Evolutionary biology

Estimating shifts in diversification rates based on higher-level phylogenies

Tanja Stadler\textsuperscript{1,2} and Jana Smrckova\textsuperscript{3}

\textsuperscript{1}Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland
\textsuperscript{2}Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland
\textsuperscript{3}Department of Zoology, Faculty of Science, University of South Bohemia, 37005 Ceske Budejovice, Czech Republic

Macroevolutionary studies recently shifted from only reconstructing the past state, i.e. the species phylogeny, to also infer the past speciation and extinction dynamics that gave rise to the phylogeny. Methods for estimating diversification dynamics are sensitive towards incomplete species sampling. We introduce a method to estimate time-dependent diversification rates from phylogenies where clades of a particular age are represented by only one sampled species. A popular example of this type of data is phylogenies on the genus- or family-level, i.e. phylogenies where one species per genus or family is included. We conduct a simulation study to validate our method in a maximum-likelihood framework. Further, this method has already been introduced into the Bayesian package MrBayes, which led to new insights into the evolution of Hymenoptera.

1. Introduction

A key goal in macroevolution is to identify changes in the rates of diversification and to find causal explanations for variations in the species diversity we observe today. It has been shown that species phylogenies based on only extant taxa and no extinct lineages can be used to infer both the speciation and extinction rates, and thus in particular the diversification rate defined as speciation rate minus extinction rate \cite{1–3}.

Many organismal clades contain a vast amount of species, which makes construction of complete species phylogenies an arduous task. Although for some species groups complete or near-complete phylogenies have been already inferred \cite{4,5}, many others are still available as phylogenies on a higher taxonomic level only, meaning that only one species per higher taxonomic unit such as genus or family is included \cite{6–10}. We call such phylogenies ‘higher-level phylogenies’.

Computational methods were developed to estimate diversification rates from higher-level phylogenies \cite{7,11–14}. While Paradis \cite{11} and Stadler & Bokma \cite{14} have devised a method for estimating constant speciation and extinction rates in higher-level phylogenies, Rabosky \textit{et al} \cite{12}, Alfaro \textit{et al} \cite{7} and Rabosky \cite{15} refined these approaches by also allowing for the computation of speciation and extinction rate variation across clades. Additionally, rate variation through time may be induced by external variables, such as climate, break-up of continents, sea-level changes or development of key innovations or competition. In this paper, we introduce a framework allowing for the estimation of changes in diversification rates through time from higher-level phylogenies. Our mathematical equations derived here have been...
implemented into MrBAYES, and used to infer a phylogeny of Hymenoptera through a Bayesian approach [16].

In what follows, we present a maximum-likelihood method to estimate changes in diversification (= speciation – extinction) rates and turnover (= extinction/speciation) for higher-level phylogenies where all phylogenetic relationships are resolved up to a certain point in time, and each clade, descending a lineage at that point in time, is collapsed to one tip. We show in a simulation approach that shifts in diversification rates can be estimated reliably based on our likelihood framework. We then explain how we can transform a phylogeny on the genus- or family-level into a higher-level phylogeny to analyse empirical data.

2. Methods

(a) Birth–death–skyline model

We extend the constant rates birth–death process (crBDP; [17–19]), to the birth–death–skyline process, following Stadler [20] and Stadler et al. [21].

The crBDP starts with a single lineage at time \( t_0 \) in the past (stem age) and gives birth to descendant lineages with a constant rate of speciation \( \lambda \) and lineages die with a constant rate of extinction \( \mu \). At the present time, the process is stopped. The reconstructed phylogenetic tree is acquired by pruning all lineages that went extinct.

The birth–death–skyline process generalizes the crBDP by allowing for rate changes through time: time between the present with the information of number of species in the higher-level phylogeny. First, we simulated species trees at time \( t_0 \) is split up through 0

(b) Higher-level trees

A higher-level phylogeny is obtained from a complete species phylogeny by pruning the extant descendants of every lineage at time \( t_0 \) in the past to one sampled lineage together with the information on the number species represented by each sampled lineage (see e.g. fig. 1d in [14]). Let the branching times in the phylogeny be \( x_1 > x_2 > \ldots > x_{n-1} \), and let the number of species represented by tip \( i \) be \( n_i \). We derived the probability of a higher-level phylogeny, in order to estimate maximum-likelihood diversification rates and turnover.

(c) Simulations

We investigate the accuracy of parameter estimation based on simulated higher-level phylogenies. First, we simulated species trees with 2000 tips under different diversification scenarios with one rate shift at 2 Mya before and after the rate shift, together with the rate of the shifts, have been obtained using TreePar v. 3.3 function bd.shifts.optim [25] with the ‘groups’ option. This function employs equation (3.1) below.

The code required to perform our analyses is provided in the electronic supplementary material.

3. Results

(a) Probability of a higher-level phylogeny

The probability density of a higher-level phylogeny \( T \) with \( n \) tips is derived in the electronic supplementary material, theorem 3,

\[
P(T | \lambda_0, \mu_0, x_1) = \frac{p(x_1 | \lambda_0, \mu_0)^2}{(1 - p(x_1 | \lambda_0, \mu_0))^2} \prod_{i=2}^{n-1} \lambda_i p_i(x_1 | \lambda_i, \mu_i) \prod_{i=1}^{n-1} p_i(h_i | \lambda_i, \mu_i),
\]

where \( p_i(t | \lambda_i, \mu_i) \) is the probability that a lineage at time \( t \) in the past has \( k \) descendants at present time 0, and \( \lambda_i \) is the speciation rate at time \( x_i \). This is equivalent to (appendix, theorem 1)

\[
P(T | \lambda, \mu, h, x_1) = \frac{1}{F(x_1)^2} \frac{1}{f(t)} \int \left( \frac{1}{f}(t) \right)^{m-n} \prod_{i=1}^{n-1} \left( \frac{1}{f}(t) \right)^{m-n},
\]

with, for \( t \) in \( (t_i, t_{i+1}) \), and re-defining \( t_{i+1} := t \) for convenient notation

\[
f(t) = 1 \sum_{k=0}^{j} G(t),
\]

\[
G(t) = \lambda_k \int_{h_k}^{\mu_k} \left( e^{(\lambda_k - \mu_k)(t_{i+1} - t_i)} - 1 \right) e^{\sum_{r=i}^{i+1}(\lambda_k - \mu_k)(t_{r+1} - t_r)} dt,
\]

and \( F(t) = \lambda_0 e^{(\lambda_0 - \mu_0)(t_{i+1} - t_i)} e^{\sum_{r=i}^{i+1}(\lambda_0 - \mu_0)(t_{r+1} - t_r)} \),

i.e. we do not need to specify how many species belong to each tip, but we need to know only the total number of species \( m \). In fact, even if we knew how many species belong to each tip, that would not improve parameter estimates.

(b) Diversification rate estimates

In all simulations, diversification rates, turnover and times of the shifts were estimated correctly when analysing fully sampled species-level phylogenies (figures 1 and 2; electronic supplementary material, figure S1–S4). With the complete
4. Discussion

Not accounting for the sampling of higher-level taxa can lead to severe biases in parameter estimation, in particular an underestimation of extinction rate and turnover. We have formulated an inference framework for estimating shifts in diversification rate in higher-level phylogenies where all higher levels have the same age. Our simulations reveal that the method can estimate past diversification patterns from these higher-level phylogenies on extant species. Any phylogeny can be converted into a higher-level phylogeny by collapsing all clades descending from a lineage at time \( x_{cut} \) where \( x_{cut} \) is a time point prior to which the phylogeny is fully resolved.

It has been shown that incorporating fossils will dramatically improve the quality of diversification rate estimates: in particular, the extinction rate can be estimated far better [26]. While Silvestro et al. [27] applied the model presented here to only fossil data, Zhang et al. [16] combined equation (3.1) of this paper with a fossilization process. Thus, coherent analysis of fossils and extant species data became possible. The resulting so-called fossilized birth–death process [28] has been implemented into MrBayes and a higher-level phylogeny with fossils of Hymenoptera has been inferred using this method.

The analyses of Zhang et al. [16] revealed that not accounting for the higher-level phylogeny structure, but assuming each species was sampled uniformly at random, has a drastic effect on tree inference. Thus, it is important to use appropriate diversification models not only for quantifying diversification rates, but also for inferring the phylogenies in the first place.

Here, we accounted for only one rate shift, even though our mathematical expression allows for an arbitrary number. The reason for this is that maximizing over multiple rate shift times is numerically very hard, and often leads the optimization tool to be stuck in local optima. In Stadler [20], a greedy approach for finding rate shifts was implemented, meaning the optimizer first searches for the best rate shift time, and with fixing this first rate shift time, it finds the

Figure 1. Results of a simulation study for trees with constant extinction rate (0.1) and decreasing diversification rate (from 0.9 to 0.4) with a rate shift at 2 myr BP. Central tendency in boxplots is median, vertical lines indicate original values used in simulations. Rows show results for trees with increasing \( x_{cut} \). Plots in column (a) depict the estimated maximum-likelihood turnover and diversification rate parameters. Parameters between the present and the time of the shift and before the shift are denoted as \( t_m \), \( t_p \), and \( t_a \) and \( t_e \) respectively. Plots in column (b) estimated maximum-likelihood shift times in myr, (c) sizes of the analysed trees and (d) \( p \)-values of the likelihood ratio test (using the Chi-squared distribution with 3 degrees of freedom, for speciation rate, extinction rate and shift time) comparing a single rate shift with a constant rate model.
second best rate shift time, etc. However, with the implementation of our mathematical equation into the Bayesian framework MrBayes, we can directly infer the posterior distribution of rate shifts and thus do not rely on a greedy approximation.

The birth–death–skyline process still makes a number of limiting assumptions, in particular that all species are assumed to be indistinguishable, hence all have the same speciation and extinction rates and the same probabilities of being sampled at any given time. Accordingly, a birth–death–skyline model cannot allow us to explicitly test scenarios of heritable rates [29] or clade-dependent rates (Medusa: [7], BAMM: [15]), although it can be used as a null model for testing more complex patterns of diversification. It remains a future challenge to combine complex models of rate variation across clades and through time, for inference of diversification rates based on higher-level phylogenies with fossils.

Authors’ contributions. Both authors contributed to the design of the paper, performed analyses and wrote the paper. Both authors approve the final version of the manuscript, and agree to be held accountable for the content therein.

Funding. This study is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 333529).

Competing interests. We declare we have no competing interests.

Acknowledgements. We thank Nick Matzke for providing the scripts for clade pruning.

References


