Syst. Biol. 64(3):432–440, 2015

DOI:10.1093/sysbio/syv001
Advance Access publication January 8, 2015

Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies

[1]*Institute of Integrative Biology, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland;* [2]*Institute for Marine and Antarctic Studies, University of Tasmania, Private Bag 49, Hobart, Tasmania 7001, Australia;* [3]*Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Christchurch 8140, New Zealand; and* [4]*Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland*

**Correspondence to be sent to: Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland; E-mail: tanja.stadler@bsse.ethz.ch.*

Received 8 July 2014; reviews returned 10 October 2014; accepted 2 January 2015
Associate Editor: Laura Kubatko

Abstract.—Tens of thousands of phylogenetic trees, describing the evolutionary relationships between hundreds of thousands of taxa, are readily obtainable from various databases. From such trees, inferences can be made about the underlying macroevolutionary processes, yet remarkably these processes are still poorly understood. Simple and widely used evolutionary null models are problematic: Empirical trees show very different imbalance between the sizes of the daughter clades of ancestral taxa compared to what models predict. Obtaining a simple evolutionary model that is both biologically plausible and produces the imbalance seen in empirical trees is a challenging problem, to which none of the existing models provide a satisfying answer. Here we propose a simple, biologically plausible macroevolutionary model in which the rate of speciation decreases with species age, whereas extinction rates can vary quite generally. We show that this model provides a remarkable fit to the thousands of trees stored in the online database *TreeBase*. The biological motivation for the identified age-dependent speciation process may be that recently evolved taxa often colonize new regions or niches and may initially experience little competition. These new taxa are thus more likely to give rise to further new taxa than a taxon that has remained largely unchanged and is, therefore, well adapted to its niche. We show that age-dependent speciation may also be the result of different within-species populations following the same laws of lineage splitting to produce new species. As the fit of our model to the tree database shows, this simple biological motivation provides an explanation for a long standing problem in macroevolution. [Birth–death process; diversification; macroevolution; Stochastic models.]

Macroevolutionary models generate phylogenetic trees representing processes by which an ancestor species evolves a diversity of species through speciation and extinction. Exploring the behavior of such models contributes toward explaining how present biodiversity evolved (Mooers and Heard 1997). Every model is a simplification of a complex system, but such abstractions may help identify patterns and raise new hypotheses. Comparing these models with empirical data enables us to test such hypotheses, and thus helps to understand evolutionary processes and identify particular deterministic forces (Hey 1992). The macroevolutionary models range from simple to complex, and even from the behavior and properties of the simplest ones much can be understood and learned (Hartmann et al. 2010; Stadler 2013).

The most basic macroevolutionary null model is the Yule model (Yule 1924), under which all extant species at a particular point in time are equally likely to undergo a speciation event. The Yule model is appealing for its simplicity but fails to reproduce empirical data, as empirical phylogenies of extant species are generally far less *balanced* (meaning that sister clades are of very different sizes) (Yule 1924; Losos and Adler 1995; Aldous 1996, 2001; Heard 1996; Mooers and Heard 1997; Steel and McKenzie 2001; Pinelis 2003; Blum and François 2006). Likewise, all *species-speciation-exchangeable* models (Stadler 2013), which include, in particular, environmental-dependent or diversity-dependent diversification models, produce the same distribution of tree shapes (i.e., a phylogeny ignoring branch lengths) as the Yule model (Stadler 2013). Therefore, these models are clearly missing important macroevolutionary features. Identifying more general macroevolutionary models that give rise to empirical tree balance will indicate which macroevolutionary dynamics may play major roles in shaping biodiversity.

Under the Yule model, each "ranked" labeled tree shape (i.e., a tree shape with an ordering of internal vertices and unique leaf labels) is equally likely (Aldous 2001), and this leads to highly balanced trees. By contrast, the *Proportional to Distinguishable Arrangements* model (PDA) (Aldous 1996, 2001; Semple and Steel 2003) assumes that each labeled tree shape (disregarding the order of speciation events) is equally likely. The PDA model produces trees that are highly unbalanced and has been biologically motivated by explosive radiation events and the colonization of new niches (Steel and McKenzie 2001). For an example of a perfectly balanced and unbalanced tree see Supplementary Material for Figures 1a and b available on Dryad at http://dx.doi.org/10.5061/dryad.31227, respectively.

The Yule and PDA models lie at opposite ends of the tree balance spectrum with empirical trees generally somewhere in between. Aldous (2001) introduced β-splitting models which span and extend the range of tree balance. In these models, the tree balance can be selected by altering a single parameter, β. Aldous (1996) found evidence that empirical trees support a value of $\beta \approx -1$, however, no biological explanation supports this value.

Mechanistic models that vary speciation rates across species have been suggested before, however, most of them rely on problematic assumptions. Trait-dependent speciation models can match empirical trees, but no obvious trait has been linked to tree shape. Models that

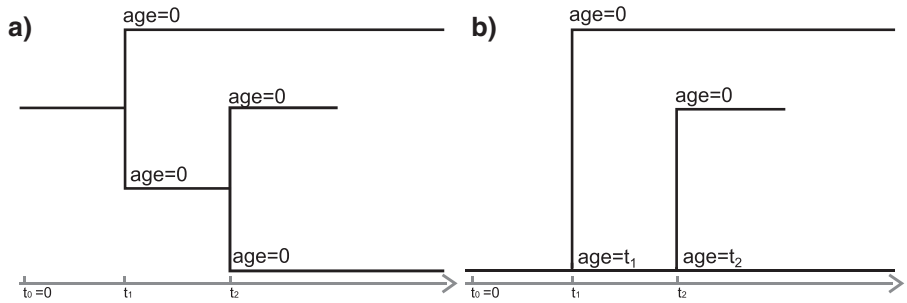

FIGURE 1. Speciation modes. a) The symmetric speciation mode captures the notion of allopatric speciation. All daughter lineages descending from a speciation event start with age 0. b) The asymmetric speciation mode is inspired by peripatric speciation and the two daughter lineages are not treated equally. On every split, one descendant has its age set to 0whereas the other keeps the ancestral "mother" age.

randomly "evolve" speciation rates along lineages may produce imbalances resembling those seen in empirical trees, (Heard 1996; Heard and Mooers 2002; Blum and François 2006) however, it is not clear how to interpret the "evolution" of speciation rates biologically: At each speciation event, the daughter lineages inherit the fraction f and $(1-f)$ of the mother speciation rate (Blum and François 2006).

In our study, we find both analytically and in simulations that trees matching empirical data can be produced across wide parameter ranges by allowing speciation rates to vary across species. We propose an evolutionary model that is both biologically plausible and provides a good match to empirical tree shapes. Under this model, a species is assumed to have a speciation rate that depends on the age of the species and these age-dependent rates are the same across all the species. Age dependence may thus link macroevolutionary processes to ecological dynamics of a species: The species populations changing through time due to ecological pressures (such as predators or pathogens) may give rise to varying macroevolutionary dynamics.

When considering age-dependent speciation (or extinction), we need to define when a species is new, that is, when to reset an age to 0. Two ways to do so are to either set both species descending from a speciation event to age 0 (symmetric speciation mode) or to set one species descending from a speciation event to age 0, whereas the second species inherits the age of the ancestor (asymmetric speciation mode, see Fig. 1 and Methods section).

We analyzed our age-dependent models both mathematically and through simulations before comparing the model tree shape distribution to empirical tree shapes.

METHODS

Age-Dependent Speciation–Extinction Model and Analytic Considerations

Under an *age-dependent model*, the rate of speciation or extinction of a lineage depends on its age. Mathematically, this can be formulated by assigning a probability density $g_s(t)$ to the time for which a newly formed lineage persists until its next speciation event, after which it is replaced by two lineages (one or both of which will constitute "new" lineages depending on the speciation mode). The extinction process is dictated by another probability density $g_e(t)$. The constant rate birth–death (crBD) model is a special case with exponentially distributed waiting times, meaning rates are age independent. Few phylogenetic studies (e.g., Jones 2011) have relaxed the age-independence assumptions.

Lifetime distributions, which are often used to model the time to failure of machinery (or death of organisms), are of particular interest to us, as they are derived from principles that may also apply to a speciation and extinction process. In this article, we focus on the Weibull distribution. The Weibull distribution is often used in reliability theory (e.g., in engineering applications) and is suitable for modeling situations where there are many parts with the same failure distribution and the time to the first failure is the end of a lifetime. This is a "chain model" in the sense that the strength of the chain is determined by the strength of the weakest link.

The Weibull distribution can be justified as a distribution for the time to speciation as follows. If we view a species as a collection of populations, and a speciation event to occur at the first time that one of these populations founds a new lineage, then if we regard these populations as behaving independently and identically with respect to this process, the time T to a speciation event is the minimum of independent and identically distributed (i.i.d.) random variables (one for each population). If the rate at which a population gives rise to a new lineage is constant, then the times to speciation in each population will be exponential distributed, and hence, so too will their minimum T. More generally, regardless of the distribution of the population-specific process, if the number of populations is large, extreme value theory (Fisher and Tippett 1928; McFadden 1978) shows that the distribution of T is of a specific family that generalizes the exponential. For a large range of distributions of the i.i.d. random variables one obtains the two-parameter Weibull distribution. Notice that the Weibull distribution includes as a special case (when the shape parameter

$\Phi=1$) the 1D Exponential distribution (induced e.g., through the minimum of i.i.d. exponentially distributed random variables), which corresponds to a constant speciation rate that is independent of species age.

Assuming a Weibull-distributed time until species extinction can also be justified as follows. Consider a simple model in which extinction of a species occurs when a first "catastrophic" event occurs. Assume further that a large number of possible processes could cause such an event, and the times until each such event occurs are i.i.d. random variables. Then extreme value theory tells us that, under some fairly general conditions, the time until species extinction will follow a Weibull distribution. Note, however, that if we model extinction to occur once the last population of a species dies, and if we regard these populations as behaving independently and identically with respect to this death process, extinction is the maximum of i.i.d. random variables (one for each population). The maximum of i.i.d. random variables with distributions such as Exponential, Normal, Gamma, Lognormal, or Weibull, is the Gumbel distribution.

The two-parameter Weibull distribution probability density is $g(t)=(\Phi/\Psi)(t/\Psi)^{\Phi-1}e^{-(t/\Psi)^{\Phi}}$, where Φ is the so-called shape parameter and Ψ the scale parameter. The mean time until an event is $\mu=\Psi\Gamma(1+\frac{1}{\Phi})$, with Γ being the gamma function. An important characteristic of the Weibull distribution is the rate at which events occur; in our application this is the rate at which a taxon undergoes speciation or extinction, and an $\Phi<1$ corresponds to a speciation rate that declines with species age. Mathematically, it is given by $\lambda(t)=g(t)/\int_t^{\infty}g(t)dt=\frac{\Phi}{\Psi}(\frac{t}{\Psi})^{\Phi-1}$. In the following, we will use the parameters Φ and μ (rather than the common parameters Φ and $\Psi=\mu/\Gamma(1+\frac{1}{\Phi})$) to define the Weibull distribution.

An age-dependent model that assumes a Weibull-distributed speciation process but no extinction induces the Yule and the PDA model at $\Phi_s=1$ and $\Phi_s=0$, respectively (see Supplementary Material for a mathematical proof available on Dryad at http://dx.doi.org/10.5061/dryad.31227). By focusing on the range $0<\Phi_s\leq 1$, we investigate models inducing intermediate balanced trees that are similar to empirical trees.

Under the age-dependent model, one needs to distinguish if, upon speciation, both descending species are "daughter" species with age 0 (*symmetric speciation mode*, Fig. 1a) or one of them remains as the "mother" species, inheriting the age (*asymmetric speciation mode*, Fig. 1b). The symmetric speciation mode may be inspired by allopatric speciation, where the divided subpopulations are exposed to comparable evolutionary forces. The asymmetric model may relate to peripatric speciation. It can be interpreted as an isolation of a small part of the population that suffers stronger speciation forces than the remaining part.

In both of the above-mentioned scenarios, the "new" species of age 0 has a significantly different population size than the mother species, and thus macroevolutionary dynamics may be very different (e.g., increased extinction risk due to small population size). We emphasize that both modes could be related to any speciation mechanism (even parapatric and sympatric speciation), depending on how much the two originated species differ from the "mother" species. Furthermore, the species age is only set to 0 at the time of a branching event (i.e., at cladogenesis), thus, a new species evolving via anagenesis is assumed to inherit the age of the mother species.

Comparing Tree Distributions

Tree shape analysis gives an indication of the difference of the daughter clade sizes of ancestral taxa throughout the tree and provides a simple method by which complex phylogenetic trees can be compared. Broadly used statistics for summarizing tree shape are: Colless (Colless 1982), Sackin (Sackin 1972; Shao and Sokal 1990), and β (Aldous 2001) (the value that maximizes the likelihood in the β-splitting model (Aldous 2001)). In this study, all three statistics were used, with a focus on the results for β, noting that Colless and Sackin outcomes were qualitatively equivalent to β. For the Yule model, the expected mean β is 0 whereas less balanced trees have $\beta<0$; in particular, the PDA model has $\beta=-1.5$.

Branch length analyses tell us about the relative timing of speciation events in a phylogenetic tree, and are, therefore, important for measuring the speciation and extinction rates and overall tree age. A popular statistic summarizing branch lengths is the γ statistic (Pybus and Harvey 2000). A value of $\gamma=0$ indicates that the underlying model is consistent with a Yule model, whereas $\gamma<0$ reflects that branching events are closer to the root than under the Yule model, e.g., caused by a decreasing speciation rate through time. A value of $\gamma>0$ reflects that branching events are generally closer to the tips than under the Yule model. A $\gamma>0$ may be caused by a speciation rate which has increased over time or reflect the contribution of extinction, since a crBD model (with positive extinction rate) induces $\gamma>0$ due to the so-called "pull-of-the-present effect"(Nee et al. 1994).

Simulations

To investigate which parameters of our age-dependent model control tree shape and branch lengths, we simulated phylogenetic trees under different parameter combinations. To illustrate the importance of speciation mode and to test our implementation, the waiting time until speciation and extinction was fixed (meaning a Dirac delta distribution was used, which is a Weibull distribution with shape $\Phi\to\infty$). The main simulation study exploring the effect of age-dependent speciation and extinction rates on the tree distribution assumes Weibull-distributed times until speciation and extinction. The Weibull distribution was parameterized

through its mean μ and shape Φ (rather than scale Ψ and shape Φ). The speciation mean was set to 1 ($\mu_s = 1$), meaning we defined one time unit as the mean time to speciation. Changing the mean time to speciation to a value different from 1 would scale branches in the simulated trees, however, as we only look at tree shape distribution (via Colless, Sackin, and β statistic) and relative branch lengths (via γ statistic), our results are invariant toward the setting of μ_s. We then explored the influence of the remaining three parameters on the induced trees: Speciation shape Φ_s, inverse extinction mean $1/\mu_e$ (referred to as turnover, being here the mean time to speciation divided by mean time to extinction), and extinction shape Φ_e. In addition, two modes of speciation, symmetric and asymmetric, were considered. We simulated trees with a fixed number of extant species (typically 100) and then pruned all extinct tips to obtain simulated phylogenies (comparable to empirical phylogenies). All simulations with nonzero extinction rate used the general sampling approach (Hartmann et al. 2010). In short, the general sampling approach simulates trees with a fixed number of tips, and, compared with standard approaches, does not stop once the number of tips is reached but also considers trees which grow bigger and then shrink due to extinctions to the desired size.

To run the simulations, the central high-performance computer cluster of ETH Zurich (Brutus) was used. To summarize tree shape and branch lengths, we employed widely used summary statistics on simulated and empirical trees (see section "Comparing tree distributions").

Trees for different parameter settings were simulated. First, we assumed a constant rate of speciation (Weibull shape $\Phi_s = 1$) and age-dependent extinction for turnover 0.5 and 0.9. Then, we simulated age-dependent speciation for different Φ_s, without extinction, a constant extinction rate and a Weibull-distributed time until extinction for turnover 0.5 and 0.9.

For the particular case of small tree sizes without extinction, it is possible to characterize the tree shape distribution analytically. This is outlined in detail in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad, "Calculating tree shape probabilities under the symmetric age-dependent model without extinction." However, the probability of particular tree shapes involves nested integrals which appear to possess no explicit analytical solution, and is impractical for large trees.

Data Set

A total of 9243 empirical trees were cached from the *TreeBase* (Sanderson et al. 1994) repository (accessed 16 September 2012) using the R package *treebase* (Boettiger and Temple Lang 2012). From these, only ultrametric binary trees containing branch length information were used for branch length analysis. For tree shape analysis, only fully binary rooted trees with or without branch lengths were selected. Within these, only trees of the kind "species trees" were used. Protein and gene trees were ignored to achieve a straightforward interpretation. Data available from Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.31227.

Macroevolutionary Parameter Estimation Based on Empirical Trees

To find the Φ_s best explaining empirical tree imbalance, a large simulation study was performed (R scripts available fromDryad Digital Repository: http://dx.doi.org/10.5061/dryad.31227). We determined the Φ_s which gave rise to simulated trees most similar to our empirical trees. In detail, 100 trees were simulated each for symmetric and asymmetric speciation modes and number of tips 5, 10, 20, 50, 100, 200, 500, 100, 2000, and Φ_s 0.1, 0.2…0.9, 1. In total 18,000 trees where simulated. No extinction was assumed, since we observed in the simulations above that extinction does not change tree shape much. For each simulated and empirical tree, all three shape statistics (Sackin, Colless, β) were calculated. For the simulated trees, the values of the shape statistics where summarized by the median over the 100 trees with the same speciation mode, same Φ_s, and same number of tips. To have a proxy for the median tree shape statistics of tree sizes which were not simulated, a linear interpolation between the simulated median tree shape statistics for each speciation mode and Φ_s was performed.

Thereafter, for every single empirical tree (i) and for all trees (ii) the best Φ_s was selected. As we did not know if the empirical trees are better modeled by symmetric or asymmetric speciation mode, the selection of the best Φ_s for both modes was performed. In case (i), the respective shape statistic for each tree was compared with the median tree shape statistic of the simulated trees (or the interpolation) for the same number of tips. The Φ_s with the lowest difference between the statistic of the empirical tree and the (interpolated) median was chosen for every single tree. In case (ii), we assumed each empirical tree is described by the same Φ_s. For each Φ_s, we calculated the sum of squared difference of each empirical tree shape statistic and the corresponding (interpolated) median tree shape statistic with shape Φ_s. Since the Colless and Sackin statistics are not normalized with respect to the number of tips (and thus big trees have much bigger summary statistics), we normalized the squared difference by the square of the (interpolated) median tree shape statistic. As β is already normalized, the normalization was not conducted for that statistic. The best Φ_s was finally determined by minimizing this sum of squares.

Empirical trees on TreeBase might still include an out-group. We thus deleted from each TreeBase tree the smaller clade descending the root, as this smaller sister clade may have been an out-group. We analyzed these out-group-corrected trees with the same procedure as the complete trees.

Macroevolutionary Parameter Estimation Based on Simulated Trees

To check whether our procedure yields reliable estimates of Φ_s, we tested it on simulated phylogenies (R scripts available fromDryad Digital Repository: http://dx.doi.org/10.5061/dryad.31227). We simulated the same number of trees with the same number of tips as in the empirical set. A Weibull-distributed speciation process with $\Phi_s = 0.5$ and a constant extinction rate with a turnover of 0.5 was assumed. We simulated under both speciation modes. Additionally, a crBD process with turnover 0.5 was simulated. Thus, we obtained three sets of simulated trees. The best speciation shape was inferred using exactly the same methodology as for the empirical trees (in fact the same simulated trees were used for parameter inference based on the empirical data set and the three simulated data sets).

RESULTS

Analytic Results

We considered age-dependent speciation models without extinction, assuming Weibull-distributed times until speciation, and a symmetric speciation mode (see Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.31227). We show that this model can produce trees ranging in balance from Yule trees to PDA trees depending on a single parameter: The Weibull shape parameter. Consequently, our simple age-dependent speciation model interpolates between and extends Yule and PDA trees. We did not find an analytic way to explore the consequences of extinction and speciation mode on the tree shape distribution. The next section investigates the consequences via simulations.

Simulation

The model implementation is available as an open source R package *TreeSimGM* on CRAN. In the simulation study, we investigated the effect of extinction and speciation processes on the resulting tree distribution (measured via Sackin, Colless, β statistics, and the γ statistic).

Simulating trees under a delta function.—First, to validate the methods and illustrate the effect of the speciation modes, simulations were based on purely deterministic ("Dirac delta distributed") times until speciation/extinction (see Supplementary Materal for Fig. 1 available on Dryad at http://dx.doi.org/10.5061/dryad.31227). The Dirac delta distribution is the limit of a Weibull distribution for shape $\Phi \to \infty$. The analysis reveals that symmetric and asymmetric speciation modes can produce very different tree shapes. Note, however, that under such a distribution the model is completely deterministic and all resulting trees are the same, thus the biological relevance is limited. For illustrative purpose, we set a waiting time until speciation that is always 2 million years and a waiting time until extinction that is always 2.5 million years. The simulations for both speciation modes were stopped after time 10, see Supplementary Material for Figure 1 available on Dryad at http://dx.doi.org/10.5061/dryad.31227 for the resulting symmetric and asymmetric mode tree.

Simulating trees under the crBD model.—Second, we assumed age-independent speciation and extinction rates, that is, a crBD process, to confirm analytic predictions. We confirmed that the crBD model induces a uniform distribution on ranked tree shapes (with a fixed number of extant tips) (Edwards 1970). However, higher γ values were obtained for increased turnover (extinction rate divided by speciation rate), as expected under the pull-of-the-present effect (Nee et al. 1994). Under a crBD model, symmetric and asymmetric models are equivalent as the age of the lineages does not influence the speciation or extinction probability.

Simulating trees under the age-dependent model.—Most importantly, we investigated the impact of age-dependent rates, assuming a Weibull distribution for the time until speciation and extinction. Our simulation results are shown in Figure 2 and Supplementary Material Figures 2 and 3 (available on Dryad at http://dx.doi.org/10.5061/dryad.31227), revealing that tree shape summarized by β is mainly controlled through Φ_s (i.e., the speciation shape parameter), meaning that Φ_e (i.e., the extinction shape parameter) and turnover do not alter β, Sackin, or Colless. In fact, Lambert (2010) showed analytically that an asymmetric speciation mode and constant speciation rate gives rise to the same tree topology distribution and thus β for arbitrary lifetime distributions.

However, the mean γ is, to a large extent, controlled by Φ_e and turnover. Varying Φ_s does not change the mean γ much, though its variance increased for smaller Φ_s. Speciation mode (i.e., symmetric and asymmetric) has little influence on both tree shape or branch lengths.

Figure 2 displays these findings for constant rate speciation and age-dependent extinction (Fig. 2a,b) as well as age-dependent speciation and constant extinction (Fig. 2c,d), with turnover 0.5. Fixing speciation shape to a value different from 1 (here $\Phi_s = 0.5$) to account for age-dependent speciation led to the same qualitative behavior as $\Phi_s = 1$: Tree shape was again not affected by varying the extinction shape (see Supplementary Material for Fig. 2b available on Dryad at http://dx.doi.org/10.5061/dryad.31227), whereas γ changed similarly for both speciation modes (see Supplementary Material for Fig. 2a available on Dryad at http://dx.doi.org/10.5061/dryad.31227). Finally, changing the turnover from 0.5 to 0.9 again led to the same qualitative result (see Supplementary Material for Fig. 2c,d available on Dryad at http://dx.doi.org/10.5061/dryad.31227).

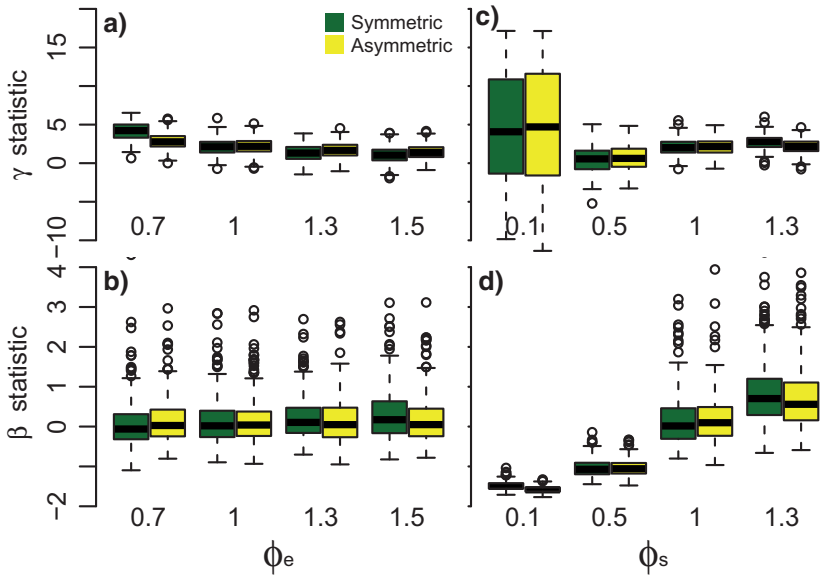

FIGURE 2. γ and β statistics for symmetric (green) and asymmetric (yellow) speciation modes under different age-dependent speciation and extinction models. a,b) Simulation results under a constant speciation rate (i.e., $\Phi_s = 1$) and varying age-dependent extinction rate (i.e., varying Φ_e): γ (a) and β (b) statistic. c,d) Simulation results for varying age-dependent speciation rate (i.e., varying Φ_s) and constant extinction rate (i.e., $\Phi_e = 1$): γ (c) and β (d) statistic. Turnover was fixed to 0.5. Three hundred trees with 100 extant species were simulated for each parameter combination. Boxes enclose 50% and lines 95% of the results and the median is indicated by the horizontal bar for each model.

Fit to Empirical Data

Within the initial 9243 empirical trees downloaded from *TreeBase*, 2759 tree shapes were fully resolved rooted species trees. From the simulations, it was observed that tree shape summarized by β is mainly influenced by Φ_S (Fig. 2, and see Supplementary Material for Figs. 2 and 3 available on Dryad at http://dx.doi.org/10.5061/dryad.31227), and extinction is not changing the tree shape much. Thus, the best Φ_S was determined in a simulation study, assuming no extinction. Note that mean speciation time μ_S only changes the time scale and thus absolute branch lengths but not the tree shape. Thus, the particular setting of μ_S does not change the inference results, and μ_S cannot be estimated based on the tree shape statistic. We used $\mu_S = 1$ as above.

Figure 3 left displays the summary statistics for each tree colored according to the Φ_S best explaining that tree. On the right, we show the histogram with the frequency of trees supported by each speciation shape ranging from $\Phi_S = 0.1, 0.2 \ldots 1$. The histograms all have a peak at intermediate speciation shapes, with the bars for the extremes, $\Phi_S = 0.1$ and $\Phi_S = 0.9$ being high as well, due to all trees which are more extreme being summarized by our most extreme values. The x corresponds to the median Φ_S. It was estimated to be 0.3 based on the Colless and Sackin statistic and 0.7 based on the β statistic, for both speciation modes. As a comparison we also calculated the best Φ_S for all trees sharing the same speciation shape, indicated by *. For Colless, we estimated 0.4 and for Sackin and β we estimated 0.5, for both speciation modes. Thus, these analyses robustly infer a speciation shape between the two classic models, the crBD model and the PDA model. The results are robust toward removing the smaller sister clade at the root, that corrects for the original tree potentially containing a monophyletic out-group (see Supplementary Material for Figs. 6 and 7 available on Dryad at http://dx.doi.org/10.5061/dryad.31227).

In summary, we obtained a strong signal for $0 < \Phi_S < 1$. This implies that young species have a higher chance of speciating than old species.

Among the downloaded species trees, 1710 trees contained branch length information. Of these, only 156 were ultrametric and could be used for calculating γ. Due to the small number of empirical trees suitable for the γ statistics and because the γ statistic is very sensitive to incomplete sampling, a general extinction process was not inferred.

Testing our procedure for parameter estimation over the three sets of simulated phylogenies (see Supplementary Material for Figs. 8, 9, and 10 available on Dryad at http://dx.doi.org/10.5061/dryad.31227) validated our implementation (and consequently our findings), and further elucidated some properties of the phylogenies resulting from age-dependent processes (here Weibull-distributed waiting times). First, we observed again that the extinction process and the speciation mode do not influence tree shape statistics much for the parameters considered. Indeed, the three simulated tree data sets were obtained using a turnover of 0.5, whereas the trees used for inference had no extinction: Nevertheless we obtain reliable estimates of Φ_S. Furthermore, using either speciation mode in the inference results in the same reliable

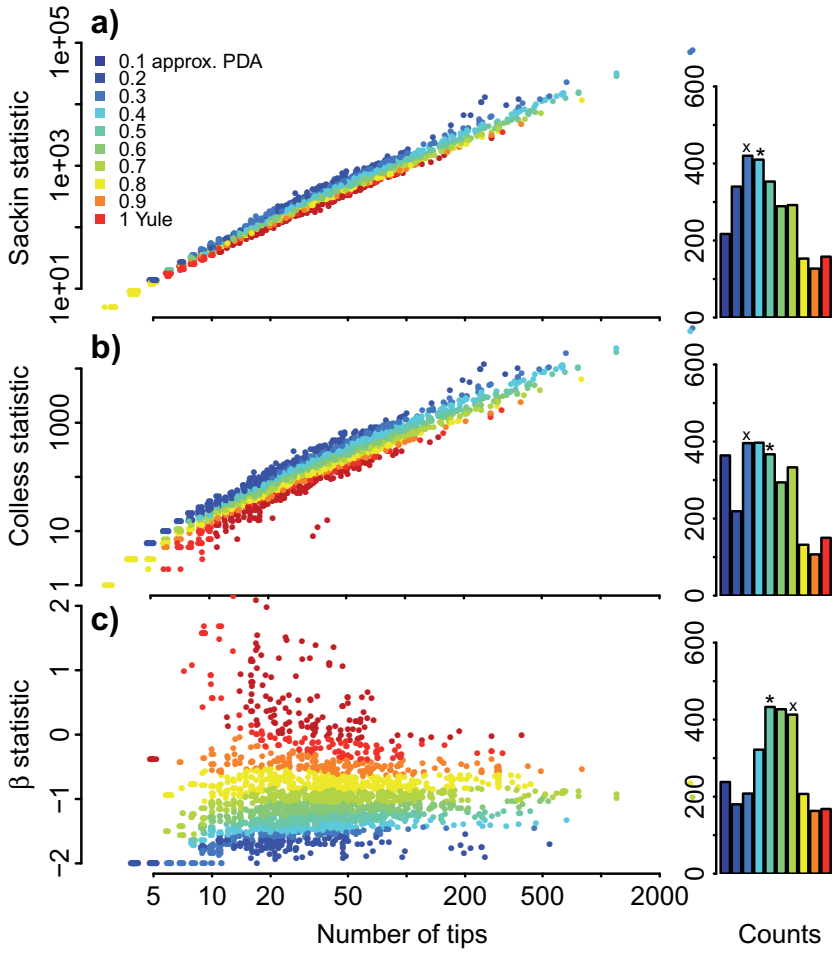

FIGURE 3. Empirical tree shape statistics and best Φ_s based on three different tree balance statistics and an asymmetric speciation mode (symmetric mode yields almost the same results, see Supplementary Data for Fig. 4 available on Dryad at http://dx.doi.org/10.5061/dryad). The empirical tree summary statistics (left) are coloured by their corresponding best Φs fit, with (a) Sackin, (b) Colless, and (c) β statistics. A histogram displaying the number of trees for each Φ_s ranging from 0.1 (the approximated-PDA model, blue) to $\Phi_s = 1$ (Yule model, red) is shown on the right. The x in the histogram denotes the median Φ_s and the * denotes the best Φ_s when assuming all trees share the same shape parameter.

estimates of Φ_s. Second, the median of the inferred Φ_s over all trees (indicated by x) was always lower than the inferred Φ_s assuming the same speciation shape for all trees (indicated by *) for Sackin and Colless. The opposite holds for the β statistic (see Supplementary Material for Figs. 8, 9, and 10 available on Dryad at http://dx.doi.org/10.5061/dryad.31227). The same pattern was also observed in the inferred Φ_s based on empirical data (Fig. 3 and see Supplementary Data for Fig. 4 available on Dryad at http://dx.doi.org/10.5061/dryad.31227).

DISCUSSION & CONCLUSIONS

Mechanistic models previously proposed in the literature give rise to phylogenies which are very different from empirical trees, for example, the Yule model (Blum and François 2006). Yule trees, for example, give rise to a tree shape distribution with a β statistic of around 0, whereas our empirical data analysis suggests a β statistic around −1 (Fig. 3). This empirical observation was stated by David Aldous already in 1996 along with his comment "I do not know a natural candidate for such a process" (Aldous 1996, p. 12). We now provide a simple biologically plausible model, namely an age-dependent speciation model, that produces trees with $\beta \approx -1$. Our model contains the well-known PDA and Yule model as special cases. The β-splitting models (or the more general Markov splitting model (Aldous 1996)) also contain the PDA and Yule models, however, these models are different from the age-dependent speciation model and do not have an obvious biological interpretation.

It was previously shown that species-speciation-exchangeable models (i.e., age-independent speciation models such as time-dependent or diversity-dependent speciation and extinction models) all produce the same tree shape distribution (uniform on ranked trees) (Stadler 2013). Our simulations extend this observation, revealing that tree shape in general is influenced mainly by the speciation process and is largely invariant to the extinction process. However, branching

times are influenced to a large extent by extinction. Speciation mode (symmetric or asymmetric) has almost no influence on tree shape or branching times for the parameter ranges investigated (though extreme scenarios of age dependence such as the Dirac delta lifetime distribution can lead to very different trees, see Supplementary Material for Fig. 1 available on Dryad at http://dx.doi.org/10.5061/dryad.31227).

In contrast to our study, Venditti et al. (2010) found support for a Yule model instead of an age-dependent model for speciation. However, the former study only considered branch length distributions and ignored tree shapes. Our simulations show that speciation mainly influences tree shape whereas extinction influences branch lengths. As branch lengths are not changed much when using the two-parameter age-dependent speciation model compared with the one-parameter Yule model, the simpler Yule model is expected to be preferred when ignoring tree shape, reconciling the findings of our study and Venditti et al. (2010).

The classic macroevolutionary models (such as the Yule model, species-speciation-exchangeable models, and the age-dependent model) assume that speciation is instantaneous. Relaxing this assumption by assuming a fixed time until speciation is completed predicts even more balanced trees (Losos and Adler 1995). However, assuming a constant rate until speciation completion may produce trees with the induced β distribution having a peak at -1 (Fig. S6-1 in Etienne and Rosindell 2012). However, the distribution obtained from this protracted speciation model is very wide, and the median β value is larger than -1. In contrast, our model, assuming a speciation rate decreasing with species age, gives rise to trees with an induced β distribution having most values at $\beta \approx -1$ (Fig. 2d). Note though that the γ statistic rather than the βstatistic was the main focus of the Etienne and Rosindell (2012) study, and future work should further investigate how well the protracted speciation model can produce realistic β values.

There may be several possible mechanisms for a decreasing speciation rate supported by the empirical data. Here is a particularly simple one. Suppose the time to speciation is the time until the first population within a species founds a new lineage, and the times to lineage foundation in the populations are i.i.d. random variables. Then under fairly general conditions, extreme value theory shows that the time until speciation is described by a Weibull distribution asymptotically (i.e., as the number of populations grows). Other mechanisms are also conceivable for age-dependent speciation. A newly evolved species having a novel trait opening up a new niche space may be exposed to more possible niches (e.g., resources), thus future new species would have more possible niches to occupy and establish themselves, and so on. Alternatively, a new species may have fewer competitors or predators, and only once the species is established the competitors and predators become coadapted. However, these mechanisms cannot be determined through phylogenetic analysis (as the different mechanisms being modeled via age-dependent speciation give rise to the same phylogenies). Our methods cannot test different mechanisms for age dependence, however, our results indicate that future empirical studies should investigate possible causes of decreasing speciation rate with species age.

While we looked at all empirical trees simultaneously, a future avenue may be to look within clades and determine clade-specific age-dependent processes, highlighting differences of macroevolutionary processes across the tree of life. Moreover, although we have concentrated on the Weibull distribution here, other distributions may also be of interest. Furthermore, we simulated species trees assuming all species of a clade are sampled. A future topic to explore will be the effect of incomplete species sampling. We note that our results are robust toward possible out-group removal in empirical trees. With our implementation being available within the R package *TreeSimGM* on CRAN, and further scripts on how the package was used in this study being available on Dryad, http://dx.doi.org/10.5061/dryad.31227, simulations under any distributions with subsequent species sampling is straightforward to do. Thus, for future studies, our implementation may be a powerful tool to obtain a more complete picture of macroevolutionary processes.

In summary, we have presented a biologically plausible mechanistic model that is capable of adequately describing empirical tree shape distribution. By employing age-dependent decreasing speciation rates, our model provides a robust fit to empirical data. This mechanistic model has a simple biological interpretation— the probability of a speciation event occurring on a lineage decreases as the time since the last speciation event increases. Based on empirical tree topologies, we identified and quantified age-dependent speciation. The reconstruction of large-dated phylogenetic trees will soon allow us to compare the model presented here with empirical branch lengths, shedding light on extinction processes. However, it is important to stress that it seems impossible to determine a predominant speciation mode from a phylogenetic tree: For the parameter range best explaining the empirical phylogenies, the symmetric and asymmetric mode produces very similar trees *in silico*.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.31227.

Funding

T.S. is supported in part by the European Research Council under the 7th Framework Programme of the European Commission (PhyPD: Grant Agreement Number 335529).

Acknowledgments

The authors thank the editors Frank Anderson, Laura Kubatko, and four anonymous reviewers for their extensive and constructive suggestions.

References

Aldous D.J. 1996. Probability distributions on cladograms. In: Aldous D.J., Pemantle R., editors. Random discrete structures. New York: Springer. p. 1–18.

Aldous D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to Today. Stat. Sci. 16:23–34.

Blum M.G.B., François O. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55:685–691.

Boettiger C., Temple Lang D. 2012. Treebase: an R package for discovery, access and manipulation of online phylogenies. Methods Ecol. Evol. 3:1060–1066.

Colless D.H. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool. 31:100–104.

Edwards A.W.F. 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. B. 32:155–174.

Etienne R.S., Rosindell J. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61:204–213.

Fisher R.A., Tippett L.H.C. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Cambridge Phil. Soc. 24:180–190.

Hartmann K., Wong D., Stadler T. 2010. Sampling trees from evolutionary models. Syst. Biol. 59:465–476.

Heard S.B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50:2141–2148.

Heard S.B., Mooers A.Ø. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. Syst. Biol. 51: 889–897.

Hey J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640.

Jones G. 2011. Calculations for multi-type age-dependent binary branching processes. J. Math. Biol. 63:33–56.

Lambert A. 2010. The contour of splitting trees is a Levy process. Ann. Prob. 38:348–395.

Losos J.B., Adler F.R. 1995. Stumped by trees? A generalized null model for patterns of organismal diversity. Amer. Nat. 145:329–342.

McFadden D. 1978. Modeling the choice of residual location. In: Karlqvist A., Lundqvist L., Snickars F., Weibull J.W., editors. Spacial interaction theory and planning models. Amsterdam: Studies in Regional Science and Urban Economics. p. 75–96.

Mooers A.O., Heard S.B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Bio. 72:31–54.

Nee S., Holmes E.C., May R.M., Harvey P.H. 1994. Extinction rates can be estimated from molecular phylogenies. Phil. Trans. R. Soc. Lond. B. 344:77–82.

Pinelis I. 2003. Evolutionary models of phylogenetic trees. Proc. R. Soc. Lond. B. 270:1425–1431.

Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B. 267:2267–2272.

Sackin M.J. 1972. Good and bad phenograms. Syst. Zool. 21:225–226.

Sanderson M.J., Donoghue M.J., Piel W.H., Eriksson T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am. J. Bot. 81:183.

Semple C., Steel M. 2003. Phylogenetics. New York: Oxford University Press.

Shao K.T., Sokal R.R. 1990. Tree balance. Syst. Zool. 39:266–276.

Stadler T. 2013. Recovering speciation and extinction dynamics based on phylogenies. J. Evolution. Biol. 6:1203–1219.

Steel M., McKenzie A. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170:91–112.

Venditti C., Meade A., Pagel M. 2010. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature 463:349–352.

Yule G.U. 1924. A Mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Phil. Trans. R. Soc. Lond. B. 213:21–87.