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 5, 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 a fair comparison, using trees with the same amount of information, which is the number of taxa and the same (expected) number of DNA mutations. To obtain these twin trees, **[RJCB: I suggest to first start with the equal number of taxa, and the calculation of the speciation rates first.]** 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}:

$$m_{MBD} = m_{BD} \tag{1}$$

We first generate a set of MBD trees. For each of them we can measure the amount of mutations m_{MBD}.

[GL: I think this should definitely go to the methods] [RJCB: I put it there for use, hopefully at a spot you liked :-)]

The expected number of mutations m of a phylogeny **[RJCB: I think 'expected number of mutations' would be more correct. Do you agree?]** with crown age T in fact is given by **[RJCB: above stood 'crown age -T'. I feel that a crown age is a positive number, but I know you have had a good reason. Perhaps better would be to write something explicit like: t_now - t_crown = T. Looking forward for a better suggestion than mine :-)]**

$$m = L \cdot \rho \cdot \int_0^T n(t)\ dt \tag{2}$$

where L is the number of DNA nucleotides, ρ is the per-site per-species mutation rate and $n(t)$ the number of species at each time.

[GL: This feels kind of a repetitions of what we wrote before the formula. I comment it. We can think of reinsert it afterwards, if needed (see comment above).] [RJCB: I suggest to remove such commented-out lines. Although I sometimes get attached to my sentences, they clog up the document by non-info and I usually delete them anyways in the end. I cannot remember ever regretting this (would I, I could find it in the git history). Would you agree?]

Since we cannot know $n_{BD}(t)$ before running simulations we need to replace it with a proxy. For this reason we will use the average number of species in time according to the BD model. It's well known that this is

equal to [GL: insert proper citation] [RJCB: I see you use angle bracket as a notation for the expected value. I usually see 'E(x)' as the expected value for 'x', and this is used at the beloved https://en.wikipedia.org/wiki/Expected_value. What are the reasons you prefer the notation with the angle brackets?]

$$< n_{BD} > (t) = n_0 \cdot e^{(\mu_{BD} - \lambda_{BD})t} \tag{3}$$

where $n_0 = n_{BD}(-T) = n_{MBD}(-T)$ is the initial number of species at the crown age. From 1, 2 and 3 follows:

$$m_{MBD} = L \cdot \rho \cdot \int_0^T < n_{BD} > (t)\ dt = L \cdot \rho \cdot n_0 \cdot \left[\frac{e^{(\mu_{BD} - \lambda_{BD})T} - 1}{\mu_{BD} - \lambda_{BD}} \right] \tag{4}$$

If we set $\mu_{BD} = \mu_{MBD}$ and reverse this relation we can extrapolate the value of λ_{BD} to use to generate BD trees.

[RJCB: I suggest $n_{BD} = n_{MBD}$ and only change ρ_{BD} to reach $< m_{MBD} >=< m_{BD} >$] [GL: @Richel: Don't you think it might make more sense to set $\mu_{BD} = \mu_{MBD}$? What changes in the two model is the way we use to generate new species, not the way to remove them. Maybe one thing that's possible to do would be to make λ_{BD} a function of time, a bit like Tho is doing in his comparison between DD and TD4 (which, by the way, seem to yield very different results). In case you are wondering the theory can be found in Caesar's master thesis.] [RJCB: I fully agree to use the same extinction rates! The 'mu' used in the context of mutations (now 'rho') messed me up. I hope this is clear now. To recap: (1) calculate the speciation rate of the

twin tree as you wrote down excellently, (2) simulate a twin tree with same number of taxa, (3) calculate the mutation rates of the trees, so their alignments contain as much information]

[GL: My doubt is if we need to use m_{MBD} for the single tree or the same quantity averaged on the full MBD dataset $< m_{MBD} >$. Do you think is better to use the individual m_{MBD} for each tree or the average across the whole dataset?] [RJCB: I think a per-tree calculation of the mutation rates is the best we can do. As there was some noise between us above, I think the iteration after the next will allow us to get a better idea about this]

- 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.

- 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

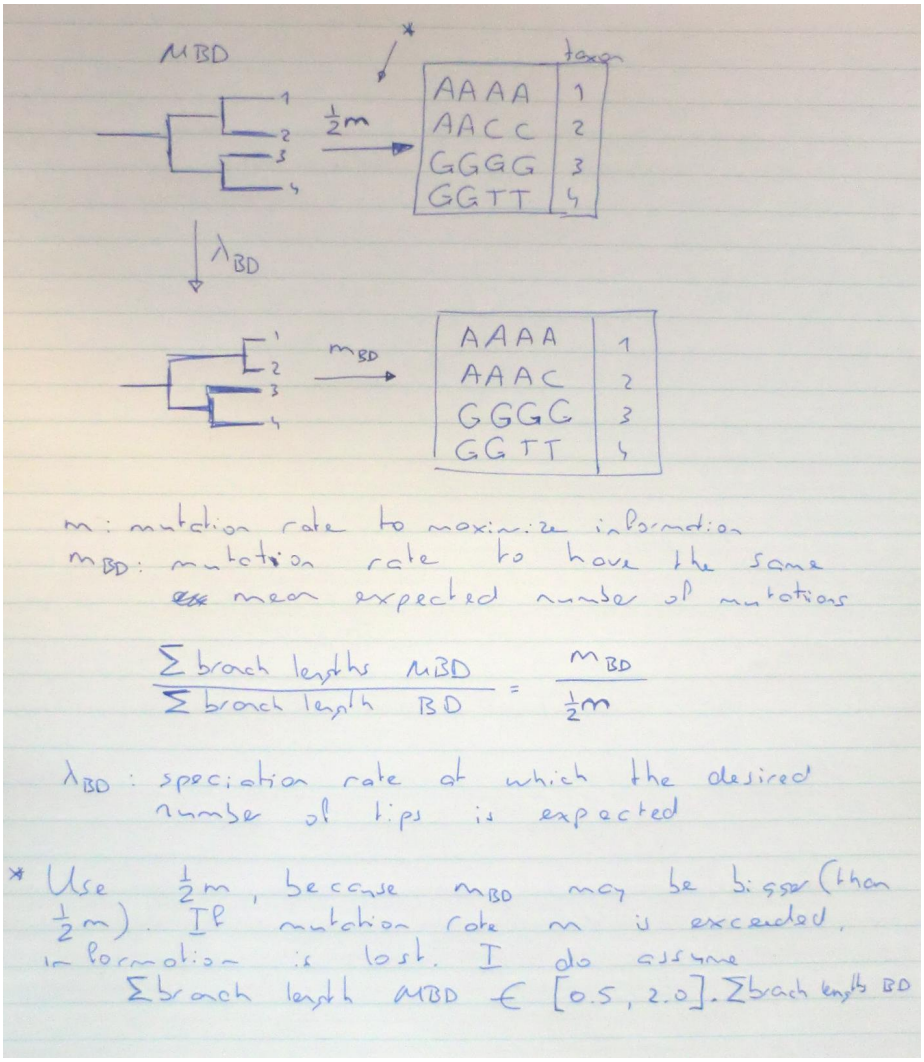

Figure 1: How to create twin trees and alignments. From a focal MBD tree, a twin tree is produced as such: (1) estimate the λ_{BD} to get the same expected number of tips, (2) simulate a BD tree with that amount of tips (discard trees with different number of tips), (3) estimate a mutation rate to get 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)

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.