The distribution of branch lengths in phylogenetic trees

Emmanuel Paradis

The distribution of branch lengths in phylogenetic trees

Emmanuel Paradis

Institut des Sciences de l'Évolution, Université Montpellier – CNRS – IRD – EPHE, CC 064, Place Eugène Bataillon, F-34095 Montpellier cédex 05, France

Abstract

A lot of effort has been devoted to analyse the distribution of branching times observed in a phylogenetic tree. On the other hand, the distribution of branch lengths has not received similar attention. In this paper, the distribution of branch lengths is studied. It is shown that different types of branches within a tree have distinct distributions. Some equations to predict these distributions are derived with respect to diversification parameters and whether the size of the tree is known or not. A simulation study validated these predictions. The inferred distributions are used to develop graphical and statistical tools to assess the goodness-of-fit of diversification models. An application is presented on a recently published dated phylogeny of Carnivora. Some future developments are discussed.

Keywords: diversity, extinction, phylogeny, speciation

Tel: +33 (0) 4 67 14 46 85

Fax: +33 (0) 4 67 14 36 14

E-mail: Emmanuel.Paradis@ird.fr

1. Introduction

Phylogenetic trees have become a foundamental tool to study macroevolutionary processes. Several information can be extracted from a phylogeny: its general tree shape or various summaries of branch lengths (Mooers and Heard, 1997; Pybus and Harvey, 2000; François and Mioland, 2007). These information can then be used to infer the speciation and extinction rates of the studied group thus quantifying the tempo of its diversification through time. An important class of models considers the homogeneous models which assume that the rates of speciation and extinction may vary with respect to time but at a given time they are the same for all species present in the clade. From a biological point of view, these models are justified when diversification is driven by global environmental variables such as climate. An important result from Lambert and Stadler (2013) is that under a homogeneous model of diversification, the distribution of branching times depends only on the rates of speciation and extinction. Therefore, the branching times of a phylogenetic tree summarize the information required to estimate these rates under the assumption of homogeneity. However, when this assumption does not hold, one has to consider other quantities such as the distribution of branch lengths. For instance, Venditti et al. (2010) analysed the distribution of internal branch lengths for 101 phylogenies in order to assess the constancy of speciation rates accross a wide range of taxonomic groups. Besides, the distribution of branch lengths had retained some attention, for instance for the study of phylogenertic diversity (Faith, 1992; Mooers et al., 2012; Stadler and Steel, 2012). These previous studies have mostly considered models of constant diversification, i.e., with constant rates of speciation and extinction.

The main objective of this paper is to develop an approach to infer the distribution of the branch lengths of a phylogenetic tree considered as the result of a diversification process with known rates of speciation and extinction. Contrary to previous contributions, the present approach handles time-dependent variation in speciation and extinction rates. Two cases are considered: when the size of the tree is unknown, and when it is known so that the predictions of the mean branch length can be conditioned on tree size. A simulation study was performed to assess the accuracy of the predictions of the equations derived in this work. Some applications of

these equations are considered with a focus on a method to assess the goodness of the fit of a diversification model using graphical tools and statistical tests based on the cumulative distribution function. An analysis of a phylogeny of Carnivora (Mammalia) is presented to illustrate their practical use.

2. Models

2.1. THE GENERAL CASE UNCONDITIONAL ON TREE SIZE

We consider a dated phylogenetic tree with N species. We assume that this tree is rooted and fully dichotomous; consequently, it has $N-1$ ($=m$) internal nodes. Several quantities can be extracted and analysed from such a tree (Fig. 1). The branching times (also known as node times, node heights, or node depths), measured from the root, are denoted $t_1,\ldots,t_m$. In this section, the root is the origin of the time scale, so that $t_1 = 0$, and the present is denoted T. Alternatively, we could consider the branching intervals (known as coalescent intervals in population genetics) which summarise the same information as the branching times and are defined as $t_{i+1} - t_i$ ($i = 1,\ldots,m-1$). The branch lengths provide more information than the branching times because different sets of branch lengths can result in the same branching times. However, one needs the topology of the tree to fully determine the branching times from the branch lengths.

Birth–death models are a class of continuous-time models where each species (or individual) is continuously exposed to speciation (birth) and extinction (death). The time-dependent birth–death model assumes that the probability of a speciation event varies with time and follows a function $\lambda(t)$, and similarly for the probability of extinction $\mu(t)$. This formulation implies no lineage-specific variation: at a given time, all species have the same probabilities of speciation and extinction. We first remind some general mathematical background.

Under a time-dependent birth–death model, we can write the probability that a lineage, originating from a single species at time t, has exactly one descendant at time T (Kendall, 1948):

$$\Pr(t, n_T = 1) = \frac{\mathrm{e}^{-\rho(t,T)}}{W(t)^2},$$

with

$$\rho(t,T) = \int_t^T \mu(u) - \lambda(u)\mathrm{d}u,$$

and

$$W(t) = \mathrm{e}^{-\rho(t,T)}\left[1 + \int_t^T \mathrm{e}^{\rho(t,u)}\mu(u)\mathrm{d}u\right].$$

The probability of a lineage being extinct during the same time interval is:

$$\Pr(t, n_T = 0) = 1 - \frac{\mathrm{e}^{-\rho(t,T)}}{W(t)},$$

from which we can write the probability that a lineage is not extinct:

$$\Pr(t, n_T \geq 1) = 1 - \Pr(t, n_T = 0) = \frac{\mathrm{e}^{-\rho(t,T)}}{W(t)}.$$

We easily deduce the probability that a lineage has two or more lineages which we will use below:

$$\Pr(t, n_T \geq 2) = 1 - \Pr(t, n_T = 0) - \Pr(t, n_T = 1).$$

The number of species living at time t is a random variable with expectation given by:

$$\mathbb{E}(n_t) = \mathrm{e}^{-\rho(0,t)}.$$

These equations are the basis to derive the distribution of branching times of an ultrametric phylogenetic tree (Nee et al., 1994). If we are interested in the distribution of branch lengths, we need to consider separately different types of branches. Traditionally, two types are distinguished: terminal and internal branches. In this paper, two kinds of terminal branches are distinguished: those connected to the same internal node, and thus defining a cherry (McKenzie and Steel, 2000), and those that are sister-group of a clade of two or more species. These two types of terminal branches are termed *cherry* and *outer* branch, respectively (Fig. 2). The cherry branches are by definition duplicated in a tree.

A cherry is the result of a speciation at time t leading to two species surviving at present (time

T) with no other (recorded) speciation event. Under the assumption that these events do not covary, the density $\xi_1(t)$ of a cherry is given by the product of the probabilities of these three events (speciation at t and survival of both species from t to T):

$$\xi_1(t) = \lambda(t)\Pr(t, n_T = 1)^2. \tag{1}$$

Note that this is a density function, not a probability function because we cannot assume that $\xi_1(t)$ sums to one over all possible values of t.

We can use a similar reasoning and apply it to derive the density of outer branches. We consider here three events: a speciation at time t, the survival of a single species from t to T, and the survival of a clade with at least two species from t to T. There is an additional factor 2 because there are two possible combinations (i.e., left and right children from the speciation at t). Thus, the density $\xi_2(t)$ of a single terminal branch originating at time t and being sister of a clade with two or more species is:

$$\xi_2(t) = 2\lambda(t)\Pr(t, n_T = 1)\Pr(t, n_T \geq 2). \tag{2}$$

These two density functions $\xi_1(t)$ and $\xi_2(t)$, after being multiplied by $\mathbb{E}(n_t)$ and properly normalized, lead to probability density functions (pdf) of the distribution of both types of terminal branch length. The respective expected means can be calculated with:

$$\mathbb{E}(l) = \int_0^T u\tilde{\xi}(u)\mathrm{d}u, \tag{3}$$

where $\tilde{\xi}$ is the normalized function and l is the branch length. The normalizing factor is computed with:

$$\int_0^T \mathbb{E}(n_u)\xi(u)\mathrm{d}u,$$

where ξ is replaced by either ξ_1 or ξ_2 depending on the type of branch. The cumulative density function (CDF) of branch lengths, which is by definition the probability that l is smaller than or equal to a given value (t) is:

$$\Pr(l \leq t) = \int_0^t \tilde{\xi}(u)\mathrm{d}u,$$

87 The expected mean terminal branch length over the tree (i.e., pooling both types) is computed by the mean of both means weighted by their normalizing factors.

We now turn to the distribution of internal branch lengths. The problem is more difficult as it 90 requires to derive the probabilities of branches connecting two internal nodes. The first (oldest) node connects two sister-clades, one of them includes two or more species and the other one includes one or more species. The density function of such a node being observed at time t is 93 denoted as g_1. This node is the result of three events: a speciation at time t, the survival of a clade with two or more species from t until T, and the survival of a clade with one or more species during the same time. Thus:

$$g_1(t) = 2\lambda(t)\Pr(t, n_T \geq 2)\Pr(t, n_T \geq 1).$$

96 The factor 2 is, again, because there are two possible combinations for the sister-clades. The second (youngest) node connects two clades both made of one or more species, and its density function is:

$$g_2(t) = \lambda(t)\Pr(t, n_T \geq 1)^2.$$

99 The density of internal branches is found by multiplying the density functions of these two nodes and the probability of neither extinction, nor recorded speciation between t and t'. The probability of this last event is precisely given by the function g_2. It is thus necessary to integrate 102 between t and t' the sum of the two functions $\mu(t)$ and $g_2(t)$ (Cox and Oakes, 1984). We finally find that the density of an internal branch starting at time t and ending at time t' is:

$$f(t,t') = g_1(t)g_2(t')\exp\left[-\int_t^{t'} g_2(u) + \mu(u)\mathrm{d}u\right].$$

As before, we can use this density, after proper normalization, as a pdf of internal branch 105 lengths. In this case, calculating the expected mean requires a double integration on t and t' (with

the constraint $0 \leq t < t' < T$):

$$\mathbb{E}(l) = \int_0^T \int_t^T (t' - t)\tilde{f}(t, t')\mathrm{d}t'\mathrm{d}t. \tag{4}$$

Note that $t' - t$ is the internal branch length (l), and $\tilde{f}$ is, as above, the normalized density.

2.2. CONDITIONING ON TREE SIZE

The above development assumes that T (the age of the root of the tree) is known whereas N is unknown. In this section, it is assumed that N is known and fixed whereas T is a random variable. To derive the distributions of branch lengths, a different approach is needed. The approach used here is inspired from the coalescent approach developed by Stadler (2008, 2009, 2011a). The coalescent (Kingman, 1982) considers the way a sample of gene lineages from a population coalesce backwards in time. In population genetics, the distribution of the times to coalescence is determined by the genealogical structure within the population (Wakeley, 2009). On the other hand, in a phylogenetic tree, the coalescence events follow the speciation and extinction events (see Stadler, 2011a, for a clear exposition of this rationale). This approach leads in a straightforward way to the distribution of branching times in a phylogeny conditioned on N (see below); however, we also need to know the distribution of cherry and outer terminal branches.

Consider a coalescent tree of size N; each node of this tree is the result of a coalescent event between two lineages. These lineages may be either a lineage not yet coalesced, or the result of a previous coalescent event (i.e., lineages already clustered). For simplicity, we may call the first kind of lineage 'singleton' and the second kind 'cluster'. Let α_i and ω_i denote the respective number of these two kinds of lineages, where the subscript i denote the coalescent event ($i = 1, \ldots, N-1$) with $i = 0$ designing the initial state (i.e., present time). Clearly, we have $\alpha_0 = N$ and $\omega_0 = 0$ (no coalescent event have yet occurred). The first event ($i = 1$) is the coalescence of two singleton lineages so that $\alpha_1 = N - 2$ and $\omega_1 = 1$. The second coalescence event may join, either two singleton lineages, so $\alpha_2 = N - 4$ and $\omega_2 = 2$, or a singleton lineage with the cluster lineage built at the previous step, so $\alpha_2 = N - 3$ and $\omega_2 = 1$. We can continue along the successice steps of the coalescent process, and find that, under the assumption that lineages coalesce randomly, α_i and ω_i follow a random bivariate discrete process with the three

following possible transitions with their respective probabilities on the right-hand side:

$$\left.\begin{aligned}\alpha_{i+1} &= \alpha_i - 2\\ \omega_{i+1} &= \omega_i + 1\end{aligned}\right\} \quad \frac{\alpha_i(\alpha_i-1)}{2K_i}, \tag{5}$$

$$\left.\begin{aligned}\alpha_{i+1} &= \alpha_i - 1\\ \omega_{i+1} &= \omega_i\end{aligned}\right\} \quad \frac{\alpha_i\omega_i}{K_i}, \tag{6}$$

$$\left.\begin{aligned}\alpha_{i+1} &= \alpha_i\\ \omega_{i+1} &= \omega_i - 1\end{aligned}\right\} \quad \frac{\omega_i(\omega_i-1)}{2K_i}, \tag{7}$$

with $K_i = (\alpha_i+\omega_i)(\alpha_i+\omega_i-1)/2$. The first transition (5) is the coalescence of two singleton lineages; the second transition (6) is the coalescence of one singleton lineage with one cluster lineage; and the third transition (7) is the coalescence of two cluster lineages. K_i denotes the number of combinations among the $\alpha_i+\omega_i$ lineages not yet coalesced. Because α and ω are discrete values, it is straightforward to compute their pdf for each coalecent event (see code provided with this paper). Trivially, the first coalescent event ($i=1$) is of the first type (since $\omega_0 = 0$), and at the end of the coalescent process ($i = N-1$), we have $\alpha = 0$ and $\omega = 1$ (Fig. 3). We note that at each step of a coalescent the number of lineages not yet coalesced is given by $N-i$ and is equal to $\alpha_i+\omega_i$, so the distributions of these two variables are symmetric.

From this random process it is possible to derive the probability that a branching time leads to a cherry or to an outer branch: the first possible transition (5) leads to the creation of two cherry branches, while the second one (6) leads to the creation of an outer branch. The probability that the coalescent event i results in a cherry is thus given by eq. 5 summed over all possible values of α_i weighted by their probabilities as derived above:

$$\sum_{x=0}^{N} \Pr(\alpha_i = x)\frac{\alpha_i(\alpha_i-1)}{2K_i}.$$

Similarly, the probability the same coalescent event results in an outer branch is given by eq. 6 summed over all possible values of α_i and ω_i:

$$\sum_{\alpha_i=0}^{N}\sum_{\omega_i=0}^{N}\frac{\alpha_i\omega_i}{K_i}.$$

150 However, we do not have yet the distribution of the lengths of these two kinds of branches. For this we need the distribution of branching times. In population genetics, this distribution is related to the mutation rate and the demographic structure and dynamics of the population
153 (Wakeley, 2009), but this cannot be applied to phylogenetic trees. Using the description of the coalescent process generating a phylogenetic tree in Stadler (2011a), it is possible to obtain this distribution in a straightforward way. To use this approach, we have to assume that the
156 birth–death process is time-reversible which is a reasonable assumption if the probabilities of speciation and extinction are not affected by the past states of species (in other words, if the diversification process is Markovian). The successive events considered here are speciations and
159 extinctions. When looking at the tree backwards in time (i.e., from present to the past), a speciation leads to the coalescence of two lineages, whereas an extinction leads to the appearance of a lineage (which does not survive until present). Thus, at time t during this process, the clade
162 is constituted of some lineages surviving until present and some lineages that go extinct before present and which are not recorded in the observed phylogenetic tree. Let S_t and E_t denote the number of each type of lineages at time t. Note that, by contrast with the previous section, time is
165 now measured from present ($t = 0$). Under the assumption of independence of events (speciation, extinction and coalescence), the changes through time of these two variables can be modelled by the following pair of differential equations:

$$\begin{aligned} \frac{\mathrm{d}S_t}{\mathrm{d}t} &= -\lambda(t) n_t \left(\frac{S_t}{n_t}\right)^2, \\ \frac{\mathrm{d}E_t}{\mathrm{d}t} &= n_t \left[\mu(t) - \lambda(t)\frac{2S_tE_t}{n_t^2} - \lambda(t)\left(\frac{E_t}{n_t}\right)^2\right], \end{aligned}$$

168 with $n_t = S_t + E_t$. The initial state is given by $S_0 = N$ and $E_0 = 0$. These equations are solved until $S_t = 1$. This can be done numerically after setting the value of N and the functions $\lambda(t)$ and $\mu(t)$, for instance with the package deSolve (Soetaert et al., 2010). The solutions give the
171 expected values of S_t and E_t through time (Fig. 4). Thus, the value t for which $S_t = N - i$ is the expected value of the ith coalescent time of the phylogenetic tree. We then just have to weight these expected values with the probabilities obtained above to derive the expected mean cherry
174 and outer branch lengths.

Like we have seen in the previous section, the calculations for internal branches are more complicated because we have to consider the times of two nodes in the tree. Let $\mathcal{A}_i^j$ be the probability that a node created during the coalescent at step j is available for coalescence at step i with $j < i$. Clearly, a node is created at each step of the coalescent so this node is available for coalescence at the next step, thus we have $\mathcal{A}_{i+1}^i = 1$. For a node created before step i, the probability that it does not coalesce at step i is given by the ratio of the number of combinations excluding this node on the total number of combinations:

$$\frac{\binom{N-i-1}{2}}{\binom{N-i}{2}} = \frac{(N-i-1)(N-i-2)}{2} \times \frac{2}{(N-i)(N-i-1)} = \frac{N-i-2}{N-i}.$$

We may thus write how $\mathcal{A}_i^j$ changes along the coalescent process:

$$\mathcal{A}_{i+1}^j = \begin{cases} 1 & \text{if } j = i, \\ \mathcal{A}_i^j \dfrac{N-i-2}{N-i} & \text{otherwise.} \end{cases}$$

We now use $\mathcal{A}_i^j$ to derive the frequency distribution of node ages at all steps of the coalescent, and thus compute the expected length of an internal branch. The reasoning is the same than for terminal branches: a recursive calculation is done for all coalescent events, computing the probabilities of different types of events. The difference here is that instead of considering the branching times, we consider their difference weighted by the probabilities that an internal branch is made for each interval.

3. Simulation Study

3.1. METHOD

To validate the above predictions, some trees were simulated using three algorithms. The first algorithm is described in Paradis (2011) and simulates a tree in continuous time where λ and μ are allowed to vary with time under any user-specified model; this is a time-forward algorithm where T is fixed and N is a random variable. It is implemented in the R package ape (Paradis et al., 2004) with the function `rbdtree`. The second algorithm is due to Stadler (2011b) and simulates trees in continuous time using a time-backward method with both N and T fixed; it is

implemented in the R package TreeSim (Stadler, 2014) with the function `sim.bd.taxa.age`. The third algorithm is similar to the second one except that T is random: the function `rphylo` in ape was used. All algorithms require to specify the values of the parameters λ and μ. Different values of these parameters were used (see Table 1). For each set of values of λ and μ, 10,000 trees were simulated using each algorithm. For the second and third algorithms, N was fixed to 200 (this value did not influence significantly the results reported here).

The agreement between the model predictions and the simulations were assessed by comparing the predicted means with the observed means of each type of branch. Additionally for some simulated trees, the distribution of branch lengths was plotted and compared with the distribution inferred from eqs. 1 and 2.

3.2. RESULTS

The distribution of terminal branch lengths differed markedly between the two types of branch: the cherry branches had an exponential-like distribution whereas the outer branches had a unimodal distribution (Figs. 5 and 6). The exact shape of these distributions depended on the values of speciation and extinction rates. The histograms displayed on Figs. 5 and 6 show the observed distributions for two trees simulated with the same net diversification rate ($\lambda - \mu = 0.1$) but with different values of the rates λ and μ. The curves superimposed on the histograms are the densities predicted from eqs. 1 and 2.

When the trees were simulated with fixed N and fixed T (package TreeSim), the observed mean branch lengths were well predicted by eqs. 3 and 4 when the value of μ was small relative to the value of λ. However, the difference between the observed and the predicted values increased when the values of μ increased (Table 1). Similarly, the predictions of the mean internal branch length were not very accurate except when μ was small; however, the predicted values varied in the same way as the observed ones.

When the trees were simulated with random N (package ape), the predicted values for terminal branches were close to the observed ones. For both types of terminal branches, the observed means were very close to the predicted values: the difference between both values was less than 0.1 with one exception (the cherry branches with $\lambda = 0.1$ and $\mu = 0.09$) where the

difference was less than 0.5 (Table 2). The prediction of mean internal branch lengths was not accurate though, as before, the predicted and observed values varied in the same way.

When the trees were simulated with fixed N and random T, the observed mean cherry or outer branch length (Table 3) were slightly different from the predicted means not conditioned on N (Table 1 or 2). However, when the prediction was conditioned on $N = 200$ (within parentheses in Table 3), the predicted and observed means were in good agreement for all values of λ and μ. The same observation was made for internal branch lengths.

4. Application

The ability to predict the distribution of branch lengths may have several practical applications, such as predicting variation in phylogenetic diversity (Faith, 1992), or deriving likelihood functions for estimation of diversification parameters. Another potential application could be in specifying informative priors for Bayesian phylogenetic analyses: assuming distributions on speciation and extinction rates (for instance, obtained from data on species diversity), it may then be possible to derive prior distributions on the different kinds of branch lengths to use as input for a Bayesian approach on phylogenetic inference.

4.1. AN EXAMPLE WITH GOODNESS-OF-FIT TESTS

Here, we detail an application of the above equations to assess the goodness of fit of diversification models. Methods to assess the goodness-of-fit of a model must be distinguished from methods that compare models (e.g., ratios of mean squares, likelihood-ratio tests, information criteria). The latter seek to test the adequacy of two or more models relatively to each others. By contrast, goodness-of-fit methods assess the adequacy of a model in an absolute manner by comparing the predictions of the model with the data. These methods may be graphical (e.g., plot of residuals in a regression analysis) or statistical (e.g., testing the null hypothesis that the model does not fit the data).

The predicted distributions of branch lengths can be used in different ways to assess the adequacy of a specific birth–death model to a phylogeny. It is possible to draw a histogram of the observed distribution of branch lengths and compare it to the expected density. We must be

careful that the shape of the histogram can be influenced by the definition of the intervals. Another method, which avoids this problem, is to compare the observed and expected quantiles of the distribution of branch lengths (a method known as the QQ-plot). Appendix A details a statistical procedure based on this approach.

Computer code written in R (R Core Team, 2014) is provided with this article to perform all calculations described in this paper.

4.2. APPLICATION TO A CARNIVORA PHYLOGENY

Nyakatura and Bininda-Emonds (2012) published a dated phylogeny with 286 species of Carnivora and 8 other mammal species as outgroup which were removed from the tree before analysis. The "best estimate" tree provided by Nyakatura and Bininda-Emonds was used. The dated phylogeny was analysed with the method from Nee et al. (1994) to estimate λ and μ; the 95% confidence intervals (CIs) were calculated with the method described in Paradis (2003). From the above estimates, the expected distributions of terminal branch lengths were derived and compared to the observed ones. Giving that the prediction of internal branch lengths is less precise than for terminal branches (see previous section), only the latter were considered.

The estimated speciation and extinction rates from the dated phylogeny were: $\hat{\lambda} = 0.134$ (95% CI: 0.097–0.182) and $\hat{\mu} = 0.032$ (95% CI: 0.008–0.068). The distribution of terminal branch lengths showed a unimodal shape for both types (Fig. 7). For the cherry branches, there were less short branches than expected as revealed by the histogram and the QQ-plot (Fig. 7A). For the outer branches, the distribution of branches of length between 0 and 10 Ma seemed to agree with the expectation of the fitted constant-rate model; however, longer branches were more frequent than expected (Fig. 7B).

5. Discussion

The present paper proposes to infer the distributions of terminal branch lengths by distinguishing two types: the cherry branches which lead to two sister-species, and the outer branches which lead to a single species sister of a clade with at least two species. The simulations confirmed that these two kinds of terminal branches have distinct distributions. Besides, both types of terminal

branches are not expected to be observed with the same frequencies. In a phylogenetic tree, cherry branches are, on average, twice more frequent than outer ones; so, the mixture of both is dominated by the exponential nature of the distribution of the first type.

The simulation results were substantially different depending on the algorithm used to simulate the trees. It is interesting to note that the mean length of cherry branches did not vary much with respect to the simulation algorithm used (first column in Tables 1–3), while a substantial difference was observed for the outer (second column) and the internal branches (third column). It is noteworthy that the predictions matched well the simulation results when the same conditioning (or lack of) was used. The only notable discrepancy relates to the mean internal branch lengths with random N which were not well predicted when the extinction rate was high (Table 2). On the other hand, the predictions conditioned on the value of N were particularly more accurate for all types of branches (Table 3). This shows the advantage of conditioning the predictions when the number of species in the phylogeny is known.

Two approaches were used in this paper. The first one considers time-forward equations where tree size is a random variable. The second approach is based on time-backward equations with known tree size. Both approaches have their respective advantages: the first one makes possible to derive the complete distribution of branch lengths, while the second one gives more accurate predictions of mean branch lengths. It seems possible to use this second approach to derive the complete distribution of branch lengths conditioned on N, but this has not been yet worked out. This is currently under study. Nevertheless, it appeared here that the predicted distributions from the unconditioned case is already accurate for large trees (see Fig. 6). So it will be particularly interesting to assess whether conditioning on N improves these predictions.

The equations presented in this paper may have several applications. The branch lengths of phylogenetic trees are often used to quantify the diversity of a clade because two distantly related species (thus separated by a long path on a phylogeny) likely represent more biological diversity than two closely related species (Faith, 1992). This question is still under current research (Volkmann et al., 2014). The probability density function of branch lengths with respect to diversification parameters is thus likely to be helpful to predict the distribution of phylogenetic

biodiversity when reconstructed phylogenetic trees are not available. For instance, speciation and extinction rates can be estimated from species diversity of clades (e.g., Magallón and Sanderson, 2001; Paradis et al., 2013; Stadler et al., 2014). These estimates could then be used together with the eqs. 1 and 2 to predict the phylogenetic diversity of such groups even without phylogenies. In a recent paper, Crawford and Suchard (2013) derived expressions for the distributions of phylogenetic diversity and of phenotypic disparity under the assumption of diversification under a Yule model (so $\mu = 0$ and λ constant) so that these quantities may be estimated event without a known phylogeny.

The application with the Carnivora phylogeny had mainly an illustrative purpose. It showed how to use graphical tools to assess the excess or deficit of short and long branches within a tree. It should be kept in mind that such diagnostics of fit are done with respect to a fitted model. For simplicity, a simple model was used in the present analyses, but it is straightforward to use more complicated models. The results showed that the Carnivora tree has a deficit of short cherry branches compared to what is expected under a null model. It is clear that such deficit may result in a an apparent decline of speciation rates close to present, a pattern that has been frequently observed in real phylogenies (Moen and Morlon, 2014). The tools introduced here may thus contribute to investigate this widely debated issue.

In a previous work, Venditti et al. (2010) quantified the distribution of internal branch lengths using standard statistical distributions and found that the exponential distribution was the best fit for the majority of phylogenies. They interpreted this result as evidence that speciations are the results of single rare events leading to reproductive isolation. Interestingly for the present work, they excluded all terminal branches under the motivation that these do not record information on waiting times between speciation events. The developments presented here may bring a new perspective on such analyses by including all information from the phylogenies.

The possibility to derive separate distributions for cherry and outer terminal branches may have some implications for the study of tree shape and its balance (Tarver and Donoghue, 2011). In particular, it seems possible to derive under which conditions unbalanced trees are generated. It must be pointed out that the main objective of this paper was to set a framework to derive the

distributions of different types of branch lengths within a tree. For simplicity, it was implicitly assumed in the above developments that diversification is homegeneous (λ and μ can vary through time but they have the same values for all species at a given time). However, it is possible to generalise this approach to situations with heterogeneous rates (e.g., when some contemporaneous clades do not diversify at the same rates). The fact that the above equations (1 and 2) can consider a single branch makes possible to derive distributions which consider heterogeneous rates. A possible application could be, for instance, with the coalescent approach used with fixed tree size, to relate diversification parameters to ancestral character states, and thus analyse a trait-dependent diversification model. This is currently under development.

Acknowledgements

I am grateful to two anonymous reviewers for their constructive comments. This is publication ISEM 201x-xxx.

Appendix A.

A statistical test of the fit of a birth–death model to a phylogenetic tree may be performed by comparing the expected CDF with the empirical (i.e., observed) cumulative distribution function (ECDF). A traditional approach to perform such a comparison is to define discrete intervals over the possible values of the observed variables (branch length in our situation), count the number of observations falling within each interval, compute with the CDF the expected numbers for the same intervals, and compare the two tables of counts with a χ^2-test. In practice, this approach presents some difficulties when the counts may be small which is the case in phylogenetic trees where typically long branches are less frequent than the short ones (see figures above). A more powerful approach is to use tests that compare directly the curves defined by the ECDF and the CDF. One of these tests is the Kolmogorov–Smirnov test which considers the largest difference between these two curves. Other tests use all information from both curves. Two of these tests have been used to assess the distribution of branching times (Paradis, 1998): the Cramér-von Mises test and the Anderson–Darling test. Recently, some alternative new tests based on the same principle have been developed (Zhang, 2002; Esteban et al., 2007; Zhao et al., 2009). The

difficulty with these ECDF-based tests is that their exact distributions are not known, and therefore the critical values must be found by simulations of the null hypothesis (Stephens, 1974).

Here we consider two tests: the Cramér–von Mises test:

$$W^2 = \sum_{i=1}^{n} \left(z_i - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n},$$

where n is the number of points considered for the test and z_i is the predicted CDF for the ith point. The second test was proposed by Zhang (2002):

$$Z_C = \sum_{i=1}^{n} \left\{ \log \left[\frac{1/z_i - 1}{(n-0.5)/(i-0.75) - 1} \right] \right\}^2.$$

For both tests, the null distribution was determined by simulating a large number of trees (typically 10^4). The procedure is as follows.

1. Obtain the estimates $\hat{\lambda}$ and $\hat{\mu}$ from the observed phylogeny using the maximum likelihood method from Nee et al. (1994).
2. Compute the expected CDF of branch lengths (z_i's) using either eq. 1 (for the cherry branches) or eq. 2 (for the outer branches) with $\hat{\lambda}$ and $\hat{\mu}$ as speciation and extinction rates and T equal to the age of the root node of the observed phylogeny.
3. Compute the observed statistics W^2 and Z_C.
4. Generate 9999 trees using the algorithm from Stadler (2011b) with $\hat{\lambda}$ and $\hat{\mu}$ as speciation and extinction rates, and the number of species (N) and the age of the root node (T) both from the observed phylogeny.
5. Compute the statistic (W^2 and Z_C) for the 9999 trees. These give the null distributions of these tests. The P-values are the numbers of these 9999 values which are greater than the observed values computed at step 3 divided by 10,000 (since the observed values are considered as following the null hypothesis).

Note that in this procedure the tested model is that λ and μ are constant through time. If a time-dependent model were tested, it would be necessary to use the appropriate estimators of $\lambda(t)$ and $\mu(t)$ in step 2 and use them in the simulation in step 4.

When applied to the Carnivora phylogeny, the goodness-of-fit tests rejected very strongly the null hypothesis that the distribution of cherry branch lengths followed the constant-rate model (Table 4). For the outer branches, the Cramér–von Mises test was not significant ($P = 0.055$) and Zhang's Z_C was slightly significant ($P = 0.020$). The plots of the predicted CDFs against the ECDFs show more clearly the contrast between the two types of branches: the two curves are remarkably different for the cherry branches while they are more similar for the outer ones (Fig. 8).

Appendix B. Supplementary material

Supplementary data associated with this article can be found in the online version

References

Cox, D. R., Oakes, D., 1984. Analysis of Survival Data. Chapman & Hall, London.

Crawford, F. W., Suchard, M. A., 2013. Diversity, disparity, and evolutionary rate estimation for unresolved yule trees. Syst. Biol. 62 (3), 439–455.

Esteban, M. D., Marhuenda, Y., Morales, D., Sánchez, A., 2007. New goodness-of-fit tests based on sample quantiles. Commun. Stat. Simulat. Comput. 36 (3), 631–642.

Faith, D. P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61 (1), 1–10.

François, O., Mioland, C., 2007. Gaussian approximations for phylogenetic branch length statistics under stochastic models of biodiversity. Math. Biosci. 209 (1), 108–123.

Kendall, D. G., 1948. On the generalized "birth-and-death" process. Ann. Math. Stat. 19, 1–15.

Kingman, J. F. C., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Lambert, A., Stadler, T., 2013. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. Theor. Pop. Biol. 90, 113–128.

Magallón, S., Sanderson, M. J., 2001. Absolute diversification rates in angiosperm clades. Evolution 55 (9), 1762–1780.

McKenzie, A., Steel, M., 2000. Distributions of cherries for two models of trees. Math. Biosci. 164, 81–92.

Moen, D., Morlon, H., 2014. Why does diversification slow down? Trends Ecol. Evol. 29 (4), 190–197.

Mooers, A., Gascuel, O., Stadler, T., Li, H. Y., Steel, M., 2012. Branch lengths on birth-death trees and the expected loss of phylogenetic diversity. Syst. Biol. 61 (2), 195–203.

Mooers, A. Ø., Heard, S. B., 1997. Inferring evolutionary process from phylogenetic tree shape. Quart. Rev. Biol. 72 (1), 31–54.

Nee, S., May, R. M., Harvey, P. H., 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. Lond. B 344, 305–311.

Nyakatura, K., Bininda-Emonds, O. R. P., 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 10, 12.

Paradis, E., 1998. Testing for constant diversification rates using molecular phylogenies: a general approach based on statistical tests for goodness of fit. Mol. Biol. Evol. 15 (4), 476–479.

Paradis, E., 2003. Analysis of diversification: combining phylogenetic and taxonomic data. Proc. R. Soc. Lond. B 270, 2499–2505.

Paradis, E., 2011. Time-dependent speciation and extinction from phylogenies: a least squares approach. Evolution 65 (3), 661–672.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20 (2), 289–290.

Paradis, E., Tedesco, P. A., Hugueny, B., 2013. Quantifying variation in speciation and extinction rates with clade data. Evolution 67 (12), 3617–3627.

Pybus, O. G., Harvey, P. H., 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B 267, 2267–2272.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Soetaert, K., Petzoldt, T., Setzer, R. W., 2010. Solving differential equations in R: package deSolve. Journal of Statistical Software 33 (9), 1–25.
URL http://www.jstatsoft.org/v33/i09

Stadler, T., 2008. Lineages-through-time plots of neutral models for speciation. Math. Biosci. 216 (2), 163–171.

Stadler, T., 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261 (1), 58–66.

Stadler, T., 2011a. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. USA 108 (15), 6187–6192.

Stadler, T., 2011b. Simulating trees with a fixed number of extant species. Syst. Biol. 60 (5), 676–684.

Stadler, T., 2014. TreeSim: Simulating trees under the birth-death model. R package version 2.1.
URL http://CRAN.R-project.org/package=TreeSim

Stadler, T., Rabosky, D. L., Ricklefs, R. E., Bokma, F., 2014. On age and species richness of higher taxa. Am. Nat. 184 (4), 447–455.

Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J. Theor. Biol. 297, 33–40.

Stephens, M. A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Statist. Assoc. 69, 730–737.

453 Tarver, J. E., Donoghue, P. C. J., 2011. The trouble with topology: phylogenies without fossils provide a revisionist perspective of evolutionary history in topological analyses of diversity. Syst. Biol. 60 (5), 700–712.

456 Venditti, C., Meade, A., Pagel, M., 2010. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature 463, 349–352.

Volkmann, L., Martyn, I., Moulton, V., Spillner, A., Mooers, A. O., 2014. Prioritizing populations
459 for conservation using phylogenetic networks. PLoS ONE 9 (2), e88945.

Wakeley, J., 2009. Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village, CO.

462 Zhang, J., 2002. Powerful goodness-of-fit tests based on the likelihood ratio. J. R. Statist. Soc. B 64 (2), 281–294.

Zhao, J., Xu, X., Ding, X., 2009. Some new goodness-of-fit tests based on stochastic sample
465 quantiles. Commun. Stat. Simulat. Comput. 38 (3), 571–589.

Fig. 1. (a) A tree with $N = 7$, (b) its branching times (or node times, or node depths, or node heights) ordered in decreasing size, (c) its branching intervals (or coalescent intervals) ordered through time, and (d) its branch lengths ordered with the seven terminal branches first (note the cherry ones which are duplicated) and then the five internal ones.

Fig. 2. A tree showing cherry and outer terminal branches.

Fig. 3. The probability densities of the number of singletons (α) and of clusters (ω) in a coalescent tree of size $N = 10$. All graphs are on the same scale. The column on the right-hand side gives the different coalescent events (i).

Fig. 4. Expected numbers of surviving (S_t) and extinct (E_t) lineages through time for a phylogeny with $N = 200$ species, speciation rate $\lambda = 0.1$ and extinction rate $\mu = 0.05$.

Fig. 5. Distributions of cherry (left) and outer (right) terminal branch lengths. The histograms show the observed distributions for two trees simulated with $N = 200$, $T = 50$, and (a) $\lambda = 0.1$, $\mu = 0$, or (b) $\lambda = 0.2$, $\mu = 0.1$. The curves show the predicted distributions for each type of branch and each combination of parameters.

Fig. 6. Same as Fig. 5 except that $N = 2000$.

Fig. 7. (a) Distribution of cherry branches (histogram), predicted density (curve) under a model of constant speciation and extinction rates, and QQ-plot of the observed and expected quantiles for the Carnivora phylogeny (Nyakatura and Bininda-Emonds, 2012). (b) Same thing for outer branches.

Fig. 8. Observed (empirical) and predicted CDFs for both types of terminal branches for the Carnivora phylogeny (Nyakatura and Bininda-Emonds, 2012).

Table 1: Mean branch length of trees simulated with $N = 200$ and $T = 50$. The values within parentheses are the predicted means using the equations presented in this paper. The simulations were replicated 10,000 times.

λ	μ	Terminal		Internal
		cherry	outer	
0.1	0	3.321 (3.333)	8.294 (8.327)	4.902 (4.934)
	0.05	3.609 (3.859)	9.630 (10.422)	5.990 (6.469)
	0.09	3.519 (4.486)	9.777 (12.248)	6.349 (7.302)
0.2	0.1	1.936 (1.931)	5.364 (5.341)	3.890 (3.501)
	0.15	2.065 (2.137)	6.092 (6.407)	4.721 (4.004)
	0.19	2.053 (2.375)	6.288 (7.477)	5.052 (4.244)

Table 2: Same as Table 1 except that N was random.

λ	μ	Terminal		Internal
		cherry	outer	
0.1	0	3.334 (3.333)	8.339 (8.327)	4.933 (4.934)
	0.05	3.836 (3.859)	10.460 (10.422)	6.829 (6.469)
	0.09	4.165 (4.486)	12.228 (12.248)	8.524 (7.302)
0.2	0.1	1.937 (1.931)	5.354 (5.341)	3.824 (3.501)
	0.15	2.129 (2.137)	6.424 (6.407)	5.123 (4.004)
	0.19	2.300 (2.375)	7.506 (7.477)	6.297 (4.244)

Table 3: Same as Table 1 except that T was random. The predicted values were calculated conditioned on the value of $N = 200$.

λ	μ	Terminal		Internal
		cherry	outer	
0.1	0	3.334 (3.332)	8.337 (8.657)	5.008 (5.191)
	0.05	3.862 (3.885)	10.697 (11.002)	7.683 (7.942)
	0.09	4.640 (4.656)	15.511 (15.686)	17.046 (17.111)
0.2	0.1	1.930 (1.958)	5.347 (5.513)	3.854 (3.968)
	0.15	2.141 (2.166)	6.482 (6.624)	5.610 (5.725)
	0.19	2.393 (2.425)	8.506 (8.563)	11.016 (10.958)

Table 4: Results of the goodness-of-fit tests of the constant-rate model to the Carnivora data using two types of branches (cherry and outer) from Nyakatura and Bininda-Emonds's (2012) tree. W^2: Cramér–von Mises test; Z_C: test from Zhang (2002).

	W^2	P	Z_C	P
Cherry	18.85	0.0021	131.50	0.0014
Outer	9.15	0.0553	40.23	0.0204

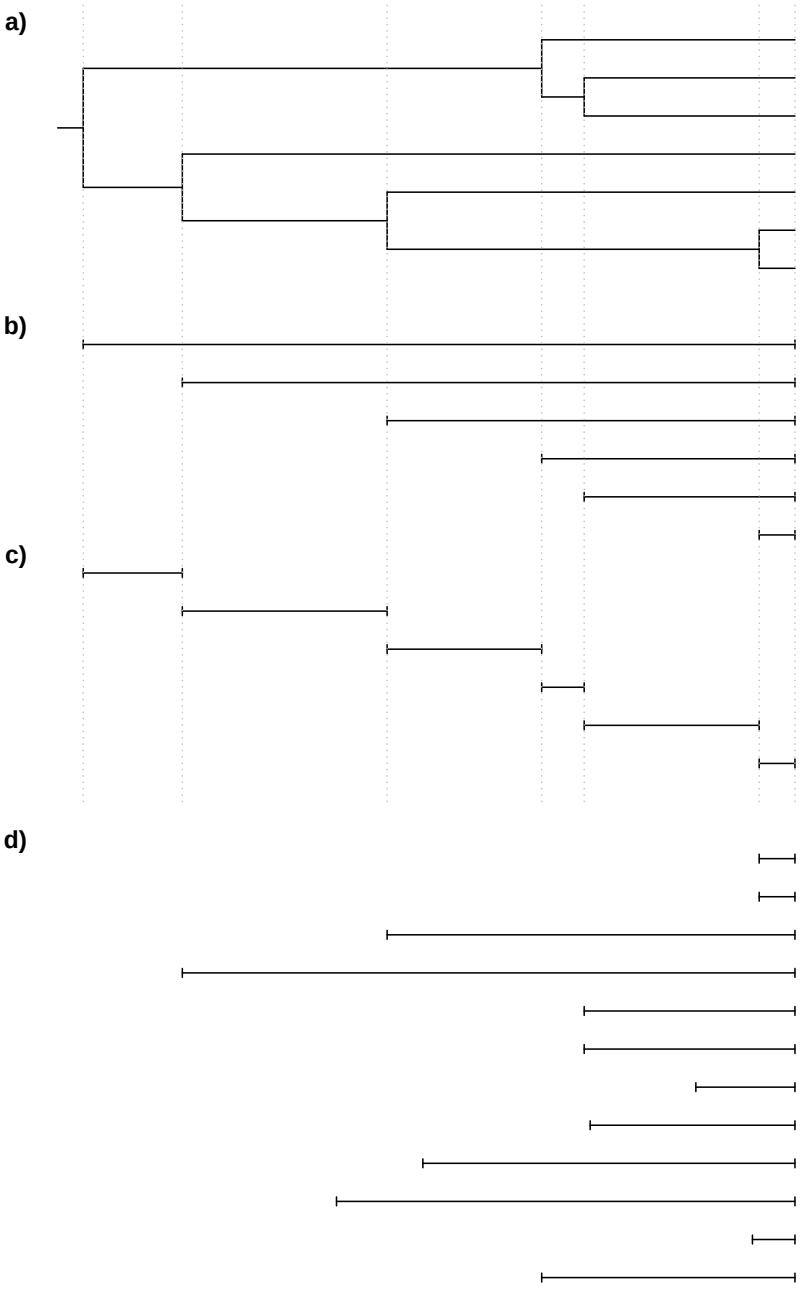

Figure 1: (a) A tree with $N = 7$, (b) its branching times (or node times, or node depths, or node heights) ordered in decreasing size, (c) its branching intervals (or coalescent intervals) ordered through time, and (d) its branch lengths ordered with the seven terminal branches first (note the cherry ones which are duplicated) and then the five internal ones.

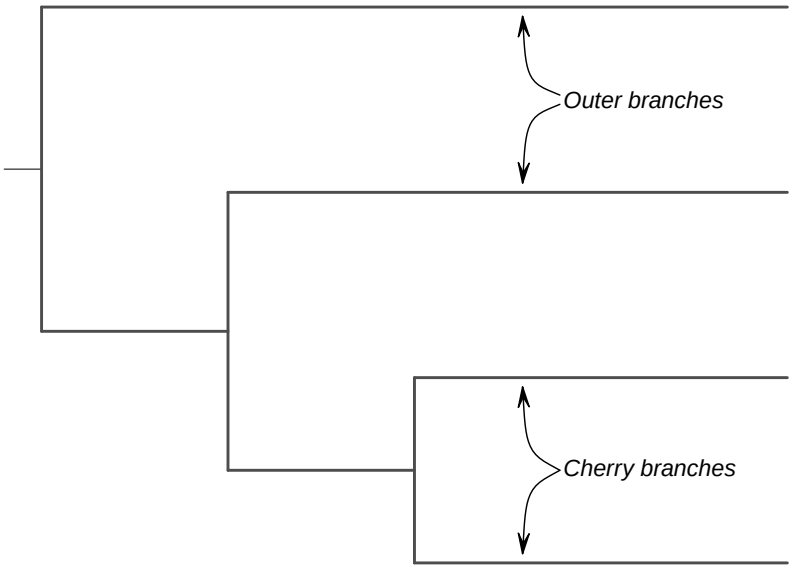

Figure 2: A tree showing cherry and outer terminal branches.

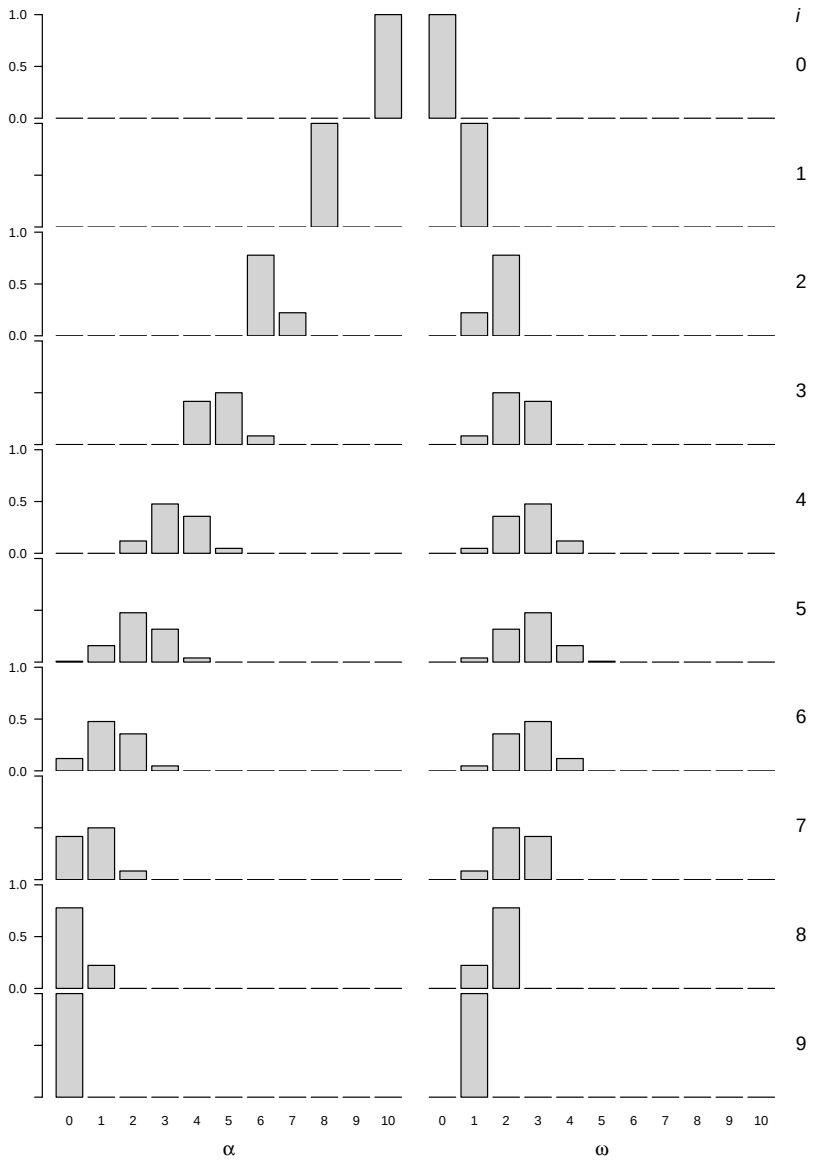

Figure 3: The probability densities of the number of singletons (α) and of clusters (ω) in a coalescent tree of size $N = 10$. All graphs are on the same scale. The column on the right-hand side gives the different coalescent events (i).

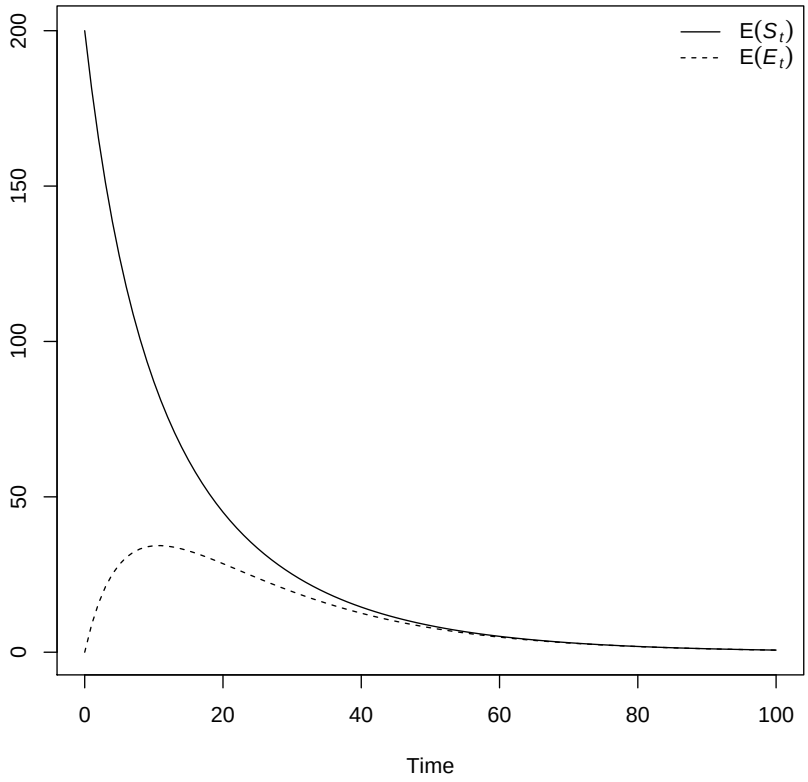

Figure 4: Expected numbers of surviving (S_t) and extinct (E_t) lineages through time for a phylogeny with $N = 200$ species, speciation rate $\lambda = 0.1$ and extinction rate $\mu = 0.05$.

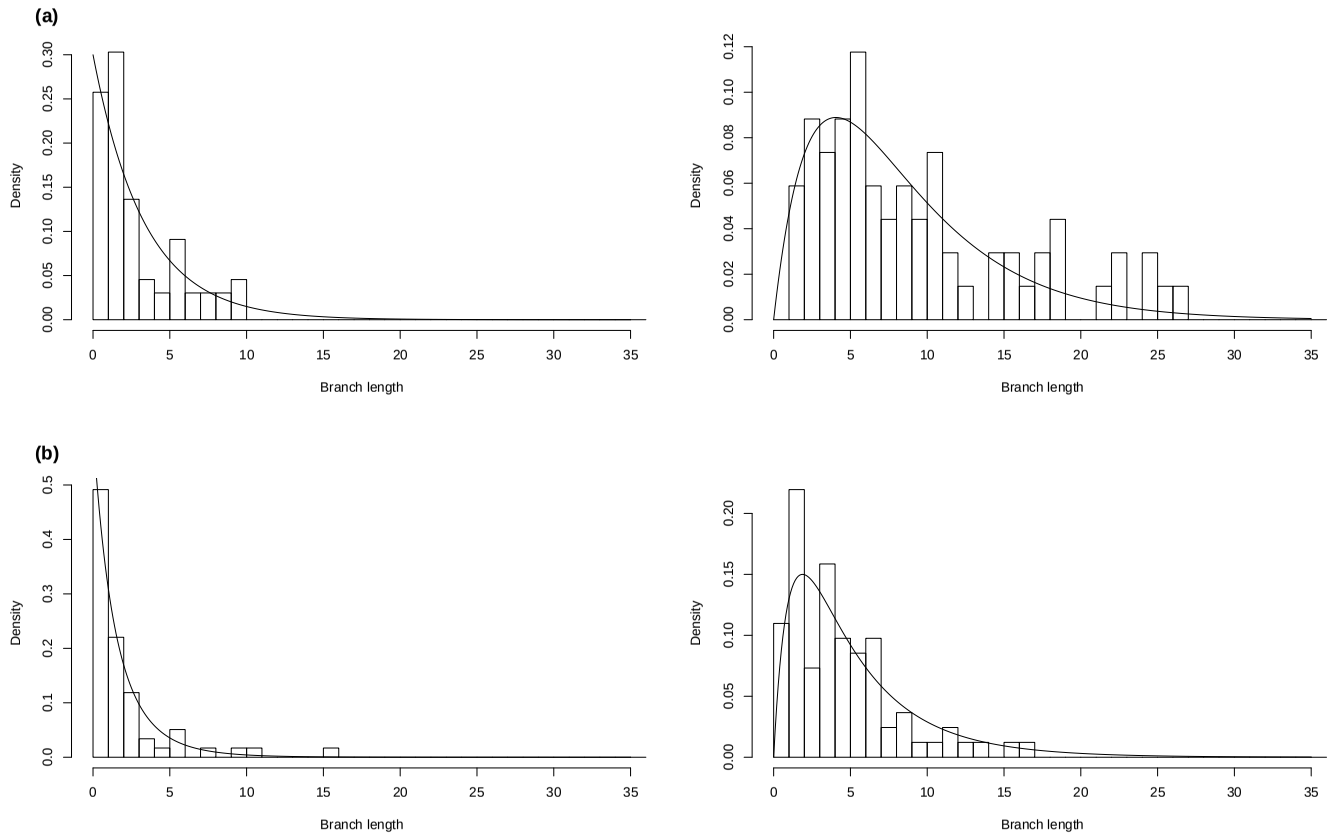

Figure 5: Distributions of cherry (left) and outer (right) terminal branch lengths. The histograms show the observed distributions for two trees simulated with $N = 200$, $T = 50$, and (a) $\lambda = 0.1$, $\mu = 0$, or (b) $\lambda = 0.2$, $\mu = 0.1$. The curves show the predicted distributions for each type of branch and each combination of parameters.

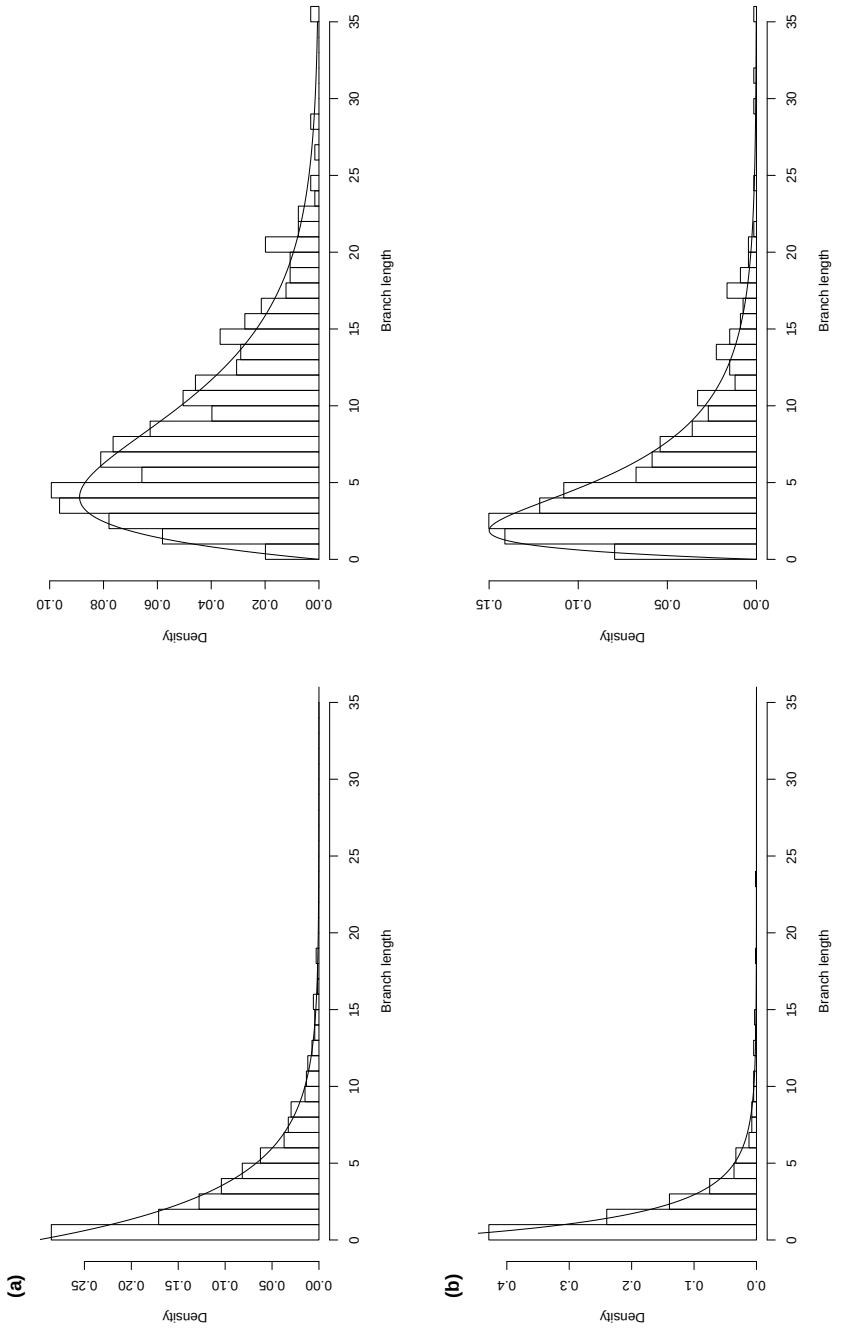

Figure 6: Same as Fig. 5 except that $N = 2000$.

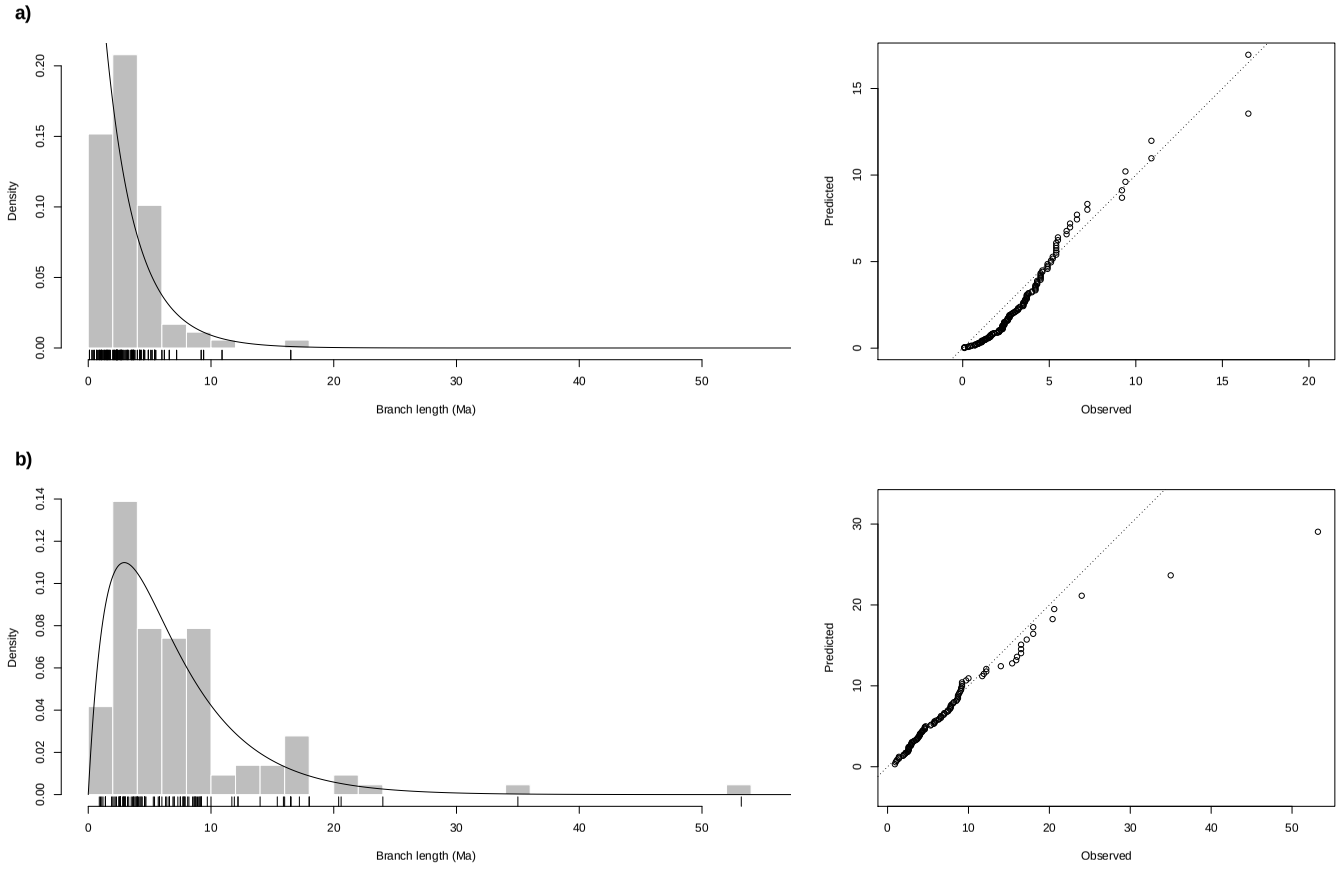

Figure 7: (a) Distribution of cherry branches (histogram), predicted density (curve) under a model of constant speciation and extinction rates, and QQ-plot of the observed and expected quantiles for the Carnivora phylogeny (Nyakatura and Bininda-Emonds, 2012). (b) Same thing for outer branches.

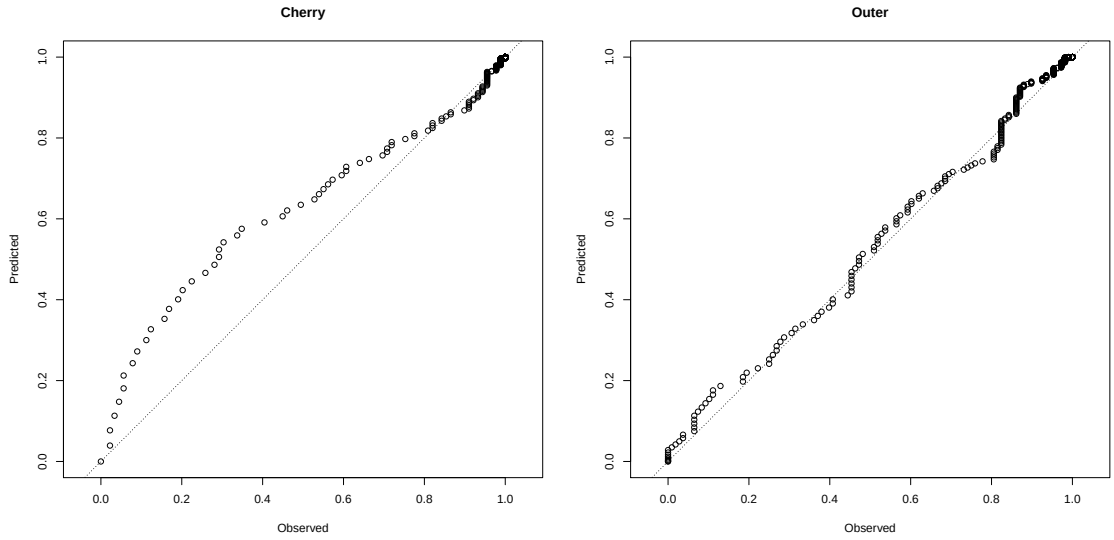

Figure 8: Observed (empirical) and predicted CDFs for both types of terminal branches for the Carnivora phylogeny (Nyakatura and Bininda-Emonds, 2012).