AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Computing the skewness of the phylogenetic mean pairwise distance in linear time

Constantinos Tsirogiannis[1,2*] and Brody Sandel[1,2]

Abstract

Background: The phylogenetic Mean Pairwise Distance (MPD) is one of the most popular measures for computing the phylogenetic distance between a given group of species. More specifically, for a phylogenetic tree $\mathcal{T}$ and for a set of species R represented by a subset of the leaf nodes of $\mathcal{T}$, the MPD of R is equal to the average cost of all possible simple paths in $\mathcal{T}$ that connect pairs of nodes in R.

Among other phylogenetic measures, the MPD is used as a tool for deciding if the species of a given group R are closely related. To do this, it is important to compute not only the value of the MPD for this group but also the expectation, the variance, and the skewness of this metric. Although efficient algorithms have been developed for computing the expectation and the variance the MPD, there has been no approach so far for computing the skewness of this measure.

Results: In the present work we describe how to compute the skewness of the MPD on a tree $\mathcal{T}$ optimally, in $\Theta(n)$ time; here n is the size of the tree $\mathcal{T}$. So far this is the first result that leads to an exact, let alone efficient, computation of the skewness for any popular phylogenetic distance measure. Moreover, we show how we can compute in $\Theta(n)$ time several interesting quantities in $\mathcal{T}$, that can be possibly used as building blocks for computing efficiently the skewness of other phylogenetic measures.

Conclusions: The optimal computation of the skewness of the MPD that is outlined in this work provides one more tool for studying the phylogenetic relatedness of species in large phylogenetic trees. Until now this has been infeasible, given that traditional techniques for computing the skewness are inefficient and based on inexact resampling.

Keywords: Algorithms for phylogenetic trees, Mean pairwise distance, Skewness

Background

Communities of co-occuring species may be described as "clustered" if species in the community tend to be close phylogenetic relatives of one another, or "overdispersed" if they are distant relatives [1]. To define these terms we need a function that measures the phylogenetic relatedness of a set of species, and also a point of reference for how this function should behave in the absence of ecological and evolutionary processes. One such function is the mean pairwise distance (MPD); given a phylogenetic tree $\mathcal{T}$ and a subset of species R that are represented by leaf nodes of $\mathcal{T}$, the MPD of the species in R is equal to average cost of all possible simple paths that connect pairs of nodes in R.

To decide if the value of the MPD for a specific set of species R is large or small, we need to know the average value (expectation) of the MPD for all sets of species in $\mathcal{T}$ that consist of exactly $r = |R|$ species. To judge how much larger or smaller is this value from the average, we also need to know the standard deviation of the MPD for all possible sets of r species in $\mathcal{T}$. Putting all these values together, we get the following index that expresses how clustered are the species in R [1]:

$$\text{NRI} = \frac{\text{MPD}(\mathcal{T}, R) - \text{expec}_{\text{MPD}}(\mathcal{T}, r)}{sd_{\text{MPD}}(\mathcal{T}, r)},$$

where $\text{MPD}(\mathcal{T}, R)$ is the value of the MPD for R in $\mathcal{T}$, and $\text{expec}(\mathcal{T})$ and $sd_{\text{MPD}}(\mathcal{T}, r)$ are the expected value and the standard deviation respectively of the MPD calculated over all subsets of r species in $\mathcal{T}$.

In a previous paper we presented optimal algorithms for computing the expectation and the standard deviation of the MPD of a phylogenetic tree $\mathcal{T}$ in $O(n)$ time, where n

*Correspondence: constant@cs.au.dk
[1] MADALGO, Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation, Aarhus University, Aarhus, Denmark
[2] Department of Bioscience, Aarhus University, Aarhus, Denmark

is the number of the edges of $\mathcal{T}$ [2]. This enabled exact computations of these statistical moments of the MPD on large trees, which were previously infeasible using traditional slow and inexact resampling techniques. However, an important problem remained unsolved; quantifying our degree of confidence that the NRI value observed in a community reflects non-random ecological and evolutionary processes.

This degree of confidence can be expressed as a statistical P value, that is the probability that we would observe an NRI value as extreme or more so if the community was randomly assembled. Traditionally, estimating P is accomplished by ranking the observed MPD against the distribution of randomized MPD values [3]. If the MPD falls far enough into one of the tails of the distribution (generally below the 2.5 percentile or above the 97.5 percentile, yielding $P < 0.05$), the community is said to be significantly overdispersed or significantly clustered. However, this approach relies on sampling a large number of random subsets of species in $\mathcal{T}$, and recomputing the MPD for each random subset. Therefore, this method is slow and imprecise. This problem is exacerbated when it is necessary to consider multiple trees at once, arising for example from a Bayesian posterior sample of trees [4,5]. In such cases, sufficient resampling from all trees in the sample can be computationally limiting.

We can approximate the P value of an observed NRI by assuming a particular distribution of the possible MPD values and evaluating its cumulative distribution function at the observed MPD. Because the NRI measures the difference between the observed values and expectation in units of standard deviations, this yields a very simple rule if we assume that possible MPD values are normally distributed: any NRI value larger than 1.96 or smaller than -1.96 is significant. Unfortunately, the distribution of MPD values is often skewed, such that this simple rule will lead to incorrect P value estimates [6,7]. Of particular concern, this skewness introduces a bias towards detecting either significant clustering or significant overdispersion [8]. If the distribution of MPD values for a particular tree can be reasonably approximated using a skew-normal distribution, calculating the skewness analytically would enable us to remove this bias and improve the accuracy of P value estimates. In the last part of the paper, we describe experiments on large randomly generated trees, supporting this argument. Further, when a large sample of trees should be considered, the full distribution of MPD values can be considered as a mixture of skew-normal distributions [9,10], greatly simplifying and speeding up the process of calculating P values across the entire set of trees.

However, so far there has been no result in the related literature that shows how to compute the needed skewness measure efficiently. Hence, given a phylogenetic tree $\mathcal{T}$ and an integer r there is the need to design an efficient and exact algorithm that can compute the skewness of the MPD for r species in $\mathcal{T}$. This would provide the last critical piece required for the adoption of a fully analytical and efficient approach for analysing ecological communities using the MPD and the NRI.

Our results

In the present work we show how we can compute efficiently the skewness of the MPD. More specifically, given a tree $\mathcal{T}$ that consists of n edges and a positive integer r, we prove that we can compute the skewness of of the MPD over all subsets of r leaf nodes in $\mathcal{T}$ optimally, in $\Theta(n)$ time. For the calculation of this skewness value we consider that every subset of exactly r species in $\mathcal{T}$ is picked uniform with probability out of all possible subsets that have r species. The main contribution of this paper is a constructive proof that leads straightforwardly to an algorithm that computes the skewness of the MPD in $\Theta(n)$ time. This is clearly optimal, and it outperforms even the best algorithms that are known so far for computing lower-order statistics for other phylogenetic measures; for example the best known algorithm for computing the variance of the popular Phylogenetic Diversity (PD) runs in $O(n^2)$ time [2].

More than that, we prove how we can compute in $\Theta(n)$ time several quantities that are related with groups of paths in the given tree; these quantities can be possibly used as building blocks for computing efficiently the skewness (and other statistical moments) of phylogenetic measures that are similar to the MPD. Such an example is the measure which is the equivalent of the MPD for computing the distance between two subsets of species in $\mathcal{T}$ [11].

The rest of this paper is, almost in its entirety, an elaborate proof for computing the skewness of the MPD on a tree $\mathcal{T}$ in $\Theta(n)$ time. In the next section we define the problem that we want to tackle, and we present a group of quantities that we use as building blocks for computing the skewness of the MPD. We prove that all of these quantities can be computed in linear time with respect to the size of the input tree. Then, we provide the main proof of this paper; there we show how we can express the value of the skewness of the MPD in terms of the quantities that we introduced earlier. The proof implies a straightforward linear time algorithm for the computation of the skewness as well. In the last section we provide experimental results that indicate that computing the skewness of the MPD can be a useful tool for improving the estimation of P values when a skew-normal distibution is assumed. There we describe experiments that we conducted on large randomly generated trees to compare two different methods for estimating P values; one method is based on random sampling of a large number of tip sets, and the other method relies in calculating the mean, variance, and skewness of the MPD for the given tree.

Description of the problem and basic concepts

Definitions and notation Let $\mathcal{T}$ be a phylogenetic tree, and let E be the set of its edges. We denote the number of the edges in $\mathcal{T}$ by n, that is $n = |E|$. For an edge $e \in E$, we use w_e to indicate the weight of this edge. We use S to denote the set of the leaf nodes of $\mathcal{T}$. We call these nodes the *tips* of the tree, and we use s to denote the number of these nodes.

Since a phylogenetic tree is a rooted tree, for any edge $e \in E$ we distinguish the two nodes adjacent to e into a *parent* node and a *child* node; among these two, the parent node of e is the one for which the simple path from this node to the root does not contain e. We use $\mathrm{Ch}(e)$ to indicate the set of edges whose parent node is the child node of e, which of course implies that $e \notin \mathrm{Ch}(e)$. We indicate the edge whose child node is the parent node of e by $\mathrm{parent}(e)$. For any edge $e \in E$, tree $\mathcal{T}(e)$ is the subtree of $\mathcal{T}$ whose root is the child node of edge e. We denote the set of tips that appear in $\mathcal{T}(e)$ as $S(e)$, and we denote the number of these tips by $s(e)$.

Given any edge $e \in E$, we partition the edges of $\mathcal{T}$ into three subsets. The first subset consists of all the edges that appear in the subtree of e. We denote this set by $\mathrm{Off}(e)$. The second subset consists of all edges $e' \in E$ for which e appears in the subtree of e'. We use $\mathrm{Anc}(e)$ to indicate this subset. For the rest of this paper, we define that $e \in \mathrm{Anc}(e)$, and that $e \notin \mathrm{Off}(e)$. The third subset contains all the tree edges that do not appear neither in $\mathrm{Off}(e)$, nor in $\mathrm{Anc}(e)$; we indicate this subset by $\mathrm{Ind}(e)$.

For any two tips $u, v \in S$, we use $p(u, v)$ to indicate the simple path in $\mathcal{T}$ between these nodes. Of course, the path $p(u, v)$ is unique since $\mathcal{T}$ is a tree. We use $cost(u, v)$ to denote the cost of this path, that is the sum of the weights of all the edges that appear on the path. Let u be a tip in S and let e be an edge in E. We use $cost(u, e)$ to represent the cost of the shortest simple path between u and the child node of e. Therefore, if $u \in S(e)$ this path does not include e, otherwise it does. For any subset $R \subseteq S$ of the tips of the tree $\mathcal{T}$, we denote the set of all pairs of elements in R, that is the set of all combinations that consist of two distinct tips in R, by $\Delta(R)$. Given a phylogenetic tree $\mathcal{T}$ and a subset of its tips $R \subseteq S$, we denote the Mean Pairwise Distance of R in $\mathcal{T}$ by $\mathrm{MPD}(\mathcal{T}, R)$. Let $r = |R|$. This measure is equal to:

$$\mathrm{MPD}(\mathcal{T}, R) = \frac{2}{r(r-1)} \sum_{\{u,v\} \in \Delta(R)} cost(u, v) \ .$$

Aggregating the costs of paths

Let $\mathcal{T}$ be a phylogenetic tree that consists of n edges and s tips, and let r be a positive integer such that $r \leq s$. We use $\mathrm{sk}(\mathcal{T}, r)$ to denote the skewness of the MPD on $\mathcal{T}$ when we pick a subset of r tips of this tree with uniform probability. In the rest of this paper we describe in detail how we can compute $\mathrm{sk}(\mathcal{T}, r)$ in $O(n)$ time, by scanning $\mathcal{T}$ only a constant number of times. Based on the formal definition of skewness, the value of $\mathrm{sk}(\mathcal{T}, r)$ is equal to:

$$\mathrm{sk}(\mathcal{T}, r) = E_{R \in \mathrm{Sub}(S,r)} \left[\left(\frac{\mathrm{MPD}(\mathcal{T}, R) - \mathrm{expec}(\mathcal{T}, r)}{\mathrm{var}(\mathcal{T}, r)} \right)^3 \right]$$
$$= \frac{E_{R \in \mathrm{Sub}(S,r)} \left[\mathrm{MPD}^3(\mathcal{T}, R) \right] - 3 \cdot \mathrm{var}(\mathcal{T}, r)^2 - \mathrm{expec}(\mathcal{T}, r)^3}{\mathrm{var}(\mathcal{T}, r)^3} \ , \qquad (1)$$

where $\mathrm{expec}(\mathcal{T}, r)$ and $\mathrm{var}(\mathcal{T}, r)$ are the expectation and the variance of the MPD for subsets of exactly r tips in $\mathcal{T}$, and $E_{R \in \mathrm{Sub}(S,r)}[\,\cdot\,]$ denotes the function of the expectation over all subsets of exactly r tips in S. In a previous paper, we showed how we can compute the expectation and the variance of the MPD on $\mathcal{T}$ in $O(n)$ time [2]. Therefore, in the rest of this work we focus on analysing the value $E_{R \in \mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T}, R)]$ and expressing this quantity in a way that can be computed efficiently, in linear time with respect to the size of $\mathcal{T}$.

To make things more simple, we break the description of our approach into two parts; in the first part, we define several quantities that come from adding and multiplying the costs of specific subsets of paths between tips of the tree. We also present how we can compute all these quantities in $O(n)$ time in total by scanning $\mathcal{T}$ a constant number of times. Then, in the next section, we show how we can express the skewness of the MPD on $\mathcal{T}$ based on these quantities, and hence compute the skewness in $O(n)$ time as well. Next we provide the quantities that we want to consider in our analysis; these quantities are described in Table 1. In this table but also in the rest of this work, for any tip $u \in S$, we consider that $\mathrm{SQ}(u) = \mathrm{SQ}(e)$, and $\mathrm{TC}(u) = \mathrm{TC}(e)$, such that e is the edge whose child node is u.

We provide now the following lemma.

Lemma 1. *Given a phylogenetic tree $\mathcal{T}$ that consists of n edges, we can compute all the quantities that are presented in Table 1 in $O(n)$ time in total.*

Proof. Each of the quantities (I)-(X) in Table 1 can be computed by scanning a constant number of times the input tree $\mathcal{T}$, either bottom-up or top-to-bottom. For computing quantity (XI) we follow a more involved divide-and-conquer approach.

We showed in a previous paper how we can compute quantity (I) and the quantities in (III) for all $e \in E$ in $O(n)$ time in total [2].

Table 1 The quantities that we use for expressing the skewness of the MPD

I) $\mathrm{TC}(\mathcal{T}) = \sum_{\{u,v\} \in \Delta(S)} cost(u,v)$	II) $\mathrm{CB}(\mathcal{T}) = \sum_{\{u,v\} \in \Delta(S)} cost^3(u,v)$
III) $\forall e \in E,\ \mathrm{TC}(e) = \sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} cost(u,v)$	IV) $\forall e \in E,\ \mathrm{SQ}(e) = \sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} cost^2(u,v)$
V) $\forall e \in E,\ \mathrm{Mult}(e) = \sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} \mathrm{TC}(u) \cdot \mathrm{TC}(v)$	VI) $\forall u \in S,\ \mathrm{SM}(u) = \sum_{v \in S \setminus \{u\}} cost(u,v) \cdot \mathrm{TC}(v)$
VII) $\forall e \in E,\ \mathrm{TC}_{\mathrm{sub}}(e) = \sum_{u \in S(e)} cost(u,e)$	VIII) $\forall e \in E,\ \mathrm{SQ}_{\mathrm{sub}}(e) = \sum_{u \in S(e)} cost^2(u,e)$
IX) $\forall e \in E,\ \mathrm{PC}(e) = \sum_{u \in S} cost(u,e)$	X) $\forall e \in E,\ \mathrm{PSQ}(e) = \sum_{u \in S} cost^2(u,e)$
XI) $\forall e \in E,\ \mathrm{QD}(e) = \sum_{u \in S(e)} \left(\sum_{v \in S(e) \setminus \{u\}} cost(u,v) \right)^2$	

For an edge $e \in E$, the quantity in (VII) can be written as:

$$\mathrm{TC}_{\mathrm{sub}}(e) = \sum_{u \in S(e)} cost(u,e) = \sum_{l \in \mathrm{Off}(e)} w_l \cdot s(l)\ .$$

We can compute this quantity for every $e \in E$ in linear time as follows; in the first scan we compute for every edge e the number of leaves $s(e)$ in $\mathcal{T}(e)$. This can be done in $O(n)$ time by computing in a bottom-up manner $s(e)$ as the sum of the numbers of tips $s(e')$, $\forall e' \in \mathrm{Ch}(e)$. Then, we can compute $\mathrm{TC}_{\mathrm{sub}}(e)$ by scanning bottom-up the tree using the following formula:

$$\mathrm{TC}_{\mathrm{sub}}(e) = \sum_{l \in \mathrm{Ch}(e)} w_l \cdot s(l) + \mathrm{TC}_{\mathrm{sub}}(l)\ .$$

For quantity (VIII), for any $e \in E$ we have that:

$$\begin{aligned}
\mathrm{SQ}_{\mathrm{sub}}(e) &= \sum_{u \in S(e)} cost^2(u,e) \\
&= \sum_{l \in \mathrm{Off}(e)} w_l \sum_{k \in \mathrm{Off}(l)} 2 \cdot w_k \cdot s(k) + \sum_{l \in \mathrm{Off}(e)} w_l^2 \cdot s(l) \\
&= \sum_{l \in \mathrm{Off}(e)} 2 \cdot w_l \cdot \mathrm{TC}_{\mathrm{sub}}(l) + w_l^2 \cdot s(l)\ .
\end{aligned}$$

Then $\mathrm{SQ}_{\mathrm{sub}}(e)$ can be computed for every edge $e \in E$ by scanning $\mathcal{T}$ bottom up and evaluating the formula:

$$\mathrm{SQ}_{\mathrm{sub}}(e) = \sum_{l \in \mathrm{Ch}(e)} 2 \cdot w_l \cdot \mathrm{TC}_{\mathrm{sub}}(l) + w_l^2 \cdot s(l) + \mathrm{SQ}_{\mathrm{sub}}(l)\ .$$

For every edge e in $\mathcal{T}$, quantity (IV) can be written as:

$$\begin{aligned}
\sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} cost^2(u,v) &= 2 \sum_{l,k \in E} w_l \cdot w_k \cdot \mathrm{NumPath}(e,l,k) \\
&\quad + \sum_{l \in E} w_l^2 \cdot \mathrm{NumPath}(e,l)\ .
\end{aligned}$$

In the last expression, value $\mathrm{NumPath}(e,l,k)$ is equal to the number of simple paths that connect two tips in $\mathcal{T}$ and which also contain all three edges e, l and k. The quantity $\mathrm{NumPath}(e,l)$ is equal to the number of simple paths that connect two tips in $\mathcal{T}$ and which also contain both edges e and l. Therefore, for any $e \in E$ we have:

$$\begin{aligned}
\sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} cost^2(u,v) &= 2(s - s(e)) \sum_{l \in \mathrm{Off}(e)} w_l \sum_{k \in \mathrm{Off}(l)} w_k \cdot s(k) \\
&\quad + 2 \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l)) \sum_{k \in \mathrm{Off}(e)} w_k \cdot s(k) \\
&\quad + 2 \cdot s(e) \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l)) \sum_{\substack{k \in \mathrm{Anc}(e) \\ k \in \mathrm{Off}(l)}} w_k \\
&\quad + 2 \cdot s(e) \sum_{l \in \mathrm{Ind}(e)} w_l \sum_{k \in \mathrm{Off}(l)} w_k \cdot s(k) \\
&\quad + 2 \sum_{l \in \mathrm{Ind}(e)} w_l \cdot s(l) \sum_{k \in \mathrm{Off}(e)} w_k \cdot s(k) \\
&\quad + 2 \cdot s(e) \sum_{l \in \mathrm{Anc}(e)} w_l \sum_{k \in \mathrm{Ind}(l)} w_k \cdot s(k) \\
&\quad + (s - s(e)) \sum_{l \in \mathrm{Off}(e)} w_l^2 \cdot s_l + s(e) \\
&\quad \sum_{l \in \mathrm{Anc}(e)} w_l^2 \cdot (s - s(l)) + s(e) \sum_{l \in \mathrm{Ind}(e)} w_l^2 \cdot s(l) \\
&= (s - s(e)) \cdot \mathrm{SQ}_{\mathrm{sub}}(e) \\
&\quad + \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l)) (2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) + w_l \cdot s(e)) \\
&\quad + 2 \cdot s(e) \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l)) \left(\sum_{\substack{k \in \mathrm{Anc}(e) \\ k \in \mathrm{Off}(l)}} w_k \right) \\
&\quad + s(e) \sum_{l \in \mathrm{Ind}(e)} w_l (2 \cdot \mathrm{TC}_{\mathrm{sub}}(l) \\
&\quad + w_l \cdot s(l)) + 2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) \sum_{l \in \mathrm{Ind}(e)} w_l \cdot s(l) \\
&\quad + 2 \cdot s(e) \sum_{l \in \mathrm{Anc}(e)} w_l \sum_{k \in \mathrm{Ind}(l)} w_k \cdot s(k)\ .
\end{aligned} \tag{2}$$

We explain now how we can compute the six quantities in (2) in $O(n)$ time, assuming that we have already computed $\mathrm{TC}_{\mathrm{sub}}(e)$ and $s(e)$ for every $e \in E$. To make the description simpler, we show in detail how we can compute the second and fourth quantities that appear in the last expression; it is easy to show that the rest of the quantities in (2) can be calculated in a similar manner.

For any $e \in E$, we denote the second quantity as follows:

$$\mathrm{SUM}_1(e) = \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l))(2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) + w_l \cdot s(e)) \ .$$

We also define the following quantities:

$$\mathrm{SUM}_{1A}(e) = \sum_{l \in \mathrm{Anc}(e)} w_l (s - s(l)) \ ,$$

and

$$\mathrm{SUM}_{1B}(e) = \sum_{l \in \mathrm{Anc}(e)} w_l^2 (s - s(l)) \ .$$

We can calculate $\mathrm{SUM}_1(e)$ for every edge e by traversing the tree top-to-bottom and evaluating the following expressions:

$$\mathrm{SUM}_{1A}(e) = w_e(s - s(e)) + \mathrm{SUM}_{1A}(\mathrm{parent}(e)) \ .$$

$$\mathrm{SUM}_{1B}(e) = w_e^2(s - s(e)) + \mathrm{SUM}_{1B}(\mathrm{parent}(e)) \ .$$

$$\mathrm{SUM}_1(e) = 2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) \cdot \mathrm{SUM}_{1A}(e) + \mathrm{SUM}_{1B}(e) \cdot s(e) \ .$$

To compute the fourth quantity in (2), we use the following quantity:

$$\mathrm{SUM}_2(e) = \sum_{l \in \mathrm{Off}(e)} w_l (2 \cdot \mathrm{TC}_{\mathrm{sub}}(l) + w_l \cdot s(l)) \ .$$

This quantity can be evaluated in $O(n)$ time for every $e \in E$ with a bottom-up scan of the tree. We also consider the following value which we can precompute in $O(n)$ time:

$$\mathrm{SUM}_2(\mathcal{T}) = \sum_{e \in E} w_e (2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) + w_e \cdot s(e)) \ .$$

For every edge $e \in E$ we calculate in a top-to-bottom manner the formula:

$$\mathrm{SUM}_3(e) = w_e(2 \cdot \mathrm{TC}_{\mathrm{sub}}(e) + w_e \cdot s(e)) + \mathrm{SUM}_3(\mathrm{parent}(e)) \ .$$

Then for each tree edge e, the fourth quantity in (2) can be computed in constant time as follows:

$$\begin{aligned} & s(e) \sum_{l \in \mathrm{Ind}(e)} w_l (2 \cdot \mathrm{TC}_{\mathrm{sub}}(l) + w_l \cdot s(l)) \\ & = s(e) \cdot (\mathrm{SUM}_2(\mathcal{T}) - \mathrm{SUM}_2(e) - \mathrm{SUM}_3(e)) \ . \end{aligned}$$

The remaining quantities in (2) can be computed in a quite similar manner as the two quantities that we already described.

Quantity (II) in Table 1 is equal to:

$$\begin{aligned} \mathrm{CB}(\mathcal{T}) &= \sum_{\{u,v\} \in \Delta(S)} cost^3(u,v) \\ &= \sum_{e \in E} w_e \sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} cost^2(u,v) = \sum_{e \in E} w_e \cdot \mathrm{SQ}(e) \ . \end{aligned}$$

We have already presented how to compute $\mathrm{SQ}(e)$ for every edge e in $\mathcal{T}$ in $O(n)$ time in total, hence we can also compute $\mathrm{CB}(\mathcal{T})$ in $O(n)$ time by simply summing up the values $w_e \cdot \mathrm{SQ}(e)$ for every edge e in the tree. For quantity (V) it holds that:

$$\begin{aligned} \mathrm{Mult}(e) &= \sum_{\substack{\{u,v\} \in \Delta(S) \\ e \in p(u,v)}} \mathrm{TC}(u) \cdot \mathrm{TC}(v) \\ &= \sum_{u \in S(e)} \mathrm{TC}(u) \sum_{v \in S - S(e)} \mathrm{TC}(v) \\ &= \left(\sum_{u \in S(e)} \mathrm{TC}(u) \right) \left(\sum_{v \in S} \mathrm{TC}(v) - \sum_{u \in S(e)} \mathrm{TC}(u) \right) . \end{aligned}$$

Since we have already computed $\mathrm{TC}(v)$ for every tip $v \in S$, we can trivially evaluate $\sum_{v \in S} \mathrm{TC}(v)$ in $O(n)$ time. Hence, to compute quantity (V) it remains now to calculate the values $\mathrm{SUM}_4(e) = \sum_{u \in S(e)} \mathrm{TC}(u)$ for every edge $e \in E$. We can do this in $O(n)$ time as follows: at each tip $u \in S$ we store the value $\mathrm{TC}(u)$ that we have already computed. Then we scan $\mathcal{T}$ bottom-up and we calculate $\mathrm{SUM}_4(e)$ by summing up the values $\mathrm{SUM}_4(l)$ for all edges $l \in \mathrm{Ch}(e)$.

Let u be a tip in S, and let e be the edge which is adjacent to u. Then, quantity (VI) is equal to:

$$\begin{aligned} \mathrm{SM}(u) &= \sum_{v \in S \setminus \{u\}} cost(u,v) \cdot \mathrm{TC}(v) \\ &= \sum_{l \in \mathrm{Anc}(e)} w_l \sum_{v \in S \setminus S(l)} \mathrm{TC}(v) + \sum_{l \in \mathrm{Ind}(e)} w_l \sum_{v \in S(l)} \mathrm{TC}(v) \\ &= \sum_{l \in \mathrm{Anc}(e)} w_l \left(\sum_{v \in S} \mathrm{TC}(v) - \sum_{x \in S(l)} \mathrm{TC}(x) \right) \\ &\quad + \sum_{l \in E} w_l \sum_{v \in S(l)} \mathrm{TC}(v) - \sum_{l \in \mathrm{Anc}(e)} w_l \sum_{v \in S(l)} \mathrm{TC}(v) \ . \end{aligned}$$

In the last expression, value $\sum_{v \in S} \mathrm{TC}(v)$ can be computed in $O(n)$ time, given that we have already computed $\mathrm{TC}(v)$ for every $v \in S$. Value $\sum_{l \in E} w_l \sum_{v \in S(l)} \mathrm{TC}(v)$ and values $\sum_{x \in S(l)} \mathrm{TC}(x)$ for any $l \in E$ can be calculated with a bottom-up scan of $\mathcal{T}$ in a similar way as we computed $\mathrm{TC}_{\mathrm{sub}}(e)$ for all $e \in E$. The remaining sums that involve edges in $\mathrm{Anc}(e)$ can be computed in linear time for every edge e with a similar mechanism as with $\mathrm{SUM}_3(e)$ that we described earlier in this proof. For any edge $e \in E$, quantities $\mathrm{PC}(e)$ and $\mathrm{PSQ}(e)$ in Table 1 are equal to:

$$\mathrm{PC}(e) = \sum_{u \in S} cost(u,e) = \mathrm{TC}_{\mathrm{sub}}(e) + \sum_{l \in \mathrm{Ind}(e)} w_l \cdot s(l) + \sum_{l \in \mathrm{Anc}(e)} w_l(s - s(l)) \; ,$$

and:

$$\begin{aligned}\mathrm{PSQ}(e) = \sum_{u \in S} cost^2(u,e) &= \mathrm{SQ}_{\mathrm{sub}}(e) \\ &+ 2 \sum_{l \in \mathrm{Ind}(e)} w_l \cdot \mathrm{TC}_{\mathrm{sub}}(l) + w_l^2 \cdot s(l) \\ &+ 2 \sum_{l \in \mathrm{Anc}(e)} w_l \sum_{k \in \mathrm{Ind}(l)} w_k \cdot s_k \\ &+ 2 \sum_{l \in \mathrm{Anc}(e)} w_l (s - s_l) \left(\sum_{\substack{k \in \mathrm{Anc}(e) \\ k \in \mathrm{Off}(l)}} w_k \right) + w_l^2 (s - s(l)).\end{aligned}$$

From the two last expressions, and given the description that we provided for other similar quantities in Table 1, it easy to conclude that $\mathrm{PC}(e)$ can be evaluated for every edge e in $O(n)$ time by scanning $\mathcal{T}$ a constant number of times. Having computed $\mathrm{PC}(e)$ for all edges $e \in E$, the quantity $\mathrm{PSQ}(e)$ can be computed in a similar manner.

Next we describe a divide-and-conquer approach for computing in $\Theta(n)$ time quantity (XI) in Table 1 for every $e \in E$. Before we start our description, we define one more quantity that will help us simplify the rest of this proof. For an edge $e \in E$ and a tip $u \in S(e)$ we define that $\mathrm{TC}_e(u)$ is equal to:

$$\mathrm{TC}_e(u) = \sum_{v \in S(e) \setminus \{u\}} cost(u,v) \; .$$

For any edge $e \in E$ it is easy to show that:

$$\sum_{u \in S(e)} \mathrm{TC}_e(u) = \sum_{u \in S(e)} \mathrm{TC}(u) - \mathrm{TC}(e) \tag{3}$$

Therefore, according to (3) we can compute the sum $\sum_{u \in S(e)} \mathrm{TC}_e(u)$ for all edges $e \in E$ in linear time in total, given that we have already computed $\mathrm{TC}(e)$ for every $e \in E$, and $\mathrm{TC}(u)$ for every $u \in S$.

Next we continue our description for computing $\mathrm{QD}(e)$ using a divide-and-conquer approach. We start with the base case; for every edge tree e that is adjacent to a leaf node we have:

$$\mathrm{QD}(e) = \sum_{u \in S(e)} \left(\sum_{v \in S(e) \setminus \{u\}} cost(u,v) \right)^2 = 0.$$

For any edge $e \in E$ that is not adjacent to a leaf node, we can calculate $\mathrm{QD}(e)$ using the values of the respective quantities of the edges in $\mathrm{Ch}(e)$:

$$\begin{aligned}\mathrm{QD}(e) = &\sum_{l \in \mathrm{Ch}(e)} \mathrm{QD}(l) \\ &+ 2 \sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \sum_{v \in S(e) \setminus S(l)} cost(u,v) \cdot \mathrm{TC}_l(u) \\ &+ \sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \left(\sum_{v \in S(e) \setminus S(l)} cost(u,v) \right)^2 .\end{aligned} \tag{4}$$

The first sum in (4) can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time for each edge e, given that we have already computed the values $\mathrm{QD}(l)$ for every $l \in \mathrm{Ch}(e)$. We leave the description for calculating the second sum in (4) for the end of this proof. The third sum in this expression is equal to:

$$\begin{aligned}&\sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \left(\sum_{v \in S(e) \setminus S(l)} cost(u,v) \right)^2 \\ &= \sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \left(\sum_{v \in S(e) \setminus S(l)} cost(u,l) + cost(v,l) \right)^2 \\ &= \sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \sum_{v \in S(e) \setminus S(l)} \sum_{x \in S(e) \setminus S(l)} cost^2(u,l) \\ &\quad + cost(u,l) \cdot cost(v,l) + cost(u,l) \cdot cost(x,l) \\ &\quad + cost(v,l) \cdot cost(x,l) \; .\end{aligned} \tag{5}$$

The first term of the sum in (5) can be expressed as:

$$\begin{aligned}&\sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} \sum_{v \in S(e) \setminus S(l)} \sum_{x \in S(e) \setminus S(l)} cost^2(u,l) \\ &= \sum_{l \in \mathrm{Ch}(e)} \sum_{u \in S(l)} (s(e) - s(l))^2 \cdot cost^2(u,l) \\ &= \sum_{l \in \mathrm{Ch}(e)} (s(e) - s(l))^2 \cdot \mathrm{SQ}_{\mathrm{sub}}(l) \; ,\end{aligned} \tag{6}$$

and can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time, given that we have already computed $\mathrm{SQ}_{\mathrm{sub}}(l), \forall l \in \mathrm{Ch}(e)$.

The next two parts of the sum in (5) are equal to:

$$
\begin{aligned}
&\sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} \sum_{x\in S(e)\setminus S(l)} cost(u,l)\cdot cost(v,l)\\
&\quad + cost(u,l)\cdot cost(x,l)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} 2\cdot (s(e)-s(l))\cdot cost(u,l)\cdot cost(v,l)\\
&= 2\sum_{l\in \mathrm{Ch}(e)} (s(e)-s(l)) \sum_{u\in S(l)} cost(u,l)\left(w_l(s(e)-s(l)) \right.\\
&\quad + \sum_{k\in \mathrm{Ch}(e)} (w_k\cdot s(k)) - w_l\cdot s(l)\\
&\quad \left. + \left(-\sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\right) - \mathrm{TC}_{\mathrm{sub}}(l)\right)\\
&= 2\sum_{l\in \mathrm{Ch}(e)} (s(e)-s(l))\cdot \mathrm{TC}_{\mathrm{sub}}(l)\cdot \left(w_l(s(e)-s(l)) \right.\\
&\quad + \left(\sum_{k\in \mathrm{Ch}(e)} w_k\cdot s(k)\right) - w_l\cdot s(l) + \sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\\
&\quad \left. - \mathrm{TC}_{\mathrm{sub}}(l)\right) .
\end{aligned}
\tag{7}
$$

The last expression can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time as well, if we have already computed the sum $\sum_{k\in \mathrm{Ch}(e)} w_k \cdot s(k)$ and the quantity $\mathrm{TC}_{\mathrm{sub}}(e)$ for every edge e in the tree. We can rewrite the remaining term in (5) as:

$$
\begin{aligned}
&\sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} \sum_{x\in S(e)\setminus S(l)} cost(v,l)\cdot cost(x,l)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \left(\sum_{v\in S(e)\setminus S(l)} cost(v,l)\right)^2\\
&= \sum_{l\in \mathrm{Ch}(e)} s(l)\cdot \left(\sum_{v\in S(e)\setminus S(l)} cost(v,l)\right)^2\\
&= \sum_{l\in \mathrm{Ch}(e)} s(l)\cdot \left(w_l(s(e)-s(l)) + \sum_{k\in \mathrm{Ch}(e)} w_k\cdot s(k) - w_l\cdot s(l) \right.\\
&\quad \left. + \sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k) - \mathrm{TC}_{\mathrm{sub}}(l)\right)^2 .
\end{aligned}
\tag{8}
$$

The last expression can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time in a similar way as the previous terms of the sum in (5).

We left for the end the description of the calculation of the second sum in (4). We can express this sum as follows:

$$
\begin{aligned}
&\sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} cost(u,v)\cdot \mathrm{TC}_l(u)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} (cost(u,l)+cost(v,l))\cdot \mathrm{TC}_l(u)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} cost(u,l)\cdot \mathrm{TC}_l(u)\\
&\quad + \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} cost(v,l)\cdot \mathrm{TC}_l(u)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} (s(e)-s(l))\cdot cost(u,l)\cdot \mathrm{TC}_l(u)\\
&\quad + \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} cost(v,l)\cdot \mathrm{TC}_l(u) .
\end{aligned}
\tag{9}
$$

We start with the second sum in (9). For this sum we get:

$$
\begin{aligned}
&\sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \sum_{v\in S(e)\setminus S(l)} cost(v,l)\cdot \mathrm{TC}_l(u)\\
&= \sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \left(w_l(s(e)-s(l)) + \left(\sum_{k\in \mathrm{Ch}(e)} w_k\cdot s(k)\right)\right.\\
&\quad \left. - w_l\cdot s(l) + \left(\sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\right) - \mathrm{TC}_{\mathrm{sub}}(l)\right)\cdot \mathrm{TC}_l(u).
\end{aligned}
$$

Because of (3), the last expression can be written as:

$$
\begin{aligned}
&\sum_{l\in \mathrm{Ch}(e)} \sum_{u\in S(l)} \left(w_l(s(e)-s(l)) + \sum_{k\in \mathrm{Ch}(e)} (w_k\cdot s(k)) - w_l\cdot s(l)\right.\\
&\quad \left. + \left(\sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\right) - \mathrm{TC}_{\mathrm{sub}}(l)\right)\cdot \mathrm{TC}_l(u)\\
&= \sum_{l\in \mathrm{Ch}(e)} \left(w_l(s(e)-s(l)) + \left(\sum_{k\in \mathrm{Ch}(e)} w_k\cdot s(k)\right) - w_l\cdot s(l)\right.\\
&\quad \left. + \left(\sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\right) - \mathrm{TC}_{\mathrm{sub}}(l)\right)\left(\sum_{u\in S(e)} \mathrm{TC}(u) - \mathrm{TC}(e)\right),
\end{aligned}
$$

which takes $\Theta(|\mathrm{Ch}(e)|)$ time to be computed for each edge e.

To compute the first sum in (9) efficiently, we need to precompute for every edge $l\in E$ the following quantity:

$$\sum_{u\in S(e)} cost(u,e)\cdot \mathrm{TC}_e(u) .$$

To do this, we follow again a divide-and-conquer approach. We get the base case for this computation for the edges of $\mathcal{T}$ that are adjacent to tips. For any such edge e we have:

$$\sum_{u\in S(e)} cost(u,e)\cdot \mathrm{TC}_e(u) = 0\ .$$

For any other edge $e \in E$ we can compute this quantity based on the respective quantities of the edges in $\mathrm{Ch}(e)$. In particular, we have that:

$$\begin{aligned}\sum_{u\in S(e)} cost(u,e)\cdot \mathrm{TC}_e(u) &= \sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)} cost(u,l)\cdot \mathrm{TC}_l(u)\\ &\quad + \sum_{l\in \mathrm{Ch}(e)} w_l \sum_{u\in S(l)} \mathrm{TC}_l(u) + \sum_{l\in \mathrm{Ch}(e)}\\ &\quad \sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,e)\cdot cost(u,v)\\ &= \sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)} cost(u,l)\cdot \mathrm{TC}_l(u) + \sum_{l\in \mathrm{Ch}(e)}\\ &\quad \times w_l \left(\sum_{u\in S(l)} \mathrm{TC}(u) - \mathrm{TC}(l)\right) + \sum_{l\in \mathrm{Ch}(e)}\\ &\quad \sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,e)\cdot cost(u,v)\ .\end{aligned} \tag{10}$$

The first two sums in the last expression can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time, given that we have computed already for every $l \in \mathrm{Ch}(e)$ the quantity $\mathrm{TC}(l)$ and the sum $\sum_{u\in S(l)} \mathrm{TC}(u)$ (can be done with a single bottom-up scan of the tree). The last sum in (10) can be expressed as:

$$\begin{aligned}&\sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,e)\cdot cost(u,v)\\ &\quad = \sum_{l\in \mathrm{Ch}(e)} w_l \sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,v)\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,l)\cdot cost(u,v)\\ &\quad = \sum_{l\in \mathrm{Ch}(e)} w_l \sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,v)\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost^2(u,l)\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)}\sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,l)\cdot cost(v,l)\ .\end{aligned} \tag{11}$$

The two last sums in (11) are identical with the quantities that we analysed in (6) and in (7). Finally, the first sum in (11) is equal to:

$$\begin{aligned}&\sum_{l\in \mathrm{Ch}(e)} w_l \sum_{u\in S(l)}\sum_{v\in S(e)\backslash S(l)} cost(u,v)\\ &\quad = \sum_{l\in \mathrm{Ch}(e)} w_l (s(e)-s(l))\cdot \mathrm{TC}_{\mathrm{sub}}(l)\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)} w_l^2\cdot s(l)\cdot (s(e)-s(l))\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)} w_l \left(s(l)\left(\sum_{k\in \mathrm{Ch}(e)} s(k)\cdot w_k\right) - s^2(l)\cdot w_l\right)\\ &\qquad + \sum_{l\in \mathrm{Ch}(e)} w_l\cdot s(l)\left(\left(\sum_{k\in \mathrm{Ch}(e)} \mathrm{TC}_{\mathrm{sub}}(k)\right) - \mathrm{TC}_{\mathrm{sub}}(l)\right),\end{aligned} \tag{12}$$

which can also be computed in $\Theta(|\mathrm{Ch}(e)|)$ time.

All the sums that we analysed from (4) up to (12) can be computed in $\Theta(|\mathrm{Ch}(e)|)$ time for every edge e in the tree. From this we conclude that for every edge $e \in E$ we can evaluate $\mathrm{QD}(e)$ in (4) in $\Theta(|\mathrm{Ch}(e)|)$ time from the respective values of the edges in $\mathrm{Ch}(e)$. Since $\sum_{e\in E} |\mathrm{Ch}(e)| = \Theta(|E|)$, we prove that we can compute $\mathrm{QD}(e)$ for all the edges in $\mathcal{T}$ in $\Theta(n)$. □

Computing the skewness of the MPD

In the previous section we defined the problem of computing the skewness of the MPD for a given phylogenetic tree $\mathcal{T}$. Given a positive integer $r \leq s$, we showed that to solve this problem efficiently it remains to find an efficient algorithm for computing $E_{R\in \mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$; this is the mean value of the cube of the MPD among all possible subsets of tips in $\mathcal{T}$ that consist of exactly r elements. To compute this efficiently, we introduced in Table 1 eleven different quantities which we want to use in order to express this mean value. In Lemma 1 we proved that these quantities can be computed in $O(n)$ time, where n is the size of $\mathcal{T}$.

Next we prove how we can calculate the value for the mean of the cube of the MPD based on the quantities in Table 1. In particular, in the proof of the following lemma we show how the value $E_{R\in \mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$ can be written analytically as an expression that contains the quantities in Table 1. This expression can then be straightforwardly evaluated in $O(n)$ time, given that we have already computed the aforementioned quantities. Because the full form of this expression is very long (it consists of a large number of terms), we have chosen not to include it in the definition of the following lemma. We chose to do so because we considered that including the entire expression would not make this work more readable. In any case, the full expression can be easily infered from the proof of the lemma.

Lemma 2. *For any given natural $r \leq s$, we can compute $E_{R \in Sub(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$ in $\Theta(n)$ time.*

Proof. The expectation of the cube of the MPD is equal to:

$$E_{R\in \mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$$

$$= \frac{8}{r^3(r-1)^3} \cdot E_{R\in \mathrm{Sub}(S,r)} \left[\sum_{\{u,v\}\in\Delta(R)} \sum_{\{x,y\}\in\Delta(R)} \sum_{\{c,d\}\in\Delta(R)} cost(u,v) \cdot cost(x,y) \cdot cost(c,d) \right].$$

From the last expression we get:

$$E_{R\in \mathrm{Sub}(S,r)} \left[\sum_{\{u,v\}\in\Delta(R)} \sum_{\{x,y\}\in\Delta(R)} \sum_{\{c,d\}\in\Delta(R)} cost(u,v) \cdot cost(x,y) \cdot cost(c,d) \right]$$

$$= \sum_{\{u,v\}\in\Delta(S)} \sum_{\{x,y\}\in\Delta(S)} \sum_{\{c,d\}\in\Delta(S)} cost(u,v) \cdot cost(x,y) \cdot cost(c,d) \cdot E_{R\in \mathrm{Sub}(S,r)}[\,AP_R(u,v,x,y,c,d)]\,, \qquad (13)$$

where $AP_R(u,v,x,y,c,d)$ is a random variable whose value is equal to one in the case that $u,v,x,y,c,d \in R$, otherwise it is equal to zero. For any six tips $u,v,x,y,c,d \in S$, which may not be all of them distinct, we use $\theta(u,v,x,y,c,d)$ to denote the number of distinct elements among these tips. Let t be an integer, and let $(t)_k$ denote the k-th falling factorial power of t, which means that $(t)_k = t(t-1)\ldots(t-k+1)$. For the expectation of the random variables that appear in the last expression it holds that:

$$E_{R\in \mathrm{Sub}(S,r)}\left[AP_R(u,v,x,y,c,d)\right] = \frac{(r)_{\theta(u,v,x,y,c,d)}}{(s)_{\theta(u,v,x,y,c,d)}} \qquad (14)$$

Notice that in (14) we have $2 \leq \theta(u,v,x,y,c,d) \leq 6$. The value of the function $\theta(\cdot)$ cannot be smaller than two in the above case because we have that $u \neq v$, $x \neq y$, and $c \neq d$. Thus, we can rewrite (13) as:

$$\sum_{\{u,v\}\in\Delta(S)} \sum_{\{x,y\}\in\Delta(S)} \sum_{\{c,d\}\in\Delta(S)} \frac{(r)_{\theta(u,v,x,y,c,d)}}{(s)_{\theta(u,v,x,y,c,d)}} \cdot cost(u,v) \cdot cost(x,y) \cdot cost(c,d) \qquad (15)$$

Hence, our goal now is to compute a sum whose elements are the product of costs of triples of paths. Recall that for each of these paths, the end-nodes of the path are a pair of distinct tips in the tree. Although the end-nodes of each path are distinct, in a given triple the paths may share one or more end-nodes with each other. Therefore, the distinct tips in any triple of paths may vary from two up to six tips. Indeed, in (15) we get a sum where the triples of paths in the sum are partitioned in five groups; a triple of paths is assigned to a group depending on the number of distinct tips in this triple. In (15) the sum for each group of triples is multiplied by the same factor $(r)_{\theta(u,v,x,y,c,d)}/(s)_{\theta(u,v,x,y,c,d)}$, hence we have to calculate the sum for each group of triples separately.

However, when we try to calculate the sum for each of these groups of triples we see that this calculation is more involved; some of these groups of triples are divided into smaller subgroups, depending on which end-nodes of the paths in each triple are the same. To explain this better, we can represent a triple of paths schematically as a graph; let $\{u,v\}, \{x,y\}, \{c,d\} \in \Delta(S)$ be three pairs of tips in $\mathcal{T}$. As mentioned already, the tips within each pair are distinct, but tips between different pairs can be the same.

We represent the similarity between tips of these three pairs as a graph of six vertices. Each vertex in the graph corresponds to a tip of these three pairs. Also, there exists an edge in this graph between two vertices if the corresponding tips are the same. Thus, this graph is tripartite; no vertices that correspond to tips of the same pair can be connected to each other with an edge. Hence, we have a tripartite graph where each partite set of vertices consists of two vertices–see Figure 1 for an example.

For any triple of pairs of tips $\{u,v\}$, $\{x,y\}$, $\{c,d\} \in \Delta(S)$ we denote the tripartite graph that corresponds to this triple by $G[u,v,x,y,c,d]$. We call this graph the *similarity* graph of this triple. Based on the way that similarities may occur between tips in a triple of paths, we can partition the five groups of triples in (15) into smaller subgroups. Each of these subgroups contains triples whose similarity graphs are isomorphic. For a tripartite graph that consists of three partite sets of two vertices each, there can be eight different isomorphism classes. Therefore, the five

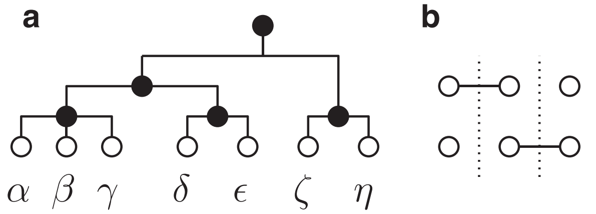

Figure 1 Representing triples of paths as graphs. (a) A phylogenetic tree $\mathcal{T}$ and **(b)** an example of the tripartite graph induced by the triplet of its tip pairs $\{\alpha,\gamma\}$, $\{\delta,\gamma\}$, $\{\epsilon,\delta\}$, where $\{\alpha,\gamma,\delta,\epsilon\} \subset S$. The dashed lines in the graph distinguish the partite subsets of vertices; the vertices of each partite subset correspond to tips of the same pair.

groups of triples in (15) are partitioned into eight subgroups. Figure 2 illustrates the eight isomorphism classes that exist for the specific kind of tripartite graphs that we consider. Since we refer to isomorphism classes, each of the graphs in Figure 2 represents the combinatorial structure of the similarities between three pairs of tips, and it does not correspond to a particular planar embedding, or ordering of the tips.

Let X be any isomorphism class that is illustrated in Figure 2. We denote the set of all triples of pairs in $\Delta(S)$ whose similarity graphs belong to this class by $\mathcal{B}_X$. More formally, the set $\mathcal{B}_X$ can be defined as follows :

$$\mathcal{B}_X = \{\{\{u,v\},\{x,y\},\{c,d\}\} : \{u,v\},\{x,y\},\{c,d\} \in \Delta(S)$$
$$\text{and } G[\,u,v,x,y,c,d]\ \text{ belongs to class } X \text{ in Figure 2}\}\,.$$

We introduce also the following quantity:

$$\mathrm{TRS}(X) = \sum_{\{\{u,v\},\{x,y\},\{c,d\}\}\in\mathcal{B}_X} cost(u,v)\cdot cost(x,y)\cdot cost(c,d)\,.$$

Hence, we can rewrite (15) as follows:

$$\frac{(r)_2}{(s)_2}\cdot\mathrm{TRS}(A) + 3\cdot\frac{(r)_3}{(s)_3}\cdot\mathrm{TRS}(B) + 6\cdot\frac{(r)_3}{(s)_3}\cdot\mathrm{TRS}(C)$$
$$+6\cdot\frac{(r)_4}{(s)_4}\cdot\mathrm{TRS}(D) + 3\cdot\frac{(r)_4}{(s)_4}\cdot\mathrm{TRS}(E) + 6\cdot\frac{(r)_4}{(s)_4}\cdot\mathrm{TRS}(F)$$
$$+6\cdot\frac{(r)_5}{(s)_5}\cdot\mathrm{TRS}(G) + 6\cdot\frac{(r)_6}{(s)_6}\cdot\mathrm{TRS}(H) \tag{16}$$

Notice that some of the terms $\frac{(r)_i}{(s)_i}\cdot\mathrm{TRS}(X)$ in (16) are multiplied with an extra constant factor. This happens for the following reason; the sum in $\mathrm{TRS}(X)$ counts each triple once for every different combination of three pairs of tips. However, in the triple sum in (15) some triples appear more than once. For example, every triple that belongs in class B appears three times in (15), hence there is an extra factor three in front of $\mathrm{TRS}(B)$ in (16).

To compute efficiently $E_{R\in\mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$, it remains to compute efficiently each value $\mathrm{TRS}(X)$ for every isomorphism class X that is presented in Figure 2. Next we show in detail how we can do that by expressing each quantity $\mathrm{TRS}(X)$ as a function of the quantities that appear in Table 1.

For the triples that correspond to the isomorphism class A we have:

$$\mathrm{TRS}(A) = \sum_{\{u,v\}\in\Delta(S)} cost^3(u,v) = \mathrm{CB}(\mathcal{T})\,.$$

For $\mathrm{TRS}(B)$ we get:

$$\mathrm{TRS}(B) = \sum_{\{u,v\}\in\Delta(S)} cost^2(u,v)\left(\sum_{x\in S\setminus\{u\}} cost(u,x)\right.$$
$$\left. + \sum_{y\in S\setminus\{v\}} cost(v,y) - 2\cdot cost(u,v)\right)$$
$$= \sum_{\{u,v\}\in\Delta(S)} cost^2(u,v)(\mathrm{TC}(u)+\mathrm{TC}(v)-2\cdot cost(u,v))$$
$$= \sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u) - 2\cdot\mathrm{CB}(\mathcal{T})\,.$$

The quantity $\mathrm{TRS}(C)$ is equal to:

$$\frac{1}{6}\sum_{u\in S}\sum_{v\in S\setminus\{u\}} cost(u,v)\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$
$$= \sum_{e\in E} w_e \sum_{u\in S(e)}\sum_{v\in S-S(e)}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)\,. \tag{17}$$

For any $e\in E$ we have that:

$$\sum_{u\in S(e)}\sum_{v\in S-S(e)}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$
$$= \sum_{u\in S(e)}\sum_{v\in S\setminus\{u\}}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$
$$-2\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)\,. \tag{18}$$

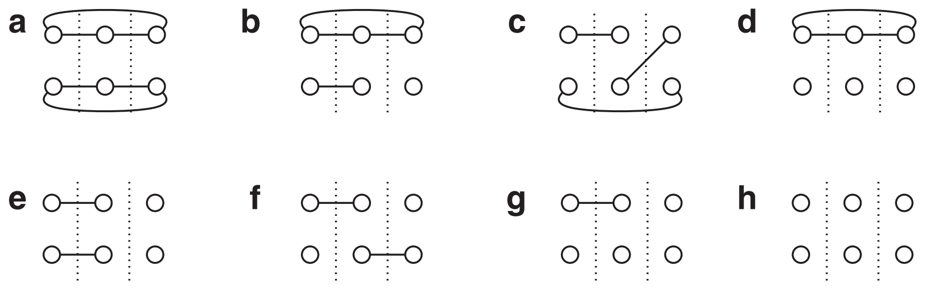

Figure 2 Isomorphism classes. The eight isomorphism classes of a tripartite graph of 3 × 2 vertices that represent schematically the eight possible cases of similarities between tips that we can have when we consider three paths between pairs of tips in a tree $\mathcal{T}$.

The first of the two sums in (18) can be written as:

$$\sum_{u\in S(e)}\sum_{v\in S\setminus\{u\}}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$

$$= \sum_{u\in S(e)}\sum_{v\in S\setminus\{u\}}\sum_{x\in S\setminus\{u,v\}} cost(u,v)\cdot cost(x,v)$$

$$= \sum_{u\in S(e)}\sum_{v\in S\setminus\{u\}} \left(cost(u,v)\cdot \mathrm{TC}(v) - cost^2(u,v)\right)$$

$$= \sum_{u\in S(e)} \mathrm{SM}(u) - \mathrm{SQ}(u)\,. \tag{19}$$

According to Lemma 2, we can compute $\mathrm{SM}(u)$ and $\mathrm{SQ}(u)$ for all tips $u \in S$ in linear time with respect to the size of $\mathcal{T}$. Given these values, we can compute $\sum_{u\in S(e)} \mathrm{SM}(u) - \mathrm{SQ}(u)$ for every edge $e \in E$ in $\mathcal{T}$ with a single bottom-up scan of the tree. For any edge e in E, the second sum in (18) is equal to:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$

$$= \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S(e)\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$

$$+ \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(u,x)\cdot cost(x,v)\,. \tag{20}$$

We can express the first sum in (20) as:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S(e)\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)$$

$$= \frac{1}{2}\sum_{u\in S(e)}\left(\sum_{v\in S(e)\setminus\{u\}} cost(u,v)\right)^2$$

$$- \frac{1}{2}\sum_{u\in S(e)}\sum_{v\in S(e)\setminus\{u\}} cost^2(u,v)$$

$$= \frac{1}{2}\mathrm{QD}(e) - \frac{1}{2}\sum_{u\in S(e)}\sum_{v\in S(e)\setminus\{u\}} cost^2(u,v)\,. \tag{21}$$

The last sum in (21) is equal to:

$$\sum_{u\in S(e)}\sum_{v\in S(e)\setminus\{u\}} cost^2(u,v) = \left(\sum_{u\in S(e)} \mathrm{SQ}(u)\right) - \mathrm{SQ}(e)\,. \tag{22}$$

The value of the sum $\sum_{u\in S(e)} \mathrm{SQ}(u)$ can be computed for every edge e in $\Theta(n)$ time in total as follows; for every tip $u \in S$ we store $\mathrm{SQ}(u)$ together with this tip, and then scan bottom-up the tree adding those values that are in the subtree of each edge. For the remaining part of (20) we get:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(u,x)\cdot cost(x,v)$$

$$= \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} \left(cost(u,e) + cost(x,e)\right)$$

$$\times \left(cost(v,e) + cost(x,e)\right)$$

$$= \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(u,e)\cdot cost(v,e)$$

$$+ \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(x,e)\cdot\left(cost(u,e)+cost(v,e)\right)$$

$$+ \sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost^2(x,e)\,. \tag{23}$$

The first sum in (23) is equal to:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(u,e)\cdot cost(v,e)$$

$$= \frac{1}{2}\cdot(s - s(e))\left(\mathrm{TC_{sub}}^2(e) - \mathrm{SQ_{sub}}(e)\right)\,. \tag{24}$$

For the second sum in (23) we have:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost(x,e)\cdot\left(cost(u,e) + cost(v,e)\right)$$

$$= (s(e) - 1)\cdot \mathrm{TC_{sub}}(e)\sum_{x\in S\setminus S(e)} cost(x,e)$$

$$= (s(e) - 1)\cdot \mathrm{TC_{sub}}(e)\cdot\left(\mathrm{PC}(e) - \mathrm{TC_{sub}}(e)\right)\,. \tag{25}$$

The last sum in (23) can be written as:

$$\sum_{\{u,v\}\in\Delta(S(e))}\sum_{x\in S\setminus S(e)} cost^2(x,e)$$

$$= \frac{s(e)(s(e) - 1)}{2}\left(\mathrm{PSQ}(e) - \mathrm{SQ_{sub}}(e)\right)\,. \tag{26}$$

Combining the analyses that we did from (17) up to (26) we get:

$$\begin{aligned}\mathrm{TRS}(C) = \sum_{e\in E} w_e \Bigg(&\sum_{u\in S(e)} \mathrm{SM}(u) - \mathrm{QD}(e) - \mathrm{SQ}(e)\\ &- (s - s(e))\left(\mathrm{TC_{sub}}^2(e) - \mathrm{SQ_{sub}}(e)\right)\\ &- 2(s(e) - 1)\cdot \mathrm{TC_{sub}}(e)\cdot\left(\mathrm{PC}(e) - \mathrm{TC_{sub}}(e)\right)\\ &- s(e)(s(e) - 1)\cdot\left(\mathrm{PSQ}(e) - \mathrm{SQ_{sub}}(e)\right)\Bigg)\,.\end{aligned}$$

The value of TRS(D) can be expressed as:

$$\begin{aligned}
&\sum_{u\in S}\ \sum_{\substack{v,x,y\in S\setminus\{u\}\\ v,x,y\text{are distinct}}} cost(u,v)\cdot cost(u,x)\cdot cost(u,y)\\
&= \frac{1}{6}\left(\sum_{u\in S}\mathrm{TC}^3(u) - 2\cdot\mathrm{TRS}(A) - 3\cdot\mathrm{TRS}(B)\right)\\
&= \frac{1}{6}\cdot\sum_{u\in S}\mathrm{TC}^3(u) + \frac{2}{3}\cdot\mathrm{CB}(\mathcal{T}) - \frac{1}{2}\cdot\sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u)\,.
\end{aligned}$$

For TRS(E) we get:

$$\begin{aligned}
&\sum_{\{u,v\}\in\Delta(S)}\ \sum_{\{x,y\}\in\Delta(S\setminus\{u,v\})} cost^2(u,v)\cdot cost(x,y)\\
&= \sum_{\{u,v\}\in\Delta(S)} cost^2(u,v)(\mathrm{TC}(\mathcal{T})-\mathrm{TC}(u)-\mathrm{TC}(v)+cost(u,v))\\
&= \mathrm{TC}(\mathcal{T})\sum_{e\in E} w_e\cdot\mathrm{TC}(e) - \sum_{u\in S}(\mathrm{SQ}(u)\cdot\mathrm{TC}(u)) + \mathrm{CB}(\mathcal{T})\,.
\end{aligned}$$

We can rewrite TRS(F) as follows:

$$\begin{aligned}
&\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\\
&\left(\mathrm{TC}(u)\cdot\mathrm{TC}(v) - cost^2(u,v) - \sum_{x\in S\setminus\{u,v\}} cost(u,x)\cdot cost(x,v)\right)\\
&= \sum_{\{u,v\}\in\Delta(S)} cost(u,v)\cdot\mathrm{TC}(u)\cdot\mathrm{TC}(v) - \mathrm{CB}(\mathcal{T}) - 3\cdot\mathrm{TRS}(C)\\
&= \sum_{e\in E} w_e\cdot\mathrm{Mult}(e) - \mathrm{CB}(\mathcal{T}) - 3\cdot\mathrm{TRS}(C)\,.
\end{aligned}$$

For the value of TRS(G) we have:

$$\begin{aligned}
\mathrm{TRS}(G) = \frac{1}{2}&\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\sum_{x\in S\setminus\{u,v\}}(cost(u,x)\\
&+ cost(v,x))\,(\mathrm{TC}(\mathcal{T}) - \mathrm{TC}(u) - \mathrm{TC}(v)\\
&- \mathrm{TC}(x) + cost(u,v) + cost(u,x) + cost(v,x))\,.
\end{aligned} \tag{27}$$

We now break the sum in (27) into five pieces and express each piece of this sum in terms of the quantities in Table 1. The first piece of the sum is equal to:

$$\begin{aligned}
&\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\sum_{x\in S\setminus\{u,v\}}(cost(u,x)+cost(v,x))\cdot\mathrm{TC}(\mathcal{T})\\
&= \frac{1}{2}\cdot\mathrm{TC}(\mathcal{T})\left(\sum_{u\in S}\mathrm{TC}^2(u) - 2\cdot\sum_{\{u,v\}\in\Delta(S)} cost^2(u,v)\right)\\
&= \frac{1}{2}\cdot\mathrm{TC}(\mathcal{T})\left(\sum_{u\in S}\mathrm{TC}^2(u) - 2\cdot\sum_{e\in E} w_e\cdot\mathrm{TC}(e)\right)\,.
\end{aligned}$$

The second piece that we take from the sum in (27) can be expressed as:

$$\begin{aligned}
&-\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\sum_{x\in S\setminus\{u,v\}}(cost(u,x)\\
&+\ cost(v,x))\,(\mathrm{TC}(u)+\mathrm{TC}(v))\\
&= -\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\,(\mathrm{TC}(u)+\mathrm{TC}(v)\\
&-\ 2\cdot cost(u,v))\,(\mathrm{TC}(u)+\mathrm{TC}(v))\\
&= -\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\,\big(\mathrm{TC}^2(u)\\
&+ \mathrm{TC}^2(v) + 2\cdot\mathrm{TC}(u)\cdot\mathrm{TC}(v)\\
&-\ 2\cdot cost(u,v)\cdot(\mathrm{TC}(u)+\mathrm{TC}(v)))\\
&= -\frac{1}{2}\sum_{u\in S}\mathrm{TC}^3(u) - \sum_{\{v,x\}\in\Delta(S)} cost(v,x)\cdot\mathrm{TC}(v)\cdot\mathrm{TC}(x)\\
&+\ \sum_{\{y,z\}\in\Delta(S)} cost^2(y,z)\,(\mathrm{TC}(y)+\mathrm{TC}(z))\\
&= -\frac{1}{2}\sum_{u\in S}\mathrm{TC}^3(u) - \sum_{e\in E} w_e\cdot\mathrm{Mult}(e)\\
&+ \sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u)\,.
\end{aligned} \tag{28}$$

The next piece that we select from (27) is equal to:

$$\begin{aligned}
&-\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)\sum_{x\in S\setminus\{u,v\}}(cost(u,x)\\
&+ cost(v,x))\cdot\mathrm{TC}(x)\\
&= -\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)(\mathrm{SM}(u)\\
&+ \mathrm{SM}(v) - cost(u,v)\cdot\mathrm{TC}(u) - cost(u,v)\cdot\mathrm{TC}(v))\\
&= -\frac{1}{2}\sum_{u\in S}\mathrm{SM}(u)\cdot\mathrm{TC}(u)\\
&+ \frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost^2(u,v)\,(\mathrm{TC}(u)+\mathrm{TC}(v))\\
&= -\frac{1}{2}\sum_{u\in S}\mathrm{SM}(u)\cdot\mathrm{TC}(u) + \frac{1}{2}\sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u)\,.
\end{aligned} \tag{29}$$

For the fourth piece of the sum in (27) we get:

$$\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost^2(u,v) \sum_{x\in S\setminus\{u,v\}} (cost(u,x)+cost(v,x))$$
$$= \frac{1}{2}\cdot \mathrm{TRS}(B) = \frac{1}{2}\sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u) - \mathrm{CB}(\mathcal{T})\,. \tag{30}$$

The last piece of the sum in (27) can be expressed as:

$$\frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v) \sum_{x\in S\setminus\{u,v\}} (cost(u,x)+cost(v,x))^2$$
$$= \frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v) \sum_{x\in S\setminus\{u,v\}} (cost^2(u,x)$$
$$+ cost^2(v,x)) + 3\cdot\mathrm{TRS}(C)$$
$$= \frac{1}{2}\sum_{\{u,v\}\in\Delta(S)} cost(u,v)(\mathrm{SQ}(u)+\mathrm{SQ}(v)$$
$$- 2\cdot cost^2(u,v)) + 3\cdot\mathrm{TRS}(C)$$
$$= \frac{1}{2}\sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u) - \mathrm{CB}(\mathcal{T}) + 3\cdot\mathrm{TRS}(C)\,. \tag{31}$$

Combining our analyses from (27) up to (31) we get:

$$\mathrm{TRS}(G) = \frac{1}{2}\cdot\mathrm{TC}(\mathcal{T})\cdot\sum_{u\in S}\mathrm{TC}^2(u)$$
$$- \mathrm{TC}(\mathcal{T})\cdot\sum_{e\in E} w_e\cdot\mathrm{TC}(e) - \frac{1}{2}\cdot\sum_{u\in S}\mathrm{TC}^3(u)$$
$$- \sum_{e\in E} w_e\cdot\mathrm{Mult}(e) - \frac{1}{2}\cdot\sum_{u\in S}\mathrm{SM}(u)\cdot\mathrm{TC}(u)$$
$$+ \frac{5}{2}\cdot\sum_{u\in S}\mathrm{SQ}(u)\cdot\mathrm{TC}(u)$$
$$- 2\cdot\mathrm{CB}(T) + 3\cdot\mathrm{TRS}(C)\,.$$

We can express TRS(H) using the values of the other isomorphism classes:

$$\mathrm{TRS}(H) = \frac{1}{6}\cdot\Bigg(\sum_{\{u,v\}\in\Delta(S)}\ \sum_{\{x,y\}\in\Delta(S)}\ \sum_{\{c,d\}\in\Delta(S)} cost(u,v)$$
$$\cdot\, cost(x,y)\cdot cost(c,d) - \mathrm{TRS}(A)$$
$$- 3\cdot\mathrm{TRS}(B) - 6\cdot\mathrm{TRS}(C) - 6\cdot\mathrm{TRS}(D)$$
$$- 3\cdot\mathrm{TRS}(E) - 6\cdot\mathrm{TRS}(F) - 6\cdot\mathrm{TRS}(G)\Bigg)$$
$$= \frac{1}{6}\cdot\mathrm{TC}^3(\mathcal{T}) - \frac{1}{6}\cdot\mathrm{TRS}(A) - \frac{1}{2}\cdot\mathrm{TRS}(B)$$
$$- \mathrm{TRS}(C) - \mathrm{TRS}(D) - \frac{1}{2}\cdot\mathrm{TRS}(E)$$
$$- \mathrm{TRS}(F) - \mathrm{TRS}(G)\,.$$

We get the value of $E_{R\in\mathrm{Sub}(S,r)}[\,\mathrm{MPD}^3(\mathcal{T},R)]$ by plugging into (16) the values that we got for all eight isomorphism classes of triples. For any isomorphism class X we showed that the value TRS(X) can be computed by using the quantities in Table 1. The lemma follows from the fact that each quantity that appears in this table is used a constant number of times for computing value TRS(X) for any class X, and since we showed that we can precompute all these quantities in $\Theta(n)$ time in total. □

Theorem 3. *Let $\mathcal{T}$ be a phylogenetic tree that contains s tips, and let r be a natural number with $r \leq s$. The skewness of the mean pairwise distance on $\mathcal{T}$ among all subsets of exactly r tips of $\mathcal{T}$ can be computed in $\Theta(n)$ time.*

Table 2 The sizes of tip samples that we considered for our experiments, together with the number of sets that we sampled for each tip size in order to derive the "true" values

Size of each tip sample	Number of sampled sets
10	10^5
20	10^5
40	$5\cdot 10^4$
80	$3\cdot 10^4$
160	$2\cdot 10^4$
320	10^4

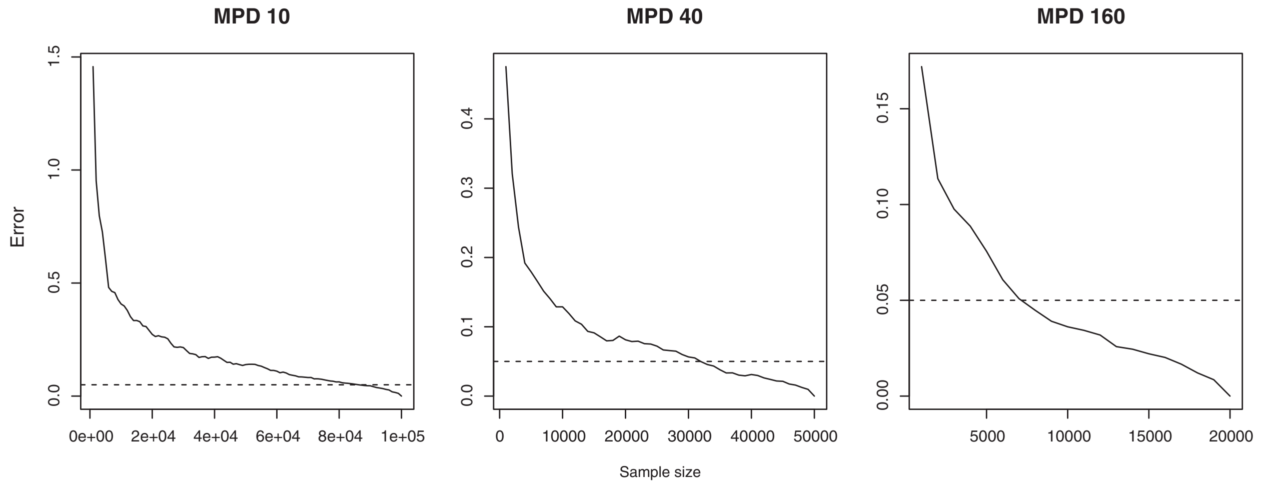

Figure 3 Error in calculating the distribution that is used as point of reference. Error in *P* value estimation as the number of tree set resamples increases of tip set sizes of 10, 40 and 160. The dotted lines show errors of 0.05 MPD units, illustrating that the number of resamplings used here were sufficient to estimate percentiles to within 0.05 distance units in each case.

Proof. According to the definition of skewness, as it is also presented in (1), we need to prove that we can compute in $\Theta(n)$ time the expectation and the variance of the MPD, and the value of the expression $E_{R\in\text{Sub}(S,r)}[\,\text{MPD}^3(\mathcal{T},R)]$. In a previous paper we showed that the expectation and the variance of the MPD can be computed in $\Theta(n)$ time. By combining this with Lemma 2 we get the proof of the theorem. □

Experiments: improved *P* value estimation incorporating skewness

Earlier in this paper, we mentioned that distributions of MPD values are often found to be skewed, suggesting that it is necessary to incorporate this skewness into analytical *P* value estimation. However, it is unclear whether good *P* value estimates are possible with only the first three moments of the distribution, or if more detailed distributional information is required.

We investigate this question here by considering random phylogenetic trees produced by a pure birth process [12], though results were qualitatively identical when using trees generated by a combined birth-death process (and skewness did not vary as a function of the death rate). We took two approaches for estimating the position of the 2.5 and 97.5 percentile of MPD distribution given a particular tree instance. For any tree $\mathcal{T}$ that we

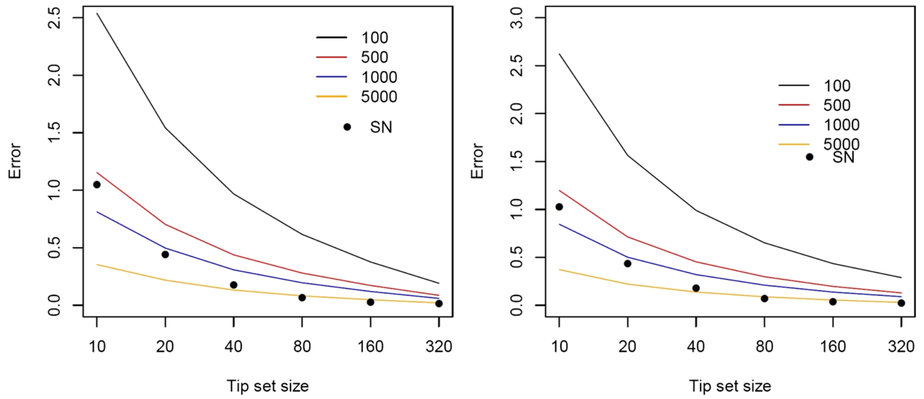

Figure 4 Comparison of approximation methods. Errors in *P* value approximation using different resampling replicates (indicated by the coloured lines), compared to that obtained by assuming a skew-normal distribution of MPD values (indicated as SN). Errors were strongly influenced by tip set size *r*, and weakly by tree size; on the left side appear the results for a 500 tip tree, and on the right for a 2000 tip tree). In most cases, *P* value approximation based on the skew-normal distribution performed better than the most commonly-used standard of 1000 set resamplings (blue line), and the relative performance of the skew-normal approach improved with increasing tip set size.

constructed, we first calculated the distribution of the MPD values using as a point of reference extensive sampling of sets of tips (much more extensive than is usually employed in practice). In particular, for specific values of r we sampled from $\mathcal{T}$ a large number of sets that consist of exactly r tips (see Table 2 for the values of r and numbers of sets that we sampled). We simply calculated the percentiles of these distributions, and call these the *reference* values, recognizing that they neveretheless contain some error, being incomplete samples from the tree. Complete sampling from large trees is computationally infeasible, but we estimate that the error in the calculated percentiles was less than 0.05 distance units in all cases (corresponding to an error of approximately 0.01% relative to the mean MPD–see Figure 3).

The two approaches that we used to estimate the percentile positions reflect two alternatives that might be employed by practising researchers. In the first approach, for each value r that we considered, we sampled again several sets of tips, yet much fewer than the ones we used to calculate the reference values (100, 500, 1000 or 5000 sets). We then compare the absolute difference between percentiles estimated in this manner and the reference values. We refer to this difference as the *error* between the estimated percentile values and the reference values. The second approach uses the mean, variance and skewness of the MPD distribution to determine the position of the 2.5 and 97.5 percentile of the skew-normal distribution with these moments [13]. The mean, variance and skewness were computed in this case based on all the MPD values that we used to calculate reference percentiles. Although we have implemented algorithms for computing the exact values of the mean and variance of the MPD, we have not implemented so far the algorithm that computes the skewness of the MPD; that is the algorithm outlined in the previous sections of this paper. As with the previous approach, the error of this approximation method was calculated by taking the absolute difference between each estimated percentile position and the corresponding reference value.

The experiment described above was repeated across 100 replicate trees of each of two sizes (500 and 2000 tips), and across a range of tip set sizes (10, 20, 40, 80, 160 and 320). Errors were weakly related to tree size but decreased strongly with tip set size–see Figure 4. This decrease was more pronounced for estimates based on skew-normal approximation than resampling. Notably, the skew-normal approximation yielded smaller errors than the most commonly used standard of 1000 resamplings for all but the smallest tip set sizes.

Thus, we conclude that the errors introduced by assuming a skew-normal distribution of MPD values appear to be comparable to or smaller than those introduced by standard resampling procedures, while also showing better scaling with increased tip sample size. Finally, the computation of P values using skew-normal approximation is typically faster than with resampling, particularly in cases involving large samples of trees.

Conclusions

Given a rooted tree $\mathcal{T}$ and a non-negative integer r, we proved that we can compute the skewness of the MPD among all subsets of r leaves in $\mathcal{T}$ in $O(n)$ time. An interesting problem for future research would be to implement the algorithm that is outlined by our proof, and show its efficiency in practice. Also, it would be interesting to derive a similar result for the so-called *Community Distance* measure; this is the equivalent of the MPD when distances between two sets of species are considered [11].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Both authors have contributed both in developing the research results presented in this paper, as well as in the writing process. Both authors read and approved the final manuscript.

Acknowledgements

We want to thank Dr. Erick Matsen for his insightful comments on the paper and the productive discussion that we had regarding the motivation for deriving an algorithm for the MPD skewness.

Received: 9 December 2013 Accepted: 27 May 2014
Published: 14 June 2014

References

1. Webb CO, Ackerly DD, McPeek MA, Donoghue MJ: **Phylogenies and community ecology.** *Annu Rev Ecol Systemat* 2002, **33:**475–505.
2. Tsirogiannis C, Sandel B, Cheliotis D: **Efficient computation of popular phylogenetic tree measures.** In *Proceedings of 12th International Workshop on Algorithmms in Bioinformatics (WABI)*. Edited by Raphael B, Tang J: Springer-Verlag Berlin Heidelberg; 2012:30–43.
3. Pontarp M, Canbäck B, Tunlid A, Lunberg P: **Phylogenetic analysis suggests that habitat filtering is structuring marine bacterial communities across the globe.** *Microb Ecol* 2012, **64:**8–17.
4. Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO: **The global diversity of birds in space and time.** *Nature* 2012, **491:**444–448.
5. Barnagaud J-Y, Daniel Kissling W, Sandel B, Eiserhardt WL, Şekercioğlu CH, Enquist BJ, Tsirogiannis C, Svenning J-C: **Ecological traits influence the phylogenetic structure of bird species co-occurrences worldwide.** *Ecol Lett* 2014. To appear.
6. Cooper N, Rodríguez J, Purvis A: **A common tendency for phylogenetic overdispersion in mammalian assemblages.** *Proc Biol Sci* 2008, **275:**2031–2037.
7. Vamosi JC, Vamosi SM: **Body size, rarity, and phylogenetic community structure: Insights from diving beetle assemblages of alberta.** *Divers Distributions* 2007, **13:**1–10.
8. Harmon-Threat AN, Ackerly DD: **Filtering across spatial scales: phylogeny, biogeography and community structure in bumble bees.** *PLoS ONE* 2013, **8:**60446.
9. Lin TI, Lee JC, Yen SY: **Finite mixture modelling using the skew normal distribution.** *Statistica Sinica* 2007, **17:**209–227.
10. Lee SX, McLachlan GJ: **On mixtures of skew normal and skew t-distributions.** *Adv Data Anal Classif* 2013, **7:**241–266.
11. Swenson NG: **Phylogenetic beta diversity metrics. trait evolution and inferring the functional beta diversity of communities.** *PLoS ONE* 2011, **6**(6):21264.

12. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W: **Geiger: investigating evolutionary radiations.** *Bioinformatics* 2008, **24:**129–131.
13. Azzalini A: **The r 'sn' package: the skew-normal and skew-t distributions (version 1.0-0).** [http://cran.r-project.org/web/packages/sn/index.html] 2014. Accessed 2014-05-7.

doi:10.1186/1748-7188-9-15
Cite this article as: Tsirogiannis and Sandel: Computing the skewness of the phylogenetic mean pairwise distance in linear time. *Algorithms for Molecular Biology* 2014 **9**:15.