On the Distribution of the Distances Between Pairs of Leaves in Phylogenetic Trees

Arnau Mir
Department of Math and Computer Science
University of Balearic Islands
Palma de Mallorca, Spain
arnau.mir@uib.es

Francesc Rosselló
Department of Math and Computer Science
University of Balearic Islands
Palma de Mallorca, Spain
cesc.rossello@uib.es

***Abstract*—The distance, or path length, between two nodes in a phylogenetic tree (rooted or unrooted) is defined as the length of the unique undirected path connecting these nodes. In this paper we study the distribution of the distances between pairs of leaves in fully resolved phylogenetic trees with a fixed number of leaves. More precisely, we prove both in the unrooted and the rooted cases that, when the trees are equiprobably chosen, this distribution approximates a gamma distribution.**

***Keywords*-phylogenetic trees; statistical distribution;**

I. INTRODUCTION

Over the last years there has been an increasing interest in the study of the statistical behaviour of topological features in phylogenetic trees under different evolution models [2], [3], [9], [10]. The motivations for such studies are the assessment of the validity of an evolutionary model for a given set of phylogenetic trees, and the objective evaluation of how atypical a given phylogenetic tree is.

One feature whose behaviour has been studied is the topological distance, or path length, between pairs of leaves. Steel and Penny [11] computed the mean value and the variance of this distance d between two leaves in a fully resolved unrooted phylogenetic tree with n leaves. The statistical analysis of this random variable was continued in [6], where Steel and Penny's results were generalized to rooted phylogenetic trees, and in [7], [8], where the median and the mode of d were computed, both in the rooted and the unrooted cases. Let us mention that the study of the distance between pairs of leaves has a further motivation, as it may be used in the study of the statistical properties of the nodal distance between phylogenetic trees, an interesting and mostly open problem in phylogenetics [6], [11].

In this paper, instead of focusing on the exact computation of the statistical measures for d, we focus on its distribution, and we prove that, in the fully resolved case, it is approximately a gamma distribution, in the sense that the mean quadratic error between the distribution of d and a gamma distribution of the same mean and mode has limit 0 as $n \to \infty$.

The rest of this paper is organized as follows. In Section II, we prove our main result for unrooted fully resolved trees. Then, in Section III we briefly describe how this result translates to the rooted case, and in Section IV we report on some experimental results showing the fast convergence between d and the corresponding gamma distribution. The paper ends with a Conclusions section.

II. THE UNROOTED CASE

Throughout this paper, by a *phylogenetic tree* on a set S we mean a *fully resolved* (that is, with all its internal nodes of degree 3) unrooted tree with its leaves bijectively labelled in the set S. Although in practice S may be any set of taxa, to fix ideas we shall always take $S = \{1,\dots,n\}$, where n is the number of tree leaves. For simplicity, we shall always identify a leaf of a phylogenetic tree with its label.

Let $\mathscr{T}_n^u$ be the set of (isomorphism classes of) phylogenetic trees with n leaves. It is well known [4] that $|\mathscr{T}_1^u| = |\mathscr{T}_2^u| = 1$ and $|\mathscr{T}_n^u| = (2n-5)!! = (2n-5)(2n-7)\cdots 3\cdot 1$, for every $n \geqslant 3$.

Let $k,l \in S = \{1,\dots,n\}$ be any two different labels of trees in $\mathscr{T}_n^u$. The *distance*, or *path length*, $d_T^u(k,l)$ between the leaves k and l in a phylogenetic tree $T \in \mathscr{T}_n^u$ is the length of the unique path between them. Let's consider the random variable

$$d_{kl}^u = \text{distance between } k \text{ and } l \text{ in one tree in } \mathscr{T}_n^u.$$

The possible values of d_{kl}^u are $\Omega^u = \{1,2,\dots,n-1\}$.

Our goal is to approximate the distribution of the variable d_{kl}^u on $\mathscr{T}_n^u$ when the tree and their leaves are chosen equiprobably. In this case, $d_{kl}^u = d_{12}^u$, and thus we can reduce our problem to study the distribution of the variable $d_n^u := d_{12}^u$.

For every $i \in \Omega^u$, let

$$c_{i,n}^u = \frac{|\{T \in \mathscr{T}_n^u \mid d_T^u(1,2) = i\}|}{(2n-5)!!}$$

denote the fraction of trees in $\mathscr{T}_n^u$ where the leaves 1 and 2 are at distance i. The sequence $(c_{i,n}^u)_{i=1,\dots,n-1}$ is the distribution of the variable d_n^u. From [11, p. 140] and [8], we have the following result.

Lemma 1: (a) $c_{n-1}^u = \frac{(n-2)!}{(2n-5)!!}$ and, for every $i = 1,\dots,n-2$,

$$c_{i,n}^u = \frac{(i-1)(2n-i-4)!}{(2(n-i-1))!!\cdot(2n-5)!!}.$$

 ISBN: 978-1-61208-137-3

(b) The mean of d_n^u is

$$\mu_n = \sum_{i=2}^{n-1} i c_{i,n}^u = \frac{2^{n-2}(n-2)!}{(2n-5)!!}$$

(c) The mode of d_n^u is

$$m_n = \left\lceil \frac{1+\sqrt{8n-15}}{2} \right\rceil.$$

Let $\gamma(k,\theta)$ denote a gamma distribution with parameters k (shape) and θ (scale), and let $f_{\gamma(k,\theta)}$ be its density function. Recall that the mean of $\gamma(k,\theta)$ is $k\cdot\theta$ and its mode is $(k-1)\cdot\theta$. Our goal in this section is to prove the following result:

Theorem 1: Consider the gamma distribution $\gamma(k_n,\theta_n)$, with parameters k_n and θ_n given by

$$k_n \cdot \theta_n = \mu_n, \quad (k_n-1)\cdot\theta_n = m_n,$$

where

$$\begin{aligned} k_n &= \frac{2^{n-2}\cdot(n-2)!}{2^{n-2}\cdot(n-2)! - (2n-5)!!\lceil(\sqrt{8n-15}+1)/2\rceil}, \\ \theta_n &= \frac{2^{n-2}(n-2)!}{(2n-5)!!} - \left\lceil \frac{1+\sqrt{8n-15}}{2} \right\rceil. \end{aligned}$$

In other words, $\gamma(k_n,\theta_n)$ is the gamma distribution with the same mean and mode as d_n^u on $\mathscr{T}_n^u$. Let MQE_n^u be the mean quadratic error between the random variable d_n^u and this gamma distribution:

$$MQE_n^u = \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_n,\theta_n)}(i))^2.$$

Then, $\lim_{n\to\infty} MQE_n^u = 0$.

Proof: Let $g_n(x)$ be the following function:

$$g_n(x) = \frac{(x-1)\cdot 2^{x-1}\cdot\Gamma(2n-x-3)\cdot\Gamma(n-1)}{\Gamma(n-x)\cdot\Gamma(2n-3)}.$$

This function satisfies that $g_n(i) = c_{i,n}^u$ for every $i = 1,\ldots,n-1$, and therefore it can be seen as the extension to $\mathbb{R}^+$ of the discrete distribution of d_n^u.

The sequence $(g_n(i))_{n=1,\ldots,n-1}$ reaches its maximum at m_n of d_n^u. We want to approximate $g_n(m_n)$. To do that, we shall use the following expansion of the logarithm of the Gamma function:

$$\ln\Gamma(x) \approx \frac{\ln(2\pi)}{2} + \left(x-\frac{1}{2}\right)\ln\left(x-\frac{1}{2}\right) - \left(x-\frac{1}{2}\right), \quad (1)$$

for large values of x. Using this expansion and using that

$$m_n = \sqrt{2n} + \frac{1}{2} + O((1/n)^{1/2}),$$

the expansion of the value of $\ln g_n(m_n) = \ln g_n\left(\sqrt{2n}+\frac{1}{2}+O\left(\left(\frac{1}{n}\right)^{1/2}\right)\right)$ is

$$\ln g_n(m_n) = \frac{1}{2}\left(-1+\ln\left(\frac{1}{2n}\right)\right) + O\left(\left(\frac{1}{n}\right)^{1/2}\right).$$

So, we can approximate $g_n(m_n)$ by

$$g_n(m_n) = \frac{\mathrm{e}^{-1/2}}{\sqrt{2n}} + O(1/n).$$

Next, we study the value of $f_{\gamma(k_n,\theta_n)}(m_n)$.

Lemma 2: The expansions of the parameters μ_n, k_n and θ_n are the following:

$$\begin{aligned} \mu_n &= \sqrt{\pi}\sqrt{n} + O(1/n), \\ k_n &= \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}} + O(1/n), \\ \theta_n &= (\sqrt{\pi}-\sqrt{2})\sqrt{n} - \frac{1}{2} + O(1/n). \end{aligned}$$

Proof: The parameter θ_n can be written as:

$$\theta_n = \frac{1}{4}\left(\frac{2^n(n-2)!}{(2n-5)!!} - 2\left(\sqrt{8n-15}+1\right)\right).$$

If we expand the previous expression, we obtain:

$$\theta_n = \frac{1}{4}\left(\frac{2^{2n-2}((n-2)!)^2}{(2n-4)!} - 4\sqrt{2n} - 2\right) + O\left(\frac{1}{\sqrt{n}}\right).$$

Using that $(n-2)! = \Gamma(n-1)$ and the expansion (1), we have:

$$\begin{aligned} \theta_n &= \frac{1}{4}\left(4\sqrt{\pi}\sqrt{n}\mathrm{e}^{O\left(\frac{1}{n}\right)} - 4\sqrt{2n} - 2\right) + O\left(\frac{1}{\sqrt{n}}\right), \\ &= \sqrt{\pi n} - \sqrt{2n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right), \\ &= \left(\sqrt{\pi}-\sqrt{2}\right)\sqrt{n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Thus, the expression for the parameter θ_n is obtained.

Next, we will proceed similarly with the parameter μ_n. This parameter can be written as:

$$\mu_n = \frac{2^{n-2}(n-2)!}{(2n-5)!!} = \frac{2^{2n-4}(n-2)!^2}{(2n-4)!}.$$

For the second time, using that $(n-2)! = \Gamma(n-1)$ and the expansion (1), we have:

$$\mu_n = \sqrt{\pi}\sqrt{n} + O\left(\frac{1}{n}\right).$$

Finally, using that $k_n = \frac{\mu_n}{\theta_n}$ and the previous expansions for the parameters μ_n and θ_n, we can obtain:

$$\begin{aligned} k_n &= \frac{\sqrt{\pi}\sqrt{n} + O\left(\frac{1}{n}\right)}{\left(\sqrt{\pi}-\sqrt{2}\right)\sqrt{n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right)} = \frac{\sqrt{\pi} + O\left(\frac{1}{n\sqrt{n}}\right)}{\sqrt{\pi}-\sqrt{2}+O\left(\frac{1}{n}\right)}, \\ &= \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}} + O\left(\frac{1}{n}\right), \end{aligned}$$

as we claimed.

 ISBN: 978-1-61208-137-3

Using these expansions, we obtain the following expression for the value of $f_{\gamma(k_n,\theta_n)}(m_n) = \frac{m_n^{k_n-1}\cdot e^{-\frac{m_n}{\theta_n}}}{\Gamma(k_n)\cdot\theta_n^{k_n}}$:

$$f_{\gamma(k_n,\theta_n)}(m_n) = \frac{\alpha}{\beta(n)},$$

where:

$$\begin{aligned}\alpha &= 2^{(1/2)\cdot(\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})-1)},\\ \beta(n) &= e^{\sqrt{2}/(\sqrt{\pi}-\sqrt{2})}\cdot\Gamma\left(\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})\right)\\ &\quad\cdot(\sqrt{\pi}-\sqrt{2})^{\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})}\cdot n^{-1/2}+O\left(\frac{1}{n}\right).\end{aligned}$$

We conclude that:

$$(g_n(m_n)+f_{\gamma(k_n,\theta_n)}(m_n))^2 = \frac{C}{n}+O\left(\frac{1}{n\sqrt{n}}\right),$$

where the constant C could be found using the expansions of $g_n(m_n)$ and $f_{\gamma(k_n,\theta_n)}(m_n)$.

Finally, an upper bound for the mean quadratic error MQE_n^u is found:

$$\begin{aligned}MQE_n^u &= \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_n,\theta_n)}(i))^2\\ &\leqslant \frac{n}{n-1}\cdot((g_n(m_n)+f_{\gamma(k_n,\theta_n)}(m_n))^2\\ &= \frac{C}{n-1}+O\left(\frac{1}{n\sqrt{n}}\right),\end{aligned}$$

and the right hand side term in this inequality tends to zero as n goes to infinity, as we claimed. This finishes the proof of Theorem 1.

III. The rooted case

By a *rooted phylogenetic tree* on S we mean a *fully resolved* (which in this case means with all its internal nodes of out-degree 2) rooted tree with its leaves bijectively labelled in the set S. As in the previous section, for simplicity we consider only the sets of labels $S_n = \{1,\ldots,n\}$, with n the number of leaves of the tree. Let $\mathscr{T}_n^r$ be the set of (isomorphism classes of) rooted phylogenetic trees on S_n. It is well known [4, Ch. 3] that $|\mathscr{T}_n^r| = |\mathscr{T}_{n+1}^u|$ for every $n \geqslant 1$.

Let $k,l \in S_n$ be any two different labels and let $T \in \mathscr{T}_n^r$. The *distance*, or *path length*, $d_T^r(k,l)$ between the leaves k and l in T is the length of the unique *undirected* path between them. We consider now the random variable

$$d_{kl}^r = \text{distance between } k \text{ and } l \text{ in one tree in } \mathscr{T}_n^r,$$

which takes values in $\Omega^r = \{2,3,\ldots,n\}$. Arguing as in Section II, when the trees and the leaves are chosen equiprobably, we are reduced to study the variable $d_n^r := d_{12}^r$.

For every $i \in \Omega^r$, let

$$c_{i,n}^r = |\{T \in \mathscr{T}_n^r \mid d_T^r(1,2) = i\}|.$$

The sequence $(c_{i,n}^r)_{i=2,\ldots,n}$ is the distribution of the variable d_n^r.

We have the following result connecting $c_{i,\cdot}^r$ with $c_{i,\cdot}^u$. For the sake of completeness, we sketch a direct proof, although it could be deduced from the explicit computations provided in [6], [11].

Lemma 3: $c_{i,n}^r = c_{i,n+1}^u$, for every $n \geqslant 2$ and $i = 2,\ldots,n$,

Proof: Consider the usual bijection $\Phi : \mathscr{T}_n^r \to \mathscr{T}_{n+1}^u$ that sends a rooted tree $T \in \mathscr{T}_n^r$ to the unrooted tree $\Phi(T) \in \mathscr{T}_{n+1}^u$ obtained by adding a new leaf labeled $n+1$ and a new edge connecting the root of T with this leaf (cf. [4, Ch. 3]). Then, $d_T^r(1,2) = d_{\Phi(T)}^u(1,2)$, and therefore Φ induces a bijection

$$\{T \in \mathscr{T}_n^r \mid d_T^r(1,2) = i\} \to \{T \in \mathscr{T}_{n+1}^u \mid d_T^u(1,2) = i\}.$$

This lemma allows one to translate Theorem 1 into the rooted case as follows:

Theorem 2: Let $\gamma(k_{n+1},\theta_{n+1})$ be the gamma distribution with parameters k_{n+1} and θ_{n+1} given by

$$\begin{aligned}k_n &= \frac{2^{n-2}\cdot(n-2)!}{2^{n-2}\cdot(n-2)!-(2n-5)!!\lceil(\sqrt{8n-15}+1)/2\rceil},\\ \theta_n &= \frac{2^{n-2}(n-2)!}{(2n-5)!!}-\left\lceil\frac{1+\sqrt{8n-15}}{2}\right\rceil.\end{aligned}$$

Let MQE_n^r be the mean quadratic error between the random variable d_n^r and this gamma distribution:

$$MQE_n^r = \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_{n+1},\theta_{n+1})}(i))^2.$$

Then, $\lim_{n\to\infty} MQE_n^r = 0$.

IV. Experimental results

Figure 1 shows the data plot of $(c_{i,n}^u)_{i=1,\ldots,n-1}$ and the gamma density function with parameters $k = \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}}$ and $\theta_n = (\sqrt{\pi}-\sqrt{2})\sqrt{n}$ as functions of i, for $n = 5000$ leaves. The figure confirms that the distribution of d_n^u approximates well this gamma density function.

Figure 2 shows the data plot of minus the logarithm of the mean quadratic error function $(-\ln(MQE_n))$ as a function of the number n of leaves. The curve hints at the existence of parameters α and β such that $-\ln(MQE_n) \approx \alpha+\beta\ln(n)$, that is, $MQE_n \approx e^{-\alpha}\cdot n^{-\beta}$. If we adjust the values of the α and β using the least squares method, we obtain $\alpha \approx 4.659$ and $\beta \approx 1.44$. This confirms the theoretical result in Section II.

V. Conclusion

In this paper, we have proven that the distribution of the distance between a fixed pair of leaves in an equiprobably, randomly chosen, fully resolved phylogenetic tree with n leaves approximates a gamma distribution as n goes to ∞. This result holds in the rooted and the unrooted case.

Our result is purely numerical, and it remains to be seen whether there is some deep meaning in the relationship between the distribution of the distances in phylogenetic

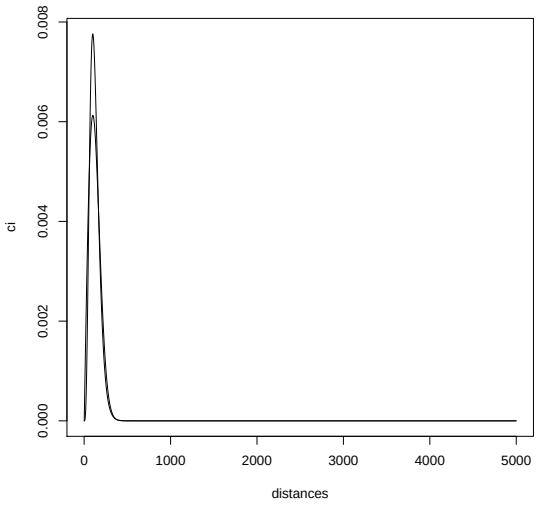

Figure 1. Data plot of $(c^u_{i,n})_{i=1,\ldots,n-1}$ and the corresponding gamma density function for $n = 5000$ leaves. The higher curve corresponds to the gamma density function, the lower one to d^u_n.

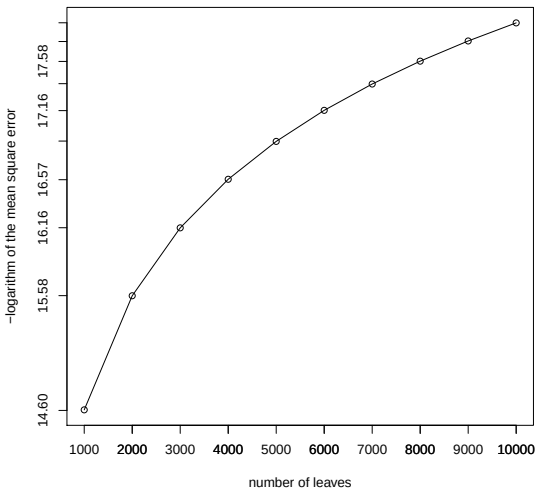

Figure 2. Data plot of $-\ln(MQE_n)$ as a function of the number n of leaves.

trees and a gamma distribution. Another unanswered question is whether the distances between pairs of leaves in the phylogenetic trees contained in some phylogenetic database, see TreeBASE ([1]) or PhylomeDB ([5]) are well approximated by using a gamma distribution. A negative answer would give information on the random model for real-life phylogenetic trees. We are working currently in this topic.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Government, through projects MTM2009-07165 and TIN2008-04487-E/TIN.

REFERENCES

[1] M. J. Sanderson, M. J. Donoghue, W. Piel, and T. Eriksson, TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81 (1994), pp. 183-189. http://www.treebase.org/ (last visited: March 11, 2011)

[2] M. Blum, N. Bortolussi, E. Durand, and O. François, AP-Treeshape: statistical analysis of phylo- genetic tree shape. Bioinformatics 22 (2006), pp. 363-364.

[3] H. Chang and M. Fuchs, Limit theorems for patterns in phylogenetic trees. Journal of Mathematical Biology 60 (2010), pp. 481-512.

[4] J. Felsenstein: Inferring Phylogenies. Sinauer Associates Inc. (2004)

[5] J. Huerta-Cepas, A. Bueno, J. Dopazo, and T. Gabaldón, PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Research 36 (2008), D491-6. http://phylomedb.org/ (last visited: March 11, 2011)

[6] A. Mir and F. Rosselló, The mean value of the squared path-difference distance for rooted phylogenetic trees. Journal of Mathematical Analysis and Applications 371 (2010), pp. 168-176

[7] A. Mir and F. Rosselló, The median of the distance between two leaves in a phylogenetic tree. Advances in Bioinformatics, Proc. IWPACBB 2010 (M.P. Rocha *et al*, eds.), Advances in Intelligent and Soft Computing, vol. 74 (Springer, 2010), pp. 131-135.

[8] A. Mir and F. Rosselló, The mode of the distance between two leaves in a phylogenetic tree. X Jornadas de Bioinformática (Málaga, Spain, october 2010).

[9] N. Rosenberg, The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Annals of Combinatorics 10 (2006), pp. 129-146.

[10] M. Steel and A. Mooers, The expected length of pendant and interior edges of a Yule tree. Applied Mathematics Letters 23 (2010), pp. 1315-1319

[11] M.A. Steel and D. Penny, Distributions of tree comparison metrics—some new results. Systematic Biology 41 (1993), pp. 126-141

 ISBN: 978-1-61208-137-3