On the Number of Ranked Species Trees Producing Anomalous Ranked Gene Trees

Filippo Disanto and Noah A. Rosenberg

Abstract—Analysis of probability distributions conditional on species trees has demonstrated the existence of anomalous ranked gene trees (ARGTs), ranked gene trees that are more probable than the ranked gene tree that accords with the ranked species tree. Here, to improve the characterization of ARGTs, we study enumerative and probabilistic properties of two classes of ranked labeled species trees, focusing on the presence or avoidance of certain subtree patterns associated with the production of ARGTs. We provide exact enumerations and asymptotic estimates for cardinalities of these sets of trees, showing that as the number of species increases without bound, the fraction of all ranked labeled species trees that are ARGT-producing approaches 1. This result extends beyond earlier existence results to provide a probabilistic claim about the frequency of ARGTs.

Index Terms—Enumeration, gene trees, labeled histories, mathematical phylogenetics, species trees

1 Introduction

Recent research in phylogenetics has conducted detailed probabilistic explorations of the properties of different gene tree structures using models of gene lineage evolution conditional on species trees [1], [2], [5], [6], [11]. These phylogenetic modeling investigations uncover new phylogenetic phenomena, facilitate mathematical and simulation-based analyses of complex data spaces for phylogenetic studies, enable development and theoretical analysis of species tree inference algorithms, and assist in identifying strengths, limitations, and protocols for proposed methods [8], [19], [23], [24], [30].

A ranked labeled gene tree, or gene tree labeled history, consists of a rooted labeled gene tree topology together with the temporally ordered sequence in which coalescences in the gene tree take place [15], [25]. Ranked gene trees arise in a model of random bifurcation in which each lineage is equally likely to be the next to bifurcate, or, backward in time, each pair of lineages is equally likely to be the next to coalesce. This simple branching assumption, originating from the classical *Yule model* [31] and providing the model of tree topology in coalescent models for gene lineage evolution [16], [29], generates a convenient uniform distribution on the set of ranked gene trees [13], [18].

Given a genealogical history of a set of gene lineages, the ranked gene tree is an elemental tree structure, in the sense that other structures—such as unranked rooted gene trees, unranked unrooted gene trees, and the list of clades included in a tree—are uniquely specified by a ranked gene tree, whereas many ranked gene trees might be compatible with a given choice for one of these other structures. Thus, as properties of other structures can often be derived from properties of ranked gene trees [3], [12], [22], [27], ranked gene trees represent a natural class of objects for phylogenetic modeling.

Degnan et al. [10] initiated the probabilistic study of ranked gene trees in species tree models, providing a formula under the standard multispecies coalescent model [8], [11], [17], [19], [21] for the probability conditional on a labeled species tree that a particular ranked gene tree is produced (see also [26]). Under the model, [10] termed ranked labeled gene trees that are more likely to be generated than the ranked labeled gene tree that matches the ranked labeled species tree *anomalous ranked gene trees* (ARGTs). ARGTs represent a surprising outcome of genealogical descent in which an unexpected ranked gene tree exceeds the model ranked species tree in probability.

Degnan et al. [9] obtained a full characterization of the set of unranked labeled species trees for which at least one ranking produces ARGTs. That is, they identified all unranked labeled species trees for which a ranking and a set of branch lengths can be selected so that the most likely ranked gene tree conditional on the ranked species tree together with its branch lengths disagrees with the ranked species tree. They found that the set of unranked labeled species trees with at least one ARGT-producing ranking is precisely the set of unranked labeled species trees that do not have a caterpillar or pseudocaterpillar shape.

Though the constructive proof of [9] identifies specific ARGT-producing rankings for a given unranked labeled species tree, the set of *ranked* labeled species trees that are ARGT-producing remains incompletely characterized. For small trees, [9, Table 1] reported the numbers of ranked labeled species trees that give rise to ARGTs, but general results have not been presented to assess the fraction of ranked labeled species trees that are ARGT-producing.

Here, we show that as the number of species increases without bound, the fraction of all ranked labeled species trees that are ARGT-producing—that is, the fraction for which some set of species tree branch lengths gives rise to ARGTs—approaches 1. In other words, we extend beyond the proof of [9] to argue that not only does each unranked

• *The authors are with the Department of Biology, Stanford University, Stanford, CA. E-mail: {fdisanto, noahr}@stanford.edu.*

Manuscript received 2 Apr. 2014; revised 3 June 2014; accepted 17 July 2014. Date of publication 28 July 2014; date of current version 4 Dec. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2343977

1545-5963 © 2014 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

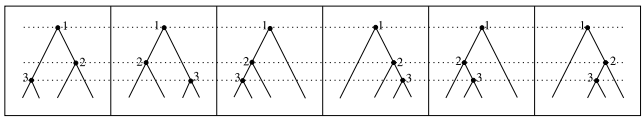

Fig. 1. The six ordered ranked trees of size $n = 3$ internal nodes. Left-right orientation determines different trees.

species tree have at least one ARGT-producing ranking, nearly all ranked species trees are ARGT-producing. We obtain the result through a combinatorial approach, counting the number of ranked labeled species trees with n internal nodes that are identified by the proof of [9] as ARGT-producing, and we show that the ratio of the cardinality of this set and the total number of ranked labeled species trees on n nodes, or $(n+1)!\,n!/2^n$, approaches 1 as n approaches infinity.

2 Preliminaries

2.1 Ranked Trees, Ranked Species Trees, and Ordered Ranked Trees

It is convenient here to index tree and subtree sizes by the number of internal nodes, rather than by the usual index, the number of leaves.

A *ranked* tree t of size n is a binary rooted tree with n *internal* nodes (and $n+1$ leaves), each one bijectively associated with a number in $\{1, 2, \ldots, n\}$. The labeling of the internal nodes must be *increasing*, in the sense that each path from the root of t to a leaf contains an increasing sequence of numbers. The increasing labeling gives a time ordering of the coalescence events occurring along the branches of the tree. The most recent event is the one that carries the greatest label. Ranked trees are considered in a graph-theoretic sense. Therefore, unless specified otherwise, they do not carry any left-right orientation.

A ranked *species* tree is a ranked tree equipped with a labeling for its taxa. Thus, two ranked species trees can be the same when treated as ranked trees but different in their leaf labeling. The set of ranked species trees is denoted by $\mathcal{S}$, and $\mathcal{S}_n$ denotes the set of ranked species trees of size n. It is well-known ([22, Corollary 3.2]) that the cardinality of $\mathcal{S}_n$ is

$$|\mathcal{S}_n| = \frac{(n+1)!\,n!}{2^n}. \tag{1}$$

An *ordered* ranked tree is a ranked tree provided with a left-right orientation of its subtrees. The set of ordered ranked trees is denoted by $\mathcal{R}$, and $\mathcal{R}_n$ is the subset of $\mathcal{R}$ consisting of those trees of size n. The cardinality of $\mathcal{R}_n$ is ([14, Example II.17])

$$|\mathcal{R}_n| = n!. \tag{2}$$

In Fig. 1, we depict the six ordered ranked trees of size 3. Note that in each tree, the labeling of the internal nodes increases from the root toward the leaves.

2.2 Maximally Probable (MP) and Non-Maximally Probable (NMP) Subtrees

Following [9, Proposition 6], given a ranked tree t and an internal node k, we say that k generates a *maximally probable* subtree MP-subtree for short, if we can assign the name L to one of the two subtrees appended to node k and the name R to the other such subtree in such a way both (i) and (ii) hold for that assignment:

(i) $m \geq q \geq 0$, where $m = |L|$ and $q = |R|$.

(ii) Looking back in time, the sequence of coalescences in the subtree of node k has the form

$$\ell^{m-q}\{\ell r, r\ell\}^q, \tag{3}$$

where ℓ and r stand for coalescence events belonging to subtrees L and R, respectively.

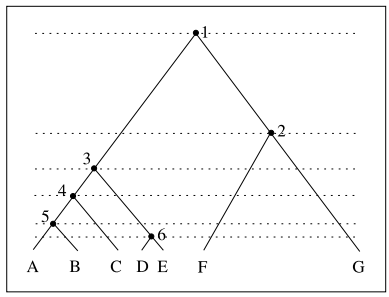

Fig. 2. A ranked species tree that is non-maximally probable at internal node 3. This tree is maximally probable at the root.

The notation $\{a, b\}^q$ in (3) indicates the set of words of length q over the alphabet $\{a, b\}$, where $a = \ell r$ and $b = r\ell$. Thus, by $\ell^{m-q}\{\ell r, r\ell\}^q$, for $m \geq q$, it is meant that the first $m - q$ entries are in L, after which q pairs of entries appear. Each pair has one event in L and the other in R, and the sequences of these events within pairs are not necessarily the same. The suggestive labels L and R can refer to the *left* and *right* subtrees of k, but the definition of maximally probable does not require specification of which subtree is denoted L and which is denoted R.

Given a ranked tree t and an internal node k, we say that k generates a *non-maximally probable* subtree (NMP-subtree for short) when it does not generate an MP-subtree. It is equivalent for a ranked species tree t to avoid NMP-subtrees and to contain only MP-subtrees. The subset of trees in $\mathcal{S}$ containing only MP-subtrees is denoted $\mathcal{S}^{(mp)}$. By $\mathcal{S}_n^{(mp)}$, we indicate trees in $\mathcal{S}^{(mp)}$ of size n.

The tree in Fig. 2 contains exactly one NMP-subtree, that is, the one generated at node 3. Indeed, observe that the only possible assignment of L and R that satisfies (i) gives a sequence of coalescences $r\ell\ell$ that does not match (3); none of the other nodes generates an NMP-subtree. For instance, at node 1, we can assign L to the subtree generated by node 3 and R to the subtree generated by node 2, and the resulting sequence of coalescence events is $\ell\ell\ell\ell r$.

Note that for a node k to generate an NMP-subtree it is necessary to satisfy the following *1-2 condition*: one of the two subtrees appended to k has size at least 1 and the other has size at least 2. Trees for which the 1-2 condition is not satisfied for any internal node are either *caterpillar* or *pseudocaterpillar* (Fig. 3), using the definition that a tree has a *caterpillar* shape when each internal node has at least one leaf stemming from it, and a *pseudocaterpillar* shape when it is not a caterpillar and, still, no node has the 1-2 condition.

We define $\mathcal{S}^{(cat)}$ as the set of caterpillar and pseudocaterpillar ranked species trees. The subset $\mathcal{S}_n^{(cat)}$ contains such

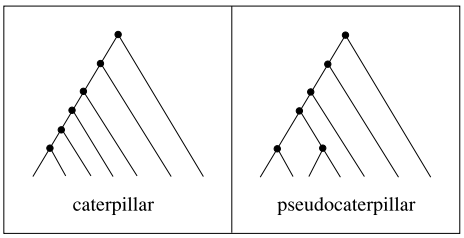

Fig. 3. Caterpillar and pseudocaterpillar trees. These trees do not contain NMP-subtrees.

trees of size n. Caterpillar and pseudocaterpillar trees are not NMP, and they contain no NMP-subtrees.

2.3 Anomalous Ranked Gene Trees

We recall that an *anomalous* ranked gene tree is a ranked gene tree that does not match the ranked species tree and that has probability under the multispecies coalescent model greater than that of the matching ranked gene tree [9], [10]. We say that a ranked species tree produces ARGTs if there exist values for the speciation times such that the ranked species tree together with the speciation times has at least one ARGT.

When we disregard the ranking of the coalescences in the species tree, the set of unranked species trees that produce ARGTs has a known complete characterization. In particular, as shown in [9, Theorem 1], each unranked species tree t that is neither a caterpillar nor a pseudocaterpillar can be ranked in such a way that it is NMP at a particular subtree $H(t)$. Further, being NMP at a subtree implies that speciation times can be chosen to produce an ARGT at that subtree. Thus, each unranked species tree t other than caterpillars and pseudocaterpillars produces ARGTs.

Here, we focus on *ranked* species trees that produce ARGTs. That is, the ranking of the species tree is given and it cannot be carefully selected as in the unranked case studied by [9] and [10]. Formally, from [9, Propositions 9, 2, and 3], we borrow two facts:

(iii) If a ranked species tree t contains an NMP-subtree, that is, $t \in \mathcal{S} \setminus \mathcal{S}^{(mp)}$, then t produces ARGTs at the NMP-subtree.

(iv) If a ranked species tree t is either a caterpillar or a pseudocaterpillar, that is, $t \in \mathcal{S}^{(cat)}$, then t does not produce ARGTs.

As stated in [9], (iii) is only a sufficient condition for production of ARGTs and not a complete characterization of the set of ranked species trees that generate ARGTs. Because (iii) connects NMP-subtrees to the problem of counting ranked species trees that produce ARGTs, our interest is in counting ranked species trees containing or avoiding NMP-subtrees.

2.4 A Subtree Specified by the 1-2 Condition

Property (iii) states that being NMP at a given subtree implies producing ARGTs at that particular subtree. It is of interest to investigate not only the presence of ARGT-producing subtrees but also their position in the species tree. Here we introduce the set of ranked species trees t for which (iii) ensures production of ARGTs at the largest subtree $H(t)$ that satisfies the 1-2 condition. In particular, for any ranked species tree t, there is no NMP-subtree that properly contains $H(t)$. It is by examining the ranking of $H(t)$ that [9] showed that with the exception of caterpillars and pseudocaterpillars, each unranked species tree produces ARGTs.

The subtree $H(t)$ can be defined by a recursive query procedure: starting from the root of the tree t, if the current node satisfies the 1-2 condition, then stop and set $H(t)$ equal to the subtree rooted at the current node. Otherwise, at the current node, the tree splits into two subtrees that either both have size smaller than 2, or exactly one of them has size smaller than 1. In the first case, stop the procedure and set $H(t)$ empty. In the second case, query the node whose subtree has size at least 2. Observe that $H(t)$ is empty if and only if t is either a caterpillar or a pseudocaterpillar. The symbol $\mathcal{S}^{(H)}$ denotes the set of ranked species trees t that are MP at $H(t)$. The tree in Fig. 2 belongs to $\mathcal{S}^{(H)}$ but not to $\mathcal{S}^{(mp)}$; the subtree $H(t)$ is, in this case, the subtree generated by the root.

As was observed in [9],

$$\mathcal{S}^{(cat)} \subseteq \mathcal{S}^{(mp)} \subseteq \mathcal{S}^{(H)}. \tag{4}$$

Thus, $|\mathcal{S}_n| - |\mathcal{S}_n^{(H)}|$ bounds from below the cardinality of $\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}$, also providing a lower bound for the ultimate quantity of interest, the number of ranked species trees that produce ARGTs.

3 Results

We now present enumerative results for the classes of ranked species trees that we have introduced. In Section 3.2, we show that the probability that a randomly selected ranked species tree of size n produces ARGTs approaches 1 as n becomes large. In Section 3.3, we obtain the enumeration of the set $\mathcal{S}_n^{(H)}$. Section 3.4 provides a recursion to enumerate $\mathcal{S}_n^{(mp)}$. The recursion enables a closed formula that bounds from below the number of ARGT-producing ranked species trees of size n. First, in Section 3.1, we obtain a result that allows us to switch our perspective between ranked species trees and ordered ranked trees.

3.1 Equivalence between Ranked Species Trees and Ordered Ranked Trees

Observe that the subtree patterns defining $S^{(cat)}$, $S^{(mp)}$, and $S^{(H)}$ do not depend on the leaf labeling, and only consider the ranking of the internal nodes. To simplify our computations, we focus on ordered ranked trees instead of ranked species trees, using an equivalence to convert results about ordered ranked trees into results about ranked species trees. If P is a tree property that does not concern labeling of taxa but only concerns the ranking of the coalescence events—such as avoiding NMP-subtrees, for instance—then the two sets of trees can be treated as equivalent. More precisely, we have the following:

Proposition 1. *If P is a tree property that depends only on the ranking of the coalescence events, then*

$$\frac{|\{t \in \mathcal{R}_n : P(t)\}|}{n!} = \frac{|\{t \in \mathcal{S}_n : P(t)\}|}{(n+1)!\, n!/2^n}. \quad (5)$$

Proof. Define an equivalence relation $\approx_o$ on the set of ordered ranked trees of the same size, so that $t_a \approx_o t_b$ when t_b can be obtained from t_a by switching pairs of subtrees appended to corresponding nodes—in other words, if, ignoring left-right orientation, t_a and t_b represent the same ranked tree. Similarly, define the equivalence relation $\approx_s$ on the set of ranked species trees of the same size, so that $t_c \approx_s t_d$ if t_c and t_d represent the same ranked tree once labels for the leaves have been removed.

On the set of ranked trees of size n, consider the probability distribution induced by the Yule model of random branching. Under this model, the probability of a ranked tree t depends on two parameters: the size n and the number of subtrees of size 1 (i.e., cherries), denoted by $c(t)$. We have $P_{\text{Yule}}(t) = 2^{n-c(t)}/n!$, as in [22, Theorem 3.4] (see also [18], [28]).

Observe that for a fixed ordered ranked tree t of size n, the cardinality of the equivalence class $[t]_{\approx_o}$ is given by $2^{n-c(t)}$ because switching left and right subtrees at the root of a subtree of size greater than 1 is the only way to produce a different ordered ranked tree. Similarly, if we fix a ranked species tree t, then the cardinality of $[t]_{\approx_s}$ is $(n+1)!/2^{c(t)}$. Indeed, each of the $(n+1)!$ possible permutations of the leaf labels of t gives exactly $2^{c(t)}$ equivalent labelings of the taxa.

It follows that if we fix the size n, then the uniform distribution over the set of ordered ranked trees and the uniform distribution over the set of ranked species trees induce the same probability distribution—the Yule distribution—over the set of ranked trees. In particular, the probability of a ranked tree under the Yule model is given by the cardinality of the corresponding equivalence class in $\approx_o$ divided by $n!$, or by the cardinality of the equivalence class in $\approx_s$ divided by $(n+1)!\, n!/2^n$.

Finally, observe that the property P respects the equivalence classes defined under $\approx_o$ and $\approx_s$ in the sense that an ordered ranked tree (resp. ranked species tree) t satisfies P if and only if all the ordered ranked trees (resp. ranked species trees) in the equivalence class $[t]_{\approx_o}$ (resp. $[t]_{\approx_s}$) satisfy P.

We can then write

$$\begin{aligned}\frac{|\{t \in \mathcal{R}_n : P(t)\}|}{n!} &= \sum_{[t]_{\approx_o} : P(t)} \frac{|[t]_{\approx_o}|}{n!} \\ &= \sum_{[t]_{\approx_s} : P(t)} \frac{|[t]_{\approx_s}|}{(n+1)!\, n!/2^n} \\ &= \frac{|\{t \in \mathcal{S}_n : P(t)\}|}{(n+1)!\, n!/2^n}.\end{aligned}$$

□

In the framework of ordered trees, we define $\mathcal{R}^{(mp)}$, $\mathcal{R}^{(H)}$, and $\mathcal{R}^{(cat)}$ as corresponding versions of the classes $\mathcal{S}^{(mp)}$, $\mathcal{S}^{(H)}$, and $\mathcal{S}^{(cat)}$, respectively. Indeed, our definitions for sets $\mathcal{S}^{(x)}$ did not depend on the left-right orientation of subtrees. Therefore, the same definitions apply to ordered ranked trees to define the associated $\mathcal{R}^{(x)}$. To determine the cardinality of a set $\mathcal{S}_n^{(x)} \subseteq \mathcal{S}_n$, our approach consists of finding the cardinality of the corresponding ordered set $\mathcal{R}_n^{(x)} \subseteq \mathcal{R}_n$ and then applying (5) to obtain

$$|\mathcal{S}_n^{(x)}| = \frac{(n+1)!}{2^n}\, |\mathcal{R}_n^{(x)}|. \quad (6)$$

3.2 Probability that a Ranked Species Tree Produces ARGTs

We are now ready to show that the probability that a randomly selected ranked species tree of size n produces ARGTs approaches 1 as n becomes large. It is useful to introduce the sequence α_n, defined as

$$\alpha_n = \sum_{q=1}^{n-1} \frac{2^{\min(q,n-q)}}{\binom{n}{q}}. \quad (7)$$

Considering $q = 1$ and $q = n-1$ in the sum, we find

$$\alpha_n \geq 4/n. \quad (8)$$

We also have

$$\alpha_n \leq 2 \sum_{q=1}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{n}{q}} \leq 2 \sum_{q=1}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{2\lfloor n/2 \rfloor}{q}} = 2 s_{\lfloor n/2 \rfloor}.$$

The sequence

$$s_n = \sum_{q=1}^{n} \frac{2^q}{\binom{2n}{q}}$$

can be bounded by

$$s_n \geq 1/n, \quad (9)$$

considering only the $q = 1$ term in the sum. Furthermore, s_n has the following property.

Lemma 1. *The sequence s_n satisfies the recursion*

$$9(2n+1)s_{n+1} - 4(2n+3)s_n = \frac{10n+9}{n+1} + \frac{n(2^{n+1})}{\binom{2n}{n}}. \quad (10)$$

Proof. Using the Wilf-Zeilberger summation approach [20], define $F(q,n) = 2^q/\binom{2n}{q}$ and

$$R(q,n) = \frac{(2n+1-q)[3q(2n+1) - 2(2n+1)(5n+6)]}{2(n+1)(2n+1)}.$$

It is easily verified that

$$\begin{aligned}&9(2n+1)F(q,n+1) - 4(2n+3)F(q,n) \\ &\quad = F(q+1,n)R(q+1,n) - F(q,n)R(q,n). \quad (11)\end{aligned}$$

Indeed, the identity follows by noting the ratios

$$\frac{F(q,n+1)}{F(q,n)} = \frac{(2n+2-q)(2n+1-q)}{2(n+1)(2n+1)}$$

and

$$\frac{F(q+1,n)}{F(q,n)} = \frac{2(q+1)}{2n-q}.$$

Summing both sides of (11) from $q=1$ to $n+1$, the right-hand side telescopes, giving a final contribution of $F(n+2,n)R(n+2,n) - F(1,n)R(1,n)$. Therefore,

$$\begin{aligned} &9(2n+1)s_{n+1} - 4(2n+3)\left[s_n + \frac{2^{n+1}}{\binom{2n}{n+1}}\right] \\ &\quad = \frac{10n+9}{n+1} - \frac{2^{2+n}(2n+1)(7n+6)(n-1)!\,(n+2)!}{(2n+2)!}, \end{aligned}$$

from which simple calculations lead to (10). □

Starting from (10), it can be shown by induction on n that for n large,

$$s_n \le \frac{n+10}{n(n-1)}. \tag{12}$$

Consider $n \ge 23$. We can easily verify (12) for $n = 23$. For the inductive step, we begin from a binomial inequality, which holds for $n \ge 1$ [4]:

$$\binom{2n}{n} \ge \frac{2^{2n-1}}{\sqrt{n}}. \tag{13}$$

We then have

$$\frac{n(2^{n+1})}{\binom{2n}{n}} \le \frac{4n^{3/2}}{2^n} \le \frac{n^4}{2^n} \le \frac{1}{n},$$

where the last inequality holds because $n \ge 23$. We can thus write

$$\begin{aligned} &9(2n+1)s_{n+1} - 4(2n+3)\left[\frac{n+10}{n(n-1)}\right] \\ &\quad \le 9(2n+1)s_{n+1} - 4(2n+3)s_n \le \frac{10n+9}{n+1} + \frac{1}{n}, \end{aligned}$$

from which

$$s_{n+1} \le \frac{18n^3+100n^2+203n+119}{9n(n-1)(n+1)(2n+1)}.$$

Finally, note that

$$\begin{aligned} &\frac{(n+1)+10}{(n+1)n} - s_{n+1} \\ &\quad \ge \frac{(n+1)+10}{(n+1)n} - \frac{18n^3+100n^2+203n+119}{9n(n-1)(n+1)(2n+1)} \\ &\quad = \frac{89n^2-311n-218}{9n(n-1)(n+1)(2n+1)}. \end{aligned}$$

This last quantity is positive for $n \ge 5$, completing the inductive proof of (12).

TABLE 1
Asymptotic Equivalence of α_n and $4/n$, with α_n Computed from (7)

	n				
	50	100	250	500	1000
α_n	0.08753	0.04172	0.01626	0.00806	0.00402
$4/n$	0.08000	0.04000	0.01600	0.00800	0.00400

Therefore, from (9) and (12), we have for $n \ge 23$,

$$\frac{1}{n} \le s_n \le \frac{n+10}{n(n-1)}, \tag{14}$$

producing, for n large, the asymptotic equivalence

$$s_n \sim 1/n.$$

Thus, for n large, by (8),

$$\frac{4}{n} \le \alpha_n \le 2s_{\lfloor n/2 \rfloor} \sim \frac{2}{\lfloor n/2 \rfloor}, \tag{15}$$

so that

$$\alpha_n \sim 4/n. \tag{16}$$

Table 1 illustrates this asymptotic equivalence of α_n and $4/n$ for a variety of values of n. It is from this asymptotic equivalence in (16) that the main result of this section follows.

Proposition 2. *The probability that a randomly selected ranked species tree with n internal nodes produces ARGTs approaches 1 as $n \to \infty$.*

Proof. Consider the number $c'_n + c''_n$ of ordered ranked trees of size n that are MP at their root. Here c'_n is the number of ordered ranked trees t of size n that are MP at their root and that have $H(t) = t$, and c''_n is the number of ordered ranked trees of size n that have $H(t) \ne t$ (and that are therefore MP at the root). The remaining $n! - c'_n - c''_n$ ordered ranked trees of size n are NMP at the root. Observe that, if $n \ge 3$, then

$$c'_{n+1} = \sum_{q=1}^{n-1} q!\,(n-q)!\,2^{\min(q,n-q)} = n!\,\alpha_n. \tag{17}$$

This result holds because each tree t counted in c'_{n+1} is built by appending two ordered ranked trees of sizes $1 \le q \le n-1$ and $n-q$ to a shared root. Once these subtrees are chosen, we choose one of the $2^{\min(q,n-q)}$ orderings that create an MP-subtree at the root of t to merge the rankings of the subtrees of sizes q and $n-q$. This value is obtained by noting from the definition of MP-subtrees that for t to be MP at the root, the coalescence sequence for t must have the form (3) once names L and R have been assigned to the two subtrees of the root in such a way that $|R| = \min(q, n-q)$. The number of sequences satisfying (3) is $2^{|R|} = 2^{\min(q,n-q)}$.

Moreover, we have

$$c''_{n+1} = 2n!, \tag{18}$$

because each tree counted in c''_{n+1} has a leaf—a subtree of size 0—appended to the root, and its other subtree of the

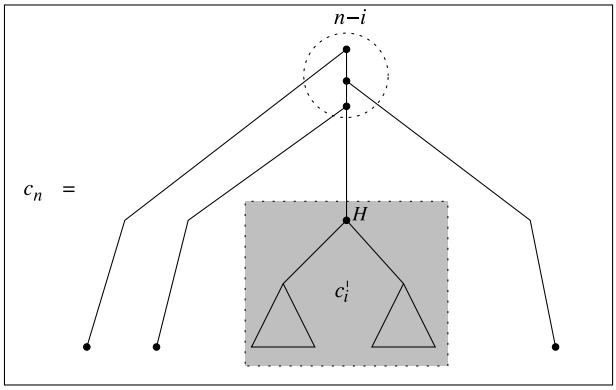

Fig. 4. Decomposition of an ordered ranked tree t of size n that is in $\mathcal{R}_n^{(H)}$, is neither a caterpillar nor a pseudocaterpillar, and has subtree $H(t)$ maximally probable. The highlighted subtree is used for c_i' in computing (19).

root has size n. The factor of 2 arises because the leaf can appear on either side of the root.

By (6), because ranked species trees that are NMP at their root produce ARGTs, $(n+1)!\,(n! - c_n' - c_n'')/2^n$ gives a lower bound for the number of ranked species trees of size n producing ARGTs. Dividing by the number of ranked species trees of size n (1), by (17) and (18), we obtain

$$1 - \frac{c_n' + c_n''}{n!} = 1 - \frac{(n-1)!\,\alpha_{n-1} + 2(n-1)!}{n!} = 1 - \frac{\alpha_{n-1} + 2}{n}.$$

By (16), this value nears 1 as n becomes large. □

3.3 Ranked Species Trees t that Are NMP at the Subtree $H(t)$

We have shown that the fraction of ranked species trees t that are NMP at subtree $H(t)$ approaches 1 as $n \to \infty$. In this section, we extend beyond this result to enumerate the set of ranked species trees that are NMP at $H(t)$. We achieve the result by counting ordered ranked trees t that are MP at $H(t)$.

Let c_n be the number of ordered ranked trees t of size n that are neither caterpillar nor pseudocaterpillar and that have the property that the subtree $H(t)$ is MP at its root. For $n \geq 4$, the smallest number of internal nodes for which a tree can be neither a caterpillar nor a pseudocaterpillar, we have

$$c_n = \sum_{i=4}^{n} (c_i') 2^{n-i}, \tag{19}$$

where c_i' is, as in the proof of Proposition 2, the number of trees t of size i that are MP at their root and that have $H(t) = t$. The result is obtained by noting that each tree t counted in c_n is constructed from a tree in c_i', with $4 \leq i \leq n$, which reaches the root of t through a branch to which $n - i$ leaves are appended (Fig. 4). The leaves can be placed on either the right or the left of the branch, producing the factor 2^{n-i}.

Observe that the number of caterpillar or pseudocaterpillar ordered ranked trees is given by

$$|\mathcal{R}_n^{(cat)}| = 3 \cdot 2^{n-2}. \tag{20}$$

In particular, we have 2^{n-1} caterpillar ordered ranked trees obtained by the possible left-right orientations of the leaves stemming from $n-1$ of the coalescences (all coalescences except the root of the cherry). Similarly, we have 2^{n-2} pseudocaterpillar ordered ranked trees, considering the two possible left-right orientations of all coalescences except the roots of the two cherries. Therefore, $|\mathcal{R}_n^{(H)}| = c_n + |\mathcal{R}_n^{(cat)}| = c_n + 3 \cdot 2^{n-2}$.

The sequence c_n' can be computed as in (17). Using (19), we obtain

$$c_n = \sum_{i=4}^{n} (i-1)!\,\alpha_{i-1}\, 2^{n-i} = 2^n \left(\sum_{i=4}^{n} (i-1)!\,\alpha_{i-1}\, 2^{-i} \right). \tag{21}$$

Using $|\mathcal{R}_n^{(H)}|$ with (6), we can compute the number of ranked species trees t that are MP at subtree $H(t)$.

Proposition 3. *The number of ranked species trees t with n internal nodes that are MP at the subtree $H(t)$ is*

$$\begin{aligned} |\mathcal{S}_n^{(H)}| &= \frac{(n+1)!\,|\mathcal{R}_n^{(H)}|}{2^n} \\ &= \frac{(n+1)!\,(c_n + 3 \cdot 2^{n-2})}{2^n} \\ &= \frac{(n+1)!}{2^n} \\ &\quad \times \left[2^n \left(\sum_{i=4}^{n} (i-1)!\,\alpha_{i-1}\, 2^{-i} \right) + 3 \cdot 2^{n-2} \right], \end{aligned} \tag{22}$$

where α_n can be computed as in (7).

The number of ranked species trees t with n internal nodes that are NMP at the subtree $H(t)$ is

$$\begin{aligned} |\mathcal{S}_n| - |\mathcal{S}_n^{(H)}| &= \frac{(n+1)!}{2^n} \\ &\quad \times \left[n! - 2^n \left(\sum_{i=4}^{n} (i-1)!\,\alpha_{i-1}\, 2^{-i} \right) - 3 \cdot 2^{n-2} \right]. \end{aligned} \tag{23}$$

Bounds. By Proposition 3, the exact number of ranked species trees t that are NMP at the subtree $H(t)$ can be computed. From Proposition 2, the probability that a randomly selected ranked species tree t is NMP at $H(t)$ approaches 1 as n grows large. Here we provide upper and lower bounds for the speed of convergence.

Observe that (19) implies that $c_n \geq c_n'$. Using (17) and (8), we can write

$$\begin{aligned} \frac{|\mathcal{R}_n^{(H)}|}{n!} \geq \frac{c_n'}{n!} &= \frac{(n-1)!\,\alpha_{n-1}}{n!} \\ &\geq \frac{(n-1)!\,4/(n-1)}{n!} = \frac{4}{n(n-1)} \geq \frac{4}{n^2}. \end{aligned} \tag{24}$$

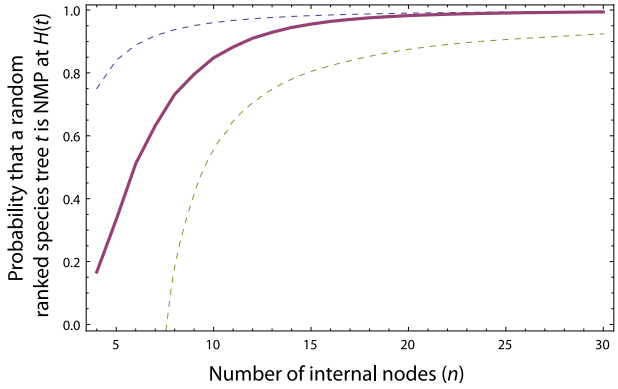

Fig. 5. The probability that subtree $H(t)$ is non-maximally probable in a randomly selected ranked species tree t with n nodes, or $1-\mathcal{R}_n^{(H)}/n!$. The probability is confined by lower bound $1-2(n^2-4n+46)/\,[n(n-2)(n-4)]$ and upper bound $1-4/n^2$.

On the other hand, given that $|\mathcal{R}_n^{(H)}|-c'_n$ counts a set of trees for which $H(t)\neq t$, we must have $|\mathcal{R}_n^{(H)}|-c'_n\leq c''_n=2(n-1)!$, where c''_n is as in (18) and corresponds to the number of ordered ranked trees t with $H(t)\neq t$. Dividing by $n!$ and using inequalities (14) and (15) gives

$$\begin{aligned}\frac{|\mathcal{R}_n^{(H)}|}{n!}&\leq\frac{c''_n+c'_n}{n!}=\frac{2(n-1)!+\alpha_{n-1}(n-1)!}{n!}\\&=\frac{2+\alpha_{n-1}}{n}\\&\leq\frac{1}{n}\left[2+2\left(\frac{\lfloor(n-1)/2\rfloor+10}{\lfloor(n-1)/2\rfloor(\lfloor(n-1)/2\rfloor-1)}\right)\right]\\&\leq\frac{1}{n}\left[2+2\left(\frac{(n-1)/2+10}{((n-2)/2)((n-2)/2-1)}\right)\right]\\&=\frac{2(n^2-4n+46)}{n(n-2)(n-4)}.\end{aligned}\tag{25}$$

When n becomes large, the value

$$\frac{|\mathcal{S}_n\setminus\mathcal{S}_n^{(H)}|}{|\mathcal{S}_n|}=1-\frac{|\mathcal{R}_n^{(H)}|}{n!},$$

that is, the probability that a randomly selected ranked species tree t is NMP at $H(t)$, approaches 1 at most as fast as $1-4/n^2$ (24) and at least as fast as $1-2(n^2-4n+46)/\,[n(n-2)(n-4)]$ (25).

Fig. 5 plots the exact value of $1-|\mathcal{R}_n^{(H)}|/n!$ with its bounds. The probability that a randomly selected ranked species tree t is NMP at $H(t)$—and that it therefore produces ARGTs at $H(t)$—approaches 1 quickly. Moreover, the upper bound appears to approximate the probability more accurately than does the lower bound.

3.4 Ranked Species Trees that Are NMP for at Least One Subtree

The set $\mathcal{S}_n\setminus\mathcal{S}_n^{(mp)}$—ranked species trees of size n containing at least one NMP-subtree—is a superset of $\mathcal{S}_n\setminus\mathcal{S}_n^{(H)}$, and it thus expands the class of ARGT-producing ranked gene trees beyond the set $\mathcal{S}_n\setminus\mathcal{S}_n^{(H)}$. In this section we provide a recursion to compute the cardinality of $\mathcal{S}_n\setminus\mathcal{S}_n^{(mp)}$. We also determine a more accurate lower bound for the number of ranked species trees that are ARGT-producing.

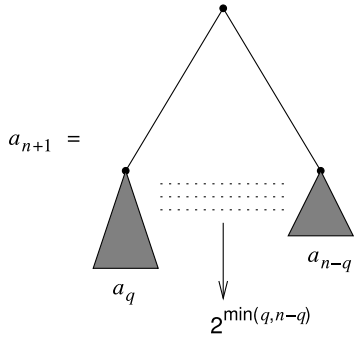

Fig. 6. Decomposition of an ordered ranked tree of size $n+1$ that is in $\mathcal{R}_{n+1}^{(mp)}$ and has no non-maximally probable subtrees. The two subtrees of the root are taken from $\mathcal{R}^{(mp)}$, with sizes q and $n-q$. According to the definition of maximally probable subtrees, once names L and R are assigned to the two subtrees in such a way that $|R|=\min(q,n-q)$, the number of possible rankings to obtain a sequence of coalescences satisfying (3) is $2^{|R|}$.

We first focus on the class $\mathcal{R}_n^{(mp)}$ of ordered ranked trees of size n avoiding NMP-subtrees. Next, using (6), we convert the result to obtain $|\mathcal{S}_n\setminus\mathcal{S}_n^{(mp)}|$. Let $a_n=|\mathcal{R}_n^{(mp)}|$. Each tree in $\mathcal{R}_{n+1}^{(mp)}$ is obtained by appending to the same root two trees belonging to $\mathcal{R}^{(mp)}$, one of size q and the other of size $n-q$, with $0\leq q\leq n$. As was already noticed in the proof of Proposition 2, when merging the rankings of the two subtrees of the root, exactly $2^{\min(q,n-q)}$ among the $\binom{n}{q}$ possible choices create an MP-subtree at the root. Recall that once we have assigned the names L and R to the two subtrees of the root in such a way that $|R|=\min(q,n-q)$, the number of possible rankings to obtain a sequence of coalescences of the form (3) is $2^{|R|}$. The decomposition is illustrated in Fig. 6.

The recursion to compute a_n is thus

$$a_{n+1}=\sum_{q=0}^{n}(a_qa_{n-q})2^{\min(q,n-q)},\tag{26}$$

where $a_0=1$. Taking a_n and using property (6), we can obtain the cardinality of $\mathcal{S}_n^{(mp)}$.

Proposition 4. *The number of ranked species trees with n internal nodes that contain only MP-subtrees is*

$$|\mathcal{S}_n^{(mp)}|=\frac{(n+1)!\,|\mathcal{R}_n^{(mp)}|}{2^n}=\frac{(n+1)!\,a_n}{2^n}.\tag{27}$$

The number of ranked species trees with n internal nodes that contain at least one NMP-subtree is

$$|\mathcal{S}_n|-|\mathcal{S}_n^{(mp)}|=\frac{(n+1)!}{2^n}(n!-a_n).\tag{28}$$

An explicit formula for $|\mathcal{S}_n\setminus\mathcal{S}_n^{(mp)}|$ requires a solution of recursion (26). Although we have not obtained such a solution, we can use the recursion to find a closed-form upper bound for a_n and therefore a lower bound for the number of ranked species trees that produce ARGTs. For large n, this bound is more accurate than the bound given by the number of ranked species trees t that are NMP at subtree $H(t)$ (Proposition 3).

Bounds. Fix a parameter β, $1/2 < \beta < 1$. Observe that

$$\sum_{q=0}^{\infty} \frac{q}{2^{q(2\beta-1)}} = \sum_{q=1}^{\infty} \frac{q}{2^{q(2\beta-1)}}$$

converges to a constant

$$k_\beta = \frac{1}{2^{2\beta-1}(1-1/2^{2\beta-1})^2}. \tag{29}$$

This result is obtained by noting that

$$\sum_{q=0}^{\infty} \frac{z^q}{v^q} = \frac{1}{1-z/v},$$

differentiating both sides with respect to z, setting $v = 2^{2\beta-1}$, and choosing $z = 1$. For instance, when $\beta = 6/11$, we have $k_\beta \approx 251.762$.

When $0 < q \le n/2$, by (13), $\binom{n}{q} \ge \binom{2q}{q} \ge 2^{2q-1}/\sqrt{q}$. For every n, we then have a bound:

$$\begin{aligned}\sum_{q=0}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{n}{q}^\beta} &= 1 + \sum_{q=1}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{n}{q}^\beta} \le 1 + \sum_{q=1}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{2q}{q}^\beta} \\ &\le 1 + \sum_{q=1}^{\infty} \frac{2^q}{\binom{2q}{q}^\beta} \le 1 + \sum_{q=1}^{\infty} \frac{2^q}{\left(\frac{2^{2q}}{2\sqrt{q}}\right)^\beta} \\ &= 1 + 2^\beta \sum_{q=1}^{\infty} \frac{q^{\beta/2}}{2^{q(2\beta-1)}} \le 1 + 2^\beta k_\beta.\end{aligned} \tag{30}$$

Choose n to be a positive integer such that $2(1+2^\beta k_\beta) \le (n+1)^\beta$. Let $c_\beta \ge 1$ be a constant such that for all i, $0 \le i \le n$, we have $a_i \le c_\beta^i (i!)^\beta$. Note that the existence of c_β is ensured because we could set, for instance, $c_\beta = \max\{a_i : 0 \le i \le n\}$. Thus, for such a constant we have both of the following conditions:

$$a_i \le c_\beta^i (i!)^\beta \text{ for all } i, 0 \le i \le n \tag{31}$$

$$2(1+2^\beta k_\beta) \le c_\beta (n+1)^\beta. \tag{32}$$

We can now prove by induction that if conditions (31) and (32) are both satisfied for a certain n, then they also hold for $n+1$. For the second condition, the result is trivial. For the first condition, we use (26):

$$\begin{aligned} a_{n+1} &\le 2 \sum_{q=0}^{\lfloor n/2 \rfloor} a_q a_{n-q} 2^q \le 2 c_\beta^n \sum_{q=0}^{\lfloor n/2 \rfloor} (q!)^\beta (n-q)!^\beta 2^q \\ &= \frac{2 c_\beta^{n+1} (n+1)!^\beta}{c_\beta (n+1)^\beta} \sum_{q=0}^{\lfloor n/2 \rfloor} \frac{2^q}{\binom{n}{q}^\beta} \\ &\le c_\beta^{n+1} (n+1)!^\beta \, \frac{2(1+2^\beta k_\beta)}{c_\beta (n+1)^\beta} \\ &\le c_\beta^{n+1} (n+1)!^\beta. \end{aligned} \tag{33}$$

We have therefore proven the following result.

Proposition 5. *Choose β with $1/2 < \beta < 1$. Take a positive constant c_β and define*

$$X = X(\beta, c_\beta) = \left[2(1+2^\beta k_\beta)/c_\beta\right]^{1/\beta} - 1.$$

Suppose it can be verified that for every integer n with $0 \le n \le X$,

$$a_n \le c_\beta^n (n!)^\beta. \tag{34}$$

Then for every $n \ge 0$, $a_n \le c_\beta^n (n!)^\beta$, and therefore, the number of ranked species trees with n internal nodes that produce ARGTs is at least

$$|\mathcal{S}_n| - |\mathcal{S}_n^{(mp)}| \ge \frac{(n+1)!}{2^n}[n! - (c_\beta)^n (n!)^\beta]. \tag{35}$$

The upper bound for a_n contained in Proposition 5 shows that for n large, the number of ranked species trees that contain only MP-subtrees is much smaller than the number of ranked species trees t that are MP at $H(t)$. Indeed, from (24) we have that $|\mathcal{R}_n^{(H)}| \ge 4n!/n^2$, and therefore, for any $1/2 < \beta < 1$,

$$\frac{|\mathcal{S}_n^{(H)}|}{|\mathcal{S}_n^{(mp)}|} \ge \frac{(4\,n!)/n^2}{c_\beta^n (n!)^\beta} = \frac{4(n!)^{1-\beta}}{n^2 c_\beta^n} \to \infty.$$

The constants c_β in Proposition 5 can be evaluated numerically. If we fix, for instance, $\beta = 6/11$, then we have $k_\beta \approx 251.762$ as noted above. In this case, setting $c_\beta = c_{6/11} = 5$, we have $X(\beta, c_\beta) \approx 9449.7$. We can then computationally verify that condition (34) is satisfied for every n, $0 \le n \le 9449$. Thus, with $\beta = 6/11$ and $c_\beta = 5$, (34) holds for every $n \ge 0$. An efficient implementation of recursion (26) can be achieved by saving each a_j once computed, to minimize the number of calls to the recursive steps.

4 Conclusions

We have examined three nested classes of ranked species trees (4) characterized by the presence or absence of particular subtree patterns: $\mathcal{S}_n \setminus \mathcal{S}_n^{(H)}$, a class of ranked species trees proven by Degnan et al. [9] to produce ARGTs; $\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}$, a larger class that by extension of their proof was identified as producing ARGTs; and the still larger class $\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}$ that excludes caterpillar and pseudocaterpillar ranked species trees proven by [9] *not* to produce ARGTs.

Extending beyond the result of [9] that for each unranked species tree—with the exception of caterpillars and pseudocaterpillars—at least one ranking exists that gives rise to ARGTs, we have demonstrated that as $n \to \infty$, *almost all* ranked species trees with n internal nodes give rise to ARGTs (Proposition 2). We have additionally provided a closed-form for the cardinality $|\mathcal{S}_n \setminus \mathcal{S}_n^{(H)}|$ (23) and a recursion as well as a closed-form lower bound for $|S_n \setminus \mathcal{S}_n^{(mp)}|$ (28, 35).

TABLE 2
Cardinalities of Sets of Ranked Species Trees with n Internal Nodes

Class of trees	Equation	n = 4	5	6	7	8	9	10
$\mathcal{S}_n$	(1)	180	2,700	56,700	1,587,600	57,153,600	2,571,912,000	141,455,160,000
$\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}$	(36)	90	2,160	52,920	1,557,360	56,881,440	2,569,190,400	141,425,222,400
$\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}$	(28)	30	900	30,240	1,083,600	46,176,480	2,278,886,400	132,773,256,000
$\mathcal{S}_n \setminus \mathcal{S}_n^{(H)}$	(23)	30	900	28,980	1,002,960	41,821,920	2,047,096,800	119,964,952,800
$\mathcal{S}_n^{(cat)}$	(36)	90	540	3,780	30,240	272,160	2,721,600	29,937,600

$\mathcal{S}_n^{(cat)}$ is the set containing caterpillar and pseudocaterpillar ranked species trees. $\mathcal{S}_n \setminus \mathcal{S}_n^{(H)}$ is the set of trees t for which the subtree $H(t)$ is NMP. $\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}$ is the set of trees containing at least one NMP-subtree. $\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}$ is the set of trees that are neither caterpillar nor pseudocaterpillar. $\mathcal{S}_n$ is the set of ranked species trees.

For illustration, Table 2 shows the cardinalities for small n, alongside the total number of ranked species trees $|S_n|$, the upper bound $|\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}|$ on the number of ranked species trees with ARGTs, and the lower bound $\mathcal{S}_n^{(cat)}$ on the number of ranked species trees without ARGTs. The row for $\mathcal{S}_n \setminus \mathcal{S}_n^{(H)}$ extends a corresponding enumeration in Table 1 of [9], correcting an error in the $n = 7$ case ($n = 8$ in [9], which indexed cases by the number of leaves rather than the number of internal nodes). It can be observed from the table that the quantities in the central row increase quite quickly with n when considered as a fraction of $|\mathcal{S}_n|$.

The problem of characterizing the set of ranked species trees that produce ARGTs is analogous to the corresponding problem of characterizing the set of unranked species trees that produce anomalous unranked gene trees in the unranked case [7], [23], [24]. In that context, every species tree with four or more species, as well as the caterpillar species tree with four species, produces anomalous unranked gene trees [7]. Our work extends the analogy: for large n, not only does almost every unranked species tree have a ranking that produces ARGTs, almost every *ranked* species trees produces ARGTs. The related characterization in the unranked case has been useful in facilitating the development of species tree inference methods and the design of simulation-based tests relying on unranked gene trees [23], and we expect our results to serve in a similar role in the ranked case.

We note that we have not fully completed the characterization of ranked species trees that produce ARGTs, a problem that was left open by [9]. We have, however, shown that the work of [9] implies that among all ranked species trees with n internal nodes, the fraction that produce ARGTs approaches 1—and approaches it quickly. Our recursion for $|\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}|$ as well as (23) and (35) provide lower bounds for the number of ranked species trees with n internal nodes that are ARGT-producing. An upper bound is provided by the cardinality of the set of ranked species trees excluding only the caterpillars and pseudocaterpillars, or

$$|\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}| = [(n+1)!/2^n](n! - 3 \cdot 2^{n-2}), \tag{36}$$

where $|\mathcal{S}_n^{(cat)}| = [(n+1)!/2^n](3 \cdot 2^{n-2})$. For the unsolved complete characterization of ranked species trees that produce ARGTs, the exact value must lie in a narrow range bounded between $|\mathcal{S}_n \setminus \mathcal{S}_n^{(mp)}|$ and $|\mathcal{S}_n \setminus \mathcal{S}_n^{(cat)}|$.

Acknowledgments

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

References

[1] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Determining species tree topologies from clade probabilities under the coalescent," *J. Theor. Biol.*, vol. 289, pp. 96-106, 2011.
[2] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent," *J. Math. Biol.*, vol. 62, pp. 833–862, 2009.
[3] J. K. M. Brown, "Probabilities of evolutionary trees," *Syst. Biol.*, vol. 43, pp. 78–91, 1994.
[4] P. S. Bullen, A Dictionary of Inequalities. Harlow, U.K.: Addison Wesley, 1998.
[5] J. H. Degnan, "Anomalous unrooted gene trees," *Syst. Biol.*, vol. 62, pp. 574–590, 2013.
[6] J. H. Degnan, M. DeGiorgio, D. Bryant, and N. A. Rosenberg, "Properties of consensus methods for inferring species trees from gene trees," *Syst. Biol.*, vol. 58, pp. 35–54, 2009.
[7] J. H. Degnan and N. A. Rosenberg, "Discordance of species trees with their most likely gene trees," *PLoS Genet.*, vol. 2, pp. 762–768, 2006.
[8] J. H. Degnan and N. A. Rosenberg, "Gene tree discordance, phylogenetic inference and the multispecies coalescent," *Trends Ecol. Evol.*, vol. 24, pp. 332–340, 2009.
[9] J. H. Degnan, N. A. Rosenberg, and T. Stadler, "A characterization of the set of species trees that produce anomalous ranked gene trees," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 6, pp. 1558–1568, Nov./Dec. 2012.
[10] J. H. Degnan, N. A. Rosenberg, and T. Stadler, "The probability distribution of ranked gene trees on a species tree," *Math. Biosci.*, vol. 235, pp. 45–55, 2012.
[11] J. H. Degnan and L. A. Salter, "Gene tree distributions under the coalescent process," *Evolution*, vol. 59, pp. 24–37, 2005.
[12] F. Disanto, A. Schlizio, and T. Wiehe, "Yule-generated trees constrained by node imbalance," *Math. Biosci.*, vol. 246, pp. 139–147, 2013.
[13] A. W. F. Edwards, "Estimation of the branch points of a branching diffusion process," *J. Roy. Stat. Soc. Ser. B*, vol. 32, pp. 155–174, 1970.
[14] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*. Cambridge, MA, USA: Cambridge Univ. Press, 2009.
[15] E. F. Harding, "The probabilities of rooted tree-shapes generated by random bifurcation," *Adv. Appl. Probab.*, vol. 3, pp. 44–77, 1971.
[16] J. Hein, M. H. Schierup, and C. Wiuf, *Gene Genealogies, Variation and Evolution*. Oxford, U.K.: Oxford Univ. Press, 2005.
[17] W. P. Maddison, "Gene trees in species trees," *Syst. Biol.*, vol. 46, pp. 523–536, 1997.
[18] R. D. M. Page, "Random dendograms and null hypotheses in cladistic biogeography," *Syst. Zool*, vol. 40, pp. 54–62, 1991.
[19] P. Pamilo and M. Nei, "Relationships between gene trees and species trees," *Mol. Biol. Evol.*, vol. 5. pp. 568–583, 1988.
[20] M. Petkovšek, H. S. Wilf, and D. Zeilberger, A=B. Wellesley, MA, USA: Peters, 1996.

[21] N. A. Rosenberg, "The probability of topological concordance of gene trees and species trees," *Theor. Popul. Biol.*, vol. 61, pp. 225–247, 2002.
[22] N. A. Rosenberg, "The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees," *Ann. Combinatorics*, vol. 10, pp. 129–146, 2006.
[23] N. A. Rosenberg, "Discordance of species trees with their most likely gene trees: A unifying principle," *Mol. Biol. Evol.*, vol. 30, pp. 2709–2713, 2013.
[24] N. A. Rosenberg and R. Tao, "Discordance of species trees with their most likely gene trees: The case of five taxa," *Syst. Biol.*, vol. 57, pp. 131–140, 2008.
[25] Y. S. Song, "Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees," *Ann. Combinatorics*, vol. 10, pp. 147–163, 2006.
[26] T. Stadler and J. H. Degnan, "A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree," *Algorithms Mol. Biol.* vol. 7, article 7, 2012.
[27] M. Steel and A. McKenzie, "Properties of phylogenetic trees generated by Yule-type speciation models," *Math. Biosci.*, vol. 170, pp. 91–112, 2001.
[28] F. Tajima, "Evolutionary relationship of DNA sequences in finite populations," *Genetics*, vol. 105, pp. 437–460, 1983.
[29] J. Wakeley, *Coalescent Theory: An Introduction*. Greenwood Village, CO, USA: Roberts, 2009.
[30] Y. Wu, "Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood," *Evolution*, vol. 66, pp. 763–775, 2012.
[31] G. U. Yule, "A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F. R. S," *Philos. Trans. Roy. Soc. London Ser. B*, vol. 213, pp. 21–87, 1924.

Filippo Disanto received the PhD degree in theoretical computer science from both the University of Siena and the University of Paris VII in 2010. After receiving the PhD degree, he was a postdoc at CNRS in Montpellier and at the Institut für Genetik, University of Cologne. Since November 2013, he has been a postdoc in the Rosenberg Laboratory, Stanford University. His main research interests include combinatorics and its applications.

Noah A. Rosenberg received the PhD degree in biological sciences from Stanford University in 2001 and completed postdoctoral training at the University of Southern California. He was on the faculty of the University of Michigan from 2005 to 2011, and he is currently a professor in the Department of Biology at Stanford University. His research interests include human evolutionary genetics, population-genetic theory, and mathematical phylogenetics.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.