Fitting diversification models on undated or partially dated trees

Nicolas Lartillot

Peer Community In Evolutionary Biology

Fitting diversification models on undated or partially dated trees

Nicolas Lartillot based on reviews by Amaury Lambert, Dominik Schrempf and 1 anonymous reviewer

Open Access

A recommendation of:

Gilles Didier. **Probabilities of tree topologies with temporal constraints and diversification shifts** *(2020), bioRxiv, 376756, ver. 4 peer-reviewed by Peer Community in Evolutionary Biology.* 10.1101/376756

Published: 10 January 2020

Submitted: 30 January 2019, Recommended: 09 January 2020

Cite this recommendation as:

Nicolas Lartillot (2020) Fitting diversification models on undated or partially dated trees. *Peer Community in Evolutionary Biology, 100088.* 10.24072/pci.evolbiol.100088

Phylogenetic trees can be used to extract information about the process of diversification that has generated them. The most common approach to conduct this inference is to rely on a likelihood, defined here as the probability of generating a dated tree T given a diversification model (e.g. a birth-death model), and then use standard maximum likelihood. This idea has been explored extensively in the context of the so-called diversification studies, with many variants for the models and for the questions being asked (diversification rates shifting at certain time points or in the ancestors of particular subclades, trait-dependent

diversification rates, etc). However, all this assumes that the dated tree T is known without error. In practice, trees (that is, both the tree topology and the divergence times) are inferred based on DNA sequences, possibly combined with fossil information for calibrating and informing the divergence times. Molecular dating is a delicate exercise, however, and much more so in fact than reconstructing the tree topology. In particular, a mis-specificied model for the relaxed molecular clock, or a mis-specifiied prior, can have a substantial impact on the estimation of divergence dates - which in turn could severely mislead the inference about the underlying diversification process. This thus raises the following question: would that be possible to conduct inference and testing of diversification models without having to go through the dangerous step of molecular dating? In his article ""Probabilities of tree topologies with temporal constraints and diversification shifts"" [1], Gilles Didier introduces a recursive method for computing the probability of a tree topology under some diversification model of interest, without knowledge of the exact dates, but only interval constraints on the dates of some of the nodes of the tree. Such interval constraints, which are derived from fossil knowledge, are typically used for molecular dating: they provide the calibrations for the relaxed clock analysis. Thus, what is essentially proposed by Gilles Didier is to use them in combination with the tree topology only, thus bypassing the need to estimates divergence times first, before fitting a diversification model to a phylogenetic tree. This article, which is primarily a mathematical and algorithmic contribution, is then complemented with several applications: testing for a diversification shift in a given subclade of the phylogeny, just based on the (undated) tree topology, with interval constraints on some of its internal nodes; but also, computing the age distribution of each node and sampling on the joint distribution on node ages, conditional on the interval constraints. The test for the presence of a diversification shift is particularly interesting: an application to simulated data (and without any interval constraint in that case) suggests that the method based on the undated tree performs about as well as the classical method based on a dated tree, and this, even granting the classical approach a perfect knowledge of the dates - given that, in practice, one in fact relies on potentially biased estimates. Finally, an application to a well-known example (rate shifts in cetacean

phylogeny) is presented. This article thus represents a particularly meaningful contribution to the methodology for diversification studies; but also, for molecular dating itself: it is a well known problem in molecular dating that computing and sampling from the conditional distributions on node ages, given fossil constraints, and more generally understanding and visualizing how interval constraints on some nodes of the tree impact the distribution at other nodes, is a particularly difficult exercise. For that reason, the algorithmic routines presented in the present article will be useful in this context as well.

References

[1] Didier, G. (2020) Probabilities of tree topologies with temporal constraints and diversification shifts. bioRxiv, 376756, ver. 4 peer-reviewed and recommended by PCI Evolutionary Biology. doi: 10.1101/376756

Revision round #2

2019-11-23

Dear Gilles,

Your revised manuscript has been reviewed by Pr. Amaury Lambert. As you will see, only minor points remain to be fixed: if you could just have a look at them. In particular, I agree that it would be important to refer to the alternative method proposed by Amaury Lambert directly in the main text, referring to his review. The reviewing process is public, and thus it is probably a good thing to refer the Readers to it directly from the manuscript, so as to invite them to read it and compare the two algorithms.

We are very close to final acceptance. Once you have submitted your final version, I will proceed with the recommendation.

witth best regards,

nicolas lartillot

Preprint DOI: https://doi.org/10.1101/376756

Reviewed by Amaury Lambert, 2019-11-19 21:29

Download the review (PDF file)

Author's reply:

Dear Nicolas,

I uploaded the revised manuscript on BioRXiv (it is available on their site). I fixed the issues pointed out in the last review, except the remark about the paragraph at the end of page 7. Since Amaury Lambert does not plan to publish his alternative method, I briefly exposed it page 13 of the revision. The revised manuscript includes the modifications suggested in the e-mail from PCIEvolBiol (mentions of PCIEvolBiol in the aknowledgements section and in a footnote in the first page). I (temporarily) gave up using the PCI template since the combination font/text width led to issues in typesetting the formulae. Many thanks for your work.

with best regards,

Gilles

Download author's reply (PDF file)

Revision round #1

2019-04-24

Dear Gilles Didier,

There is a general consensus among the reviewers that this manuscript represents an important contribution in the field of diversification studies. The algorithmic and computational results are potentially useful, and their derivation is tight and rigorous.

On the other hand, there is also a general feeling that, as it stands, the manuscript is very technical and does not sufficiently emphasize the intuitions behind the

mathematical developments or the potential applications to specific research questions in diversification studies. In the end, there is a legitimate concern that this highly technical presentation will make the manuscript not accessible to most readers of the targeted audience and will not do justice to the practical significance of the work.

The reviewers have made several suggestions to improve the overall presentation and make it less arduous, among which: - getting rid of the combinatorial factors related to the labelling of the tree, by labelling it from the start; - doing the recursion only in terms of the constraints on node ages, leaving the piece-wise constant aspect of the model hidden in the details -- in fact, the whole derivation could even be conducted under a homogeneous birth-death, then just suggesting that the calculation could be generalized to arbitrary piecewise constant. or even other time-varying, versions of the process, without major modifications. - using both simpler and more explicit notations; - relying a graphical example for explaining the intuition behind the quadratic recursive algorithm (e.g. continuing on the example given in figure 3).

I agree with those suggestions. I would even go further, and suggest a different way to organize the manuscript: in the main text, a more general and more intuitive description of the main algorithmic ideas could be given, relying more heavily on a graphical example such as the one given in figure 3, and leaving all technical aspects of the derivation (much of the current main text) in an appendix. Then, as suggested by one of the reviewers, more emphasis could be put on the applications. This would give the reader with two options: either a fast track (to get the general idea and appreciate the significance of the work in terms of its potential applications), or the complete story, for the more theoretically inclined readers.

The english also needs improvement.

Of note: one of the reviewers point out an alternative integration method, which might have a better complexity as a function of tree size. This should probably be examined and discussed.

Concerning the application to testing for diversification shifts, I would have some additional comments:

(1) in practice, the shift time is not known, but one may have good fossil data giving an upper and/or lower bound for the age of the last common ancestor of the subclade. Similarly, the time of origin of the entire clade is not known either, but some interval constraint derived from fossil information might be available concerning the age of the root. I was wondering if the test could be designed so as to rely on this practically more relevant fossil information instead of relying on the knowledge of the shift time (and of the time of origin, which is fixed and assumed known, right?).

(2) comparing Lambda*N with Lambda*P is theoretically interesting, but not so useful in practice (since exact knowledge of divergence times, such as assumed by Lambda*P, is lacking). In real-world applications, one would instead want to compare Lambda*N with a plug-in version of Lambda*P relying on an explicit dating of the tree obtained using relaxed clock approaches. In this context, a key question is whether Lambda*N shows more robustness, without loosing so much in sensitivity. This point could be discussed.

(3) ideally, a empirical example could be presented based on a previously published case (this relates to the suggestion of one of the reviewers, to put more emphasis on the applications).

Additional requirements of the managing board:

As indicated in the 'How does it work?' section and in the code of conduct, please make sure that (if adequate): -Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad (to pay) or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data. -Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused. -Details on experimental

procedures are available to readers in the text or as appendices. -Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: “XXX is one of the PCI XXX recommenders.”

All the best. The Managing Board of PCI Evol Biol.

Preprint DOI: https://doi.org/10.1101/376756

Reviewed by Amaury Lambert, 2019-04-09 10:54

Download the review (PDF file)

Reviewed by anonymous reviewer, 2019-04-15 17:09

This manuscript focuses on the calculation of the joint probability density of a tree topology and internal node ages under the piecewise-constant birth-death and sampling model of diversification with shifts in the birth and death parameters during the course of evolution and time constraints. Being able to evaluate this density in an efficient manner is important as it is a the core of macro-evolutionary approaches that characterize the fluctuation of species diversity across taxa and time.

My comments are mainly about elements used to derive the main results and not about the main results themselves. First, I did not quite get what the times s*i, i=0..k exactly correspond to. They are not arbitrary values since s*0 and s*k are obviously not arbitrary times. When leaving s*0 aside, they are neither random variables corresponding the times of sampling events or the times at which lineages die since, in Figure 1 left, there are four of these events but only two values of s (i.e., s*1 and s*2). Giving a precise definition for these times would help.

On page 6, the time tau*n*i are not defined previously. Also, the relationship between the node ages and tree topology are not well defined in the current manuscript in my opinion. Indeed, a tree topology induces a partial ordering of

internal node ages which interactions with the time constraints is not explicitly dealt with.

On Figure 3, my understanding is that the set of all start sets A with node b in A is {{b,a}}. I do not understand why the author considers then that the trees to the right of the summation sign represent the set of all start-sets with b in A (beside the fact that a set of start-sets is not a set of trees if I am not mistaken).

A brief illustration of how the quadratic computation works on the toy example of Figure 3 would probably be very helpful. Moreover, it was not clear to me whether the computation time would stay quadratic when increasing the number of shifts. It would be interesting to mention whether the proposed approach remains computationally efficient (or not) whenever the number of shifts increases.

On Section 7 onward, the author keeps referring to P(\tau, ... | \tau). I think the \tau on the left of the conditional should be removed. Also, corrolary 1 gives the cumulative density function for the age of a particular node. It is not obvious to me that obtaining the derivative of this function is straightforward (in order the get the pdf). I would recommend adding some explanations here. I also did not understand why birth, death and sampling parameters are considered here as three separate parameters as the birth-death-sampling process only has two identifiable parameters (see Equation 6 in Stadler, Journal of Theoretical Biology, 2009. 261: 58-66).

Moreover, it is not clear how one can derive the joint density of tree topology and *all* internal node ages from the results presented in this study. This joint density is needed in case one wants to use the piecewise constant birth-death-sampling model in standard phylogenetic inference using MCMC.

The caption of Figure 5 makes references to three row while only two are displayed here.

On page 15, Lemma ?? needs fixing.

Reviewed by Dominik Schrempf, 2019-02-19 09:08

Download the review (PDF file)

Author's reply:

Dear Nicolas,

Please find joined my detailed response to the reviewers comments and the revised version with the modifications in blue color.

Best wishes,

Gilles

Download author's reply (PDF file)