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

June 21, 2019

Abstract

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

Starting from heritable material and explicit assumptions, Bayesian phylogenetics allows to infer a jointly-estimated phylogeny and parameter estimates distribution. One of these assumptions in the speciation model, which mathematically describes the branching process of a phylogeny in time. The most used speciation model assumes that speciation events are independent, where we know that certain events can trigger speciation events in multiple species.

This research answers the question what the impact is of using a species tree model that assumes speciation is independent, when it is used on 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-occuring speciation events, which valudates

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 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 disallow for environment al changes that trigger speciations in multiple clades at a same point in time.

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

The (constant-rate) birth-death (BD) model embodies the common assumption that only a single speciation event can occur at any given time. The multiple-birth-death (MBD) model relaxes this assumption allowing, in additi on to standard BD events, also events in which large-scale environmental changes lead to speciation bursts. 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 BD tree priors might, in principle, prove to be good enough at detecting such events despite the lower level of complexity. If this is the case one should always be more keen to adopt the simplest model.

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

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

2 Methods

2.1 Model

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

2.2 Tree simulations

We simulate the speciation process in continuous time using the Doob-Gillespie algorithm, using the `mbd_sim` function in the `mbd` package (Laudanno 2018). We let parameters vary using all possible combinations of values as shown in Table 1. For each parameter setting, we generate 1000 independent reconstructed trees of the same crown age. We have picked the parameters in such a way that in the most speciose setting, the simulated trees have usually less than 200 taxa.

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.

2.3 Inference error estimation

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

In brief: `pirouette` starts from a 'true' (but unobservable in nature) starting phylogeny, from which a DNA sequence alignment (which is observable in nature) is simulated. From each sequence alignment, a Bayesian inference is run, to obtain a posterior distribution of jointly-estimated trees and model parameter estimates. By comparing the true tree and the posterior trees, an inference error distribution is generated. We use the twinning option available in `pirouette` to measure a minimum and full error, in which the minimum error is caused by stochasticity in the full pipeline, where the full error is the added error from using an invalid but standard speciation model.

In our context, the alignments are 1000 nucleotides in length, with a known root sequence of four 250 mono-nucloetide blocks, following the simplest nucleotide substitution model (JC69) and clock model (strict), with a mutation rate of $\frac{1}{2} \cdot c$, in which c is the crown age. With this mutation rate, each nuclotide is expected to mutate (both silently and non-silently) in half of the histories from the root sequence in the past to the sequence in the tips in the present.

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

For the error measurement, we use the nLTT statistic and a burn-in of 0.1.

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

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

3 Results

The inference error made for each of the parameter combinations is shown in Fig. 1. For both extinction rates, we find, as expected, that the error increases for increased ν or q. Also in line with our predictions, we find no difference between the two extinction rates.

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-occurent speciation event is dependent on the ratio between λ and a combination of ν and q.

We selected our parameters in such a way that the simulated trees had usually less than 200 taxa. One could argue that starting from trees with more taxa would result in a clearer inference, which we agree upon. We chose to

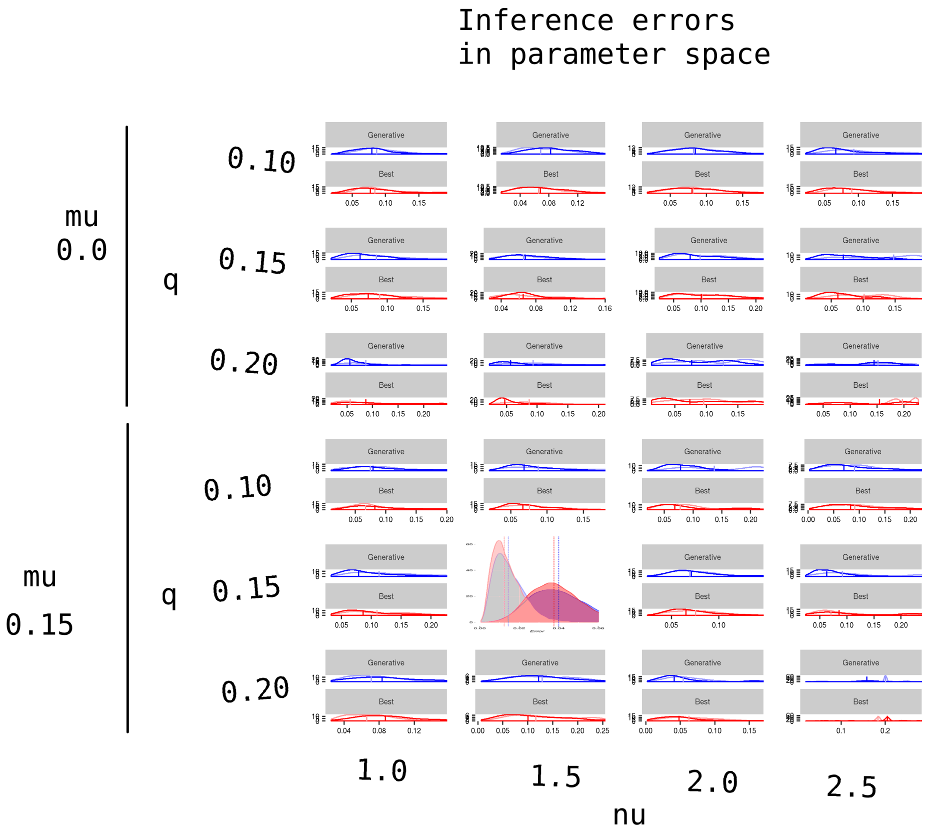

Figure 1: The inference error distribution for the different biological parameter settings. In all cases, $\lambda = 0.2$ and crown age equals 6.

use more replicates over more taxa, as we could easily add more replicates in a scheduled way.

References

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

Bilderbeek, R.J.C., Laudanno, G. & Etienne, R.S. (2019) Quantifying the importance of a tree prior in bayesian phylogenetics.

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

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

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

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

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

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

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

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

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

Verheyen, E., Rüber, L., Snoeks, J. & Meyer, A. (1996) Mitochondrial phylogeography of rock-dwelling cichlid fishes reveals evolutionary influence of historical lake level fluctuations of lake tanganyika, africa. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **351**, 797–805.