

Counting Coalescent Histories

NOAH A. ROSENBERG

ABSTRACT

Given a species tree and a gene tree, a valid coalescent history is a list of the branches of the species tree on which coalescences in the gene tree take place. I develop a recursion for the number of valid coalescent histories that exist for an arbitrary gene tree/species tree pair, when one gene lineage is studied per species. The result is obtained by defining a concept of m -extended coalescent histories, enumerating and counting these histories, and taking the special case of $m = 1$. As a sum over valid coalescent histories appears in a formula for the probability that a random gene tree evolving along the branches of a fixed species tree has a specified labeled topology, the enumeration of valid coalescent histories can considerably reduce the effort required for evaluating this formula.

Key words: genealogy, labeled history, labeled topology, lineage sorting, phylogeny.

1. INTRODUCTION

PROBABILITY DISTRIBUTIONS under mathematical models for the shapes of gene trees that evolve along the branches of species trees can provide information about the nature of evolutionary descent among species, and can inform strategies for the accurate inference of species phylogenies from genetic data (Degnan, 2005; Degnan and Rosenberg, 2006; Degnan and Salter, 2005; Maddison, 1997; Pamilo and Nei, 1988; Rannala and Yang, 2003; Rosenberg, 2002; Slatkin and Pollack, 2006; Takahata, 1989).

In a significant advance, Degnan and Salter (2005) derived a general formula under a standard model of gene tree evolution for the conditional probability that a random gene tree has a specified labeled topology, given a species tree with fixed labeled topology and branch lengths and assuming that a single gene lineage is sampled from each species. The innovation of the Degnan-Salter formula—which is applicable to any number of taxa—was the recognition that the desired probability could be decomposed into a sum over a collection of coalescent histories, each of which specifies a possible list of branches of the species tree on which coalescences in the gene tree can take place. For each coalescent history, the probability that the gene tree has both the required list of coalescences and the specified labeled topology is computed, and these probabilities are summed over all coalescent histories compatible with the species tree and gene tree labeled topologies.

Degnan and Salter (2005) developed a method by which coalescent histories are proposed from a broad set of hypothetical histories, and are then evaluated to determine their compatibility with the specified species tree and gene tree labeled topologies. Coalescent histories that could conceivably provide the list of species tree branches on which gene tree coalescences can occur—“valid” coalescent histories—are

Department of Human Genetics, Bioinformatics Program, and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan.

identified in the larger set of proposed histories, and the sum of probabilities proceeds only over those coalescent histories that are valid.

This method of computation expends considerable resources evaluating whether or not each proposed coalescent history is valid. Consequently, the speed of evaluation of the Degnan-Salter formula would be dramatically reduced if the set of valid coalescent histories could simply be enumerated. This article shows how the enumeration of valid coalescent histories compatible with given species tree and gene tree labeled topologies can be performed, and a recursive formula is developed for counting the number of these histories. Section 2 gives some definitions used in obtaining the results. Section 3 considers the case that the species tree and gene tree have the same labeled topology, and Section 4 considers the more general case where these labeled topologies are not necessarily identical.

2. DEFINITIONS

For n taxa, consider a bifurcating rooted species tree labeled topology S , whose taxa are listed in the label set X_S . Consider also a bifurcating rooted gene tree labeled topology G , with label set $X_G \subset X_S$. We assume that exactly one gene lineage is studied from each species.

Trees are viewed as evolving in time beginning from a single edge—the edge “above the root.” For convenience, “above” and “below” a node will refer to parts of the tree that are more ancient and more recent than the node, respectively. The terms “edges” and “branches” are used interchangeably. The “taxa” of a tree may refer either to the leaves of the tree or to the labels associated with these leaves.

We refer to a tree that contains an internal node that descends from all other internal nodes as a caterpillar tree (Figure 1A). A tree with at least five taxa that contains a four-taxon symmetric subtree whose root is descended from all internal nodes not in the subtree is a pseudocaterpillar tree (Fig. 1B). A tree each of whose two subtrees immediately descended from the root is a caterpillar tree is a bicaterpillar tree (Fig. 1C).

The internal nodes of species trees are enumerated according to a postorder traversal (Sedgewick and Flajolet, 1996), and the internal edge immediately above a given node is given the same number as the node. We allow the species tree to be “extended,” so that the edge immediately above the root can be artificially subdivided into multiple edges (Fig. 2).

Looking backwards in time, lineages from the gene tree coalesce in a pairwise manner along the internal edges of the species tree. A valid coalescent history for a gene tree and a species tree is a list of coalescences in the gene tree together with the edges of the species tree on which they occur (Degnan and Salter, 2005). Thus, a valid coalescent history is a property of a gene tree/species tree pair. A valid m -extended coalescent history is a valid coalescent history for a gene tree and a species tree, when the edge above the root of the species tree is subdivided into m components ($m \geq 1$). For convenience, the word “valid” will often be omitted, and it will be implied that coalescent histories are valid. An example of a valid 4-extended coalescent history is shown in Figure 2.

3. LABELED TOPOLOGY OF THE SPECIES TREE AND GENE TREE IDENTICAL

In this section, it is assumed that the gene tree under consideration has the same labeled topology as the species tree. We denote by $A_{S,m}$ the number of valid m -extended coalescent histories for which the gene tree and species tree both have labeled topology S . The number of valid coalescent histories is obtained from $A_{S,1}$. In this section, we refer to coalescent histories as being a property of a labeled topology or simply of a “tree” rather than of a gene tree/species tree pair, as the gene tree and species tree have the same labeled topology.



FIG. 1. Caterpillar, bicaterpillar, and pseudocaterpillar trees with seven taxa. (A) A caterpillar tree. (B) A pseudocaterpillar tree. (C) A bicaterpillar tree whose two caterpillar subtrees have four and three taxa.

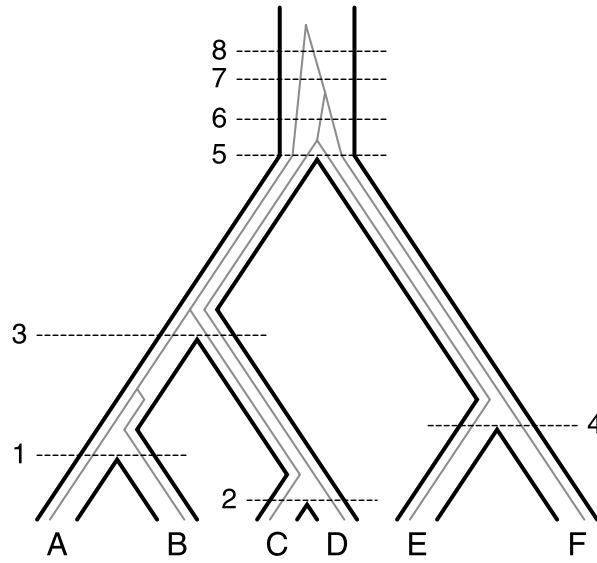


FIG. 2. A 4-extended coalescent history for a species tree and gene tree with six taxa. The species tree (thick lines) has labeled topology $((((A,B),(C,D)),(E,F)))$, and the gene tree (thin lines) has labeled topology $((((A,B),C),((D,E),F)))$. The internal nodes of the species tree are numbered according to a postorder traversal. The edge above the root of the species tree is divided into four branches, numbered 5, 6, 7, and 8. The 4-extended coalescent history for the gene tree—(A,B), ((A,B),C), (D,E), ((D,E),F), and (((A,B),C),((D,E),F)). If given as a vector, the 4-extended coalescent history is (1,3,5,6,8).

For matching species tree and gene tree labeled topologies, we first develop a recursion that counts the number of valid m -extended coalescent histories, and we describe how these coalescent histories can be enumerated. Using the recursion, closed-form expressions are then obtained for the number of coalescent histories possessed by caterpillar, pseudocaterpillar, and bicaterpillar trees. The number of coalescent histories for a caterpillar tree with n taxa is the Catalan number C_{n-1} ; this result had previously been obtained by Degnan (2005), who constructed an equivalence with the problem of counting the number of sequences of $2n - 2$ total $+1$ and -1 values for which all partial sums are nonnegative and for which the sum of all $2n - 2$ values is zero. We show that the number of coalescent histories for pseudocaterpillar trees with $n \geq 5$ taxa is $(5n - 12)C_{n-1}/(4n - 6)$, and that for bicaterpillar trees whose two caterpillar subtrees have l and $n - l$ taxa, the number of coalescent histories is $C_l C_{n-l}$.

Using the recursion and the closed-form values for caterpillar, pseudocaterpillar, and bicaterpillar trees, we obtain a variety of lower and upper bounds on numbers of coalescent histories. We first give a lower bound on the speed at which the number of m -extended coalescent histories increases with the addition of taxa to a tree; this bound had previously been obtained by Degnan and Salter (2005) for the case of $m = 1$. We then provide crude closed-form bounds on the number of m -extended coalescent histories for a given tree, as well as for any tree with a fixed number of taxa n . Finally, we show that $(5n - 12)n\sqrt{n\pi}/[32(4n - 6)]$ provides a lower bound on the ratio of the numbers of coalescent histories for the n -taxon trees with the largest and smallest numbers of these histories.

Theorem 3.1. For $m \geq 1$, the number of valid m -extended coalescent histories for a tree S with $n \geq 2$ taxa is

$$A_{S,m} = \sum_{k=1}^m A_{S_L,k+1} A_{S_R,k+1}, \tag{1}$$

where S_L and S_R respectively denote the left and right subtrees of S . If S has $n = 1$ taxon, then $A_{S,m} = 1$ for all m .

Proof. If $n = 1$, no coalescences of gene lineages occur, and trivially there is only one m -extended coalescent history, regardless of the value of m . For $n \geq 2$, the final coalescence joins the ancestor of the left subtree of the gene tree and the ancestor of the right subtree. This coalescence can occur on any of the m branches above the root of S .

Suppose the final gene tree coalescence happens on the k th species tree branch above the root, with $1 \leq k \leq m$. The left and right subtrees of S , S_L and S_R , then each have $k + 1$ branches above their respective roots on which coalescences can take place. Because the coalescences of the gene lineages from S_L and S_R can occur in any order with respect to each other, the number of m -extended coalescent histories of S with final coalescence on branch k above the root of S is the product of the number of $(k + 1)$ -extended coalescent histories for S_L and the corresponding number for S_R . Summing over possible values of k , the result follows. ■

Corollary 3.2. For $m \geq 2$, the number of valid m -extended coalescent histories for a tree S with $n \geq 1$ taxa satisfies $A_{S,m} \geq A_{S,m-1}$.

Proof. This is trivial for $n = 1$, and follows directly from Equation (1) for $n \geq 2$. ■

Remarks. Although we focus on bifurcating species trees and gene trees, Theorem 3.1 is straightforward to extend to multifurcating trees: the summand in Equation (1) simply becomes a product of the number of $(k + 1)$ -extended coalescent histories over all subtrees immediately descended from the root of S .

For a given tree S , repeated iteration of Equation (1) leads to a closed-form function of m for the number of valid m -extended coalescent histories associated with S . The nature of the recursion guarantees that $A_{S,m}$ is a polynomial in m of degree $n - 1$, where n is the number of taxa for the tree S . Tables 1–4 show these polynomials for all trees with $n \leq 9$ taxa. For each m , all species tree labeled topologies with the same unlabeled topology have the same number of m -extended coalescent histories; consequently, the tables index the polynomials by unlabeled topology.

Theorem 3.1 makes it possible to enumerate all m -extended coalescent histories for a tree S . Suppose the left and right subtrees of S , S_L and S_R , have l and r taxa, respectively, with $l + r = n$. In the postorder traversal used to number branches of trees, the branches of the left subtree of S are labeled from 1 to $l - 1$, with $l - 1$ being the label for the branch immediately above the root of S_L . Similarly, the branches of the right subtree of S are labeled from l to $n - 2$, with $n - 2$ being the label for the branch immediately above the root of S_R . We extend this scheme by sequentially labeling the m branches above the root of S by $n - 1$ to $n + m - 2$, with $n - 1$ being the branch closest to the root (Fig. 3A).

Considering the $(k + 1)$ -extended left subtree of S as a separate tree, in a postorder traversal, the branches above its root are labeled $l - 1$ to $l - 1 + k$ (Fig. 3B). Similarly, the branches above the root for the $(k + 1)$ -extended right subtree of S are labeled $r - 1$ to $r - 1 + k$ (Fig. 3C). As a result, the following recursive procedure—considering each value of k with $1 \leq k \leq m$, each $(k + 1)$ -extended coalescent history for the left subtree S_L , and each $(k + 1)$ -extended coalescent history for the right subtree S_R —enumerates all possible m -extended coalescent histories for a tree S .

1. Consider a $(k + 1)$ -extended coalescent history of S_L . For each coalescence that occurs on a branch numbered l or higher, add $r - 1$ to the number denoting the branch on which the coalescence occurs.
2. Consider a $(k + 1)$ -extended coalescent history of S_R . For each coalescence, add $l - 1$ to the number denoting the branch on which the coalescence occurs.
3. Concatenate the modified $(k + 1)$ -extended coalescent history of S_L , the modified $(k + 1)$ -extended coalescent history of S_R , and the number $n + k - 2$ denoting the branch for the final coalescence joining the ancestor of the gene lineages of S_L and the ancestor of the gene lineages of S_R .

The number of coalescent histories for caterpillar, pseudocaterpillar, and bicaterpillar trees

It is of interest to identify the properties of trees that give rise to larger and smaller numbers of valid coalescent histories. Tables 1–4 illustrate that for a given number of taxa, trees in which the two subtrees immediately descended from the root have similar numbers of taxa usually have fewer coalescent histories, and those in which one of these subtrees has only one taxon usually have more coalescent histories.

TABLE 1. NUMBER OF COALESCENT HISTORIES FOR TREES WITH $n \leq 7$ TAXA

| <i>Number of taxa</i> | <i>Unlabeled topology</i> | <i>Number of coalescent histories</i> | <i>Number of m-extended coalescent histories</i> |
|-----------------------|---------------------------|---------------------------------------|---|
| 1 | | 1 | 1 |
| 2 | | 1 | m |
| 3 | | 2 | $\frac{1}{2}m(m+3)$ |
| 4 | | 5 | $\frac{1}{6}m(m+4)(m+5)$ |
| | | 4 | $\frac{1}{6}m(2m^2+9m+13)$ |
| 5 | | 14 | $\frac{1}{24}m(m+5)(m+6)(m+7)$ |
| | | 13 | $\frac{1}{12}m(m+5)(m^2+7m+18)$ |
| | | 10 | $\frac{1}{8}m(m^3+10m^2+31m+38)$ |
| 6 | | 42 | $\frac{1}{120}m(m+6)(m+7)(m+8)(m+9)$ |
| | | 42 | $\frac{1}{120}m(m+6)(m+7)(2m^2+19m+69)$ |
| | | 37 | $\frac{1}{120}m(3m^4+60m^3+445m^2+1560m+2372)$ |
| | | 28 | $\frac{1}{120}m(4m^4+75m^3+490m^2+1305m+1486)$ |
| | | 26 | $\frac{1}{120}m(8m^4+105m^3+530m^2+1215m+1262)$ |
| | | 25 | $\frac{1}{60}m(3m^4+45m^3+245m^2+585m+622)$ |
| 7 | | 132 | $\frac{1}{720}m(m+7)(m+8)(m+9)(m+10)(m+11)$ |
| | | 138 | $\frac{1}{360}m(m+7)(m+8)(m+9)(m^2+12m+56)$ |
| | | 130 | $\frac{1}{240}m(m+7)(m^4+26m^3+253m^2+1184m+2436)$ |
| | | 112 | $\frac{1}{360}m(m+6)(2m^4+51m^3+464m^2+1821m+3422)$ |
| | | 113 | $\frac{1}{360}m(4m^5+99m^4+1000m^3+5205m^2+14,656m+19,716)$ |
| | | 106 | $\frac{1}{120}m(m+7)(m^4+20m^3+150m^2+525m+894)$ |
| | | 84 | $\frac{1}{720}m(5m^5+153m^4+1775m^3+9555m^2+23,420m+25,572)$ |
| | | 84 | $\frac{1}{360}m(5m^5+117m^4+1100m^3+5145m^2+11,675m+12,198)$ |
| | | 74 | $\frac{1}{48}m(m^5+21m^4+169m^3+675m^2+1366m+1320)$ |
| | | 70 | $\frac{1}{360}m(5m^5+117m^4+1055m^3+4575m^2+9740m+9708)$ |
| | | 65 | $\frac{1}{36}m(m^5+18m^4+130m^3+474m^2+895m+822)$ |

TABLE 2. NUMBER OF COALESCENT HISTORIES FOR TREES WITH $n = 8$ TAXA
























| Unlabeled topology | Number of coalescent histories | Number of m -extended coalescent histories |
|---|--------------------------------|--|
|  | 429 | $\frac{1}{5040}m(m+8)(m+9)(m+10)(m+11)(m+12)(m+13)$ |
|  | 462 | $\frac{1}{5040}m(m+8)(m+9)(m+10)(m+11)(2m^2+29m+165)$ |
|  | 453 | $\frac{1}{1680}m(m+8)(m+9)(m^4+32m^3+385m^2+2246m+5792)$ |
|  | 416 | $\frac{1}{5040}m(4m^6+189m^5+3661m^4+37,485m^3+219,751m^2+728,406m+1,107,144)$ |
|  | 442 | $\frac{1}{5040}m(8m^6+315m^5+5243m^4+47,565m^3+253,337m^2+781,200m+1,140,012)$ |
|  | 408 | $\frac{1}{840}m(m+8)(m+9)(m^4+25m^3+238m^2+1084m+2460)$ |
|  | 354 | $\frac{1}{5040}m(5m^6+231m^5+4319m^4+41,685m^3+220,220m^2+638,484m+879,216)$ |
|  | 372 | $\frac{1}{2520}m(5m^6+189m^5+2996m^4+25,515m^3+123,515m^2+337,176m+448,044)$ |
|  | 353 | $\frac{1}{1680}m(5m^6+175m^5+2513m^4+19,285m^3+85,050m^2+214,900m+271,112)$ |
|  | 322 | $\frac{1}{2520}m(5m^6+189m^5+2933m^4+24,045m^3+111,650m^2+293,286m+379,332)$ |
|  | 326 | $\frac{1}{504}m(2m^6+63m^5+833m^4+5964m^3+24,815m^2+59,745m+72,882)$ |
|  | 264 | $\frac{1}{1680}m(2m^6+91m^5+1659m^4+15,295m^3+73,563m^2+171,094m+181,816)$ |
|  | 276 | $\frac{1}{1680}m(4m^6+147m^5+2247m^4+18,165m^3+80,241m^2+178,248m+184,628)$ |
|  | 260 | $\frac{1}{1680}m(6m^6+203m^5+2779m^4+19,985m^3+80,059m^2+167,132m+166,636)$ |
|  | 224 | $\frac{1}{1680}m(8m^6+259m^5+3311m^4+21,385m^3+75,677m^2+142,996m+132,684)$ |
|  | 226 | $\frac{1}{1680}m(16m^6+413m^5+4417m^4+25,025m^3+80,479m^2+142,562m+126,768)$ |
|  | 212 | $\frac{1}{840}m(6m^6+168m^5+1904m^4+11,235m^3+37,184m^2+67,137m+60,446)$ |
|  | 210 | $\frac{1}{1680}m(5m^6+175m^5+2457m^4+17,605m^3+67,690m^2+135,100m+129,768)$ |
|  | 210 | $\frac{1}{840}m(5m^6+140m^5+1638m^4+10,220m^3+35,665m^2+66,920m+61,812)$ |
|  | 185 | $\frac{1}{1680}m(15m^6+385m^5+4011m^4+22,015m^3+68,110m^2+116,200m+100,064)$ |
|  | 196 | $\frac{1}{2520}m(10m^6+315m^5+4039m^4+26,880m^3+97,825m^2+188,265m+176,586)$ |
|  | 182 | $\frac{1}{2520}m(20m^6+525m^5+5621m^4+31,605m^3+99,575m^2+171,990m+149,304)$ |
|  | 169 | $\frac{1}{1260}m(20m^6+420m^5+3731m^4+17,955m^3+49,805m^2+78,015m+62,994)$ |

TABLE 3. NUMBER OF COALESCENT HISTORIES FOR TREES WITH $n = 9$ TAXA, FOR WHICH THE NUMBERS OF TAXA IN THE TWO SUBTREES IMMEDIATELY DESCENDED FROM THE ROOT ARE 8 AND 1

| Unlabeled topology | Number of coalescent histories | Number of m -extended coalescent histories |
|--------------------|--------------------------------|---|
| | 1430 | $\frac{1}{40,320}m(m+9)(m+10)(m+11)(m+12)(m+13)(m+14)(m+15)$ |
| | 1573 | $\frac{1}{20,160}m(m+9)(m+10)(m+11)(m+12)(m+13)(m^2+17m+114)$ |
| | 1584 | $\frac{1}{40,320}m(m+9)(m+10)(m+11)(3m^4+114m^3+1633m^2+11,394m+35,240)$ |
| | 1511 | $\frac{1}{10,080}m(m+9)(m^6+57m^5+1335m^4+16,587m^3+118,656m^2+482,820m+903,632)$ |
| | 1663 | $\frac{1}{50,400}m(m+9)(m^6+48m^5+975m^4+10,818m^3+70,722m^2+268,980m+486,608)$ |
| | 1518 | $\frac{1}{20,160}m(m+9)(m+10)(m+11)(3m^4+90m^3+1037m^2+5790m+16,264)$ |
| | 1368 | $\frac{1}{40,320}m(5m^7+324m^6+8834m^5+131,376m^4+1,156,925m^3+6,173,916m^2+19,276,476m+28,409,904)$ |
| | 1488 | $\frac{1}{20,160}m(5m^7+276m^6+6566m^5+87,444m^4+707,105m^3+3,536,904m^2+10,535,604m+15,124,176)$ |
| | 1478 | $\frac{1}{13,440}m(5m^7+260m^6+5754m^5+70,952m^4+532,805m^3+2,492,420m^2+7,021,436m+9,740,688)$ |
| | 1316 | $\frac{1}{20,160}m(m+9)(5m^6+231m^5+4403m^4+44,709m^3+260,624m^2+875,868m+1,467,216)$ |
| | 1408 | $\frac{1}{10,080}m(m+9)(5m^6+195m^5+3215m^4+28,941m^3+153,056m^2+475,116m+758,736)$ |
| | 1155 | $\frac{1}{20,160}m(3m^7+192m^6+5138m^5+74,172m^4+621,047m^3+3,025,428m^2+8,355,092m+11,203,728)$ |
| | 1248 | $\frac{1}{10,080}m(3m^7+162m^6+3752m^5+48,258m^4+370,937m^3+1,704,108m^2+4,524,188m+5,928,432)$ |
| | 1229 | $\frac{1}{20,160}m(9m^7+456m^6+9758m^5+115,332m^4+819,581m^3+3,533,124m^2+8,938,172m+11,360,208)$ |
| | 1136 | $\frac{1}{1680}m(m^7+49m^6+1001m^5+11,095m^4+73,094m^3+292,376m^2+691,984m+838,880)$ |
| | 1213 | $\frac{1}{1680}m(2m^7+83m^6+1477m^5+14,651m^4+88,333m^3+329,602m^2+739,268m+864,424)$ |
| | 1118 | $\frac{1}{10,080}m(9m^7+396m^6+7378m^5+75,852m^4+470,281m^3+1,793,064m^2+4,088,572m+4,833,888)$ |
| | 1020 | $\frac{1}{40,320}m(15m^7+780m^6+17,038m^5+202,776m^4+1,426,495m^3+6,025,740m^2+14,886,052m+18,567,504)$ |

(continued)

TABLE 3. (Continued)













| Unlabeled topology | Number of coalescent histories | Number of m -extended coalescent histories |
|---|--------------------------------|---|
|  | 1074 | $\frac{1}{20,160}m(m+9)(15m^6 + 525m^5 + 7777m^4 + 62,391m^3 + 287,896m^2 + 755,076m + 1,051,504)$ |
|  | 1022 | $\frac{1}{13,440}m(15m^7 + 620m^6 + 10,878m^5 + 105,896m^4 + 624,575m^3 + 2,276,540m^2 + 4,989,172m + 5,727,984)$ |
|  | 980 | $\frac{1}{10,080}m(5m^7 + 240m^6 + 4886m^5 + 54,768m^4 + 366,905m^3 + 1,491,840m^2 + 3,579,564m + 4,380,192)$ |
|  | 994 | $\frac{1}{5040}m(5m^7 + 210m^6 + 3752m^5 + 37,170m^4 + 222,635m^3 + 821,940m^2 + 1,819,848m + 2,104,200)$ |
|  | 1010 | $\frac{1}{2520}m(5m^7 + 180m^6 + 2807m^5 + 24,675m^4 + 133,175m^3 + 449,715m^2 + 923,673m + 1,010,970)$ |

TABLE 4. NUMBER OF COALESCENT HISTORIES FOR TREES WITH $n = 9$ TAXA, FOR WHICH THE NUMBERS OF TAXA IN THE TWO SUBTREES IMMEDIATELY DESCENDED FROM THE ROOT ARE 7 AND 2, 6 AND 3, OR 5 AND 4

| Unlabeled topology | Number of coalescent histories | Number of m -extended coalescent histories |
|---|--------------------------------|--|
|  | 858 | $\frac{1}{40,320}m(7m^7 + 444m^6 + 11,718m^5 + 165,312m^4 + 1,327,263m^3 + 5,927,796m^2 + 13,297,172m + 13,864,848)$ |
|  | 924 | $\frac{1}{20,160}m(7m^7 + 372m^6 + 8442m^5 + 105,588m^4 + 776,643m^3 + 3,279,528m^2 + 7,137,788m + 7,319,472)$ |
|  | 906 | $\frac{1}{40,320}m(21m^7 + 1044m^6 + 21,770m^5 + 248,472m^4 + 1,679,909m^3 + 6,655,236m^2 + 13,942,460m + 13,981,008)$ |
|  | 832 | $\frac{1}{1440}m(m^7 + 48m^6 + 952m^5 + 10,110m^4 + 62,719m^3 + 229,362m^2 + 454,728m + 440,160)$ |
|  | 884 | $\frac{1}{5040}m(7m^7 + 282m^6 + 4837m^5 + 45,759m^4 + 259,168m^3 + 885,003m^2 + 1,678,908m + 1,581,396)$ |
|  | 816 | $\frac{1}{20,160}m(21m^7 + 900m^6 + 16,198m^5 + 159,012m^4 + 926,569m^3 + 3,231,480m^2 + 6,214,012m + 5,902,368)$ |
|  | 708 | $\frac{1}{40,320}m(35m^7 + 1644m^6 + 31,542m^5 + 317,688m^4 + 1,807,155m^3 + 5,937,036m^2 + 10,741,108m + 9,710,352)$ |

(continued)

TABLE 4. (Continued)

| Unlabeled topology | Number of coalescent histories | Number of m -extended coalescent histories |
|--------------------|--------------------------------|---|
| | 744 | $\frac{1}{20,160}m(35m^7 + 1356m^6 + 22,218m^5 + 198,492m^4 + 1,037,295m^3 + 3,212,664m^2 + 5,593,252m + 4,933,728)$ |
| | 706 | $\frac{1}{384}m(m^7 + 36m^6 + 538m^5 + 4368m^4 + 20,953m^3 + 60,444m^2 + 99,852m + 84,912)$ |
| | 644 | $\frac{1}{20,160}m(35m^7 + 1356m^6 + 21,798m^5 + 188,328m^4 + 948,675m^3 + 2,840,124m^2 + 4,815,412m + 4,167,312)$ |
| | 652 | $\frac{1}{10,080}m(35m^7 + 1140m^6 + 15,750m^5 + 119,784m^4 + 544,215m^3 + 1,501,500m^2 + 2,398,480m + 1,991,256)$ |
| | 660 | $\frac{1}{40,320}m(21m^7 + 1044m^6 + 21,490m^5 + 235,872m^4 + 1,479,709m^3 + 5,280,156m^2 + 10,110,940m + 9,481,968)$ |
| | 690 | $\frac{1}{20,160}m(21m^7 + 864m^6 + 15,190m^5 + 147,420m^4 + 845,929m^3 + 2,843,316m^2 + 5,247,340m + 4,810,320)$ |
| | 650 | $\frac{1}{40,320}m(63m^7 + 2412m^6 + 38,710m^5 + 340,704m^4 + 1,785,847m^3 + 5,580,708m^2 + 9,787,540m + 8,672,016)$ |
| | 560 | $\frac{1}{3360}m(7m^7 + 258m^6 + 3920m^5 + 31,878m^4 + 151,613m^3 + 428,792m^2 + 691,180m + 573,952)$ |
| | 565 | $\frac{1}{1680}m(7m^7 + 213m^6 + 2765m^5 + 19,845m^4 + 85,568m^3 + 224,462m^2 + 342,580m + 273,760)$ |
| | 530 | $\frac{1}{20,160}m(63m^7 + 2052m^6 + 28,070m^5 + 209,664m^4 + 931,847m^3 + 2,500,428m^2 + 3,878,420m + 3,134,256)$ |
| | 588 | $\frac{1}{40,320}m(35m^7 + 1500m^6 + 26,950m^5 + 262,248m^4 + 1,487,395m^3 + 4,913,580m^2 + 8,928,500m + 8,087,952)$ |
| | 546 | $\frac{1}{20,160}m(35m^7 + 1320m^6 + 20,650m^5 + 173,796m^4 + 853,615m^3 + 2,476,740m^2 + 4,064,900m + 3,416,304)$ |
| | 588 | $\frac{1}{2880}m(5m^7 + 180m^6 + 2770m^5 + 23,616m^4 + 120,325m^3 + 366,540m^2 + 630,020m + 549,984)$ |
| | 546 | $\frac{1}{10,080}m(35m^7 + 1080m^6 + 14,350m^5 + 106,344m^4 + 474,355m^3 + 1,281,840m^2 + 2,001,020m + 1,624,656)$ |
| | 518 | $\frac{1}{40,320}m(105m^7 + 3540m^6 + 50,050m^5 + 385,896m^4 + 1,763,545m^3 + 4,837,140m^2 + 7,621,180m + 6,224,304)$ |
| | 481 | $\frac{1}{4032}m(21m^7 + 600m^6 + 7238m^5 + 48,300m^4 + 194,201m^3 + 477,036m^2 + 687,596m + 524,400)$ |

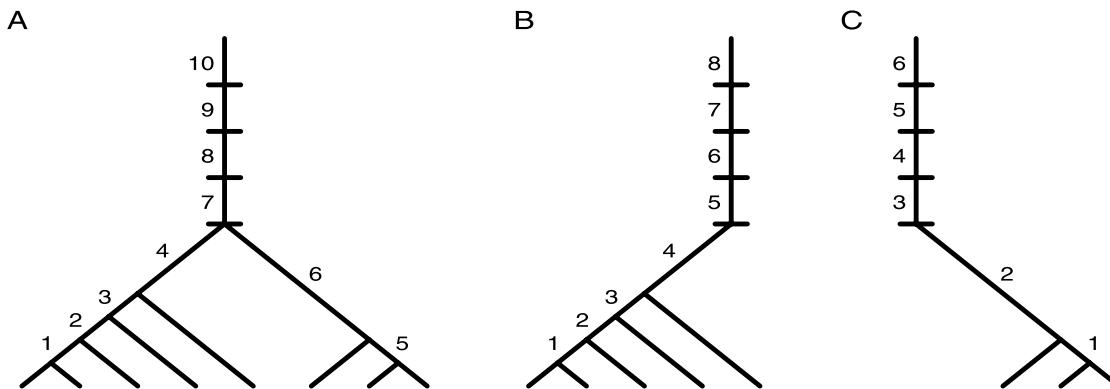


FIG. 3. Recursive procedure for enumerating coalescent histories. The tree (A) represents a 4-extended species tree, that is, a species tree with four branches above the root. The trees (B, C) represent the 5-extended left and right subtrees of the species tree in (A), respectively. To enumerate the 4-extended coalescent histories for the tree in (A), for each k from 1 to 4, we consider all possible concatenations of a $(k + 1)$ -extended coalescent history for the tree in (B), a $(k + 1)$ -extended coalescent history for the tree in (C), and the branch $k + 6$ in (A) denoting the location of the final coalescence. Note that as described in the text, because the labels for corresponding branches in (B) or (C) may differ from those in (A), the k -extended coalescent histories for (B) and (C) must be suitably modified before concatenation.

Although caterpillar trees tend to have a large number of coalescent histories, the greatest number of coalescent histories tends to occur not for caterpillar trees, but for caterpillar-like trees in which a caterpillar subtree is replaced by a subtree with a relatively high degree of symmetry. To see why this is the case, notice that a high-symmetry subtree contains several sets of gene lineages that can coalesce with each other in any of several possible sequences. The caterpillar-like part of the species tree contains many branches on which these coalescences can occur. The combination of many possible sequences of coalescences with many available branches on which the coalescences can take place contributes to a large number of coalescent histories. Thus, a hybrid of a caterpillar tree and a highly symmetric subtree will tend to have more coalescent histories than a pure caterpillar tree, which has many possible branches on which coalescences can happen, but only one possible sequence of coalescences. It will also tend to have more coalescent histories than a highly symmetric tree, which has many possible sequences of coalescences, but few branches on which the coalescences can take place.

We now evaluate the number of coalescent histories for caterpillar, pseudocaterpillar, and bicaterpillar trees. We then verify that for a given number of taxa $n \geq 7$, the pseudocaterpillar tree has more coalescent histories than the caterpillar tree, and the caterpillar tree in turn has more coalescent histories than the bicaterpillar trees in which both caterpillar components have at least two taxa.

Lemma 3.3. *Suppose $m \geq 1$ and $n \geq 2$. Then*

$$\sum_{k=2}^{m+1} \frac{k}{n-1} \binom{k+2n-3}{n-2} = \frac{m}{n} \binom{m+2n-1}{n-1}. \tag{2}$$

Proof. This lemma is proven by induction on m . For $m = 1$ both sides equal $(2n)!/[n!(n+1)!]$. Assuming the result holds for a given m ,

$$\begin{aligned} \sum_{k=2}^{m+2} \frac{k}{n-1} \binom{k+2n-3}{n-2} &= \frac{m}{n} \binom{m+2n-1}{n-1} + \frac{m+2}{n-1} \binom{m+2n-1}{n-2} \\ &= \frac{m+1}{n} \binom{m+2n}{n-1}. \end{aligned} \quad \blacksquare$$

Theorem 3.4. *Suppose $m \geq 1$ and $n \geq 2$. The number of valid m -extended coalescent histories for a caterpillar tree Z_n with n taxa is*

$$A_{Z_n, m} = \frac{m}{n-1} \binom{m+2n-3}{n-2}. \quad (3)$$

Proof. The result is proven by induction on n . For $n = 2$ and any m , $A_{Z_2, m} = m$ follows from Theorem 3.1, using $S = Z_2$ and $S_L = S_R = Z_1$. Assuming the result holds for a given n for all m , Theorem 3.1 yields

$$\begin{aligned} A_{Z_{n+1}, m} &= \sum_{k=1}^m A_{Z_n, k+1} A_{Z_1, k+1} \\ &= \sum_{k=2}^{m+1} \frac{k}{n-1} \binom{k+2n-3}{n-2} \\ &= \frac{m}{n} \binom{m+2n-1}{n-1}, \end{aligned}$$

where the last equality follows from Lemma 3.3. Thus, the result with a given n for all m implies the result with $n+1$ for all m . ■

Corollary 3.5 (Degnan, 2005). *Suppose $n \geq 2$. The number of valid coalescent histories for a caterpillar tree Z_n with n taxa is the Catalan number*

$$C_{n-1} = \frac{1}{n} \binom{2n-2}{n-1}. \quad (4)$$

Proof. The result follows directly from Theorem 3.4, taking $m = 1$. ■

Lemma 3.6. *Suppose $m, n_1, n_2 \geq 0$. Then*

$$\sum_{k=0}^m \binom{k+n_1}{n_2} = \binom{n_1+m+1}{n_2+1} - \binom{n_1}{n_2+1}. \quad (5)$$

Proof. This lemma is proven by induction on m . For $m = 0$ both sides equal $n_1!/[n_2!(n_1-n_2)!]$. Assuming the result holds for a given m ,

$$\begin{aligned} \sum_{k=0}^{m+1} \binom{k+n_1}{n_2} &= \binom{n_1+m+1}{n_2+1} - \binom{n_1}{n_2+1} + \binom{n_1+m+1}{n_2} \\ &= \binom{n_1+m+2}{n_2+1} - \binom{n_1}{n_2+1}. \end{aligned} \quad \blacksquare$$

Theorem 3.7. *Suppose $m \geq 1$ and $n \geq 5$. The number of valid m -extended coalescent histories for a pseudocaterpillar tree Y_n with n taxa is*

$$A_{Y_n, m} = \frac{m[2m^2 + (5n-11)m + (5n^2 - 22n + 21)]}{(n-1)(m+2n-3)(m+2n-4)} \binom{m+2n-3}{n-2}. \quad (6)$$

Proof. The result is proven by induction on n . For $n = 5$, $A_{Y_5,m}$ can be obtained by repeated application of Theorem 3.1, and equals (Table 1)

$$\frac{m(2m^2 + 14m + 36)}{4(m + 7)(m + 6)} \binom{m + 7}{3}.$$

Assuming the result holds for a given n and all m , Theorem 3.1 yields

$$\begin{aligned} A_{Y_{n+1},m} &= \sum_{k=1}^m A_{Y_n,k+1} A_{Z_1,k+1} \\ &= \sum_{k=2}^{m+1} \frac{k[2k^2 + (5n - 11)k + (5n^2 - 22n + 21)]}{(n - 1)(k + 2n - 3)(k + 2n - 4)} \binom{k + 2n - 3}{n - 2}. \end{aligned}$$

The summand is equivalent to

$$2 \binom{k + 2n - 2}{n - 1} - 7 \binom{k + 2n - 3}{n - 2} + 9 \binom{k + 2n - 4}{n - 3} - 6 \binom{k + 2n - 5}{n - 4}.$$

Lemma 3.6 can therefore be applied four times to sum each of the four terms from $k = 0$ to $k = m + 1$. Subtracting the values for $k = 0$ and $k = 1$ from the resulting sum and simplifying,

$$A_{Y_{n+1},m} = \frac{m[2m^2 + (5n - 6)m + (5n^2 - 12n + 4)]}{n(m + 2n - 1)(m + 2n - 2)} \binom{m + 2n - 1}{n - 1}.$$

Thus, the result with a given n for all m implies the result with $n + 1$ for all m . ■

Corollary 3.8. *Suppose $m \geq 1$ and $n \geq 5$, and (m, n) is not $(1, 5)$, $(1, 6)$ or $(2, 5)$. Then $A_{Y_n,m} > A_{Z_n,m}$, where $A_{Y_n,m}$ and $A_{Z_n,m}$ respectively denote the numbers of m -extended coalescent histories for pseudocaterpillar and caterpillar trees with n taxa.*

Proof. The result follows directly from the values of $A_{Y_n,m}$ and $A_{Z_n,m}$. ■

Corollary 3.9. *Suppose $n \geq 2$. The number of valid coalescent histories for a pseudocaterpillar tree Y_n with $n \geq 5$ taxa is $(5n - 12)C_{n-1}/(4n - 6)$.*

Proof. The result follows directly from Theorem 3.7, taking $m = 1$. ■

Theorem 3.10. *Suppose $n \geq 2$ and $1 \leq l \leq n - 1$. The number of valid coalescent histories for a bicaterpillar tree whose two caterpillar subtrees contain l and $n - l$ taxa is the product of Catalan numbers $C_l C_{n-l}$.*

Proof. Denote the bicaterpillar tree with caterpillar subtrees containing l and $n - l$ taxa by $W_{l,n-l}$. Applying Theorem 3.1 in two consecutive steps,

$$\begin{aligned} A_{W_{l,n-l},1} &= A_{Z_l,2} A_{Z_{n-l},2} \\ &= A_{Z_{l+1},1} A_{Z_{n-l+1},1} \\ &= C_l C_{n-l}, \end{aligned}$$

where the last step follows from Corollary 3.5. ■

Corollary 3.11. *For bicaterpillar trees with $n = l + (n - l)$ taxa, where $1 \leq l \leq \lfloor n/2 \rfloor$ and $n \geq 2$, the number of valid coalescent histories decreases with increasing l , so that (i) the n -taxon bicaterpillar tree*

with the greatest number of valid coalescent histories is the caterpillar tree, that is, the bicaterpillar tree whose two caterpillar subtrees have 1 and $n - 1$ taxa; (ii) the n -taxon bicaterpillar tree with the smallest number of valid coalescent histories is the tree whose two caterpillar subtrees have $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ taxa.

Proof. For $l \geq 2$ we can simplify to obtain

$$\begin{aligned} \frac{A_{W_{l,n-l,1}}}{A_{W_{l-1,n-l+1,1}}} &= \frac{C_l C_{n-l}}{C_{l-1} C_{n-l+1}} \\ &= \frac{f(l)}{f(n-l+1)}, \end{aligned}$$

where $f(x) = 2 - 3/(x+1)$. Because $f(x)$ is a positive and increasing function for $x > 1/2$, $f(l)/f(n-l+1)$ is increasing for $l \in [2, \lfloor n/2 \rfloor]$. At $l = \lfloor n/2 \rfloor$, if n is even, equaling $2k$, $f(l)/f(n-l+1) = (2k^2 + 3k - 2)/(2k^2 + 3k + 1)$; if n is odd, equaling $2k + 1$, $f(l)/f(n-l+1) = (2k^2 + 5k - 3)/(2k^2 + 5k + 3)$. In both cases, at $l = \lfloor n/2 \rfloor$, $f(l)/f(n-l+1) < 1$, from which it follows that for $l \in [2, \lfloor n/2 \rfloor]$, $f(l)/f(n-l+1) < 1$. As a result, $A_{W_{l,n-l,1}}$ decreases with increasing l , its smallest value occurring at $l = \lfloor n/2 \rfloor$ and its largest value at $l = 1$. ■

Bounds on the number of coalescent histories

The results in the previous sections enable computation of bounds on the number of coalescent histories. We provide loose lower and upper bounds on the number of m -extended coalescent histories for a given tree, as well as for any tree with a fixed number of taxa n . For a given number of taxa, we also give a lower bound on the ratio of the numbers of coalescent histories for the trees with the largest and smallest numbers of these histories.

Theorem 3.12. *Suppose $n_2 > n_1 \geq 1$ and $m \geq 1$, and that cases with $(n_1, m) = (1, 1)$ are excluded. If S_2 is a tree with n_2 taxa, S_1 is a tree with n_1 taxa, and S_2 displays S_1 , then*

$$A_{S_2, m} \geq 2^{n_2 - n_1} A_{S_1, m}. \quad (7)$$

Proof. If $(n_2, n_1) = (2, 1)$, the result follows from the fact that $A_{S, m} = 1$ when S has one taxon and $A_{S, m} = m$ when S has two taxa (Table 1). We now use strong induction. Suppose that for all $m \geq 2$, if $n_1 = n_2 - 1 \geq 1$, Equation (7) holds for all trees S with $n_2 \leq N$ taxa. We will show that this implies it holds for all $m \geq 1$ for trees S^* with $n_2 = N + 1$ taxa, when $n_1 = n_2 - 1$.

Consider a tree S^* that has $N + 1$ taxa and that displays S . Two cases are possible: (i) one of the two subtrees of S^* , say the right subtree S_R^* , is the same as the right subtree of S , or S_R , and the other subtree—the left subtree S_L^* —contains the “extra” taxon of S^* ; (ii) the “extra” taxon of S^* descends directly from the root of S^* , with the other subtree of S^* descended from the root being S .

In case (i), using Theorem 3.1 and the inductive hypothesis,

$$\begin{aligned} A_{S^*, m} &= \sum_{k=1}^m A_{S_L^*, k+1} A_{S_R, k+1} \\ &\geq \sum_{k=1}^m 2 A_{S_L, k+1} A_{S_R, k+1} \\ &= 2 A_{S, m}. \end{aligned}$$

In case (ii), using Theorem 3.1 and Corollary 3.2,

$$\begin{aligned}
 A_{S^*,m} &= \sum_{k=1}^m A_{S,k+1} \\
 &= \sum_{k=1}^m \sum_{j=1}^{k+1} A_{S_L,j+1} A_{S_R,j+1} \\
 &\geq \sum_{k=1}^m \sum_{j=k}^{k+1} A_{S_L,j+1} A_{S_R,j+1} \\
 &\geq \sum_{k=1}^m 2A_{S_L,k+1} A_{S_R,k+1} \\
 &= 2A_{S,m}.
 \end{aligned}$$

This verifies the result in the case that $n_1 = n_2 - 1$; repeated application of Equation (7) with $n_1 = n_2 - 1$ establishes the result for $n_2 - n_1 > 1$. ■

Corollary 3.13 (Degnan and Salter, 2005). *For $n \geq 3$ the number of coalescent histories for a tree with n taxa exceeds that of any of its displayed trees with $n - 1$ taxa by a multiplicative factor of at least 2.*

Proof. This result follows directly from Theorem 3.12 with $m = 1$. ■

Theorem 3.14. *For $m \geq 1$, the number of valid m -extended coalescent histories for a tree S with $n \geq 2$ taxa is (i) greater than or equal to $2^{n-2}m$, and (ii) less than or equal to $\prod_{r=2}^n (m + n - r)^{d_r(S)}$, where $d_r(S)$ is the number of internal nodes of S from which exactly r taxa descend.*

Proof. (i) No matter where other coalescences occur, each coalescence of the gene tree—except the final coalescence that joins the left and right subtrees—has at least two branches of the species tree on which it can happen: the most recent branch from which all of its taxa descend, and the branch immediately ancestral to this branch. When all other coalescences happen on their first or second possible branches, the final coalescence that joins the left and right subtrees of the gene tree has m possible branches on which it can occur. Because there are $n - 2$ coalescences other than the final coalescence, the minimum number of valid m -extended coalescent histories is $2^{n-2}m$.

(ii) No matter where other coalescences of gene lineages occur, a coalescence from which r lineages descend has at least $r - 2$ internal branches of the species tree below it, so that the maximum number of species tree internal branches on which the coalescence can happen is $(n - 2) - (r - 2) = n - r$; in addition, the coalescence has the potential to occur on any of the m branches above the root of S . Thus, a coalescence from which r lineages descend has at most $m + n - r$ branches on which it can take place. Taking a product over all $n - 1$ coalescences, the result follows. ■

Corollary 3.15. *The number of valid coalescent histories for a tree S with $n \geq 2$ taxa is (i) greater than or equal to 2^{n-2} , and (ii) less than or equal to $\prod_{r=2}^{n-1} (n - r + 1)^{d_r(S)}$.*

Proof. This follows from Theorem 3.14, with $m = 1$. ■

Corollary 3.16 (Degnan and Salter, 2005). *The number of valid coalescent histories for any tree with $n \geq 2$ taxa is (i) greater than or equal to 2^{n-2} , and (ii) less than or equal to $(n - 1)^{n-2}$.*

Proof. Result (i) follows directly from Corollary 3.15i, and (ii) follows from the fact that the product in Corollary 3.15ii has $n - 2$ terms—one for each internal node of the tree other than the root—and from the fact that each term is at most $n - 1$. ■

Corollary 3.17. *The number of valid coalescent histories for any tree with $n \geq 4$ taxa is less than or equal to*

$$(n-1)^{\lfloor n/2 \rfloor} (n-2)^{\lfloor n/3 \rfloor} (n-3)^{n-2-\lfloor n/2 \rfloor-\lfloor n/3 \rfloor} \quad (8)$$

Proof. That (8) provides an upper bound for $n = 4, 5,$ or 6 taxa can be verified from Table 1. Consider a tree S with $n \geq 7$ taxa. The product in Corollary 3.15ii can be bounded above by maximizing the exponent for the term with $r = 2$ and then the exponent for the $r = 3$ term. Because at most $\lfloor n/2 \rfloor$ nodes of S have exactly 2 descendants, $d_2(S) \leq \lfloor n/2 \rfloor$. Similarly, $d_3(S) \leq \lfloor n/3 \rfloor$.

The total number of terms in the product in Corollary 3.15ii is $\sum_{r=2}^{n-1} d_r(S) = n-2$. Suppose the exponent for the term with $r = 2$ is set to its upper bound of $\lfloor n/2 \rfloor$ and the exponent for the $r = 3$ term is set to its upper bound of $\lfloor n/3 \rfloor$. The number of remaining terms after these exponents are set is $n-2-\lfloor n/2 \rfloor-\lfloor n/3 \rfloor$, a nonnegative number for $n \geq 7$. Each of these remaining terms is at most $n-3$, from which the result follows. ■

Theorem 3.18. *For a given $n \geq 2$, the ratio of the numbers of valid coalescent histories for the n -taxon tree with the greatest number of coalescent histories and the n -taxon tree with the smallest number of coalescent histories is greater than*

$$\left(\frac{\sqrt{\pi}}{32}\right) \left(\frac{5n-12}{4n-6}\right) n\sqrt{n}. \quad (9)$$

Proof. For $n = 2, 3,$ or 4 the result can be verified from Table 1. Suppose $n \geq 5$. The number of coalescent histories for the n -taxon tree with the greatest number of coalescent histories is greater than or equal to the number of coalescent histories for the pseudocaterpillar tree, or $(5n-12)C_{n-1}/(4n-6)$. The number of coalescent histories for the n -taxon tree with the smallest number of coalescent histories is less than or equal to the number of coalescent histories for the bicaterpillar tree whose caterpillar subtrees have $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ taxa, or $C_{\lfloor n/2 \rfloor} C_{\lceil n/2 \rceil}$. Therefore, the desired ratio must be greater than or equal to

$$\frac{5n-12}{4n-6} \frac{C_{n-1}}{C_{\lfloor n/2 \rfloor} C_{\lceil n/2 \rceil}}. \quad (10)$$

We can expand the Catalan terms using Equation (4) and use Stirling's approximation for factorials, $e^{1/(12n+1)}(n/e)^n \sqrt{2\pi n} < n! < e^{1/(12n)}(n/e)^n \sqrt{2\pi n}$ (Feller, 1968, p. 54), so that the lower bound is applied to factorials in the numerator of expression 10 and the upper bound is applied to factorials in the denominator.

For even n , we obtain

$$\frac{C_{n-1}}{C_{\lfloor n/2 \rfloor} C_{\lceil n/2 \rceil}} > \frac{\sqrt{\pi}}{16} \frac{(n+2)^2 \sqrt{n}}{2n-1} \exp\left(\frac{162n^2-15n-1}{3n(6n+1)(24n+1)}\right), \quad (11)$$

and for odd n ,

$$\frac{C_{n-1}}{C_{\lfloor n/2 \rfloor} C_{\lceil n/2 \rceil}} > \frac{\sqrt{\pi}}{32} \frac{(n+1)^2(n+3)\sqrt{n-1}}{n^2} \exp\left(\frac{162n^2-339n+176}{3(n-1)(6n-5)(24n-23)}\right). \quad (12)$$

In both the even and odd cases, it is straightforward to show that for $n \geq 5$, the lower bounds for $C_{n-1}/(C_{\lfloor n/2 \rfloor} C_{\lceil n/2 \rceil})$ exceed $(\sqrt{\pi}/32)n\sqrt{n}$.

Remarks. Theorem 3.18 gives a lower bound on the the ratio of the numbers of coalescent histories for the n -taxon trees with the largest and smallest numbers of these histories. A very loose upper bound on this ratio can be obtained for $n \geq 2$ from Corollary 3.16 as $[(n-1)/2]^{n-2}$. For $n \geq 4$, Corollary 3.17 enables this upper bound to be sharpened slightly to $[(n-1)/2]^{\lfloor n/2 \rfloor} [(n-2)/2]^{\lfloor n/3 \rfloor} [(n-3)/2]^{n-2-\lfloor n/2 \rfloor-\lfloor n/3 \rfloor}$.

4. LABELED TOPOLOGY OF THE SPECIES TREE AND GENE TREE NOT NECESSARILY IDENTICAL

In this section, we relax the assumption that the gene tree under consideration has the same labeled topology as the species tree. We denote by $B_{G,S,m}$ the number of valid m -extended coalescent histories in which the gene tree has labeled topology G and the species tree has labeled topology S ; the number of valid coalescent histories for G and S is obtained from $B_{G,S,1}$. In this section coalescent histories are a property of a gene tree/species tree pair, as the gene tree and species tree labeled topologies are not necessarily identical. The gene tree is assumed to have a subset of the taxa contained in the species tree, but the two trees need not have identical sets of taxa.

For an arbitrary gene tree labeled topology and species tree labeled topology, we generalize the recursion in Theorem 3.1 to count the number of valid m -extended coalescent histories when the gene tree and species tree do not necessarily have the same labeled topology, showing that the generalized recursion reduces to Equation (1) when $G = S$. We also give a general procedure for enumerating m -extended coalescent histories. We begin with two definitions, examples of which are illustrated in Figure 4.

Definition 4.1. For a tree S with $n \geq 2$ taxa and a tree G whose taxa are a subset of those of S , let $T(G, S)$ denote the minimal displayed tree of S that is induced by the taxa of G .

Definition 4.2. For a tree S with $n \geq 2$ taxa and a tree G whose taxa are a subset of those of S , let $d(G, S)$ denote the number of branches of S that separate the root of S from the root of $T(G, S)$. For a given tree S , considering all possible trees G whose taxa form a subset of those of S , the value of $d(G, S)$ ranges from 0 to the largest number of branches separating a leaf from the root of S .

Theorem 4.3. For $m \geq 1$, the number of valid m -extended coalescent histories for a species tree labeled topology S with $n \geq 2$ taxa and a gene tree labeled topology G whose taxa are a subset of those of S is

$$B_{G,S,m} = \sum_{k=1}^m B_{G_L,T(G_L,S),k+d(G_L,S)} B_{G_R,T(G_R,S),k+d(G_R,S)}, \tag{13}$$

where G_L and G_R respectively denote the left and right subtrees of G . If $n = 1$ or G has only one taxon, then $B_{G,S,m} = 1$ for all m .

Proof. If $n = 1$, no coalescences occur, and trivially there is only one m -extended coalescent history, regardless of the value of m . Exactly as in the case where $G = S$, for $n \geq 2$, the final coalescence that joins the ancestor of the left subtree of G and the ancestor of the right subtree of G can occur on any of the m branches above the root of S .

Suppose the final gene tree coalescence happens on branch k , with $1 \leq k \leq m$. The left subtree of the gene tree, G_L , has $k + d(G_L, S)$ branches of the species tree above the root of $T(G_L, S)$ on which its

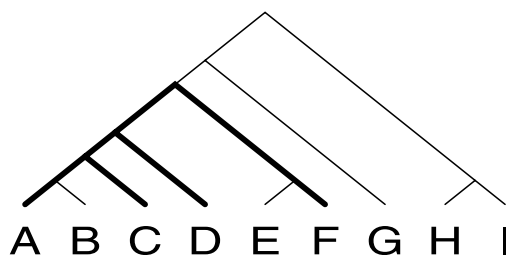


FIG. 4. Displayed trees. For a tree S with nine taxa, the minimal displayed tree of taxa A, C, D, and F is shown with thick lines. The number of internal branches of S that separate the root of the displayed tree from the root of S is 2.

coalescences can occur. The root of G_L must be located above the root of $T(G_L, S)$ and the number of branches above this root on which coalescences of G_L can take place is k plus the number of branches separating the root of $T(G_L, S)$ and the root of S , or $d(G_L, S)$. A similar result holds for G_R . Because the coalescences of G_L and G_R can occur in any order with respect to each other, the number of m -extended coalescent histories for (G, S) with final coalescence on branch k above the root of S is the product of the number of $(k + d(G_L, S))$ -extended coalescent histories for the pair (G_L, S) and the number of $(k + d(G_R, S))$ -extended coalescent histories for (G_R, S) . Summing over possible values of k , the result follows. ■

Remarks. As in the case of $G = S$, Theorem 4.3 is straightforward to extend to multifurcating gene trees by replacing the summand in Equation (13) with a product of the number of $(k + 1)$ -extended coalescent histories over all subtrees immediately descended from the root of G . The result is applicable without modification if only the species tree is multifurcating. Note also that the closed form expressions in results 3.14–3.17 for the upper bounds on the number of m -extended coalescent histories—but not the lower bounds—apply if the gene tree labeled topology G and the species tree labeled topology S are not necessarily identical but have the same taxa.

Theorem 4.3 contains Theorem 3.1 as a special case. If $G = S$, then $T(G_L, S) = G_L$ and $T(G_R, S) = G_R$. Further, because $G_L = S_L$ and $G_R = S_R$, where S_L and S_R respectively denote the labeled topologies for the left and right subtrees of S , the distances $d(G_L, S)$ and $d(G_R, S)$ both equal 1. As a result, the recursion for $B_{G,S,m}$ in Equation (13) when $G = S$ is the same as the recursion for $A_{S,m}$ in Equation (1), and $B_{G,S,m}$ reduces to $A_{S,m}$.

By a similar procedure to the approach in the case of $G = S$, Theorem 4.3 makes it possible to recursively enumerate all m -extended coalescent histories for a gene tree labeled topology G and a species tree labeled topology S . We use the same postorder traversal of S as in Section 3. For each k with $1 \leq k \leq m$, a k -extended coalescent history for (G, S) is obtained by combining a $(k + d(G_L, S))$ -extended coalescent history for (G_L, S) and a $(k + d(G_R, S))$ -extended coalescent history for (G_R, S) , and specifying that the final coalescence of G occurs on branch k above the root of S . The full list of k -extended coalescent histories for (G, S) is obtained by considering all pairs involving a $(k + d(G_L, S))$ -extended coalescent history for (G_L, S) and a $(k + d(G_R, S))$ -extended coalescent history for (G_R, S) ; the full list of m -extended coalescent histories for (G, S) is obtained by applying this procedure for all k with $1 \leq k \leq m$.

ACKNOWLEDGMENTS

I am grateful to J. Degnan and L. Nakhleh for helpful conversations. This research was supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences and by an Alfred P. Sloan Research Fellowship.

REFERENCES

- Degnan, J.H. 2005. Gene tree distributions under the coalescent process. [Ph.D. dissertation]. University of New Mexico, Albuquerque.
- Degnan, J.H., and Rosenberg, N.A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2, 762–768.
- Degnan, J.H., and Salter, L.A. 2005. Gene tree distributions under the coalescent process. *Evolution* 59, 24–37.
- Feller, W. 1968. *An Introduction to Probability Theory and Its Applications, Volume 1*, 3rd ed. Wiley, New York.
- Maddison, W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46, 523–536.
- Pamilo, P., and Nei, M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.
- Rannala, B., and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164, 1645–1656.
- Rosenberg, N.A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Pop. Biol.* 61, 225–247.
- Sedgewick, R., and Flajolet, P. 1996. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, Boston.

- Slatkin, M., and Pollack, J.L. 2006. The concordance of gene trees and species trees at two linked loci. *Genetics* 172, 1979–1984.
- Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122, 957–966.

Address reprint requests to:
Noah A. Rosenberg
Department of Human Genetics
Bioinformatics Program
and the Life Sciences Institute
University of Michigan
100 Washtenaw Ave.
Ann Arbor, MI 48109-2218

E-mail: rnoah@umich.edu