

1 The error in Bayesian phylogenetic reconstruction
2 when speciation co-occurs

3 Giovanni Laudanno¹, Richèl J.C. Bilderbeek¹, and Rampal S.
4 Etienne¹

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

7 July 1, 2019

8 **Abstract**

9 There exist millions of species on Earth, all originating from a common
10 ancestor billions of years ago. The field of phylogenetics uses heritable
11 material (e.g. DNA) to determine the evolutionary history of species.

12 Starting from heritable material and explicit assumptions, Bayesian
13 phylogenetics allows to infer a jointly-estimated phylogeny and parameter
14 estimates distribution. One of these assumptions is the speciation model,
15 which mathematically describes the branching process of a phylogeny in
16 time. Speciation models commonly assume that speciation events are
17 independent and disjointed. Yet, such assumption may overlook the com-
18 plexity of certain processes, where environmental changes promote spe-
19 ciation events in multiple lineages. This new layer of complexity can be
20 captured developing a novel ad hoc speciation model. However, the in-
21 troduction of a new model could be not necessary if current models are

capable of describing the process, under simpler assumptions. Here we investigate the extent of the discrepancy produced by current BEAST2's tree priors, when trying to infer back trees where speciation can co-occur.

We let the impact of co-occurring speciation vary on datasets of simulated trees and show the corresponding error produced during the inference process.

We show that the extent of the inference error grows with the amount of co-occurring speciation events to establish the limits of BEAST2's standard tree priors.

These results allow phylogeneticists to judge under which circumstances the commonly used speciation model can be safely used.

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior, pirouette, BEAST2, babette

1 Introduction

[GL: Introduce "BD" and "MBD" acronyms at some point.]

Modern computational techniques allow to infer phylogenies from genetic data such as DNA, RNA or proteins. BEAST (Drummond & Rambaut 2007) and its descendant BEAST2 (Bouckaert *et al.* 2014) are widely used tools to perform such task. They return posterior distributions of phylogenies and estimated parameters by running a Bayesian analysis, given genetic data and tree priors. A tree prior is a mathematical description of the a priori characteristics that we expect to be reflected in posterior phylogenies. The choice of a specific prior is, by definition, arbitrary but the consequences of such choice can be vet analyzing the so-obtained posteriors.

BEAST2 gives to the user the option to set up several possible phylogenetic priors (e.g. substitution/clock/speciation models). However, currently available

priors might be not suitable to analyze some specific datasets. For this reason BEAST2 provides users with the possibility to introduce new tree priors, to infer phylogenies based on different assumptions on how the speciation process takes place.

Current phylogenetic tools assume that only a single speciation event can occur at any given time. While this assumption has been proved to be useful to construct a wide variety of successful models (e.g Maddison *et al.* 2007, Valente *et al.* 2015, Etienne *et al.* 2012, Etienne *et al.* 2014), they do not take into account the possibility for environmentally-driven contemporary speciations on multiple lineages.

We explore such case introducing the multiple-birth death model (MBD), currently absent in BEAST2. The multiple birth hypothesis aims to describe how large-scale environmental changes (also known as species pumps [GL: add some ref?]) can lead phylogenies to sport an impressive amount of speciation events in relatively short times.

Such hypothesis can be useful to better understand the history of systems of particular interests for evolutionary biologists, such as the diversification of cichlid fish in the African Great Lakes (Malawi, Tanganyika and Victoria), where water level fluctuations are thought to play an important role in promoting diversification (Verheyen *et al.* 1996, Sturmbauer *et al.* 2001, Janzen *et al.* 2016, Janzen *et al.* 2017).

However, the introduction of new tree priors is not always desirable (Bilderbeek *et al.* 2019). Current standard birth death (BD) tree priors might, in principle, prove to be good enough at detecting such events despite their lower level of complexity. If this is the case one should always choose to adopt the simplest model.

We used the R package `pirouette` (Bilderbeek & Laudanno 2019) to perform

75 such test, starting on phylogenies simulated under the MBD regime using the
76 `mbd` package (Laudanno 2018). From such phylogenies we measure the inference
77 error made adopting a standard BD tree prior in the inference process.

78 With this work we aim, using such inference error distributions, to test
79 whether or not it is advantageous to implement a new prior model that can
80 allow the construction of trees where multiple speciations can co-occur at the
81 same time.

82 2 Methods

83 We proceed in the following way: we build simulated datasets generated un-
84 der the multiple birth model. Then we run a `pirouette` analysis, which will
85 lead to error distributions between the inference posterior and the original sim-
86 ulated trees. Importantly, the `pirouette` analysis includes also a 'twin' parallel
87 pipeline, which will provide a measure of the baseline error due to pure stochas-
88 ticity, unavoidably occurring in the process.

89 2.1 Model

90 In the MBD model, parameters λ and μ correspond, respectively, to the common
91 per-species speciation and extinction rates already present in the standard BD
92 model. Additionally, MBD relies on two additional parameters, ν and q . The
93 first, ν , is the rate at which an environmental change is triggered. When such
94 event is triggered, each species present in the phylogeny at that moment has a
95 probability q to speciate at that time. This kind of speciation is of a different
96 nature respect to the one triggered by λ . In fact, whereas parameter λ can
97 be seen as describing a sympatric process, ν induces the rise of geographical
98 barriers interrupting the gene flow [GL: @richel: please check if biology
99 here is accurate] and leading to an allopatric speciation. Even though multiple

100 speciations can co-occur, polytomies are not allowed in such process as each
101 species can speciate only once at the time. A likelihood expression for the
102 process is provided in Laudanno 2018.

103 2.2 Tree simulations

104 We simulate the speciation process in continuous time using the Doob-Gillespie
105 algorithm, using the `mbd_sim` function from the `mbd` package (Laudanno 2018).
106 We let parameters vary using all possible combinations of values as shown in Ta-
107 ble 1. For each parameter setting, we generate 1000 independent reconstructed
108 trees of the same crown age. **[GL: Do we need to add more information**
109 **on the Doob-Gillespie algorithm or it is overkilling?]**

110 2.3 Inference error estimation

111 From each MBD tree, we measure the impact of using the simpler BD tree
112 prior on the inference, using the `pirouette` R package, as described in detail in
113 Bilderbeek & Laudanno 2019.

114 In brief: `pirouette` starts from a 'true' (but unobservable in nature) start-
115 ing phylogeny, from which a DNA sequence alignment (which is observable in
116 nature) is simulated. From each sequence alignment, a Bayesian inference is run,
117 to obtain a posterior distribution of jointly-estimated trees and model parameter
118 estimates. By comparing the true tree and the posterior trees, an inference error
119 distribution is generated. We use the twinning option available in `pirouette`
120 to measure a minimum and full error, in which the minimum error is caused by
121 stochasticity in the full pipeline, where the full error is the added error from
122 using an invalid but standard speciation model **[GL: I am not sure I get**
123 **this. What do you mean?]**.

124 In our context, the alignments are 1000 nucleotides in length, with a known

125 root sequence of four 250 mono-nucleotide blocks, following the simplest nu-
126 cleotide substitution model (JC69) and clock model (strict), with a mutation
127 rate of $\frac{1}{2} \cdot c$, in which c is the crown age. With this mutation rate, each nu-
128 cleotide is expected to mutate (both silently and non-silently) in half of the
129 histories from the root sequence in the past to the sequence in the tips in the
130 present.

131 For the Bayesian inference, we use a JC69 site model, a strict clock model
132 and a BD tree prior. Additionally, we use an MRCA prior equal to the crown
133 age with a normal distribution of width $\sigma = 0.01$. We pick an MCMC setup of
134 10M states, which is sampled each 1k moves.

135 For the error measurement, we use the nLTT statistic Janzen *et al.* 2015 and
136 a burn-in fraction of 10%.

137 For the twinning, we let the twin trees follow a BD model.

138 For the experiments, we use both a hand-picked generative model and a
139 set of candidate models. We set a BD tree prior, JC69 site model and strict
140 clock model as the generative model. We used all other combinations of four
141 tree priors, two clock models and five speciation models, resulting in a set of 39
142 candidate models.

Parameter	Values
λ	(0.2)
μ	(0, 0.15)
ν	(1.0, 1.5, 2.0, 2.5)
q	(0.1, 0.15, 0.2)
crown age	6

Table 1: Parameters used to simulate MBD trees. For each parameter setting 1000 trees are simulated.

3 Results

The inference error made for each of the parameter combinations is shown in Fig. 1. For both extinction rates, we find that the error increases as ν and q increase. Also in line with our predictions, we find no difference between the two extinction rates [GL: I don't actually have predictions on that.].

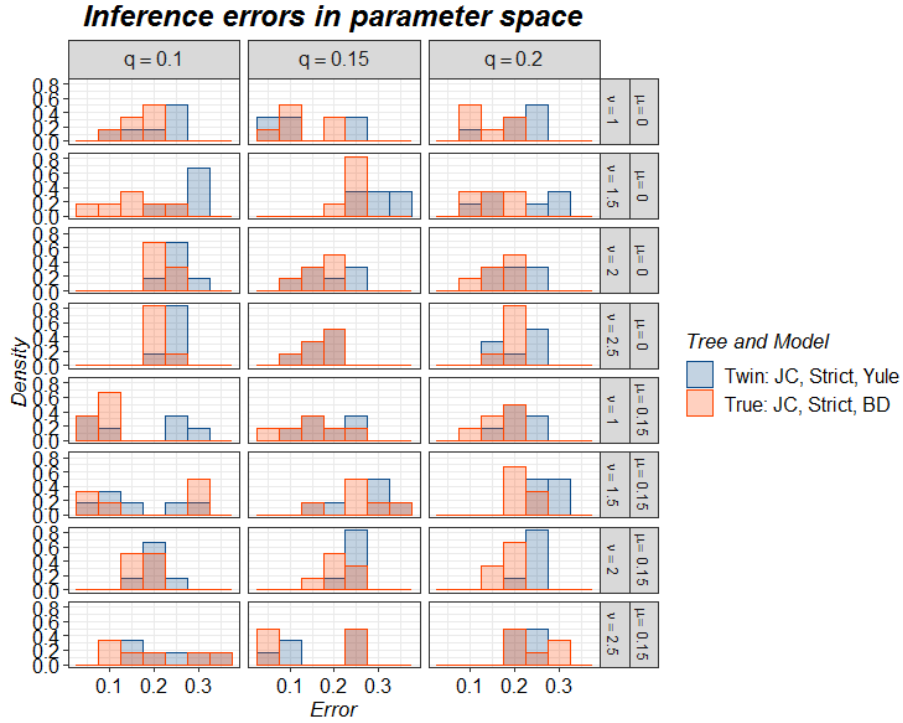


Figure 1: The inference error distribution for the different biological parameter settings. In all cases, $\lambda = 0.2$ and crown age equals 6.

4 Discussion

From the four MBD parameters λ , μ , ν and q , we investigated 1, 2, 4 and 3 different values respectively. We chose to use only one λ , as the proportion of species created in a co-occurrent speciation event is dependent on the ratio

152 between λ and a combination of ν and q .

153 References

- 154 Bilderbeek, R.J. & Laudanno, G. (2019) *pirouette: create a posterior from a*
155 *phylogeny*.
- 156 Bilderbeek, R.J.C., Laudanno, G. & Etienne, R.S. (2019) Quantifying the im-
157 portance of a tree prior in bayesian phylogenetics.
- 158 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard,
159 M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform
160 for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.
- 161 Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis
162 by sampling trees. *BMC evolutionary biology*, **7**, 214.
- 163 Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A.
164 & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies
165 closer to agreement with the fossil record. *Proc R Soc Lond B: Biol Sci*, **279**,
166 1300–1309.
- 167 Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of
168 speciation from phylogenies. *Evolution*, **68**, 2430–2440.
- 169 Janzen, T., Alzate, A., Muschick, M., Maan, M.E., van der Plas, F. & Etienne,
170 R.S. (2017) Community assembly in lake tanganyika cichlid fish: quantifying
171 the contributions of both niche-based and neutral processes. *Ecology and*
172 *Evolution*, **7**, 1057–1067.

173 Janzen, T., Alzate, A., Muschick, M., van der Plas, F. & Etienne, R.S. (2016)
174 Stochastic processes dominate community assembly in cichlid communities in
175 lake tanganyika. *bioRxiv*, p. 039503.

176 Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian compu-
177 tation of diversification rates from molecular phylogenies: introducing a new
178 efficient summary statistic, the nlrt. *Methods in Ecology and Evolution*, **6**,
179 566–575.

180 Laudanno, G. (2018) *MBD: Multiple Birth Death Diversification*. R package
181 version 0.1.

182 Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary char-
183 acter’s effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

184 Sturmbauer, C., Baric, S., Salzburger, W., Rüber, L. & Verheyen, E. (2001) Lake
185 level fluctuations synchronize genetic divergences of cichlid fishes in african
186 lakes. *Molecular Biology and Evolution*, **18**, 144–154.

187 Valente, L.M., Phillimore, A.B. & Etienne, R.S. (2015) Equilibrium and non-
188 equilibrium dynamics simultaneously operate in the galápagos islands. *Ecol-
189 ogy Letters*, **18**, 844–852.

190 Verheyen, E., Rüber, L., Snoeks, J. & Meyer, A. (1996) Mitochondrial phylo-
191 geography of rock-dwelling cichlid fishes reveals evolutionary influence of his-
192 torical lake level fluctuations of lake tanganyika, africa. *Philosophical Trans-
193 actions of the Royal Society of London Series B: Biological Sciences*, **351**,
194 797–805.