Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 209 (2007) 108–123
Mathematical Biosciences
www.elsevier.com/locate/mbs

Gaussian approximations for phylogenetic branch length statistics under stochastic models of biodiversity

Olivier François *, Céline Mioland

TIMC-TIMB (Dept Math Biology), INP Grenoble, UJF-CNRS, Faculté de Médecine, F38706 La Tronche, France

Received 20 July 2006; received in revised form 16 December 2006; accepted 10 January 2007
Available online 4 February 2007

Abstract

Stein's method for Gaussian approximations and derived results are used to study the distribution of two phylogenetic branch length statistics: the total height of cherries and the sum of external branch lengths. The Gaussian approximations are obtained under a particular model of phylogenetic tree recently introduced by Popovic. Under an appropriate normalization the model is shown to behave similarly as the coalescent, and the approximations given here are also valid in this context.

Keywords: Phylogenetic models; Branching processes; Tree shape statistics; Branch length statistics; Gaussian approximations

1. Introduction

Macro-evolutionary biology aims to uncover and explain the underlying processes that have led to the biological diversity we observe today. Phylogenetic trees not only provide a clear snapshot of biodiversity, but also make it possible to infer how the diversity has arisen (see for example [39,33,22,34,27]). To this aim, variation in diversification rates have been investigated through their signatures in the shape of phylogenetic trees [29,17,31].

* Corresponding author.
E-mail address: olivier.francois@imag.fr (O. François).

0025-5564/$ - see front matter
doi:10.1016/j.mbs.2007.01.005

Previous work on this problem has largely involved the development of tree-balance statistics which measure the topological distribution of species diversity as a single number [26,1]. Mooers and Heard [29] wrote an exhaustive review about tree balance in systematic biology, and Aldous gave an introduction in a more mathematical setting [3]. See [9] for a large-scale study of phylogenetic tree shape. Among these statistics, the most widespread may be the Sackin's and the Colless' indices [43,44,13,10]. McKenzie and Steel [28] also considered the *number of cherries*, i.e., pairs of leaves that are adjacent to a common ancestor node, and they described the distribution of this statistic under the Yule and the uniform models.

Tree shape statistics cannot capture the full amount of information contained in a phylogenetic tree because they ignore branch lengths. Recently, statistics that account for both temporal and topological aspects of phylogenetic trees have proven useful for measuring phylogenetic diversity [30]. The γ statistic of Pybus and Harvey [40] can be used to test whether the diversification rate vary through time. The index of *phylogenetic diversity* was introduced by Faith as the sum of the lengths over all branches in the tree [18]. Using this index, Nee and May [35] showed that a large amount of phylogenetic diversity is conserved after mass extinction under the coalescent model of evolution.

Null models for generating binary phylogenetic trees are useful for testing evolutionary hypotheses. Here we consider three such models sharing the same topological properties: the Yule model, the Hey model and the critical branching process conditioned on having a prescribed number of species at present time. The last model was studied by Popovic [38] and after by Aldous and Popovic [4] as a variant of a model considered earlier by Wollenberg [51]. Here, we focus on a point process representation of the Popovic model, and we derive useful properties of tree shape and branch length statistics from this representation. The main mathematical tool used in our approach is Stein's method for Gaussian approximations ([46], reviewed by Rinott and Rotar [41]).

This article is organized as follows. In Section 2 we describe the point process representation of trees, and prove a new result for the number of cherries under the Harding–Yule distribution of topologies. In Section 3 we recall results about the Popovic model, and exhibit connections with the coalescent (Hey) model. In Section 4 we use Stein's method to establish convergence to a Gaussian distribution for two statistics: the total height of cherries and the sum of external (pendant) branch lengths. In the last section we briefly discuss two examples.

2. A new perspective on tree shape

2.1. The Harding distribution

The ability of phylogenies to inform studies of differential diversification has been strengthened by the use of stochastic models of trees. Stochastic branching processes are indeed frequently employed to generate an expected distribution of differences against which observed differences can be tested. One of the most traditional way of representing a tree topology is the equal-rates Markov (ERM) or Yule model [49,20].

If the ERM branching process is initiated with a single species, and observed for a period of time t, with a branching rate λ, the probability of observing n species is geometric [21]

$$P(N_t = n) = p_t(1 - p_t)^{n-1}, \quad p_t = \mathrm{e}^{-\lambda t}, \quad t > 0.$$

Assuming that the N_t species are partitioned between the left (L_t) and right (R_t) descendents of the ancestral node, Harding [20] obtained that

$$P(L_t = k | N_t = n) = \frac{1}{n-1}, \quad k = 1, \ldots, n-1.$$

For an observed partition into ℓ and r species ($\ell + r = n$) among two sister clades, the probability of observing ℓ species in the left clade is then independent on ℓ. Extending this approach to a larger class of tree shape (called dendrograms), Aldous [2] introduced the concept of a *split distribution*. The split distribution describes the size of the left sister clade given the size of the parent clade, and can be propagated at each internal node of the tree recursively. The Harding uniform distribution is the split distribution that characterizes the ERM tree shape. The Harding distribution also appears in other models of trees, such as the Hey model [24] or the coalescent [25]. Trees with the Harding distribution have equivalent topologies.

2.2. A point process representation of trees

In this study, we make use of another representation of trees also yielding the Harding distribution. The tree model is essentially a graphical representation. It is based on internal node heights instead of branch lengths, and it assumes that the node heights are independent and identically distributed random variables.

A tree with n tips has $(n-1)$ internal nodes including the root. We denote by $H_1, \ldots, H_{n-1}$ the heights of the internal nodes represented as vertical bars. The tips are located at the left and right sides of each bar (see Fig. 1). These heights form a *point process* from which the tree can be reconstructed. The maximal height corresponds to the root which separates the tree into two clades. The tree topology can be deduced by iterating the search of the maximum within each sister clade. Because the location of the maximum is distributed uniformly over $[n-1]$, an elementary proof shows that this recursive procedure produces a tree with the Harding distribution.

At first glance, relating the previous construction to some existing branching processes used in the literature may seem difficult. We postpone the discussion of this question until Section 3 where the point process theory of Popovic will be summarized and used to prove connections with critical branching processes and coalescent models.

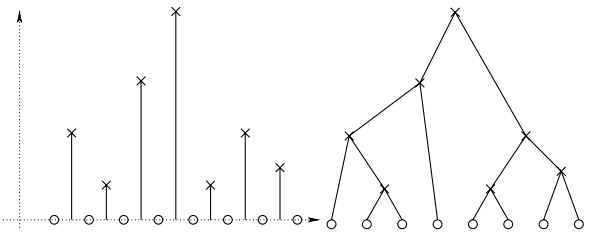

Fig. 1. Left: A point process realization of node heights for $n = 8$ tips. Right: The tree deduced from the realization.

2.3. Application to cherries

In this section, we use the point process representation of trees in order to improve results of McKenzie and Steel about the number of cherries in trees enjoying the Harding property [28]. A *cherry* is a pair of tips adjacent to a common node. Let us denote by C_n the number of cherries in a tree with n tips. The number of cherries can reveal the degree of imbalance of a tree. Under the Harding distribution, McKenzie and Steel proved a central limit theorem for C_n using a combinatorial argument (Polya urn). In addition, they proved that the expected number of cherries is equal to $\lambda = n/3$, $(n \geqslant 3)$, and the variance is equal to $\sigma^2 = 2n/45$, $(n \geqslant 5)$. Blum and François [8] and Rosenberg [42] also gave a proof of this result using a different method (distributional recursions and combinatorics). Here, we provide a stronger result (a *Berry–Essen* theorem) for the cherries.

Theorem 2.1. *Let C_n denote the number of cherries in a tree with n tips following the Harding distribution. Let λ be its mean number, and σ^2 be its variance. When n tends to infinity, we have*

$$\left| P(C_n \leqslant w) - \Phi\left(\frac{w-\lambda}{\sigma}\right) \right| \leqslant Cn^{-\frac{1}{4}}$$

where C is a positive constant and Φ denotes the standard normal cumulative distribution function.

The proof is based on Stein's method for Gaussian approximations [46]. Let us consider the point process representation of the tree. For a tree with n tips, a cherry corresponds to a local minimum in the sequence of heights $H_1, \ldots, H_{n-1}$ (see Fig. 1 again). More specifically the pair of tips $(i, i+1)$ forms a cherry when the height H_i is the minimum of the heights H_{i-1}, H_i, H_{i+1}. Then, the number of cherries corresponds to the number of minima of a random function defined on a graph.

Before concluding the argument, we recall some basic terminology. A graph $(\mathcal{V}, \mathcal{E})$ consists of a finite set $\mathcal{V}$ of vertices and a set $\mathcal{E}$ of edges. The distance $\delta(v, v')$ is the number of edges in the shortest path from vertex v to vertex v'. The degree of a vertex is the number of edges to which it belongs. A regular graph is one in which all vertices have the same degree, which is denoted by d. Let $s(u, v)$ be the number of common neighbors of vertices u and v. To conclude the proof of the theorem, we can use a result by Baldi et al. [5].

Theorem 2.2. (Baldi et al. [5]). *Let $(\mathcal{V}, \mathcal{E})$ be a regular graph and Y a random function on $\mathcal{V}$ whose values are independently distributed with a common continuous distribution and let W be the number of local maxima of Y. Then the mean and variance of W are given by*

$$\lambda = E[W] = \frac{|\mathcal{V}|}{d+1}$$

and

$$\sigma^2 = \mathrm{Var}[W] = \sum\nolimits_{\substack{u,v \\ \delta(u,v)=2}} s(u,v)(2d+2-s(u,v))^{-1}(d+1)^{-2},$$

and, for all $w \in \mathbb{R}^+$,

$$\left|P(W \leqslant w) - \Phi\left(\frac{w-\lambda}{\sigma}\right)\right| \leqslant C\sigma^{-\frac{1}{2}}$$

where C is an absolute constant.

In order to apply the above theorem, the point process representation described in Fig. 1 must be slightly transformed into an equivalent representation. Instead of a linear arrangement, the points are now organized in a circular fashion separating regularly spaced tips on a ring. The new representation includes n tips and n height points. To see that it is equivalent to Popovic's representation, the circle can be cut at the maximum height point, and unfolded to obtain a linear segment with $n-1$ height points. Then the number of cherries defined from n height points on the ring corresponds to the number of cherries defined from n tips on the segment. As a result, the approximation theorem can then be applied to the regular graph $\mathcal{V} = [n] = \{1, \ldots, n\}$ with edges $\mathcal{E} = \{(1,2),(2,3),\ldots,(n-1,n),(n,1)\}$ and $Y_i = H_i$ for all $i \in [n]$ to conclude the proof. □

Comments. The point process representation can also be used to prove a strong law of large numbers for the cherries. Actually, we have

$$C_n = \sum_{i=1}^{n-1} X_i$$

where $X_i = \mathbf{1}_{(H_{i-1}>H_i<H_{i+1})}$ for $i = 2,\ldots,n-2$, $X_1 = \mathbf{1}_{(H_1<H_2)}$ and $X_{n-1} = \mathbf{1}_{(H_{n-2}>H_{n-1})}$. There are no difficulties to extend the sequence (X_i) to obtain a stationary sequence on $\mathbb{Z}$. Then the application of Birkoff's ergodic Theorem (see [15], p. 337) leads to

$$\frac{C_n}{n} \to \frac{1}{3}, \quad \text{a.s.}$$

3. The Popovic model and its connection to the coalescent

3.1. Popovic's representation

In a recent study, Popovic showed that the point process representation can be used to model the asymptotic genealogy of a critical branching process [38]. These results were extended by Aldous and Popovic [4]. In this section, we recall Lea Popovic's main results and show that her point process representation can also be connected to the Moran–Hey model after an appropriate renormalization.

Motivated as a null model for comparison with phylogenetic data, Popovic studied a branching process model for a phylogenetic tree on n extant species. The origin of the clade is a random time in the past, denoted by t, which distribution is uniform (improper) over the interval $(0,\infty)$. Started from time t in the past, the process of extinctions and speciations is a continuous-time branching process conditioned on having n extant species at the present time.

According to this process each individual lives for a time distributed as an exponential random variable of rate $\lambda > 0$. During its lifetime, it gives birth at times of an independent Poisson process of rate λ. After birth all individuals behave independently of each other. Time can be rescaled so

that $\lambda = 1$. Using Popovic's notations, we call $\mathcal{T}_{n,t}$ the complete tree originated at time t in the past and having n species at the present time. In such models, the conventional notation "time s" means time s before present.

Every realization of a complete tree also determines a realization of a *lineage tree* of the extant species. This tree is denoted by $\mathcal{A}_{n,t}$ and corresponds to the genealogy of the n species: This is the smallest subtree of $\mathcal{T}_{n,t}$ that contains all the divergence times for pairs of lineages of extant species, without recording which ancestral species contain the lineage.

Aldous and Popovic [4] wrote that "*it is perhaps remarkable that there is a useful exact description of the lineage tree $\mathcal{A}_{n,t}$ based on a point process representation*" (cf. Fig. 1 again).

Theorem 3.1 (Popovic, [38] Lemma 3)**.** *Fix $n \geqslant 2$ and $t > 0$. The point process $\{(i + \frac{1}{2}, H_i), 1 \leqslant i \leqslant n-1\}$ where the (H_i) are i.i.d. with density function*

$$f_t(s) = (1 + t^{-1})(1 + s)^{-2}, \quad 0 < s < t$$

represents the lineage tree $\mathcal{A}_{n,t}$ within the complete tree $\mathcal{T}_{n,t}$.

In the sequel, we shall refer to this point process representation as the *Popovic model*. An obvious consequence of the above result is that the lineage tree $\mathcal{A}_{n,t}$ topology is characterized by the Harding distribution.

3.2. Connection with coalescent models

In this section, we establish connections between the Popovic model and the coalescent. The *coalescent* is a model of the genealogy of n individuals that arises from a diffusion limit in the Wright–Fisher dynamics [48,36]. A similar model called the *Hey model* is also used in macroevolution studies [24,35].

While the original definition of the coalescent is as a stochastic process on partitions of $[n]$, it is convenient to think it as a random rooted binary tree with lengths attached to edges. In this case, time is measured backward. At time T_n, two randomly chosen lineages coalesce and form an ancestral lineage which replaces them. At T_{n-1}, two lineages are chosen from the $n-1$ remaining, etc. At T_2 all lineages have coalesced, and T_2 represents the time to the *most recent common ancestor* of the sample. In the coalescent the inter-coalescence times $(T_i - T_{i+1})$ are independent random variables having exponential distribution with rate $\lambda_i = i(i-1)/2$.

The coalescent also appears as a limit of the Moran model after rescaling time by the total population size [32,14]. In the Moran model [16] the number of coexisting species is fixed at n. At successive discrete times, one randomly chosen species goes extinct and another randomly chosen speciates. Aldous and Popovic noticed that their critical branching model is qualitatively similar to the Moran model. Here we show that the branching times (node heights) and the external branch lengths have the same distributions under the Popovic model and under the coalescent model for large t. This result is obtained after multiplying the edge lengths by a factor $n/2$ in the coalescent.

3.2.1. Height of subtrees

In the Popovic model with n tips, the height distribution is given by the Theorem 3.1. The Laplace transform of the height H_i can be described as

$$E[e^{-sH_i}] = \frac{1+t}{t}\int_0^t \frac{\exp(-su)}{(1+u)^2}du, \quad s>0.$$

To see that the coalescent shares the same distribution, let us choose a height randomly among the node heights. To do this, we pick one ancestral node uniformly among the $n-1$ internal nodes. The height of the corresponding subtree is

$$H_n^c = T_K$$

where K is the label of the node sampled randomly from $\{2,\ldots,n\}$. Computing the Laplace transform of $nH_n^c/2$, we obtain

$$E\left[\exp\left(-\frac{nsH_n^c}{2}\right)\right] = \frac{1}{n-1}\sum_{k=2}^{n}\prod_{j=k}^{n}\left(1+\frac{ns}{j(j-1)}\right)^{-1}$$

As n tends to infinity, this function is equivalent to

$$E\left[\exp\left(-\frac{nsH_n^c}{2}\right)\right] \sim \int_0^1 \exp\left(-s\frac{1-z}{z}\right)dz = \int_0^\infty \frac{\exp(-su)}{(1+u)^2}du$$

For large n, the height $h_n = nH_n^c/2$ has density

$$f(s) = \frac{1}{(1+s)^2}, \quad s>0$$

which is the limiting density in the Popovic model when t goes to infinity.

3.2.2. External branch length

Because this will be a subject of interest in the next section, we check here that the same kind of results can be obtained for the external (i.e., pendant) branches of the coalescent. Under the Popovic model, the length of an external branch corresponds to the minimum between two consecutive heights H_{i-1} and H_i.

Lemma 3.2. *Let $n \geqslant 2$ and $t>0$. Under the Popovic model, an external branch length has density function*:

$$\forall\ 0<s<t, \quad \tilde{f}_t(s) = 2\frac{n-1}{n}\frac{1+t}{t}\frac{1}{(1+s)^2} - 2\frac{n-2}{n}\left(\frac{1+t}{t}\right)^2\frac{s}{(1+s)^3}$$

When t and n tend to infinity, the density function converges to

$$\tilde{f}(s) = \frac{2}{(1+s)^3}, \quad s>0.$$

Under the coalescent model, the external branch lengths were studied in [7]. Let us consider an external branch in a tree with n tips, and denote by B_n its length. The Laplace transform of B_n is given by

$$E[e^{-uB_n}] = \binom{n}{2}^{-1}\sum_{k=2}^{n}(k-1)\prod_{j=k}^{n}\left(1+\binom{j}{2}^{-1}u\right)^{-1}, \quad u>0$$

R.C. Griffiths (personal communication) described the limit of the Laplace transform of $nB_n/2$ when n goes to infinity as

$$2\int_0^1 z\exp\left(-u\frac{(1-z)}{z}\right)\mathrm{d}z = \int_0^\infty \mathrm{e}^{-uy}\cdot\frac{2}{(1+y)^3}\mathrm{d}y$$

Consequently the function $\tilde{f}$ defined in Lemma 3.2 is the limiting density function of $nB_n/2$ when $n\to\infty$. This result was also stated in [12].

3.2.3. Simulations

According to the previous results, the Popovic model is expected to behave as the coalescent model. For this to be true, the coalescent branch lengths need to be multiplied by $n/2$ and t,n

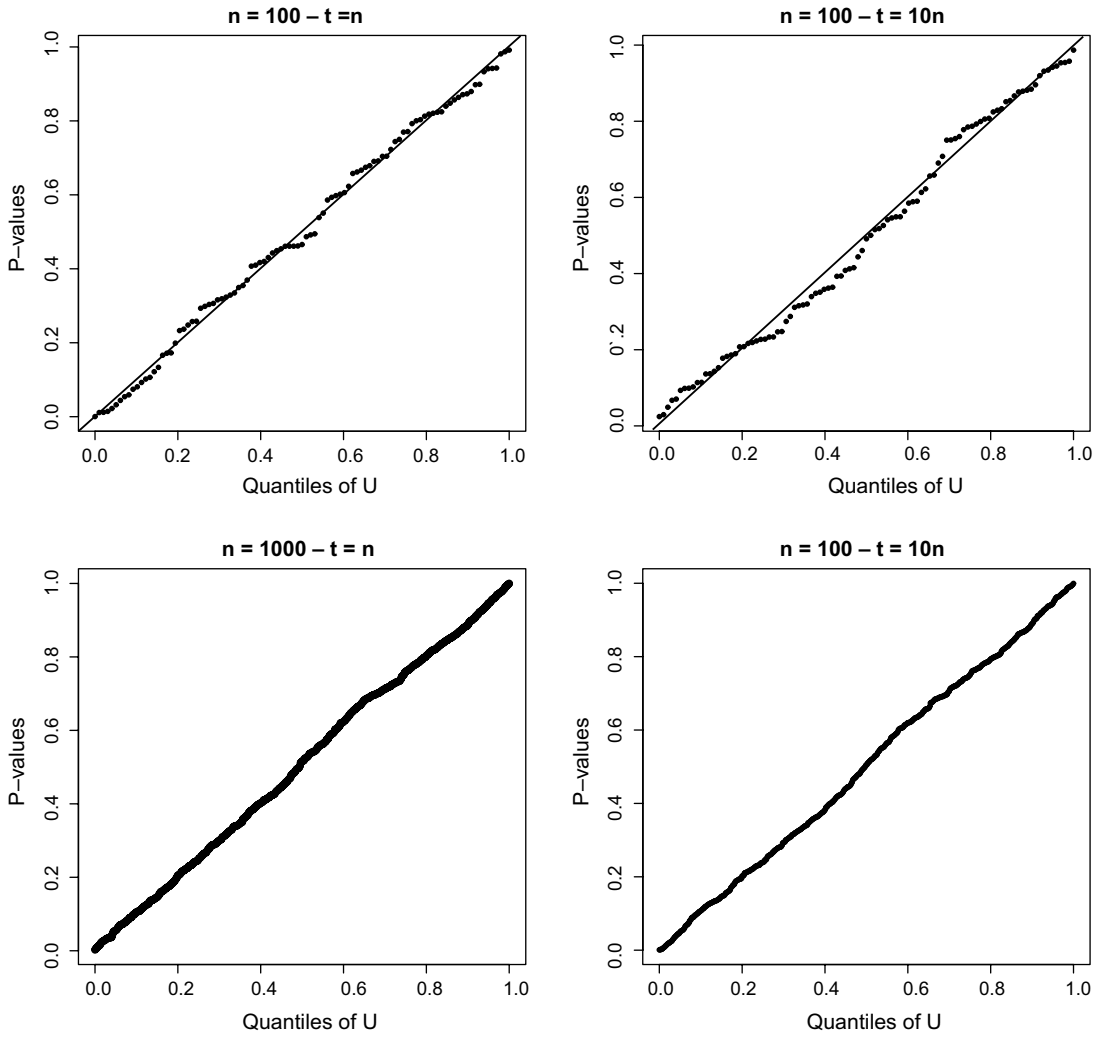

Fig. 2. U-plots for coalescent times resized by $n/2$ against a sample of the Popovic distribution with the same sample size $n = 100, 1000$ and $t = n$, $10\times n$.

should be large. In this paragraph, we present a brief simulation study which confirm this expectation, and which allows us to estimate how large n and t should be to get the approximation accurate.

We simulated coalescent times corresponding to $n = 100$ and 1000 tips. After multiplying by $n/2$, we compared these branching times with i.i.d. replicates from the Popovic model. We used two values of t, $t = n$ and $10n$. We computed the cdf of the Popovic model as $F_t(x) = (1 - t^{-1})/(1 - x^{-1})$, for x in $(0, t)$. Fig. 2 reports U-plots of the Popovic distribution against one particular sample of coalescent branching times (T_i), $i = 2, \ldots, n$. (U-plots display the P-values $P = F_t(t_i)$ against the quantiles of the uniform distribution). The graphics display a convincing agreement between the two distribution as $n \geqslant 100$. Note that $t = n$ roughly corresponds to the height of the root in the coalescent tree after the normalization ($t_{\mathrm{mrca}} \times n/2 = 2 \times n/2 = n$). Although t is not large compared to the time since the most recent common ancestor of the n tips, the fit is nevertheless accurate. Of course, the fit was even given stronger support when many samples were used (not reported). The same experiments were also conducted for the external branch lengths, and the results were consistent with those presented here.

4. Branch length statistics

One goal of this study is to devise branch length statistics for use in testing null models of phylogenetic trees such as the coalescent, the Hey model or other branching processes. In this section, we introduce two natural such statistics, for which the asymptotic distributions will be characterized under the Popovic model. According to the results of Section 3, the tests deduced from these distributions may also be applied to coalescent models. The method for obtaining the asymptotic distributions follows the same lines as Section 2, and makes use of Stein's approximation.

Given a tree with n tips and its branch lengths, the statistics under consideration are the *total height of cherries*, S_{c}, and the *sum of external branch lengths*, S_{e}. The first index is a natural extension of the number of cherries, while the second one is a more classical statistic which also plays an important role in testing the neutrality of mutations in population genetics [19,14].

More formally, we let X_i denote the indicator of a minimum height at i in the sequence H_{i-1}, H_i, H_{i+1} (see Section 2). Then we let

$$Y_i = H_i X_i, \quad i \in [n-1]$$

and we define the total height of cherries as

$$S_{\mathrm{c}} = \sum_{i=1}^{n-1} Y_i$$

Turning to the external branch lengths, we let Z_i denote the length of the branch ending at tip i. According to Section 3, we have

$$Z_i = \begin{cases} H_1 & \text{if } i = 1 \\ \min(H_{i-1}, H_i) & \text{if } i = 2, \ldots, n-1 \\ H_{n-1} & \text{if } i = n \end{cases}$$

Summing over the n tips, we obtain the total length of external branches as

$$S_e = \sum_{i=1}^{n} Z_i$$

Under the Popovic model, the two previous statistics involve local dependencies. In the definition of the height of cherries, Y_i and Y_j are independent as soon as $|i-j|>2$, while in the definition of external branch lengths Z_i and Z_j are independent when $|i-j|>1$. This short-range dependency property is the basis for applying Stein's method. For each statistic we first compute its mean and variance in the next lemmas.

Lemma 4.1. *Let $n \geqslant 3$ and $t>0$. Consider a realization of the Popovic model. Then, as $t \to \infty$, we have*

$$E[S_c] = \frac{n+3}{6} + o(1)$$

and

$$\mathrm{Var}[S_c] = \frac{53n}{180} + 2\log(t+1) - \frac{229}{36} + o(1)$$

Proof. These results can be obtained from direct computations using the definition of S_c. The exact values are

$$E[S_c] = \frac{t^2+2t-2(1+t)\log(1+t)}{t^2} + \frac{(n-3)}{3}\,\frac{t^3-3t^2-6t+6(t+1)\log(t+1)}{2t^3}$$

and

$$\mathrm{Var}[S_c] = A_t n + B_t$$

where

$$\begin{aligned} A_t = \frac{1}{180t^6}[&53t^6+740t^5+575t^4-780t^3-660t^2 \\ &+(-360t^5-1020t^4+600t^3+2820t^2+1560t)\log(t+1) \\ &+(-360t^3-1620t^2-2160t-900)(\log(t+1))^2], \end{aligned}$$

and

$$\begin{aligned} B_t = \frac{1}{36t^6}[&-229t^6-708t^5-191t^4+804t^3+516t^2 \\ &+(72t^6+552t^5+588t^4-1464t^3-2844t^2-1272t)\log(t+1) \\ &+(72t^4+792t^3+2124t^2+2160t+756)(\log(t+1))^2]. \end{aligned}$$

Regarding the external branch lengths, we obtain the following result. □

Lemma 4.2. *Let $n \geqslant 3$ and $t > 0$. Consider a realization of the Popovic model. Then, as $t \to \infty$, we have*

$$E[S_e] = 2\log(1+t) + n - 4 + o(1)$$

and

$$\mathrm{Var}[S_e] = (2\log(t+1) - 4)n + 2t - 8\log(t+1) + 6 + o(1)$$

Proof. The exact values are

$$(E[S_e] = \frac{1}{t^2}(2t^2 + 6t - 2n(t+1) + 4)\log(t+1) + (n-4)t^2 + (2n-4)t)$$

and

$$\mathrm{Var}[S_e] = A_t n + B_t$$

with

$$\begin{aligned} A_t = \frac{2}{t^4}[&-2t^4 - 2t^3 \\ &+ (t^4 + 8t^3 + 13t^2 + 6t)\log(1+t) \\ &+ (-2t^3 - 10t^2 - 14t - 6)(\log(1+t))^2], \end{aligned}$$

and

$$\begin{aligned} B_t = \frac{2}{t^4(1+t)}[&t^6 + 3t^5 + 6t^4 + 3t^3 - 2t^2 \\ &+ (-4t^5 - 28t^4 - 62t^3 - 52t^2 - 14t)\log(1+t) \\ &+ (10t^4 + 46t^3 + 78t^2 + 58t + 16)(\log(1+t))^2]. \end{aligned}$$ □

The main result in this article is a Gaussian approximation for the distributions of S_c and S_e. It is stated in the following theorem.

Theorem 4.3. *Let $t > 0$. In a lineage tree with n tips distributed according to the Popovic model, the total height of cherries S_c and the total external branch length S_e converge to a Gaussian distribution*

$$\frac{S_c - E[S_c]}{\mathrm{Var}[S_c]} \to \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

and

$$\frac{S_e - E[S_e]}{\mathrm{Var}[S_e]} \to \mathcal{N}(0,1) \quad \text{as } n \to \infty$$

As for the number of cherries in Section 2, we obtain the result by considering variables associated to a dependency graph. Before giving a proof, we need to recall some definitions. Let us consider a set of random variables $\{X_i, i \in \mathcal{V}\}$ indexed by the vertices of a graph $G = (\mathcal{V}, \mathcal{E})$. G is said to be a *dependency graph* if for any pair of disjoints sets A_1, A_2 in $\mathcal{V}$ such that no edge

in $\mathcal{E}$ has one endpoint in A_1 and the other in A_2, the sets of random variables $\{X_i, i \in A_1\}$ and $\{X_i, i \in A_2\}$ are independent.

The main argument of the proof is a result about sums of locally dependent variables, i.e. variables associated to a dependency graph. This theorem is based on Stein's method for normal approximations.

Theorem 4.4. (Baldi and Rinott [6]) *Let $\{X_i, i \in \mathcal{V}\}$ be random variables having a dependency graph $G = (\mathcal{V}, \mathcal{E})$. Set $W = \sum_{i \in \mathcal{V}} X_i$ and $\sigma^2 = Var[W]$. Let D denote the maximal degree of G and suppose $|X_i| \leqslant B$ a.s. Define*

$$Q = \frac{|\mathcal{V}| D^2 B^3}{\sigma^3}$$

Then

$$\left| P\left(\frac{W - E[W]}{\sigma} \leqslant w \right) - \Phi(w) \right| \leqslant 32(1 + \sqrt{6}) Q^{1/2} \tag{1}$$

To conclude the proof of our theorem, we need to consider a dependency graph for each set, $\{Y_i, i \in [n-1]\}$ and $\{Z_i, i \in [n]\}$. For the Y_i's, a vertex $3 \leqslant i \leqslant n-3$ has 4 neighbors $\{i-2, i-1, i+1, i+2\}$ and $i = 2, n-2$ (resp. $i = 1, n-1$) have 3 (resp. 2 neighbors). Hence we have $D = 4$. B can be taken equal to t and $|\mathcal{V}| = n-1$. For the Z_i's, a vertex $2 \leqslant i \leqslant n-2$ has 2 neighbors $\{i-1, i+1\}$ and $i = 1, n-1$ have a single neighbor. Here we have $D = 2$. In both cases, we obtain a stronger result than a simple TCL (namely a Berry–Essen convergence theorem). □

Remark. Another simple statistic is the sum of branching times

$$S_n = \sum_{i=1}^{n-1} H_i.$$

The sum of branching times is related to the sum of branch lengths L_n (phylogenetic diversity) as $L_n = S_n + \max H_i$. Under the Popovic model, the sum S_n has mean

$$E[S_n] = \left(\frac{1+t}{t} \log(1+t) - 1 \right)(n-1)$$

and variance

$$\mathrm{Var}[S_n] = \left(1 + t + \frac{1+t}{t} \log(1+t) - \left(\frac{1+t}{t} \log(1+t) \right)^2 \right)(n-1),$$

and the Gaussian approximation is valid for L_n according to the central-limit theorem for i.i.d. random variables. Note that the finite sample size distribution of the length of a coalescent tree is also well-known (see, e.g., [23]).

To conclude this article, we illustrate the above mathematical results with a brief application to two phylogenetic trees. These data sets are available from the R package *ape* [37]. We also used the *apTreeshape* package to reanalyze them [11].

The first data set describes an estimated clock-like phylogeny of 193 HIV-1 group M sequences sampled in the Democratic Republic of Congo [50,47]. Yusim et al. [50] argued that this set of sequences has several properties that make it suitable for analysis using coalescent theory, enabling them to make inferences about the demographic history of the HIV-1 epidemic. Using a variable population size model, they found strong evidence of demographic expansion. Here we use the external branch length statistic to reject the constant population size model, and we also test the null hypothesis that the tree has the Harding topology.

From the R data, the time to the most recent ancestor of the HIV-1 tree can be computed as $t_{\mathrm{mrca}} = 0.2091$. This time is 10-fold less than the usual average coalescent value ($t_{\mathrm{mrca}} \approx 2$). To account for an efficient population size, we rescaled the branch lengths of the HIV-1 tree by multiplying them by a factor of 10. Then applying the $n/2$ normalization, we found a value of the sum of external branch lengths equal to $S_{\mathrm{e}} = 16338$. The Gaussian approximation (used with $t = n$) and simulations of the rescaled coalescent model provided a one-sided P-value around $P \approx 0$. The number of cherries provided a P-value around $P \approx 0.13$ (Gaussian approximation). The shape statistic s used in [9] was equal to $s = 3.48$, and using it to test the Harding topology led to a P-value $P \approx 0.0005$. The constant population size coalescent model should therefore be rejected. Fig. 3 (Left) displays the distribution of the normalized branching times, and we can observe a poor agreement with the power-laws characterizing the Popovic model. The bell-shaped curve observed in Fig. 3 (Left) is typical of star-like phylogenetic trees, and may indeed reflect population expansion. Because the Harding topology is also implausible for this data set, a possible explanation of the departure from a clock-like phylogeny may be that the HIV-1 has evolved under strong positive selection.

The second data set describes the phylogenetic relationships of the families of birds as reported by Sibley and Ahlquist [45] who inferred this phylogeny from an extensive number of DNA/DNA hybridization experiments. One unresolved node (node 64) was resolved arbitrarily

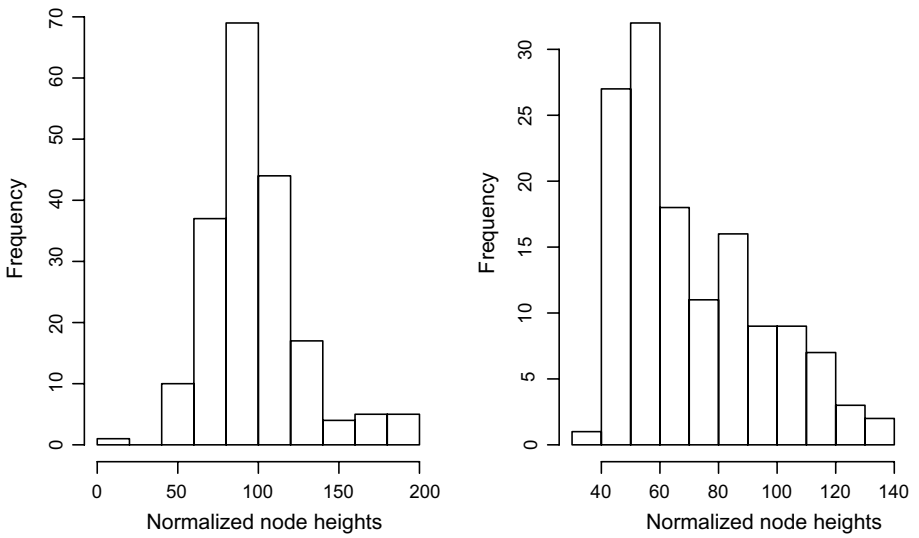

Fig. 3. Histograms of normalized branching times. Left: HIV-1 phylogeny. Right: Bird families.

without major influence on the output results. The histogram of branch lengths seemed more consistent with a (shifted) power-law distribution (Fig. 3 Right). Normalizing as in the first example ($t_{\mathrm{mrca}} = 28$), we summed the external branch lengths to find a value equal to $S_{\mathrm{e}} = 8511$. The Gaussian approximation (used with $t = n$) and simulations of the rescaled coalescent model provided a one-sided P-value around $P \approx 0$. The test based on the shape statistic s provided a P-value $P \approx 9.57\mathrm{e} - 10$. The tree is underbalanced compared to the Harding distribution, and the external branches are much longer than expected under the Hey model, which is again strongly rejected.

Acknowledgements

The authors are grateful to Michael Blum for his comments on a preliminary version of the manuscript. They also thank an anonymous reviewer for her/his constructive comments which contributed to improve the manuscript. This work was supported by grants from the Agence Nationale de la Recherche project MAEV "Modèles Aléatoires pour l'Evolution du Vivant".

References

[1] P.M. Agapow, A. Purvis, Power of eight tree shape statistics to detect non-random diversification: a comparison by simulation of two models of cladogenesis, Syst. Biol. 51 (2002) 866–872.
[2] D.J. Aldous, Probability distributions on cladograms, in: D.J. Aldous, R. Pemantle (Eds.), Random Discrete Structures, IMA Volumes Math. Appl, vol. 76, Springer, Berlin, 1996, pp. 1–18.
[3] D.J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to Today, Stat. Sci. 16 (2001) 23–34.
[4] P. Aldous, L. Popovic, A critical branching process model for biodiversity, Adv. Appl. Prob. 37 (2005) 1094–1115.
[5] P. Baldi, Y. Rinott, C. Stein, A normal approximation for the number of local maxima of a random function on a graph, in: Probability, Statistics, and Mathematics, Academic Press, Boston, 1989, pp. 59–81.
[6] P. Baldi, Y. Rinott, On normal approximations of distributions in terms of dependency graphs, Ann. Probab. 17 (1989) 1646–1650.
[7] M.G.B. Blum, O. François, Minimal clade size and external branch length under the neutral coalescent, Adv. Appl. Prob. 37 (2005) 647–662.
[8] M.G.B. Blum, O. François, On statisticals tests of phylogenetic tree imbalance: the Sackin and other indices revisited, Math. Biosci. 195 (2005) 141–153.
[9] M.G.B. Blum, O. François, Which random processes describe the Tree of Life? A large scale study of phylogenetic tree imbalance, Syst. Biol. 55 (2006) 685–691.
[10] M.G.B. Blum, O. François, S. Janson, The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, Ann. Appl. Probab. 16 (2006) 2195–2214.
[11] N. Bortolussi, E. Durand, M.G.B. Blum, O. François, ApTreeshape: Statistical analysis of phylogenetic treeshape, Bioinformatics 22 (2006) 363–364.
[12] A. Caliebe, R. Neininger, M. Krawczak, U. Rösler, The length of external branches in coalescent trees, in: 33rd European Mathematical Genetics Meeting, EMGM05, Annals of Human Genetics, vol. 69, 2005, p. 764.
[13] D.H. Colless, Review of phylogenetics: the theory and practice of phylogenetic systematics, Syst. Zool. 31 (1982) 100–104.
[14] R. Durrett, Probabilistic Models of DNA Sequences, Springer Verlag, New York, 2002.
[15] R. Durrett, Probability: Theory and Examples, third ed., Duxbury Press, Belmont, CA, 2005.
[16] W.J. Ewens, Mathematical Population Genetics, Springer, New York, 1978.

[17] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2003.
[18] D.P. Faith, Conservation evaluation and phylogenetic diversity, Biol. Conservat. 61 (1992) 1.
[19] Y.X. Fu, W.H. Li, Statistical tests of neutrality of mutations, Genetics 133 (1993) 93–709.
[20] E.F. Harding, The probabilities of rooted tree- shapes generated by random bifurcation, Adv. Appl. Prob. 3 (1971) 4–77.
[21] T.E. Harris, The Theory of Branching Processes, Springer Verlag, Berlin, 1964.
[22] P.H. Harvey, A.J.L. Brown, J.M. Smith, S. Nee, New Uses for New Phylogenies, Oxford University Press, Oxford, UK, 1996.
[23] J. Hein, M.H. Schierup, C. Wiuf, Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory, Oxford University Press, Oxford, UK, 2005.
[24] J. Hey, Using phylogenetic trees to study speciation and extinction, Evolution 46 (1992) 627–640.
[25] J.F.C. Kingman, The coalescent, Stoch. Proc. Appl. 13 (1982) 235–248.
[26] M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution 47 (1993) 1171–1181.
[27] G.M. Mace, J.L. Gittleman, A. Purvis, Preserving the Tree of Life, Science 300 (2003) 1707–1709.
[28] A. McKenzie, M. Steel, Properties of phylogenetic trees generated by Yule-type speciation models, Math. Biosci. 170 (2000) 91–112.
[29] A.Ø. Mooers, S.B. Heard, Inferring evolutionary process from phylogenetic tree shape, Quart. Rev. Biol. 72 (1997) 1–54.
[30] A.Ø. Mooers, L.J. Harmon, M.G.B. Blum, D.H.J. Wong, S.B. Heard, Some models of phylogenetic tree shape, in: O. Gascuel, M. Steel (Eds.), Reconstructing Evolution: New Mathematical and Computational Advances, Oxford University Press, Oxford, 2007, in press.
[31] B.R. Moore, K.M.A. Chan, M.J. Donoghue, Detecting diversification rate variation in supertrees, in: O.R.P. Bininda-Edmonds (Ed.), Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, Computational Biology, vol. 3, Kluwer Academic Publishers, Netherlands, 2004, pp. 487–533.
[32] P.A.P. Moran, Random processes in genetics, Proc. Camb. Philos. Soc. 54 (1958) 60–72.
[33] S. Nee, R.M. May, P.H. Harvey, The reconstructed evolutionary process, Philos. Trans. Roy. Soc. Lond. B 344 (1994) 305–311.
[34] S. Nee, T.G. Barraclough, P.H. Harvey, Temporal changes in biodiversity: detecting patterns and identifying causes, in: K. Gaston (Ed.), Biodiversity, Oxford University Press, Oxford, UK, 1996, pp. 230–252.
[35] S. Nee, R.M. May, Extinction and the loss of evolutionary history, Science 278 (1997) 692–694.
[36] M. Nordborg, Coalescent theory, in: D.J. Balding et al. (Eds.), Handbook of Statistical Genetics, Wiley, New York, 2003, pp. 179–208.
[37] E. Paradis, J. Claude, K. Strimmer, APE: analyses of phylogenetics and evolution in R language, Bioinformatics 20 (2004) 289–290.
[38] L. Popovic, Asymptotic genealogy of a critical branching process, Ann. Appl. Probab. 14 (2004) 2120–2148.
[39] A. Purvis, A. Hector, Getting the measure of biodiversity, Nature 405 (2000) 212–219.
[40] O.G. Pybus, P.H. Harvey, Testing macro-evolutionary models using incomplete molecular phylogenies, Proc. R. Soc. Lond. B 267 (2000) 2267–2272.
[41] Y. Rinott, V. Rotar, Normal approximations by Stein's method, Decisions Econ. Finance 23 (2000) 15–29.
[42] N.A. Rosenberg, The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees, Ann. Combinatorics 10 (2006) 129–146.
[43] M.J. Sackin, Good and bad phenograms, Syst. Zool. 21 (1972) 225.
[44] K. Shao, R.R. Sokal, Tree balance, Syst. Zool. 39 (1990) 266.
[45] C.G. Sibley, J.E. Ahlquist, Phylogeny and Classification of Birds: A Study in Molecular Evolution, Yale University Press, New Haven, 1990.
[46] C. Stein, Approximate computation of expectations, in: S.S. Gupta (ed.), Institute of Mathematical Statistics Lecture Notes-Monograph Series vol. 7, Hayward California, 1986.
[47] K. Strimmer, O.G. Pybus, Exploring the demographic history of DNA sequences using the generalized skyline plot, Mol. Biol. Evol. 18 (2001) 2298–2305.

[48] S. Tavaré, Ancestral inference in population genetics, in: Lectures on Probability Theory and Statistics, Lecture Notes Math. 1837, Springer, Berlin, 2004, pp. 1–188.
[49] G.U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, Philos. Trans. Roy. Soc. Lond. Ser. B 213 (1924) 21–87.
[50] K. Yusim, M. Peeters, O.G. Pybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, B. Korber, Using HIV-1 sequences to infer historical features of the AIDS epidemic and HIV evolution, Philos. Trans. R. Soc. Lond. B 356 (2001) 855–866.
[51] K. Wollenberg, J. Arnold, J.C. Avise, Recognizing the forest for the trees: testing temporal patterns of cladogenesis using a null model of stochastic diversification, Mol. Biol. Evol. 13 (1996) 833–849.