



Mean deep coalescence cost under exchangeable probability distributions



C.V. Than^{*}, N.A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:

Received 10 July 2012

Received in revised form 6 February 2014

Accepted 11 February 2014

Available online 5 March 2014

Keywords:

Deep coalescence
Evolutionary models
Exchangeability
Lineage sorting
Phylogeny

ABSTRACT

We derive formulas for mean deep coalescence cost, for either a fixed species tree or a fixed gene tree, under probability distributions that satisfy the exchangeability property. We then apply the formulas to study mean deep coalescence cost under two commonly used exchangeable models—the uniform and Yule models. We find that mean deep coalescence cost, for either a fixed species tree or a fixed gene tree, tends to be larger for unbalanced trees than for balanced trees. These results provide a better understanding of the deep coalescence cost, as well as allow for the development of new species tree inference criteria.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The deep coalescence cost is a measure of the relationship between an ordered pair of rooted, binary, labeled trees [17,18,24]. It arises in evolutionary models, in which one of the two trees, the “gene tree”, can be viewed as evolving over time through a process dependent on the other tree, the “species tree” [7,19]. Computations of the deep coalescence cost have been of interest in a variety of problems, such as in algorithms to identify a species tree that produces the minimal deep coalescence cost summed over a given set of gene trees [24,25], and in studies of trees that describe the duplication and loss of genes in a set of species [2,28].

Recent papers have established a series of mathematical properties of the deep coalescence cost. Lin et al. [16] proved that the minimal deep coalescence cost tree is Pareto. That is, for any given set of gene trees, if a cluster (i.e. the leaf set of a subtree) appears in every input gene tree, then the cluster must appear in every species tree that produces the minimal deep coalescence cost for the set of gene trees. Zhang [28] provided an elegant relationship between the deep coalescence cost and the duplication and loss of genes in a set of species, and showed that computing a minimal deep coalescence cost for a set of gene trees is NP-hard. Most recently, considering all possible gene trees, we have obtained the maximum deep coalescence cost for a fixed species tree, characterizing the set of gene trees that achieve this maximum [26]. Further, considering all possible species trees, although a general characterization of the maximum deep coalescence cost for a fixed gene tree remains open, we have solved the problem in certain cases and have identified some features of the general solution [26].

Here we pursue an analogous investigation of the *mean* deep coalescence cost, under a general probability model for trees in which leaf labels are exchangeable. After presenting key concepts in Section 2, in Section 3 we introduce a

^{*} Correspondence to: ZBIT, Universität Tübingen, Germany. Tel.: +49 176 8485 1432.

E-mail addresses: thvcuong@gmail.com (C.V. Than), noahr@stanford.edu (N.A. Rosenberg).

convenient formula for the deep coalescence cost that is used in subsequent computations. In Section 4, we obtain the mean deep coalescence cost for a fixed species tree, considering all possible gene trees. In Section 5, we obtain the mean deep coalescence cost for a fixed gene tree, considering all possible species trees. We further establish the upper bound on this mean, considering all possible gene trees, and characterize the gene trees that achieve the upper bound. In Sections 4 and 5, although our main results apply to general exchangeable models, we focus on two of the most popular exchangeable models used in evolutionary biology, the Yule and uniform models. The paper concludes with a short discussion.

2. Background

We consider only binary, rooted trees. We also assume that tree leaves are bijectively labeled by a (nonempty) taxon set X . The set of all binary, rooted trees on X is denoted by $R(X)$. It is well-known that the number of trees in $R(X)$ is

$$b(|X|) = (2|X| - 3)!! = \frac{(2|X| - 2)!}{2^{|X|-1}(|X| - 1)!}, \quad (1)$$

with the convention that $b(1) = (-1)!! = 1$ [5,8,20]. For brevity, we let $|X| = n$ throughout the paper, except when explicitly stated otherwise.

We denote by $V(T)$ and $E(T)$ the sets of nodes and edges of a tree $T \in R(X)$. The set $V(T)$ minus the set of leaves of T is called the set of internal nodes of T , and it is denoted by $\dot{V}(T)$. An edge of T is called an internal edge if both of its endpoints are in $\dot{V}(T)$. The set of internal edges of T is denoted by $\dot{E}(T)$. For $T \in R(X)$, $|V(T)| = 2n - 1$ and $|E(T)| = 2n - 2$. Thus, $|\dot{V}(T)| = |V(T)| - n = n - 1$ and $|\dot{E}(T)| = |E(T)| - n = n - 2$.

For a node $v \in V(T)$, let $T(v)$ be the subtree of T rooted at v , and let $C_T(v)$ be the set of leaves of $T(v)$. The set $C_T(v)$ is called the cluster induced by v . The set of clusters induced by T is

$$\mathcal{C}(T) = \{C_T(v) \mid v \in V(T)\}. \quad (2)$$

We consider the edges of T to be directed away from the root of T . A non-root node v uniquely determines edge $e = (u, v)$, where u is the parent of v . Hence, for convenience we also refer to $T(v)$ and $C_T(v)$ as $T(e)$ and $C_T(e)$, respectively.

2.1. External path length and the Sackin index

For a node $v \in V(T)$, let $\ell_T(v)$ be the length of the path from the root of T to v (i.e. the number of edges in the path). The length $\ell_T(v)$ is called the depth of v . The external path length of T is defined as

$$\text{epl}(T) = \sum_{x \in X} \ell_T(x). \quad (3)$$

Eq. (3) can be expressed as (e.g. [3]):

$$\text{epl}(T) = \sum_{e \in E(T)} |C_T(e)| = -n + \sum_{v \in V(T)} |C_T(v)|. \quad (4)$$

The reason for Eq. (4) is that a leaf x appears in exactly $\ell_T(x)$ clusters $C_T(e)$, and hence by summing $\ell_T(x)$ over all the leaves of T , we obtain the sum of the sizes of all clusters $C_T(e)$. More formally, we have

$$\ell_T(x) = \sum_{e \in E(T)} [x \in C_T(e)],$$

where $[\cdot]$ is the Iverson bracket, equaling 1 if the logical condition in square brackets is true, and 0 otherwise. Thus,

$$\text{epl}(T) = \sum_{x \in X} \ell_T(x) = \sum_{x \in X} \sum_{e \in E(T)} [x \in C_T(e)] = \sum_{e \in E(T)} \sum_{x \in X} [x \in C_T(e)] = \sum_{e \in E(T)} |C_T(e)|.$$

The Sackin index of tree T is the average depth of a leaf in T [13,22]:

$$\bar{\ell}(T) = \frac{\text{epl}(T)}{n} = \frac{1}{n} \sum_{x \in X} \ell_T(x). \quad (5)$$

This index, and hence the external path length, can be considered as a measure of the amount of unbalance of a tree: in general, $\bar{\ell}(T)$ is larger for more asymmetric trees (e.g. [13]). In fact, for a fixed n , $\text{epl}(T)$ (and hence $\bar{\ell}(T)$) is largest if T is a caterpillar tree [14,15], and smallest when T is a complete binary tree [15], in which $2n - 2^k$ leaves, where $k = \lceil \log_2 n \rceil$,

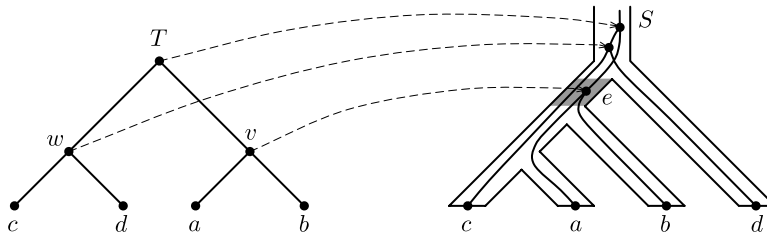


Fig. 1. An illustration of the calculation of $dc(T, S)$. Each node of the gene tree T (left) is mapped to its MRCA in the species tree S (right). In this example, only one internal node of T (node w) is mapped to a node below edge e , and hence there is $|C_S(e)| - 1 - 1 = 1$ extra lineage in e . In total, two extra lineages are required to reconcile T within S .

have depth k and appear as far left in T as possible, and the remaining $2^k - n$ leaves have depth $k - 1$ and appear as far right in T as possible. For example, tree $((((a, b), (c, d)), ((e, f), g)))$ is a complete binary tree for $n = 7$.

2.2. Deep coalescence cost

The deep coalescence cost for reconciling a gene tree $T \in R(X)$ within a species tree $S \in R(X)$, $dc(T, S)$, is computed by using a most recent common ancestor (MRCA) mapping between the nodes of T and the nodes of S [17]. For a node $v \in V(T)$, let $MRCA_S(v)$ be the farthest node from the root of S such that $C_T(v) \subseteq C_S(MRCA_S(v))$. For an edge e of S , let c_e be the size of the set $\{v \in V(T) \mid MRCA_S(v) \text{ is a node in } C_S(e)\}$, that is, the set of internal nodes of T that are mapped to nodes of $C_S(e)$. We define the number of extra lineages in an edge e as

$$xl(T, e) = |C_S(e)| - c_e - 1, \tag{6}$$

and compute the cost $dc(T, S)$ as the sum

$$dc(T, S) = \sum_{e \in E(S)} xl(T, e). \tag{7}$$

An example of how the deep coalescence cost is computed appears in Fig. 1.

It is possible to compute $xl(T, e)$ and $dc(T, S)$ without relying on an MRCA mapping. For a nonempty subset A of X , we say that a subtree $T(v)$ of T is A -maximal if:

1. The leaf set of $T(v)$ (i.e. $C_T(v)$) is a subset of A .
2. For any subtree t of T of which $T(v)$ is a proper subtree, the leaf set of t is not a subset of A .

By this definition, the only X -maximal subtree of T is tree T itself. Conversely, if A is a nonempty, proper subset of X , then an A -maximal subtree of T must be a proper subtree of T , that is, it must be induced by some non-root node v of T . It can be seen that $T(v)$ is A -maximal, where $A \subsetneq X$ and $A \neq \emptyset$, if and only if [26]:

1. The leaf set of $T(v)$, $C_T(v)$, is contained in A .
2. The leaf set of $T(u)$, $C_T(u)$, where u is the parent of v , is not contained in A .

For an edge e of S , denote by $ms(T, C_S(e))$ the number of $C_S(e)$ -maximal subtrees of T . We then have [24]:

$$xl(T, e) = ms(T, C_S(e)) - 1. \tag{8}$$

2.3. Exchangeable probability models on trees

Consider a probability distribution on the trees in $R(X)$, and let $\mathbb{P}(T)$ denote the probability of a tree $T \in R(X)$. The probability that a subset A of X is an induced cluster of a tree randomly sampled from $R(X)$ is:

$$\mathbb{P}_X(A) = \sum_{T \in R(X)} \mathbb{P}(T) [A \in \mathcal{C}(T)]. \tag{9}$$

Note that $\mathbb{P}_X(X) = 1$ and $\mathbb{P}_X(\{x\}) = 1$ for every $x \in X$, as both X and single-leaf clusters $\{x\}$ are induced by every tree in $R(X)$.

Let π be a permutation of X . For a tree $T \in R(X)$, let $\pi(T)$ denote the tree obtained from T by relabeling each leaf x of T by $\pi(x)$. We say that a probability distribution has the exchangeability property [1] if for every π

$$\mathbb{P}(T) = \mathbb{P}(\pi(T)).$$

It follows from exchangeability that $\mathbb{P}_X(A)$ is the same for all subsets $A \subseteq X$ with the same number of elements. Consider two subsets A and A' of the same cardinality. Let π be a permutation of X that maps the elements of A to the elements of A' . Then $A \in \mathcal{C}(T)$ implies $A' \in \mathcal{C}(\pi(T))$. It follows that $[A \in \mathcal{C}(T)] \leq [A' \in \mathcal{C}(\pi(T))]$. Because $\mathbb{P}(T) = \mathbb{P}(\pi(T))$, we have

$$\begin{aligned} \mathbb{P}_X(A) &= \sum_{T \in R(X)} \mathbb{P}(T) [A \in \mathcal{C}(T)] \leq \sum_{T \in R(X)} \mathbb{P}(\pi(T)) [A' \in \mathcal{C}(\pi(T))] \\ &= \sum_{T' \in R(X)} \mathbb{P}(T') [A' \in \mathcal{C}(T')] \quad (T' = \pi(T)) \\ &= \mathbb{P}_X(A'). \end{aligned}$$

Exchanging the roles of A and A' , we also have $\mathbb{P}_X(A') \leq \mathbb{P}_X(A)$. Hence, $\mathbb{P}_X(A) = \mathbb{P}_X(A')$.

Both the uniform (also called PDA for “proportional to distinguishable arrangements”) and Yule models have the exchangeability property [1]. In the uniform model, every tree in $R(X)$, regardless of its shape and leaf labeling, has the same probability $1/b(n)$. Let A be a subset of X , and let z be an element not in X . Then a tree $T \in R(X)$ of which A is a cluster can be constructed by replacing leaf z in a tree in $R(X \setminus A \cup \{z\})$ by a tree in $R(A)$. Hence, the number of such trees $T \in R(X)$ is $b(n - i + 1)b(i)$, where $i = |A|$, and the probability $\mathbb{P}_X(A)$ under the uniform model is

$$p_n^u(i) = \frac{b(i)b(n - i + 1)}{b(n)} = \binom{n - 1}{i - 1} \binom{2n - 2}{2i - 2}^{-1}. \tag{10}$$

The Yule model [11,27] generates trees from the root by repeatedly splitting a leaf, chosen uniformly at random, into two leaves. The Yule model is equivalent to the coalescent process [12], ignoring branch lengths [1]. In the Yule model, the probability of a tree in $R(X)$, regardless of its leaf labeling, is [4,23]

$$\mathbb{P}(T) = \frac{2^{n-1}}{n!} \prod_{v \in \dot{V}(T)} \frac{1}{|C_T(v)| - 1}. \tag{11}$$

Unlike in the uniform model, the probability of a tree in the Yule model depends on its shape. Rosenberg [21] proved that the probability $\mathbb{P}_X(A)$ of an i -subset A is

$$p_n^y(i) = \begin{cases} \frac{2n}{i(i + 1)} \binom{n}{i}^{-1} & \text{if } 1 \leq i \leq n - 1, \\ 1 & \text{if } i = n. \end{cases} \tag{12}$$

3. A formula for $dc(T, S)$

In this section, we derive a formula that relates $dc(T, S)$ to the external path length of S . The formula is the key for obtaining the results in the rest of the paper.

Lemma 1. *Let A be a nonempty subset of X , and let T be a tree in $R(X)$. Then the number of A -maximal subtrees of T is*

$$ms(T, A) = |A| - \sum_{v \in \dot{V}(T)} [C_T(v) \subseteq A] = 2|A| - \sum_{v \in V(T)} [C_T(v) \subseteq A]. \tag{13}$$

Proof. Define the indicator variable

$$\mathbb{I}_A(T(v)) = \begin{cases} 1 & \text{if } T(v) \text{ is } A\text{-maximal,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$ms(T, A) = \sum_{v \in \dot{V}(T)} \mathbb{I}_A(T(v)).$$

If v is the root of T , then $\mathbb{I}_A(T(v)) = \mathbb{I}_A(T) = [X \subseteq A]$, because T is A -maximal if and only if $A = X$. For a non-root node $v \in V(T)$, let u be the parent of v (i.e. $(u, v) \in E(T)$). Then $[C_T(v) \subseteq A] - [C_T(u) \subseteq A] = 1$ if $C_T(v) \subseteq A$ and $C_T(u) \not\subseteq A$, and 0 if either $C_T(v) \not\subseteq A$ (which implies $C_T(u) \not\subseteq A$) or $C_T(u) \subseteq A$ (which implies $C_T(v) \subseteq A$). Thus,

$$\mathbb{I}_A(T(v)) = [C_T(v) \subseteq A] - [C_T(u) \subseteq A].$$

It follows that

$$ms(T, A) = [X \subseteq A] + \sum_{(u,v) \in E(T)} ([C_T(v) \subseteq A] - [C_T(u) \subseteq A]). \tag{14}$$

Because each internal node of T , including the root, has exactly two children,

$$\begin{aligned} \sum_{(u,v) \in E(T)} ([C_T(v) \subseteq A] - [C_T(u) \subseteq A]) &= \sum_{(u,v) \in E(T)} [C_T(v) \subseteq A] - 2 \sum_{u \in \hat{V}(T)} [C_T(u) \subseteq A] \\ &= -[X \subseteq A] + \sum_{v \in V(T)} [C_T(v) \subseteq A] - 2 \sum_{v \in \hat{V}(T)} [C_T(v) \subseteq A] \\ &= |A| - [X \subseteq A] - \sum_{v \in \hat{V}(T)} [C_T(v) \subseteq A]. \end{aligned} \tag{15}$$

Eqs. (14) and (15) imply the first part of Eq. (13). The second part of it follows because exactly $|A|$ leaves of T are in A . \square

By the preceding lemma,

$$xl(T, e) = ms(T, C_S(e)) - 1 = 2|C_S(e)| - 1 - \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)].$$

Substituting this equation into Eq. (7), we obtain

$$\begin{aligned} dc(T, S) &= \sum_{e \in E(S)} xl(T, e) = -(2n - 2) + 2 \sum_{e \in E(S)} |C_S(e)| - \sum_{e \in E(S)} \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)] \\ &= -(2n - 2) + 2epl(S) - \sum_{e \in E(S)} \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)], \end{aligned}$$

where in the last step we use Eq. (4). We have just proven the following important result.

Theorem 2. For a gene tree T and species tree S in $R(X)$,

$$dc(T, S) = -(2n - 2) + 2epl(S) - \sum_{e \in E(S)} \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)]. \tag{16}$$

4. Mean deep coalescence cost for a fixed species tree

In this section and the next section, we assume that the probability distribution on $R(X)$ has the exchangeability property. Recall that exchangeability implies that $\mathbb{P}_X(A)$ is the same for every subset A of X with a given size. We denote this shared probability by $p_n(i)$, where $|X| = n$ and $|A| = i$.

Let $S \in R(X)$ be a given species tree. The mean deep coalescence cost of S , averaging over all gene trees, is defined as

$$\overline{dc}_S(S) = \sum_{T \in R(X)} \mathbb{P}(T) dc(T, S).$$

We derive in this section a formula that expresses $\overline{dc}_S(S)$ in terms of the external path length of S and cluster probabilities $p_n(i)$. We then present formulas for $\overline{dc}_S(S)$ in the Yule and uniform models, as well as a comparison between the two formulas.

Theorem 3. If the probability distribution on $R(X)$ has the exchangeability property, then

$$\overline{dc}_S(S) = -(2n - 2) + 2epl(S) - \sum_{e \in E(S)} \sum_{i=1}^{|C_S(e)|} p_n(i) \binom{|C_S(e)|}{i}. \tag{17}$$

Proof. From the definition of $\overline{dc}_S(S)$ and Theorem 2, we have

$$\begin{aligned} \overline{dc}_S(S) &= \sum_{T \in R(X)} \mathbb{P}(T) \left(-(2n - 2) + 2epl(S) - \sum_{e \in E(S)} \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)] \right) \\ &= -(2n - 2) + epl(S) - \sum_{e \in E(S)} \sum_{T \in R(X)} \sum_{v \in V(T)} \mathbb{P}(T) [C_T(v) \subseteq C_S(e)]. \end{aligned} \tag{18}$$

Let $\mathcal{P}_0(X)$ be the collection of all nonempty subsets of X . Then

$$\begin{aligned} \sum_{T \in R(X)} \sum_{v \in V(T)} \mathbb{P}(T) [C_T(v) \subseteq C_S(e)] &= \sum_{T \in R(X)} \sum_{A \in \mathcal{P}_0(X)} \mathbb{P}(T) [A \in \mathcal{C}(T)] [A \subseteq C_S(e)] \\ &= \sum_{A \in \mathcal{P}_0(X)} [A \subseteq C_S(e)] \sum_{T \in R(X)} \mathbb{P}(T) [A \in \mathcal{C}(T)] \\ &= \sum_{A \in \mathcal{P}_0(X)} \mathbb{P}_X(A) [A \subseteq C_S(e)]. \end{aligned} \tag{19}$$

We now partition $\mathcal{P}_0(X)$ into families F_i of i -element subsets of X , $1 \leq i \leq n$. The number of elements $A \in F_i$ that are subsets of $C_S(e)$ is $\binom{|C_S(e)|}{i}$. Note that this quantity is 0 for $i > |C_S(e)|$, consistent with the fact that $A \not\subseteq C_S(e)$ if $|A| = i > |C_S(e)|$. Exchangeability implies that $\mathbb{P}_X(A) = p_n(i)$ for every $A \in F_i$. Therefore, the right-hand side of Eq. (19) is equal to

$$\sum_{i=1}^n \sum_{A \in F_i} \mathbb{P}_X(A) [A \subseteq C_S(e)] = \sum_{i=1}^n p_n(i) \binom{|C_S(e)|}{i} = \sum_{i=1}^{|C_S(e)|} p_n(i) \binom{|C_S(e)|}{i}.$$

The theorem now follows by substituting Eq. (19) into Eq. (18). \square

4.1. $\overline{dc}_S(S)$ in the Yule and uniform models

The following corollary provides a formula for $\overline{dc}_S(S)$ in the Yule model. It is obtained by plugging the probability $p_n(i) = \frac{2n}{i(i+1)} \binom{n}{i}^{-1}$ (Eq. (12)) for $1 \leq i \leq n - 1$ into Eq. (17). A corresponding formula for $\overline{dc}_S(S)$ in the uniform model is given in Corollary 5.

Corollary 4. Assume the Yule model on $R(X)$. Then the mean deep coalescence cost for a fixed species tree $S \in R(X)$ is

$$\overline{dc}_S^y(S) = -(2n - 2) + 2\text{epl}(S) - \sum_{e \in E(S)} \sum_{i=1}^{|C_S(e)|} \frac{2n}{i(i+1)} \binom{n}{i}^{-1} \binom{|C_S(e)|}{i}. \tag{20}$$

Corollary 5. Assume the uniform model on $R(X)$. Then the mean deep coalescence cost for a fixed species tree $S \in R(X)$ is

$$\overline{dc}_S^u(S) = -2n(2n - 2) + 2\text{epl}(S) + \frac{(2n - 2)!!}{(2n - 3)!!} \sum_{e \in E(S)} \frac{(2n - 2|C_S(e)| - 1)!!}{(2n - 2|C_S(e)| - 2)!!}. \tag{21}$$

Proof. From the probability $p_n(i) = b(i)b(n - i + 1)/b(n)$ (Eq. (10)) and the formula for $\overline{dc}_S(S)$ in Theorem 3, we have

$$\overline{dc}_S^u(S) = -(2n - 2) + 2\text{epl}(S) - \sum_{e \in E(S)} \sum_{i=1}^{|C_S(e)|} \binom{|C_S(e)|}{i} \frac{b(i)b(n - i + 1)}{b(n)}. \tag{22}$$

The inner sum of the last term in Eq. (22) can be simplified by the following claim. Let k be a positive integer smaller than or equal to n . Then

$$\sum_{i=1}^k \binom{k}{i} b(i)b(n - i + 1) = b(n + 1) - \frac{(2n - 2)!!}{(2n - 2k - 2)!!} b(n - k + 1). \tag{23}$$

Let Z be a set of $n + 1$ taxa, and let A be a fixed, k -element subset of Z . Denote by $\mathcal{T}_Z(A)$ the set of trees $T \in R(Z)$ that satisfy the following property:

(PA) The leaf set of the left subtree T_ℓ of T contains only elements of A .

Note that we do not require the leaf set of T_ℓ to contain every element of A , only that it *not* contain any element of $Z \setminus A$. Thus, some elements of A can appear in the right subtree T_r of T . We will prove Eq. (23) by counting the size of $\mathcal{T}_Z(A)$ in two different ways.

Property PA implies that the left subtree T_ℓ of a tree $T \in \mathcal{T}_Z(A)$ has at most k leaves. Let $F_i = \{T \in \mathcal{T}_Z(A) \mid T_\ell \text{ has exactly } i \text{ leaves}\}$, $1 \leq i \leq k$. A tree $T \in F_i$ can be formed by a two-step process: (1) choosing an i -element subset B of A ; and (2) choosing a tree in $R(B)$ for T_ℓ and a tree in $R(Z \setminus B)$ for T_r . There are $\binom{k}{i}$ choices for the subset B in step (1), and therefore,

$$|F_i| = \binom{k}{i} |R(B)| |R(Z \setminus B)| = \binom{k}{i} b(i)b(n - i + 1).$$

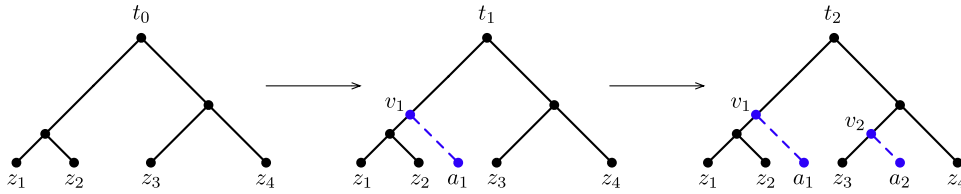


Fig. 2. An illustration for the proof of [Corollary 5](#). A binary, rooted tree is constructed on the taxon set $Z = A \cup \{z_1, z_2, z_3, z_4\}$, where $A = \{a_1, a_2\}$. The left subtree contains at least one leaf in $Z \setminus A$. We first choose a binary, rooted tree t_0 on $Z \setminus A$, and then sequentially add leaves a_1 and a_2 to t_0 .

Because $\{F_i \mid 1 \leq i \leq k\}$ is a partition of $\mathcal{T}_Z(A)$,

$$|\mathcal{T}_Z(A)| = \sum_{i=1}^k |F_i| = \sum_{i=1}^k \binom{k}{i} b(i)b(n-i+1). \tag{24}$$

On the other hand, we can also count the number of trees in $R(Z) \setminus \mathcal{T}_Z(A)$, that is, trees whose left subtrees have at least one leaf not in A . Enumerate the elements of A by a_1, \dots, a_k . A tree in $R(Z) \setminus \mathcal{T}_Z(A)$ can then be constructed as follows ([Fig. 2](#)). We first choose a tree $t_0 \in R(Z \setminus A)$. Having built t_i , for $i = 0, \dots, k-1$, we create t_{i+1} by bisecting an edge of t_i with a 2-degree node v_{i+1} , and then attaching leaf a_{i+1} to v_{i+1} .

It is clear that $t_k \in R(Z)$. We show that $t_k \notin \mathcal{T}_Z(A)$. The left subtree of t_0 has at least one leaf not in A , as $t_0 \in R(Z \setminus A)$. By the process just described, leaf a_{i+1} is added to a subtree of t_i in each step i . Consequently, the left subtree of t_k contains all the leaves of the left subtree of t_0 . Hence, it has at least one leaf not in A , that is, $t_k \notin \mathcal{T}_Z(A)$.

There are $b(n-k+1)$ choices for the tree $t_0 \in R(Z \setminus A)$. Tree t_i has $n-k+i+1$ leaves, and so it has $2n-2k+2i$ edges. Therefore, there are $2n-2k+2i$ trees t_{i+1} that can be built from t_i by joining leaf a_{i+1} to an edge of t_i . It follows that the number of trees in $R(Z) \setminus \mathcal{T}_Z(A)$ is

$$b(n+1) - |\mathcal{T}_Z(A)| = b(n-k+1) \prod_{i=0}^{k-1} (2n-2k+2i) = \frac{(2n-2)!!}{(2n-2k-2)!!} b(n-k+1). \tag{25}$$

Eqs. (24) and (25) imply [Eq. \(23\)](#).

[Eq. \(22\)](#) can now be simplified using [Eq. \(23\)](#):

$$\begin{aligned} \overline{dc}_s^u(S) &= -(2n-2) + 2\text{epl}(S) - \sum_{e \in E(S)} \left(\frac{b(n+1)}{b(n)} - \frac{(2n-2)!!}{b(n)} \frac{b(n-|C_S(e)|+1)}{(2n-2|C_S(e)|-2)!!} \right) \\ &= -2n(2n-2) + 2\text{epl}(S) + \frac{(2n-2)!!}{(2n-3)!!} \sum_{e \in E(S)} \frac{(2n-2|C_S(e)|-1)!!}{(2n-2|C_S(e)|-2)!!}. \quad \square \end{aligned}$$

4.2. Numerical investigation of $\overline{dc}_s(S)$























































We note from [Eq. \(17\)](#) that $\overline{dc}_s(S)$ of a species tree S depends on the shape of S , but not on the specific leaf labeling of S . We now investigate the relationship between $\overline{dc}_s(S)$ and the Furnas rank of S [9], in which tree shapes are assigned consecutive positive integers, $\text{rank}_F(S)$, starting from 1. The rank of a tree depends first on the number of leaves in its left subtree: trees with fewer leaves in their left subtrees have smaller ranks. Next, for trees with the same number of leaves in their left subtrees, their relative order is determined by the ranks of their left subtrees. Finally, for trees with identical left subtrees, the ranks of their right subtrees determine their relative rank. Generally, trees with higher rank_F appear more balanced than trees with smaller rank_F (e.g. [13]). In particular, rank 1 is always assigned to caterpillar trees, and the largest ranks are assigned to trees in which for each internal node, the numbers of leaves in the left and right subtrees of the node differ by at most one. For brevity, we refer to the k -leaf caterpillar and the k -leaf tree shape with the highest Furnas rank as T_k^c and T_k^b , respectively.

The values of $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$ for all species tree shapes with up to nine leaves are given in [Table 1](#), and plots of the two functions for the 46 species tree shapes with nine leaves appear in [Fig. 3\(a\)](#). We first observe that both $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$ follow a similar pattern, with slightly higher values for $\overline{dc}_s^u(S)$. They generally decrease as $\text{rank}_F(S)$ increases, with T_9^c having a high value, and T_9^b having a low value. However, the plots in [Fig. 3\(a\)](#) have jumps that usually occur between consecutive trees S_1 and S_2 in which the left subtree of S_1 has one leaf fewer than the left subtree of S_2 . Note then that the left and right subtrees of S_1 are T_k^b and T_{9-k}^b , where k is the number of leaves in S_1 , while the left and right subtrees of S_2 are T_{k+1}^c and T_{8-k}^c .

[Fig. 3\(b\)](#) plots the Sackin index for the 46 species tree shapes against their Furnas ranks. The figure shows that the Sackin index exhibits a pattern very similar to the pattern observed in either $\overline{dc}_s^y(S)$ or $\overline{dc}_s^u(S)$. This observation can be partly explained by the fact that the formula for $\overline{dc}_s(S)$ in [Eq. \(17\)](#) contains the term $2\text{epl}(S) = 2n\bar{\ell}(S)$, where $\bar{\ell}(S)$ is the

Table 1

The values of $\overline{dc}_i^u(S)$ and $\overline{dc}_i^v(S)$ for an arbitrarily chosen labeling of fixed species tree shapes with $1 \leq n \leq 9$ leaves. For each species tree shape S , these quantities appear as an ordered pair $(\overline{dc}_i^v(S); \overline{dc}_i^u(S))$. Species tree shapes are ordered according to their Furnas ranks.

$n \leq 5$		0.00; 0.00		0.00; 0.67		1.94; 2.00		1.56; 1.60		3.23; 3.37		3.08; 3.20
	$n = 6$		6.28; 6.67		5.59; 5.94		5.24; 5.49		4.56; 4.76		4.83; 5.02	
$n = 7$		9.32; 10.00		8.58; 9.21		7.78; 8.26		7.03; 7.47		7.23; 7.65		6.48; 6.74
	$n = 8$		12.95; 14.00		12.16; 13.17		11.11; 11.97		10.32; 11.15		11.00; 11.76	
$n = 9$			10.63; 11.32		10.00; 10.63		8.79; 9.30		8.95; 9.43		9.47; 9.91	
	$n = 9$		17.16; 18.67		16.34; 17.82		15.13; 16.46		14.31; 15.61		14.82; 16.05	
$n = 9$			14.03; 15.11		13.34; 14.36		12.00; 12.90		12.12; 13.01		12.51; 13.35	
	$n = 9$		14.60; 15.66		13.91; 14.91		12.57; 13.45		12.70; 13.56		12.27; 13.04	
$n = 9$			12.41; 13.06		11.89; 12.45		11.20; 11.70		13.09; 13.64		11.58; 12.04	

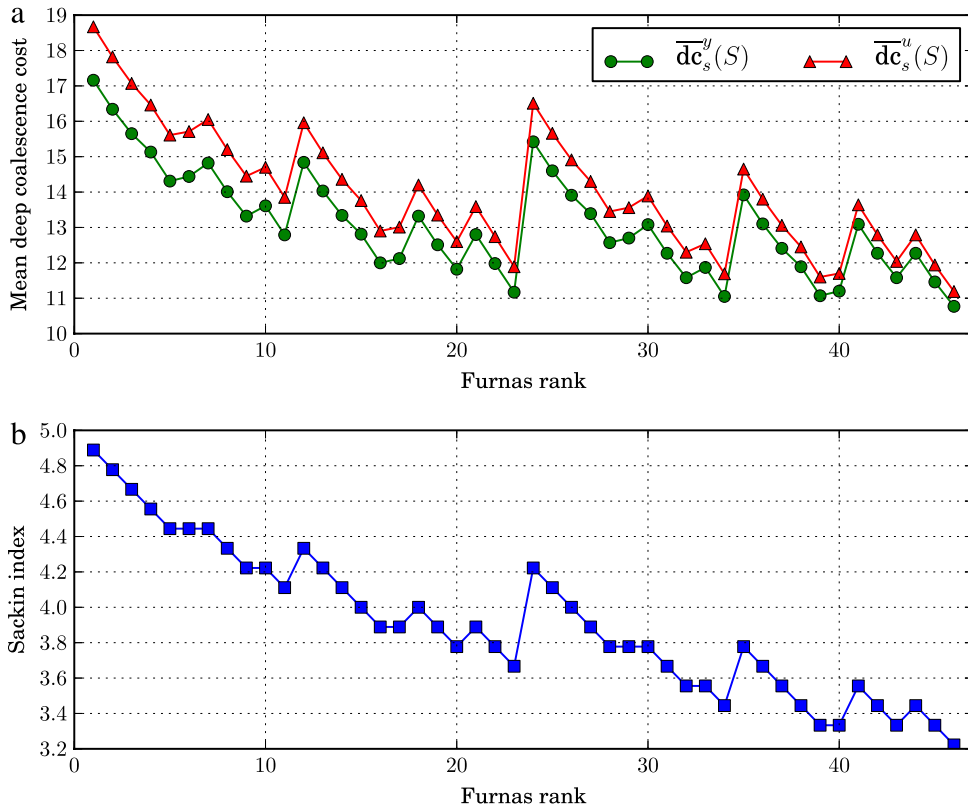


Fig. 3. Mean deep coalescence costs and the Sackin index for the 46 species tree shapes with nine leaves. Figure (a): $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$, computed using Eqs. (20) and (21). Figure (b): the Sackin index, computed using Eq. (5). The species tree shapes are ordered according to their Furnas ranks.

Sackin index of S . The formula and the plots in Fig. 3(a) and (b) demonstrate that a close connection exists between $\overline{dc}_s(S)$ and $\bar{\ell}(S)$. Note, however, that the connection is complicated, as it depends on the sum $\sum_{e \in E(S)} \sum_{i=1}^{|C_S(e)|} p_n(i) \binom{|C_S(e)|}{i}$.

5. Mean deep coalescence cost for a fixed gene tree

In this section, we deal with the problem of evaluating the mean deep coalescence cost for a given gene tree $T \in R(X)$, averaging over all possible species trees. By Eq. (16),

$$\begin{aligned}
 \overline{dc}_t(T) &= \sum_{S \in R(X)} \mathbb{P}(S) dc(T, S) \\
 &= \sum_{S \in R(X)} \mathbb{P}(S) \left(-(2n - 2) + 2\text{epI}(S) - \sum_{e \in E(S)} \sum_{v \in V(T)} [C_T(v) \subseteq C_S(e)] \right) \\
 &= -(2n - 2) + 2\mathbb{E}[\text{epI}(S)] - \sum_{v \in V(T)} \sum_{S \in R(X)} \sum_{e \in E(S)} \mathbb{P}(S) [C_T(v) \subseteq C_S(e)].
 \end{aligned}
 \tag{26}$$

With the assumption that the probability distribution on $R(X)$ has the exchangeability property, we can write $\mathbb{E}[\text{epI}(S)]$ and the sum

$$f(T, v) = \sum_{S \in R(X)} \sum_{e \in E(S)} \mathbb{P}(S) [C_T(v) \subseteq C_S(e)]$$

in terms of cluster probabilities, as in the following two lemmas.

Lemma 6. *If the probability distribution on $R(X)$ has the exchangeability property, then*

$$\mathbb{E}[\text{epI}(S)] = -n + \sum_{i=1}^n \binom{n}{i} i p_n(i).
 \tag{27}$$

Proof. Using the formula for $\text{epl}(S)$ in Eq. (4), we have

$$\begin{aligned} \mathbb{E}[\text{epl}(S)] &= \sum_{S \in R(X)} \mathbb{P}(S) \text{epl}(S) \\ &= \sum_{S \in R(X)} \mathbb{P}(S) \left(-n + \sum_{v \in V(S)} |C_S(v)| \right) = -n + \sum_{S \in R(X)} \sum_{v \in V(S)} \mathbb{P}(S) |C_S(v)|. \end{aligned}$$

Let $\mathcal{P}_0(X)$ be the collection of all nonempty subsets of X , and let F_i be the collection of subsets of X that have exactly i elements, $1 \leq i \leq n$. Then

$$\begin{aligned} \sum_{S \in R(X)} \sum_{v \in V(S)} \mathbb{P}(S) |C_S(v)| &= \sum_{S \in R(X)} \mathbb{P}(S) \sum_{A \in \mathcal{P}_0(X)} [A \in \mathcal{C}(S)] |A| \\ &= \sum_{i=1}^n \sum_{A \in F_i} |A| \sum_{S \in R(X)} \mathbb{P}(S) [A \in \mathcal{C}(S)]. \end{aligned}$$

The term $\sum_{S \in R(X)} \mathbb{P}(S) [A \in \mathcal{C}(S)]$ is the probability that A is a cluster of a tree in $R(X)$, and by assumption, it is $p_n(i)$ for every $A \in F_i$. Hence,

$$\begin{aligned} \mathbb{E}[\text{epl}(S)] &= -n + \sum_{i=1}^n \sum_{A \in F_i} |A| \sum_{S \in R(X)} \mathbb{P}(S) [A \in \mathcal{C}(S)] \\ &= -n + \sum_{i=1}^n \sum_{A \in F_i} i p_n(i) = -n + \sum_{i=1}^n \binom{n}{i} i p_n(i). \quad \square \end{aligned}$$

Lemma 7. If the probability distribution on $R(X)$ has the exchangeability property, then

$$f(T, v) = \sum_{S \in R(X)} \sum_{e \in E(S)} \mathbb{P}(S) [C_T(v) \subseteq C_S(e)] = -1 + \sum_{i=|C_T(v)|}^n \binom{n - |C_T(v)|}{i - |C_T(v)|} p_n(i). \tag{28}$$

Proof. Let $\mathcal{P}_0(X)$ be the collection of all nonempty subsets of X . Observing that $\{C_S(e) \mid e \in E(S)\} = \mathcal{C}(S) \setminus \{X\}$, we rewrite $f(T, v)$ as

$$\begin{aligned} f(T, v) &= \sum_{S \in R(X)} \sum_{A \in \mathcal{P}_0(X) \setminus \{X\}} \mathbb{P}(S) [A \in \mathcal{C}(S)] [C_T(v) \subseteq A] \\ &= \sum_{A \in \mathcal{P}_0(X) \setminus \{X\}} [C_T(v) \subseteq A] \sum_{S \in R(X)} \mathbb{P}(S) [A \in \mathcal{C}(S)] = \sum_{A \in \mathcal{P}_0(X) \setminus \{X\}} \mathbb{P}_X(A) [C_T(v) \subseteq A] \\ &= -1 + \sum_{A \in \mathcal{P}_0(X)} \mathbb{P}_X(A) [C_T(v) \subseteq A] = -1 + \sum_{i=1}^n p_n(i) \sum_{A \in F_i} [C_T(v) \subseteq A], \end{aligned}$$

where F_i is the collection of all i -element subsets of X . Note that in the last step, we use the assumption that $\mathbb{P}_X(A) = p_n(i)$ for every A in F_i .

The sum $\sum_{A \in F_i} [C_T(v) \subseteq A]$ is the number of elements of F_i that contain $C_T(v)$ as a subset. It is clear that if $1 \leq i \leq |C_T(v)| - 1$, then the summand is zero. If $i \geq |C_T(v)|$, then the sum is equal to $\binom{n - |C_T(v)|}{i - |C_T(v)|}$. For if $C_T(v) \subseteq A$, then elements in $A \setminus C_T(v)$ are chosen from $X \setminus C_T(v)$. Hence,

$$f(T, v) = -1 + \sum_{i=|C_T(v)|}^n \binom{n - |C_T(v)|}{i - |C_T(v)|} p_n(i). \quad \square$$

Theorem 8. If the probability distribution on $R(X)$ has the exchangeability property, then

$$\overline{\text{dc}}_t(T) = -(2n - 1) + 2 \sum_{i=1}^n \binom{n}{i} i p_n(i) - \sum_{v \in V(T)} \sum_{i=|C_T(v)|}^n \binom{n - |C_T(v)|}{i - |C_T(v)|} p_n(i). \tag{29}$$

Proof. The theorem is a direct consequence of Eqs. (26)–(28). \square

We next provide an upper-bound for $\overline{dc}_t(T)$, considering all gene trees T (Theorem 10). The proof of the theorem relies on the following lemma.

Lemma 9. *Let ϕ be a strictly decreasing function on $[2, n]$. Then for any tree $T \in R(X)$, where $|X| = n$,*

$$\sum_{v \in \dot{V}(T)} \phi(|C_T(v)|) \geq \sum_{k=2}^n \phi(k), \tag{30}$$

with equality if and only if T is a caterpillar tree.

Proof. We assign integers $2, \dots, n$ to the $n - 1$ internal nodes of T in postorder. For each $v \in \dot{V}(T)$, let D_v be the set of nodes in $\dot{V}(T)$ that are proper descendants of v . Then node v is labeled only after all the nodes in D_v have been labeled. Since T is binary and rooted, $|D_v| = |C_T(v)| - 2$, and it follows that at least $|D_v|$ numbers $2, \dots, |C_T(v)| - 1$ have been used for labeling. Consequently, the next available number for labeling v is at least $|C_T(v)|$. In other words, if we denote by $\kappa(v)$ the number assigned to v by the postorder labeling procedure, then we always have $\kappa(v) \geq |C_T(v)|$ for any $v \in \dot{V}(T)$. Since ϕ is decreasing, $\phi(|C_T(v)|) \geq \phi(\kappa(v))$, and hence

$$\sum_{v \in \dot{V}(T)} \phi(|C_T(v)|) \geq \sum_{v \in \dot{V}(T)} \phi(\kappa(v)).$$

As each $v \in \dot{V}(T)$ is assigned a unique integer among $2, \dots, n$, we have $\{\kappa(v) \mid v \in \dot{V}(T)\} = \{2, \dots, n\}$. Thus, the right-hand side of the last equation is equal to $\sum_{k=2}^n \phi(k)$. Eq. (30) now follows.

In order for the equality in Eq. (30) to hold, we must have $\phi(|C_T(v)|) = \phi(\kappa(v))$ for every $v \in \dot{V}(T)$. Because ϕ is strictly decreasing, $|C_T(v)| = \kappa(v)$ for every $v \in \dot{V}(T)$. This in turn implies that T induces for each $k = 2, \dots, n$ a subtree that has k leaves, which occurs if and only if T is a caterpillar tree. \square

Theorem 10. *Assume that the probability distribution on $R(X)$ has the exchangeability property, and that $p_n(i) > 0$ for every $1 \leq i \leq n$. Then*

$$\overline{dc}_t(T) \leq -(n - 1) + \sum_{i=2}^n \left[\binom{n}{i} i - \binom{n-1}{i-2} \right] p_n(i), \tag{31}$$

with equality if and only if T is a caterpillar tree.

Proof. For a positive integer $2 \leq k \leq n$, define

$$\phi(k) = \sum_{i=k}^n \binom{n-k}{i-k} p_n(i).$$

If $2 \leq k \leq n - 1$, then

$$\begin{aligned} \phi(k) &> \sum_{i=k+1}^n \binom{n-k}{i-k} p_n(i) = \sum_{i=k+1}^n \frac{n-k}{i-k} \binom{n-k-1}{i-k-1} p_n(i) \\ &\geq \sum_{i=k+1}^n \binom{n-k-1}{i-k-1} p_n(i) = \phi(k+1), \end{aligned}$$

that is, the function ϕ is strictly decreasing on $[2, n]$. By Lemma 9,

$$\begin{aligned} \sum_{v \in \dot{V}(T)} \phi(|C_T(v)|) &\geq \sum_{k=2}^n \phi(k) = \sum_{k=2}^n \sum_{i=k}^n \binom{n-k}{i-k} p_n(i) \\ &= \sum_{i=2}^n p_n(i) \sum_{k=2}^i \binom{n-k}{i-k} = \sum_{i=2}^n p_n(i) \sum_{k=2}^i \binom{n-k}{n-i} \\ &= \sum_{i=2}^n \binom{n-1}{i-2} p_n(i), \end{aligned} \tag{32}$$

where in the last step we use the identity $\sum_{r=p}^q \binom{r}{p} = \binom{q+1}{q-p}$ (e.g. Eq. (5.10), p. 160, [10]).

From Eqs. (29) and (32), we have

$$\begin{aligned}\overline{dc}_t(T) &= -(2n-1) + 2 \sum_{i=1}^n \binom{n}{i} i p_n(i) - \left(n \sum_{i=1}^n \binom{n-1}{i-1} p_n(i) + \sum_{v \in V(T)} \phi(|C_T(v)|) \right) \\ &\leq -(2n-1) + \sum_{i=1}^n \binom{n}{i} i p_n(i) - \sum_{i=2}^n \binom{n-1}{i-2} p_n(i) \\ &= -(n-1) + \sum_{i=2}^n \left[\binom{n}{i} i - \binom{n-1}{i-2} \right] p_n(i).\end{aligned}$$

The equality in the last equation holds if and only if the equality in Eqs. (32) holds, which by Lemma 9, occurs if and only if T is a caterpillar tree. \square

5.1. $\overline{dc}_t(T)$ in the Yule and uniform models

We derive in this section formulas for $\overline{dc}_t(T)$ and its upper bound in the Yule and uniform models.

Corollary 11. Let $H_n = \sum_{i=1}^n 1/n$, and assume the Yule model on $R(X)$. Then the mean deep coalescence cost for a fixed gene tree $T \in R(X)$ is

$$\overline{dc}_t^y(T) = -(2n-2) + 4n(H_n - 1) - \sum_{v \in V(T)} \sum_{i=|C_T(v)|}^{n-1} \frac{2n}{i(i+1)} \binom{n}{i}^{-1} \binom{n-|C_T(v)|}{i-|C_T(v)|}. \quad (33)$$

Further,

$$\overline{dc}_t^y(T) \leq -(2n-2) + \left(2n-2 + \frac{8}{n+2} \right) (H_n - 1). \quad (34)$$

Proof. Eqs. (33) and (34) can be obtained by substituting the formula for $p_n(i)$ in the Yule model (Eq. (12)) into Eqs. (29) and (31), respectively. The derivation of (34) involves several additional steps:

$$\begin{aligned}\overline{dc}_t^y(T) &\leq -(n-1) + (n - (n-1)) + \sum_{i=2}^{n-1} \left(\binom{n}{i} i - \binom{n-1}{i-2} \right) \frac{2n}{i(i+1)} \binom{n}{i}^{-1} \\ &= -(n-2) + 2n(H_n - 3/2) - \sum_{i=2}^{n-1} \frac{2(i-1)}{(i+1)(n-i+1)} \\ &= -(n-2) + 2n(H_n - 3/2) - \frac{2n}{n+2} \sum_{i=2}^{n-1} \frac{1}{n-i+1} + \frac{4}{n+2} \sum_{i=2}^{n-1} \frac{1}{i+1} \\ &= -(2n-2) + \left(2n-2 + \frac{8}{n+2} \right) (H_n - 1). \quad \square\end{aligned}$$

As for the derivation of $\overline{dc}_t(T)$ and its upper bound in the uniform model, we make use of the following lemma, whose proof will be presented shortly.

Lemma 12. Let k and n be positive integers, where $k \leq n$. Then

$$\sum_{i=k}^n \binom{n-k}{i-k} b(i) b(n-i+1) = \frac{(2n-2)!!}{(2k-2)!!} b(k). \quad (35)$$

Corollary 13. Assume the uniform model on $R(X)$. Then the mean deep coalescence cost for a fixed gene tree $T \in R(X)$ is

$$\overline{dc}_t^u(T) = -(2n-1) + 2n \frac{(2n-2)!!}{(2n-3)!!} - \frac{(2n-2)!!}{(2n-3)!!} \sum_{v \in V(T)} \frac{(2|C_T(v)|-3)!!}{(2|C_T(v)|-2)!!}. \quad (36)$$

Further,

$$\overline{dc}_t^u(T) \leq -2(2n-1) + (n+1) \frac{(2n-2)!!}{(2n-3)!!}. \quad (37)$$

Proof. Recall that in the uniform model, $p_n(i) = b(i)b(n - i + 1)/b(n)$, where $b(i) = (2i - 3)!!$ (Eq. (10)). Thus, we can rewrite the second term in the formula for $\overline{dc}_t(T)$ in Eq. (29) as

$$2 \sum_{i=1}^n \binom{n}{i} i p_n(i) = 2n \sum_{i=1}^n \binom{n-1}{i-1} \frac{b(i)b(n-i+1)}{b(n)} = 2n \frac{(2n-2)!!}{(2n-3)!!},$$

where in the last step we use Lemma 12. Similarly, by applying the lemma on the third term of the formula for $\overline{dc}_t(T)$ in Eq. (29), we have

$$\begin{aligned} \sum_{v \in V(T)} \sum_{i=|C_T(v)|}^n \binom{n-|C_T(v)|}{i-|C_T(v)|} p_n(i) &= \sum_{v \in V(T)} \sum_{i=|C_T(v)|}^n \binom{n-|C_T(v)|}{i-|C_T(v)|} \frac{b(i)b(n-i+1)}{b(n)} \\ &= \frac{(2n-2)!!}{(2n-3)!!} \sum_{v \in V(T)} \frac{(2|C_T(v)|-3)!!}{(2|C_T(v)|-2)!!}. \end{aligned}$$

Eq. (36) now follows.

We can obtain an upper bound of $\overline{dc}_t^u(T)$ either by using the upper bound in Eq. (31) with $p_n(i)$ replaced by $b(i)b(n - i + 1)/b(n)$, or by applying Lemma 9 directly on the formula for $\overline{dc}_t^u(T)$ that we have just derived. The latter approach is employed here.

For a positive integer k , let $\phi(k) = (2k - 3)!!/(2k - 2)!!$. Then

$$\frac{\phi(k+1)}{\phi(k)} = \frac{(2k-1)!!(2k-2)!!}{(2k-3)!!(2k)!!} = \frac{2k-1}{2k} < 1,$$

that is, $\phi(k)$ is strictly decreasing on the set of positive integers. Thus, by Lemma 9,

$$\begin{aligned} \overline{dc}_t^u(T) &= -(2n-1) + 2n \frac{(2n-2)!!}{(2n-3)!!} - \frac{(2n-2)!!}{(2n-3)!!} \left(n + \sum_{v \in V(T)} \phi(|C_T(v)|) \right) \\ &\leq -(2n-1) + n \frac{(2n-2)!!}{(2n-3)!!} - \frac{(2n-2)!!}{(2n-3)!!} \sum_{k=2}^n \phi(k) \\ &= -(2n-1) + (n+1) \frac{(2n-2)!!}{(2n-3)!!} - \frac{(2n-2)!!}{(2n-3)!!} \sum_{k=0}^{n-1} (-1)^k \binom{-1/2}{k}, \end{aligned}$$

where $\binom{x}{k}$ is defined as $x(x-1)\cdots(x-k+1)/k!$ for real x and nonnegative k . Then using the identity $\sum_{k=0}^n (-1)^k \binom{x}{k} = (-1)^n \binom{x-1}{n}$ (e.g. Eq. (5.16), p. 165, [10]), we have

$$\begin{aligned} \overline{dc}_t^u(T) &\leq -(2n-1) + (n+1) \frac{(2n-2)!!}{(2n-3)!!} - (-1)^{n-1} \frac{(2n-2)!!}{(2n-3)!!} \binom{-3/2}{n-1} \\ &= -2(2n-1) + (n+1) \frac{(2n-2)!!}{(2n-3)!!}. \quad \square \end{aligned}$$

Proof of Lemma 12. We now return to the proof of Lemma 12. The proof presented here employs a similar idea as in the proof of Eq. (23): counting a certain class of binary, rooted trees in two different ways. Let X be a taxon set of cardinality n , and let z be a distinguishing taxon name not appearing in X . Further, let A be a fixed subset of X of cardinality k . Denote by $\mathcal{F}_X(A, z)$ the set of trees $T \in R(X \cup \{z\})$ that have the following two properties:

- (PL) Every element of A appears in the left subtree T_ℓ of T .
- (PR) Taxon z appears in the right subtree T_r of T .

Note that T_ℓ can have leaves from $X \setminus A$, but none of the elements of A appears in T_r . For a tree $T \in \mathcal{F}_X(A, z)$, the leaf set of T_ℓ has at least k leaves (by PL) and at most n leaves (by PR). Let $F_i = \{T \in \mathcal{F}_X(A, z) \mid T_\ell \text{ has exactly } i \text{ leaves}\}$, $k \leq i \leq n$. A tree $T \in F_i$ can be formed in two steps: (1) choosing an i -element subset B of X that contains A ; (2) choosing a tree in $R(B)$ to be T_ℓ and a tree in $R((X \setminus B) \cup \{z\})$ to be T_r . The elements of $B \setminus A$ are chosen from $X \setminus A$, and hence, there are $\binom{n-k}{i-k}$ different choices for the set B in step (1). Therefore,

$$|F_i| = \binom{n-k}{i-k} |R(B)| |R((X \setminus B) \cup \{z\})| = \binom{n-k}{i-k} b(i)b(n-i+1).$$

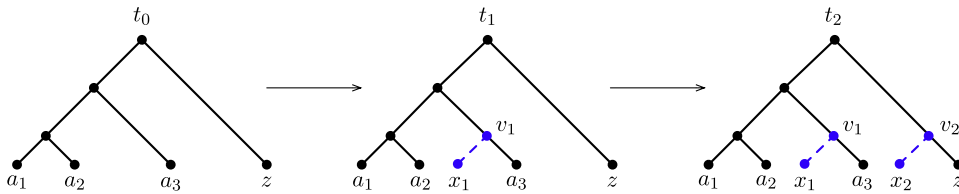


Fig. 4. An illustration for the proof of Lemma 12. A binary, rooted tree is constructed on the taxon set $X = A \cup \{x_1, x_2, z\}$, where $A = \{a_1, a_2, a_3\}$. The left subtree contains all the elements of A and the right subtree contains z . We first choose a binary, rooted tree t_0 on $A \cup \{z\}$, and then sequentially add leaves x_1 and x_2 to t_0 .

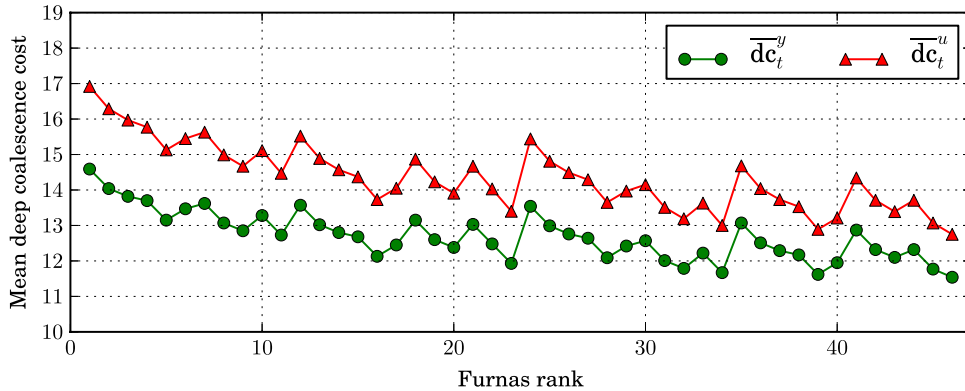


Fig. 5. Mean deep coalescence costs for the 46 gene tree shapes with nine leaves. The values of $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$ are computed using Eqs. (33) and (36). The gene tree shapes are ordered according to their Furnas ranks.

Because $\{F_i \mid k \leq i \leq n\}$ is a partition of $\mathcal{F}_X(A, z)$, we have

$$|\mathcal{F}_X(A, z)| = \sum_{i=k}^n |F_i| = \sum_{i=k}^n \binom{n-k}{i-k} b(i)b(n-i+1). \tag{38}$$

On the other hand, we can construct a tree $T \in \mathcal{F}_X(A, z)$ as follows. Enumerate the elements of $X \setminus A$ by x_1, \dots, x_{n-k} . We choose a tree $t_0 \in R(A \cup \{z\})$ such that the left subtree of t_0 is a tree in $R(A)$ (and so z is the only leaf in the right subtree of t_0). Having built t_i , for $i = 0, \dots, n-k-1$, we create t_{i+1} by bisecting an edge of t_i with a 2-degree node v_{i+1} , and then attaching leaf x_{i+1} to v_{i+1} (Fig. 4). It can be seen that $t_{n-k} \in R(X \cup \{z\})$ and satisfies both properties PL and PR. Thus, $t_{n-k} \in \mathcal{F}_X(A, z)$.

Since the left subtree of t_0 is a tree in $R(A)$, there are $|R(A)| = b(k)$ different choices for t_0 . Tree t_i has $k+i+1$ leaves, and so it has $2k+2i$ edges. Therefore, there are $2k+2i$ possible trees t_{i+1} that can be constructed from t_i by joining leaf x_{i+1} to an edge of t_i . It follows that

$$|\mathcal{F}_X(A, z)| = b(k) \prod_{i=0}^{n-k-1} (2k+2i) = \frac{(2n-2)!!}{(2k-2)!!} b(k). \tag{39}$$

Eqs. (38) and (39) imply Eq. (35). \square

5.2. Numerical investigation of $\overline{dc}_t(T)$

The values of \overline{dc}_t^y and \overline{dc}_t^u for all gene tree shapes with up to nine leaves are given in Table 2, and plots of the two functions appear in Fig. 5. Both $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$ generally decrease with increasing values of the Furnas rank. As is demonstrated by the bounds in Eqs. (34) and (37), the highest values of the mean deep coalescence cost occur for the caterpillar tree. The lowest values occur for trees with high Furnas ranks. In the same manner as in the case of $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$, jumps occur in $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$ at Furnas ranks where the number of leaves in the left subtree changes. Also as in the case of means over species trees, the mean deep coalescence costs across gene trees are larger for the uniform model than for the Yule model.

Both for the Yule and uniform models, we have

$$\sum_S \mathbb{P}(S) \overline{dc}_s(S) = \sum_S \mathbb{P}(S) \sum_T \mathbb{P}(T) dc(T, S) = \sum_T \mathbb{P}(T) \sum_S \mathbb{P}(S) dc(T, S) = \sum_T \mathbb{P}(T) \overline{dc}_t(T).$$

In other words, the mean across all 46 species trees of the values in Fig. 3 must equal the mean across gene trees of the values in Fig. 5, as each quantity represents a mean over both gene trees and species trees. However, the values of $\overline{dc}_t^y(T)$

Table 2

The values of $\overrightarrow{dc}_i^r(T)$ and $\overleftarrow{dc}_i^r(T)$ for an arbitrarily chosen labeling of fixed gene tree shapes with $1 \leq n \leq 9$ leaves. For each gene tree shape T , these quantities appear as an ordered pair $(\overrightarrow{dc}_i^r(T); \overleftarrow{dc}_i^r(T))$. Gene tree shapes are ordered according to their Furnas ranks.

$n \leq 5$																			
	0.00; 0.00	0.00; 0.00	0.67; 0.67	1.94; 2.00	1.56; 1.60	3.73; 3.94	3.30; 3.49	3.12; 3.26											
$n = 6$																			
	5.95; 6.44	5.48; 5.94	5.28; 5.68	5.18; 5.52	4.71; 5.02	4.98; 5.27													
$n = 7$																			
	8.53; 9.46	8.03; 8.91	7.82; 8.63	7.71; 8.46	7.21; 7.90	7.50; 8.18	7.14; 7.78												
$n = 8$																			
	11.42; 12.97	10.90; 12.37	10.68; 12.07	10.56; 11.88	10.04; 11.29	10.35; 11.59	9.96; 11.16	9.75; 10.86											
$n = 9$																			
	14.59; 16.92	14.04; 16.29	13.82; 15.97	13.70; 15.77	13.15; 15.13	13.47; 15.45	13.07; 14.99	12.85; 14.67	12.73; 15.11										
	13.02; 14.89	12.80; 14.57	12.68; 14.37	12.13; 13.73	12.45; 14.05	13.15; 14.87	12.60; 14.23	13.03; 14.67	12.48; 14.03										
	12.99; 14.81	12.76; 14.49	12.64; 14.29	12.09; 13.65	12.42; 13.97	12.57; 14.15	12.01; 13.51	12.22; 13.63	11.67; 13.00										
	12.29; 13.73	12.17; 13.53	11.62; 12.89	11.95; 13.21	12.87; 14.34	12.32; 13.71	12.10; 13.39	11.77; 13.07	11.54; 12.75										

and $\overline{dc}_t^u(T)$ in Fig. 5 show a number of differences from the corresponding values of $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$ in Fig. 3(a). First, the range of the values is greater for $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$, with both the largest and the smallest values falling outside the range of $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$. Second, as might be expected from the difference in range, the variability across trees in mean deep coalescence costs is greater for $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$ than for $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$. Despite this greater variability, however, the difference between corresponding values for the Yule and uniform models is smaller for $\overline{dc}_s^y(S)$ and $\overline{dc}_s^u(S)$ than for $\overline{dc}_t^y(T)$ and $\overline{dc}_t^u(T)$.

6. Discussion

Our results on the mean deep coalescence cost have a series of parallels with our previous results on the maximum deep coalescence cost [26]. We have observed that both for fixed species trees and for fixed gene trees, the mean and maximum deep coalescence costs are highest when the fixed tree is a caterpillar. More generally, the mean and maximum tend to decrease with increasing tree balance, as measured by the Furnas rank. Both the mean and maximum appear to vary to a greater extent across fixed species trees than across fixed gene trees.

The mean deep coalescence cost provides a natural way of normalizing deep coalescence costs in evolutionary studies. As we have previously argued [26], in searching for a species tree that has minimal deep coalescence cost summed across a given set of fixed gene trees, a method might be applied that penalizes candidate species trees that naturally generate higher deep coalescence costs (e.g. caterpillars) less severely than candidate species trees with lower deep coalescence costs (e.g. balanced trees). Adapting the minimization criterion through normalizations involving either the maximum or mean deep coalescence cost might help to eliminate a bias toward inference of balanced trees that has been both predicted under evolutionary models [25] and observed in the analysis of genetic sequences [6]. Investigation of such normalized criteria involving our results on the maximum and mean provides an important direction for future work on the deep coalescence cost.

Acknowledgments

The authors acknowledge grant support from the National Science Foundation (DBI-1146722) and the Burroughs Wellcome Fund.

References

- [1] D. Aldous, Probability distributions on cladograms, in: D. Aldous, R. Pemantle (Eds.), *Random Discrete Structures*, in: IMA Volumes in Mathematics and its Applications, vol. 76, Springer-Verlag, New York, 1996, pp. 1–18.
- [2] M.S. Bansal, J.G. Burleigh, O. Eulenstein, Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models, *BMC Bioinformatics* 11 (2010) S42.
- [3] M.G.B. Blum, O. François, On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited, *Math. Biosci.* 195 (2005) 141–153.
- [4] J.K.M. Brown, Probabilities of evolutionary trees, *Syst. Biol.* 43 (1994) 78–91.
- [5] L.L. Cavalli-Sforza, A.W.F. Edwards, Phylogenetic analysis: models and estimation procedures, *Am. J. Hum. Genet.* 19 (1967) 233–257.
- [6] M. DeGiorgio, J. Syring, A.J. Eckert, A.I. Liston, R. Cronn, D.B. Neale, N.A. Rosenberg, An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines, *BMC Evol. Biol.* (2014) in press.
- [7] J.H. Degnan, N.A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent, *Trends Ecol. Evol.* 24 (2009) 332–340.
- [8] J. Felsenstein, The number of evolutionary trees, *Syst. Zool.* 27 (1978) 27–33.
- [9] G.W. Furnas, The generation of random, binary unordered trees, *J. Classification* 1 (1984) 187–233.
- [10] R.L. Graham, D.E. Knuth, O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, second ed., Addison-Wesley, Reading, Massachusetts, 1994.
- [11] E.F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, *Adv. Appl. Probab.* 3 (1971) 44–77.
- [12] J.F.C. Kingman, On the genealogy of large populations, *J. Appl. Probab.* 19A (1982) 27–43.
- [13] M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, *Evolution* 47 (1993) 1171–1181.
- [14] R. Klein, D. Wood, On the path length of binary trees, *J. Assoc. Comput. Mach.* 36 (1989) 280–289.
- [15] D.E. Knuth, *The Art of Computer Programming Vol. 1: Fundamental Algorithms*, third ed., Addison-Wesley, Reading, Massachusetts, 1997.
- [16] H.T. Lin, J.G. Burleigh, O. Eulenstein, The Deep Coalescence Consensus Tree Problem is Pareto on Clusters, in: *Lecture Notes in Bioinformatics*, vol. 6674, 2011, pp. 172–183.
- [17] W.P. Maddison, Gene trees in species trees, *Syst. Biol.* 46 (1997) 523–536.
- [18] W.P. Maddison, L.L. Knowles, Inferring phylogeny despite incomplete lineage sorting, *Syst. Biol.* 55 (2006) 21–30.
- [19] P. Pamilo, M. Nei, Relationships between gene trees and species trees, *Mol. Biol. Evol.* 5 (1988) 568–583.
- [20] J. Phipps, The numbers of classifications, *Can. J. Bot.* 54 (1976) 686–688.
- [21] N.A. Rosenberg, The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model, *Evolution* 57 (2003) 1465–1477.
- [22] M.J. Sackin, “Good” and “bad” phenograms, *Syst. Zool.* 21 (1972) 225–226.
- [23] M. Steel, A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, *Math. Biosci.* 170 (2001) 91–112.
- [24] C. Than, L. Nakhleh, Species tree inference by minimizing deep coalescences, *PLoS Comput. Biol.* 5 (2009) e1000501.
- [25] C.V. Than, N.A. Rosenberg, Consistency properties of species tree inference by minimizing deep coalescences, *J. Comput. Biol.* 18 (2011) 1–15.
- [26] C.V. Than, N.A. Rosenberg, Mathematical properties of the deep coalescence cost, *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10 (2013) 61–72.
- [27] G.U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S., *Philos. Trans. R. Soc. Lond. Ser. B* 213 (1925) 21–87.
- [28] L. Zhang, From gene trees to species trees II: species tree inference by minimizing deep coalescence events, *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8 (2011) 1685–1691.