despite this additional source of information, it is very likely that models with heterogeneous rates are asymptotically unidentifiable in the absence of any constraint. Working with biologically interpretable speciation and extinction rates has helped regularizing this problem, for example, by favoring rare rate shifts with large effects corresponding to the invasion of new ecological space [55–57] or by favoring frequent shifts with small effects corresponding to heritable rates, formalized by regularization in the form of autocorrelated Bayesian priors [59,60].

Concluding remarks

Identifiability issues naturally arise in approaches that try to infer the potentially unlimited complexity of historical processes from limited contemporary data, and inference of past diversification history from phylogenies of present-day species is no exception. These identifiability issues are one of the reasons why scientists adhere to hypothesis-driven research, use parsimony or regularization principles, or integrate multiple data types. Phylogenetic-based diversification analyses have already adopted these methods in the past, and need to pursue this effort to provide increasingly robust tools for understanding past diversification histories from the data that are available today (see Outstanding questions).

Outstanding questions

Does the inclusion of fossils render the BD model with time-variable speciation λ and extinction μ rates identifiable in the absence of constraints on the temporal variation of λ and μ?

Does the inclusion of fossils provide realistic estimates of extinction when applying the BD model to empirical data?

Will model selection on biologically motivated *a priori* hypotheses typically identify a model that is close to the truth (Figure 1)?

How can we exhaustively explore congruent classes? Exploring congruence classes can be useful to select a given model among models with identical likelihood, for example, according to a given simplicity criterion. It would also help to understand how many equivalently complex models exist in a congruence class, and how different they are in their specified speciation–extinction dynamics.

Will reasonable priors or regularization techniques that are based on standard biological, statistical, or information-theoretic arguments render λ and μ fully identifiable? Will they provide solutions that are close to the truth?

What is the best technical/computational approach to include regularization techniques in BD models?

Acknowledgments

We thank Amaury Lambert and the Morlon (current and past) research team for discussions. We also thank Knud Jønsson for providing us with the Madagascan vanga tree. H.M. acknowledges funding from the ERC (grant number ERC-CoG PANDA).

Declaration of interests

The authors have no interests to declare.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tree.2022.02.004

References

1. Van Valen, L. (1973) A new evolutionary law. *Evol. Theory* 1, 1–30
2. Gould, S.J. *et al.* (1977) The shape of evolution: a comparison of real and random clades. *Paleobiology* 3, 23–40
3. Raup, D.M. (1985) Mathematical models of cladogenesis. *Paleobiology* 11, 42–52
4. Alroy, J. (2008) Dynamics of origination and extinction in the marine fossil record. *Proc. Natl. Acad. Sci.* 105, 11536–11542
5. Silvestro, D. *et al.* (2014) Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Syst. Biol.* 63, 349–367
6. Simpson, G.G. (1944) *Tempo and Mode in Evolution*, Columbia University Press
7. Stanley, S.M. (1979) *Macroevolution: Pattern and Process*, The John Hopkins University Press
8. Quental, T.B. and Marshall, C.R. (2013) How the Red Queen drives terrestrial mammals to extinction. *Science* 341, 290–292
9. Ricklefs, R.E. (2007) Estimating diversification rates from phylogenetic information. *Trends Ecol. Evol.* 22, 601–610
10. Morlon, H. (2014) Phylogenetic approaches for studying diversification. *Ecol. Lett.* 17, 508–525
11. Stadler, T. (2013) Recovering speciation and extinction dynamics based on phylogenies. *J. Evol. Biol.* 26, 1203–1219
12. Pennell, M.W. and Harmon, L.J. (2013) An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci.* 1289, 90–105
13. Nee, S. *et al.* (1994) The reconstructed evolutionary process. *Phil. Trans. R. Soc. B* 344, 305–311
14. Rabosky, D.L. and Lovette, I.J. (2008) Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Proc. R. Soc. B* 62, 1866–1875
15. Morlon, H. *et al.* (2011) Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci.* 108, 16327–16332
16. Stadler, T. (2011) Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl. Acad. Sci.* 108, 6187–6192
17. Etienne, R.S. *et al.* (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. R. Soc. B* 279, 1300–1309
18. Condamine, F.L. *et al.* (2019) Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* 22, 1900–1912
19. May, M.R. *et al.* (2016) A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods Ecol. Evol.* 7, 947–959
20. Valente, L. *et al.* (2020) A simple dynamic model explains the diversity of island birds worldwide. *Nature* 579, 92–96
21. Bininda-Emonds, O.R.P. *et al.* (2007) The delayed rise of present-day mammals. *Nature* 446, 507–512
22. Jetz, W. *et al.* (2012) The global diversity of birds in space and time. *Nature* 491, 444–448
23. Louca, S. and Pennell, M.W. (2020) Extant timetrees are consistent with a myriad of diversification histories. *Nature* 580, 502–505

24. Pagel, M. (2020) Evolutionary trees can't reveal speciation and extinction rates. *Nature* 580, 461–462
25. Stadler, T. (2009) On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* 261, 58–66
26. Legried, B. and Terhorst, J. (2021) A class of identifiable phylogenetic birth-death models. *bioRxiv* Published online October 18, 2021. https://doi.org/10.1101/2021.10.04.463015
27. Stadler, T. (2013) How can we improve accuracy of macroevolutionary rate estimates? *Syst. Biol.* 62, 321–329
28. Lewitus, E. and Morlon, H. (2018) Detecting environment-dependent diversification from phylogenies: a simulation study and some empirical illustrations. *Syst. Biol.* 67, 576–593
29. Magee, A.F. *et al.* (2020) Locally adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. *PLoS Comput. Biol.* 16, e1007999
30. Helmstetter, A.J. *et al.* (2021) Pulled diversification rates, lineages-through-time plots and modern macroevolutionary modelling. *Syst. Biol.* syab083 Published online October 6, 2021. https://doi.org/10.1093/sysbio/syab083
31. Heath, T.A. *et al.* (2014) The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci.* 111, E2957–E2966
32. Gupta, A. *et al.* (2020) The probability distribution of the reconstructed phylogenetic tree with occurrence data. *J. Theor. Biol.* 488, 110115
33. Manceau, M. *et al.* (2021) The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data. *J. Theor. Biol.* 509, 110400
34. Gavryushkina, A. *et al.* (2014) Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.* 10, e1003919
35. Andréoletti, J. *et al.* (2020) A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences. Published online October 27, 2020. https://doi.org/10.1101/2020.10.27.356758
36. Stadler, T. (2010) Sampling-through-time in birth–death trees. *J. Theor. Biol.* 267, 396–404
37. Didier, G. *et al.* (2012) The reconstructed evolutionary process with the fossil record. *J. Theor. Biol.* 315, 26–37
38. Silvestro, D. *et al.* (2018) Closing the gap between palaeontological and neontological speciation and extinction rate estimates. *Nat. Commun.* 9, 5237
39. Louca, S. *et al.* (2021) Fundamental identifiability limits in molecular epidemiology. *Mol. Biol. Evol.* 38, 4010–4024
40. May, M.R. *et al.* (2021) Inferring the total-evidence timescale of marattialean fern evolution in the face of model sensitivity. *Syst. Biol.* 70, 1232–1255
41. Marshall, C.R. (2017) Five palaeobiological laws needed to understand the evolution of the living biota. *Nat. Ecol. Evol.* 1, 1–6
42. Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer
43. Nee, S. *et al.* (1992) Tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl. Acad. Sci.* 89, 8322–8326
44. Simpson, G.G. (1953) *The Major Features of Evolution*, Columbia University Press
45. Rabosky, D.L. and Lovette, I.J. (2008) Density-dependent diversification in North American wood warblers. *Proc. R. Soc. B* 275, 2363–2371
46. Hastie, T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science and Business Media
47. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B. Stat. Methodol.* 67, 301–320
48. dos Reis, M. *et al.* (2016) Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* 17, 71–80
49. Aristide, L. and Morlon, H. (2019) Understanding the effect of competition during evolutionary radiations: an integrated model of phenotypic and species diversification. *Ecol. Lett.* 22, 2006–2017
50. Hagen, O. *et al.* (2021) gen3sis: a general engine for eco-evolutionary simulations of the processes that shape Earth's biodiversity. *PLoS Biol.* 7, e3001340
51. Hurlbert, A.H. and Stegen, J.C. (2014) When should species richness be energy limited, and how would we know? *Ecol. Lett.* 17, 401–413
52. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B. Stat. Methodol.* 58, 267–288
53. Hastie, T. and Tibshirani, R. (1990) Generalized additive models. In *CRC Monographs on Statistics and Applied Probability*, CRC Press
54. Wood, S.N. (2017) *Generalized Additive Models: An Introduction with R, Second Edition*, CRC Press
55. Alfaro, M.E. *et al.* (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci.* 106, 13410–13414
56. Rabosky, D.L. (2014) Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* 9, e89543
57. Barido-Sottani, J. *et al.* (2020) A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. *Syst. Biol.* 69, 973–986
58. Ronquist, F. *et al.* (2021) Universal probabilistic programming offers a powerful approach to statistical phylogenetics. *Commun. Biol.* 4, 1–10
59. Maliet, O. *et al.* (2019) A model with many small shifts for estimating species-specific diversification rates. *Nat. Ecol. Evol.* 3, 1086–1092
60. Maliet, O. and Morlon, H. (2022) Fast and accurate estimation of species-specific diversification rates using data augmentation. *Syst. Biol.* 71, 353–366
61. Coelho, M.T.P. *et al.* (2019) A parsimonious view of the parsimony principle in ecology and evolution. *Ecography* 42, 968–976
62. Aho, K. *et al.* (2014) Model selection for ecologists: the world-views of AIC and BIC. *Ecology* 95, 631–636
63. Wood, S.N. *et al.* (2016) Smoothing parameter and model selection for general smooth models. *J. Am. Stat. Assoc.* 111, 1548–1563
64. Jønsson, K.A. *et al.* (2012) Ecological and evolutionary determinants for the adaptive radiation of the Madagascan vangas. *Proc. Natl. Acad. Sci.* 109, 6620–6625
65. Harmon, L.J. *et al.* (2010) Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64, 2385–2396