The error in Bayesian phylogenetic reconstruction when speciation co-occurs

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

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

September 25, 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 events are independent, where we know that speciation can coincide, for example, when trigger by a larger geographic change.

Here we explore the effects of ignoring speciation co-occurence with an extensive simulation study.

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

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior

1 Introduction

The computational tools that are currently available to the phylogeneticists 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 offshoot 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. With these priors and an alignment of DNA, RNA or protein sequences, they create a sample of the 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.

The model priors in Bayesian phylogenetic reconstruction can be grouped into three categories: (1) site model, specifying nucleotide substitutions, (2) clock model, specifying the rate of mutation per lineage in time, and (3) tree model, constituting the speciation model underlying branching events (speciation) and branch termination (extinction). The choice of 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 events are independent, whilst we know multiple examples of adaptive radiations in multiple species families can co-occur, for example **[RJCB: @gio: please add examples]**.

The (constant-rate) birth-death (BD) model is a commonly used tree prior, but it ignores the co-occurence of speciation. It even assume that two speciation events at exactly the same time has zero likelihood! The multiple birth-death (MBD) model, an extension of the BD model, does incorporate the idea that speciation can co-occur.

[RJCB: explain model here, example is below in the comments]

Unfortunately, a tree prior according to this model, providing the probability of a species tree under the MBD model, is unavailable in current Bayesian phylogenetic tools. Whilst a likelihood equation has been derived (**[RJCB: cite yourself here]**), it has not been implemented as tree prior yet. There are various reasons for this. First, the computation of the MBD likelihood 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.

Here we aim to explore the effect of using the BD prior on MBD simulated phylogenies. In brief, we simulate phylogenies with co-occuring speciation events using the MBD process. 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 (MBD) species tree. Furthermore, while we evidently know the clock and site models used in the simulation, using a different clock and/or site model prior in inference may compensate or increase this difference between inferred and simulated tree. To study this, we also explore the effect of a different clock and site model prior in inference.

The MBD model has **[RJCB: @gio: how many?]** parameters, depicted in table 2.

[RJCB: @gio: describe parameter values used here, example is below]

We use **[RJCB: @gio: parameter setting here]** as our control for which the MBD model reduces to the BD model.

We simulate protracted birth-death trees, using the MBD package (Etienne 2015) in the R programming language (R Core Team 2013). The first tree has a random number generator seed of 1, which is incremented by 1 for each simulated tree. For each combination of **[RJCB: @gio: parameter values here]**, we generate incipient species trees with a crown age of 15 million years. Only trees with the desired number of good taxa are kept.

We create one data set to explore parameter space, All the trees with the correct number of good species are kept. Based on the species tree, we simulate a DNA alignment that has the same history as this species tree, using the `phangorn` package (Schliep 2011). We set the nucleotides of the DNA alignment to follow a Jukes-Cantor (Jukes *et al.* 1969) nucleotide substitution model, in which all nucleotide-to-nucleotide transitions are equally likely. The DNA sequence of the root ancestor consists of four equally sized single-nucleotide blocks of adenine, cytosine, guanine and thymine respectively. For example, for a DNA sequence length of 12, this would be AAACCCGGGTTT. The order of nucletides does not matter in this study, because we do not consider several partitions of the sequence with their own parameters. Only the frequency of occurrence matters. In our Bayesian inference (see below) we use the same site model as the (obviously correct) site model prior, but we also explore the effect of assuming a more complex site model prior. We predict with the more complex substitution model, that there will be more noise and hence our inference error will increase. On the other hand, we dare not rule out that the inference error will decrease, due to more flexibility in the more complex prior. We set the mutation rate in such a way to maximize the information contained in the alignment. To do so, we set the mutation rate such that we expect on average one (possibly silent) mutation per nucleotide between crown age and present, which equates to $\frac{1}{15}$ mutations per million years. 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 result in 1 mutation per alignment per 10^3 years (which is coincidentally the same as Möller *et al.* 2018). The simulation of these DNA aligments follows a strict clock model, which we will specify as one of the two clock models assumed in the Bayesian inference (see below).

[RJCB: must rewrite, use pirouette as a starting point] From an alignment, we run a Bayesian analysis and create a posterior distribution of trees and parameters using the `pirouette` (Bilderbeek 2018) package that sets the input parameters similar to BEAUti 2 and then runs BEAST2. For our site model, we assume either a Jukes-Cantor or GTR nucleotide substitution model. The Jukes-Cantor model is the correct one, as it is used for simulating that alignment, where the GTR model is the site model that is picked as a default by most users. For our clock model, we assume either a strict or relaxed log-normal clock model. Also here, the strict clock model is the correct one, as it is used for simulating the alignment, but the relaxed log-normal clock model is the one most commonly used. We set the BD model as a tree prior, as gauging the effect of this incorrect assumption is the goal of this study. We assume an MRCA prior with a tight normal distribution around the crown age, by choosing the crown age as mean, and a standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown ages inferred have the same resolution (of 10^{-3} time units) as the alignment. We ran the MCMC chain to generate 1111 states, of which we remove the first 10% (also called the 'burn-in'). Of the remaining 1000 MCMC states, 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. If the ESS is less than 200, we decrease autocorrelation by doubling the MCMC sampling interval of that simulation, until the ESS exceeds 200.

We compare each posterior phylogeny to the (sampled) species tree using the nLTT statistic (Janzen *et al.* 2015), from the `nLTT` package (Janzen 2015). The nLTT statistic equals 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 simulated species tree with each of the posterior species trees yields a distribution of nLTT statistics.

The input trees generated with a **[RJCB: @gio: parameter that is set to reduce the MBD model to BD]** allow us to measure the noise of the experiment.

We produce one data set as a comma-separated file. The general data set has ?144 **[RJCB: recalc]** different combinations of biological parameter com-

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

Table 1: Glossary

binations, site and clock models. The data set to investigate sampling has ?552 **[RJCB: recalc]** different combinations of biological parameter combinations, site models, clock models and sampling methods. The experiment is computationally intensive: pilot experiments show that the experiment takes roughly 100 days of CPU time and 20 days of wall clock time (which includes the queued waiting for computational resources) per replicate. Due to this, we choose to perform ten replicates, so that the complete experiment will take an acceptable time of roughly seven months.

For both data sets, we display the nLTT statistics distribution per biological parameter combination as a violin plot. We only show the nLTT distributions that were generated under the (correct) assumptions of a Jukes-Cantor site model and a strict clock model, separated per sampling method used. We display the nLTT statistic distributions separated per site or clock model in the supplementary information.

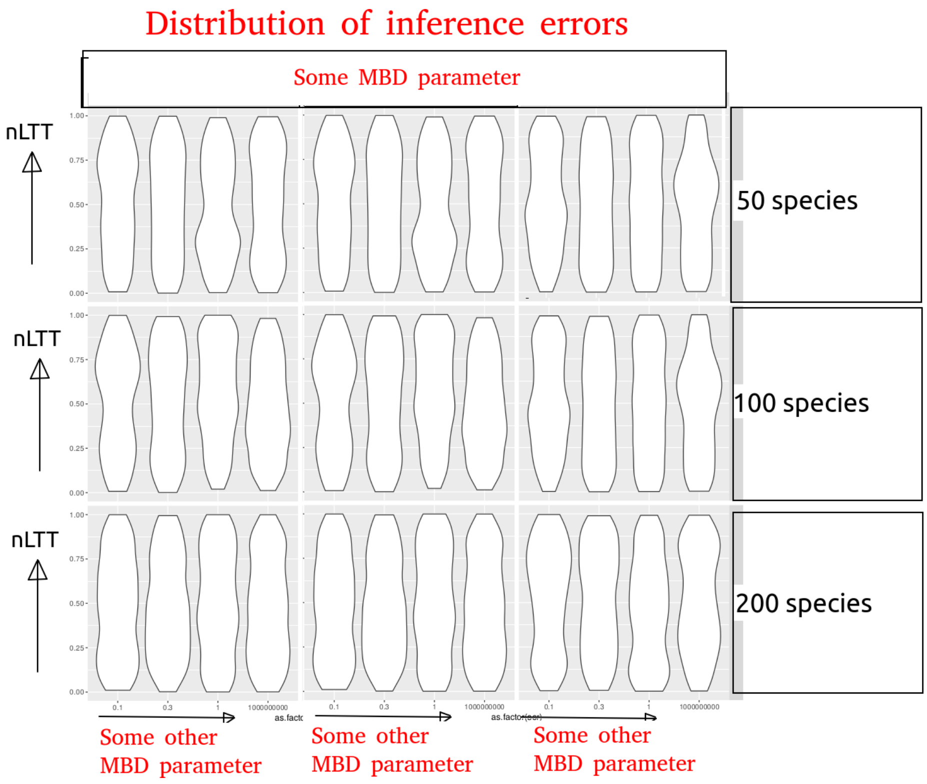

Figure 1: nLTT statistic distribution per biological parameter set, using the general data set, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

2 Results

3 Glossary

References

Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular biology and evolution*, **30**, 239–243.

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.

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.

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

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.

Jukes, T.H., Cantor, C.R. *et al.* (1969) Evolution of protein molecules. *Mammalian protein metabolism*, **3**, 132.

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.

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

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.

Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, **17**, 57–86.

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

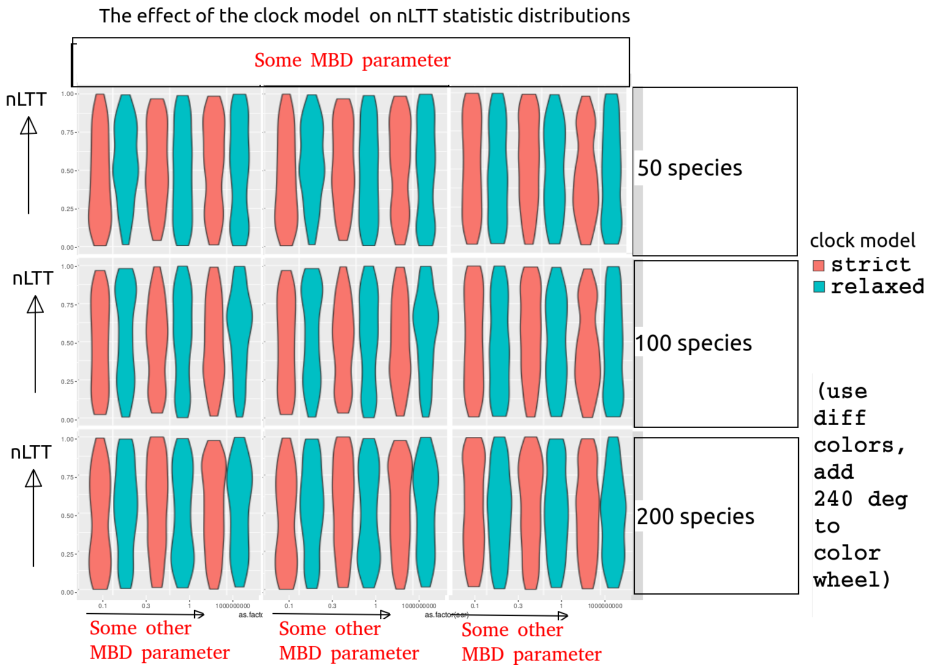

Figure 2: nLTT statistic distribution per biological parameter set per clock model, using the general data set, under the (correct) assumption of a Jukes-Cantor site model.

A Acknowledgements

[RJCB: put this section here, as the journal does not request for this]

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.

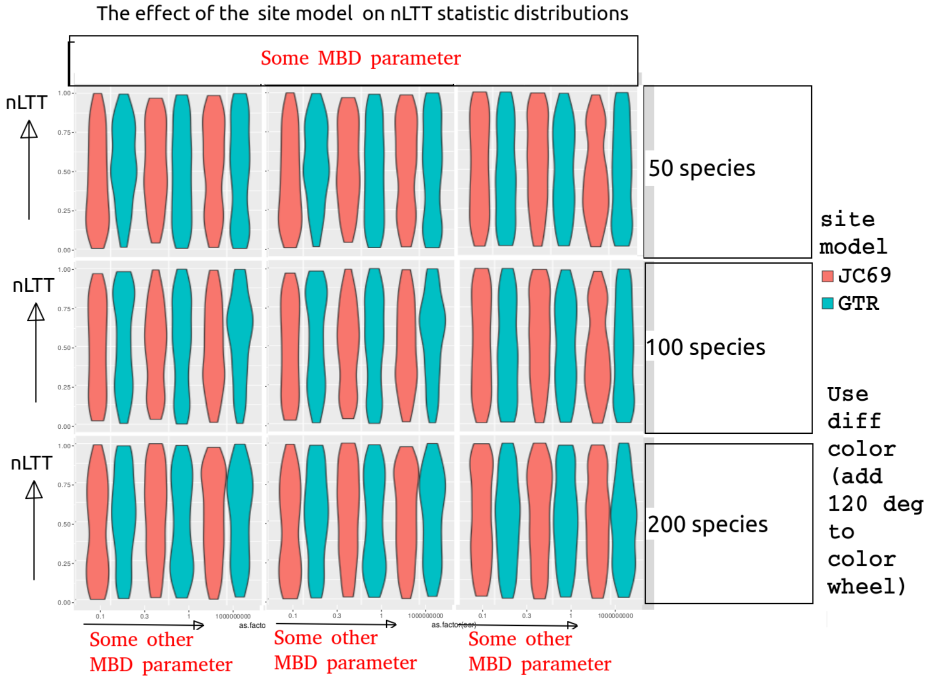

Figure 3: nLTT statistic distribution per biological parameter set per site model, using the general data set, under the (correct) assumption of a strict clock model.

	Description	Values
λ	**[RJCB: @gio: MBD params here]**	0.1, 0.3, 1.0, 10^9
λ	Speciation rate	0.1, 0.3, 1.0, 10^9
μ	Extinction rate	0.0, 0.1, 0.2
n	Number of good taxa	50, 100, 200
t_c	Crown age	15
σ_c	Standard deviation around crown age	0.001
M_s	Sampling method	S, L, R
M_c	Clock model	S, RLN
M_t	Site model	JC69, GTR
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 and randomly sampled species tree	1, 2, etc.
R_a	RNG seed alignment simulation	R_i
R_b	RNG seed BEAST2	R_i

Table 2: Overview of the simulation parameters. Above the horizontal line is the biological parameter set. The RNG seed R_i is 1 for the first simulation, 2 for the next, and so on. The clock models are abbreviated as 'S' for a strict and 'RLN' for a relaxed log-normal model. The site models are abbreviated as 'JC69' for Jukes-Cantor (Jukes *et al.* 1969) and 'GTR' for the generalized time-reversible model (Tavaré 1986).

n	Description
12 **[RJCB: recalc]**	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 **[RJCB: recalc]** 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 parameters estimated by BEAST2

B Authors' contributions

[RJCB: put this section here, as the journal does not request for this]

RSE **[RJCB: @gio: I assume this this is true?]** conceived the idea for this experiment. GL created and tested the MBD package. RJCB created and tested the experiment. GL and RJCB wrote the first draft of the manuscript. RSE contributed substantially to revisions.