Report on "Probabilities of tree topologies with temporal constraints and diversification shifts" by Gilles Didier.

Referee's summary. In the context of macro-evolution, a popular approach in the last ten to twenty years has consisted in extracting the signal contained in the species phylogeny to infer the process of diversification that has generated it. The standard way of dealing with this question is to use likelihood methods under stochastic models of lineage-based diversification, also called birth-death models, where species are seen as elementary particles that can split (speciate) or die (become extinct). The main difficulty lies in computing the likelihood of the 'reconstructed tree' under birth-death models, that is the genealogy of species that are extant and sampled, to the exclusion of unknown or extinct species. In the case when the birth and death rates are only time-dependent and species are sampled at the present independently with a fixed probability, the likelihood of the reconstructed tree is explicit, and in product form, since the divergence times of the reconstructed tree are actually independent and identically distributed.

In this paper, the author is interested in computing the likelihood of the topology of the tree, also called tree shape, without the information on the precise dates of the nodes, but only bounds on these dates. This question is particularly interesting when one is not interested in the precise datation but only in the phylogenetic relations between species when the node dates are subject to bounds independently inferred from the fossil record.

The paper is devoted to the exposition of the mathematical method, with one application to the Hominoid tree. The method is recursive. Since the exposition is quite technical, I will try to rephrase it with a more basic terminology, which in passing also provides a slightly more efficient computing procedure.

Let $\mathcal{T}$ denote the labelled tree shape with $L = |L_{\mathcal{T}}|$ leaves that we are interested in. For any node n of $\mathcal{T}$, denote by $a(n)$ the mother node of n and by $H_n \in (0, T)$ the divergence time of n (taken equal to T if n is a leaf). Here the time is measured from 0 (crown) to T (present time). In particular, the likelihood of the labelled tree shape $\mathcal{T}$ is given by

$$\mathbb{P}(\mathcal{T}) = \frac{b_L}{c_L}\,\mathbb{P}(H_{a(n)} < H_n, n \in \mathcal{T}),$$

where $H_{a(n)} := 0$ if n is the root, $\mathbb{P}(H_{a(n)} < H_n, n \in \mathcal{T})$ is the probability of a(ny) plane orientation of the unlabelled tree shape conditioned on its number of leaves, which *does not depend on the law of* H (as earlier said the (H_n), where n ranges over internal nodes are iid random variables), b_L is the probability that the birth-death-sampling process has L sampled descendants at time T and c_L is a combinatorial constant equal to $2^{1-L}L!$ as shown in the following argument. Let τ be the tree shape obtained from $\mathcal{T}$ by forgetting its labels and let $s(\tau)$ be the number of symmetric nodes of τ (e.g., cherries). Then by using the coalescent point process planar view of reconstructed trees of birth-death processes (Lambert & Stadler *TPB* 2013, Lambert *Brazil. J. Probab. Statist.* 2017),

$$c_L = \frac{\#\text{ distinct labellings of } \tau}{\#\text{ distinct plane orientations of } \tau} = \frac{L!\,2^{-s(\tau)}}{2^{L-1-s(\tau)}} = 2^{1-L}L!$$

Now assume for simplicity that the temporal constraints are of the form $H_n < u_n$, for all $n \in \mathcal{T}$ (with u_n possibly equal to T so that the constraint is not really one) and that the

nodes have been labelled $n = 1, \dots, L$ in such a way that $u_1 \leq \cdots \leq u_L$. Define the event $\mathcal{U} := \{H_n < u_n, n = 1, \dots, L\}$. What we are interested in is the probability of the labelled tree shape **with temporal constraints**

$$\mathbb{P}(\mathcal{T}, \mathcal{U}) = \frac{b_L}{c_L} \mathbb{P}(H_{a(n)} < H_n < u_n, n \in \mathcal{T}). \tag{1}$$

The recursive procedure relies on the following decomposition

$$\mathbb{P}(H_{a(n)} < H_n < u_n, n \in \mathcal{T}) \\ = \sum_{\mathcal{A} \in \Upsilon} \mathbb{P}(H_{a(n)} < H_n < u_1, n \in \mathcal{A}) \, \mathbb{P}(u_1 < H_{a(n)} < H_n < u_n, a(n) \notin \mathcal{A}),$$

where Υ is the set of so-called *start-sets*, that is, sets $\mathcal{A}$ of internal nodes of $\mathcal{T}$ such that $1 \in \mathcal{A}$ and: $n \in \mathcal{A} \Rightarrow a(n) \in \mathcal{A}$. The first term equals $\mathbb{P}(H < u_1)^{|\mathcal{A}|}$ times the probability of the (plane-oriented) tree shape $\mathcal{A}$ unlabelled and conditioned on its number of leaves, which depends neither on the law of H nor on u_1. The second term in the sum can be factorized as a product over (maximal) subtrees with roots $n \notin A$, and all terms of this product are of the form $\mathbb{P}(u_1 < H_{a(n)} < H_n < u_n, n \in \mathcal{T}')$ for some subtree $\mathcal{T}'$ with number of leaves strictly less than L, hence the recursivity of the procedure.

The problem with this decomposition is that the cardinal of Υ is possibly exponential in L (depending on the shape of $\mathcal{T}$). A more efficient factorization is as follows

$$\mathbb{P}(H_{a(n)} < H_n < u_n, n \in \mathcal{T}) = \sum_{k=1}^{L-1} V_k(\mathcal{T}),$$

where

$$V_k(\mathcal{T}) = \sum_{\mathcal{A} \in \Upsilon : |L_{\mathcal{A}}| = k} \mathbb{P}(H_{a(n)} < H_n < u_1, n \in \mathcal{A}) \, \mathbb{P}(u_1 < H_{a(n)} < H_n < u_n, a(n) \notin \mathcal{A}) \tag{2}$$

Now let b_1 and b_2 we the two daughters of the root r of $\mathcal{T}$ and let $\mathcal{B}_i$ denote the subtree of $\mathcal{T}$ rooted at b_i. For each start set $\mathcal{A} \in \Upsilon$ such that $|L_{\mathcal{A}}| = k$, $\mathcal{A} \cap \mathcal{B}_i$ is a (possibly empty, if $b_i \notin \mathcal{A}$) subtree of $\mathcal{A}$ with root b_i, for $i = 1, 2$. Then

$$\{H_{a(n)} < H_n < u_1, n \in \mathcal{A}\} = \{H_r < H_n < u_1, n \in \mathcal{A} \backslash \{r\}\} \cap \bigcap_{i=1}^{2} \{H_{a(n)} < H_n < u_1, a(n), n \in \mathcal{A} \cap \mathcal{B}_i\}$$

so that

$$\mathbb{P}(H_{a(n)} < H_n < u_1, n \in \mathcal{A}) = \frac{\mathbb{P}(H < u_1)}{k - 1} \prod_{i=1}^{2} \mathbb{P}(H_{a(n)} < H_n < u_1, a(n), n \in \mathcal{A} \cap \mathcal{B}_i),$$

where each probability in the product is taken equal to 1 whenever its subtree is empty. Then by factorizing $\mathbb{P}(u_1 < H_{a(n)} < H_n < u_n, a(n) \notin \mathcal{A})$ into the subtrees not intersecting $\mathcal{A}$ descending from b_1 vs. from b_2, one easily gets the following analogue to Eq (5) in the manuscript

$$V_k = \frac{\mathbb{P}(H < u_1)}{k - 1} \sum_{i=0}^{k} V_i(\mathcal{B}_1) \, V_{k-i}(\mathcal{B}_2), \tag{3}$$

where $V_0(\mathcal{B}_i) = \mathbb{P}(u_1 < H_{a(n)} < H_n < u_n, a(n) \in \mathcal{B}_i)$, which is basically $\mathbb{P}(H > u_1)^{|\mathcal{B}_i|}$ times the probability of the (plane-oriented, unlabelled) tree shape $\mathcal{B}_i$ **with one less temporal constraint**. This last property allows the author to propose a recursive procedure to compute the V_k's (called W_k in the manuscript, where the combinatorial information on labelling has to be dealt with additionally) and finally the probability of the tree with temporal constraints.

The author shows in addition that this algorithm has complexity $O(\Delta|\mathcal{T}|^2)$, where Δ denotes the number of effective temporal constraints.

Opinion and main comments. A method rigorously computing the likelihood of a birth-death-sampled tree with constraints on node dates was, I believe, strongly needed. The method proposed here is rigorous and contains very good ideas applied to reduce its computational cost. Also, the author shows very nicely how his method can be applied to the direct sampling of divergence times in a phylogeny subject to temporal constraints, and to testing the presence of diversification shifts. I have four comments.

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 $F(x) = \mathbb{P}(H < x)$ 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

   $$Q_m(x) := \mathbb{P}(x < H_{a(n)} < H_n < u_n, n, a(n) \in \mathcal{T}_m),$$

   where $\mathcal{T}_m$ is the subtree rooted at m. Then Q_m is piecewise polynomial in x and can be computed symbolically and recursively, using

   $$Q_m(x) = \int_x^{u_m} Q_{m_1}(y)\, Q_{m_2}(y)\, dy,$$

   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 $\mathcal{T}$ times the combinatorial constant c_L (equal to $2^{1-L}L!$ as seen in beginning of report).

2. The proof of the complexity of the algorithm would gain from a slightly more rigorous induction reasoning. Please specify from the start that $\mathcal{T}$ is fixed and that the induction hypothesis (already well specified to be on Δ) is that the complexity is $O(\Delta|\mathcal{T}|^2)$. I was originally misled believing that the induction hypothesis was that the complexity is $O(\Delta|\mathcal{T}|^2)$ for any tree $\mathcal{T}$ and supposed to be applied to the smaller subtrees $\mathcal{T}_m$. It might help to use the notation Θ^k (obvious if I define $\Theta^0 = \Theta$ and $\Theta^1 = \Theta'$).

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 $\mathcal{T}$ 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.

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.

(Signed: Amaury Lambert)