Responses to reviews

September 4, 2019

I wish to warmly thank the reviewers and the editor for their readings, their comments and their useful suggestions. I do apologize for the delay in submitting the revised version of my manuscript. I was swamped with other obligations. Then, I hesitated for a while before deciding to submit a revision, since the method proposed by Amaury Lambert is far more elegant than the one of my manuscript (presenting a "complicated" approach while there is a simpler one which performs the exact same task would be pure sadism). I eventually convince myself that the approach of the manuscript may still deserve interest because (i) from a computational point of view, it has the exact same order of complexity as that of Amaury Lambert and (ii) the general idea of this approach can actually be applied in a wider range of situations. More precisely, the approach of Amaury Lambert requires the property that the divergence times are independent an identically distributed (and their density) while my approach only requires the Markov property and to be able to compute the probability of a tree topology and that of observing N descendants from a single lineage after a given time (being able to factorize the tree probability is only required in order to obtain a quadratic computation). Please find below my detailed answers to yours comments.

Sincerely,
Gilles

1 Associate editor's comments

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;*

I failed to get rid of the combinatorial factors mainly because of the shift case.

- *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.*

Very good suggestion. I presented the approach under general diversification models and birth-death-sampling models. I added an appendix to explicitly show how it can be applied to birth-death models with time varying rates and past and extant sampling events.

- *using both simpler and more explicit notations;*

I changed some of the notations and added a table in order to help the reader (Appendix A).

- *relying a graphical example for explaining the intuition behind the quadratic recursive algorithm (e.g. continuing on the example given in figure 3).*

The quadratic computation is not easy to represent in a graphical way. I presented the idea of the algorithm in a less formal way in the main text and moved the proof of the theorem in Appendix C.

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 manuscript was re-organized in this sense. In particular, the technical proofs are now in Appendix C.

The English also needs improvement.

I tried to improve the text. Thanks again to Dominik Schrempf for kindly annotating the pdf.

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.

Since the alternative integration method, which has the same computational complexity as the method presented in the submission, is not yet published, I was embarrassed to discuss it in the text.

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 LambdaN with LambdaP is theoretically interesting, but not so useful in practice (since exact knowledge of divergence times, such as assumed by LambdaP, is lacking). In real-world applications, one would instead want to compare LambdaN with a plug-in version of LambdaP relying on an explicit dating of the tree obtained using relaxed clock approaches. In this context, a key question is whether LambdaN 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).

I do plan to improve the approach presented in Section 9, which is quite basic and just illustrates the interest of the computations presented in the paper. Actually, the null model of the revised version is slightly different from that of the initial submission.

About your point 1, both the time of origin and that of the diversification shift are assumed exactly known, which may sound unrealistic. It is possible to deal with uncertainty about the time of origin by integrating the probability of the special pattern of Theorem 3 with regard to the origin over a given interval of time. If one assume a uniform probability on the interval, there is an explicit formula for this integral (we did something like this in the fossil case in [4]). Though it may be possible to directly integrate with regard to the shift time over a given interval, I am rather thinking about plugging the computation of the probability of Theorem 3 into a Bayesian choice model framework in order to better explore both the diversification parameters and the shift time.

I am not sure to well understand your point 2. Do you mean that it should be more relevant to compute λ_P from the tree topology by estimating the divergence times from a (simulated) sequence alignment? I agree that the fact that comparing λ_N with λ_P is somewhat unfair since it uses only a small part of the information available for λ_P. I made the choice to evaluate all the methods by fullfilling all their assumptions about the diversification model and the information available. In this way, one can expect that their performance is somewhat optimal.

I added an empirical example based on the phylogeny of Cetaceans for which the new likelihood ratio λ_N detects (actually is maximal) at the same clade detected by a previous method (MEDUSA) by using all the diversification times.

2 Review of the anonymous reviewer

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.

In the definition of the piecewise-constant-birth-death-sampling model. We now read "s_k are the starting and ending times of the diversification process respectively and $s_1, \ldots, s_{k-1}$ are rate shift and mass extinction events times"

On page 6, the time τ_{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.

Sorry for that. The notation "τ" is now defined and I added "The temporal constraints induced by the tree topology, i.e., that we have $\tau_n \leq \tau_m$ if n is an ancestor of m, are implicitly assumed in the probability above."

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).

By construction the set of start-sets containing are the set of the subsets of internal nodes which contain b and satisfy the "start-set property". These two properties are both granted not only for $\{a, b\}$ but also for $\{a, b, c\}$, $\{a, b, c, d\}$, $\{a, b, c, d\}$ and $\{a, b, c, d\}$. I added a new figure to illustrate the definition of start-sets and start-trees.

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.

The number of rate shifts was on the algorithm in the initial version of the manuscript

On Section 7 onward, the author keeps referring to $P(\tau, ...|\tau)$. I think the τ on the left of the conditional should be removed. Also, corollary 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.

We actually consider the joint probability of the sets of constraints at the right of conditional and of the additional constraint $\tau_m < t$ in order to get the distribution of the divergence time of m at time t. Removing the set of constraints at the left would be misleading. I did not try to devise an algorithm to compute the density of divergence times rather than their distribution (I feel that it should be possible). The densities plotted in the manuscript are obtained by finite difference approximations. This point was already stated below Corollary 1 and is now also stated in the captions of Figures 5 and 6.

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).

You are perfectly right. This property is now pointed out in the revised manuscript. All the parameters are still provided for convenience for the reader.

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.

The computation of the joint density of the tree topology and all its divergence times is provided in Stadler (2011) ([29] in the manuscript). The caption of Figure 5 was ambiguous and was modified.

On page 15, Lemma ?? needs fixing.

Fixed!

Review of Amaury Lambert

1. The author notes in Section 7.1 that it seems not straightforward to use the independence of divergence times to compute the likelihood of the tree subject to temporal constraints. However, Equation (1) in the present report shows that this likelihood can be obtained by integrating an explicit, product density over a domain expressed as an intersection of half-spaces. In my opinion, it would be good to explain first why this direct integration is slower than the method proposed in the paper. Let me present hereafter an alternative method. To perform the integration, we can assume that the H_i 's are iid uniform in $(0,1)$ modulo replacing u_n by $F^{-1}(u_n)$, where $F(x) = P(H < x)$ is explicit (see again Lambert & Stadler 2013). Now for each node m and for any $x \in [0,T]$ denote by $Q_m(x)$ the probability

$$Q_m(x) := P(x < H_{a(n)} < H_n < u_n, n, a(n) \in T_m),$$

where T_m is the subtree rooted at m. Then Q_m is piecewise polynomial in x and can be computed symbolically and recursively, using

$$Q_m(x) = \int_x^{u_m} Q_{m_1}(y) Q_{m_1}(y) dy$$

where m_1 and m_2 are the two daughters of m (replace Q_{m_i} with 1 if there is no such daughter m_i). Finally compute $Q_r(0)$. Notice that taking all u's equal to T (no time constraint) yields the probability of the labelled tree shape T times the combinatorial constant c_L (equal to $2^{1-L}L!$ as seen in beginning of report).

Your alternative method is perfectly sound, and far more elegant than the one I proposed! Actually, my remark in Section 7.1 came from the fact I found not straightforward to extend the approach of Gehrard (2008, that which relies on the distribution of the k^{th} divergence time) to the case of multiple time constraints.

2. The proof of the complexity of the algorithm would gain from a slightly more rigorous induction reasoning. Please specify from the start that T is fixed and that the induction hypothesis (already well specified to be on Δ) is that the complexity is $O(\Delta|T|^2)$. I was originally misled believing that the induction hypothesis was that the complexity is $O(\Delta|T|^2)$ for any tree T and supposed to be applied to the smaller subtrees T_m . It might help to use the notation Θ^k (obvious if I define $\Theta^0 = \Theta$ and $\Theta^1 = \Theta^0$).

The proof of Theorem 4 was not well written. In particular the induction hypothesis was not clearly stated, it is actually that computing the probabilities for all the subtrees is $O(\Delta|T|^2)$. I rewrote this part.

3. The modification of the algorithm I propose avoids computing binomial coefficients related to distinct labellings, by relying on the fact that the tree shape T is labelled from the start. It also enables rates to be time-dependent in a general way, not only piecewise constant, and so avoids hashing the algorithm in as many pieces as there are intervals where the rates are constant. Last, it saves a lot of algebraic formulae, by expressing nearly everything in terms of the rv H. I am not sure the gain is really worth changing the framework in terms of computational efficiency, but I think this alternative framework gives a condensed way of explaining the method. I am adding this comment because I found the exposition of the method in the paper very technical and difficult to follow and I would welcome any solution helping the reader to grasp rapidly the main ideas of the method.

The method that you propose is more than a modification of the algorithm presented in the manuscript. It is a whole different approach.

If I am not wrong, the complexity order of your method is exactly the same as that proposed in the manuscript since all polynomials Q_n have degree $L_{T_n} - 1$ (thus L_{T_n} coefficients) and are defined with about Δ pieces, except that Δ does not include the change times of the model in your case.

The revised version deals with time-dependent rates and without hashing the algorithm with regard to the intervals of the model. I feel that the new Appendix B suggests that in all cases where one has an explicit formula for the distribution of the H's and the iid property, it should be possible to get explicit formulae for the ending configurations of the patterns used in my computation, thus to apply the whole method.

Anyway, I agree that your method is more elegant and may be more efficient from a computational point of view (only up a constant factor though). The only advantage of my approach is that since it does not require the "independence property", its general idea may be possibly used in a wider range of situations (e.g., with the fossilized birth death model or with models when the iid property is not granted).

4. On a more general note, I am not sure a standard biology journal (other than journals devoted to 'mathematical biology') would accept a paper where so much emphasis is put on the technicalities of a method; I wonder how much PCI Evol Biol is immune to this tradition.

The aim of the manuscript is to present the general idea of the computation and to show that it can be used in various situations.

Review of Dominik Schrempf

I wish to thank you for your corrections and your suggestions in the annotated pdf. As you suggested in on of your comments, I added a figure in order to illustrate the start sets. Please find below my answers to your other comments.

Abstract, page 1, page 4: size of phylogeny. Could you please be more specific and move the definition of size from page 4 before the first use of this term?

"Size of phylogeny" is not mentioned in the revised abstract, which now refers to "number of species", which is vague but true whatever any possible interpretation, instead. In the introduction, the first occurrence of "size of the phylogeny" is now followed by "(i.e., its total number of nodes)".

Abstract: divergence time. When reading the manuscript for the first time, the exact meaning of this term was not clear to me. It could refer to the divergence time between 2, or any number of leaves on a tree (possibly also the height of a tree). It could also refer to the divergence time between inner nodes on a tree. It may be good to be more specific. Would it be precise to just state that the divergence times are the branch lengths of the tree?

I added "(i.e., the times of the speciation events corresponding to the internal nodes of a phylogenetic tree)" at the beginning of the revision in order to make the term less ambiguous.

Abstract and Page 2: exact divergence time distributions. Exact sounds very strong in this context. Do you mean exact when assuming the piecewise-constant, birth, death, and sampling model?

"Exact" was for emphasizing that the distributions are actually computed and not estimated by sampling the divergence times. It was not very useful and was removed

Chapter 2, first paragraph: I am confused about the meaning of ρ_i. According to Tanja Stadlers paper [29], it is the survival probability when not at the present, and the sampling probability when at the present. Could you please explain the meaning of ρ_i in more detail here? Especially, the sentence: "The samplings of ancestral lineages .. are interpreted as extinction events .." confuses me. How can ancestral lineages be extinct?

"Ancestral" was misleading. I rewrote the explanation of the sampling as "Sampling lineages at a time s_i anterior to the present time has to be interpreted as a mass extinction events while sampling at the present time accounts for our incomplete knowledge of extant species".

Chapter 3, first sentence: I do not understand this sentence. Trees obtained from the piecewise constant birth, death and sampling process are rooted and binary, but I can think of diversification processes that do not yield binary trees.

The "binary" limitation comes from the models considered in which speciations always result in two lineages and exclude polytomies. I added "(since a speciation event gives rise to a single new lineage under the models considered here)". Dealing with polytomies requires to use different models.

Section 3.1.: Why does the reconstructed birth-death-sampling process not have extinction? Do I misunderstand something here?

Since the reconstructed phylogenetic tree is obtained from the sampled extant taxa only, extinct taxa cannot be observed or reconstructed.

Chapter 4; Figure 2: Could you please explain what a special lineage is before referencing Figure 2?

We now read " A *pattern* is a part of the observed diversification process starting from a single lineage at a given time and ending with a certain number of lineages at another given time, these ending lineages being either observable or *special*, where "special" means "known to be alive at the ending time and distinguished for some reason" " at the beginning of Section 4.

Section 5.1: To me it is not clear what the events $\tau_{n_i} < u_i$ are. As far as I can see, we restrict the node (event?) n_i , to be younger than u_i . Isn't it clearer to write: node ages τ_{n_i} , instead of events?

"Event" was to be taken in the probabilistic sense. Since it can be confusing with the more restricted sense of "event of the diversification process", I avoided to use it in the revision.

Section 5.1: At this point, I also got confused about the nomenclature (which is very consistent but involves many different, new symbols). Let me summarize:

- *Bold symbols denote probabilities. Especially, the symbol* $\mathbf{P}$ *denotes "probability of".*
- *The symbol* $\mathbf{T}$ *for instance, is just* $\mathbf{P}(\mathcal{T}|n)$*, where n are the number of tips of the topology* $\mathcal{T}$*.*
- *Subscripts denote model parameters (mostly* Θ*). Why are the times* u_i*, and* u_i' *not part of* Θ*?*

This is well the logic behind the notations. The revision contains a table of the symbols and notations used with brief descriptions in order to help the reader (Appendix A).

Theorem 2: So the time constraints are not part of the model?

No. The model accounts for how the diversification process works. The time constraints are specific to a phylogenetic tree which is assumed to be a realization of the diversification process. For instance the same model may apply to two clades living at the same period, each one with specific time constraints on their nodes.

Theorem 2: "$\Omega_{\mathcal{T},n}$ if $s_1 = o$". What is $\Omega_{\mathcal{T},n}$? It does not form part of the definitions from before. I guess you mean $\Omega_{\mathcal{T}}$? What if $s_1 \neq o$?

Thanks for seing this typo. I meant well $\Omega_{\mathcal{T}}$. Your second question is good and makes me realize that the definition of $\mathcal{S}$ was ambiguous. The corresponding part of Theorem 2 was rewritten.

Theorem 2: "$s'_{k'} = s_k$". Why introduce a new symbol, when it is just the same as a symbol that was already introduced before?

It was completely useless (and now removed)!

Theorem 2: "Θ'". Let the earliest time constraint be at time u. Is it correct, that if $u < s_1$, just another slice is introduced at u? Can you please describe in words what is being done here?

In the case where $u < s_1$, the model Θ' is obtained from Θ by replacing its starting time s_0 by u. Otherwise, Θ' is obtained from Θ by discarding its first slice. The revised version deal with piecewise constant models in a different way.

Proof of Theorem 2: "being a divergence time assignation of $\mathcal{T}$": I do not understand this sentence.

It was replaced by "by assuming that the divergence times of $\mathcal{T}$ are consistent with the temporal constraints", which is less ambiguous.

Proof of Theorem 3: "The set of subsets of internal nodes with divergence time anterior to t, and consistent with the assumptions of the Theorem is thus exactly $\Omega^{\times}_{\mathcal{T},m}$". $\Omega^{\times}_{\mathcal{T},m}$ is the set of all start-sets A of $\mathcal{T}$ such that m is a tip of $\Gamma_{\mathcal{T},A}$. I argue that $\Omega^{\times}_{\mathcal{T},m}$ is the "set of subsets of internal nodes with divergence time anterior to t, and consistent with the assumptions of the Theorem" together with start-sets including nodes in $\mathcal{T}_m$.

By re-reading it, I found the beginning of the proof confusing and replaced the second sentence by "Reciprocally, if the divergence times of m and of its direct ancestor are respectively posterior and anterior to t, then a diversification shift at time t may occur for the clade originating at m." The point here is that if the assumptions of the theorem are granted then the set of subsets of nodes with divergence times anterior to t is $\Omega^{\times}_{\mathcal{T},m}$ and that these assumptions are granted only if the set of subsets of nodes with divergence times anterior to t is $\Omega^{\times}_{\mathcal{T},m}$.

Figure 7 and paragraph afterwards: The abbreviation receiver operating characteristic (ROC) was not defined.

Fixed.