The error when inferring phylogenies with incipient species by a birth-death model

Richèl J.C. Bilderbeek[1] and Rampal S. Etienne[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

April 26, 2018

Abstract

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

1 Introduction

The computational tools a contemporary phylogeneticist has at his/her disposal goes beyond the wildest imagination of those living three decades ago. Advances in computational power allowed the first cladograms to be inferred from DNA alignments in 1981 (Felsenstein 1981), where the first Bayesian tools emerged in 1996 (Rannala & Yang 1996), the latter providing for 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 incorporate known or hypothesized structure of a phylogenetic tree-to-be-inferred through model priors. From these priors and an alignment of DNA, RNA or protein sequences, a posterior is created. A posterior is a collection of phylogenies and parameter estimates (of the model prior), in which likelier combinations are present more often. Each of these tools use efficient tree sampling routines to create an informative posterior fast.

In a Bayesian analysis, the model priors are explicitly specified. Those model priors can be grouped in three groups: (1) site model, which governs the nucleotide subsititution model, (2) clock model, specifying the rate of mutation per lineage in time, and (3) tree prior, embodying the speciation model behind 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 (Möller *et al.* 2018).

The contemporary phylogenetic tools provide only for tree priors that assume speciation is instantaneous, where we know that, in animals, speciation is a gradual process. When big populations sizes can be assumed big (thus the effect of sampling to be small), the (constant-rate) birth-death (BD) model is a commonly used tree prior, which ignores the 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.

The effect of using the (incorrect) BD tree prior for a PBD process is unknown. A potential problem in species conservation is that the number of species is underestimated (see Fennessy *et al.* 2016 for a clear example). Additionally, protracted speciation may be one explanation in the observed decline of speciation rates in time (Etienne & Rosindell 2012). Also, a BD model places the most recent common ancestor (MRCA) of a young species duo closer to the present, as the BD model allows for a speciation event being recognized immediately, where the PBD model accounts for speciation needing time.

There are multiple possiblities why the PBD model is relatively unexplored. Biologically, the PBD model is predicted to have an effect strongest in the present (as earlier speciation events are nearly always recognized), so in research that investigates (mostly) older species, a BD model would suffice. Computationally, the BD model is simpler, thus more light-weight, model. Methodological, there is no computational tool where the PBD model fits in: every contemporary framework assumes either an analysis at the species or subspecies level. In the PBD model, incipient species are the cause there is no such thing as a 'true' species tree, as incipient species may give rise to paraphylies.

This research's goal is to explore the effect of using an overly simplistic BD prior on PBD simulated phylogenies. We provide a data set, that quantifies the inference error made in general, and explores the effect of the way species trees are sampled from an incipient species tree. In brief, we simulate protracted phylogenies using the PBD process, from which we sample a species tree. From that species tree, we simulate a DNA sequence alignment. Then, BEAST2 uses these alignments to infer a posterior of phylogenies, using a BD prior. The difference between the (BD) posterior phylogenies and (PBD) species tree is quantified.

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 assume that the speciation initiation and extinction rates of an incipient and good species are equal ($b = b_i = b_g$ and $\mu = \mu_i = \mu_g$), as this enables use to do more replicates **[NOTE: I am unconvinced. I think we should also explore $b_i \neq b_g$ and $\mu_i \neq \mu_g$]**. We only simulate a PBD process for phylogenies in which speciation initiation exceeds extinction rate ($b > \mu$), and in which their difference is not too big ($b - \mu < 0.8$), to prevent overly taxon-poor and taxon-rich phylogenies respectively. The parameter values chosen are a superset of Etienne *et al.*

2014, as these parameters result in reasonably sized phylogenies and allows us to compare results. For the speciation initiation rate *b*, we'll 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. For $\lambda = \infty$ (where we assume that in this context $10^9 \approx \infty$), the PBD model equals a BD model, which allows us to measure the baseline error. 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, a protracted birth-death tree is simulated, using the PBD package (Etienne 2015) in the R programming language (R Core Team 2013), with the same crown age as Etienne *et al.* 2014 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 an incipient species tree, we sample a species tree. To do so, from each species a sub-species is chosen to represent the good species as a whole. To clarify, it may be that an incipient species branched of from its mother lineage. Both of these subspecies are recognized as the good species of the mother lineage, and both can be picked as an (equally good) representative of the good mother species. In this research, we use three sampling scenario's, in which we pick the most recent common ancestor (MRCA), most distant common ancestor (MDCA) or random subspecies. The scenario in which sampling has an effect on the branch length distributions of the species tree, is when a species in the proces of speciation, gives rise to a new incipient lineage that finishes speciation before the ancestral completes speciation itself.

From a species tree, we simulate a DNA alignment that has the same history as the phylogeny, using the phangorn package (Schliep 2011). 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. Although this may seem as a simplification, in our Bayesian inference (see below) we use this exact site model as the (obviously correct) site model prior. The mutation rate is set in such a way to maximize chronologic information. To do so, the mutation rate is set to expect on average one (possibly silent) mutation per nucleotide between crown age and present, which equates to $\frac{1}{15}$ mutations per million year. The DNA sequence length is chosen to provide a resolution of 10^3 years, that is, to have one expected nucleotide change per 10^3 years per lineage on average. As one nucleotide is expected to have on average one (possibly silent) mutation per 15 million years, $15 \cdot 10^3$ nucleotides results in 1 mutation per alignment per 10^3 years (which is coincidentally the same as Möller *et al.* 2018). The simulation of these DNA aligment 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, using the phylogetic tool BEAST2 Bouckaert *et al.* 2014 using the pirouette (Bilderbeek 2018) package. For our site model, we assume a Jukes-Cantor nucleotide substitution model, as we used that 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: Möller *et al.* 2018 did not use a fixed clock rate, I do not see why**]. The tree prior assumed is the BD model, as this simplification is the goal of this research. Additionally, we assume a MRCA prior with a normal distribution with a mean of the crown age, and a standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown ages used have the same resolution (of 10^{-3} time units) as the alignment. The MCMC chain is run to generate 1111 states, of which the first 10% (also called the 'burn-in') is removed. 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 (Drummond & Bouckaert 2015). 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.

Each posterior's phylogeny is compared to the (sampled) species tree by the nLTT statistic (Janzen *et al.* 2015), using the nLTT package (Janzen 2015). The nLTT statistic equates 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 synonymously. Comparing the one (sampled) species tree with each of the posterior's species trees, a distribution nLTT statistics is created.

Two data sets are produced by this research. The first data set is a general balanced 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 MRCA and MDCA subspecies. We predict that these two most extreme sampling methods result in the most pronounced differences.

Each data set is stored as a comma-seperated file. As a theoretical study such as this could theoretically (pun intended) produce an infinitely big data set, we placed an upper limit for this data set's size. This size is chosen as such the the R programming language (R Core Team 2013) can contain a data file in memory twice. Each row will contain a parameter set and the generated nLTT

Term	Definition
Phylogenetics	The inference of evolutionary relationships of groups of organisms using genetics
Model prior	Knowledge or assumptions about the onotogeny of evolutionary histories
Posterior	A collection of phylogenies and parameter estimates, in which likelier combinations are present more
Protracted speciation	The process in which speciation takes two events to complete: a speciation initiation and a speciation completion event
Speciation initiation	The start of a speciatiation event, in which initially the new species-to-be is not recognized as such
Speciation completion	The end of a speciatiation event, in which the new species is recognized as such

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

statistics (see 3 for the exact data specification). From these constraints, this allows for $2 \cdot 10^3$ rows. As there are 48 combinations of biological parameter **[NOTE: calculate exact number when Rampal decides for $b_i = b_g$ only or for also allowing $b_i \neq b_g$**], there will be 168 [NOTE: recalculate, as before] replicates per biological parameter set.

Our results show the general effect of the biological parameters (b, λ, μ) using the balanced data set, and the effect of sampling using the data set conditioned on sampling having an effect. In both cases, the nLTT statistics distribution is plotted per biological parameter set using a violin plot, as such a plot maintains information about distribution. We predict that nLTT statistic values increase with an increasing protractednedness (that is, a low speciation completion rate), but we cannot predict the the extent of this error, as it has never been measured.

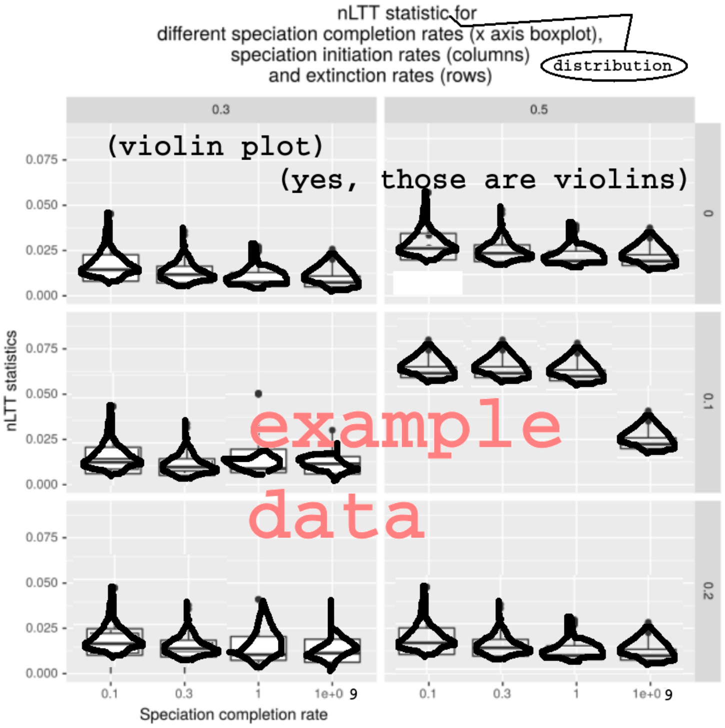

Figure 1: nLTT statistic distribution per biological parameter set, using the balanced data set

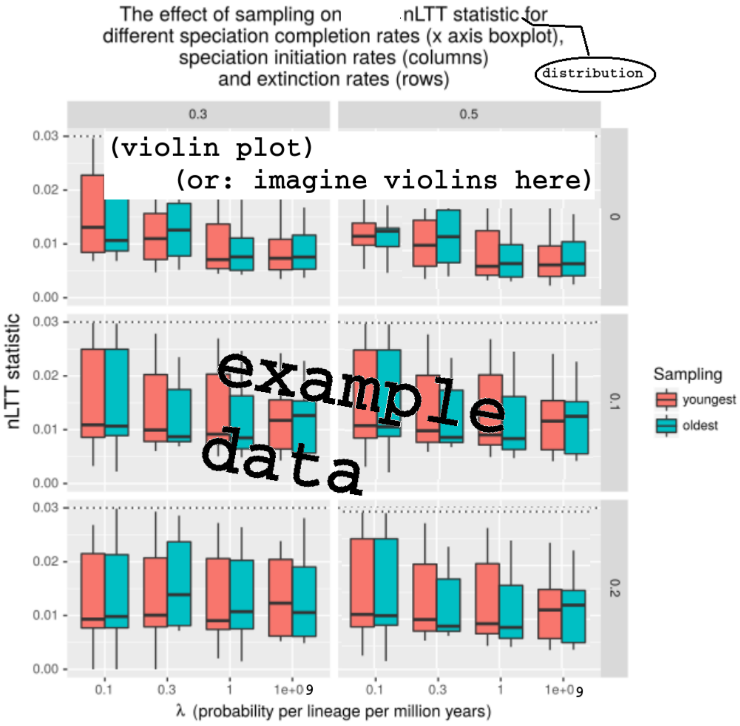

Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect

3 Results

4 Glossary

5 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.

6 Authors' contributions

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

References

Bilderbeek, R.J. (2018) *pirouette: create a posterior from a phylogeny.*

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.

Cantor, J. & Jukes, T. (1969) Mammalian protein metabolism. *Evolution of protein molecules Academic Press, New York, NY*, pp. 21–132.

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

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

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

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

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

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

Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M.A., Vamberger, M., Fritz, U. & Janke, A. (2016) Multi-locus analyses reveal four giraffe species instead of one. *Current Biology*.

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

Huelsenbeck, J.P. & Ronquist, F. (2001) Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

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

Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, **6**, 566–575.

Parameter	Description	Values
$b = b_g = b_i$	Speciation initiation rate	0.1, 0.5, 1.0
λ	Speciation completion rate	0.1, 0.3, 1.0, ∞
$\mu = \mu_g = \mu_i$	Extinction rate	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	MRCA, MDCA 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.

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

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

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

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

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	mrcatime
11	clockRate

Table 4: Overview of the 11 BEAST2 estimated parameters