J. Math. Anal. Appl. 371 (2010) 168–176

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

The mean value of the squared path-difference distance for rooted phylogenetic trees

Arnau Mir, Francesc Rosselló *

Research Institute of Health Science (IUNICS) and Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca, Spain

ARTICLE INFO

Article history:
Received 13 June 2009
Available online 6 May 2010
Submitted by J.J. Nieto

Keywords:
Phylogenetic trees
Path-difference metric
Nodal distance
Hypergeometric series

ABSTRACT

The path-difference metric is one of the oldest distances for the comparison of fully resolved phylogenetic trees, but its statistical properties are still quite unknown. In this paper we compute the mean value of the square of the path-difference metric between two fully resolved rooted phylogenetic trees with n leaves, under the uniform distribution. This complements previous work by Steel and Penny, who computed this mean value for fully resolved unrooted phylogenetic trees.

1. Introduction

The definition and study of metrics for the comparison of rooted phylogenetic trees is a classical problem in phylogenetics [11, Ch. 30], motivated, among other applications, by the need to compare alternative phylogenetic trees for a given set of organisms obtained from different datasets or using different reconstruction algorithms [14], and by the desire to assess phylogenetic tree reconstruction methods [29]. Many metrics for the comparison of rooted phylogenetic trees on the same set of taxa have been proposed so far, including the Robinson–Foulds, or partition, metric [22,23], the nearest-neighbor interchange metric [27], the subtree transfer distance [3], the geodesic distance [16], the triples metric [8] and the transposition distance [2]. For other applications of metrics in computational biology, see, for instance, [11,12,15,17,18,28,30].

Some of the first metrics for the comparison of rooted phylogenetic trees, defined around 40 years ago, were based on the comparison of the vectors of lengths of (undirected) paths connecting pairs of taxa in the corresponding trees. These metrics comprise, for instance, the Euclidean distance between these vectors [9,10], the Manhattan distance between them [29], or the correlation between them [20]. Similar metrics have also been defined later for unrooted phylogenetic trees [5,21,26]. Let us point out here that, in the rooted case, these path-lengths based metrics were defined only for *binary* phylogenetic trees, as they do not satisfy the separation axiom of metrics (distance 0 means isomorphism) for rooted phylogenetic trees with nodes or arbitrary out-degree. Recently, a metric for rooted phylogenetic trees based on the comparison of path lengths between leaves that overcomes this drawback, and hence that can be used in a safe and sound way on arbitrary rooted trees, has been introduced in the literature [7]. On the other hand, a well-known theorem by Smolenskii [24] guarantees that, in the unrooted case, the metrics defined through the comparison of path-lengths vectors always satisfy the desired separation axiom.

In contrast with other metrics [6,13,25,26], and despite their tradition and popularity, the statistical properties of these path-lengths based metrics are mostly unknown. For instance, the diameter of none of these metrics (either in the rooted

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

0022-247X/$ – see front matter
doi:10.1016/j.jmaa.2010.05.001

or in the unrooted case) is known yet. Steel and Penny [26] studied the distribution of one of these distances for unrooted trees: the one defined through the Euclidean distance between path-lengths vectors, which these authors called the *path-difference metric* (other published names for this metric are the *cladistic difference* [9] and, generically, a *nodal distance* [5,21]). In particular, in the aforementioned paper, Steel and Penny computed the mean value of the square of this path-difference metric for fully resolved unrooted trees.

In this paper we compute the mean value of the square of the path-difference metric for fully resolved rooted phylogenetic trees with n leaves. Although the raw argument underlying our computation is the same as in Steel and Penny's paper, the details in the rooted case are much harder than in the unrooted case, because of the asymmetric role of the root. We have proved that this mean value grows in $O(n^3)$; more specifically, it is

$$2\binom{n}{2}\left(4(n-1)+2-\frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}}-\left(\frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}}\right)^2\right).$$

This turns out to be the mean value obtained by Steel and Penny for unrooted phylogenetic trees, but with $n+1$ leaves. A similar relationship between combinatorial values for rooted and unrooted phylogenetic trees arises in other problems; for instance, a simple argument shows that the number of rooted phylogenetic trees with n leaves is the number of unrooted phylogenetic trees with $n+1$ leaves [11, Ch. 3]; also, as we shall see in this paper (Corollary 11), the mean value of the length of the undirected path between two given leaves in a rooted phylogenetic tree with n leaves is equal to the corresponding mean value for unrooted phylogenetic trees. But we have not been able to find a clever argument that proves directly this relationship between the mean values of the squared path-difference metric, or of the path-length between two leaves, in the rooted and unrooted cases, and thus we have needed to compute them.

The knowledge of the mean value for a metric is useful in the assessment of a comparison through this metric, because it "provides an indication as to whether or not the measured similarity could have come about by chance" [26]. In the specific case of path-lengths based metrics for (rooted and unrooted) phylogenetic trees, there exist software packages (see, for instance, [21]) that estimate this mean value in a given experiment by computing the mean of the distances between many pairs of randomly generated phylogenetic trees. The explicit computation of this mean value, as a function of the number of leaves, makes this simulation unnecessary.

2. Preliminaries

2.1. Phylogenetic trees

In this paper, by a *phylogenetic tree* on a set S of taxa we mean a *fully resolved*, or *binary* (that is, with all its internal nodes of out-degree 2), rooted tree with its leaves bijectively labeled in the set S. 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 a given set of n taxa, when this set is known or nonrelevant.

We shall represent a path from u to v in a phylogenetic tree T by $u \rightsquigarrow v$. Whenever there exists a path $u \rightsquigarrow v$, we shall say that v is a *descendant* of u and also that u is an *ancestor* of v. Given a node v of a phylogenetic tree T, the *subtree of T rooted at v* is the subgraph of T induced on the set of descendants of v. It is a phylogenetic tree on the set of descendant leaves of v, and with root this node v.

The *lowest common ancestor* (LCA) of a pair of nodes u, v of a phylogenetic tree T, in symbols $LCA_T(u, v)$, is the unique common ancestor of them that is a descendant of every other common ancestor of them. The *path difference* $d_T(u, v)$ between two nodes u and v is the sum of the lengths of the paths $LCA_T(u, v) \rightsquigarrow u$ and $LCA_T(u, v) \rightsquigarrow v$; equivalently, it is the length of the only path connecting u and v in the undirected tree associated to T. It is well known (for a proof, see [7]) that the vector of path differences $d(T) = (d_T(i, j))_{1 \leqslant i < j \leqslant n}$ between all pairs of leaves characterizes up to isomorphism a phylogenetic tree with n leaves: this property is false if we remove the binarity assumption on the trees.

Let $\mathcal{T}_n$ be the set of (isomorphism classes of) phylogenetic trees with n leaves. It is well known [11, Ch. 3] that $|\mathcal{T}_1| = 1$ and, for every $n \geqslant 2$,

$$|\mathcal{T}_n| = (2n-3)!! = (2n-3)(2n-5)\cdots 3 \cdot 1.$$

An *ordered m-forest* on a set S is an ordered sequence of m phylogenetic trees $(T_1, T_2, \ldots, T_m)$, each T_i on a set S_i of taxa, such that these sets S_i are pairwise disjoint and their union is S. Let $\mathcal{F}_{m,n}$ be the set of (isomorphism classes of) ordered m-forests on any given set S with $|S| = n$. The cardinal of $\mathcal{F}_{m,n}$ is computed (although not explicitly) along the proof of Theorem 3 in [26].

Lemma 1. *For every $m \geqslant 1$, $|\mathcal{F}_{m,m}| = m!$ and*

$$|\mathcal{F}_{m,n}| = \frac{m(n!)\prod_{l=1}^{n-m-1}(n+l)}{(2(n-m))!!} = \frac{(2n-m-1)!m}{(n-m)!2^{n-m}} \quad \text{for every } n > m.$$

Proof. The exponential generating function for the number of rooted phylogenetic trees with n leaves is $B(x) = 1 - \sqrt{1-2x}$. Then, the exponential generating function for the number of ordered forests consisting of a given number of trees (marked by the variable y) and a given global number of leaves (marked by the variable y) is

$$F(x, y) = \sum_{m \geqslant 1} y^m B(x)^m = \frac{1}{1 - yB(x)} - 1.$$

This implies that the number $|\mathcal{F}_{m,n}|$ of ordered m-forests on a set of n leaves is equal to $\frac{\partial^n}{\partial x^n}(B(x)^m)|_{x=0}$. This derivative can be easily computed, yielding the values given in the statement. □

2.2. Hypergeometric functions

The (*generalized*) *hypergeometric function* ${}_pF_q$ is defined [4] as

$${}_pF_q\left(\begin{matrix} a_1, & \ldots, & a_p \\ b_1, & \ldots, & b_q \end{matrix}; z\right) = \sum_{k \geqslant 0} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k} \cdot \frac{z^k}{k!},$$

where $(a)_k := a \cdot (a+1) \cdots (a+k-1)$.

The following lemmas will be used in the next section.

Lemma 2.

$${}_2F_1\left(\begin{matrix} n-1, & 2-n \\ & -n \end{matrix}; \frac{1}{2}\right) = \frac{2^{n-1}}{n}.$$

Proof. To compute the value of ${}_2F_1\left(\begin{smallmatrix} n-1, & 2-n \\ & -n \end{smallmatrix}; \frac{1}{2}\right)$ we shall use Formula 15.1.26 in [1] (see also http://functions.wolfram.com/07.23.03.0028.01):

$${}_2F_1\left(\begin{matrix} a, & 1-a \\ & c \end{matrix}; \frac{1}{2}\right) = \frac{2^{1-c}\sqrt{\pi}\,\Gamma(c)}{\Gamma(\frac{a+c}{2})\Gamma(\frac{c-a+1}{2})}.$$

We cannot apply this expression to $a = n-1$ and $c = -n$, because $\Gamma(-n) = \infty$. So, instead, we use a standard pass to limit argument:

$${}_2F_1\left(\begin{matrix} n-1, & 2-n \\ & -n \end{matrix}; \frac{1}{2}\right) = \lim_{\varepsilon \to 0} {}_2F_1\left(\begin{matrix} n-1, & 2-n \\ & -n+\varepsilon \end{matrix}; \frac{1}{2}\right) = \lim_{\varepsilon \to 0} \frac{2^{1+n-\varepsilon}\sqrt{\pi}\,\Gamma(-n+\varepsilon)}{\Gamma(\frac{\varepsilon-1}{2})\Gamma(\frac{2-2n+\varepsilon}{2})} = \frac{2^{n-1}}{n}. \qquad \square$$

Lemma 3.

$${}_3F_2\left(\begin{matrix} 1-n, & 2-n, & n-1 \\ -n, & -n & \end{matrix}; \frac{1}{2}\right) = \frac{2^{n-1}}{n^2}\left(-1 + \frac{(2n-1)!!}{2^{n-2}(n-1)!}\right).$$

Proof. The hypergeometric series ${}_3F_2\left(\begin{smallmatrix} 1-n, & 2-n, & n-1 \\ -n, & -n & \end{smallmatrix}; \frac{1}{2}\right)$ can be written as a function of the hypergeometric function ${}_2F_1$ as follows[1]:

$${}_3F_2\left(\begin{matrix} 1-n, & 2-n, & n-1 \\ -n, & -n & \end{matrix}; \frac{1}{2}\right) = {}_2F_1\left(\begin{matrix} n-1, & 2-n \\ & -n \end{matrix}; \frac{1}{2}\right) - \frac{(n-1)(n-2)}{2n^2}\,{}_2F_1\left(\begin{matrix} n, & 3-n \\ & 1-n \end{matrix}; \frac{1}{2}\right). \tag{1}$$

We already know from the previous lemma that ${}_2F_1\left(\begin{smallmatrix} n-1, & 2-n \\ & -n \end{smallmatrix}; \frac{1}{2}\right) = \frac{2^{n-1}}{n}$. It remains to compute ${}_2F_1\left(\begin{smallmatrix} n, & 3-n \\ & 1-n \end{smallmatrix}; \frac{1}{2}\right)$. To do it, we shall use the following formula[2]:

$${}_2F_1\left(\begin{matrix} a, & 3-a \\ & c \end{matrix}; \frac{1}{2}\right) = \frac{2^{3-c}\sqrt{\pi}\,\Gamma(c)}{(a-1)(a-2)}\left(\frac{c-2}{\Gamma(\frac{a+c}{2}-1)\Gamma(\frac{c-a+1}{2})} - \frac{2}{\Gamma(\frac{a+c-3}{2})\Gamma(\frac{c-a}{2})}\right).$$

[1] See http://functions.wolfram.com/07.27.03.0118.01.

[2] See http://functions.wolfram.com/07.23.03.0030.01.

Again, we cannot apply this formula to $a=n-1$ and $c=-n$, and thus we use a pass to limit argument:

$$\begin{aligned}
{}_2F_1\left(\begin{matrix} n, & 3-n \\ 1-n & \end{matrix}; \frac{1}{2}\right) &= \lim_{\varepsilon\to 0} {}_2F_1\left(\begin{matrix} n, & 3-n \\ 1-n+\varepsilon & \end{matrix}; \frac{1}{2}\right)\\
&= \lim_{\varepsilon\to 0} \frac{2^{2+n-\varepsilon}\sqrt{\pi}\,\Gamma(1-n+\varepsilon)}{(n-1)(n-2)}\left(\frac{(-n-1-\varepsilon)}{\Gamma(\frac{\varepsilon-1}{2})\Gamma(1-n+\frac{\varepsilon}{2})} - \frac{2}{\Gamma(\frac{\varepsilon-2}{2})\Gamma(\frac{1+\varepsilon-2n}{2})}\right)\\
&= \frac{2^{2+n}\sqrt{\pi}}{(n-1)(n-2)}\left(\lim_{\varepsilon\to 0}\frac{(-n-1-\varepsilon)\Gamma(1-n+\varepsilon)}{\Gamma(\frac{\varepsilon-1}{2})\Gamma(1-n+\frac{\varepsilon}{2})} - \lim_{\varepsilon\to 0}\frac{2\Gamma(1-n+\varepsilon)}{\Gamma(\frac{\varepsilon-2}{2})\Gamma(\frac{1+\varepsilon-2n}{2})}\right)\\
&= \frac{2^{2+n}\sqrt{\pi}}{(n-1)(n-2)}\left(\frac{(n+1)}{4\sqrt{\pi}} - \frac{(-1)^{n+2}\binom{-1/2}{n}n!}{(n-1)!\sqrt{\pi}}\right)\\
&= \frac{2^{2+n}\sqrt{\pi}}{(n-1)(n-2)}\left(\frac{(n+1)}{4} - \frac{(2n-1)!!}{(n-1)!2^n}\right).
\end{aligned}$$

Replacing ${}_2F_1\left(\begin{smallmatrix} n-1, & 2-n \\ -n & \end{smallmatrix}; \frac{1}{2}\right)$ and ${}_2F_1\left(\begin{smallmatrix} n, & 3-n \\ 1-n & \end{smallmatrix}; \frac{1}{2}\right)$ in Eq. (1) by their values given above, we obtain

$${}_3F_2\left(\begin{matrix} 1-n, & 2-n, & n-1 \\ -n, & -n & \end{matrix}; \frac{1}{2}\right) = \frac{2^{n-1}}{n} - \frac{2^{n+2}}{2n^2}\left(\frac{(n+1)}{4} - \frac{(2n-1)!!}{(n-1)!2^n}\right) = \frac{2^{n-1}}{n^2}\left(-1 + \frac{(2n-1)!!}{2^{n-2}(n-1)!}\right),$$

as we claimed. □

Lemma 4. *For every real numbers a, b,*

$${}_4F_3\left(\begin{matrix} 1, & a, & a+1/2, & b \\ 2, & 2a, & b+1/2 & \end{matrix}; 1\right) = \frac{(2b-1)}{(a-1)(b-1)}\left(-1 + {}_3F_2\left(\begin{matrix} a-1, & a-1/2, & b-1 \\ 2a-1, & b-1/2 & \end{matrix}; 1\right)\right).$$

Proof. By definition,

$${}_4F_3\left(\begin{matrix} 1, & a, & a+1/2, & b \\ 2, & 2a, & b+1/2 & \end{matrix}; 1\right) = \sum_{k\geqslant 0}\frac{k!(a)_k(a+1/2)_k(b)_k}{(k+1)!(2a)_k(b+1/2)_k}\cdot\frac{1}{k!} = \sum_{k\geqslant 1}\frac{(a)_{k-1}(a+1/2)_{k-1}(b)_{k-1}}{k!(2a)_{k-1}(b+1/2)_{k-1}} = (*).$$

Taking into account that

$$\begin{aligned}
(a)_{k-1} &= \frac{(a-1)_k}{a-1}, \qquad (a+1/2)_{k-1} = \frac{(a-1/2)_k}{a-1/2}, \qquad (b)_{k-1} = \frac{(b-1)_k}{b-1},\\
(2a)_{k-1} &= \frac{(2a-1)_k}{2a-1}, \qquad (b+1/2)_{k-1} = \frac{(b-1/2)_k}{b-1/2},
\end{aligned}$$

the expression $(*)$ can be written as

$$\begin{aligned}
(*) &= \sum_{k\geqslant 1}\frac{(a-1)_k(a-1/2)_k(b-1)_k(2a-1)(b-1/2)}{(a-1)(a-1/2)(b-1)(2a-1)_k(b-1/2)_k}\cdot\frac{1}{k!}\\
&= \frac{(2b-1)}{(a-1)(b-1)}\left(-1 + {}_3F_2\left(\begin{matrix} a-1, & a-1/2, & b-1 \\ 2a-1, & b-1/2 & \end{matrix}; 1\right)\right)
\end{aligned}$$

yielding the formula in the statement. □

3. Mean total areas

For every $s\in\mathbb{Z}^+$, the *total s-area* of a phylogenetic tree T is

$$D^{(s)}(T) = \sum_{1\leqslant i<j\leqslant n} d_T(i,j)^s.$$

This value (or, rather, its s-th root) measures the total amount of evolutive history captured by the phylogenetic tree. Let

$$\mu\big(D^{(s)}\big)_n = \frac{\sum_{T\in\mathcal{T}_n} D^{(s)}(T)}{|\mathcal{T}_n|}$$

be the mean value of $D^{(s)}(T)$ for $T \in \mathcal{T}_n$ under the uniform distribution on $\mathcal{T}_n$. In this section we compute $\mu(D^{(1)})_n$ and $\mu(D^{(2)})_n$. To simplify the notations, for every $s \in \mathbb{Z}^+$ let

$$S_n^{(s)} = \sum_{T \in \mathcal{T}_n} d_T(1,2)^s.$$

Lemma 5. *For every $s \in \mathbb{Z}^+$ and for every $1 \leqslant i < j \leqslant n$,*

$$\sum_{T \in \mathcal{T}_n} d_T(i,j)^s = S_n^{(s)}.$$

Proof. Let $\sigma_{i,j}$ be the involutive permutation that interchanges 1 and i, and 2 and j and leaves the other elements fixed and, for every $T \in \mathcal{T}_n$, let $T_{\sigma_{i,j}}$ be the phylogenetic tree obtained by applying to the leaves in T the permutation $\sigma_{i,j}$. On the one hand, it is clear that $d_T(i,j) = d_{T_{\sigma_{i,j}}}(1,2)$, and, on the other hand, since the mapping $\mathcal{T}_n \to \mathcal{T}_n$ defined by $T \mapsto T_{\sigma_{i,j}}$ is bijective, we have the equality of multisets

$$\{d_T(1,2) \mid T \in \mathcal{T}_n\} = \{d_{T_{\sigma_{i,j}}}(1,2) \mid T \in \mathcal{T}_n\}.$$

Combining these two observations we obtain

$$\sum_{T \in \mathcal{T}_n} d_T(i,j)^s = \sum_{T \in \mathcal{T}_n} d_{T_{\sigma_{i,j}}}(1,2)^s = \sum_{T \in \mathcal{T}_n} d_T(1,2)^s. \qquad \square$$

Corollary 6. *For every $n \geqslant 2$ and for every $s \in \mathbb{Z}^+$,*

$$\mu\big(D^{(s)}\big)_n = \binom{n}{2} \frac{S_n^{(s)}}{(2n-3)!!}.$$

Proof. Using the previous lemma,

$$\mu\big(D^{(s)}\big)_n = \frac{\sum_{T \in \mathcal{T}_n} D^{(s)}(T)}{|\mathcal{T}_n|} = \frac{\sum_{1 \leqslant i < j \leqslant n} \sum_{T \in \mathcal{T}_n} d_T(i,j)^s}{(2n-3)!!} = \frac{\binom{n}{2} \sum_{T \in \mathcal{T}_n} d_T(1,2)^s}{(2n-3)!!}. \qquad \square$$

For every $i = 1, \ldots, n-1$, let c_i be the cardinal of the set

$$\{T \in \mathcal{T}_n \mid d_T(1,2) = i\}.$$

Then, $S_n^{(s)} = \sum_{i=2}^n i^s c_i$. Our first goal is to find a suitable expression for these coefficients c_i.

Proposition 7. $c_i = \frac{(i-1)(2n-i-2)!}{(2(n-i))!!}$.

Proof. Let $T \in \mathcal{T}_n$ be any tree such that $d_T(1,2) = i$; to simplify the notations, let us denote by x the node $LCA_T(1,2)$. Then, on the one hand, the paths $x \rightsquigarrow 1$ and $x \rightsquigarrow 2$ have, respectively, j and $i-2-j$ intermediate nodes, for some $j = 0, \ldots, i-2$, and each such intermediate node is the parent of the root of a rooted subtree of T. Let $\{i_1, \ldots, i_k\}$ be the union of the (pairwise disjoint) sets of leaves of these subtrees: notice that $i-2 \leqslant k \leqslant n-2$, because each subtree has some leaf and the leaves 1, 2 cannot belong to these subtrees. On the other hand, x is the leaf of the phylogenetic tree T_0 with leaves $(\{1, \ldots, n\} \setminus \{1, 2, i_1, \ldots, i_k\}) \cup \{x\}$ obtained by collapsing the subtree of T rooted at x into a single leaf x.

So, the tree T is determined by a subset $\{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$, with $i-2 \leqslant k \leqslant n-2$, a phylogenetic tree T_0 on $(\{1, \ldots, n\} \setminus \{1, 2, i_1, \ldots, i_k\}) \cup \{x\}$ (and hence with $n-k-1$ leaves), an ordered $(i-2)$-forest $(T_1, \ldots, T_{i-2})$ on $\{i_1, \ldots, i_k\}$, and an index $j \in \{0, 1, \ldots, i-2\}$. The tree T is obtained by starting in the leaf x of T_0 two new paths $(x, v_j, \ldots, v_1, 1)$ and $(x, v_{j+1}, \ldots, v_{i-2}, 2)$ of lengths $j+1$ and $i-j-1$, respectively, and then adding to each intermediate node v_l in these paths an arc with head the root of the tree T_l (cf. Fig. 1).

This shows that c_i can be computed as

$$\begin{aligned} c_i = \sum_{k=i-2}^{n-2} & (\text{number of ways of choosing the } k \text{ nodes } i_1, \ldots, i_k) \\ & \cdot \big(\text{number of ordered } (i-2)\text{-forests trees on } \{i_1, \ldots, i_k\}\big) \\ & \cdot (\text{number of ways of choosing } j \text{ between 0 and } i-2) \\ & \cdot (\text{number of phylogenetic trees with } n-k-1 \text{ leaves}) \end{aligned}$$

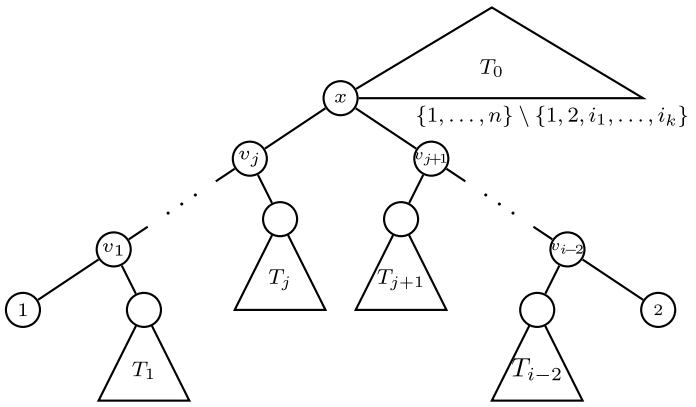

Fig. 1. The structure of a tree T with $d_T(1,2)=i$.

$$= \sum_{k=i-2}^{n-2} \binom{n-2}{k} \cdot |\mathcal{F}_{i-2,k}| \cdot (i-1) \cdot (2n-2k-5)!!$$

$$= (i-1)!\binom{n-2}{i-2}(2n-2i-1)!!$$

$$+ (i-1)(i-2) \sum_{k=0}^{n-i-1} \binom{n-2}{k+i-1} \frac{(k+i-1)!\prod_{l=1}^{k}(k+i-1+l)}{(2(k+1))!!}(2n-2k-2i-3)!!.$$

Applying the *hypergeometric series lookup algorithm* given in [19, p. 36], we obtain

$$\sum_{k=0}^{n-i-1} \binom{n-2}{k+i-1} \frac{(k+i-1)!\prod_{l=1}^{k}(k+i-1+l)}{(2(k+1))!!}(2n-2k-2i-3)!!$$

$$= \frac{1}{2}\binom{n-2}{i-1}(i-1)!(2n-2i-3)!! \cdot {}_4F_3\left(\begin{matrix}1, & i/2, & i/2+1/2, & i-n+1\\ 2, & i, & i-n+3/2 & \end{matrix}; 1\right),$$

and hence

$$c_i = (i-1)!\binom{n-2}{i-2}(2n-2i-3)!!$$

$$\cdot \left(2n-2i-1+\frac{1}{2}(i-2)(n-i){}_4F_3\left(\begin{matrix}1, & i/2, & i/2+1/2, & i-n+1\\ 2, & i, & i-n+3/2 & \end{matrix}; 1\right)\right).$$

If we apply Lemma 4 with $a=i/2$ and $b=i-n+1$, we obtain

$$c_i = (i-1)!\binom{n-2}{i-2}(2n-2i-1)!! \cdot {}_3F_2\left(\begin{matrix}i/2-1, & i/2-1/2, & i-n\\ i-1, & i-n+1/2 & \end{matrix}; 1\right). \tag{2}$$

The value of ${}_3F_2\left(\begin{matrix}i/2-1,\ i/2-1/2,\ i-n\\ i-1,\ \ i-n+1/2\end{matrix}; 1\right)$ is computed in [4, Form. (2), p. 9]:

$${}_3F_2\left(\begin{matrix}i/2-1, & i/2-1/2, & i-n\\ i-1, & i-n+1/2 & \end{matrix}; 1\right) = \frac{(i-2)!\Gamma(1/2+i-n)\Gamma(n-i/2)\Gamma(3/2-i/2)}{\Gamma(i/2)(n-2)!\Gamma(3/2+i/2-n)\Gamma(1/2)}.$$

Now, using that

$$\Gamma(x) = (x-1)\Gamma(x-1),$$

we can write

$$\Gamma(3/2-i/2) = \frac{(-1)^{n-i}\prod_{k=1}^{n-i}(2k+i-3)}{2^{n-i}}\Gamma(3/2+i/2-n),$$

$$\Gamma(1/2) = \frac{(-1)^{n-i}\prod_{k=1}^{n-i}(2k-1)}{2^{n-i}}\Gamma(1/2+i-n),$$

$$\Gamma(n-i/2) = \frac{\prod_{k=1}^{n-i}(2n-i-2k)}{2^{n-i}}\Gamma(i/2).$$

Then, using these formulas, the expression for ${}_3F_2\left(\begin{smallmatrix} i/2-1, & i/2-1/2, & i-n \\ i-1, & i-n+1/2 & \end{smallmatrix}; 1\right)$ can be simplified, yielding

$$ {}_3F_2\left(\begin{matrix} i/2-1, & i/2-1/2, & i-n \\ i-1, & i-n+1/2 & \end{matrix}; 1\right) = \frac{(i-2)!(2n-i-2)!}{(n-2)!(2n-2i-1)!!2^{n-i}}. $$

Replacing this expression in Eq. (2), we finally obtain

$$ c_i = \frac{(i-1)(2n-i-2)!}{(2(n-i))!!}, $$

as we claimed. □

Proposition 8. $S_n^{(1)} = 2^{n-1}(n-1)! = (2n-2)!!$.

Proof. By Proposition 7 we know that

$$ S_n^{(1)} = \sum_{i=2}^{n} \frac{i(i-1)(2n-i-2)!}{(2(n-i))!!} = \sum_{i=0}^{n-2} \frac{(i+2)(i+1)(2n-i-4)!}{(2(n-i-2))!!}. $$

If we compute this sum using again the algorithm given in [19, p. 36], we obtain

$$ S_n^{(1)} = \frac{2^{3-n}(2n-4)!}{(n-2)!}\, {}_2F_1\left(\begin{matrix} 3, & 2-n \\ 4-2n & \end{matrix}; 2\right). $$

Replacing in this equality ${}_2F_1\left(\begin{smallmatrix} 3, & 2-n \\ 4-2n & \end{smallmatrix}; 2\right)$ by its definition, we obtain

$$ S_n^{(1)} = 2^{2-n} \sum_{k=0}^{n-2} \frac{(k+1)(k+2)(2n-k-4)!}{(n-k-2)!} 2^k = \sum_{k=0}^{n-2} \frac{(n-k-1)(n-k)(n+k-2)!}{k!} 2^{-k}. $$

This sum can be computed using again the algorithm given in [19, p. 36], yielding

$$ S_n^{(1)} = n!\, {}_2F_1\left(\begin{matrix} n-1, & 2-n \\ -n & \end{matrix}; 1/2\right). $$

By Lemma 2, we conclude that

$$ S_n^{(1)} = \frac{n!2^{n-1}}{n} = 2^{n-1}(n-1)! = (2n-2)!!, $$

as we claimed. □

Proposition 9. $S_n^{(2)} = 2\cdot(2n-1)!! - (2n-2)!!$.

Proof. By Proposition 7, we have

$$ S_n^{(2)} = \sum_{i=2}^{n} i^2 c_i = \sum_{i=2}^{n} \frac{i^2(i-1)(2n-i-2)!}{(2(n-i))!!} = \sum_{i=0}^{n-2} \frac{(i+2)^2(i+1)(2n-i-4)!}{(2(n-i-2))!!}. $$

Using again the algorithm given in [19, p. 36], the value of the sum $S_n^{(2)}$ is:

$$ S_n^{(2)} = \frac{2^{4-n}(2n-4)!}{(n-2)!}\, {}_3F_2\left(\begin{matrix} 3, & 3, & 2-n \\ 2, & 4-2n & \end{matrix}; 2\right) = \sum_{k=0}^{n-2} \frac{(k+2)^2(k+1)(2n-k-4)!}{(n-k-2)!} 2^{k-n+2} $$

$$ = \sum_{k=0}^{n-2} \frac{(n-k)^2(n-k-1)(n+k-2)!}{k!} 2^{-k}. $$

Using once again the algorithm given in [19, p. 36], the last sum can be computed as:

$$ S_n^{(2)} = n\cdot n!\, {}_3F_2\left(\begin{matrix} 1-n, & 2-n, & n-1 \\ -n, & -n & \end{matrix}; \frac{1}{2}\right). $$

By Lemma 3, we obtain

$$S_n^{(2)} = (n-1)!2^{n-1}\left(-1+\frac{(2n-1)!!}{2^{n-2}(n-1)!}\right) = 2\cdot(2n-1)!! - (2n-2)!!,$$

as we claimed. □

Applying Corollary 6, we obtain the following total areas.

Corollary 10.

$$\mu\big(D^{(1)}\big)_n = \binom{n}{2}\cdot\frac{(2n-2)!!}{(2n-3)!!}, \qquad \mu\big(D^{(2)}\big)_n = \binom{n}{2}\frac{2(2n-1)!! - (2n-2)!!}{(2n-3)!!}.$$

$S_n^{(1)}$ can also be used to compute the mean value of the length of the undirected path between two given leaves in a phylogenetic tree.

Corollary 11. *For every $i, j \in \{1,\dots,n\}$, $i \neq j$, the mean value of $d_T(i,j)$ for $T \in \mathcal{T}_n$ under the uniform distribution is*

$$\mu\big(d_T(i,j)\big)_n = \frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}}.$$

Proof.

$$\begin{aligned}\mu(d_T(i,j))_n &= \frac{\sum_{T\in\mathcal{T}_n} d_T(i,j)}{|\mathcal{T}_n|} = \frac{S_n^{(1)}}{(2n-3)!!} = \frac{(2n-2)!!}{(2n-3)!!}\\ &= \frac{(2n-2)!!^2}{(2n-3)!!(2n-2)!!} = \frac{2^{2n-2}\cdot(n-1)!^2}{(2n-2)!} = \frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}}.\end{aligned}$$ □

In the unrooted case, this mean value is proved in [26] to be $2^{2(n-2)}/\binom{2(n-2)}{n-2}$.

4. Mean path-difference distance

The *path-difference distance* between a pair of phylogenetic trees $T, T' \in \mathcal{T}_n$ is

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

Lemma 12. *The mean value of $\delta(T,T')^2$, with $T, T \in \mathcal{T}_n$, under the uniform distribution on $\mathcal{T}_n$, is*

$$\mu\big(\delta^2\big)_n = 2\binom{n}{2}\left(4(n-1)+2-\frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}} - \left(\frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}}\right)^2\right).$$

Proof. By definition

$$\begin{aligned}\mu\big(\delta^2\big)_n &= \frac{\sum_{T,T'\in\mathcal{T}_N}\sum_{1\leqslant i<j\leqslant n}(d_T(i,j)-d_{T'}(i,j))^2}{|\mathcal{T}_N|^2}\\ &= \frac{1}{|\mathcal{T}_N|^2}\left(\sum_{1\leqslant i<j\leqslant n}\ \sum_{T,T'\in\mathcal{T}_N}\big(d_T(i,j)^2 + d_{T'}(i,j)^2 - 2d_T(i,j)d_{T'}(i,j)\big)\right)\\ &= \frac{1}{|\mathcal{T}_N|^2}\sum_{1\leqslant i<j\leqslant n}\left(2|\mathcal{T}_n|\sum_{T\in\mathcal{T}_N} d_T(i,j)^2 - 2\left(\sum_{T\in\mathcal{T}_N} d_T(i,j)\right)^2\right)\end{aligned}$$

and then, using Lemma 5,

$$\mu\big(\delta^2\big)_n = 2\binom{n}{2}\left(\frac{\sum_{T\in\mathcal{T}_n} d_T(1,2)^2}{|\mathcal{T}_n|} - \left(\frac{\sum_{T\in\mathcal{T}_n} d_T(1,2)}{|\mathcal{T}_n|}\right)^2\right) = 2\binom{n}{2}\left(\frac{S_n^{(2)}}{(2n-3)!!} - \left(\frac{S_n^{(1)}}{(2n-3)!!}\right)^2\right).$$

If we replace $S_n^{(1)}$ and $S_n^{(2)}$ by their values given in Propositions 8 and 9, we obtain

$$\mu(\delta^2)_n = 2\binom{n}{2}\left(\frac{2(2n-1)!! - (2n-2)!!}{(2n-3)!!} - \left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2\right) = 2\binom{n}{2}\left(4n-2-\frac{(2n-2)!!}{(2n-3)!!} - \left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2\right).$$

Applying finally (see Corollary 11)

$$\frac{(2n-2)!!}{(2n-3)!!} = \frac{2^{2(n-1)}}{\binom{2(n-1)}{n-1}},$$

we obtain the expressions in the statement. □

The value of $\mu(\delta^2)_n$ obtained by Steel and Penny in the unrooted case was

$$\mu\left(\delta^2\right)_n = 2\binom{n}{2}\left(4(n-2)+2-\frac{2^{2(n-2)}}{\binom{2(n-2)}{n-2}} - \left(\frac{2^{2(n-2)}}{\binom{2(n-2)}{n-2}}\right)^2\right).$$

Using Stirling approximation, both mean values are equivalent to

$$2\binom{n}{2}\left((4-\pi)n - \sqrt{\pi n}\right).$$

Acknowledgments

The research reported in this paper has been partially supported by the Spanish government and the UE FEDER program project MTM2006-07773 COMGRIO. We thank G. Valiente for several comments on this work, and the reviewer, whose comments and suggestions lead to a substantial improvement of the final paper.

References

[1] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, 1964. Available on line at http://www.math.ucla.edu/~cbm/aands/.
[2] R. Alberich, G. Cardona, F. Rosselló, G. Valiente, An algebraic metric for phylogenetic trees, Appl. Math. Lett. 22 (2009) 1320–1324.
[3] B.L. Allen, M.A. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. 5 (2001) 1–13.
[4] W.N. Bayley, Generalized Hypergeometric Series, Cambridge Tracts in Math. Math. Phys., vol. 32, Stechert–Hafner Service Inc., 1964.
[5] J. Bluis, D.-G. Shin, Nodal distance algorithm: Calculating a phylogenetic tree comparison metric, in: Proc. 3rd IEEE Symp. Bioinformatics and Bioengineering, 2003, p. 87.
[6] D. Bryant, M.A. Steel, Computing the distribution of a tree metric, IEEE/ACM Trans. Comput. Biol. Bioinformatics 99 (2009) 1545–5963.
[7] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, Nodal distances for rooted phylogenetic trees, J. Math. Biol., doi:10.1007/s00285-009-0295-2, in press.
[8] D.E. Critchlow, D.K. Pearl, C. Qian, The triples distance for rooted bifurcating phylogenetic trees, Syst. Biol. 45 (3) (1996) 323–334.
[9] J.S. Farris, A successive approximations approach to character weighting, Syst. Zool. 18 (1969) 374–385.
[10] J.S. Farris, On comparing the shapes of taxonomic trees, Syst. Zool. 22 (1973) 50–54.
[11] J. Felsenstein, Inferring Phylogenies, Sinauer Associates Inc., 2004.
[12] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, J. Theoret. Biol. 257 (2009) 17–26.
[13] M.D. Hendy, C.H.C. Little, D. Penny, Comparing trees with pendant vertices labelled, SIAM J. Appl. Math. 44 (1984) 1054–1065.
[14] K. Hoef-Emden, Molecular phylogenetic analyses and real-life data, Comput. Sci. Eng. 7 (3) (2005) 86–91.
[15] L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, J. Mol. Biol. 5 (1993) 123–138.
[16] A. Kupczok, A.V. Haeseler, S. Klaere, An exact algorithm for the geodesic distance between phylogenetic trees, J. Comput. Biol. 15 (2008) 577–591.
[17] V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny, Metrics on RNA secondary structures, J. Comput. Biol. 7 (2000) 277–292.
[18] J.J. Nieto, A. Torres, D.N. Georgiou, T.E. Karakasidis, Fuzzy polynucleotide spaces and metrics, Bull. Math. Biol. 68 (2006) 703–725.
[19] M. Petkovsek, H. Wilf, D. Zeilberger, $A = B$, AK Peters Ltd., 1996. Available on line at http://www.math.upenn.edu/~wilf/AeqB.html.
[20] J.B. Phipps, Dendrogram topology, Syst. Zool. 20 (1971) 306–308.
[21] P. Puigbò, S. Garcia-Vallvé, J. McInerney, TOPD/FMTS: a new software to compare phylogenetic trees, Bioinformatics 23 (12) (2007) 1556–1558.
[22] D.F. Robinson, L.R. Foulds, Comparison of weighted labelled trees, in: Proc. 6th Australian Conf. Combinatorial Mathematics, in: Lecture Notes in Math., vol. 748, 1979, pp. 119–126.
[23] D.F. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. 53 (1/2) (1981) 131–147.
[24] Y.A. Smolenskii, A method for the linear recording of graphs, USSR Comput. Math. Math. Phys. 2 (1963) 396–397.
[25] M.A. Steel, Distribution of the symmetric difference metric on phylogenetic trees, SIAM J. Discrete Math. 1 (1988) 541–551.
[26] M.A. Steel, D. Penny, Distributions of tree comparison metrics—Some new results, Syst. Biol. 42 (1993) 126–141.
[27] M.S. Waterman, T.F. Smith, On the similarity of dendograms, J. Theoret. Biol. 73 (1978) 789–800.
[28] M.S. Waterman, T.F. Smith, W.A. Beyer, Some biological sequence metrics, Adv. Math. 20 (1976) 367–387.
[29] W.T. Williams, H.T. Clifford, On the comparison of two classifications of the same set of elements, Taxon 20 (4) (1971) 519–522.
[30] W. Xu, D.P. Miranker, A metric model of amino acid substitution, Bioinformatics 20 (2004) 1214–1221.