Appendices

1 Table of the notations

λ	cladogenetic (speciation) rate
μ	extinction rate
ψ	fossil find rate
θ	(λ, μ, ψ), the vector of parameters
ρ	sampling rate of extant taxa
α, β and ω	3 quantities depending on the rates convenient for expressing formula
$\mathbf{P}_\theta(n,t)$	probability of ending with n lineages at time t by starting with a single lineage at time 0 and without fossil meanwhile
$\mathbf{P}_{\mathbf{o},\theta}(t)$	probability for a lineage present a time 0 to be observable at t
$\mathbf{P}_\tau(\mathcal{S} \mid n)$	probability of the tree topology $\mathcal{S}$ conditioned on having n leaves
$\mathbf{P}_{\mathrm{x},\theta}(k+1,t,t')$	probability density of a fossil find at t' on a lineage, in addition to k other lineages observable at time t' by starting with single lineage at t and without any other fossils meanwhile
$\mathbf{P}_{\mathrm{y},\theta}(k,t,t')$	probability of ending with k lineages observable at t' by starting from a single lineage at t and without fossil meanwhile
$\mathbf{P}_{\mathbf{a},\theta}(\mathcal{E})$ (resp. $\mathbf{P}_{\mathbf{b},\theta}(\mathcal{E})$, $\mathbf{P}_{\mathbf{c},\theta}(\mathcal{E})$)	probability (density) of the pattern $\mathcal{E}$ of type $\mathbf{a}$ (resp. $\mathbf{b}$, $\mathbf{c}$)

2 Probability of a tree topology

Number of birth rankings consistent with a tree topology

We are interested here in tree topologies (i.e. trees without time information) resulting from realizations of general birth and death processes, in which no deaths occur, i.e. the process is not necessarily a pure-birth process but we consider only realizations in which only births occur. Most of the ideas of this section are close to those developed in Ford *et al.* (2009).

To keep things as general as possible, we define a *(pure-birth) realization* of n lineages with birth times $t_1 < t_2 < \ldots < t_{n-1}$ in the following way. The process starts with a single lineage at a time $t_0 < t_1$. At time t_i, a lineage is picked among the lineages alive to give birth to a new lineage. Each lineage is associated to a different label in an arbitrary way (i.e. not depending on its birth date, its parenthood etc.). Remark that, since by convention, we consider only realizations without death, all the lineages live until the end of the process. In particular, a lineage is still alive after having given birth to a new one. The natural (and usual) way to associate a tree topology with a realization is as follows:

- the internal nodes and the leaves of the tree are in one-to-one correspondence with the birth events and the lineages of the realization, respectively;
- the direct ancestor of the leaf associated with the lineage x is the node corresponding to the last birth event involving x, which may be either its own birth, or the last time it gave birth to another lineage;
- the direct ancestor of the internal node associated with the birth of the lineage x is the node corresponding to the last event before the birth of x that involves the parent of x.

By construction, the trees resulting from realizations are rooted, binary and (leaf) labeled (labels of leaves are those of the lineages).

The *scenario* of a realization is its sequence of birth events ordered following their occurrence times. The i^{th} event of a scenario E is noted E_i and is of the form "lineage x is borne from lineage y", x and y being referred to as the child and the parent lineages of E_i, respectively. A given scenario is *valid* if there exists a realization from which it arises. Basically a scenario is valid if and only if all its lineages but the starting one are the child lineages of the earliest event involving them.

Let us point out a major conceptual difference between species as we conceptualize them here and lineages as actors of the (modeled) diversification process. From the phylogenetic perspective, each branch

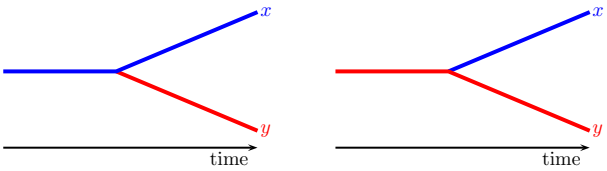

Figure 1: The two possible scenarios leading to a tree with two leaves. Note that this scenario involves two lineages, but three species as we conceptualize them here (the unlabeled species corresponds with the part of lineage that includes the stem of x and y under both scenarios).

of the tree is associated to different species and a speciation event corresponds to the pseudo-extinction of the ancestral species that gives birth to two new species. By contrast, birth and death processes model diversification and not evolution, in the sense that they do not take into account genotypic or phenotypic changes, and more importantly, no pseudo-extinction is involved in this diversification process, given that a lineage is considered to continue after it gives rise to another lineage (Fig. 1). Thus, the delimitation of species differs from that of lineages in this framework. Birth and death processes thus deal with lineages rather than with species. This explains why several evolutionary scenarios may lead to a same tree. For instance, a tree made of only two lineages x and y may result from both the scenarios "y was borne from x" (Fig. 1-left) or "x was borne from y" (Fig. 1-right). Conversely, the scenario of a realization fully determines its tree in the following way. If a tree topology $\mathcal{S}$ and a scenario E both result from a same realization, the internal nodes and the leaves of $\mathcal{S}$ are in one-to-one correspondence with the events and the lineages of E respectively. The direct ancestor of the leaf associated with the lineage x is the internal node corresponding to the event with the greatest rank in E involving x, while the direct ancestor of the node associated with the event E_i is the node corresponding to the event with the greatest rank strictly smaller than i which involves the parent lineage of E_i.

Let $\mathbf{r}$ denote the birth ranking of the lineages of a realization (i.e. $\mathbf{r}$ is the vector in which the i^{th} entry $\mathbf{r}_i$ contains the i^{th} oldest lineage). A scenario E perfectly determines the birth ranking of its lineages: the starting lineage has rank 1 while the ranks of the other ones are obtained by adding 1 to the ranks of their birth events in E.

Remark 1. *If a tree topology and a scenario come from a same realization, then the node n associated with the event "x is borne from y" is such that y and x are respectively the oldest and the second oldest lineages/leaves of the subtree rooted at n. It follows that the given of both the tree and the birth ranking of the lineages resulting from a realization is sufficient to reconstruct its scenario.*

In short, the scenario of a realization fully determines both its tree topology and the birth ranking of its lineages, while the given of both the tree topology and the ranking of a realization determines its scenario. A tree and a birth ranking are *consistent* with one another if there exists a valid scenario corresponding to both of them.

With Remark 1, the number of scenarios leading to a given tree $\mathcal{S}$ is equal to the number $\mathbf{R}_{\mathcal{T}}$ of birth rankings consistent with $\mathcal{S}$. This number depends on the tree considered. It may actually differ between two trees with the same number of leaves.

Lemma 1. *Let $\mathcal{S}$ be a rooted binary labeled tree topology, and $\mathcal{S}_l$ and $\mathcal{S}_r$ be the two subtree topologies rooted at the children of its root. A birth ranking $\mathbf{r}$ of the lineages of $\mathcal{S}$ is consistent with $\mathcal{S}$ if and only if:*

1. *the two oldest lineages of $\mathcal{S}$ are the oldest lineage of $\mathcal{S}_l$ and of $\mathcal{S}_r$,*
2. *$\mathbf{r}^{(l)}$, the restriction of $\mathbf{r}$ to the lineages of $\mathcal{S}_l$, is consistent with $\mathcal{S}_l$,*
3. *$\mathbf{r}^{(r)}$, the restriction of $\mathbf{r}$ to the lineages of $\mathcal{S}_r$, is consistent with $\mathcal{S}_r$.*

Proof. Let us first assume that $\mathbf{r}$ is consistent with $\mathcal{S}$. There exists a scenario E leading to $\mathcal{S}$ in which the i^{th} event is the birth of the lineage of rank $(i+1)$. In particular, its first event is the birth of the second oldest lineage of $\mathcal{S}$ (the oldest one starts the process). The first event corresponds to the root node of $\mathcal{S}$, which thus involves the two oldest lineages and splits $\mathcal{S}$ into the subtree containing the second oldest lineage and all its descendants and the subtree containing the oldest lineage and all its

descendants except that of the second oldest one and the second oldest one itself. It follows that the two oldest lineages of $\mathcal{S}$ are the oldest lineage of $\mathcal{S}_l$ and the oldest one of $\mathcal{S}_r$. Let $E^{(l)}$ be the scenario obtained from E by discarding its first event and all the events not involving a lineage of $\mathcal{S}_l$. Basically the tree $\mathcal{S}_l$ follows the sequence of events of $E^{(l)}$ and the corresponding birth ranking is the restriction of $\mathbf{r}$ to the lineages of $\mathcal{S}_l$. The same holds for $\mathcal{S}_r$.

Reciprocally, let $\mathbf{r}^{(l)}$ and $\mathbf{r}^{(r)}$, two birth rankings consistent with $\mathcal{S}_l$ and $\mathcal{S}_r$ respectively and $\mathbf{r}$ be obtained by merging $\mathbf{r}^{(l)}$ and $\mathbf{r}^{(r)}$ in such a way that the two first lineages of $\mathbf{r}$ are chosen among $\mathbf{r}_1^{(l)}$ and $\mathbf{r}_1^{(r)}$. There exist two scenarios $E^{(l)}$ and $E^{(r)}$ leading to the pair $(\mathbf{r}^{(l)}, \mathcal{S}_l)$ and the pair $(\mathbf{r}^{(r)}, \mathcal{S}_r)$ respectively. Let now E be the scenario where the first event is "$\mathbf{r}_2$ borne from $\mathbf{r}_1$" and, for all $i > 1$, the event E_i is the birth event of the lineage $\mathbf{r}_{i+1}$, which belongs either to $E^{(l)}$ or to $E^{(r)}$. Since the scenarios $E^{(l)}$ and $E^{(r)}$ are valid, E is valid and determines both the tree $\mathcal{S}$ and the birth ranking $\mathbf{r}$. □

Theorem 1. *The number of birth rankings consistent with a rooted binary labeled tree topology $\mathcal{S}$ is*

$$\mathbf{R}_{\mathcal{S}} = \mathbf{R}_{\mathcal{S}_l}\mathbf{R}_{\mathcal{S}_r} 2 \binom{\mathbf{L}_{\mathcal{S}_l} + \mathbf{L}_{\mathcal{S}_r} - 2}{\mathbf{L}_{\mathcal{S}_l} - 1}$$

where $\mathcal{S}_l$, $\mathcal{S}_r$, $\mathbf{L}_{\mathcal{S}_l}$ and $\mathbf{L}_{\mathcal{S}_r}$ are the two subtree topologies rooted at the children of the root of $\mathcal{S}$ and their numbers of leaves/lineages respectively.

Proof. From Lemma 1, there are as many rankings consistent with $\mathcal{S}$ as ways of merging a ranking of $\mathcal{S}_l$ with one of $\mathcal{S}_r$, by taking the two first lineages among $\mathbf{r}_1^{(l)}$ and $\mathbf{r}_1^{(r)}$. There are:

- $\mathbf{R}_{\mathcal{S}_l}$ rankings consistent with $\mathcal{S}_l$,
- $\mathbf{R}_{\mathcal{S}_r}$ rankings consistent with $\mathcal{S}_r$,
- 2 ways of setting the two first lineages of a ranking of $\mathcal{S}$ among $\mathbf{r}_1^{(l)}$ and $\mathbf{r}_1^{(r)}$,
- $\binom{\mathbf{L}_{\mathcal{S}_l}-1}{\mathbf{L}_{\mathcal{S}_l}+\mathbf{L}_{\mathcal{S}_r}-2}$ ways of merging the lineages of $\mathbf{r}^{(l)}$ and $\mathbf{r}^{(r)}$ except for the two oldest ones (such a merging in fully determined by the ranks occupied by the lineages of $\mathbf{r}^{(l)}$).

All these possibilities may be combined independently to give a ranking consistent with $\mathcal{S}$. □

Since the number of rankings consistent with the tree made of a single lineage is 1, Theorem 1 provides a recursive way to compute $\mathbf{R}_{\mathcal{S}}$ for any tree topology $\mathcal{S}$.

Probability of a tree topology given its number of leaves

Since the labeling of the leaves/lineages is arbitrary (i.e. depends neither on the tree topology nor on their birth ranks), the following remark follows by symmetry.

Remark 2. *In a realization with n lineages arbitrarily labeled, all the birth rankings of the (labeled) lineages have equal probability. Since there are $n!$ possible rankings, this probability is $\frac{1}{n!}$.*

Until here, we made no assumptions about the realizations or about the processes leading to tree topologies. From now on, we consider only tree topologies arising from pure-birth realizations (i.e. realizations of general birth and death processes in which no death occurs). Moreover, we focus on a large class of processes, which contains the usual diversification models. A process is said *lineage-homogeneous* if, at each event time, all the lineages give birth at a same rate. Such models are called *ERM models* in (Ford *et al.*, 2009).

Lemma 2. *Being given the birth ranking of a pure-birth realization of a lineage-homogeneous process, all the tree topologies have probability $\frac{1}{(n-1)!}$.*

Proof. Since the realization contains no death and the process is lineage-homogeneous, the i^{th} lineage is borne from any of the $(i-1)$ lineages alive at its birth date with equal probability $\frac{1}{i-1}$, independently of the other events. It follows that, being given the birth ranking, the joint probability of the parenthood of all the lineages is $\frac{1}{(n-1)!}$. □

Theorem 2. *A tree topology $\mathcal{S}$ resulting from a pure-birth realization of a lineage-homogeneous process has probability*

$$\mathbf{P}_\tau(\mathcal{S} \mid \mathbf{L}_\mathcal{S}) = \frac{\mathbf{R}_\mathcal{S}}{(\mathbf{L}_\mathcal{S}-1)!\,\mathbf{L}_\mathcal{S}!}$$

conditioned on having $\mathbf{L}_\mathcal{S}$ leaves.

Proof. From Remark 2 and Lemma 2, the joint probability of a pair tree/ranking is $\frac{1}{(n-1)!n!}$. To obtain the probability of a tree $\mathcal{S}$ with n leaves, we just have to sum these joint probabilities over all the rankings consistent with $\mathcal{S}$, which gives us the result. □

3 Complexity index of a tree

The complexity of Algorithm 1 relies on the number of possible before/after assignments of the basic trees encountered during its execution (see the proof of Theorem 1). Let us put $\mathcal{A}_\mathcal{B}$ for the number of before/after assignments of a basic tree $\mathcal{B}$ with regard to a time t between those of the origin and of the oldest leaf of $\mathcal{B}$ (i.e. any internal node of $\mathcal{B}$, including its root, corresponds to a divergence date that is possibly anterior or posterior to t). Let $\mathcal{B}_l$ and $\mathcal{B}_r$ be the subtrees pending to the children of the root of $\mathcal{B}$. Any before/after assignment of $\mathcal{B}$ in which the root is set to "before" (time t), is obtained in a unique way by combining an assignment of $\mathcal{B}_l$ with one of $\mathcal{B}_r$. There is only one possible assignment of $\mathcal{B}$ in which the root is set to "after" (time t). It follows that we have

$$\mathcal{A}_\mathcal{B} = \mathcal{A}_{\mathcal{B}_l}\mathcal{A}_{\mathcal{B}_r} + 1$$

The number of before/after assignments of a basic tree is recursively computed (the tree made of a single leaf has a unique before/after assignment).

The *complexity index* of a tree $\mathcal{T}$ is mainly obtained by summing the number of possible before/after assignments of all the basic trees that have to be considered to compute the likelihood of $\mathcal{T}$. For technical reasons, we actually consider an additional term that is very similar to the number of assignments. Though it can certainly be improved, the complexity index predicts quite well the duration of a likelihood computation (see the help of *Diversification*, `https://github.com/gilles-didier/Diversification`).

4 Sampling extant taxa

Following Stadler (2010), we assume here that each extant taxon is independently discovered (or sampled) in the present with a certain probability ρ. Let us define $\mathbf{P}_{\rho,\theta}(n,t)$ as the probability of sampling $n > 0$ lineages at time t with the probability ρ, by starting from a single lineage at time 0 without any fossils dated between 0 and t, under the rates $\theta = (\lambda, \mu, \psi)$. The probabilities $\mathbf{P}_{\rho,\theta}(0,t)$ and $\mathbf{P}_{\rho,\theta}(1,t)$ were already provided in Stadler (2010).

For all positive integers n, we have

$$\begin{aligned}\mathbf{P}_{\rho,\theta}(n,t) &= \sum_{j=0}^{\infty}\binom{j+n}{n}\rho^n\,(1-\rho)^j\,\mathbf{P}_\theta(j+n,t)\\ &= \frac{(\beta-\alpha)^2\mathrm{e}^{\omega t}\rho^n\,(1-\mathrm{e}^{\omega t})^{n-1}}{(\beta-\alpha\mathrm{e}^{\omega t}-(1-\rho)(1-\mathrm{e}^{\omega t}))^{n+1}}\end{aligned}$$

The probability of sampling no lineage at t, still by starting from a single lineage at time 0 without any fossils dated between 0 and t, is

$$\begin{aligned}\mathbf{P}_{\rho,\theta}(0,t) &= \mathbf{P}_\theta(0,t) + \sum_{j=1}^{\infty}(1-\rho)^j\,\mathbf{P}_\theta(j,t)\\ &= \frac{\alpha\beta(1-\mathrm{e}^{\omega t})+(1-\rho)(\beta\mathrm{e}^{\omega t}-\alpha)}{\beta-\alpha\mathrm{e}^{\omega t}-(1-\rho)(1-\mathrm{e}^{\omega t})}\end{aligned}$$

The probabilities $\mathbf{P}_{\rho,\theta}(0,t)$ and $\mathbf{P}_{\rho,\theta}(1,t)$ are equal to $p_0(t)$ and $p_1(t)$ of Theorem 3.1 in Stadler (2010), which refer to the same probabilities but which are computed and expressed in a slightly different way.

The likelihood of a reconstructed tree with fossils and extant taxa sampled with the probability ρ may be computed in a similar way as under the assumption that all the extant taxa are known. One just needs to replace the probabilities of the form $\mathbf{P}_\theta(n, T-t)$ by $\mathbf{P}_{\rho,\theta}(n, T-t)$ in the calculus. Further tests have to be carried out to check to what extent the sampling probability influences the estimation of the diversification and fossilization rates and in what extent it can be estimated itself.

5 Proportion of lineages unobservable from the fossil record

Let us put $\mathbf{P}_{\oslash,\theta}$ for the probability for a lineage to leave no fossil, neither of itself, nor of its descendant in the hypothetical situation where the diversification process continues indefinitely. Assuming this hypothetical situation is essentially the same as considering that we are dealing with a lineage present at a time arbitrarily far from the present. We have that:

$$\mathbf{P}_{\oslash,\theta} = \frac{\mu}{\lambda+\mu+\psi} + \frac{\lambda}{\lambda+\mu+\psi}\mathbf{P}^2_{\oslash,\theta}$$

The first term of the right-hand side of the expression just above is the probability that the first event occurring on the lineage after its birth is an extinction. The second one is the probability that this event is a speciation giving birth to two lineages that left no fossils.

The preceding equation can be written as

$$\lambda\mathbf{P}^2_{\oslash,\theta} - (\lambda+\mu+\psi)\mathbf{P}_{\oslash,\theta} + \mu = 0$$

and was already considered in the section *Probability of ending with n lineages without observing fossils – Type* **a** and Didier *et al.* (2012).

If $\lambda = 0$, the unique solution of the equation just above is $\mathbf{P}_{\oslash,\theta} = \frac{\mu}{\mu+\psi}$, that is the probability that the first event occurring on the lineage is an extinction (there cannot be any speciation/birth since $\lambda = 0$). Otherwise, its roots are

$$\alpha = \frac{\lambda+\mu+\psi-\sqrt{(\lambda+\mu+\psi)^2-4\lambda\mu}}{2\lambda} \text{ and } \beta = \frac{\lambda+\mu+\psi+\sqrt{(\lambda+\mu+\psi)^2-4\lambda\mu}}{2\lambda}$$

By noting that

$$(\lambda+\mu+\psi)^2 - 4\lambda\mu = (\lambda-\mu+\psi)^2 + 4\mu\psi = (-\lambda+\mu+\psi)^2 + 4\lambda\psi,$$

it comes that α and β are real numbers with $\beta \geq 1$ and $0 \leq \alpha \leq 1$.

The case where $\psi = 0$ is plain: we then have $\mathbf{P}_{\oslash,\theta} = 1$ (there is no fossil). If $\psi > 0$, then $\beta > 1$ and we have necessarily $\mathbf{P}_{\oslash,\theta} = \alpha$, which gives us a natural interpretation of the coefficient α. The probability $\mathbf{P}_{\oslash,\theta}$ can be itself interpreted as the asymptotical proportion of lineages unobservable from the fossil record. It does not take into account the lineages observable from the present time only.

The probability $\mathbf{P}_{\oslash,\theta}$ is close to the complementary of the probability P_s of Bapst (2013). The probability P_s is defined as the probability of sampling an extinct clade of unknown size, also under the assumption that the diversification process continues indefinitely. The only difference is that $\mathbf{P}_{\oslash,\theta}$ stands for the probability of not sampling a clade extinct or not.

Remark that $\mathbf{P}_{\oslash,\theta}$ is not the complementary probability for species to leave fossils (i.e. before cladogenesis or extinction). This last probability, again under the assumption that the diversification process does not end, is $\frac{\psi}{\lambda+\mu+\psi}$.

References

Bapst, D. W. (2013). A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. *Methods in Ecology and Evolution*, **4**(8), 724–733.

Didier, G., Royer-Carenzi, M., and Laurin, M. (2012). The reconstructed evolutionary process with the fossil record. *Journal of Theoretical Biology*, **315**(0), 26 – 37.

Ford, D., Matsen, F. A., and Stadler, T. (2009). A Method for Investigating Relative Timing Information on Phylogenetic Trees. *Systematic Biology*.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, **267**(3), 396 – 404.