Bull Math Biol (2019) 81:384–407
https://doi.org/10.1007/s11538-017-0342-x

Society for Mathematical Biology

SPECIAL ISSUE: ALGEBRAIC METHODS IN PHYLOGENETICS

On the Number of Non-equivalent Ancestral Configurations for Matching Gene Trees and Species Trees

Filippo Disanto[1,2] · **Noah A. Rosenberg**[1]

Received: 15 March 2017 / Accepted: 31 August 2017 / Published online: 14 September 2017

Abstract An *ancestral configuration* is one of the combinatorially distinct sets of gene lineages that, for a given gene tree, can reach a given node of a specified species tree. Ancestral configurations have appeared in recursive algebraic computations of the conditional probability that a gene tree topology is produced under the multispecies coalescent model for a given species tree. For matching gene trees and species trees, we study the number of ancestral configurations, considered up to an equivalence relation introduced by Wu (Evolution 66:763–775, 2012) to reduce the complexity of the recursive probability computation. We examine the largest number of non-equivalent ancestral configurations possible for a given tree size n. Whereas the smallest number of non-equivalent ancestral configurations increases polynomially with n, we show that the largest number increases with k^n, where k is a constant that satisfies $\sqrt[3]{3} \leq k < 1.503$. Under a uniform distribution on the set of binary labeled trees with a given size n, the mean number of non-equivalent ancestral configurations grows exponentially with n. The results refine an earlier analysis of the number of ancestral configurations considered without applying the equivalence relation, showing that use of the equivalence relation does not alter the exponential nature of the increase with tree size.

Keywords Ancestral configurations · Combinatorics · Gene trees and species trees · Phylogenetics

✉ Filippo Disanto
filippo.disanto@unipi.it

[1] Department of Biology, Stanford University, Stanford, CA, USA

[2] Department of Mathematics, University of Pisa, Pisa, Italy

1 Introduction

Under the multispecies coalescent model for the evolution of gene trees conditional on species trees, symmetries and identities among gene tree probabilities and algebraic perspectives for examining the probability computations have contributed to advances in understanding the properties of evolutionary descent in closely related species (Allman et al. 2011). Calculations of the probabilities of gene tree topologies can proceed by one of two computational approaches: non-recursive (Degnan and Salter 2005) or recursive (Wu 2012). Both methods involve combinatorial and probabilistic components, in which probabilities are evaluated for each element of a set of objects that can be defined purely in mathematical terms. Computational complexity is affected both by the size of the underlying set of objects and by the complexity of the probability calculation.

In the recursive approach, the relevant combinatorial set consists of *ancestral configurations*, each of which represents a set of gene lineages that can be extant at a given node of the species tree (Wu 2012). We have previously studied the set of ancestral configurations possible for a given gene tree and matching species tree, showing that the largest number of ancestral configurations across labeled tree topologies of a fixed tree size n increases exponentially with n (Disanto and Rosenberg 2017).

To lower the computation time of the recursive evaluation of gene tree probabilities, Wu (2012) introduced an equivalence relation that, taking into account symmetries in tree shapes, reduces the set of ancestral configurations to a potentially much smaller set of *non-equivalent ancestral configurations*. The computation of gene tree probabilities can then make use of intermediate steps calculated for the elements of this smaller set, rather than for the full set of ancestral configurations.

Here, for gene trees and species trees with a matching labeled topology t, we study the number of *non-equivalent* ancestral configurations that can appear at the nodes of a species tree t. We determine the number of non-equivalent ancestral configurations when t belongs to special families of trees characterized by balanced and unbalanced patterns. We study the largest number of non-equivalent ancestral configurations possible for a given tree size n, showing that this number grows exponentially with k^n, where k is a constant that satisfies $\sqrt[3]{3} \leq k < 1.503$. Although tree families exist for which the number of non-equivalent ancestral configurations grows polynomially in n (Wu 2012), we show that under a uniform distribution on the set of labeled trees of size n, the mean number of non-equivalent ancestral configurations of a random labeled tree shape also grows exponentially in n. Finally, we compare our results on the number of non-equivalent ancestral configurations with corresponding results for the full set of ancestral configurations (Disanto and Rosenberg 2017). Although by definition, the non-equivalent ancestral configurations are no more numerous than ancestral configurations that do not take into account the equivalence relation—and indeed, are intended to be less numerous—the base k for the maximal number of non-equivalent ancestral configurations k^n across trees of size n is bounded below by a constant only slightly smaller than the corresponding base for the maximal number of ancestral configurations.

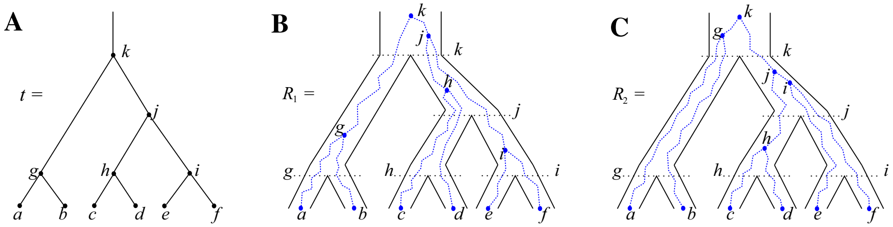

Fig. 1 A matching gene tree and species tree with labeled topology t. **a** A tree t of size 6 isomorphic to the gene tree and species tree in **b** and **c**. Tree t is uniquely determined by the labeling of its leaves and by its unlabeled shape. It is convenient to assign arbitrary labels to the internal nodes of t as well. We use letters g, h, i, j, k in this case. Each lineage (edge) of t is identified by the lowest node it intersects; for example, lineages h and i descend from lineage j. **b** A possible realization R_1 of a gene tree (*dotted lines*) in a species tree (*solid lines*). The gene tree and the species tree have a matching topology that follows (**a**). At species tree node j, the ancestral configuration is $\{c, d, i\}$. At node k, the configuration is $\{g, h, i\}$. **c** A non-equivalent realization R_2 of the gene tree in **a** in the matching species tree. At species tree nodes j and k, the configurations are $\{h, e, f\}$ and $\{a, b, j\}$, respectively

2 Preliminaries

We study the number of non-equivalent ancestral configurations of rooted binary labeled trees. We start by giving definitions and preliminary results. In Sect. 2.1, we recall some properties of rooted binary labeled trees. In Sect. 2.2, we discuss properties of the exponential growth of sequences of nonnegative numbers. Following Wu (2012), Sect. 2.3 defines ancestral configurations for a gene tree and a species tree with a matching labeled topology t. In Sect. 2.4, we recall related enumerative results of Disanto and Rosenberg (2017).

2.1 Labeled Topologies

A labeled topology t of size $|t| = n$ is a bifurcating rooted tree with n labeled leaves, also termed "taxa" (Fig. 1a). We sometimes refer to labeled topologies simply as "trees." We define a total order $a \prec b \prec c \prec \ldots$ for the set $\{a, b, c, \ldots\}$ of labels of the leaves of a tree, proceeding alphabetically. That is, without loss of generality, we assume that a tree of size n has its taxa labeled using the first n symbols that appear in the order $\prec$.

We represent labeled topologies in Newick notation (Felsenstein 2004), in which $t = (t_1, t_2)$ is the tree obtained by appending trees t_1 and t_2 to a common root node. For example, $((a, b), ((c, d), (e, f)))$ gives the Newick notation for the tree depicted in Fig. 1a. We term non-leaf nodes of a tree "internal" nodes. By "subtree" of a tree t, we mean a node of t together with all its descendants; a "root subtree" of t is a subtree—one of two possible—immediately descended from the root of t.

For two trees t_1, t_2, we say that t_1 is isomorphic to t_2 and write $t_1 \cong t_2$ when, after their leaf labels are removed, t_1 and t_2 have the same unlabeled topology. Moreover, given trees t_1 and t_2 with $|t_1| \geq |t_2|$, we say that a subtree t of t_1 is equal to t_2 up to "rescaling" labels when, respecting the order $\prec$, we can replace the labels of t to obtain t_2. For instance, the largest root subtree $((c, d), (e, f))$ of the tree depicted in Fig. 1a is equal to $((a, b), (c, d))$ up to rescaling, as we can replace the labels $c \to a, d \to b, e \to c, f \to d$. Note that alphabetical order is preserved in this replacement.

We denote by T_n the set of trees of size n and by $T = \bigcup_{n=1}^{\infty} T_n$ the set of all trees of any size. The number of trees of size $n \geq 2$ is given by

$$|T_n| = (2n-3)!! = 1 \times 3 \times 5 \times \cdots \times (2n-3) \tag{1}$$

(Felsenstein 1978), which assuming $n \geq 1$ can be rewritten

$$|T_n| = \frac{(2n-2)!}{2^{n-1}(n-1)!} = \frac{(2n)!}{2^n(2n-1)n!}. \tag{2}$$

We will have occasion to employ a uniform probability distribution over the set of trees of fixed size. In this distribution, each tree of size n has probability $1/|T_n|$.

2.2 Exponential Growth of a Sequence

As in Flajolet and Sedgewick (2009), we say that a sequence of positive numbers a_n is of exponential order k or, equivalently, has exponential growth k^n, when

$$\limsup_{n\to\infty}[(a_n)^{1/n}] = \lim_{n\to\infty}[\sup_{m\geq n}[(a_m)^{1/m}]] = k.$$

This relation holds when $a_n = k^n s(n)$, where s is a subexponential factor, so that $\limsup_{n\to\infty}[s(n)^{1/n}] = 1$. According to these definitions, a sequence a_n grows exponentially in n if its exponential order strictly exceeds 1.

The exponential order of a sequence describes its asymptotic growth. It follows from the definition that if (a_n) has exponential order k_a and (b_n) has exponential order $k_b > k_a$, then a_n/b_n converges to 0 exponentially fast as $(k_a/k_b)^n$ for $n \to \infty$. When two sequences (a_n) and (b_n) have the same exponential order, we write $a_n \bowtie b_n$. If $a_n \bowtie b_n$ and $\lim_{n\to\infty}(a_n/b_n) = 1$, we write $a_n \sim b_n$.

2.3 Ancestral Configurations

This section defines the set of ancestral configurations of a gene tree G in a species tree S. In our setting, exactly one gene lineage is selected from each species. We assume a matching labeled topology t for G and S.

Consider a realization R of a gene tree G in a species tree S, with $G = S = t$ (Fig. 1). Equivalently, R is one of the possible evolutionary scenarios for gene tree G on species tree S. Given a node κ of t, we denote by $C(\kappa, R)$ the set of gene lineages, i.e., edges of G, that are present in S at the point right before node κ looking backward in time. Following Wu (2012), we call the set $C(\kappa, R)$ the *ancestral configuration* of G at node κ of S.

For the tree t in Fig. 1a, if we consider the realization R_1 of the gene tree $G = t$ in the species tree $S = t$ depicted in Fig. 1b, then we see that $C(k, R_1) = \{g, h, i\}$ is the ancestral configuration of the gene tree at node k of the species tree. The gene lineages g, h, and i are those present in the species tree at the point right before the root node k. Similarly, the ancestral configuration of the gene tree at node j of the species tree is given by the set of gene lineages $C(j, R_1) = \{c, d, i\}$. In Fig. 1c, a different realization R_2 of the same gene tree is described. The ancestral configuration at the root k of the species tree is in this case $C(k, R_2) = \{a, b, j\}$, whereas the ancestral configuration at node j is $C(j, R_2) = \{h, e, f\}$.

We denote the set of all possible realizations of the gene tree $G = t$ in the species tree $S = t$ by $\Re(G, S)$. By considering all elements $R \in \Re(G, S)$, for a given node κ of t we define the set of all possible ancestral configurations at node κ,

$$C(\kappa) = \{C(\kappa, R) : R \in \Re(G, S)\}, \tag{3}$$

and the number of such configurations,

$$c(\kappa) = |C(\kappa)|. \tag{4}$$

In particular, $c(\kappa)$ counts the number of ways the gene lineages of G can reach the point right below node κ in S, when all possible realizations of G in S are taken into account. For example, if we set t as in Fig. 1a, then we have $C(g) = \{\{a, b\}\}$ and $C(j) = \{\{c, d, e, f\}, \{h, e, f\}, \{c, d, i\}, \{h, i\}\}$. At the root node k, the set of all possible ancestral configurations is

$$C(k) = \{\{g, j\}, \{a, b, j\}, \{g, c, d, e, f\}, \{a, b, c, d, e, f\}, \{g, h, e, f\}, \{a, b, h, e, f\},\\ \{g, c, d, i\}, \{a, b, c, d, i\}, \{g, h, i\}, \{a, b, h, i\}\}.$$

Note that two different realizations $R_1, R_2 \in \Re(G, S)$ can generate the same ancestral configuration $C(\kappa, R_1) = C(\kappa, R_2)$ at an internal node κ.

Following Disanto and Rosenberg (2017), for each internal node κ, our definition of ancestral configuration excludes the case $\{\kappa\} \in C(\kappa)$. This choice accords with the fact that each configuration at node κ is considered at the point right below node κ in the species tree, with no time for the gene lineages from the left and right subtrees of κ to coalesce together. With the exception that we say that a leaf or 1-taxon tree has 0 ancestral configurations, our definition is identical to that of Wu (2012), which assigns these cases 1 ancestral configuration.

Under our assumption of a matching gene tree and species tree $G = S = t$, the set $C(\kappa)$ defined in (3) and its cardinality $c(\kappa)$ (4) depend only on node κ and tree t. When we refer to an element of $C(\kappa)$, we use the term *configuration at node κ of t*. When κ is the root node, we use the term *root configuration* to describe an element of $C(\kappa)$. Also, considering the union of all the sets $C(\kappa)$ of configurations across all internal nodes κ of t, we can count the *total* number of configurations.

2.4 The Number of Configurations

We recall some of the results of Disanto and Rosenberg (2017) on the number of configurations possessed by a tree. These results are used to measure the decrease in the number of configurations when, as in Wu (2012), an equivalence relation is introduced in Sect. 3 to merge topologically equivalent configurations.

(i) If A, B are two sets of sets, define $A \otimes B = \{a \cup b : a \in A, b \in B\}$. For a given tree t with $|t| > 1$, the set $C(r)$ of configurations at the root r of t satisfies the following decomposition

$$C(r) = \{\{r_\ell, r_r\}\} \cup \big[C(r_\ell) \otimes \{\{r_r\}\}\big] \cup \big[\{\{r_\ell\}\} \otimes C(r_r)\big] \cup \big[C(r_\ell) \otimes C(r_r)\big],$$

where r_ℓ and r_r, respectively, denote the left and right children of r.

(ii) For a given tree t with $|t| > 1$, the number $c(r)$ of possible configurations at the root node r of t can be recursively computed as

$$c(r) = [c(r_\ell) + 1][c(r_r) + 1] = 1 + c(r_\ell) + c(r_r) + c(r_\ell)\, c(r_r), \qquad (5)$$

where we set $c(r) = 0$ when $|t| = 1$. At each node κ of t, the number of configurations $c(\kappa)$ is bounded as $c(\kappa) \le c(r)$. Thus, the total number of configurations $c = \sum_\kappa c(\kappa)$ satisfies $c(r) \le c \le (2|t| - 1)c(r)$. In particular, the quantities c and $c(r)$ are equal up to a factor that is at most polynomial in $|t|$, and they have the same exponential order when measured across families of trees of increasing size.

(iii) Denote by $M_n(r)$ and M_n, respectively, the largest number of root configurations and the largest total number of configurations that a tree of size n can have. The exponential growth of the sequences $M_n(r)$ and M_n is $M_n(r) \bowtie M_n \bowtie k_0^n$, where k_0 is a constant, $k_0 \approx 1.5028$.

(iv) A completely balanced tree of size $n = 2^h$ has $\lfloor k_0^n \rfloor - 1$ root configurations. A caterpillar tree of size n has $n - 1$ root configurations.

(v) For a tree of given size n leaves selected uniformly at random, the mean number of root configurations $c(r)$ and the mean total number of configurations c have exponential growth $\mathbb{E}_n[c(r)] \bowtie \mathbb{E}_n[c] \bowtie (4/3)^n$ with n.

3 Equivalent and Non-equivalent Configurations

Wu (2012) introduced an equivalence relation over the set of configurations at a given node of a species tree, using this equivalence relation to evaluate the probability of a gene tree topology by performing computations over the sets of non-equivalent configurations of the gene tree at species tree nodes (e.g., Eq. (7) of Wu (2012)). Following the definition of Wu (2012), in this section, we introduce the notion of equivalent configurations for gene trees and species trees with matching topology t. Under certain assumptions on t, in Sect. 3.3, we provide a recursion analogous to the one in (5) for counting non-equivalent configurations at the root of t.

3.1 An Equivalence Relation

We begin with some notation. If κ is a node of a tree t, denote by t_κ the subtree of t generated by κ (i.e., κ and all nodes below it). If X is a set of nodes of a subtree t_κ, the restriction $t_\kappa(X)$ of t_κ to X is the tree shape obtained by removing from t_κ all nodes that remain *strictly* below the nodes belonging to X. For instance, if t_j is the subtree generated by node j in the tree t in Fig. 1a and $X = \{h, e, f\}$, then $t_j(X)$ is obtained by removing nodes c and d from t_j, and thus is the caterpillar tree shape of size 3. Similarly, if $X = \{a, b, h, i\}$, then $t_k(X)$ is the balanced tree shape of size 4.

The definition of equivalent configurations given by Wu (2012) reduces to the following one when gene trees and species trees are matching. Given a tree t and a node κ, two configurations γ_1, γ_2 at node κ, $\gamma_1, \gamma_2 \in C(\kappa)$, are *equivalent* at κ—with the equivalence denoted by $\gamma_1 \sim_\kappa \gamma_2$—when the tree shape $t_\kappa(\gamma_1)$ is isomorphic to the tree shape $t_\kappa(\gamma_2)$. For instance, in Fig. 1a, we have $\{h, e, f\} \sim_j \{c, d, i\}$ and $\{a, b, j\} \sim_k \{g, h, i\}$. The set of non-equivalent configurations at a given node κ is denoted by $C^*(\kappa)$, and its cardinality is $c^*(\kappa) = |C^*(\kappa)|$.

The notion of equivalent configurations groups together at a given node configurations for which exactly the same topological constraints apply in ordering the coalescent events of their gene lineages. In other words, gene lineages of equivalent configurations at a node κ of a species tree have completely topologically equivalent transitions when they move from node κ backward in time (upward in the species tree).

For instance, consider the tree in Fig. 1a, where the configurations $\{a, b, j\}$ and $\{g, h, i\}$ at node k satisfy $\{a, b, j\} \sim_k \{g, h, i\}$. Consider the mapping $\phi(a) = h, \phi(b) = i, \phi(j) = g, \phi(g) = j, \phi(k) = k$. The transition in Fig. 1c that along the root branch of the species tree transforms the set of gene lineages $\{a, b, j\}$ into the single lineage k corresponds topologically to the transition in Fig. 1b that transforms $\{g, h, i\}$ into k. Indeed, the two trees $t_k(\{a, b, j\})$ with nodes $\{a, b, j, g, k\}$ and $t_k(\{g, h, i\})$ with nodes $\{g, h, i, j, k\}$ are isomorphic through ϕ.

As described in Fig. 2, for a given tree t, the effective computation of non-equivalent configurations can be performed recursively as in the algorithm STELLS (Wu 2012)

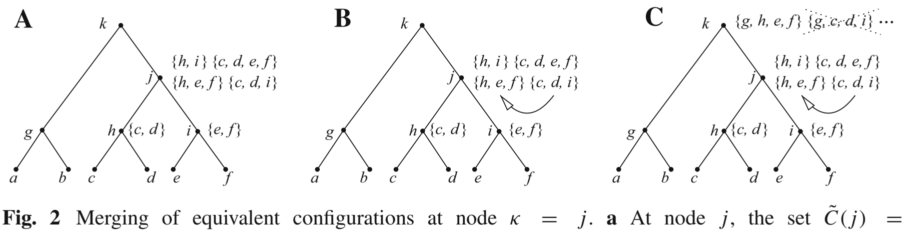

Fig. 2 Merging of equivalent configurations at node $\kappa = j$. **a** At node j, the set $\tilde{C}(j) = \{\{h, i\}, \{h, e, f\}, \{c, d, i\}, \{c, d, e, f\}\}$ of configurations is computed from the non-equivalent configurations at the child nodes h and i by using (6). **b** Two equivalent configurations appear in $\tilde{C}(j)$, namely $\{h, e, f\} \sim_j \{c, d, i\}$. Configuration $\{c, d, i\}$ is merged into $\{h, e, f\}$ (or vice versa). **c** The configurations in $C^*(j) = \{\{h, e, f\}, \{h, i\}, \{c, d, e, f\}\}$ are used to determine configurations at node k. In particular, $\{g, h, e, f\} \in \tilde{C}(k)$ and $\{g, c, d, i\} \notin \tilde{C}(k)$, as $\{c, d, i\}$ has been merged into $\{h, e, f\}$. Configuration $\{g, c, d, i\}$, which is not present in $\tilde{C}(k)$, is represented by the equivalent configuration $\{g, h, e, f\} \sim_k \{g, c, d, i\}$. Similarly, $\{a, b, c, d, i\} \notin \tilde{C}(k)$, and it is represented by $\{a, b, h, e, f\} \sim_k \{a, b, c, d, i\}$

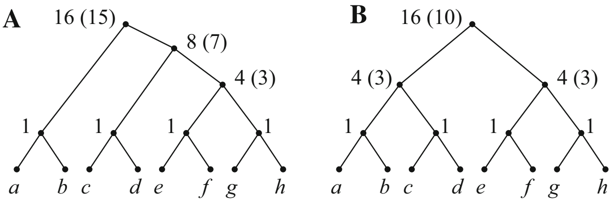

Fig. 3 Computing the number of non-equivalent configurations in two trees of size 8. By using (7), at each internal node κ, $|\tilde{C}(\kappa)|$ is computed from the number of non-equivalent configurations at the nodes descending from κ. When $|\tilde{C}(\kappa)| > c^*(\kappa)$, $c^*(\kappa)$ appears in parentheses. **a** A tree considered in Table A1 by Wu (2012). Adding $|t| = 8$ to the value $\sum_\kappa |\tilde{C}(\kappa)| = 32$ to take into account the fact that Wu (2012) counts a configuration for each leaf, whereas our definition does not do so, we produce entry 40 of the table of Wu (2012). **b** The completely balanced tree of size 8 considered in Table 1 by Wu (2012). Adding $|t| = 8$ to $\sum_\kappa |\tilde{C}(\kappa)| = 28$, we produce entry 36. The numbers $c^*(\kappa)$ satisfy recursion (10)

by scanning t from bottom to top with a postorder traversal. At each visited node κ, we first compute the set

$$\tilde{C}(\kappa) = \{\{\kappa_\ell, \kappa_r\}\} \cup \left[C^*(\kappa_\ell) \otimes \{\{\kappa_r\}\}\right] \cup \left[\{\{\kappa_\ell\}\} \otimes C^*(\kappa_r)\right] \cup \left[C^*(\kappa_\ell) \otimes C^*(\kappa_r)\right] \quad (6)$$

from the sets of non-equivalent configurations of the two child nodes κ_ℓ, κ_r (Fig. 2a with $\kappa = j$). Next, we merge all the equivalent configurations present in $\tilde{C}(\kappa)$ into a single representative, one for each class of equivalence of the relation $\sim_\kappa$, to determine the set $C^*(\kappa)$ of non-equivalent configurations at κ (Fig. 2b). Only the configurations in $C^*(\kappa)$ are used to determine configurations at the parent node of κ (Fig. 2c). Note that from (6), the cardinality of the set $\tilde{C}(\kappa) \supseteq C^*(\kappa)$ satisfies

$$c^*(\kappa) \leq |\tilde{C}(\kappa)| = 1 + c^*(\kappa_\ell) + c^*(\kappa_r) + c^*(\kappa_\ell)\, c^*(\kappa_r). \quad (7)$$

Following this procedure in Fig. 3, we report the quantities $|\tilde{C}(\kappa)|$ and $c^*(\kappa)$ at each internal node κ of two trees of size 8. When $|\tilde{C}(\kappa)| > c^*(\kappa)$, the latter value is given in parentheses. The same trees are considered in the enumerations provided in Table A1 (Fig. 3a) and Table 1 (Fig. 3b) by Wu (2012).

In the next sections, we study the number $c^*(\kappa) = |C^*(\kappa)|$ of pairwise non-equivalent configurations at a given node κ of a fixed or random tree $t \in T_n$ selected uniformly as well as the total number of non-equivalent configurations $c^* = \sum_\kappa c^*(\kappa)$ in t. To measure the strength of the equivalence relation $\sim_\kappa$, we focus on $c^*(r)$, the number of non-equivalent configurations at the root $\kappa = r$ of t, comparing our results with those in Sect. 2.4.

When there is no need to distinguish between the number of non-equivalent root configurations and the total number of non-equivalent configurations, we simply write "number of non-equivalent configurations." It is then understood that a statement applies to both root and total non-equivalent configurations. Similarly, "number of configurations" stands for both "number of root configurations" and "total number of configurations."

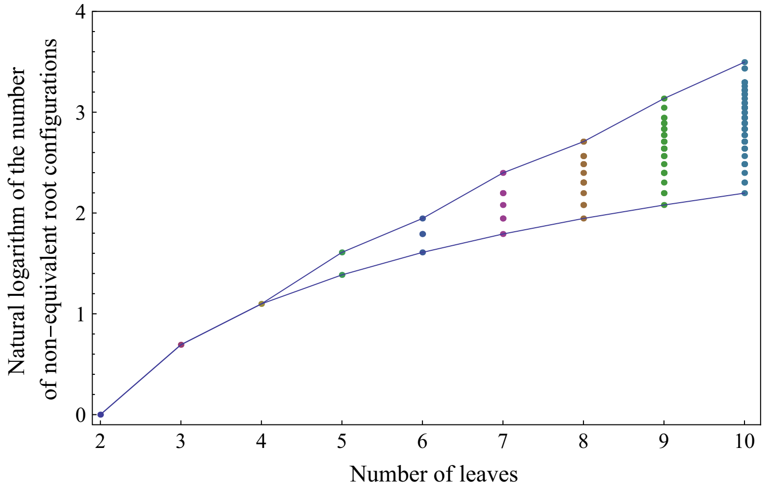

Fig. 4 Natural logarithm of the number of non-equivalent root configurations for all possible tree shapes of size $2 \leq n \leq 10$. The value for $n = 1$, $\log(0)$, is omitted. Points corresponding to the largest and smallest numbers of root configurations for each n are connected by the *top* and *bottom lines*, respectively

3.2 Non-equivalent Root Configurations in Small Trees

For small values of n, it is possible to exhaustively compute the number of non-equivalent root configurations $c^*(r)$ for representative labelings of each of the unlabeled topologies of size n. In Fig. 4, each dot corresponds to the logarithm of the number of non-equivalent root configurations for a certain tree shape of size determined by its x-coordinate. The points associated with the largest values of $c^*(r)$ are connected by the top line, whose growth appears to be linear in n. Indeed, as we show in Sect. 4, tree families exist for which the growth of the number of non-equivalent root configurations is exponential in the tree size.

The tree shapes whose labeled topologies possess the largest number of non-equivalent root configurations among trees of fixed size $n \leq 20$ appear in Fig. 5. For $12 \leq n \leq 20$, each shape in the sequence is produced by connecting the tree with three taxa and the tree of size $n-3$ already in the sequence to a shared root. This pattern is used in Sect. 4.3 to determine a lower bound for the exponential growth of the sequence $M_n^*(r)$ describing the largest number of non-equivalent root configurations among trees at fixed n.

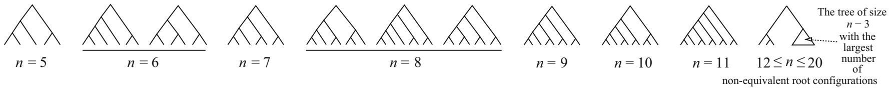

Fig. 5 Tree shapes of size $5 \leq n \leq 20$ with the largest number of non-equivalent root configurations. For $n = 4$, both unlabeled topologies have $c^*(r) = 3$. For $12 \leq n \leq 20$, the tree with the largest value of $c^*(r)$ is obtained by appending a caterpillar of size 3 and the tree of size $n-3$ with the largest value of $c^*(r)$ to a common root node. From $n = 2$ to $n = 20$, the largest values of $c^*(r)$ follow the sequence 1, 2, 3, 5, 7, 11, 15, 23, 33, 47, 69, 99, 141, 207, 297, 423, 621, 891, 1269

For values of $n \le 20$, the tree shape that minimizes the number of non-equivalent root configurations is the caterpillar topology. The number of non-equivalent root configurations in the caterpillar of size n is $n - 1$ (Wu 2012). The bottom line in Fig. 4, which connects points corresponding to the smallest number of non-equivalent root configurations for a tree with n taxa, grows with $\log(n - 1)$.

These observations show that tree topology can have a considerable impact on the number of non-equivalent configurations possible at a given tree size. Indeed, Sect. 4 investigates the effect of symmetries in a tree on its number of non-equivalent configurations. In Sect. 5, we show that although tree families (e.g., caterpillars) exist for which the growth of the number of non-equivalent configurations is polynomial in the tree size n, the expected number of non-equivalent configurations in a labeled topology selected uniformly at random in T_n grows exponentially in n.

3.3 A Recursion for the Number of Non-equivalent Root Configurations

In this section, we provide a recursive procedure for computing the number of non-equivalent root configurations in trees satisfying certain topological constraints. We later use this recursion to study the number of non-equivalent root configurations for several families of trees.

Let r be the root of a tree t. We denote by r_S and r_L the nodes descending from r that generate the smaller, t_{r_S}, and the larger, t_{r_L}, root subtrees of t (we will soon see that if the root subtrees of t have equal size, then we can choose either labeling). As depicted in Fig. 6, suppose subtree t_{r_S} can be displayed inside subtree t_{r_L} by a configuration at node r_L; that is, assume there is a configuration γ at node r_L such that

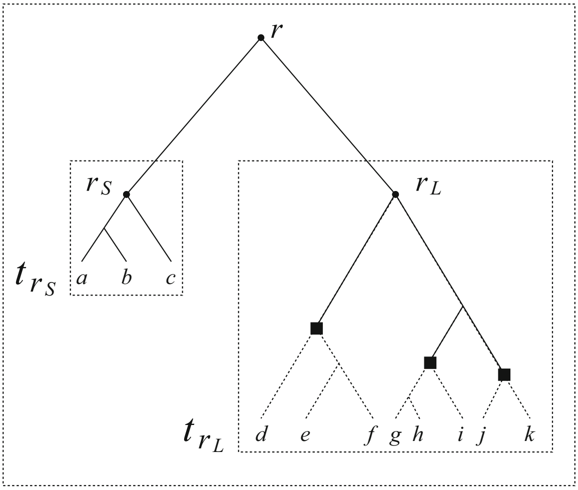

Fig. 6 A tree t in which the smaller root subtree t_{r_S} can be displayed as $t_{r_S} \cong t_{r_L}(\gamma)$ in the larger root subtree t_{r_L} through a configuration γ at node r_L. The configuration γ is determined by the *black squares*

$$t_{r_S} \cong t_{r_L}(\gamma). \tag{8}$$

Note that it immediately follows that when (8) is satisfied, if t_{r_S} and t_{r_L} have the same size, then they must have the same unlabeled shape, and it does not matter which is assigned the label t_{r_S} and which is assigned t_{r_L}. It is trivial that (8) is satisfied when $t_{r_S} \cong t_{r_L}$, by the configuration γ that simply consists of all leaves of t_{r_L}.

When condition (8) is satisfied, as shown in "Appendix 1," the number of non-equivalent configurations $c^*(r)$ at the root r of a tree t with $|t| > 1$ can be directly computed from the corresponding numbers at the children r_S and r_L:

$$\begin{aligned} c^*(r) &= [c^*(r_S)+1][c^*(r_L)+1] - \frac{c^*(r_S)}{2}[c^*(r_S)+1] \\ &= 1 + \frac{c^*(r_S)}{2} + c^*(r_L) + c^*(r_S)\,c^*(r_L) - \frac{[c^*(r_S)]^2}{2}, \end{aligned} \tag{9}$$

where $c^*(r) = 0$ if $|t| = 1$. Note that if the smaller root subtree has size $|t_{r_S}| = 1$, then condition (8) is technically not satisfied, as each configuration at node r_L has at least 2 elements (unless $|t| = 2$). However, in this case as well, with $|t_{r_S}| = 1$ and $c^*(r_S) = 0$, formula (9) holds, yielding $c^*(r) = 1 + c^*(r_L)$.

4 Non-equivalent Configurations for Special Tree Families

In this section, we study the number of non-equivalent configurations for special families of trees. We consider completely unbalanced caterpillar trees in Sect. 4.1 and completely balanced trees in Sect. 4.2. The number of non-equivalent configurations in the caterpillar family has been investigated by Wu (2012). For the completely balanced family, we show that the number of non-equivalent configurations grows exponentially in the tree size, though in a manner slower than the exponential growth of the number of configurations (see point (iv) in Sect. 2.4). By considering a particular family of unbalanced trees, in Sect. 4.3, we bound the exponential growth of the sequence $M_n^*(r)$ of the largest number of non-equivalent root configurations for a given tree size n.

4.1 Completely Unbalanced Trees

Consider the family of caterpillar trees. Recursive application of (9) shows that, as was already observed by Wu (2012), the number of non-equivalent root configurations in the caterpillar with n taxa is $n-1$. In particular, for caterpillar trees, t_{r_S} has only one leaf, and $c^*(r) = 1 + c^*(r_L)$. For a caterpillar tree of size n, subtree r_L is simply a caterpillar tree of size $n-1$. Noting that $c^*(r) = 1$ for a two-taxon caterpillar tree, we can iterate to obtain $c^*(r) = n-1$ for an n-taxon caterpillar tree. Considering all internal nodes of an n-taxon caterpillar, each of which has one fewer non-equivalent configuration than the number of leaves it subtends, the total number of non-equivalent configurations in the caterpillar of size n is $\sum_{k=2}^{n}(k-1) = n(n-1)/2$.

We have thus found a family of trees for which the growth of the number of non-equivalent configurations is polynomial in the tree size. This result suggests that

$\mathbb{E}_n[c^*(r)]$—the expected number of non-equivalent root configurations in a random tree selected uniformly among those of size n—could, in theory, grow as a subexponential function of n. We study the growth of this expectation in Sect. 5, showing that $\mathbb{E}_n[c^*(r)]$ in fact grows exponentially in n.

4.2 Completely Balanced Trees

Now consider the family of completely balanced trees $b_0, b_1, b_2, \ldots$, where b_h is the completely balanced tree of size $n = 2^h$ (Fig. 3b). Each tree b_h satisfies condition (8), as $t_{r_S} \cong t_{r_L}$. Because of this equivalence of unlabeled shapes, $c^*(r_S) = c^*(r_L)$. Therefore, denoting by γ_h the number of non-equivalent root configurations in b_h, from (9) we have the recursion

$$\gamma_{h+1} = \frac{\gamma_h^2}{2} + \frac{3\gamma_h}{2} + 1, \tag{10}$$

where $\gamma_0 = 0$. Setting $x_h = (\gamma_h + 1)/2$, this recursion can be written

$$x_{h+1} = x_h^2 + \frac{x_h + 1}{2}, \tag{11}$$

with $x_0 = 1/2$. The sequence (x_h) can be studied as in “Appendix 2.” A constant k_0^* exists for which

$$x_h \sim (k_0^*)^{(2^h)}. \tag{12}$$

The constant k_0^* can be approximated using the recursive definition of x_h, summing terms in a series

$$k_0^* = \left(\frac{1}{2}\right) \exp\left[\sum_{i=0}^{\infty} 2^{-i-1} \log\left(1 + \frac{1}{2x_i} + \frac{1}{2x_i^2}\right)\right] \approx 1.2460. \tag{13}$$

Switching back to γ_h, we obtain

$$\gamma_h = 2x_h - 1 \sim 2(k_0^*)^{(2^h)} = 2(k_0^*)^n,$$

where $n = 2^h = |b_h|$.

The following proposition summarizes our result.

Proposition 1 *Consider the family of completely balanced trees (b_h), with $n = 2^h = |b_h|$. Its sequence of the number of non-equivalent root configurations, $c^*(r)$, grows asymptotically as $c^*(r) \sim 2(k_0^*)^n$, where $k_0^* \approx 1.2460$ (13). In particular, $c^*(r)$ and the sequence of the total number of non-equivalent configurations, c^*, both have exponential growth $(k_0^*)^n$.*

Proof It remains to show that for tree family (b_h), the exponential growth of the total number of non-equivalent configurations equals the exponential growth of the number of non-equivalent root configurations. Because the sequence γ_h (10) is increasing, in the completely balanced tree b_h, the maximum number of non-equivalent configurations across all internal nodes is reached at the root of the tree, equaling $c^*(r)$. The total number of nodes (including the leaves) in b_h is $2n - 1$. We therefore have the inequality $c^*(r) \leq c^* \leq (2n-1)c^*(r)$. In particular, the quantities c^* and $c^*(r)$ are equal up to a factor that is at most polynomial in the size n. It follows that the exponential growth of c^* equals the exponential growth of $c^*(r)$. □

Comparing the constant k_0^* with the value of $k_0 \approx 1.5028$ that describes the exponential growth of the number of configurations for the completely balanced family of trees (Disanto and Rosenberg 2017), the proposition shows that in this family, the sequence of the number of non-equivalent configurations grows exponentially slower than the sequence of the number of configurations. However, the growth is still exponential in the tree size, and it is not true that non-equivalent configurations always grow polynomially—as they do for caterpillar trees.

4.3 Bounds for the Largest Number of Non-equivalent Configurations for a Given Tree Size

We now seek to bound the value of $M_n^*(r) = \max_{\{t:|t|=n\}} c_t^*(r)$, the largest number of non-equivalent root configurations among trees of size n.

Proposition 2 *Let $k_0 \approx 1.5028$ be the exponential order of the sequence $(M_n(r))$ describing the largest number of root configurations in trees of size n (point (iii) of Sect. 2.4). Then $M_n^*(r) \bowtie (k_1^*)^n$, where $\sqrt[3]{3} \leq k_1^* \leq k_0$.*

Proof For the upper bound, because non-equivalent configurations are no more numerous than configurations, $M_n^*(r) \leq M_n(r)$, and the upper bound follows.

For the lower bound, it suffices to exhibit a tree family in which the number of non-equivalent root configurations has exponential order $\sqrt[3]{3}$. For $n \geq 9$, we define the family of unlabeled topologies (u_n) by taking u_n as the tree shape of size n depicted in Fig. 5 if $n \in \{9, 10, 11\}$ and $u_n = (u_{n-3}, c_3)$—where c_3 is the caterpillar with 3 taxa—when $n \geq 12$. Note that for $n \geq 12$, the tree $t = u_n$ satisfies condition (8) with $t_{rS} = c_3$ (Fig. 6).

Let γ_n be the number of non-equivalent root configurations in u_n. For $n \geq 12$, (9) yields the recursion

$$\gamma_n = 3\gamma_{n-3}, \tag{14}$$

with $\gamma_9 = 23$, $\gamma_{10} = 33$, and $\gamma_{11} = 47$. We set $x_n = [2(n - 3\lfloor n/3 \rfloor)^2 + 8(n - 3\lfloor n/3 \rfloor) + 23]/27$ to produce a function that cycles through the values $23/27$, $33/27$, and $47/27$ as n is incremented. From (14), we have

$$\gamma_n = 3^{\lfloor \frac{n}{3} \rfloor} x_n \tag{15}$$

when $n \geq 9$. In particular, using (15), we see that (γ_n) has exponential growth $\gamma_n \bowtie \sqrt[3]{3}^n$ as desired. □

The recursive definition $u_n = (u_{n-3}, c_3)$ of the tree family (u_n) matches the pattern found by exhaustive computation for the unlabeled topologies of trees of size $12 \le n \le 20$ with the largest number of non-equivalent root configurations (Fig. 5). Applying the floor function to the expression in (15), we obtain

$$\left\lfloor 3^{\lfloor \frac{n}{3} \rfloor} \frac{2(n-3\lfloor \frac{n}{3} \rfloor)^2 + 8(n-3\lfloor \frac{n}{3} \rfloor) + 23}{27} \right\rfloor. \tag{16}$$

This formula, which equals (15) for $n \ge 9$, computes the correct values of $M_n^*(r)$ from Fig. 5 for $2 \le n \le 20$. Based on this result, it is a plausible conjecture that (16) gives the exact value for the maximum number of non-equivalent root configurations at a given $n \ge 2$.

Note that the constant k_1^* bounds from below the exponential order of the sequence M_n^* of the largest total number of non-equivalent configurations among trees of given size, as total non-equivalent configurations are at least as numerous as non-equivalent root configurations. Further, because k_0 is the exponential order of the sequence M_n of the largest total number of configurations in trees of fixed size (see point (iii) of Sect. 2.4), k_0 bounds from above the exponential order of the sequence M_n^*.

Because $\sqrt[3]{3} \approx 1.4422$, another consequence of Propositions 1 and 2 is that sequences $M_n^*(r)$ and M_n^* grow exponentially faster than the sequence of the number of non-equivalent configurations in the family of completely balanced trees. This property illustrates a remarkable effect of merging equivalent configurations. From points (iii) and (iv) of Sect. 2.4, the number of configurations for completely balanced trees follows the sequence of the largest number of configurations for trees of size n. When equivalent configurations are merged together, however, other tree families, such as the unbalanced family (u_n), possess a number of non-equivalent configurations that grows faster than the corresponding number for completely balanced trees.

5 Mean Number of Non-equivalent Root Configurations

We denote by $\mathbb{E}_n[c^*(r)]$ the expected number of non-equivalent root configurations in a random tree of size n drawn under a uniform distribution. This section shows that $\mathbb{E}_n[c^*(r)]$ grows as an exponential function of n. We first present a lower bound for $\mathbb{E}_n[c^*(r)]$. Next, we show that this lower bound is itself bounded below by a quantity that increases exponentially with n.

For the first step, we bound the expectation $\mathbb{E}_n[c^*(r)]$ by considering a certain set $T_n' \subseteq T_n$ in which each tree satisfies formula (9). For $n \ge 2$, define the quantity $x = x(n)$ as the solution of $2^{x-2} + x = n - 1$, and consider the function $w'(n)$ given by $w'(2) = 1$ and for $n \ge 3$,

$$w'(n) = \lfloor x \rfloor. \tag{17}$$

In "Appendix 3," it is shown that $w'(n)$ satisfies $w'(n) \le n/2$, and that $w'(n) = n/2$ holds only when $n = 2$, 4, or 6. For $2 \le n \le 10$, the values of $(n, w'(n))$ are (2, 1), (3, 1), (4, 2), (5, 2), (6, 3), (7, 3), (8, 3), (9, 4), and (10, 4).

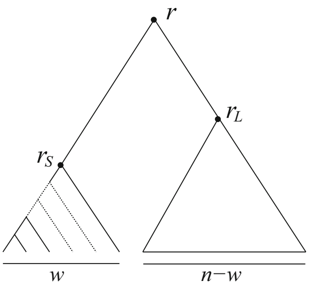

Fig. 7 Schematic representation of the unlabeled topology of a tree in set $T_{n,w}$. The smaller root subtree, t_{r_S}, is a caterpillar of size $w \in [1, w'(n)]$. The larger t_{r_L} has an unconstrained labeled topology of size $n - w$. The largest possible value of w, or $w'(n)$, is small enough for t_{r_S} to be displayed in t_{r_L}, as in (8). Note that $T_{n,w_1} \cap T_{n,w_2} = \emptyset$ if $w_1 \neq w_2$

The growth of $w'(n)$ is logarithmic. Indeed, for increasing values of n, the ratio x/n becomes small, so that $x-2 = \log_2[n(1-(x+1)/n)] \approx (\log_2 n) - (x+1)/(n \log 2)$, where the Taylor approximation $\log(1-u) \approx -u$ for u near 0 is used. We then obtain $x(n) \approx [n \log(4n) - 1]/(n \log 2 + 1) \sim (\log n)/(\log 2)$.

For a given $n \geq 2$ and a given $w \in [1, w'(n)]$, we denote by $T_{n,w}$ the set of trees of size n such that t_{r_S}, the smaller root subtree, is a caterpillar of size w, and t_{r_L}, the larger root subtree, has an unconstrained labeled topology of size $n - w$ (Fig. 7). For a given $n \geq 2$, we define the set of trees

$$T'_n = \bigcup_{w=1}^{w'(n)} T_{n,w}.$$

Four properties can be demonstrated for trees in $T_{n,w}$. (i) If $w \geq 2$, then each tree $t \in T_{n,w}$ satisfies (8) ("Appendix 4"), and thus, the number of non-equivalent root configurations in t satisfies (9). Furthermore, note that as was observed in Sect. 3.3, if $t \in T_{n,1}$, we have $c^*(r_S) = 0$, and (9) holds even though (8) does not.

(ii) For any fixed $n \geq 2$ and $w \in [1, w'(n)]$, with $w \neq n/2$, the probability of observing a given tree $\bar{t} \in T_{n-w}$ as the rescaled larger root subtree of a tree $t \in T_{n,w}$ selected uniformly at random is, as shown in "Appendix 5,"

$$\mathbb{P}[t_{r_L} = \bar{t} | t \in T_{n,w}] = \frac{1}{|T_{n-w}|}. \tag{18}$$

(iii) Because $\gamma_w = w!/(2 - \delta_{w,1})$ is the number of caterpillar trees of size $w \geq 1$ given a set of w labels, the probability $p_{n,w} = \mathbb{P}[t \in T_{n,w}]$ for a random tree of size n drawn under a uniform distribution to be in $T_{n,w}$ can be computed as $p_{n,w} = |T_{n,w}|/|T_n|$, or

$$\begin{aligned} p_{n,w} &= \binom{n}{w}\left[(1-\delta_{n,2w})\gamma_w|T_{n-w}| + \delta_{n,2w}\left(\gamma_w(|T_w|-\gamma_w) + \frac{1}{2}\gamma_w^2\right)\right]\Big/|T_n| \\ &= \frac{w!\binom{n}{w}[2(2-\delta_{w,1})(2n-2w-3)!!\,(1-\delta_{n,2w}) - \delta_{n,2w}(w!+2(2w-3)!!\,\delta_{w,1} - 4(2w-3)!!)]}{2(2n-3)!!\,(2-\delta_{w,1})^2}. \end{aligned} \tag{19}$$

Here, $\binom{n}{w}$ counts the number of ways of choosing the w taxa for the caterpillar subtree, and we have used (1) to expand $|T_n|$, $|T_w|$, and $|T_{n-w}|$.

(iv) If $w_1 \neq w_2$, then the sets T_{n,w_1} and T_{n,w_2} are disjoint, with $T_{n,w_1} \cap T_{n,w_2} = \emptyset$. Indeed, if $t \in T_{n,w_1} \cap T_{n,w_2}$, then we would have $w_1 + w_2 = n$, as t must have a caterpillar of size w_1 and a caterpillar of size w_2 as root subtrees. However, $w_1 + w_2$ cannot equal n, as either $w_1 < w_2 \leq n/2$ or $w_2 < w_1 \leq n/2$.

For a tree t of size $n \geq 2$ selected uniformly at random, the mean number $\mathbb{E}_n[c^*(r)]$ of non-equivalent root configurations can be written by conditioning on $t \in T'_n$, that is,

$$\mathbb{E}_n[c^*(r)] = \left(\sum_{w=1}^{w'(n)} p_{n,w} \mathbb{E}_n[c^*_t(r) | t \in T_{n,w}] \right) + \left(1 - \sum_{w=1}^{w'(n)} p_{n,w} \right) \mathbb{E}_n[c^*_t(r) | t \notin T'_n]. \tag{20}$$

Here, the probability $\mathbb{P}[t \in T'_n]$ has been calculated as the sum $\mathbb{P}[t \in T'_n] = \sum_{w=1}^{w'(n)} \mathbb{P}[t \in T_{n,w}]$ because $T'_n = \bigcup_{w=1}^{w'(n)} T_{n,w}$ is a disjoint union.

The expression $\mathbb{E}_n[c^*_t(r) | t \in T_{n,w}]$ in (20) can be replaced by

$$\begin{aligned} \mathbb{E}_n[c^*_t(r) | t \in T_{n,w}] &= 1 + \frac{w-1}{2} + \mathbb{E}_{n-w}[c^*(r)] + (w-1)\mathbb{E}_{n-w}[c^*(r)] - \frac{(w-1)^2}{2} \\ &= 1 + \frac{(w-1)(2-w)}{2} + w\mathbb{E}_{n-w}[c^*(r)], \end{aligned} \tag{21}$$

because for a random tree $t \in T_{n,w}$ selected under a uniform distribution, (9) applies with $c^*(r_S) = w - 1$ and $c^*(r_L) = \mathbb{E}_{n-w}[c^*(r)]$. In particular, $c^*(r_S) = w - 1$, as a caterpillar of size w has $w - 1$ non-equivalent root configurations (Sect. 4.2), and $c^*(r_L) = \mathbb{E}_{n-w}[c^*(r)]$, as the larger root subtree t_{r_L} of a random $t \in T_{n,w}$ selected uniformly has a uniform distribution over T_{n-w} if $w \neq n/2$ (18). If $w = n/2$—which can happen only for $n = 2, 4$, or 6—(21) holds because $\mathbb{E}_n[c^*_t(r) | t \in T_{2,1}] = 1$, $\mathbb{E}_n[c^*_t(r) | t \in T_{4,2}] = 3$, and $\mathbb{E}_n[c^*_t(r) | t \in T_{6,3}] = 6$, while $\mathbb{E}_1[c^*(r)] = 0$, $\mathbb{E}_2[c^*(r)] = 1$, and $\mathbb{E}_3[c^*(r)] = 2$.

Using (21) and ignoring the second term in (20) yields the inequality

$$\mathbb{E}_n[c^*(r)] \geq \sum_{w=1}^{w'(n)} p_{n,w} \left[1 + \frac{(w-1)(2-w)}{2} + w\mathbb{E}_{n-w}[c^*(r)] \right].$$

This inequality can be iterated if $n - w \geq 2$ by applying the same procedure to $\mathbb{E}_{n-w}[c^*(r)]$. It follows that for each $n \geq 1$, the integer e_n defined recursively for $n \geq 2$ by

$$e_n = \sum_{w=1}^{w'(n)} p_{n,w} \left[1 + \frac{(w-1)(2-w)}{2} + we_{n-w} \right], \tag{22}$$

where $e_1 = 0$, bounds from below the expectation $\mathbb{E}_n[c^*(r)]$. The first values of e_n and $\mathbb{E}_n[c^*(r)]$ are reported in Table 1. The values of e_n match the values of $\mathbb{E}_n[c^*(r)]$

Table 1 Sequences e_n, $\mathbb{E}_n[c^*(r)]$ and $\mathbb{E}_n[c(r)]$ for small values of n

n	e_n	$\mathbb{E}_n[c^*(r)]$	$\mathbb{E}_n[c(r)]$	n	e_n	$\mathbb{E}_n[c^*(r)]$	$\mathbb{E}_n[c(r)]$
1	0	0	0	11	$\frac{4613}{247}$	$\frac{94667}{4199}$	$\frac{9841}{323}$
2	1	1	1	12	$\frac{654726}{29393}$	$\frac{863372}{29393}$	$\frac{4840}{119}$
3	2	2	2	13	$\frac{195593}{7429}$	$\frac{1990481}{52003}$	$\frac{402752}{7429}$
4	3	3	$\frac{16}{5}$	14	$\frac{6033381}{185725}$	$\frac{9266561}{185725}$	$\frac{788741}{10925}$
5	$\frac{30}{7}$	$\frac{30}{7}$	$\frac{33}{7}$	15	$\frac{4299031}{111435}$	$\frac{21753971}{334305}$	$\frac{99454}{1035}$
6	$\frac{121}{21}$	$\frac{121}{21}$	$\frac{20}{3}$	16	$\frac{88030888}{1938969}$	$\frac{164642378}{1938969}$	$\frac{3837632}{30015}$
7	$\frac{254}{33}$	$\frac{254}{33}$	$\frac{304}{33}$	17	$\frac{9891227}{186093}$	$\frac{1959845063}{17678835}$	$\frac{52758677}{310155}$
8	$\frac{1356}{143}$	$\frac{334}{33}$	$\frac{1795}{143}$	18	$\frac{4014691853}{64822395}$	$\frac{3128723951}{21607465}$	$\frac{1157564}{5115}$
9	$\frac{8961}{715}$	$\frac{729}{55}$	$\frac{1102}{65}$	19	$\frac{1715903641}{23881935}$	$\frac{22592912099}{119409675}$	$\frac{1563215792}{5191725}$
10	$\frac{37549}{2431}$	$\frac{42039}{2431}$	$\frac{296}{13}$	20	$\frac{24415042314}{294543865}$	$\frac{72844824142}{294543865}$	$\frac{39979649}{99789}$

Values of e_n were computed by using (22). Values of $\mathbb{E}_n[c^*(r)]$ were computed by generating all possible unlabeled topologies of size n and then using STELLS (Wu 2012) to obtain the number $c_t^*(r)$ of non-equivalent root configurations for each unlabeled topology t. The probability of t under a uniform distribution over labeled topologies of size n was obtained by noting that its number of labelings $L(t)$ follows the recursion in Eq. 5.1 of Harding (1971); non-recursively, the number of labelings is $n!/2^{s(t)}$, where $s(t)$ is the number of internal nodes of t, including cherries and possibly the root, whose two descendant subtrees are isomorphic [this result is obtained by taking the quotient of the results of Theorems 3.5 and 3.3 of Rosenberg (2006)]. To compute $c_t^*(r)$, we ran STELLS on tree $(t, \cdot)$ in which the two root subtrees were t and the one-taxon tree $\cdot$. According to (7), the number of root configurations computed by STELLS is $c_t^*(r) + 1$, from which the desired $c_t^*(r)$ is obtained. Values of $\mathbb{E}_n[c(r)]$ were computed by the method of Disanto and Rosenberg (2017, Fig. 7)

for $n \leq 7$, that is, as long as $T'_n = T_n$ and the second term in (20) is 0. We also have the following result.

Proposition 3 *The expected number* $\mathbb{E}_n[c^*(r)]$ *of non-equivalent root configurations in a random tree of size* $n \geq 1$ *selected under a uniform distribution can be bounded*

$$e_n \leq \mathbb{E}_n[c^*(r)] \leq \mathbb{E}_n[c(r)], \tag{23}$$

where e_n *is defined in (22) and* $\mathbb{E}_n[c(r)]$ *is the expected number of root configurations. Furthermore, the sequence* $\mathbb{E}_n[c^*(r)]$ *grows exponentially in n, with exponential order at most* $4/3$.

Proof The upper bound follows from the fact that for any tree, $c^*(r) \leq c(r)$, and by point (v) in Sect. 2.4, $\mathbb{E}_n[c(r)]$ has exponential order $4/3$. All that remains is to show that $\mathbb{E}_n[c^*(r)]$ grows exponentially in n. To achieve this goal, we prove that the exponential order of the lower bound sequence e_n strictly exceeds one.

Truncating the sum (22) after the first four terms, for $n \geq 9$, we have

$$\begin{aligned} e_n &\geq p_{n,1}e_{n-1} + 2p_{n,2}e_{n-2} + 3p_{n,3}e_{n-3} + 4p_{n,4}e_{n-4} + (p_{n,1} + p_{n,2} - 2p_{n,4}) \\ &\geq p_{n,1}e_{n-1} + 2p_{n,2}e_{n-2} + 3p_{n,3}e_{n-3} + 4p_{n,4}e_{n-4}. \end{aligned} \tag{24}$$

The last step follows because according to (19), when $n \geq 9$, $p_{n,1} = n/(2n-3)$, $p_{n,2} = n(n-1)/[2(2n-3)(2n-5)]$, $p_{n,4} = n(n-1)(n-2)(n-3)/[2(2n-3)(2n-5)(2n-7)(2n-9)]$, and

$$p_{n,1} + p_{n,2} - 2p_{n,4} = \frac{n(2n-11)!!\,(18n^3 - 192n^2 + 645n - 681)}{2(2n-3)!!} \geq 0.$$

Define the sequence a_n by $a_n = e_n$ for $1 \leq n \leq 8$, and $a_n = p_{n,1}a_{n-1} + 2p_{n,2}a_{n-2} + 3p_{n,3}a_{n-3} + 4p_{n,4}a_{n-4}$ for $n \geq 9$. From (24), we have, for each $n \geq 1$,

$$e_n \geq a_n. \tag{25}$$

When $n \geq 9$ and $1 \leq w \leq 4$, because $w \neq n/2$ and $\delta_{n,w/2} = 0$, the probability $p_{n,w}$ in (19) can be written

$$p_{n,w} = \frac{(2n-2w-3)!!}{(2n-3)!!}\,\frac{n!}{(n-w)!}\,\frac{1}{2-\delta_{w,1}}.$$

The recursion for a_n then becomes

$$\begin{aligned} a_n = {} & \frac{n(2n-5)!!}{(2n-3)!!}a_{n-1} + \frac{n(n-1)(2n-7)!!}{(2n-3)!!}a_{n-2} + \frac{3n(n-1)(n-2)(2n-9)!!}{2(2n-3)!!}a_{n-3} \\ & + \frac{2n(n-1)(n-2)(n-3)(2n-11)!!}{(2n-3)!!}a_{n-4}. \end{aligned} \tag{26}$$

Setting $q_n = a_n(2n-3)!!/n!$, we obtain from (26)

$$q_n = q_{n-1} + q_{n-2} + \frac{3q_{n-3}}{2} + 2q_{n-4}. \tag{27}$$

Recursion (27) is homogeneous and linear with constant coefficients, and therefore (Sedgewick and Flajolet 1996, Theorems 3.3 and 4.1), the exponential order of the sequence q_n is the inverse of the unique positive solution z_0 of the characteristic equation $1 = z + z^2 + 3z^3/2 + 2z^4$.

Solving the equation numerically, we find $q_n \bowtie (1/z_0)^n$, where $z_0 \approx 0.4845$. In particular, the exponential order $1/z_0$ of the sequence q_n strictly exceeds 2. Using (2) to rewrite $(2n-3)!!$, and observing by Stirling's formula $n! \sim (n/e)^n\sqrt{2\pi n}$ that $\binom{2n}{n} \bowtie 4^n$, it follows that sequence $a_n = q_n n!/(2n-3)!!$ has exponential growth

$$a_n \bowtie q_n \frac{n!}{\frac{(2n)!}{2^n n!}} = q_n \frac{2^n}{\binom{2n}{n}} \bowtie \left(\frac{1/z_0}{2}\right)^n.$$

Therefore, the exponential order of the sequence a_n is $1/(2z_0) \approx 1.0320 > 1$. By inequality (25), the sequence e_n grows exponentially in n. □

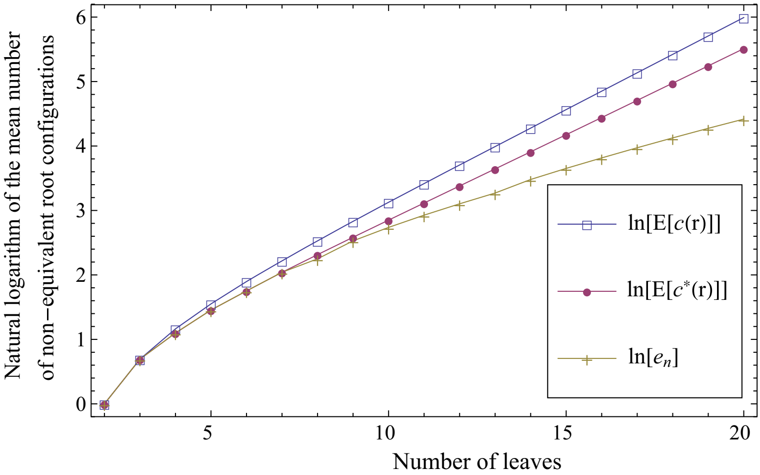

Fig. 8 Natural logarithm of the mean number $\mathbb{E}_n[c^*(r)]$ of non-equivalent root configurations for labeled topologies of size $2 \leq n \leq 20$. The value for $n = 1$, log(0), is omitted. The natural logarithms of the bounds e_n and $\mathbb{E}_n[c(r)]$ (23) determine the *lower* and *upper lines*. Exact values for the three quantities are reported in Table 1

For $n \leq 20$, the exact values of e_n, $\mathbb{E}_n[c^*(r)]$, and $\mathbb{E}_n[c(r)]$ are reported in Table 1 and plotted in Fig. 8. The figure illustrates that the numerical values of $\log \mathbb{E}_n[c^*(r)]$, though initially coincident with the values of $\log e_n$, are already closer to the values of $\log \mathbb{E}_n[c(r)]$ by $n = 20$. This observation suggests that in bounding $\mathbb{E}_n[c^*(r)]$ from below to demonstrate its exponential growth, the steps we have taken have led to a bound that is quite loose; the exponential growth of $\mathbb{E}_n[c^*(r)]$ is likely to have a comparable magnitude to that of $\mathbb{E}_n[c(r)]$, or $4/3$.

6 Discussion

For labeled gene tree topologies t that match the labeled species tree topology, we have extended the enumerative study of ancestral configurations, considering non-equivalent configurations specified by an equivalence relation that groups ancestral configurations according to symmetries in t. We have focused on the exponential growth in the tree size $|t| = n$ of the number of non-equivalent configurations present at the root of t.

We have shown that when t satisfies certain constraints, its number of non-equivalent root configurations can be recursively computed from corresponding quantities for its root subtrees. The recursion (9), which shares three of its five terms with an analogous recursion for root configurations (Disanto and Rosenberg 2017, Proposition 1), enables the study of the number of non-equivalent root configurations for special tree families. For the family of completely balanced trees, the number of non-equivalent root configurations and the total number of non-equivalent configurations grow exponentially with order $k_0^* \approx 1.2460$ in n (Proposition 1). Comparing this constant with the exponential orders of the numbers of root configurations and total configurations in the family, both of which equal $k_0 \approx 1.5028$ (Disanto and Rosenberg 2017), we see that for the completely balanced trees, the number of configurations grows exponentially faster than the number of non-equivalent configurations. Their symmetric structure collapses the set of configurations into fewer non-equivalent configurations.

A different recursively defined tree family (u_n), however, has asymptotically more non-equivalent configurations than the balanced trees, its number of root configurations growing with exponential order $\sqrt[3]{3} \approx 1.4422$ (Proposition 2). This value is close to the upper bound of $k_0 \approx 1.5028$ on the exponential order of the maximal number of configurations across all labeled topologies of size n (Disanto and Rosenberg 2017, Corollary 1). Although the unlabeled shapes that give rise to the largest numbers of non-equivalent root configurations (Fig. 5) and root configurations (Disanto and Rosenberg 2017, Fig. 3) are not in general the same, the maximal numbers of non-equivalent configurations and configurations have comparable exponential order.

As was found by Wu (2012), the growth of the number of non-equivalent configurations for some tree families (e.g., caterpillars) can be polynomial in n. Assuming a uniform distribution over the labeled topologies with size n, however, we have shown that the expected number of non-equivalent configurations for a random labeled topology of size n grows exponentially (Proposition 3). The exponential order of this growth is bounded below by $1/(2z_0) \approx 1.0320$; numerical exploration suggests that it is closer to the upper bound of $4/3$ that describes the exponential order of the mean number of configurations (Disanto and Rosenberg 2017, Proposition 5).

We focused on the situation in which the gene tree and species tree have a matching topology. In the non-matching case, in parallel to a similar result for configurations (Disanto and Rosenberg 2017), it is possible that the number of non-equivalent root configurations and the total number of non-equivalent configurations exceed the corresponding values for matching gene trees and species trees. This claim can be verified in a simple example. Let $\chi_n = ((\ldots((a_1, a_2), a_3), \ldots), a_n)$ be a caterpillar species tree, and label the unique internal node with k descendants by b_k for $2 \leq k \leq n$. For a matching caterpillar gene tree, all configurations are non-equivalent, the number of non-equivalent configurations at node b_k is $c^*(b_k) = k - 1$, the number of root configurations is $c^*(b_n) = n - 1$, and the total number of configurations is $c^* = \sum_{k=2}^{n} c^*(b_k) = n(n-1)/2$.

Continuing with χ_n as the species tree topology, consider a gene tree topology

$$\xi_n = ((\ldots((((a_1, a_2), a_3), (a_4, a_5)), a_6), \ldots), a_n)$$

with $n \geq 6$. The gene trees (ξ_n) represent a caterpillar family (Disanto and Rosenberg 2016) with seed tree $(((a_1, a_2), a_3), (a_4, a_5))$. We label the node of ξ_n ancestral to a_1 and a_2 by d_2, the node ancestral to a_1, a_2, and a_3 by d_3, the node ancestral to a_4 and a_5 by d_2^*, and the unique node ancestral to k taxa, $5 \leq k \leq n$, by d_k. Following Wu (2012), the definition of equivalent configurations in the non-matching case generalizes the definition in Sect. 3.1. Consider a gene tree G, a species tree S, a node κ of S, and two configurations γ_1, γ_2 at node κ—two possible sets of gene lineages that could be present in S at κ under different realizations of G in S. Let κ' be the most recent common ancestor of the lineages of G collected in the set γ_1, and note that κ' is also the most recent common ancestor of the lineages collected in γ_2. Following the terminology of Sect. 3.1, we say that γ_1, γ_2 are *equivalent* at κ when the unlabeled tree shape $G_{\kappa'}(\gamma_1)$ is isomorphic to the unlabeled tree shape $G_{\kappa'}(\gamma_2)$. We denote by $C^*(\kappa)$ and $c^*(\kappa)$ the set of non-equivalent configurations at κ and its cardinality, respectively.

Proceeding sequentially through the internal nodes of χ_n, the non-equivalent configurations are $C^*(b_2) = \{\{a_1, a_2\}\}$, $C^*(b_3) = \{\{a_1, a_2, a_3\}, \{d_2, a_3\}\}$, $C^*(b_4) = \{\{a_1, a_2, a_3, a_4\}, \{d_2, a_3, a_4\}, \{d_3, a_4\}\}$, and $C^*(b_5) = \{\{a_1, a_2, a_3, a_4, a_5\}, \{d_2, a_3, a_4, a_5\}, \{d_3, a_4, a_5\}\}$, with $c^*(b_2) = 1$, $c^*(b_3) = 2$, $c^*(b_4) = 3$, and $c^*(b_5) = 3$. At node b_6 of χ_n, the non-equivalent configurations are $C^*(b_6) = \{\{a_1, a_2, a_3, a_4, a_5, a_6\}, \{d_2, a_3, a_4, a_5, a_6\}, \{d_3, a_4, a_5, a_6\}, \{a_1, a_2, a_3, d_2^*, a_6\}, \{d_3, d_2^*, a_6\}, \{d_5, a_6\}\}$, and configuration $\{d_2, a_3, d_2^*, a_6\}$ is not included owing to equivalence with $\{d_3, a_4, a_5, a_6\}$.

For $7 \leq k \leq n$, $C^*(b_k)$ is obtained by augmenting configuration $\{d_{k-1}, a_k\}$ to the set of all configurations formed by adding taxon a_k to the non-equivalent configurations in $C^*(b_{k-1})$; none of the resulting configurations are equivalent, and $c^*(b_k) = c^*(b_{k-1}) + 1$. The number of non-equivalent root configurations of ξ_n for $n \geq 6$ is $c^*(b_n) = n$, and the number of total configurations is $c = 1+2+3+3+\sum_{k=6}^{n} c^*(b_k) = n(n+1)/2 - 6$. Because $n > n - 1$ and $n(n+1)/2 - 6 > n(n-1)/2$ for $n \geq 7$, non-equivalent root configurations and total non-equivalent configurations are more numerous for the non-matching ξ_n than for the matching caterpillar.

Our enumerative results on ancestral configurations can help to compare the cost of procedures for calculating gene tree probabilities recursively using ancestral configurations (Wu 2012) to those that proceed non-recursively using a different data structure, the "coalescent histories" (Degnan and Salter 2005; Rosenberg 2007, 2013; Than et al. 2007; Rosenberg and Degnan 2010; Disanto and Rosenberg 2015, 2016). In this context, it is noteworthy that the trees u_n, which have many non-equivalent root configurations, have a similar recursive structure to the lodgepole trees, which have large numbers of coalescent histories (Disanto and Rosenberg 2015).

Note that unlike for root configurations, we did not prove a general result describing the unlabeled shapes of trees that give rise to the most non-equivalent root configurations, merely evaluating the number of non-equivalent root configurations for trees u_n and noting by exhaustive computation that this value is near the maximum for small trees. We also did not produce a general relationship between non-equivalent root configurations and total non-equivalent configurations. For the family of completely balanced trees, the number of non-equivalent root configurations and the total number of non-equivalent configurations have the same exponential growth, as the maximal number of non-equivalent configurations across all internal nodes of a balanced tree is reached at its root (Proposition 1). However, we did not provide a generalization that such a maximum is applicable for arbitrary trees. Because it is the non-equivalent configurations that are employed by Wu (2012) in gene tree probability computations, their further exploration will be important for understanding the relative computational complexity of gene tree probability computations with different species trees.

Acknowledgements We thank Elizabeth Allman, James Degnan, and John Rhodes for discussions, and two reviewers for comments. Support was provided by National Institutes of Health grant R01 GM117590 and by a 2014 Rita Levi Montalcini grant to FD from the Ministero dell'Istruzione, dell'Università e della Ricerca.

Appendix 1: Proof of (9)

Let $C^*(r_S) = \{\gamma_{S,1}, \ldots, \gamma_{S,q}\}$ with $c^*(r_S) = q$, and let $C^*(r_L) = \{\gamma_{L,1}, \ldots, \gamma_{L,Q}\}$, with $c^*(r_L) = Q$. Because condition (8) is satisfied, the entire tree t_{r_S} can be displayed in t_{r_L}, each configuration $\gamma_{S,i} \in C^*(r_S)$ has exactly one corresponding configuration $\gamma_{L,i} \in C^*(r_L)$ such that $t_{r_S}(\gamma_{S,i}) \cong t_{r_L}(\gamma_{L,i})$, and $Q \geq q$.

From (6), we obtain

$$\tilde{C}(r) = \{\{r_S, r_L\}\} \cup \big[C^*(r_S) \otimes \{\{r_L\}\}\big] \cup \big[\{\{r_S\}\} \otimes C^*(r_L)\big] \cup \big[C^*(r_S) \otimes C^*(r_L)\big],$$

which can be further decomposed as

$$\begin{aligned}
\tilde{C}(r) = \{\{r_S, r_L\}\} \cup \big[\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\{r_L\}\}\big] \cup \big[\{\{r_S\}\} \otimes \big[\{\gamma_{L,1}, \ldots, \gamma_{L,q}\} \\
\cup\{\gamma_{L,q+1}, \ldots, \gamma_{L,Q}\}\big]\big] \\
\cup\big[\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \big[\{\gamma_{L,1}, \ldots, \gamma_{L,q}\} \cup \{\gamma_{L,q+1}, \ldots, \gamma_{L,Q}\}\big]\big] \\
= \{\{r_S, r_L\}\} && (28) \\
\cup\big[\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\{r_L\}\}\big] \cup \big[\{\{r_S\}\} \otimes \{\gamma_{L,1}, \ldots, \gamma_{L,q}\}\big] && (29) \\
\cup\big[\{\{r_S\}\} \otimes \{\gamma_{L,q+1}, \ldots, \gamma_{L,Q}\}\big] && (30) \\
\cup\big[\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\gamma_{L,1}, \ldots, \gamma_{L,q}\}\big] && (31) \\
\cup\big[\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\gamma_{L,q+1}, \ldots, \gamma_{L,Q}\}\big]. && (32)
\end{aligned}$$

We merge equivalent configurations to obtain $C^*(r)$ from $\tilde{C}(r)$. From (29), we remove those in $\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\{r_L\}\}$, as they are equivalent to those in $\{\{r_S\}\} \otimes \{\gamma_{L,1}, \ldots, \gamma_{L,q}\}$. Thus, we take only q among the $2q$ configurations in (29). Moreover, due to the equivalence $\gamma_{S,i} \cup \gamma_{L,j} \sim_r \gamma_{S,j} \cup \gamma_{L,i}$, we take only those configurations of the form $\gamma_{S,i} \cup \gamma_{L,j}$ with $i \leq j$ among those in $\{\gamma_{S,1}, \ldots, \gamma_{S,q}\} \otimes \{\gamma_{L,1}, \ldots, \gamma_{L,q}\}$. Thus, among the q^2 configurations in (31)—those with $1 \leq i, j \leq q$—we take only $q(q+1)/2$ non-equivalent ones. No equivalences are possible among configurations in (28), (30), and (32), and all are retained in $C^*(r)$. From (28)–(32), we then have

$$\begin{aligned}
c^*(r) = |C^*(r)| = 1 + q + (Q - q) + \frac{q(q+1)}{2} + q(Q - q) = 1 + q + Q \\
+qQ - \frac{q(q+1)}{2}.
\end{aligned}$$

Replacing q by $c^*(r_S)$ and Q by $c^*(r_L)$ gives (9).

Appendix 2: Proof of (12)

The proof follows the approach of Aho and Sloane (1973, Sect. 3) for solving certain recurrences. From (11), we have $x_{h+1} = x_h^2[1 + 1/(2x_h) + 1/(2x_h^2)]$. Taking the logarithm $y_h = \log x_h$ yields $y_{h+1} = 2y_h + \alpha_h$, where $\alpha_h = \log[1 + 1/(2x_h) + 1/(2x_h^2)]$. Following Aho and Sloane (1973), y_h has solution

$$y_h = 2^h y_0 + \sum_{i=0}^{\infty} 2^{h-i-1}\alpha_i - \sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i = 2^h\Big(y_0 + \sum_{i=0}^{\infty} 2^{-i-1}\alpha_i\Big) - \sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i. \tag{33}$$

Converting back to $x_h = \exp(y_h)$, from (33) we have

$$\begin{aligned} x_h &= \Big[x_0 \exp\Big(\sum_{i=0}^{\infty} 2^{-i-1}\alpha_i\Big)\Big]^{(2^h)} \exp\Big(-\sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i\Big) \\ &= (k_0^*)^{(2^h)} \exp\Big(-\sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i\Big), \end{aligned}$$

where the last step uses the fact that $x_0 = 1/2$.

We then have

$$\frac{x_h}{(k_0^*)^{(2^h)}} = \exp\Big(-\sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i\Big).$$

When $h \to \infty$, the sum $\sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i$ converges to zero because it can be bounded $0 \le \sum_{i=h}^{\infty} 2^{h-i-1}\alpha_i \le \alpha_h \sum_{i=h}^{\infty} 2^{h-i-1} = \alpha_h$, where because $x_h \to \infty$ as $h \to \infty$, $\alpha_h \to 0$ as $h \to \infty$. It follows that $x_h/(k_0^*)^{(2^h)}$ converges to 1, producing (12).

Appendix 3: Properties of $w'(n)$

We prove that for each $n \ge 2$, $w'(n) \le n/2$, with equality only for $n = 2$, 4, or 6. The result is verified by direct computation of $w'(n)$ for $2 \le n \le 7$. For $n \ge 8$, by definition, $w'(n) = \lfloor x \rfloor$, where x satisfies $2^{x-2} + x = n - 1$. Seeking a contradiction, suppose $\lfloor x \rfloor = w'(n) \ge n/2$. Because $x \ge \lfloor x \rfloor$, we would have $x \ge n/2$, and therefore $n - 1 = 2^{x-2} + x \ge 2^{n/2-2} + n/2 \ge 2(n/2-2) + n/2 = 3n/2 - 4$, noting that $2^u \ge 2u$ for $u \ge 2$. The inequality $n - 1 \ge 3n/2 - 4$ cannot hold if $n \ge 8$. Therefore, when $n \ge 8$, we must have $w'(n) < n/2$.

Appendix 4: Proof that Trees in $T_{n,w}$ Satisfy (8) for $w \ge 2$

We first prove that given any $w \ge 2$, a caterpillar tree t_1 of size $|t_1| = w$ can be displayed in any tree t_2 of size $|t_2| \ge 2^{w-2} + 1$ through a root configuration γ of t_2, that is, $t_1 \cong t_2(\gamma)$. The proof is by induction on w.

For $w = 2$, we have $|t_2| \ge 2$ and the result follows by taking the root configuration γ determined by the left and right descendants of the root in t_2. For the inductive step, because $|t_2| \ge 2^{w-2} + 1$, the larger root subtree of t_2 has size at least $\lceil |t_2|/2 \rceil \ge \lceil 2^{w-3} + 1/2 \rceil = 2^{w-3} + 1$. By the inductive hypothesis, the larger root subtree of t_2 can display a caterpillar of size $w - 1$ through a root configuration γ'. Taking the root configuration γ of t_2 obtained as $\gamma = \gamma' \cup \{\rho\}$, where ρ is the root of the smaller root subtree of t_2, we have $t_1 \cong t_2(\gamma)$ as desired.

Now suppose we are given a tree $t \in T_{n,w}$, with $2 \leq w \leq w'(n)$. The smaller root subtree t_{r_S} of t is by definition a caterpillar of size $w \geq 2$, and the larger root subtree t_{r_L} has size $|t_{r_L}| = n - w$. By definition, $w \leq w'(n) = \lfloor x \rfloor \leq x$, where $x = n - 2^{x-2} - 1$, and therefore, $w \leq n - 2^{w-2} - 1$. In particular, $|t_{r_L}| = n - w \geq 2^{w-2} + 1$. From what we have shown above, a root configuration γ of t_{r_L} exists such that $t_{r_S} \cong t_{r_L}(\gamma)$.

Appendix 5: Proof of (18)

Recall that for each tree $t \in T_{n,w}$, the smaller root subtree t_{r_S} is a caterpillar of size $w \in [1, w']$ and the larger root subtree t_{r_L} has size $n - w$. Because we assume $w < n/2$, t_{r_S} and t_{r_L} have different sizes and different unlabeled topologies. Given a tree $\bar{t} \in T_{n-w}$, the number of trees in $T_{n,w}$ such that $t_{r_L} = \bar{t}$ (after rescaling labels for the taxa) is $\binom{n}{w}\gamma_w$, where γ_w is the number of caterpillar labeled topologies of size w. Dividing by $|T_{n,w}| = \binom{n}{w}\gamma_w |T_{n-w}|$ yields the probability $\mathbb{P}[t_{r_L} = \bar{t} | t \in T_{n,w}] = 1/|T_{n-w}|$ as desired.

References

Aho AV, Sloane NJA (1973) Some doubly exponential sequences. Fibonacci Q. 11:429–437

Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62:833–862

Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution 59:24–37

Disanto F, Rosenberg NA (2015) Coalescent histories for lodgepole species trees. J Comput Biol 22:918–929

Disanto F, Rosenberg NA (2016) Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. IEEE/ACM Trans Comput Biol Bioinf 13:913–925

Disanto F, Rosenberg NA (2017) Enumeration of ancestral configurations for matching gene trees and species trees. J Comput Biol 24:831–850

Felsenstein J (1978) The number of evolutionary trees. Syst. Zool. 27:27–33

Felsenstein J (2004) Inferring phylogenies. Sinauer, Sunderland, MA

Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77

Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann Comb 10:129–146

Rosenberg NA (2007) Counting coalescent histories. J Comput Biol 14:360–377

Rosenberg NA (2013) Coalescent histories for caterpillar-like families. IEEE/ACM Trans Comput Biol Bioinf 10:1253–1262

Rosenberg NA, Degnan JH (2010) Coalescent histories for discordant gene trees and species trees. Theor Pop Biol 77:145–151

Sedgewick R, Flajolet P (1996) An introduction to the analysis of algorithms. Addison-Wesley, Boston

Than C, Ruths D, Innan H, Nakhleh L (2007) Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. J Comput Biol 14:517–535

Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66:763–775