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 8, 2018

Abstract

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior **[RJCB: Have you already looked up for a target journal?] [GL: Honestly I have literally no idea how to select a good journal for this kind of article.] [RJCB: I suggest you ask Rampal]**

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 (Drummond & Rambaut 2007) and its cousin BEAST2 (Bouckaert *et al.* 2014).

- BEAST is very flexible, providing the user with the option to set up all possible phylogenetic priors (e.g. site/clock/speciation model).

- However, currently available priors can be not suitable to analyze some specific datasets.

- 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 currently 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] [GL: I always had the same doubt. The problem is that a good name should be, at the same time, both descriptive and short. I'll think about that.]** 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 in systems like cichlid fish diversification 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 de-

tecting 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 (e.g 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 present also in the standard BD model. Additionally, MBD relies on two additional parameters. Parameter ν is the rate at which an environmental change is trig-

gered. 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 $N_S = 1000$ **[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 DNA sequence alignment is simulated. For each sequence alignment we then perform a Bayesian analysis to recover a posterior distribution of trees, each composed of N_P phylogenies. Such analysis is performed using 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. We want trees from the two models to contain the same amount of information, i.e. the same (expected) number of DNA mutations and the same number of taxa at the present. To obtain these twin trees, **[RJCB: I suggest to first start describing how to create the twin tree, calculation of the speciation rates, then the twin alignment.]** **[GL: @richel: both assumptions seem to make sense to me. If possible I would**

run both and see which one performs better. For now let's stick to option 1 in the manuscript.] [RJCB: Sorry for the confusion: the previous comment was about the order of describing out methods in this manuscript. As of the methods, I really meant to create twin trees with both equal number of taxa and equal expected number of mutations.]

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}.

The expected number of mutations m of a phylogeny **[RJCB: I think 'expected number of mutations' would be more correct. Do you agree?] [GL: I don't agree. It is expected if you use an expected number of species over time. If you have the correct $n(t)$ in principle you should be able to get the exact number of mutations.] [RJCB: If you have the correct $n(t)$ you can indeed calculate exactly what the number of mutations will be, if and only if the mutation rate is zero. Except for this travial case, I am curious about how you calculate this. What is the precise number of mutations on a DNA sequence of 1000 base pairs over one time unit with an expected average per-nucleotide mutation rate of once per time? I would say one expects on average 1000 mutations to occur. What would your claim be?]**

with crown age $-T$ (with $T > 0$) 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 :-)] [GL: We can specify that the crown age is $-T$ where T is a positive number (which should be already clear, I guess, if we report it as $-T$). In my opinion this makes everything nicer when we have to write down the integral which should, in principle, be $\int_{-T}^{0} f(t-(-T))$ but using integral properties can be rewritten as $\int_{0}^{T} f(t)$. Having a definite integral starting from 0 always looks better n] [RJCB: I see, in a conversion you would say 'the crown age of the phylogeny is minus ten million years', where I would the say 'the crown age of the phylogeny is ten million years' to say exactly the same thing! To try to resolve this, I checked one (actually two) of Rampal's papers ('Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record', by Rampal, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N. Pearson, Andy Purvis, Albert B. Phillimore). Rampal avoids this point altogether: '... crown age $t_c = t_p t_1$' and '... and crown ages (5, 10 and 15 Myr)'. I suggest we do the same in both writing and derivations: he is even more famous than us :-)]

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

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.

Since we cannot know $n_{BD}(t)$ before running simulations we need to re-

place 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?]** **[GL: Because the capital E is inelegant, plebeian and boor. And we want to be classy, don't we? :D]** **[RJCB: If the angle brackets are the new hip kids in mathematics town, sure, we'll use those!]**

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

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] [GL: I am still very debated about this. I actually would like to ask also Rampal's opinion.] [RJCB: Sounds like a good idea!]

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

2.3 Model selection

- So far we have simulated two datasets of trees under the two models: $\{T_i^{BD}\}_{i=1}^{N_S}$ and $\{T_i^{MBD}\}_{i=1}^{N_S}$. We used them to generate a dataset of alignments for each model: $\{X_i^{BD}\}_{i=1}^{N_S}$ and $\{X_i^{MBD}\}_{i=1}^{N_S}$. From each dataset

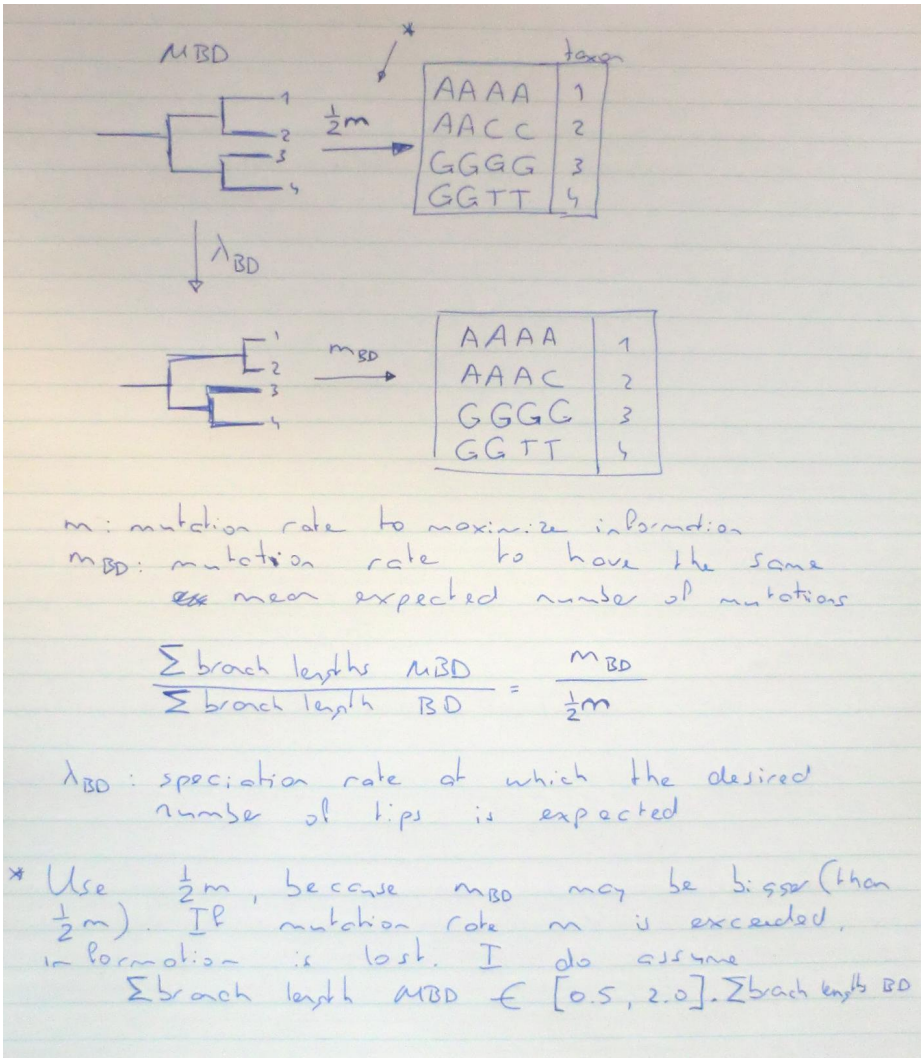

Figure 1: How to create twin trees and alignments. From a focal MBD tree, a twin tree is produced as such: (1) estimate the λ_{BD} to get the same expected number of tips, (2) simulate a BD tree with that amount of tips (discard trees with different number of tips), (3) estimate a mutation rate to get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

we produced a posterior distribution from a BD prior: $P_i(\theta|X_i^{BD}, BD)$ and $P_i(\theta|X_i^{MBD}, BD)$. [GL: 1) We might want to rename the models, e.g. BD = (0) and MBD = (1). These names with capital letters are too big and ugly;] [RJCB: I definitely get your point! I would prefer something expressive over a number, perhaps a fancy letter, like for example $\mathbb{B}$ and $\mathbb{M}$ for a Birth-Death and Multiple-Birth Death.] [GL: 2) Writing down things to tidy up manuscript and thoughts I just realized that actually the "model" is the same for the two posteriors. In fact with the word "model" Maturana et al. refer to the prior used (see eq. 1, $\pi(\theta|M)$). In such case are we sure we can actually use the Bayes factor for model selection?] [RJCB: Correct! The assumed model of the posterior is a Birth-Death model. To do model selection, we could assume a different nucleotide substitution model (GTR instead of JC69) and/or a different clock model (relaxed clock, instead of a strict clock). I hope this clears things up. I will add 'generative model' for the model (BD or MBD) that created a phylogeny and 'inferential model' for the model used in generating a posterior.]

2.3.1 Option 1: nLTT statistics

- 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 (Janzen 2015). To obtain such statistics the procedure is the following:

  - From each tree $T_{i,j}^M$ (with $j = 1, ..., N_S$) belonging to the posterior $P_i(\theta|X_i^M, BD)$ and relative to the model M, we extrapolate the lineage-

through-time (LTT), in other words we measure the number of species as a function of time $n_{i,j}(t)$. To allow a comparison we normalize dividing by the maximum number of species of each tree, i.e. the number of tips at the present $N_{i,j}(t) = \frac{n_{i,j}(t)}{n_{i,j}^{max}}$. We then define the nLTT measure as

$nLTT_{i,j} = \int_0^T |N_{i,j}(t) - N_{T_i}|dt$

[GL: I am running out of letters :(] [RJCB: Haha! I suggest to use the same equation and symbols as equation 1 in the nLTT article of Janzen, Hoehna and Etienne, 2015:]

$$\Delta nLTT = \int_0^1 |nLTT_1(t) - nLTT_2 t|dt$$

2.3.2 Option 2: Bayes Factor (BF)

[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] [GL: But after we describe the tree twinning we will have to describe the method we use for model selection anyways. BF and nLTT are mutually exclusive choices, right?] [RJCB: BF and nLTT are not mutually exclusive choices, but do complementary things: nLTT measures the difference between the 'true' tree and the inferred trees, which is the most important thing: we are interested in the error BEAST2 makes when using an overly simple (single-)BD tree prior. With the BF we can compare BEAST2 posteriors that use (the same tree prior, but) different site models (e.g. JC69, the simplest, versus

GTR, the most complex). Although we know we generated the DNA alignment with JC69, there are some hints that the GTR model is the superior choice to do inference with.] [GL: The point is: what method do we choose? According to the choice we have to describe that. I'd also like to know Rampal's opinion on that. In principle I like more the BF method cause it looks more solid while the nLTT looks a bit like the last resort to me.] [RJCB: I hope this is now more clear. If not, I'll happily elaborate :-)]

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?] [GL: I agree on the principle but I don't fully agree on the open science procedure. Sometimes you do find something good while you're looking for something else. That happened many times in the history of science. This doesn't necessarily implies that we will end up p-hacking the results. I think we need to be open to new findings that might be not forecasted by our hypothesis. I think this is the good way to do science: be crit-

ical with our results but not imposing that reality will follow the theory.] [RJCB: Looks like we are on the same page here: we both dislike making up bullshit stories after the results are known. Great!]

•

References

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

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. (2015) *nLTT: Calculate the NLTT Statistic.* R package version 1.1.

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.