

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 June 19, 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 in the speciation model,
15 which mathematically describes the branching process of a phylogeny in
16 time. The most used speciation model assumes that speciation events are
17 independent, where we know that certain events can trigger speciation
18 events in multiple species.

19 This research answers the question what the impact is of using a species
20 tree model that assumes speciation is independent, when it is used on
21 phylogenies created by a tree model in which speciation can co-occur.

Here we show the inference error made, when nature has varying degrees of co-occurring speciation over a wide range of parameter settings.

We show that the inference error correlates with the amount of co-occurring speciation events, which validates

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

In a bigger picture, these results showcase the use of a general and flexible method we used to assess the impact of using an oversimplistic tree prior, helping phylogeneticists to find the line between 'too simple' and 'too complex' speciation models.

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

1 Introduction

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 this task, which they can achieve by running a Bayesian analysis given data and tree priors.

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

48 to construct a wide variety of successful models (e.g Maddison *et al.* 2007,
49 Valente *et al.* 2015, Etienne *et al.* 2012, Etienne *et al.* 2014), they disallow
50 for environmental changes that trigger speciations in multiple clades at a same
51 point in time.

52 We explore such case introducing the multiple birth model, currently absent
53 in BEAST. The multiple birth hypothesis aims to include species pump mech-
54 anisms to investigate drivers and modes of such diversification processes whose
55 phylogenies show an impressive amount of speciation events in relatively short
56 times.

57 The (constant-rate) birth-death (BD) model embodies the common assump-
58 tion that only a single speciation event can occur at any given time. The
59 multiple-birth-death (MBD) model relaxes this assumption allowing, in ad-
60 dition to standard BD events, also events in which large-scale environmental
61 changes lead to speciation bursts. Such hypothesis can be useful to better un-
62 derstand the history of systems of particular interests for evolutionary biologists,
63 such as the diversification of cichlid fish in the African Great Lakes (Malawi,
64 Tanganyika and Victoria), where water level fluctuations are thought to play an
65 important role in promoting diversification (Verheyen *et al.* 1996, Sturmbauer
66 *et al.* 2001, Janzen *et al.* 2016, Janzen *et al.* 2017).

67 However, the introduction of new tree priors is not always desirable (Bilder-
68 beek *et al.* 2019). Current BD tree priors might, in principle, prove to be good
69 enough at detecting such events despite the lower level of complexity. If this is
70 the case one should always be more keen to adopt the simplest model.

71 We used the R package `pirouette` (Bilderbeek & Laudanno 2019) to perform
72 such test, starting on phylogenies simulated under the MBD regime using the
73 `mbd` package (Laudanno 2018). From such phylogenies we measure the inference
74 error made adopting a standard BD tree prior in the inference process.

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

79 **2 Methods**

80 **2.1 Model**

81 In the MBD model, parameters λ and μ correspond, respectively, to the com-
82 mon per-species speciation and extinction rates present also in the standard BD
83 model. Additionally, MBD relies on two additional parameters. Parameter ν
84 is the rate at which an environmental change is triggered. When such event is
85 triggered, each species present in the phylogeny at that moment has a proba-
86 bility q to speciate at that time. This kind of speciation is of a different nature
87 respect to the one triggered by λ . In fact, whereas parameter λ can be seen
88 as describing a sympatric process, ν induces the rise of geographical barriers
89 interrupting the gene flow and leading to an allopatric speciation. Even though
90 multiple speciations can co-occur, polytomies are not allowed in such process as
91 each species can speciate only once at the time. A likelihood expression for the
92 process is provided in Laudanno 2018.

93 **2.2 Tree simulations**

94 We can easily simulate such processes in continuous time using the Doob-
95 Gillespie algorithm. Simulations are performed using the function "mbd.sim"
96 from the `mbd` package (Laudanno 2018). We let parameters vary using all pos-
97 sible combinations of value included in Table 1 to simulate 1000 independent
98 trees for each parameter setting.

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	10

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

99 2.3 Pirouette

100 Once the MBD dataset has been simulated we exploit the **pirouette** package
101 (Bilderbeek & Laudanno 2019) to assess the error made by BEAST2, executing
102 the inference using a BD tree prior.

103 We briefly summarize here how the **pirouette** routine works. From each
104 MBD tree a DNA sequence alignment is simulated. For each sequence alignment
105 we then perform a Bayesian inference analysis (using BEAST2) to recover a
106 posterior distribution of trees. For each parameter setting, this process leads to
107 an inference error distribution. To evaluate the extent of this error we also run
108 a **pirouette** "twin" pipeline, through which we mimic the original pipeline in
109 every aspect but starting instead from "twin" trees. From each MBD simulated
110 tree we can produce its BD twin, inferring the most likely BD parameters (i.e.
111 λ and μ) through maximum likelihood, and use them to simulate a standard
112 BD tree (see Bilderbeek & Laudanno 2019 for more information), keeping the
113 topology of the original tree. We let the Bayesian analysis assume a BD prior in
114 both cases, to investigate the extent of the error we make under this assumption.
115 The twin pipeline serves as an estimation of the baseline error, as its error it is
116 not due by the mismatch of the generative prior with the inference prior.

117 3 Results

118 Here we present the results as show in Fig. 1...

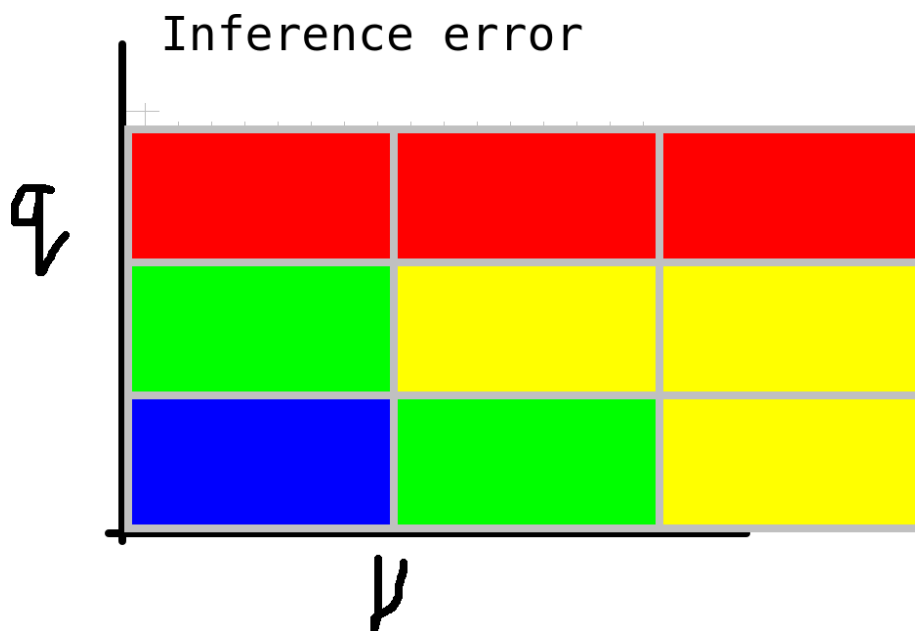


Figure 1: Here we show the inference error distribution generated from the MBD trees compared to the baseline error.

119 4 Discussion

120 We have shown how...

121 References

122 Bilderbeek, R.J. & Laudanno, G. (2019) *pirouette: create a posterior from a*
123 *phylogeny*.

124 Bilderbeek, R.J.C., Laudanno, G. & Etienne, R.S. (2019) Quantifying the im-
125 portance of a tree prior in bayesian phylogenetics.

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

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

131 Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A.
132 & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies
133 closer to agreement with the fossil record. *Proc R Soc Lond B: Biol Sci*, **279**,
134 1300–1309.

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

137 Janzen, T., Alzate, A., Muschick, M., Maan, M.E., van der Plas, F. & Etienne,
138 R.S. (2017) Community assembly in lake tanganyika cichlid fish: quantifying
139 the contributions of both niche-based and neutral processes. *Ecology and*
140 *Evolution*, **7**, 1057–1067.

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

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

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

- 148 Sturmbauer, C., Baric, S., Salzburger, W., Rüber, L. & Verheyen, E. (2001) Lake
149 level fluctuations synchronize genetic divergences of cichlid fishes in african
150 lakes. *Molecular Biology and Evolution*, **18**, 144–154.
- 151 Valente, L.M., Phillimore, A.B. & Etienne, R.S. (2015) Equilibrium and non-
152 equilibrium dynamics simultaneously operate in the galápagos islands. *Ecol-
153 ogy Letters*, **18**, 844–852.
- 154 Verheyen, E., Rüber, L., Snoeks, J. & Meyer, A. (1996) Mitochondrial phylo-
155 geography of rock-dwelling cichlid fishes reveals evolutionary influence of his-
156 torical lake level fluctuations of lake tanganyika, africa. *Philosophical Trans-
157 actions of the Royal Society of London Series B: Biological Sciences*, **351**,
158 797–805.