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

October 4, 2018

Abstract

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior **[GL: According to my fine graining approach we should at each step deepen every small section. At a certain level I think we can start to re-coarse-grain what we wrote to create the abstract.] [RJCB: I enjoy this approach! Did some minor fine-graining] [RJCB: Have you already looked up for a target journal? I know how a journal's constraints have helped me in writing an article, for example, by having a maximum number of pictures]**

1 Introduction

- There are many contemporary tools that provide the possibility to infer a phylogeny from genetic data (DNA, RNA, proteins). A popular Bayesian phylogenetic tool is called BEAST and its cousin BEAST2.

- BEAST is very flexible in setting up all possible phylogenetic priors (e.g. site/clock/speciation model).

- Current limits in current tools.

- BEAST2 gives us the possibility to introduce new tree priors to infer phylogenies based on different assumptions on how the speciation process takes place.

- One of such speciation processes is the multiple birth hypothesis, a new model (described below) and thus absent in BEAST.

- The Multiple birth hypothesis can be useful to explain a phenomenon that has always puzzled evolutionary biologists: what are the drivers of the diversification processes for those phylogenies that 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 **[RJCB: I feel MBSD (Multiple Birth Single Death) may be a better name: extinctions are still one at a time]** relaxes this assumption, allowing events in which large-scale environmental changes lead to a great number of species in relatively short time intervals. Such a hypothesis may be a better fit to describe the burst in cichlid fish diversification in systems like in the African Great Lakes: Malawi, Tanganyika and Victoria (Janzen *et al.* 2016, Janzen *et al.* 2017).

- However, it may be that current BD tree priors are good enough at detecting such events, with a (preferred) lower level of complexity. If this is the case one should always be more keen to adopt the simplest model.

- Here we present our study with the aim of exploring when using a more complex MBD tree prior is warranted.

2 Methods

2.1 Model

- Current phylogenetic tools assume that only a single speciation event can occur at any given time. While this assumption is useful to construct a wide variety of successful models (for example: Maddison *et al.* 2007, Valente *et al.* 2015, Etienne *et al.* 2012, Etienne *et al.* 2014), they disallow for environmental changes that trigger speciations in multiple clades at a same point in time.

- 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 events in which large-scale environmental changes lead to a great number of species in relatively short time intervals. Such hypothesis can be useful to describe, for example, systems like cichlid fish diversification in the African Great Lakes: Malawi, Tanganyika and Victoria (Janzen *et al.* 2016, Janzen *et al.* 2017).

- In the MBD model, parameters λ and μ correspond, respectively, to the common per-species speciation and extinction rates. Additionally, ν is the rate at which an environmental change is triggered. When such event

is triggered, all species present in the phylogeny at that moment have a probability q to speciate at that time, which is independent on λ.

- It is also possible to write down a likelihood function for such processes as in Laudanno 2018.

2.2 Simulations

- To prove our hypothesis we simulate two twin datasets. All the simulations are produced in continuous time, using the Doob-Gillespie algorithm.

- We start simulating 1000 **[RJCB: I will measure the number of trees we'll be able to simulate within a short enough time, when the experiment is set up]** MBD trees. From each MBD tree, a a DNA sequence alignment is simulated, after which that alignment starts a Bayesian analysis. We use the 'pirouette' package (Bilderbeek 2018) to call the BEAST2 tool suite from R. We let the Bayesian analysis assume a BD prior, to investigate the error this inference makes due to this.

- For each tree generated under the MBD model we aim to generate a "twin" tree under the BD model in order to perform the same analysis and compare the results. To do so we need to be sure that the comparison is fair. The method we adopt to achieve this goal is to impose that the amount of information used in the Bayesian information will be the same. These are the same number of taxa and the same (expected) number of DNA mutations. See appendix for the exact procedure.

- We explained how we set the parameters for each twin BD tree. Using this rules we generate a BD dateset. We repeat the analysis, producing alignments for each tree and subsequently using BEAST to produce a posterior for each of them.

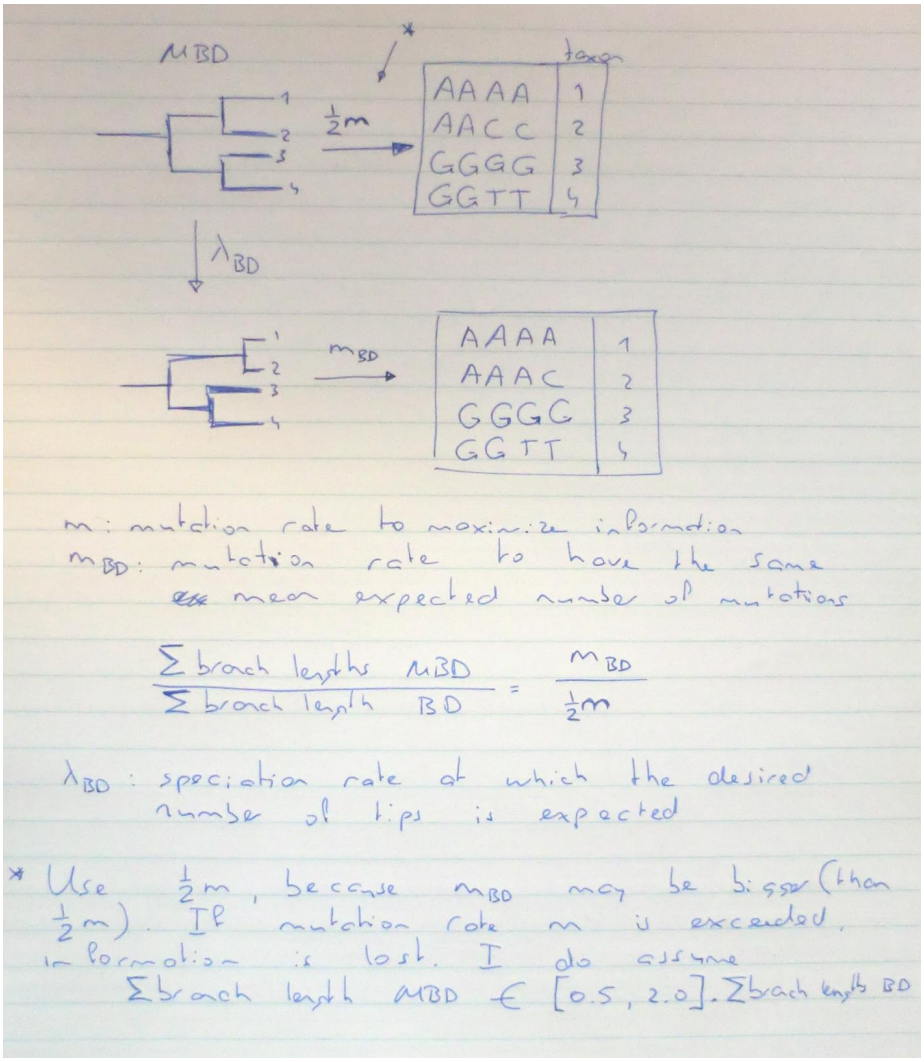

Figure 1: How to create twin trees and alignments. From a focal MBD tree, a twin tree is produced as such: (1) estimate the λ_{BD} to get the same expected number of tips, (2) simulate a BD tree with that amount of tips (discard trees with different number of tips), (3) estimate a mutation rate to get an alignment with the same expected number of mutations, (4) simulate alignments with that amount of mutations (discard those that don't, the picture shows an alignment that should be discarded)

- Now we have two datasets of posteriors to compare, one for the BD model and one for the MBD model.

- To compare the results for the two models we measure the inference error using the nLTT statistic between known/true tree and posterior/inferred trees. **[RJCB: I would love to describe this more concrete. For example, when do we say something has an effect? If we avoid making such judgements, how will we visualize?]**

  [RJCB: I removed the Bayes factor text. It is useful when letting BEAST2 pick more/overly complex models and see if that more complex model fits the data better (penalized by its increased complexity, similar to the AIC). It has its uses, but I am unsure if we already want to discuss this now or first focus on the proper tree twinning]

3 Results

- **[RJCB: I guess you know I am a fan of the Open Science Framework, in which you first register you work before you do the experiment (note: I will do some small pilots to estimate the complete time of the experiment). I think it is the proper and superior science, which helps us against writing down bullshit stories after having obtained the results (e.g. 'We expected A and indeed found it!'). It also helps me structure my work: first think deeply about the experiment, then do it (instead of the mixing up the two phases). What are your thoughts on that?]**

-

References

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

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.

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.

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.

A Creating twin trees

[RJCB: I put your work here and LaTeXed it. Maybe this moves to the Methods, but this seemed a good place to start]

The expected number of mutations, m of a phylogeny with crown age T, depends on the number of DNA nucleotides L, per-site mutation rate μ and the number of species n of time as such:

$$m = L \cdot \mu \cdot \int_{t=0}^{T} n_t \, dt \tag{1}$$

[RJCB: I suggest to use a different symbol for mutation rate, for example α or ρ. Using lambda here as well feels needlessly confusing.]

It can be easily seen considering that μ is the number of mutations per unit time per site. For this reason it's needed to multiply by time and number of sites.

To obtain twin trees with an equal amount of information, we impose that the expected number of mutations in an MBD tree, m_{MBD} equals the expected number of mutations in a BD tree, m_{BD}:

$$m_{MBD} = m_{BD} \tag{2}$$

From which follows:

$$m_{BD} = L \cdot \mu \cdot \int_{t=0}^{T} n_{BD}(t) \, dt \tag{3}$$

Obviously we cannot know $n_{BD}(t)$ before running the simulations but we can replace it with a proxy, for example the average number of species in time according to the BD model **[RJCB: I suggest $n_{BD} = n_{MBD}$ and only change**

μ_{BD} **to reach** $m_{MBD} = m_{BD}$]

[GL: START]

$$< n_{BD}(t) >= n(t=0) * exp[(mu - lambda) * t] (6)$$

The relation to use to get the equivalent lambda should therefore be

$$m_{MBD} = L{*}mu{*}integral[exp[(mu{-}lambda){*}t]dt] = (L{*}mu{*}exp[(mu{-}lambda){*}t])/(mu{-}lambda) (7)$$

Here everything is known but lambda. So solve the lambda and use that value.

My doubt is if we need to use m_{MBD} for the single tree or the same quantity averaged on the full MBD dataset $< m_{MBD} >$.

Does it make sense to you? Do you think is better to use the individual m_{MBD} for each tree or the average across the whole dataset?

[GL: END]

[RJCB: I think I see your point: in single trees stochasticity works per tree twin, in a distribution per distribution. To circumvent this: what if I would be able to measure the actual number of mutations? Then twinning trees by having an exact equal amount of mutations would be very clean. Let me know: you think I should investigate this?]