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

November 27, 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: May I suggest we aim for Molecular Phylogenetics and Evolution, the same journal as the raket paper?]**

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. With this work we aim to test whether or not the implementation of a new prior model is beneficial to study a specific kind of diversification process.

- 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 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 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 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.

- We hypothesize that the error made today, using BD tree priors, increases with an increased number or stronger effect of multiple birth events. This is straightforward: without multiple birth events or such event having no effect, the MBD model falls back to a BD model. We expect larger errors when we deviate more from the BD model's assumptions.

- Additionally, we hypothesize MBD having a stronger effect if the normal speciation process is less pronounced. The more speciations are caused by the BD process, there are relatively less multiple-birth events.

- We expect the effect of extinction rates to be neutral, as extinctions will hit lineages created by both speciation processes equally.

- Summarized, we expect the error made be correlated to the number of species created by the multiple-birth process over the total number of species created:

$$< e >= f(\frac{n_{taxa}^{MBD}}{n_{taxa}^{BD} + n_{taxa}^{MBD}}) \tag{1}$$

Where $< e >$ denotes the expected error, f is a monotonously increasing function of unknown shape, n_{taxa}^{MBD} is the number of taxa created in multiple-birth events and n_{taxa}^{BD} is the number of taxa created by the standard BD specition process.

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 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. Polytomies are not allowed in such process as each species can speciate only once at the time.

- 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 $N_S = 1000$ 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 in both cases, to investigate the extent of the error we make under this assumption.

- For each tree generated under the MBD model we aim to generate a "twin" tree under the BD model. With the word "twin" we denote a tree generated starting from the respective MBD tree, in order to perform a fair comparison with it. This operation has to be done, because we want to compare two trees that are generated by different processes. To do so we infer the parameters λ_{BD} and μ_{BD} from the MBD maximizing the likelihood under a BD model. To perform this operation we use the function "bd_ML" from the package "DDD" (Etienne *et al.* 2012).

- We then exploit such parameters to generate a BD tree using the function "tess.sim.taxa.age" from the package "TESS" (Hhna 2013). We simulate the tree in such a way the new tree has the same number of tips and the same crown age as the MBD tree. We furthermore require that the BD tree conserve the topology of the MBD tree.

  We want the MBD and twin BD trees to contain the same amount of information, i.e. the same number of DNA mutations and the same number of taxa at the present:

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

The expected number of mutations m of a phylogeny with crown age $-T$ (with $T > 0$) in fact is given by **[RJCB: So one of use likes '-T', the other likes 'T'. How to resolve this?]**

$$m = L \cdot \rho \cdot \int_0^T n(t)\ dt \tag{3}$$

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.

The parameter we'll tune is ρ ... **[RJCB: elaborate here :-)]**

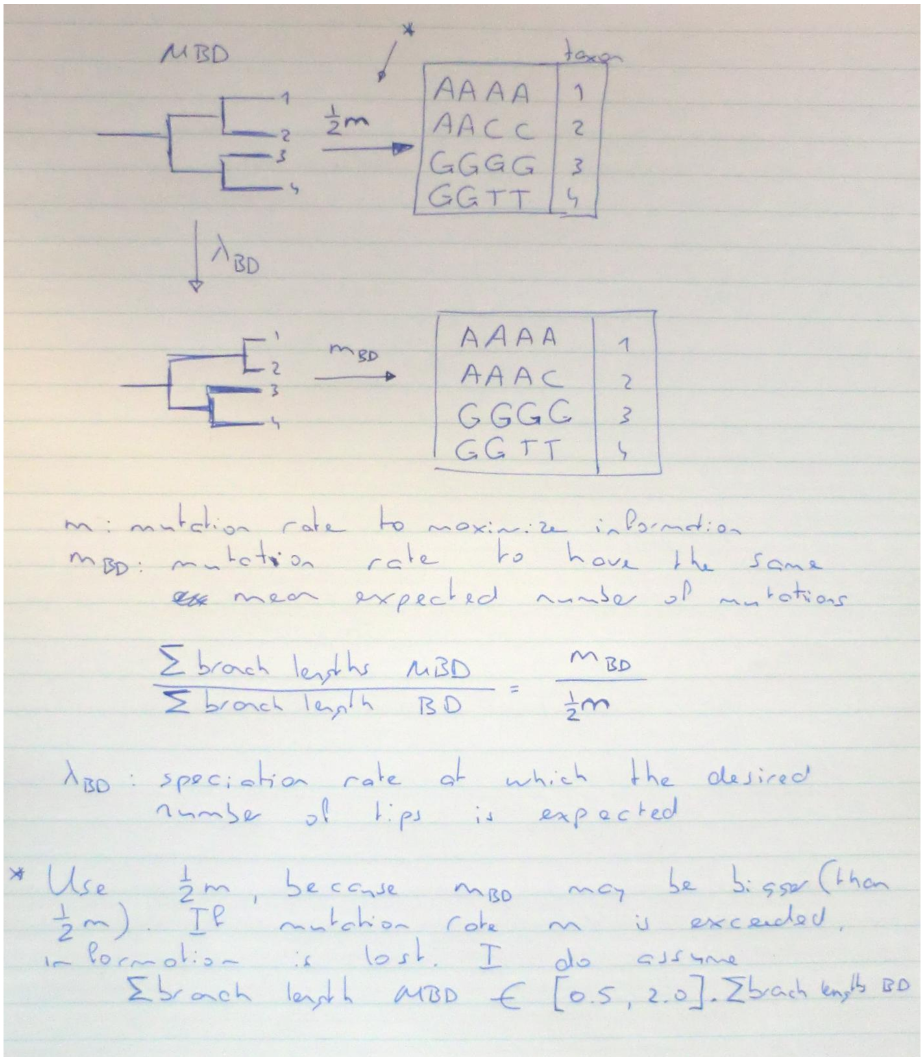
Since we cannot know $n_{BD}(t)$ before running simulations we need to replace it with a proxy. For this reason we will use the average number of species in time according to the BD model. It's well known that this is equal to **[GL: insert proper citation]**

$$< n_{BD} > (t) = n_0 \cdot e^{(\mu_{BD} - \lambda_{BD})t} \tag{4}$$

where $n_0 = n_{BD}(-T) = n_{MBD}(-T)$ is the initial number of species at the crown age. From 2, 3 and 4 follows:

$$m_{MBD} = L \cdot \rho \cdot \int_0^T < n_{BD} > (t)\ dt = L \cdot \rho \cdot n_0 \cdot \left[\frac{e^{(\mu_{BD} - \lambda_{BD})T} - 1}{\mu_{BD} - \lambda_{BD}} \right] \tag{5}$$

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.



taxon	
AAAA	1
AACC	2
GGGG	3
GGTT	4

AAAA	1
AAAC	2
GGGC	3
GGTT	4

m: mutation rate to maximize information

m_{BD}: mutation rate to have the same ~~ex~~ mean expected number of mutations

$$\frac{\sum \text{branch lengths MBD}}{\sum \text{branch length BD}} = \frac{m_{BD}}{\frac{1}{2}m}$$

λ_{BD} : speciation rate at which the desired number of tips is expected

* Use $\frac{1}{2}m$, because m_{BD} may be bigger (than $\frac{1}{2}m$). If mutation rate m is exceeded, information is lost. I do assume

$$\sum \text{branch length MBD} \in [0.5, 2.0] . \sum \text{brach lengths BD}$$

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 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 Measuring the inference error

- 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 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 would strongly prefer MBD and BD, as I feel replacing the big ugly capital letters by short pretty numbers hurts readability even more]**

- 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:]

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

2.4 Model selection

3 Results

-
-

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.

Hhna, S. (2013) Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes. *Bioinformatics*, **29**, 1367–1374.

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.