Applied Mathematics Letters 25 (2012) 2031–2036

Contents lists available at SciVerse ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

The expected value under the Yule model of the squared path-difference distance

Gabriel Cardona, Arnau Mir, Francesc Rosselló *

Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca, Spain

ARTICLE INFO

Article history:
Received 14 March 2012
Received in revised form 18 April 2012
Accepted 18 April 2012

Keywords:
Phylogenetic tree
Nodal distance
Path-difference metric
Yule model
Sackin index

ABSTRACT

The path-difference metric is one of the oldest and most popular distances for the comparison of phylogenetic trees, but its statistical properties are still quite unknown. In this work we compute the expected value under the Yule model of evolution of its square on the space of fully resolved rooted phylogenetic trees with n leaves. This complements previous work by Steel and Penny and by Mir and Rosselló, who computed this mean value for fully resolved unrooted and rooted phylogenetic trees, respectively, under the uniform distribution.

1. Introduction

The definition and study of metrics for the comparison of rooted phylogenetic trees on the same set of taxa is a classical problem in phylogenetics [1, Chapter 30]. A classical and popular family of such metrics is based on the comparison, by different methods, of the vectors of lengths of the (undirected) paths connecting all pairs of taxa in the corresponding trees [2–5]. These metrics are generically called *nodal distances*, although some of them also have specific names. For instance, the metric defined through the euclidean distance between path-length vectors is called the *path-difference metric* [6], or the *cladistic difference* [2].

In contrast with those of other metrics, the statistical properties of these nodal distances are mostly unknown. Actually, the only statistical property that has been established so far for any one of them is the expected, or mean, value of the square of the path-difference metric for unrooted [6] and rooted [7] fully resolved phylogenetic trees under the uniform distribution (that is, when all phylogenetic trees with the same number of taxa are equiprobable). The knowledge of the expected value of a metric is useful, because it provides an indication about the significance of the similarity of two individuals measured through this metric [6].

But phylogeneticists consider also other probabilistic distributions on the space of phylogenetic trees on a fixed set of taxa, defined through stochastic models of evolution [1, Chapter 33]. The most popular such model is Yule's [8,9], defined by an evolutionary process where, at each step, each currently extant species can give rise, with the same probability, to two new species. Under this model, different phylogenetic trees with the same number of leaves may have different probabilities. Formal details of this model are given in the next section.

In this work we compute the expected value of the square of the path-difference metric for rooted fully resolved phylogenetic trees under the Yule model. Besides the aforementioned application of this value in the assessment of tree comparisons, the knowledge of formulas for this expected value under different models may allow the use of the path-difference metric to test stochastic models of tree growth, a popular line of research in the last few years which so far has been mostly based on shape indices [10].

* Corresponding author.
E-mail addresses: gabriel.cardona@uib.es (G. Cardona), arnau.mir@uib.es (A. Mir), cesc.rossello@uib.es (F. Rosselló).

0893-9659/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2012.04.013

The proof of our formula for this expected value is based on several long algebraic computations. Since the space constraints prevent us from providing full detail in these computations, we give here only the overall idea of their thread, which enables the willing reader to reproduce them, and we have posted a version of this work containing all details on the arXiv preprint server [11].

2. Preliminaries

In this work, by a *phylogenetic tree* on a set S of taxa we mean a fully resolved, or binary, rooted tree with its leaves bijectively labeled in S. We understand such a rooted tree as a directed graph, with its arcs pointing away from the root. To simplify the language, we shall always identify a leaf of a phylogenetic tree with its label. We shall also use the term *phylogenetic tree with n leaves* to refer to a phylogenetic tree on the set $\{1, \dots, n\}$. We shall denote by $\mathcal{T}(S)$ the space of all phylogenetic trees on S and by $\mathcal{T}_n$ the space of all phylogenetic trees with n leaves.

Whenever there exists a directed path from u to v in a phylogenetic tree T, we shall say that v is a *descendant* of u. The *distance* $d_T(u, v)$ between two nodes u, v in a phylogenetic tree T is the length (in numbers of arcs) of the unique undirected path connecting u and v. The *depth* $\delta_T(v)$ of a node v in T is the distance from the root r of T to v. The *path-difference distance* [2,3] between a pair of trees $T, T' \in \mathcal{T}_n$ is

$$d_\nu(T, T') = \sqrt{\sum_{1\leqslant i<j\leqslant n} (d_T(i,j) - d_{T'}(i,j))^2}.$$

The *Yule*, or *equal-rate Markov*, model of evolution [8,9] is a stochastic model of phylogenetic trees' growth. It starts with a node, and at every step a leaf is chosen randomly and uniformly and it is split into two leaves. Finally, the labels are assigned randomly and uniformly to the leaves once the desired number of leaves is reached. Under this model, if T is a phylogenetic tree with n leaves and set of internal nodes $V_{\mathrm{int}}(T)$, and if for every internal node v we denote by $\ell_T(v)$ the number of its descendant leaves, then the probability of T is [12,13]

$$P_Y(T) = \frac{2^{n-1}}{n!} \prod_{v\in V_{\mathrm{int}}(T)} \frac{1}{\ell_T(v) - 1}.$$

For every $n \geqslant 1$, let $H_n = \sum_{i=1}^n 1/i$ and $H_n^{(2)} = \sum_{i=1}^n 1/i^2$. Let, moreover, $H_0 = H_0^{(2)} = 0$. H_n is called the nth *harmonic number*, and $H_n^{(2)}$, the nth *generalized harmonic number of power* 2.

3. The main results

Let N_n^2 be the random variable that chooses independently a pair of trees $T, T' \in \mathcal{T}_n$ and computes $d_\nu(T, T')^2$. In this section we establish the following result.

Theorem 1. *The expected value of N_n^2 under the Yule model is*

$$E_Y(N_n^2) = \frac{2n}{n-1}\big(2(n^2 + 24n + 7)H_n + 13n^2 - 46n + 1 - 16(n+1)H_n^2 - 8(n^2-1)H_n^{(2)}\big).$$

To prove this theorem, we shall use the auxiliary random variables D_n and $D_n^{(2)}$ that choose a tree $T \in \mathcal{T}_n$ and compute $D(T) = \sum_{1\leqslant i<j\leqslant n} d_T(i,j)$ and $D^{(2)}(T) = \sum_{1\leqslant i<j\leqslant n} d_T(i,j)^2$, respectively. The connection between $E_Y(N_n^2)$ and the expected values under the Yule model of D_n and $D_n^{(2)}$ is given by the following result.

Proposition 2. $E_Y(N_n^2) = 2\big(E_Y(D_n^{(2)}) - E_Y(D_n)^2/\binom{n}{2}\big)$.

Proof. By developing $E_Y(N_n^2)$ from its raw definition, we obtain

$$\begin{aligned}
E_Y(N_n^2) &= \sum_{T,T'\in\mathcal{T}_n} d_\nu(T,T')^2 p_Y(T)p_Y(T') = \sum_{T,T'\in\mathcal{T}_n}\left(\sum_{1\leqslant i<j\leqslant n}(d_T(i,j) - d_{T'}(i,j))^2\right)p_Y(T)p_Y(T') \\
&= 2\sum_T\left(\sum_{1\leqslant i<j\leqslant n} d_T(i,j)^2\right)p_Y(T) - 2\sum_{1\leqslant i<j\leqslant n}\left(\sum_T d_T(i,j)p_Y(T)\right)^2 \\
&= 2E_Y(D_n^{(2)}) - 2\binom{n}{2}\left(\sum_T d_T(1,2)p_Y(T)\right)^2
\end{aligned}$$

and now

$$E_Y(D_n) = \sum_{T\in\mathcal{T}_n}\sum_{1\leqslant i<j\leqslant n} d_T(i,j)p_Y(T) = \sum_{1\leqslant i<j\leqslant n}\sum_{T} d_T(i,j)p_Y(T) = \binom{n}{2}\sum_{T} d_T(1,2)p_Y(T)$$

which implies that $\left(\sum_T d_T(1,2)p_Y(T)\right)^2 = E_Y(D_n)^2/\binom{n}{2}^2$, and the formula in the statement follows. □

It is known that the expected value under the Yule model of D_n is $E_Y(D_n) = 2n(n+1)H_n - 4n^2$ [14]. As regards $E_Y(D_n^{(2)})$, its value is given by the following result. We postpone the proof to the Appendix at the end of the work.

Proposition 3. *$E_Y(D_n^{(2)}) = 8n(n+1)(H_n^2 - H_n^{(2)}) - 2n(15n+7)H_n + 45n^2 - n.$*

Then, replacing in the expression for $E_Y(N_n^2)$ given in Proposition 2 the terms $E_Y(D_n)$ and $E_Y(D_n^{(2)})$ by their values, we obtain the formula for $E_Y(N_n^2)$ given in Theorem 1.

4. Conclusions

In this work we have computed the expected value $E_Y(N_n^2)$ of the square of the path-difference metric for rooted fully resolved phylogenetic trees under the Yule model. This complements the computation of this expected value under the uniform distribution carried out in [7].

The proof of the formula for $E_Y(N_n^2)$ consists of several long algebraic manipulations of sums of sequences. Since it is not difficult to make some mistake in such long algebraic computations, to double check our result we have computed the exact value of $E_Y(N_n^2)$ $(n = 3, \dots, 7)$, by generating all trees with up to seven leaves, as well as numerical approximations to $E_Y(N_n^2)$ $(n = 10, 20, \dots, 100)$, by generating random trees until the numerical method stabilizes. These numerical experiments confirm that our formula gives the right figures. The Python scripts used in these computations, as well as a full account of the results obtained, are available on the Supplementary Material web page http://bioinfo.uib.es/~recerca/phylotrees/nodaldistYule/.

The formulas for $E_Y(N_n^2)$ and $E_U(N_n^2)$ grow in different orders: $E_Y(N_n^2)$ is in $O(n^2\ln(n))$, while $E_U(N_n^2)$ is in $O(n^3)$ [7]. Therefore, they can be used to test the Yule and the uniform models as null stochastic models of evolution for collections of phylogenetic trees reconstructed by different methods. This kind of analysis has only been performed so far through shape indices of single trees, not by means of the comparison of pairs of trees. We shall report on it elsewhere.

Acknowledgments

This work was partially supported by the Spanish government and the UE FEDER program, through projects MTM2009-07165 and TIN2008-04487-E/TIN. We thank the reviewers for their suggestions, which have allowed us to improve this work.

Appendix

In this appendix we prove Proposition 3, as well as some preliminary lemmas. To begin with, the following identities on harmonic numbers will be systematically used in the subsequent proofs, usually without any further notice.

Lemma. *For every $n \geqslant 2$:*

$$(1)\ \sum_{k=1}^{n-1} H_k = n(H_n - 1) \quad (2)\ \sum_{k=1}^{n-1} kH_k = \frac{1}{4}n(n-1)(2H_n - 1) \quad (3)\ \sum_{k=1}^{n-1} H_k/(k+1) = \frac{1}{2}(H_n^2 - H_n^{(2)})$$

$$(4)\ \sum_{k=1}^{n-1} kH_kH_{n-k} = \binom{n+1}{2}(H_{n+1}^2 - H_{n+1}^{(2)} - 2H_{n+1} + 2) \quad (5)\ \sum_{k=1}^{n-1}(H_k^2 - H_k^{(2)}) = n(H_n^2 - H_n^{(2)}) - 2n(H_n - 1).$$

$$(6)\ \sum_{k=1}^{n-1} k(H_k^2 - H_k^{(2)}) = \binom{n}{2}(H_n^2 - H_n^{(2)}) - \frac{1}{4}n(n-1)(2H_n - 1).$$

Proof. Identities (1)–(3) are well known and easily proved by induction on n; see, for instance, [15, Section 1.2.7]. Identity (4) is proved in [16, Theorem 2]. Identities (5) and (6) are easily proved using Abel's lemma on summation by parts, as is done for other identities in [17]. □

Let us consider now the random variables S_n and $S_n^{(2)}$ that choose a tree $T \in \mathcal{T}_n$ and compute their *Sackin indexes* [18] $S(T) = \sum_{i=1}^{n} \delta_T(i)$ and $S^{(2)}(T) = \sum_{1\leqslant i<j\leqslant n} \delta_T(i)^2$, respectively. It is known that the expected value under the Yule model of S_n is $E_Y(S_n) = 2n(H_n - 1)$ [19]. We shall compute now the expected values under this model of $S_n^{(2)}$ and $D_n^{(2)}$: the first will be used in the computation of the second.

Given two phylogenetic trees T, T' on disjoint sets of taxa S, S', respectively, we shall denote by $T\widehat{\ }T'$ the phylogenetic tree on $S \cup S'$ obtained by connecting the roots of T and T' to a (new) common root. Every phylogenetic tree on $\{1, \dots, n\}$ is obtained as $T_k\widehat{\ }T'_{n-k}$, for some $1 \leqslant k \leqslant n-1$, some subset $S_k \subseteq \{1, \dots, n\}$ with k elements, some tree T_k on S_k and some tree T'_{n-k} on $S_k^c = \{1, \dots, n\} \setminus S_k$. Actually, every phylogenetic tree on $\{1, \dots, n\}$ is obtained in this way *twice*. The following easy lemma on the probability of $T\widehat{\ }T'$ under the Yule model is a direct consequence of the formula for the probability of a tree; see [20, Lemma 1].

Lemma 4. *Let* $\emptyset \neq S_k \subsetneq \{1, \dots, n\}$ *with* $|S_k| = k$, *and let* $T_k \in \mathcal{T}(S_k)$ *and* $T'_{n-k} \in \mathcal{T}(S_k^c)$. *Then*

$$P_Y(T_k\widehat{\ }T'_{n-k}) = \frac{2}{(n-1)\binom{n}{k}} P_Y(T_k) P_Y(T'_{n-k}).$$

Lemma. *Let* T, T' *be two phylogenetic trees on disjoint sets of taxa* S, S', *with* $|S| = k$ *and* $|S'| = n-k$. *Then:*
(1) $S^{(2)}(T\widehat{\ }T') = S^{(2)}(T) + S^{(2)}(T') + 2(S(T) + S(T')) + n.$
(2) $D^{(2)}(T\widehat{\ }T') = D^{(2)}(T) + D^{(2)}(T') + (n-k)(S^{(2)}(T) + 4S(T)) + k(S^{(2)}(T') + 4S(T')) + 2S(T)S(T') + 4k(n-k).$

Proof. Let us assume, without any loss of generality, that $S = \{1, \dots, k\}$ and $S' = \{k+1, \dots, n\}$. Then, these identities are a direct consequence of the equalities

$$\delta_{T\widehat{\ }T'}(i)^2 = \begin{cases} (\delta_T(i)+1)^2 & \text{if } 1 \leqslant i \leqslant k \\ (\delta_{T'}(i)+1)^2 & \text{if } k+1 \leqslant i \leqslant n \end{cases} \qquad d_{T\widehat{\ }T'}(i,j)^2 = \begin{cases} d_T(i,j)^2 & \text{if } 1 \leqslant i < j \leqslant k \\ d_{T'}(i,j)^2 & \text{if } k+1 \leqslant i < j \leqslant n \\ (\delta_T(i) + \delta_{T'}(j) + 2)^2 & \text{if } 1 \leqslant i \leqslant k < j \leqslant n. \end{cases} \quad \square$$

Now we can compute explicit formulas for $E_Y(S_n^{(2)})$ and $E_Y(D_n^{(2)})$.

Proposition. $E_Y(S_n^{(2)}) = 4n(H_n^2 - H_n^{(2)}) - 6n(H_n - 1)$.

Proof. We compute $E_Y(S_n^{(2)})$ using its definition:

$$\begin{aligned} E_Y(S_n^{(2)}) &= \sum_{T\in\mathcal{T}_n} S^{(2)}(T)\cdot p_Y(T) = \frac{1}{2}\sum_{k=1}^{n-1} \sum_{\substack{S_k\subsetneq\{1,\dots,n\}\\ |S_k|=k}} \sum_{T_k\in\mathcal{T}(S_k)} \sum_{T'_{n-k}\in\mathcal{T}(S_k^c)} S^{(2)}(T_k\widehat{\ }T'_{n-k})\cdot p_Y(T_k\widehat{\ }T'_{n-k}) \\ &= \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k} \sum_{T_k\in\mathcal{T}_k} \sum_{T'_{n-k}\in\mathcal{T}_{n-k}} \Big(S^{(2)}(T_k) + S^{(2)}(T'_{n-k}) + 2(S(T_k) + S(T'_{n-k})) + n\Big)\cdot \frac{2}{(n-1)\binom{n}{k}} P_Y(T_k)P_Y(T'_{n-k}) \\ &= \frac{1}{n-1}\sum_{k=1}^{n-1}\Bigg(\sum_{T_k} S^{(2)}(T_k)P_Y(T_k) + \sum_{T'_{n-k}} S^{(2)}(T'_{n-k})P_Y(T'_{n-k}) + 2\sum_{T_k} S(T_k)P_Y(T_k) \\ &\qquad + 2\sum_{T'_{n-k}} S(T'_{n-k})P_Y(T'_{n-k}) + n\Bigg) \\ &= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(E_Y(S_k^{(2)}) + E_Y(S_{n-k}^{(2)}) + 2E_Y(S_k) + 2E_Y(S_{n-k}) + n\Big) \\ &= \frac{2}{n-1}\sum_{k=1}^{n-1} E_Y(S_k^{(2)}) + \frac{4}{n-1}\sum_{k=1}^{n-1} E_Y(S_k) + n. \end{aligned}$$

And then

$$\begin{aligned} E_Y(S_n^{(2)}) &= \frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2} E_Y(S_k^{(2)}) + \frac{2}{n-1}E_Y(S_{n-1}^{(2)}) + \frac{n-2}{n-1}\cdot\frac{4}{n-2}\sum_{k=1}^{n-2} E_Y(S_k) \\ &\quad + \frac{4}{n-1}E_Y(S_{n-1}) + \frac{n-2}{n-1}\cdot(n-1) + 2 \\ &= \frac{n-2}{n-1}E_Y(S_{n-1}^{(2)}) + \frac{2}{n-1}E_Y(S_{n-1}^{(2)}) + \frac{4}{n-1}E_Y(S_{n-1}) + 2 = \frac{n}{n-1}E_Y(S_{n-1}^{(2)}) + 8H_{n-1} - 6. \end{aligned}$$

Setting $x_n = E_Y(S_n^{(2)})/n$, this recurrence becomes

$$x_n = x_{n-1} + \frac{8H_{n-1}}{n} - \frac{6}{n}.$$

Since $S^{(2)}$ applied to a single node is 0, $x_1 = E_Y(S_1^{(2)}) = 0$, and the solution of this recursive equation with this initial condition is

$$x_n = \sum_{k=2}^{n} \left(\frac{8H_{k-1}}{k} - \frac{6}{k} \right) = 8 \sum_{k=1}^{n-1} \frac{H_k}{k+1} - 6 \sum_{k=2}^{n} \frac{1}{k} = 4(H_n^2 - H_n^{(2)}) - 6(H_n - 1)$$

from where we deduce the identity in the statement. □

Proposition 5. *$E_Y(D_n^{(2)}) = 8n(n+1)(H_n^2 - H_n^{(2)}) - 2n(15n+7)H_n + 45n^2 - n$.*

Proof. If we compute $E_Y(D_n^{(2)})$ as we did with $E_Y(S_n^{(2)})$ in the last proposition, we obtain

$$\begin{aligned}
E_Y(D_n^{(2)}) &= \sum_{T \in \mathcal{T}_n} D^{(2)}(T) \cdot p_Y(T) = \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq \{1,\dots,n\} \\ |S_k| = k}} \sum_{T_k \in \mathcal{T}(S_k)} \sum_{T'_{n-k} \in \mathcal{T}(S_k^c)} D^{(2)}(T_k \widehat{\ } T'_{n-k}) \cdot p_Y(T_k \widehat{\ } T'_{n-k}) \\
&= \frac{1}{n-1} \sum_{k=1}^{n-1} \Bigg[\sum_{T_k} D^{(2)}(T_k) P_Y(T_k) + \sum_{T'_{n-k}} D^{(2)}(T'_{n-k}) P_Y(T'_{n-k}) \\
&\quad + 2 \left(\sum_{T_k} S(T_k) P_Y(T_k) \right) \left(\sum_{T'_{n-k}} S(T'_{n-k}) P_Y(T'_{n-k}) \right) + (n-k) \sum_{T_k} S^{(2)}(T_k) P_Y(T_k) \\
&\quad + 4(n-k) \sum_{T_k} S(T_k) P_Y(T_k) + k \sum_{T'_{n-k}} S^{(2)}(T'_{n-k}) P_Y(T'_{n-k}) + 4k \sum_{T'_{n-k}} S(T'_{n-k}) P_Y(T'_{n-k}) + 4k(n-k) \Bigg] \\
&= \frac{2}{n-1} \sum_{k=1}^{n-1} \left(E_Y(D_k^{(2)}) + E_Y(S_k) E_Y(S_{n-k}) + (n-k) E_Y(S_k^{(2)}) + 4(n-k) E_Y(S_k) \right) + \frac{2}{3} n(n+1)
\end{aligned}$$

and therefore

$$\begin{aligned}
E_Y(D_n^{(2)}) &= \frac{n-2}{n-1} \cdot \frac{2}{n-2} \sum_{k=1}^{n-2} E_Y(D_k^{(2)}) + \frac{2}{n-1} E_Y(D_{n-1}^{(2)}) \\
&\quad + \frac{n-2}{n-1} \cdot \frac{2}{n-2} \sum_{k=1}^{n-2} E_Y(S_k) E_Y(S_{n-1-k}) + \frac{2}{n-1} \left(\sum_{k=1}^{n-1} E_Y(S_k) E_Y(S_{n-k}) - \sum_{k=1}^{n-2} E_Y(S_k) E_Y(S_{n-1-k}) \right) \\
&\quad + \frac{n-2}{n-1} \cdot \frac{2}{n-2} \sum_{k=1}^{n-2} (n-1-k) E_Y(S_k^{(2)}) + \frac{2}{n-1} \left(\sum_{k=1}^{n-1} (n-k) E_Y(S_k^{(2)}) - \sum_{k=1}^{n-2} (n-1-k) E_Y(S_k^{(2)}) \right) \\
&\quad + \frac{n-2}{n-1} \cdot \frac{8}{n-2} \sum_{k=1}^{n-2} (n-1-k) E_Y(S_k) + \frac{8}{n-1} \left(\sum_{k=1}^{n-1} (n-k) E_Y(S_k) - \sum_{k=1}^{n-2} (n-1-k) E_Y(S_k) \right) \\
&\quad + \frac{n-2}{n-1} \cdot \frac{2}{3} n(n-1) + \frac{2}{3} n(n+1) - \frac{n-2}{n-1} \cdot \frac{2}{3} n(n-1) \\
&= \frac{n-2}{n-1} E_Y(D_{n-1}^{(2)}) + \frac{2}{n-1} E_Y(D_{n-1}^{(2)}) + \frac{2}{n-1} \sum_{k=1}^{n-2} E_Y(S_k)(E_Y(S_{n-k}) - E_Y(S_{n-k-1})) \\
&\quad + \frac{2}{n-1} \sum_{k=1}^{n-1} E_Y(S_k^{(2)}) + \frac{8}{n-1} \sum_{k=1}^{n-1} E_Y(S_k) + 2n \\
&= \frac{n}{n-1} E_Y(D_{n-1}^{(2)}) + 8n(H_n^2 - H_n^{(2)}) - 14nH_{n-1} + 15n - 14.
\end{aligned}$$

Setting $x_n = E_Y(D_n^{(2)})/n$, this recurrence becomes

$$x_n = x_{n-1} + 8(H_n^2 - H_n^{(2)}) - 14H_{n-1} + 15 - \frac{14}{n}.$$

The solution of this recurrence with $x_1 = E_Y(D_1^{(2)}) = 0$ is

$$\begin{aligned} x_n &= \sum_{k=2}^{n} \left(8(H_k^2 - H_k^{(2)}) - 14H_{k-1} + 15 - \frac{14}{k} \right) \\ &= 8(n+1)(H_n^2 - H_n^{(2)}) - 2(15n+7)H_n + 45n - 1 \end{aligned}$$

from which we deduce the formula in the statement. □

References

[1] J. Felsenstein, Inferring Phylogenies, Sinauer Associates Inc., 2004.
[2] J.S. Farris, A successive approximations approach to character weighting, Syst. Zool. 18 (1969) 374–385.
[3] J.S. Farris, On comparing the shapes of taxonomic trees, Syst. Zool. 22 (1973) 50–54.
[4] J.B. Phipps, Dendrogram topology, Syst. Zool. 20 (1971) 306–308.
[5] W.T. Williams, H.T. Clifford, On the comparison of two classifications of the same set of elements, Taxon 20 (4) (1971) 519–522.
[6] M.A. Steel, D. Penny, Distributions of tree comparison metrics—some new results, Syst. Biol. 42 (2) (1993) 126–141.
[7] A. Mir, F. Rosselló, The mean value of the squared path-difference distance for rooted phylogenetic trees, J. Math. Anal. Appl. 371 (2010) 168–176.
[8] E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. in Appl. Probab. 3 (1971) 44–77.
[9] G.U. Yule, A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, Philos. Trans. R. Soc. Lond. Ser. B 213 (1924) 21–87.
[10] A. Mooers, S.B. Heard, Inferring evolutionary process from phylogenetic tree shape, Q. Rev. Biol. 72 (1997) 31–54.
[11] G. Cardona, A. Mir, F. Rosselló, The expected value under the Yule model of the squared path-difference distance, arXiv:1203.2503v1 [q-bio.PE].
[12] J. Brown, Probabilities of evolutionary trees, Syst. Biol. 43 (1994) 78–91.
[13] M. Steel, A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, Math. Biosci. 170 (2001) 91–112.
[14] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees. arXiv:1202.1223v1 [q-bio.PE], 2012 (submitted for publication).
[15] D. Knuth, The Art of Computer Programming, Vol. 1: Fundamental Algorithms, third ed., Addison-Wesley, 1997.
[16] C. Wei, D. Gong, Q. Wang, Chu–Vandermonde convolution and harmonic number identities, arXiv:1201.0420v1 [math.CO], 2012.
[17] Y. Chen, Q. Hou, H. Jin, The Abel–Zeilberger algorithm, Electron. J. Combin. 18 (2011) #P17.
[18] M.J. Sackin, "Good" and "bad" phenograms, Syst. Zool. 21 (1972) 225–226.
[19] M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution 47 (1993) 1171–1181.
[20] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variance of several balance indices under the Yule model. arXiv:1202.6573v1 [q-bio.PE] (submitted for publication).