COMPUTING THE DISTRIBUTION OF A TREE METRIC

DAVID BRYANT AND MIKE STEEL

Abstract. The Robinson-Foulds (RF) distance is by far the most widely used measure of dissimilarity between trees. Although the distribution of these distances has been investigated for twenty years, an algorithm that is explicitly polynomial time has yet to be described for computing this distribution (which is also the distribution of trees around a given tree under the popular Robinson-Foulds metric). In this paper we derive a polynomial-time algorithm for this distribution. We show how the distribution can be approximated by a Poisson distribution determined by the proportion of leaves that lie in 'cherries' of the given tree. We also describe how our results can be used to derive normalization constants that are required in a recently-proposed maximum likelihood approach to supertree construction.

1. Introduction

Tree comparison metrics are widely used in phylogenetics for comparing evolutionary trees [3, 9] and for performing statistical tests - for example, to test whether two trees are more 'significantly different' from each other than one might expect if one or both trees were randomly chosen [6, 7]. In order to address these statistical questions one needs to determine the distribution of the metric under some null model (see, for

1991 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. Biology and genetics, Discrete mathematics applications, Trees, Phylogenetics, Robinson-Foulds distance, Poisson approximation, normalization constant.

We thank the Alexander von Humboldt Foundation and the Marsden Fund (DB) for supporting this work.

example, [6, 7]). The *symmetric difference* or *Robinson-Foulds* metric is the most widely used measure of differences between phylogenetic trees, and its distribution is particularly attractive to study. In a landmark paper [4], the authors described this distribution of trees relative to a fixed reference tree via a system of generating functions. This allowed the authors to calculate the distribution explicitly for small trees and provided a tool for analytic results on this distribution in later work by others.

However, the approach described in [4] does not immediately appear to provide a polynomial-time algorithm for computing this distribution, and for larger trees their approach may be computationally prohibitive. In this paper, we describe how to calculate the distribution of the Robinson-Foulds metric relative to a fixed tree. We also show how the distribution can be approximated by a Poisson distribution whose parameter depends on just one aspect of tree shape - the number of 'cherries'.

Our investigation into the distribution of the metric has also been motivated by its relevance to a recent approach for 'supertree' construction that is based on maximum likelihood [10]. In particular, our algorithm allows the normalization constants in the likelihood calculations to be computed explicitly. We describe how these normalization constants depend weakly on aspects of the shape of the tree - for example, how many 'cherries' the tree has. We start by recalling some terminology.

1.1. **Terminology.** Let X be a finite set. A *phylogenetic tree* with leaf set X is a tree with its degree one vertices (leaves) labelled bijectively by elements of X and whose remaining vertices have degree at least three. We use $V(T)$ and $E(T)$ to denote the set of nodes (vertices) and edges of T. Let $\overset{\circ}{V}(T)$ denote the set of internal (non-leaf) nodes of T and let $\overset{\circ}{E}(T)$ be the set of edges in $E(T)$ that have both endpoints in $\overset{\circ}{V}(T)$, the internal edges.

A phylogenetic tree is *fully resolved* if every internal vertex has degree three. Following [4] we let $PT(n)$ denote the set of phylogenetic trees on the finite set $X = \{1, 2, \ldots, n\}$ and $BPT(n)$ the set of fully resolved ('binary') trees in $PT(n)$ (two trees in $BPT(6)$ are shown in Fig. 1). The number of trees in $BPT(n)$ is denoted $b(n)$ and is given by:

$$(1) \qquad b(n) = (2n-5)!! = \prod_{k=3}^{n}(2k-5) \qquad n \geq 3,$$

see [9]. For convenience, we let $\beta(m)$ denote the number of fully resolved trees with exactly m internal edges, so:

$$(2) \qquad \beta(m) = b(m+3) = \prod_{k=3}^{m+3}(2k-5) \qquad m \geq 0.$$

Every edge $e \in E(T)$ induces a bipartition or *split* of the leaf set X corresponding to the labels present in the two connected components remaining when the edge e is removed. Let $\pi(T, e)$ denote this bipartition, which we consider unordered. We let $c(T)$ denote the set of all bipartitions obtained by removing different edges of T. Hence $|c(T)| \leq 2n - 3$, the maximum number of edges in a phylogenetic tree, and $|c(T)| = 2n-3$ exactly when T is fully resolved. A bipartition is *trivial* if it separates a single element from all other elements; trivial bipartitions correspond to the edges in the tree that are *external*, meaning that they are incident with a leaf of the tree. A *cherry* of a fully resolved phylogenetic tree T is a pair of leaves that forms one half of a split of T (i.e. a pair of leaves whose incident edges contain a common vertex).

In Fig. 1 the pairs $(1, 2)$ and $(5, 6)$ form cherries in both trees, while the right-hand tree has an additional cherry $(3, 4)$.

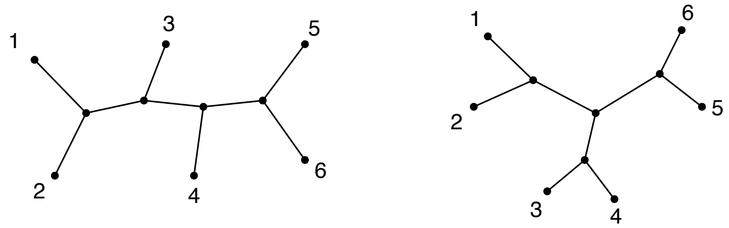

Figure 1. Two fully resolved phylogenetic trees on six leaves, with Robinson-Foulds distance two.

The *symmetric difference metric* is defined on $PT(n)$, and hence on $BPT(n)$, by:

$$d(T_1, T_2) = |c(T_1) \triangle c(T_2)|.$$

Note that this number is always even when T_1 and T_2 are both in $BPT(n)$, since, for any two trees in $PT(n)$, we have $d(T_1, T_2) = |c(T_1)| + |c(T_2)| - 2|c(T_1) \cap c(T_2)|$, and if $T_1, T_2 \in BPT(n)$ then $|c(T_1)| = |c(T_2)| = 2n - 3$. As an example, the two trees shown in Fig. 1 have a distance value of 2 since the splits $\{1, 2, 3\}|\{4, 5, 6\}$ and $\{3, 4\}|\{1, 2, 5, 6\}$ each occur in just one tree.

The metric was introduced by Bourque [1] and generalised by Robinson and Foulds [8]. As all phylogenetic trees contain all trivial splits, the maximum possible distance between two trees is $2(n - 3)$, which is twice the maximum number of internal edges.

2. Computing the distribution of the Robinson-Foulds metric

For each $T \in PT(n)$, let $b_m(T)$ denote the number of trees $T' \in BPT(n)$ for which $d(T, T') = m$. As d is a metric, $b_0(T) = 1$. A recursive formula for the generating function of $b_m(T)$ is given in [4] and [12]. This formula can be described conveniently using generating functions. Let

$$B(T, x) := \sum_{m \geq 0} b_m(T) x^m$$

and for any interior edge e of T let T/e be the tree formed by contracting e, and let T_1, T_2 be the maximal subtrees of T with e as a pendant edge. Then from [4] we have:

$$B(T, x) = xB(T/e, x) + (1 - x^2)B(T_1, x)B(T_2, x).$$

As far as we could deduce, the recursion described by this generating function identity does not provide a polynomial time algorithm for computing the $b_m(T)$ values, due to an exponential explosion in the number of subcases.

Instead we use an alternative approach, applying results of [12]. Let $q_s(T)$ denote the number of trees in $BPT(n)$ that share exactly s internal splits with T. Then for all $m = 0, 2, 4, \ldots, 2(n - 3)$, we have:

$$b_m(T) = q_{n-3-m/2}(T). \tag{3}$$

Define the polynomial

$$q(T, x) = \sum_{s=0}^{n-3} q_s(T) x^s. \tag{4}$$

Let $E \subset \mathring{E}(T)$ denote a subset of the set of internal edges of T. The forest $T - E$ has exactly $|E| + 1$ components $F_1, F_2, \ldots, F_{|E|+1}$. We use $\mathring{E}(F_i)$ as a short-hand for the edges of $\mathring{E}(T)$ that are contained in F_i.

Define

$$N_E(T) = \prod_{i=1}^{|E|+1} \beta(|\mathring{E}(F_i)|) \tag{5}$$

Note that $N_E(T)$ equals the quantity $\langle \Phi(E) \rangle$ defined in [12] (here assuming that T is fully resolved) and also equals the number of fully resolved trees containing all those splits induced by edges in E.

For $s \geq 0$ define

$$r_s(T) = \sum_{\substack{E \subseteq \mathring{E}(T) \\ |E| = s}} N_E(T),$$

the sum of N_E over all subsets $E \subseteq \mathring{E}(T)$ of cardinality s. For example, $r_0(T)$ equals $\beta(|\mathring{E}(T)|) = \beta(n-3)$. It was shown in [12] that the generating function

$$R(T, x) = \sum_{s \geq 0} r_s(T) x^s$$

satisfies the identity

$$q(T, x) = R(T, x - 1). \tag{6}$$

In what follows we derive a formula to evaluate the coefficients $r_s(T)$ so that we can compute the coefficients $b_m(T)$ via (3) and (6).

As usual, the computation applies dynamic programming, requiring us to introduce definitions for the appropriately divided sub-problems. Let v_0 be the node adjacent to leaf n. Delete leaf n and make v_0 the root of the tree, so that now every internal node has exactly two children. For each internal node v let T_v denote the subtree of T containing v and all of its descendants. Given a subset $E \subseteq \mathring{E}(T_v)$, we define $N_E(T_v)$ as in (5), where $F_1, \ldots, F_{|E|+1}$ will now be components of $T_v - E$ instead of $T - E$. We let $\kappa(v, E)$ denote the number of edges in the component of $T_v - E$ containing v. For $s, k \geq 0$, we let $\mathcal{E}(v, s, k)$ denote the set of all subsets $E \subseteq \mathring{E}(T_v)$ such that $|E| = s$ and $\kappa(v, E) = k$. Define

$$R(v, s, k) = \sum_{E \in \mathcal{E}(v,s,k)} N_E(T_v) \tag{7}$$

so that if v_0 is the root of T and $s \geq 0$, we have:

$$r_s(T) = \sum_{k=0}^{s} R(v_0, s, k). \tag{8}$$

With these definitions in mind, and recalling the notation $\beta(m)$ from (2), we now derive a recursion for $R(v, s, k)$. As is customary, an empty summation equals zero.

Lemma 1. *Suppose that $v \in \mathring{V}(T)$. Then*

$$R(v, 0, k) = \begin{cases} \beta(k) & \textit{if } k = |\mathring{E}(T_v)|; \\ 0 & \textit{otherwise.} \end{cases} \tag{9}$$

Lemma 2. *Suppose that $s \geq 1$. For all $v \in \mathring{V}(T)$ let $n_v = |\mathring{E}(T_v)|$.*

(1) *If $k > n_v$ then $R(v,s,k) = 0$.*

(2) *If $v \in \mathring{V}(T)$ has no children in $\mathring{V}$ and $s \geq 1$ then $R(v,s,k) = 0$.*

(3) *If $v \in \mathring{V}(T)$ has one child v_1 in $\mathring{V}$ then*

$$R(v,s,k) = \begin{cases} \sum_{k_1 \geq 0} R(v_1, s-1, k_1) & \textit{if } k = 0; \\ R(v_1, s, k-1)(2k+1) & \textit{otherwise;} \end{cases} \tag{10}$$

(4) *If $v \in \mathring{V}(T)$ has two children v_1, v_2 in $\mathring{V}(T)$ then*

$$R(v,s,0) = \sum_{s_1=0}^{s-2} \left(\sum_{k_1 \geq 0} R(v_1, s_1, k_1) \right) \left(\sum_{k_2 \geq 0} R(v_2, s-2-s_1, k_2) \right). \tag{11}$$

(5) *If $v \in \mathring{V}(T)$ has two children v_1, v_2 in $\mathring{V}(T)$ and $k \geq 1$ then*

$$\begin{aligned} R(v,s,k) \;&=\; \sum_{s_1=0}^{s-1} \left(\sum_{k_1 \geq 0} R(v_1, s_1, k_1) \right) R(v_2, s{-}1{-}s_1, k-1)\beta(k)/\beta(k-1) \\ &\quad + \sum_{s_2=0}^{s-1} \left(\sum_{k_2 \geq 0} R(v_2, s_2, k_2) \right) R(v_1, s{-}1{-}s_2, k-1)\beta(k)/\beta(k-1) \\ &\quad + \sum_{s_1=0}^{s} \sum_{k_1=0}^{k-2} R(v_1, s_1, k_1) R(v_2, s{-}s_1, k{-}2{-}k_1) \frac{\beta(k)}{\beta(k_1)\beta(k{-}2{-}k_1)}. \end{aligned} \tag{12}$$

Proof. Parts (1) and (2) follow from the definition of R.

(3) Let e be the edge from v_1 to v. When $k = 0$ it holds that $E \in \mathcal{E}(v,s,k)$ if and only if $E = E_1 \cup \{e\}$ for some $E_1 \in \mathcal{E}(v_1, s-1, k_1)$, where k_1 ranges from

0 to $s-1$. This gives the first case. When $k \geq 1$, the edge e connecting v and v_1 is absent from every set in $\mathcal{E}(v,s,k)$. Thus $E \in \mathcal{E}(v,s,k)$ if and only if $E \in \mathcal{E}(v_1,s,k-1)$.

$$\begin{aligned} N_E(T_v) &= N_E(T_{v_1})\frac{\beta(k)}{\beta(k-1)} \\ &= N_E(T_{v_1})(2k+1). \end{aligned}$$

(4) Let e_1, e_2 be the edges from v to v_1, v_2 respectively. Since $k=0$, for all $E \in \mathcal{E}(v,s,k)$, we have $e_1 \in E$ and $e_2 \in E$. Thus $E \in \mathcal{E}(v,s,k)$ if and only if there exists $E_1 \in \mathcal{E}(v_1,s_1,k_1)$ and $E_2 \in \mathcal{E}(v_2,s-2-s_1,k_2)$ for some $s_1,k_1,k_2 \geq 0$ such that $E = E_1 \cup E_2$. For each such set E, we have: $N_E(T_v) = N_{E_1}(T_{v_1})N_{E_2}(T_{v_2})$.

(5) Again, let e_1, e_2 be the edges from v to v_1, v_2 respectively. For each $E \in \mathcal{E}(v,s,k)$ with $k>0$, exactly one of the following cases holds:

Case 1: $e_1 \in E$ but $e_2 \notin E$. This case applies if and only there exists $E_1 \in \mathcal{E}(v_1,s_1,k_1)$ and $E_2 \in \mathcal{E}(v_2,s-1-s_1,k-1)$ for some $s_1,k_1 \geq 0$ such that $E = E_1 \cup E_2 \cup \{e_1\}$. For such a set E we have

$$N_E(T_v) = N_{E_1}(T_{v_1})N_{E_2}(T_{v_2})\frac{\beta(k)}{\beta(k-1)}.$$

Case 2: $e_1 \notin E$ but $e_2 \in E$. Identical to Case 1 with v_1 and v_2 switched.

Case 3: $e_1 \notin E$ and $e_2 \notin E$. This case applies if and only there exists $E_1 \in \mathcal{E}(v_1,s_1,k_1)$ and $E_2 \in \mathcal{E}(v_2,s-s_1,k-k_1-2)$ such that $E = E_1 \cup E_2$.

For each such set E we have:

$$N_E(T_v) = N_{E_1}(T_{v_1})N_{E_2}(T_{v_2})\frac{\beta(k)}{\beta(k_1)\beta(k-2-k_1)}.$$

□

Theorem 3. *Given a fully resolved tree T on n leaves the coefficients $b_m(T)$ can be computed in $O(n^5)$ time.*

Proof. Consider a vertex $v \in \overset{\circ}{V}(T)$. If v has one child in $\overset{\circ}{V}(T)$ then we evaluate (10) for all $s, k \leq n-3$ in $O(n^3)$ time. If v has two children in $\overset{\circ}{V}(T)$ then we evaluate (12) in $O(n^4)$ time.

Hence computing all the coefficients $r_s(T)$ takes $O(n^5)$ time. From (6), we obtain:

$$q_m(T) = \sum_{s=m}^{n-3} \binom{s}{m} r_s(T)(-1)^{s-m}, \tag{13}$$

from which we compute the values $b_m(T) = q_{n-3-m/2}(T)$. □

3. Poisson approximation

When n is large we can approximate the $q_s(T)$ values by a Poisson distribution with mean $\lambda_T := c_T/2n$ where c_T denotes the number of cherries of T (recall that a *cherry* is a pair of leaves whose incident edges contain a common vertex). More precisely, we have the following result.

Theorem 4. *For any tree $T \in BPT(n)$, let Y_T be a Poisson random variable with mean λ_T. Then the distribution $q_s(T)/b(n)$ as a function of s (the proportion of trees in $BPT(n)$ that share s nontrivial splits with T) and the distribution of Y_T have variational distance that converges to zero as $n \to \infty$. In particular,*

$$\sum_{s \geq 0} |q_s(T)/b(n) - e^{-\lambda_T}\lambda_T^s/s!| = O(n^{-1}).$$

Proof. Let X_T denote the random variable which counts the number of non-trivial splits that T shares with a tree T' selected uniformly at random from $BPT(n)$. Thus, $\mathbb{P}(X_T = s) = q_s(T)/b(n)$. Let X'_T be defined in the same ways as for X_T but counting only splits that divide the leaf set into subsets of size 2 and $n-2$. Clearly, $X'_T \leq X_T$. Moreover, the probability of the event G that T' shares a split with T that is not of the type counted by X'_T is bounded above by a term of order n^{-1} and so (since $\mathbb{P}(X_T = X'_T) \geq \mathbb{P}(X_T = X'_T|G^c)P(G^c) = 1 \cdot (1 - O(n^{-1})))$ we have:

$$\mathbb{P}(X_T \neq X'_T) = O(n^{-1}). \tag{14}$$

Now, for any two discrete random variables X and X' an elementary probability argument shows that $\sum_s |\mathbb{P}(X = s) - \mathbb{P}(X' = s)| \leq 2\mathbb{P}(X \neq X')$, and so:

$$\sum_{s \geq 0} |\mathbb{P}(X_T = s) - \mathbb{P}(X'_T = s)| \leq 2\mathbb{P}(X_T \neq X'_T). \tag{15}$$

Combining (14) and (15) gives:

$$\sum_{s \geq 0} |\mathbb{P}(X_T = s) - \mathbb{P}(X'_T = s)| = O(n^{-1}). \tag{16}$$

By the triangle inequality,

(17)
$$\sum_{s\geq 0} |\mathbb{P}(X_T = s) - \mathbb{P}(Y_T = s)| \leq \sum_{s\geq 0} |\mathbb{P}(X_T = s) - \mathbb{P}(X'_T = s)| + \sum_{s\geq 0} |\mathbb{P}(X'_T = s) - \mathbb{P}(Y_T = s)|$$

which, combined with (16), gives:

$$\sum_{s\geq 0} |\mathbb{P}(X_T = s) - \mathbb{P}(Y_T = s)| \leq \sum_{s\geq 0} |\mathbb{P}(X'_T = s) - \mathbb{P}(Y_T = s)| + O(n^{-1}). \tag{18}$$

Thus, to establish Theorem 4 it suffices to show that

$$\sum_{s\geq 0} |\mathbb{P}(X'_T = s) - \mathbb{P}(Y_T = s)| = O(n^{-1}). \tag{19}$$

Now, by Lemma 3 of [12], we have:

$$\mathbb{P}(X'_T = s) = \sum_{r=s}^{c_T} (-1)^{r+s} \binom{r}{s} \binom{c_T}{r} \frac{b(n-r)}{b(n)}. \tag{20}$$

Furthermore, letting λ denote λ_T for brevity, we have:

$$\mathbb{P}(Y_T = s) = e^{-\lambda}\lambda^s/s! = \sum_{r=s}^{\infty} (-1)^{r+s} \binom{r}{s} \frac{\lambda^r}{r!}.$$

Substituting this and (20) into the left-hand side of (19) gives the expression:

$$\sum_{s\geq 0} \left| \sum_{r=s}^{\infty} (-1)^{r+s} \binom{r}{s} \left[\binom{c_T}{r} \frac{b(n-r)}{b(n)} - \frac{\lambda^r}{r!} \right] \right| \tag{21}$$

which, after some algebra, and moving the absolute value inside the second summation, is bounded above by:

$$\Delta_n := \sum_{s\geq 0} \frac{1}{s!} \sum_{r=s}^{\infty} \frac{1}{(r-s)!} f(n,r) \tag{22}$$

where

$$f(n,r) := \left(\frac{c_T}{2n}\right)^r \cdot \left| \frac{\prod_{i=1}^{r-1}(1-i/c_T)}{\prod_{j=1}^{r}(1-(2j+3)/2n)} - 1 \right|$$

Using the fact that $c_T \leq n/2$, and a somewhat tedious case analysis, it can be shown that $f(n,r) \leq C/n$ for a constant C that is independent of r, n. It follows that

$$\Delta_n \leq \sum_{s\geq 0} \frac{1}{s!} \sum_{r=s}^{\infty} \frac{1}{(r-s)!} C/n = Ce^2/n,$$

which establishes (19) and thereby the theorem. □

Remark If T is selected uniformly at random from $BPT(n)$, then λ_T converges in probability to $\frac{1}{8}$ (since the variance of λ_T is $O(n^{-1})$ by Theorem 4(b) of [5]). Thus, Theorem 4 can be viewed as a refinement of the main result from [12] that for two trees selected uniformly at random from $BPT(n)$ the number of non-trivial splits they share is asymptotically Poisson distributed with mean $\frac{1}{8}$.

Application to Likelihood based supertrees

Rodrigo and Steel [10] recently presented a likelihood framework for constructing consensus trees and supertrees. Let $\mathcal{L}(T_i)$ denote the set of leaves of a (fully resolved) gene tree T_i. The probability of observing T_i with leaf set $\mathcal{L}(T_i) = X_i$ given an estimated species tree or supertree T has the form

$$\mathbb{P}_{T,X_i}(T_i) = \mathbb{P}_T(T_i) = \frac{1}{\mathcal{Z}_{T|\mathcal{L}(T_i)}} e^{-\beta_i d(T_i, T|\mathcal{L}T_i)} \tag{23}$$

where $T|\mathcal{L}(T_i)$ denotes the restriction of T to the leaf set T_i, and where β_i is a positive parameter that can be inferred by the data by maximum likelihood.

There are many reasons why an estimated gene tree might differ from the true tree, including sampling error, model violations, and alignment errors. Under the model of [10] the probability of observing a tree T_i on a given leaf set X_i falls off exponentially with its distance to the underlying tree T restricted to X_i. The parameter β can vary with the quantity and quality of the data, with high values of β corresponding to more confidence in the gene tree estimates. See [2] for a recent discussion of this approach.

The normalising constant

$$\mathcal{Z}_{T_i} = \mathcal{Z}_T^i \quad = \sum_{T':\mathcal{L}(T')=\mathcal{L}(T_i)} e^{-\beta_i d(T',T|\mathcal{L}T_i)} \tag{24}$$

is required so that the $\mathbb{P}_T(T_i)$ values sum to 1 over all choices of T_i. One complication with this approach is that the normalising functions $\mathcal{Z}_{T_i}$ depend on T (more precisely, although $\mathcal{Z}_{T_i}$ does not depend on how the leaves of T are labeled, it may depend on the shape of T), meaning that the constant needs to be computed in order to compare the likelihood values of two trees. This was overlooked in [10], in particular Proposition 1 of that paper may only hold in certain cases (for example, if the sets X_i are of size at most 5, or if the β_i values are sufficiently large). However, Proposition 1 of [10] can be corrected by replacing the term

$$\sum_{i=1}^{k} \beta_i d(T_i, T|X_i)$$

in the statement of that Proposition by

$$\sum_{i=1}^{k} \beta_i d(T_i, T|X_i) + \gamma_i(T),$$

where

$$\gamma_i(T) = \sum_{i=1}^{k} \log(\mathcal{Z}_{T_i}) = \log(1 + \sum_{m>0} e^{-\beta_i m} n_m(T)),$$

and where $n_m(T)$ is the number of fully resolved phylogenetic trees on leaf set X_i that have distance m from $T|X_i$.

In general, normalising constants are difficult to evaluate. When d is the Robinson-Foulds distance, however, computing the constant is straight-forward. Suppose that $|X_i| = n$ and that $b_m(T)$ has been computed for all m. Then (suppressing the index i) we have:

$$\begin{aligned} \mathcal{Z}_T &= \sum_{T' \in BPT(n)} e^{-\beta d(T,T')} \\ &= \sum_m b_m(T) e^{-\beta m}. \end{aligned} \tag{25}$$

which can be evaluated directly from the $b_m(T)$ values, and thereby in polynomial time overall in n.

It is instructive to estimate $\mathcal{Z}_T$ in two limiting cases - firstly for values of β that are close to 0, and for values of β that are large. In both cases we find that the dominant aspect of the shape of T affecting $\mathcal{Z}_T$ is the number c_T of cherries that T has. The experimental performance of these approximations is evaluated in the final section.

3.1. **Small values of β.** Our first approximation for $\mathcal{Z}_T$ makes use of Theorem 4. Fix a tree T and, as before, let $\lambda := \lambda_T = c_T/2n$, where c_T is the number of cherries in T. Starting with (25) we have

$$\begin{aligned}
\mathcal{Z}_T &= \sum_m b_m(T)e^{-\beta m} \\
&= b(n)\left(\sum_s \frac{q_s(T)}{b(n)} e^{-2\beta(n-3-s)}\right) \\
&= b(n)\left(\sum_s \frac{e^{-\lambda}\lambda^s}{s!} e^{-2\beta(n-3-s)} + O(n^{-1})\right),
\end{aligned}$$

this last line following from Theorem 4 and the inequality $0 < e^{-2\beta(n-3-s)} \leq 1$. Thus

$$\begin{aligned}
\mathcal{Z}_T &= b(n)\left(e^{-2\beta(n-3)} \sum_s \frac{e^{-\lambda}\lambda^s}{s!} e^{2\beta s} + O(n^{-1})\right) && (26)\\
&= b(n)\left(e^{-2\beta(n-3)+\lambda(e^{2\beta}-1)} + O(n^{-1})\right) && (27)
\end{aligned}$$

giving the *small-beta approximation*

$$\mathcal{Z}_T \approx b(n)\left(e^{-2\beta(n-3)+\lambda(e^{2\beta}-1)}\right). \tag{28}$$

Note that equation (27) makes use of the formula for the moment generating function of the Poisson distribution. For β close to 0, the identity $e^{-\beta m} = 1 - \beta m + O(\beta^2)$ reveals that the difference between $\mathcal{Z}_T$ and the approximation

$$b(n)\left(1 - \beta(2n - 6 - 2\frac{c_T}{2n})\right)$$

consists of terms of order β^2 and n^{-1}. Thus, for n large, as β converges to 0, $\mathcal{Z}_T$ converges to a constant, and when β is close to 0, the small difference from this constant is dominated by c_T.

3.2. **Large values of β.** When β is large, let $\epsilon = e^{-2\beta}$. Then,

$$\mathcal{Z}_T = 1 + b_2(T)\epsilon + b_4(T)\epsilon^2 + O(\epsilon^3).$$

Now, $b_2(T) = 2(n-3)$, and from Theorem 2.26 of [11] we have:

$$b_4(T) = 4\binom{n-3}{2} + 6(n - 6 + c_T).$$

Thus if we let $A_{n,\epsilon} := 1 + (2n-3)\epsilon + 2(n^2 - 4n - 6)\epsilon^2$ then

$$\mathcal{Z}_T = A_{n,\epsilon} + 6c_T\epsilon^2 + O(\epsilon^3),$$

giving the *large-β approximation*

$$\mathcal{Z}_T \approx A_{n,\epsilon} + 6c_T\epsilon^2. \tag{29}$$

Once again we see that in the limit (in this case, as β tends to infinity) $\mathcal{Z}_T$ converges to a constant, and for large values of β, the small difference from this constant is dominated by c_T.

4. Experimental results

4.1. **Features of distribution.** To study general features of the distribution, and examine the accuracy of the above approximations, we generated random trees and computed the distribution of the Robinson-Foulds distance for each tree. The trees were drawn from a uniform distribution, with the number of taxa varying from 5 to 50. One thousand replicates were performed for each number of taxa. We also constructed an unrooted caterpillar tree and a balanced unrooted tree for every set of taxa. A balanced unrooted tree is one that minimises the length of the longest path between any two leaves, an example being the right-hand tree in Fig. 1.

As predicted from the Poisson approximation, the distributions of Robinson-Foulds distances from a fixed tree were highly peaked. For all of the trees examined, at least 99% of trees are either at distance $2(n-3)$, the maximum possible, or distance $2(n-4)$.

For $T \in BPT(n)$, let $N_k(T)$ denote the number of trees in $BPT(n)$ within Robinson-Foulds distance k of T: that is,

$$N_k(T) = \sum_{m=0}^{k} b_m(T).$$

Then $N_2(T) = 2(n-3)+1$, the number of trees that share all but one split with T, together with the tree T itself. When $k > 2$, the value of $N_k(T)$ varies with the shape of T. We observed that for all k, $N_k(T)$ was minimised when T is a caterpillar. At the other extreme, $N_k(T)$ was almost always maximised when T was balanced, the exception being when T was balanced but did not have the maximum number of cherries.

4.2. **Accuracy of approximations.** For each tree, and a range of different values for β, we computed the exact normalising constant $\mathcal{Z}_T$. Fig. 2 illustrates the variation in $\mathcal{Z}_T$ over different values of β, displayed on a log-log plot. The central curve gives the average $\mathcal{Z}_T$ values for 1000 fifty-taxa trees drawn from a uniform distribution, as a function of β. The small-β and large-β approximate values for $\mathcal{Z}_T$ are also plotted.

As a function of β, the normalising constant has two distinct phases. The small-β approximation fits well for $\log(\beta) < 0.2$ (approximately) while the large-β approximation fits well for $\log(\beta) > 0.2$. By differentiating, we see that the small-β approximation has a minimum at $\beta = \frac{1}{2}\log\left(\frac{n-3}{\lambda}\right)$. To the left of this minimum, the curve is well fitted by the maximum of the two approximations. To the right of this minimum, the large-β approximation is best. To summarise, let c_T be the number of cherries of T, $\lambda = c_T/2n$, $\epsilon = e^{-2\beta}$ and $A_{n,\epsilon} := 1 + (2n-3)\epsilon + 2(n^2 - 4n - 6)\epsilon^2$. We then have the approximation

$$\mathcal{Z}_T \approx \begin{cases} \max\left\{(b(n)\left(e^{-2\beta(n-3)+\lambda(e^{2\beta}-1)}\right), A_{n,\epsilon} + 6c_T\epsilon^2\right\} & \text{if } \beta < \frac{1}{2}\log\left(\frac{n-3}{\lambda}\right); \\ A_{n,\epsilon} + 6c_T\epsilon^2 & \text{otherwise.} \end{cases} \tag{30}$$

4.3. **Importance of normalising constant.** As we observed above, to correctly compute the likelihood for a supertree under the model of [10] we need to compute $\mathcal{Z}_T$ for every distinct supertree T. Even though this calculation take polynomial time, it is still extremely expensive computationally, particularly considering that millions of candidate supertrees may be considered. We ask, then, the extent to which this computation is strictly necessary. In particular, if we ignore the normalising constant when comparing likelihoods, would the relative likelihood ordering of distinct trees change. The key question is then to determine how much the normalisation constants $\mathcal{Z}_T$ vary. If the difference is sufficiently small then there will be no impact from ignoring the differences between normalising constants.

For a given value of β define the *range* of $\mathcal{Z}_T$ to be the ratio of the largest to the smallest $\mathcal{Z}_T$ values over all fully-resolved trees with n taxa. Fig. 3 plots the range

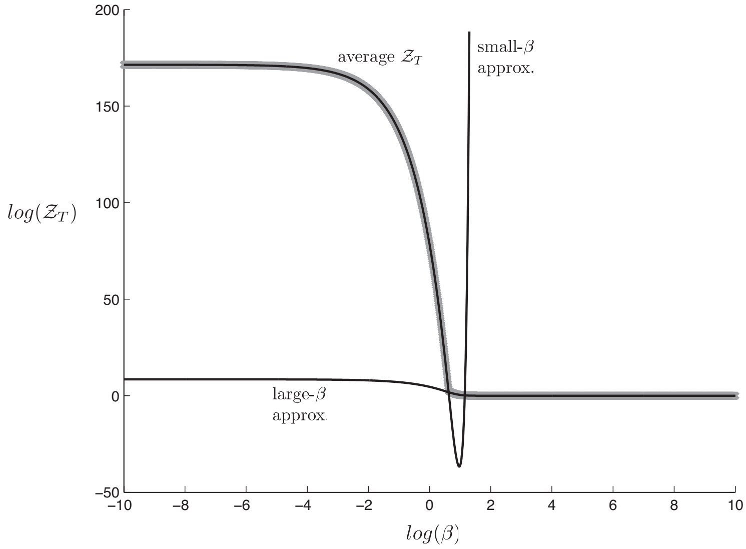

FIGURE 2. The average $\mathcal{Z}_T$ values for different values of β, plotted (in grey) on a log-log axis. The approximations (with error terms discarded) for small and large β are also plotted in black. All values were computed by drawing 1000 fifty taxa trees from a uniform distribution and computing normalising constants exactly using the algorithms described here.

of $\mathcal{Z}_T$ for the values of β used in Fig. 2, and for $n = 10, 20, 30, 40, 50$ taxa trees, on a log-log axis. The trees minimising $\mathcal{Z}_T$ were always caterpillar trees and the trees maximising $\mathcal{Z}_T$ were usually, but not always, balanced trees. The figure indicates that when β is outside the range $[0.03, 3]$ there is little variation in $\mathcal{Z}_T$ between different trees. With 50 taxa, the normalising constants differ by a maximum of 7.5 log-units.

Suppose that we are comparing the log-likelihood of two trees T_1 and T_2 with respect to a third tree T. If $d_{RF}(T, T_1) \neq d_{RF}(T, T_2)$ then

$$|\log(e^{-\beta d(T,T_1)}) - \log(e^{-\beta d(T,T_2)})| \geq 2\beta$$

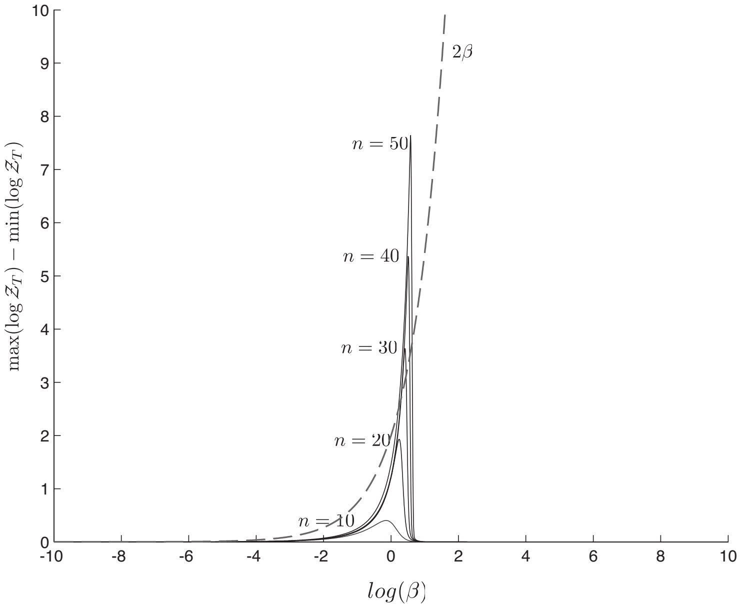

FIGURE 3. The range of the $\mathcal{Z}_T$ values computed for different β and plotted on a log-log axis. The $\mathcal{Z}_T$ values were computed by drawing 1000 trees from a uniform distribution with $n = 10, 20, 30, 40, 50$ taxa (five curves). The range is the difference between the maximum $\mathcal{Z}_T$ and minimum $\mathcal{Z}_T$ values, for each choice of β and n. The dotted line indicates the 2β value: when the range is less than 2β ignoring the normalising constant has no effect on the relative order of likelihood values.

so ignoring the normalising constant will only change the order of likelihood values if $|\log \mathcal{Z}_{T_1} - \log \mathcal{Z}_{T_2}| \geq 2\beta$. Plotting the curve for 2β on Fig. 3 we see that $|\log \mathcal{Z}_{T_1} - \log \mathcal{Z}_{T_2}| \geq 2\beta$ for some pairs of 50-taxa trees only when β lies in the interval $[1.25, 1.86]$. The corresponding interval will be even smaller for trees with fewer taxa: for 20 taxa trees there is no value of β for which ignoring $\mathcal{Z}_T$ scores leads to a switch in the order of likelihood values for two trees.

In summary, when β is approximately 1.5, and the number of taxa is greater than around 20, it is potentially important to correctly compute normalisation constants.

Outside that range, the influence of $\mathcal{Z}_T$ on likelihood rankings can be safely ignored. We note, however, that here we are only interested in relative ordering of supertrees with respect to likelihood: a Bayesian Monte-Carlo approach may well need accurate $\mathcal{Z}_T$ values for all β.

Acknowledgements. We thank the three anonymous referees for helpful comments, particularly for the suggestion to use the Poisson approximation directly when deriving small β approximations for $\mathcal{Z}_T$.

References

[1] M. Bourque, "Arbes de Steiner et reseaux dont varie l'emplagement de certains sommets," PhD thesis, Université de Montréal, Québec, Canada, 1978.

[2] J. Cotton and M. Wilkinson, "Supertrees join the mainstream of phylogenetics," *Trends in Ecology and Evolution*, vol. 24, no 1, pp. 1–3, 2009.

[3] J. Felsenstein, *Inferring phylogenies.* Sinauer Press, 2004.

[4] M.D. Hendy, C.H.C. Little, and D. Penny, "Comparing trees with pendant vertices labelled," *SIAM Journal of Applied Mathematics*, vol. 44, no. 5, pp. 1054–1065, 1984.

[5] A. McKenzie and M. Steel, "Distributions of cherries for two models of trees," *Mathematical Biosciences*, vol. 164, pp. 81–92, 2000.

[6] D. Penny, M.A. Steel and E. Watson, "Trees from languages and genes are very similar," *Systematic Biology*, vol. 42, no. 3, pp. 382–384, 1993.

[7] D. Penny, L.R. Founds, and M. D. Hendy, "Testing the theory of evolution by Comparing phylogenetic trees constructed from five different protein sequences," *Nature*, vol. 297, 197–200, 1982.

[8] D.F. Robinson and L.R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, pp. 131–147, 1981.

[9] C. Semple and M. Steel, *Phylogenetics.* Oxford University Press, 2003.

[10] M. Steel and A. Rodrigo, "Maximum likelihood supertrees," *Systematic Biology*, vol. 57, no. 2, pp. 243–250, 2008.

[11] M. Steel, "Distributions on bicoloured evolutionary trees," PhD Thesis, Massey University, Palmerston North, New Zealand, 1989.

[12] M. A. Steel, "Distribution of the symmetric difference metric on phylogenetic trees," *SIAM J. Discrete Math.*, vol. 1, no. 4, pp. 541–551, 1988.

[13] M.A. Steel and D. Penny, "Distributions of tree comparison metrics - some new results," *Systematic Biology*, vol. 42, no. 2, pp. 126–141, 1993.

DB: Mathematics Department, University of Auckland; MS: Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: d.bryant@auckland.ac.nz, m.steel@math.canterbury.ac.nz