

1 The error in Bayesian phylogenetic reconstruction
2 when speciation is not instantaneous

3 Richèl J.C. Bilderbeek¹ and Rampal S. Etienne¹

4 ¹Groningen Institute for Evolutionary Life Sciences, University of
5 Groningen, Groningen, The Netherlands

6 April 30, 2018

7 **Abstract**

8 The tools for reconstructing phylogenetic relationships between taxo-
9 nomic units (e.g. species) have become very advanced in the last three
10 decades.

11 Among the most popular tools are Bayesian approaches, such as BEAST,
12 MrBayes and RevBayes, that use efficient tree sampling routines to create
13 a posterior probability distribution of the phylogenetic tree. A feature of
14 these approaches is the possibility to incorporate known or hypothesized
15 structure of the phylogenetic tree through the tree prior. It has been
16 shown that the effect of the prior on the posterior distribution of trees
17 can be substantial.

18 Currently implemented tree priors assume that speciation is instantane-
19 ous, where we know that speciation can be a gradual process.

20 Here we explore the effects of ignoring the protractedness of the spe-
21 ciation process with an extensive simulation study.

22 We compare the inferred tree to the simulated tree, and find that

Keywords: computational biology, evolution, phylogenetics, prior choice

1 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 (Felsenstein 1981), and the first Bayesian tools emerged in 1996 (Rannala & Yang 1996), providing unprecedented flexibility in the setup of a phylogenetic model.

Currently, the most popular Bayesian phylogenetics tools are BEAST (Drummond & Rambaut 2007) and its successor BEAST2 (Bouckaert *et al.* 2014), MrBayes (Huelsenbeck & Ronquist 2001) and RevBayes (Höhna *et al.* 2016). They allow to 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 a 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.

These model priors in Bayesian phylogenetic reconstruction can be grouped in three groups: (1) site model, specifying nucleotide substitutions, (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 choice of a wrong site model (Posada & Buckley 2004), clock model (Baele *et al.* 2012) or tree prior (Möller *et al.* 2018; Yang & Rannala 2005) is known to affect the posterior.

Current phylogenetic tools use tree priors that assume speciation is instantaneous, whilst we know that, speciation 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 PBD model, is unavailable in current Bayesian phylogenetic tools. Whilst an approximate formula for this probability has been derived (Lambert *et al.* 2015) 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 2009) 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.

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

The PBD model has five biological parameters (see 2), which we explore in a factorial fashion, excluding some combinations. We only simulate a PBD process for phylogenies in which speciation initiation exceeds extinction rate ($b_i > \mu_i$ and $b_g > \mu_g$), and in which the expected number of taxa is less than 1000. As the analytic solution for the expected number of lineages is too complex to be derived (Etienne & Rosindell 2012), we use a numerical approximation instead

NOTE: maybe Lambert et al., 2014, but the (closed access) article is unavailable from home.

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 b , we use 0.1, 0.5 and 1.0 speciation initiation events per (good species)

lineage per time unit. The speciation completion rates used are 0.1, 0.3, 1.0 and 10^9 speciation completion events per (incipient species) lineage per time unit. We used $10^9 \approx \infty$ to mimic the BD model, because the PBD model reduces to the BD model for $\lambda = \infty$. This allows us to measure the baseline error, which is the difference between inferred tree and true species tree that arises purely due to noise because the generating model and the model used in inference are identical. The extinction rates used are 0.0, 0.1, 0.2 and 0.4 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 (Etienne 2015) in the R programming language (R Core Team 2013), all with a crown age of 15 million years. Each protracted birth-death tree uses a different random number generator seed, which makes all runs independent, resulting in a balanced data set.

From each incipient species tree, we construct a species tree, by sampling one incipient/good species per good species. For example, when an incipient species branched off from its mother lineage, both of these subspecies 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 scenario's, in which we pick the representative randomly or in such a way that this results in either the shortest or longest branch lengths. See the supplementary information for a visualization of these sampling methods.

Based on the sampled species tree, we simulate a DNA alignment that has the same history as this species tree, using the phangorn package (Schliep 2011). We assume that the nucleotides of the DNA alignment follow a Jukes-Cantor (Cantor & Jukes 1969) 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

128 could explore other substitution models in the simulations and in the Bayesian
 129 inference, but we chose this simple model because we are primarily interested
 130 in the effect of the choice of tree prior. If anything, our results are conservative:
 131 with a more complex substitution model, there will be more noise and hence our
 132 inference error will increase. We set the mutation rate in such a way to maximize
 133 the information contained in the alignment. To do so, we set the mutation rate
 134 such that we expect on average one (possibly silent) mutation per nucleotide
 135 between crown age and present, which equates to $\frac{1}{15}$ mutations per million years.
 136 The DNA sequence length is chosen to provide a resolution of 10^3 years, that is,
 137 to have one expected nucleotide change per 10^3 years per lineage on average. As
 138 one nucleotide is expected to have on average one (possibly silent) mutation per
 139 15 million years, $15 \cdot 10^3$ nucleotides result in 1 mutation per alignment per 10^3
 140 years (which is coincidentally the same as Möller *et al.* 2018). The simulation
 141 of these DNA alignments follows a strict clock model, which we will specify as
 142 the known clock model prior in the Bayesian inference.

143 From an alignment, we run a Bayesian analysis and create a posterior dis-
 144 tribution of trees and parameters using the *babette* (?) package that sets the
 145 input parameters similar to BEAUti 2 and then runs BEAST2. For our site
 146 model, we assume a Jukes-Cantor nucleotide substitution model, as used in the
 147 simulation of the alignment. For our clock model, we assume a strict clock with
 148 the same fixed rate as used in the simulation of the alignment [**NOTE: Möller**
 149 ***et al.* 2018 did not use a fixed clock rate, I do not see why**]. The tree
 150 prior assumed in inference is the BD model, because studying the effect of this
 151 assumption is the goal of this study. We assumed an MRCA prior with a tight
 152 normal distribution around the crown age, by choosing the crown age as mean,
 153 and a standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown
 154 ages used have the same resolution (of 10^{-3} time units) as the alignment. We

155 ran the MCMC chain to generate 1111 states, of which we removed the first 10%
 156 (also called the 'burn-in'). Of the remaining 1000 MCMC states, the effective
 157 sample size (ESS) of the posterior must at least be 200 for a strong enough
 158 inference (Drummond & Bouckaert 2015). An ESS can be increased by increas-
 159 ing the number of samples or decreasing the autocorrelation between samples.
 160 Would the ESS be less than 200, we decrease autocorrelation by doubling the
 161 MCMC sampling interval of that simulation, until the ESS exceeds 200.

162 We compared each posterior phylogeny to the (sampled) species tree by the
 163 nLTT statistic (Janzen *et al.* 2015), using the nLTT package (Janzen 2015). The
 164 nLTT statistic is equal to the area between the normalized lineages-through-
 165 time-plots of two phylogenies, which has a range from zero (for identical phy-
 166 logenies) to one. We use inference error and nLTT statistic interchangeably.
 167 Comparing the one (sampled) species tree with each of the posterior species
 168 trees yields a distribution of nLTT statistics.

169 We produce two data sets as a comma-seperated file. We set the number
 170 of replicates for each parameter combination such, that this file and a possible
 171 copy can be handled in R's memory. Each row will then contain a parameter set
 172 and the generated nLTT statistics (see 3 for the exact data specification). The
 173 abovementioned memory constraints allows for $2 \cdot 10^3$ rows. With 48 **[NOTE:**
 174 **recalculate]** combinations of biological parameter, there will be 168 **[NOTE:**
 175 **recalculate]** replicates per parameter set.

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

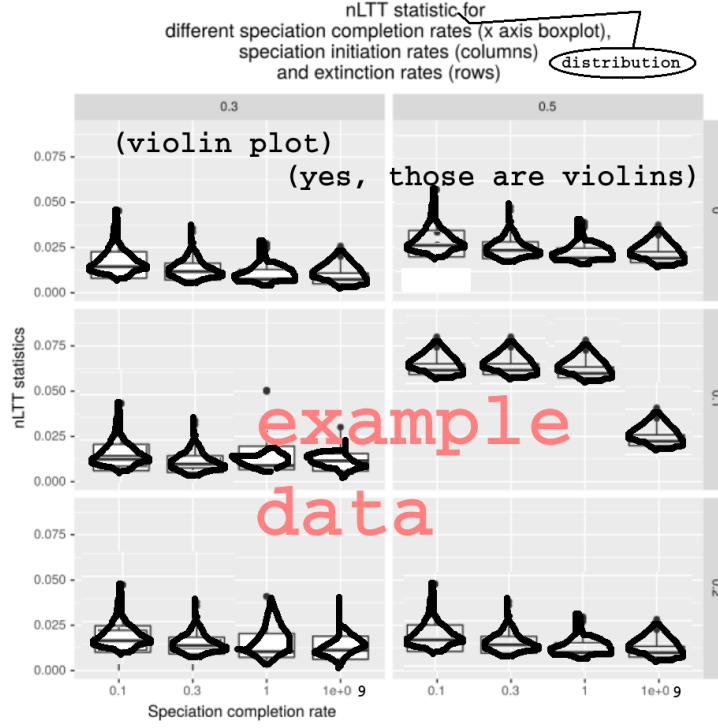


Figure 1: nLTT statistic distribution per biological parameter set, using the balanced data set

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 speciation event creating an incipient species
Speciation completion	The end of a speciation event, in which an incipient species is recognized as a good species

Table 1: Glossary [NOTE: this is requested by the journal]

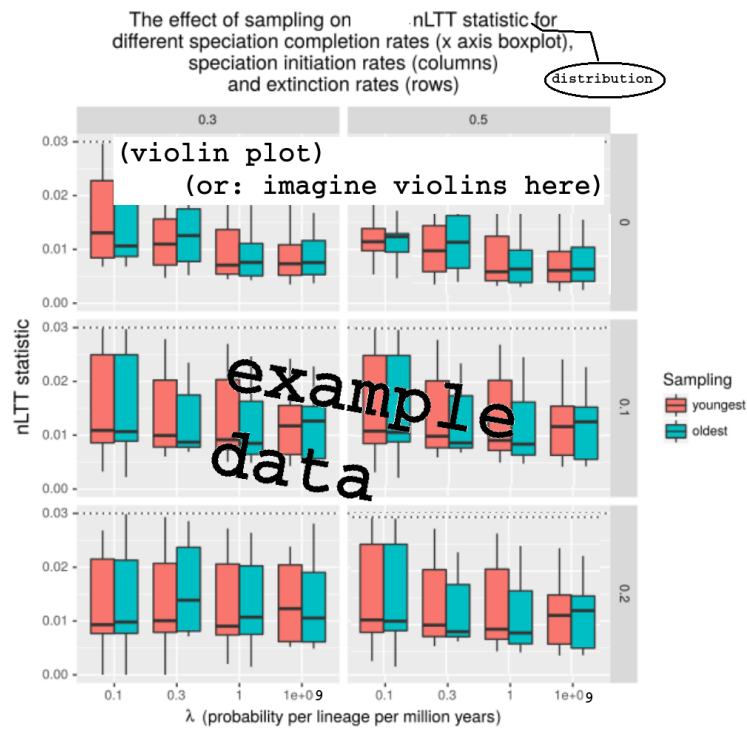


Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect

178 3 Results

179 4 Glossary

180 5 Acknowledgements

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

185 6 Authors' contributions

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

189 References

- 190 Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Ac-
191 curate model selection of relaxed molecular clocks in bayesian phylogenetics.
192 *Molecular biology and evolution*, **30**, 239–243.
- 193 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard,
194 M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform
195 for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.
- 196 Cantor, J. & Jukes, T. (1969) Mammalian protein metabolism. *Evolution of*
197 *protein molecules Academic Press, New York, NY*, pp. 21–132.

198 Drummond, A.J. & Bouckaert, R.R. (2015) *Bayesian evolutionary analysis with*
199 *BEAST*. Cambridge University Press.

200 Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis
201 by sampling trees. *BMC evolutionary biology*, **7**, 214.

202 Etienne, R.S. (2015) *PBD: Protracted Birth-Death Model of Diversification*. R
203 package version 1.1.

204 Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of
205 speciation from phylogenies. *Evolution*, **68**, 2430–2440.

206 Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of
207 the present: protracted speciation can explain observed slowdowns in diver-
208 sification. *Systematic Biology*, **61**, 204–213.

209 Felsenstein, J. (1981) Evolutionary trees from dna sequences: a maximum like-
210 lihood approach. *Journal of molecular evolution*, **17**, 368–376.

211 Heled, J. & Drummond, A.J. (2009) Bayesian inference of species trees from
212 multilocus data. *Molecular biology and evolution*, **27**, 570–580.

213 Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R.,
214 Huelsenbeck, J.P. & Ronquist, F. (2016) Revbayes: Bayesian phylogenetic
215 inference using graphical models and an interactive model-specification lan-
216 guage. *Systematic biology*, **65**, 726–736.

217 Huelsenbeck, J.P. & Ronquist, F. (2001) Mrbayes: Bayesian inference of phylo-
218 genetic trees. *Bioinformatics*, **17**, 754–755.

219 Janzen, T. (2015) *nLTT: Calculate the NLTT Statistic*. R package version 1.1.

220 Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian compu-
221 tation of diversification rates from molecular phylogenies: introducing a new

222 efficient summary statistic, the nlrt. *Methods in Ecology and Evolution*, **6**,
223 566–575.

224 Lambert, A., Morlon, H. & Etienne, R.S. (2015) The reconstructed tree in
225 the lineage-based model of protracted speciation. *Journal of mathematical*
226 *biology*, **70**, 367–397.

227 Möller, S., du Plessis, L. & Stadler, T. (2018) Impact of the tree prior on
228 estimating clock rates during epidemic outbreaks. *Proceedings of the National*
229 *Academy of Sciences*, p. 201713314.

230 Posada, D. & Buckley, T.R. (2004) Model selection and model averaging in
231 phylogenetics: advantages of akaike information criterion and bayesian ap-
232 proaches over likelihood ratio tests. *Systematic biology*, **53**, 793–808.

233 R Core Team (2013) *R: A Language and Environment for Statistical Computing*.
234 R Foundation for Statistical Computing, Vienna, Austria.

235 Rannala, B. & Yang, Z. (1996) Probability distribution of molecular evolution-
236 ary trees: a new method of phylogenetic inference. *Journal of molecular*
237 *evolution*, **43**, 304–311.

238 Schliep, K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**,
239 592–593.

240 Simonet, C., Scherrer, R., Rego-Costa, A. & Etienne, R. (2018) Robustness of
241 the approximate likelihood of the protracted speciation model. *Journal of*
242 *evolutionary biology*, **31**, 469–479.

243 Yang, Z. & Rannala, B. (2005) Branch-length prior influences bayesian posterior
244 probability of phylogeny. *Systematic Biology*, **54**, 455–470.

Parameter	Description	Values
b_g	Speciation initiation rate of a good species	0.1, 0.5, 1.0
b_i	Speciation initiation rate of an incipient species	0.1, 0.5, 1.0
λ	Speciation completion rate	0.1, 0.3, 1.0, ∞
μ_g	Extinction rate of a good species	0.0, 0.1, 0.2, 0.4
μ_i	Extinction rate of an incipient species	0.0, 0.1, 0.2, 0.4
t_c	Crown age	15
σ_c	Standard deviation around crown age	0.001
M	Sampling method	'shortest', 'longest' or random
r	Mutation rate	$\frac{1}{15}$
l_a	DNA alignment length	15K
f_i	MCMC sampling interval	1K or more
R_i	RNG seed incipient tree	1 to 20K
R_a	RNG seed alignment simulation	R_i
R_b	RNG seed BEAST2	R_i

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

n	Description
12	simulation parameters, see table 2
1000	nLTT statistic values
11	ESSes of all parameters estimated by BEAST2 (see specs below)

Table 3: 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. n denotes the number of columns a certain item will occupy, resulting in a table of 1023 columns and 20K rows.

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

Table 4: Overview of the 11 BEAST2 estimated parameters