The error in Bayesian phylogenetic reconstruction when speciation is not instantaneous

Richèl J.C. Bilderbeek

Rampal S. Etienne

The tools for reconstructing phylogenetic relationships between taxonomic units (e.g. species) have become very advanced in the last three decades.Among the most popular tools are Bayesian approaches, such as BEAST, MrBayes and RevBayes, that use efficient tree sampling routines to create a posterior probability distribution of the phylogenetic tree. A feature of these approaches is the possibility to incorporate known or hypothesized structure of the phylogenetic tree through the tree prior. It has been shown that the effect of the prior on the posterior distribution of trees can be substantial.

Currently implemented tree priors assume that speciation is instantaneous, where we know that speciation can be a gradual process.Here we explore the effects of ignoring the protractedness of the speciation process with an extensive simulation study.We compare the inferred tree to the simulated tree, and find that ....

Furthermore, we identify an important issue related to protracted speciation: because the tree produced by the protracted birth-death process is not necessarily monophyletic, we cannot speak of “the” species tree, but we have to sample among the incipient species to represent species.

**Keywords:** computational biology, evolution, phylogenetics, prior choice

# Introduction

The computational tools that are currently available go beyond the wildest imagination of those living four decades ago. Advances in computational power allowed the first cladograms to be inferred from DNA alignments in 1981 (), and the first Bayesian tools emerged in 1996 (),providing unprecedented flexibility in the setup of a phylogenetic model.

Currently, the most popular Bayesian phylogenetics tools are BEAST () and its successor BEAST2 (), MrBayes () and RevBayes (). They allow known or hypothesized structure of a phylogenetic tree-to-be-inferred to be incorporated through model priors. From these priors and an alignment of DNA, RNA or protein sequences, they create a posterior distribution of phylogenies and parameter estimates (of the models used as prior), in which more probable combinations are represented more often. Each of these tools use efficient tree sampling routines to rapidly create an informative posterior.

The model priors in Bayesian phylogenetic reconstruction can be grouped in three groups: (1) site model, specifying nucleotide subsititutions, (2) clock model, specifying the rate of mutation per lineage in time, and (3) tree prior, embodying the speciation model underlying branching events (speciation) and branch termination (extinction). The effect of choosing a (potentially wrong) prior affects the posterior. For example, recently, it was shown that the effect of choosing a tree prior biases the estimation of the molecular clock rate, for DNA sequences of 100-1000 base pairs ().

Current phylogenetic tools use tree priors that assume speciation is instantaneous, whilst we know thatspeciation is often a gradual process. The (constant-rate) birth-death (BD) model is a commonly used tree prior, but it ignores this temporal aspect of speciation. The protracted birth-death (PBD) model, an extension of the BD model, does incorporate the idea that speciation takes time. In this model, a branching event does not give rise to a new species, but to a new species-to-be, called an incipient species. Such an incipient species may go extinct, finish its speciation to become a good species, or give rise to new incipient species. Protracted speciation may explain observed declines in lineage accumulation (Etienne & Rosindell 2012).

Unfortunately, a tree prior according to this model, providing the probability of a species tree under the protracted birth-death model, is unavailable in current Bayesian phylogenetic tools. Whilst an approximate formula for this probability has been derived (Lambert et al. 201?) and the approximation is very good (Simonet et al 2018), it has not been implemented as tree prior yet. There are various reasons for this. First, the computation of this probability involves solving a set of non-linear differential equations, and while this computation is quite fast, it still takes much more time than the corresponding probability of the BD model which is a simple analytical formula. In a Bayesian MCMC chain the tree prior probability must be calculated many times, and hence the total computation will take considerably longer with a PBD tree prior. Furthermore, the approximate probability is a probability for the species tree assuming an underlying incipient species tree. It can be safely used as tree prior when only one individual per species is sampled, but if one has multiple samples per species – which is currently often the case – the methods to account for this such as the multi-species coalescent (Heled & Drummond, 200?) may not be compatible with the underlying incipient species tree. More precisely, the phylogeny under the PBD model may contain paraphylies, while the multi-species coalescent was developed exactly to avoid this by explaining them as incomplete lineage sorting. Because of the paraphylies there is no such thing as a true species tree in the PBD model. To get a species-level tree one must sample one incipient species per species. Which incipient species is sampled may therefore have an impact on the species tree.

Here we aim to explore the effect of using the BD prior on PBD simulated phylogenies, taking into account possible sampling effects. In brief, we simulate protracted phylogenies using the PBD process, from which we sample a species tree in two very different ways. Given this species tree, we simulate a DNA sequence alignment. Then, we use BEAST2 on these alignments to infer a posterior of phylogenies, using a BD prior. We quantify the difference between the (BD) posterior phylogenies and the simulated (PBD) species tree.

# Methods (but we are not allowed to keep this header)

The PBD model has five biological parameters (see [table:parameters]), which we explore in a factorial fashion, excluding some combinations. We assume that the speciation initiation and extinction rates of an incipient and good species are equal ( and ) **[NOTE: I am unconvinced. I think we should also explore and ]**. We only simulate a PBD process for phylogenies in which speciation initiation exceeds extinction rate (), and in which their difference is not too large (), to prevent overly taxon-poor and taxon-rich phylogenies respectively. The parameter values chosen are based on the parameter sets used by Etienne et al (2014) , as these parameters were shown to result in reasonably sized phylogenies and using the same set allows us to compare results. For the speciation initiation rate , we use , and speciation initiation events per (good species) lineage per time unit. The speciation completion rates used are , , and speciation completion events per (incipient species) lineage per time unit. We used to mimick the BD model, because, the PBD model reduces to the BD model for , This allows us to measure the baseline error, i.e. the difference between inferred tree and true tree that arises purely due to noise because the generating model and the model used in inference are identical. The extinction rates used are , , and extinction events per (good or incipient) lineage per time unit.

From each biological parameter set, we simulated a protracted birth-death tree using the PBD package () in the R programming language (), all with a crown age of 15 million years. Each protracted birth-death tree uses a different random number generatior seed, and thus will be unique, resulting in a balanced data set.

From each incipient species tree, we construct a species tree by sampling one incipient/good species per good speciesFor example, when an incipient species branched off from its mother lineage, both are recognized as representing the species and hence both can be picked as an (equally good) representative of the species. Here, we use three sampling scenarios, in which we pick the youngest, oldest or a random representative.

Based onthe sampled species tree, we simulate a DNA alignment that has the same history as this species tree, using the phangorn package (). We assume that the nucleotides of the DNA alignment follow a Jukes-Cantor () nucleotide substitution model, in which all nucleotide-to-nucleotide transitions are equally likely. In our Bayesian inference (see below) we use the same site model as the (obviously correct) site model prior. One could explore other substitution models in the simulations and in the Bayesian inference, but we chose this simple model because we are primarily interested in the effect of the choice of tree prior. If anything, our results are conservative: with a more complex substitution model, there will be more noise and hence our inference error will increase. We set the mutation rate in such a way to maximize chronologic information. To do so, we set the mutation rate set such that we expect on average one (possibly silent) mutation per nucleotide between crown age and present, which equates to mutations per million years. The DNA sequence length is chosen to provide a resolution of years, that is, to have one expected nucleotide change per years per lineage on average. As one nucleotide is expected to have on average one (possibly silent) mutation per 15 million years, nucleotides result in 1 mutation per alignment per years (which is coincidentally the same as ?). The simulation of these DNA aligments follows a strict clock model, which we will specify as the known clock model prior in the Bayesian inference.

From an alignment, we run a Bayesian analysis and create a posterior distribution of trees and parameters, using the babette () package that sets the input parameters similar to Beauti2 and then runs BEAST2. For our site model, we assume a Jukes-Cantor nucleotide substitution model, as used in the simulation of the alignment. For our clock model, we assume a strict clock with the same fixed rate as used in the simulation of the alignment **[NOTE: did not use a fixed clock rate, I do not see why]**. The tree prior assumed in inference is the BD model, because studying the effect ofthis assumption is the goal of this study. We assumed an MRCA prior with a tight normal distribution around the crown age by choosing the crown age as mean, and a standard deviation of time units, resulting in 95% of the crown ages used have the same resolution (of time units) as the alignment. We ran the MCMC chain to generate 1111 states, of which we removed the first 10% (also called the ’burn-in’). Of the remaining 1000 MCMC states **[NOTE: Why 1000? Why not 250? I would say 250 is preferable, as the information will be more dense]**, the effective sample size (ESS) of the posterior must at least be 200 for a strong enough inference (). An ESS can be increased by increasing the number of samples or decreasing the autocorrelation between samples. Would the ESS be less than 200, we decrease autocorrelation by doubling the MCMC sampling interval of that simulation, until the ESS exceeds 200.

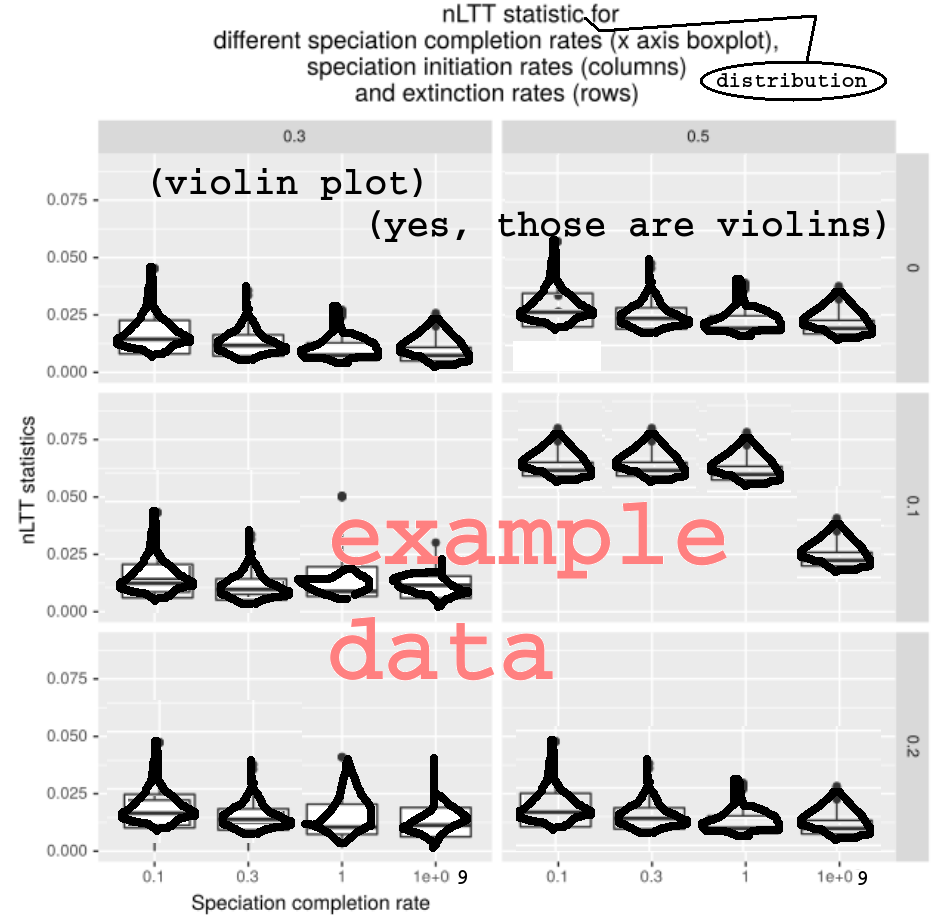
We compared each posterior phylogeny to the (sampled) species tree by the nLTT statistic (), using the nLTT package (). The nLTT statistic is equal to the area between the normalized lineages-through-time-plots of two phylogenies, which has a range from zero (for identical phylogenies) to one. We use inference error and nLTT statistic interchangeably. Comparing the one (sampled) species tree with each of the posterior species trees yields a distribution of nLTT statistics.

We produce two data sets. The first data set is a general data set to chart the effect of the biological parameters on the nLTT statistic distribution. In this data set, incipient species are sampled randomly to represent a good species. The second data set charts the effect of sampling subspecies and only uses PBD trees in which this sampling has an effect. For each of these trees, we sample both youngest and oldest incipient species, to obtain the most pronounced differences.

We stored each data set as a comma-separated file. We set the number of replicates for each parameter combination such that this file and a possible copy can be handled in R’s memory. Each row will then contain a parameter set and the generated nLTT statistics (see [table:specs] for the exact data specification). The abovementioned memory constraints allow for rows. With 48 combinations of biological parameters **[NOTE: calculate exact number when Rampal decides for only or for also allowing ]**, there will be 168 [NOTE: recalculate, as before] replicates per parameter set.

For both data sets, we plot the nLTT statistics distribution per parameter set using a violin plot, as such a plot maintains information about the distribution.

# Results



nLTT statistic distribution per biological parameter set, using the balanced data set



nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect

# Glossary

Glossary [NOTE: this is requested by the journal]

|  |  |
| --- | --- |
| Term | Definition |
| Phylogenetics | The inference of evolutionary relationships of groups of organisms using genetics |
| Model prior | Knowledge or assumptions about the ontogeny of evolutionary histories |
| Posterior | A collection of phylogenies and parameter estimates, in which more probable combinations (determined by the data and the model prior) are presented more frequently |
| Protracted speciation | The process in which speciation takes two events to complete: a speciation-initiation event and a speciation-completion event |
| Speciation initiation | The start of a speciatiation event creating an incipient species. |
| Speciation completion | The end of a speciatiation event, in which an incipient species is recognized as a good species. |

# Acknowledgements

**[NOTE: journal does not request for this. Suggest to remove, but how to acknowledge Peregrine otherwise?]** We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

# Authors’ contributions

**[NOTE: journal does not request for this]** RSE conceived the idea for this experiment. RJCB created and tested the experiment and package, and wrote the first draft of the manuscript. RSE contributed substantially to revisions.

Overview of the 12 simulation parameters. Above the horizontal line is the biological parameter set. Sampling method is random for the general data set. For the data set exploring the effect of sampling, MRCA is used for odd values of , and MDCA is used for even values of . is 1 for the first simulation, 2 for the next, etcetera.

|  |  |  |
| --- | --- | --- |
| Parameter | Description | Values |
|  | Speciation initiation rate | 0.1, 0.5, 1.0 |
|  | Speciation completion rate | 0.1, 0.3, 1.0, |
|  | Extinction rate | 0.0, 0.1, 0.2, 0.4 |
|  | Crown age | 15 |
|  | Standard deviation around crown age | 0.001 |
|  | Sampling method | MRCA, MDCA or random |
|  | Mutation rate |  |
|  | DNA alignment length |  |
|  | MCMC sampling interval | 1K or more |
|  | RNG seed incipient tree | 1 to 20K |
|  | RNG seed alignment simulation |  |
|  | RNG seed BEAST2 |  |

Specification of the data sets. Each row will contain one experiment, where the columns contain parameters, measurements and diagnostics. This table displays the content of the columns. denotes the number of columns a certain item will occupy, resulting in a table of 1023 columns and 20K rows.

|  |  |
| --- | --- |
|  | Description |
|  | simulation parameters, see table [table:parameters] |
|  | nLTT statistic values |
|  | ESSes of all parameters estimated by BEAST2 (see specs below) |

Overview of the 11 BEAST2 estimated parameters

|  |  |
| --- | --- |
| # | Description |
| 1 | posterior |
| 2 | likelihood |
| 3 | prior |
| 4 | treeLikelihood |
| 5 | TreeHeight |
| 6 | BirthDeath |
| 7 | BDBirthRate |
| 8 | BDDeathRate |
| 9 | logP.mrca |
| 10 | mrcatime |
| 11 | clockRate |