

## GENE TREE DISTRIBUTIONS UNDER THE COALESCENT PROCESS

JAMES H. DEGNAN<sup>1,2</sup> AND LAURA A. SALTER<sup>1,3</sup>

<sup>1</sup>*Department of Mathematics and Statistics, MSC03 2150, University of New Mexico, Albuquerque, New Mexico 87131*

<sup>2</sup>*E-mail: james@stat.unm.edu*

<sup>3</sup>*E-mail: salter@stat.unm.edu*

**Abstract.**—Under the coalescent model for population divergence, lineage sorting can cause considerable variability in gene trees generated from any given species tree. In this paper, we derive a method for computing the distribution of gene tree topologies given a bifurcating species tree for trees with an arbitrary number of taxa in the case that there is one gene sampled per species. Applications for gene tree distributions include determining exact probabilities of topological equivalence between gene trees and species trees and inferring species trees from multiple datasets. In addition, we examine the shapes of gene tree distributions and their sensitivity to changes in branch lengths, species tree shape, and tree size. The method for computing gene tree distributions is implemented in the computer program COAL.

**Key words.**—Coalescence, coalescent history, lineage sorting, maximum likelihood, phylogenetics, species tree.

Received June 18, 2004. Accepted October 8, 2004.

The results of population genetic and phylogenetic studies often depend upon the relationships between gene trees, which represent the evolutionary histories of genes, and species trees, which represent the evolutionary relationships for a set of taxa. For example, in population genetic studies, the probability that the gene tree topology is equivalent to the species tree topology can be used to estimate ancestral population sizes (Wu 1991). In phylogenetics, a tree inferred from a single gene reflects the evolutionary history of the taxa in the study only if the gene tree and species tree have the same topology.

Many aspects of biological evolution may result in a gene tree that is not topologically equivalent to the species tree for the same set of taxa, including horizontal gene transfer, gene duplication, hybridization, and nonneutral evolution (Hein 1993; Syvanen 1994; Maddison 1997; Sang and Zhong 2000). However, even when these forces are not present, coalescent theory predicts that the topologies of gene trees show considerable stochastic variation and that there is therefore a high probability that the topologies of the gene tree and the underlying species tree will differ (Pamilo and Nei 1988; Takahata 1989). Distributions of gene tree topologies given fixed species trees have been derived for trees with fewer than five taxa under the coalescent model (Takahata and Nei 1985; Pamilo and Nei 1988; Rosenberg 2002). This paper presents a general method for computing these probabilities for trees with an arbitrary number of taxa. This method is implemented in the computer program COAL, which computes the probability of a gene tree given a bifurcating species tree with the same number of taxa.

Although probabilities for gene trees can be simulated under the coalescent model, computing them directly will help in assessing the robustness of methods that assume the topological equivalence of gene trees and species trees (Knowles and Maddison 2002) and in understanding shapes of gene tree distributions. For example, in a recent paper (Poe and Chubb 2004), lack of congruence among gene trees was taken as evidence for a hard polytomy in early Avian evolution. Although the method proposed here does not model multifurcations (polytomies), the ability to calculate exact gene

tree probabilities for larger trees than previously possible will enable the investigation of the effects of arbitrarily short branches on gene tree distributions. These distributions can also be used to implement Maddison's (1997) proposal to infer species trees in a maximum likelihood framework. This approach is important because it incorporates information from several genes without assuming that the datasets can be concatenated and treated as a single gene, as advocated by several others, for example, Rokas et al. (2003). Other applications include testing phylogeographic hypotheses (Knowles and Maddison 2002) and inferring population parameters (such as ancestral population size,  $\theta$ ; Felsenstein 2004).

### BACKGROUND AND TERMINOLOGY

The coalescent process models divergence between gene lineages by considering the time from the present, when the genes are sampled, to the time when the genes diverged. Looking at gene lineages from the present to the past, the divergence between two lineages can also be called "a coalescence," or "coalescent event." Because it is sometimes more convenient to refer to time backwards, starting from the present, the words "before," "after," "beginning," and "until" are sometimes used so that, for example, "before" means "more recently than." The phrase "prior to" is used to mean "more ancient." For the problem considered in this paper, only one gene is sampled per population, and intra-specific variation is not modeled. In this case, coalescence between two lineages can only have occurred either at or prior to the time that their most recent ancestral population diverged.

As an example, for the species tree in Figure 1a, although B is more closely related to C than to A, the population that was ancestral to B and C (but not to A) might have had two lineages whose most recent common ancestor existed prior to the root of the tree. In this case, the version of the gene that survived in the B lineage might have been a direct descendant of the lineage that gave rise to A, and the version of the gene that survived in the C lineage might have been a direct descendant of the lineage giving rise to D, E, F, and

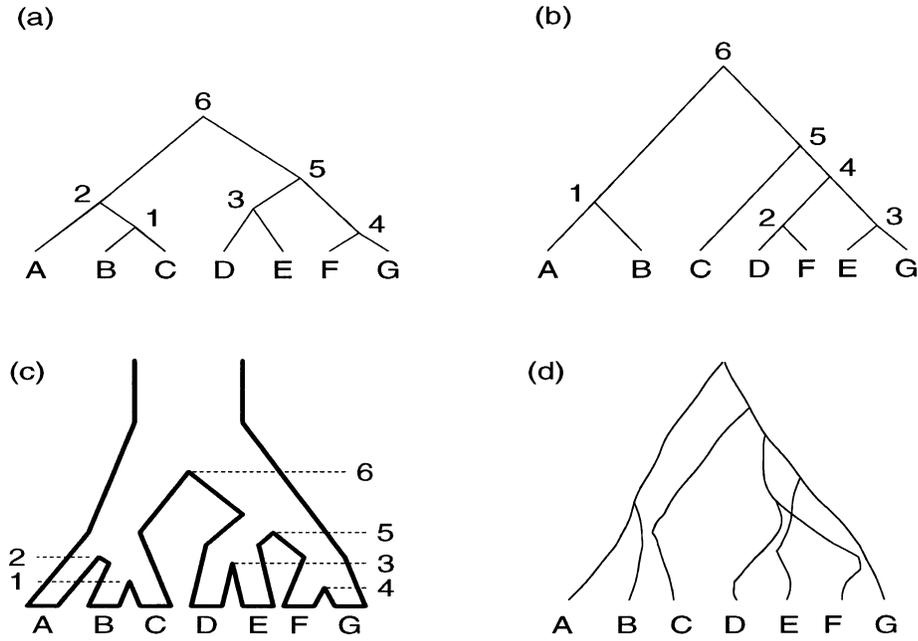


FIG. 1. (a,b) The species tree and gene tree, respectively, for the seven-taxon example, with internal nodes (equivalently, internal branches) labeled according to a postorder traversal (Rosen 1999) so that no node has a lower number than any of its descendants. (c) Another representation of the species tree. The bifurcation points (upside-down v's), which are pointed to by the dotted lines in (c), correspond to the nodes of the tree in (a). Note that (a) and (c) are drawn to the same scale, so that the branch lengths of (a) are the same as the distances between the corresponding nodes of (c). (d) The gene tree with the branch lengths changed and curved so that it can be superimposed on the species tree (c) to produce Figure 2a.

G. (See Fig. 2 for several scenarios in which this occurs.) Therefore, even though B and C share a recent ancestral population, the B and C lineages for the gene of interest are only related by a much more ancient ancestral population. Because this population must be at least as ancient as the root, it must be ancestral both to the ((DF)(EG)) clade (with which C has coalesced) and to A (with which B has coalesced). In this case, lineage sorting causes the gene tree and species tree to have different topologies. Because different genes can have separate evolutionary histories, there is no a priori reason to expect that gene trees based on different genes should have the same topologies. By sufficiently changing the branch lengths (and, for graphical purposes, allowing the branches to bend), any gene tree can be made to fit within any species tree (Fig. 1d). To see that this is always true, note that there is always the possibility that none of the gene lineages coalesce more recently than the root of the species tree. If all lineages coalesce prior to the root, then the lineages can coalesce in any order, thereby producing any desired topology for the gene tree. As will be seen below, however, gene trees with topologies radically different from that of the species tree tend to have very low probabilities.

Coalescent theory, which models coalescent times for a set of lineages, can be used to calculate the probability that two or more lineages coalesce within a fixed amount time. By treating the species tree (including branch lengths) as fixed, one can calculate the probability that, for instance, two lineages coalesce into one, or more generally that  $u$  lineages coalesce into  $v$  lineages, within the amount of time determined by the length of the branch. This probability can then be expressed as  $p_{uv}(T)$ , where  $T$  is the length of the branch.

Here  $u \geq v \geq 1$ , and if  $u = v$ , then  $p_{uv}(T)$  is the probability that no coalescences occur in the length of time  $T$ . In the coalescent model,  $T = t/(2N)$ , where  $t$  is the number of generations and  $N$  is the effective population size—the number of diploid individuals—which is assumed to be constant (Tajima 1983; Takahata and Nei 1985). An expression for this probability was given by Rosenberg (2002) and was derived earlier by several others (Tavaré 1984; Watterson 1984; Takahata and Nei 1985):

$$p_{uv}(T) = \sum_{k=v}^u e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)} \times \prod_{y=0}^{k-1} \frac{(v+y)(u-y)}{(u+y)}. \quad (1)$$

By counting the number of ways that a gene tree could have arisen on the species tree and determining the  $p_{uv}(T)$  term for each branch, one can calculate the probability of a gene tree for a fixed species tree. This is feasible to do by hand for small trees (e.g., fewer than six taxa). For larger trees, the notation and concepts that follow can be used to develop an algorithm for the computation.

This method requires keeping track of lineages from a gene tree coalescing on branches of the species tree. The internal nodes of the trees are labeled according to a postorder traversal (Rosen 1999; Fig. 1). The terms ‘node’ and ‘clade’ are used interchangeably, so that, for example, the clade ((DE)(FG)) of the species tree in Figure 1a is called either node 5 or clade 5. In addition, each branch has the number of the node incident to that branch and closer to the tips of the tree. For example, branch 5 of the species tree is the

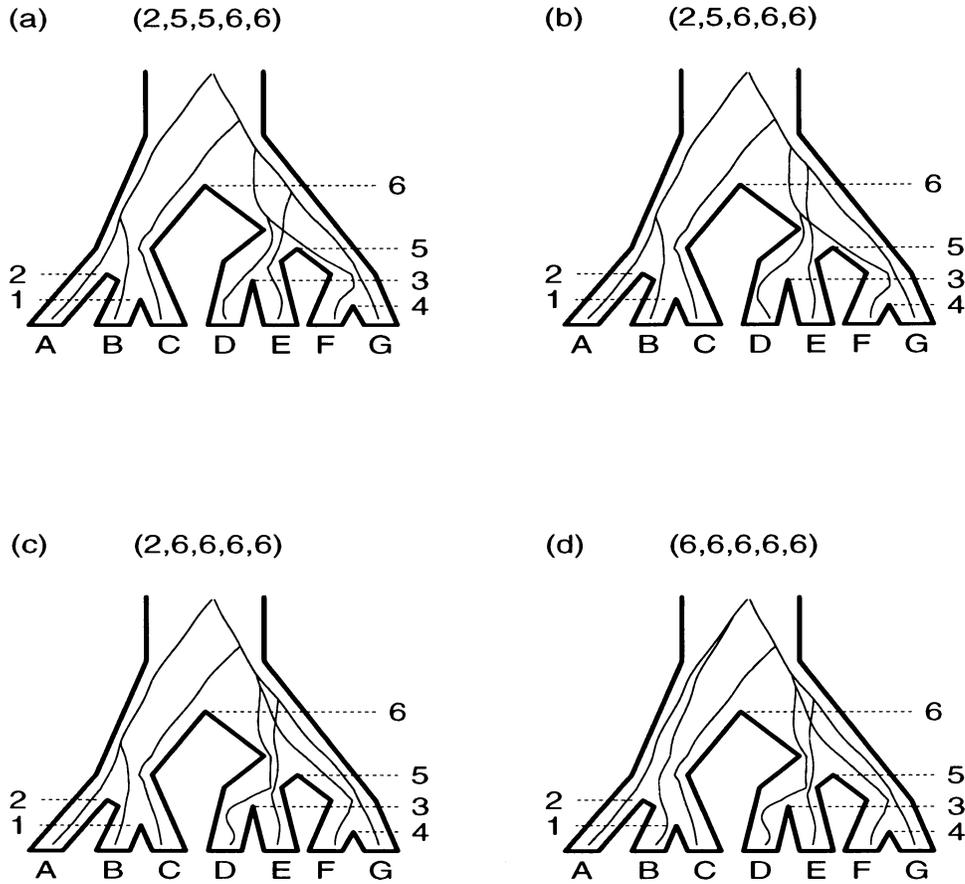


FIG. 2. Four of the 10 valid histories for the species and gene trees shown in Figures 1a and 1b, respectively. Each subfigure shows one instantiation of the coalescent history. For example, (a) shows D and F coalescing to form (DF) before E and G coalesce to form (EG), although the other order is also consistent with the gene tree topology. For the histories in (a) and (c) there are two instantiations. The histories in (b) and (d) have one and 10 instantiations, respectively.

branch connecting clade 5 and the root. Note that rooted bifurcating trees with  $n$  taxa have  $n - 1$  internal nodes (including the root) and  $n - 2$  internal branches.

The phrase “clade  $k$  coalesces on branch  $b$ ” is used to mean that the two subclades of clade  $k$  on the gene tree, which can be regarded as lineages, coalesce on branch  $b$  of the species tree. For the trees in Figure 1, “clade 3 of the gene tree coalesces on branch 5 of the species tree” means that E and G, the two lineages in (EG), coalesce on branch 5 of the species tree. Although the tree in this example does not have a sixth branch, the phrase “clade 5 coalesces on branch 6” is interpreted to mean that the ((DF)(EG)) lineage and the C lineage coalesce prior to the root.

A set of coalescences can be represented as a vector  $\mathbf{h} = (h_1, h_2, \dots, h_{n-2})$ . Each element of  $\mathbf{h}$  corresponds to an internal node (or clade) of the gene tree (not including the root), and the value of the element of the vector is the branch of the species tree on which the clade coalesces. Hence  $h_k = b$  if and only if clade  $k$  of the gene tree coalesces on branch  $b$  of the species tree, with the interpretation that if  $h_k = n - 1$ , then clade  $k$  coalesces prior to the root. A vector  $\mathbf{h}$  is called a “coalescent history.”

Not every vector  $\mathbf{h}$  corresponds to a coalescent history that is consistent with the gene and species tree topologies. To

be consistent with the gene tree and species tree, coalescences between two lineages must occur at least as anciently on the species tree as the most recent common ancestor of the lineages coalescing. In addition, a coalescent history cannot require a clade of the gene tree to coalesce more recently than one of its descendants. A useful way of counting coalescent histories is to propose a vector  $\mathbf{h}$  and then to check whether  $\mathbf{h}$  corresponds to a history. If a proposed coalescent history  $\mathbf{h}$  is consistent with the gene and species trees, then the history is “valid.” The use of “history” or “coalescent history” without qualification refers to a valid history. For the trees considered above, (2,6,5,5,5) is not a valid history because clade 5 of the gene tree cannot coalesce on branch 5 of the species tree. Clade 5 of the gene tree can only coalesce prior to the most recent common ancestor (in the species tree) of all taxa present in clade 5 of the gene tree: C, D, E, F, and G. Because this most recent common ancestor is the root of the tree, clade 5 can only coalesce prior to the root. Hence  $h_5$  is necessarily 6 for any valid history for these trees.

There are 10 valid histories (Table 1) for the species tree and gene tree used above, four of which are shown in Figure 2. A more compact display of these histories, which depicts all 10 histories simultaneously on the same tree, is provided

TABLE 1. Probabilities of each coalescent history for the seven-taxon species tree and gene tree shown in Figures 1a and 1b, respectively. In the first three branch length columns, all branches of the species tree have the indicated length. The fourth and fifth columns have all branches with length 1.0 except the indicated branch. For the fourth column, one branch that is closest to the tips is set to 0.01; for the fifth column, the branch set to 0.01 is one of the two most basal branches (the one on the right). For each pattern of branch lengths, the history with the highest probability is indicated by an asterisk. The row of totals shows the total probability of the gene tree given the species tree, which is the sum of the probabilities of each history when the species tree has the indicated branch lengths.

History	Branch lengths				
	1.0	0.5	0.2	$\lambda_1 = 0.01$	$\lambda_5 = 0.01$
(2, 5, 5, 6)	.0001135295*	.0001923529*	.0000407575	.0003055344*	.0000000009
(2, 5, 5, 6, 6)	.0000252118	.0001272950	.0000862809	.0000678507	.00000000426
(2, 5, 6, 6, 6)	.0000006937	.0000137307	.0000339637	.0000018670	.00000004206
(2, 6, 5, 6, 6)	.0000006937	.0000137307	.0000339637	.0000018670	.00000004206
(2, 6, 6, 6, 6)	.000000145	.0000015775	.0000165249	.0000000391	.0000055239*
(6, 5, 5, 6)	.0000118462	.0000746299	.0000552466	.0000318810	.0000000001
(6, 5, 5, 6, 6)	.0000023677	.0000444496	.0001052582*	.0000063719	.0000000040
(6, 5, 6, 6, 6)	.0000000579	.0000042618	.0000368301	.0000001558	.0000000351
(6, 6, 5, 6, 6)	.0000000579	.0000042618	.0000368301	.0000001558	.0000000351
(6, 6, 6, 6, 6)	.0000000011	.0000004372	.0000159996	.0000000029	.00000004117
Total	.0001544740	.0004767273	.0004616552	.0004157257	.0000068945

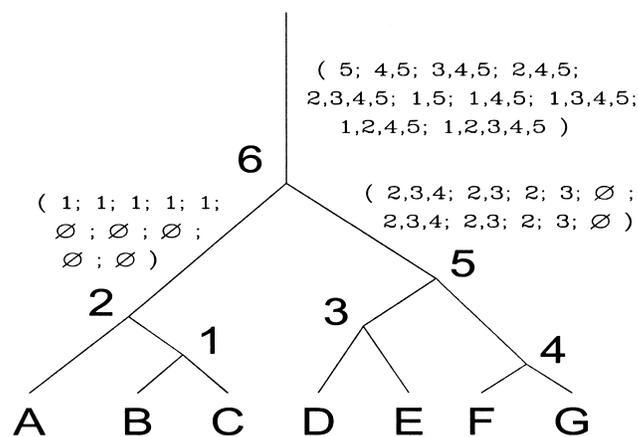


FIG. 3. The set of coalescent histories can be represented more compactly by showing, on each branch of the species tree, the clades of the gene tree that coalesced on that branch. Branches that have no coalescent events for any of the histories are left blank. The coalescent histories are separated by semicolons, and an empty set symbol indicates that for that history, nothing coalesced on that branch. As an example, for the histories ordered as in Table 1, the second coalescent history, (2,5,5,6,6), has clade 1 of the gene tree coalescing on branch 2 of the species tree, clades 2 and 3 of the gene tree coalescing on branch 5, and clades 4 and 5 coalescing prior to the root. This method of representing all coalescent histories can preserve information regarding the ordering of coalescent events within a branch. The labeling (2,3,4; 2,3; 2; 3;  $\emptyset$ ; 2,3,4; 2,3; 2; 3;  $\emptyset$ ) on branch 5 suggests that for the first, second, sixth, and seventh histories, the coalescent event forming the (DF) lineage, which is clade 2 of the gene tree, occurred before the event forming the (EG) lineage, which is clade 3 of the gene tree.

in Figure 3. The history (2,5,5,6,6) (Fig. 2a) indicates that A and B coalesce on branch 2 of the species tree to form the (AB) lineage, D and F coalesce on branch 5 to form (DF), E and G also coalesce on branch 5 to form (EG), and the remaining two clades coalesce prior to the root. If two or more lineages coalesce on the same branch (or prior to the root), then the coalescent history does not distinguish the order of these coalescences. For the history (2,5,5,6,6), two clades coalesce on branch 5, (DF) and (EG), and either order is possible: either D and F coalesce first to form (DF) before E and G coalesce, or E and G coalesce first to form (EG) before D and F coalesce. Although there are also two coalescent events on branch 6 (prior to the root), only one ordering is consistent with the gene tree because (DF) and (EG) must coalesce to form ((DF)(EG)) before C can coalesce with ((DF)(EG)). Thus there are two ways for the history (2,5,5,6,6) to occur that are consistent with the gene tree topology. Orderings of coalescent events that are consistent with the gene tree are referred to as different “instantiations” of the same history. Hence, there are two instantiations of the history (2,5,5,6,6). Calculating the probability of a gene tree requires counting how many orderings of coalescent events on a given branch are “correct,” in the sense that they are consistent with the gene tree topology. The number of instantiations of a coalescent history depends on the number of correct orderings of coalescent events for each branch. Given a fixed species tree, the probability of a gene tree can be obtained by adding the probability of each valid history. The probability of a particular history can be determined

from the probability of the events on each branch, using the  $p_{uv}(T)$  terms from equation (1). Here the probability that  $u$  lineages coalesce into  $v$  lineages on a specific branch must also reflect the possibility that there might be more than one way for these coalescences to occur. Assuming that all possible orderings, correct and incorrect, are equally likely (this follows the model of Yule [1924]), and that  $u$  lineages have coalesced into  $v$  lineages, the probability that the ordering of events on the branch is consistent with the gene tree is the number of correct orderings divided by the number of possible orderings. Multiplying the probabilities of the events on different branches yields the probability of the coalescent history for the entire tree. (These probabilities are not necessarily independent because they depend on the events of descendent branches. This will be discussed below.) Given a list of coalescent histories and the probabilities of events on each branch, the overall probability of a gene tree given a species tree can be obtained.

Under the coalescent model, for a given  $n$ -taxon species tree with topology  $\psi$  and vector of branch lengths  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_{n-2})$ , where each  $\lambda_b$  is measured in units of  $2N$  generations, the gene tree topology  $G$  is a random variable with probability mass function

$$P_{\psi,\lambda}(G = g) = \sum_{\mathbf{h} \in H_\psi(g)} \frac{w(\mathbf{h})}{d(\mathbf{h})} \prod_{b=1}^{n-2} \frac{w_b(\mathbf{h})}{d_b(\mathbf{h})} p_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b). \quad (2)$$

The sum is over all histories  $\mathbf{h}$  taken from the set  $H_\psi(g)$  of all valid coalescent histories for the particular gene tree and species tree. The product is taken over all internal branches of the species tree, labeled  $1, 2, \dots, n-2$ . The terms  $p_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b)$ , which can be determined from equation (1), are used to calculate the probability, for a particular coalescent history  $\mathbf{h}$  and a particular branch  $b$ , that  $u_b(\mathbf{h})$  lineages coalesce into  $v_b(\mathbf{h})$  lineages in the time  $\lambda_b$ , the length of branch  $b$ . Here  $2 \leq u_b(\mathbf{h}) \leq \delta_b$ , where  $\delta_b$  is the number of taxa for which branch  $b$  is an ancestor, and  $1 \leq v_b(\mathbf{h}) \leq u_b(\mathbf{h})$ . The terms  $w_b(\mathbf{h})/d_b(\mathbf{h})$  and  $w(\mathbf{h})/d(\mathbf{h})$  determine the probability for each branch and prior to the root that the coalescent events are consistent with the gene tree. Here  $w_b(\mathbf{h})$  is the number of ways that coalescent events on a branch can occur consistently with the gene tree, and  $d_b(\mathbf{h})$  is the number of possible orderings of events.

The next two sections provide details for enumerating coalescent histories and for computing the necessary terms in equation (2), and can be skipped without loss of continuity. These are followed by discussions regarding the shape of gene tree distributions, the number of coalescent histories, applications, and possible extensions.

#### ENUMERATING COALESCENT HISTORIES

To enumerate the set of valid coalescent histories  $H_\psi(g)$ , each proposed coalescent history  $\mathbf{h}$  can be identified with an integer  $h = \sum_{k=1}^{n-2} (h_k - 1) (n-1)^{n-2-k}$ . The values of  $h$  are at most  $(n-1)^{n-2} - 1$ , which corresponds to the coalescent history  $\mathbf{h} = (n-1, n-1, \dots, n-1)$ . This value of  $h$  occurs when all clades of the gene tree coalesce prior to the root. If a proposed history  $\mathbf{h}$  is valid, then  $\mathbf{h} \in H_\psi(g)$ . The problem of determining the set of valid coalescent histories

is to enumerate values of  $h$  for which  $0 \leq h \leq (n-1)^{n-2} - 1$  and that correspond to valid histories.

To check whether a proposed coalescent history  $\mathbf{h} = (h_1, h_2, \dots, h_{n-2})$  is valid, it must be the case that if  $h_k = b$ , that is, if clade  $k$  of the gene tree coalesces on branch  $b$  of the species tree, then all of the taxa in clade  $k$  must be descendants of branch  $b$  on the species tree. These restrictions on valid coalescent histories can be summarized in a matrix  $\mathbf{M} = (m_{ij})$ , where  $m_{ij} = 1$  if clade  $j$  of the gene tree only includes taxa that are also in clade  $i$  of the species tree; otherwise,  $m_{ij} = 0$ . Therefore, a necessary condition for a coalescent history  $\mathbf{h} = (h_1, h_2, \dots, h_{n-2})$  to be valid is that if  $h_k = b$ , and if  $b \leq n-2$ , then  $m_{bk} = 1$  for all  $k = 1, 2, \dots, n-2$ . Note that any clade of the gene tree can coalesce prior to the root of the species tree, and that the clade associated with the root of the gene tree can only coalesce prior to the root of the species tree. Consequently, we only need to keep track of clades  $1, 2, \dots, n-2$  of the gene tree and branches  $1, 2, \dots, n-2$  of the species tree, so the  $\mathbf{M}$  matrix is  $(n-2) \times (n-2)$ .

For even a moderate number of taxa, the maximum value of  $h$ ,  $(n-1)^{n-2} - 1$ , is unmanageably large. To reduce the number of histories that must be evaluated,  $h$  can be incremented more rapidly by skipping over large numbers of consecutively occurring histories that are not allowed. In particular, if the proposed coalescent history has the form  $\mathbf{h} = (h_1, h_2, \dots, h_k, 1, \dots, 1)$  with  $h_k = b > 1$  and  $m_{bk} = 0$ , then that history is prohibited by the  $\mathbf{M}$  matrix, as are all remaining vectors that have  $h_k = b$ . If the proposed histories are enumerated sequentially, then the next  $(n-1)^{n-2-k} - 1$  histories are invalid, and therefore do not need to be checked. This greatly reduces the number of histories that must be evaluated.

As an example of filling in the  $\mathbf{M}$  matrix for Figures 1a and 1b, consider  $m_{52}$ . Because all taxa in clade 2 of the gene tree are present in clade 5 of the species tree, a valid coalescent history might have clade 2 of the gene tree coalesce on branch 5 of the species tree. Therefore,  $m_{52} = 1$ . Filling in the matrix yields

$$\mathbf{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}. \quad (3)$$

Recalling that any clade can coalesce prior to the root (i.e., branch 6), this  $\mathbf{M}$  matrix specifies that clade 1 can only coalesce on branch 2 or 6; clades 2, 3, and 4 can only coalesce on branch 5 or 6; and clade 5 can only coalesce prior to the root.

A further restriction for a coalescent history  $\mathbf{h}$  to be valid is that if  $i$  and  $j$  are clades of the gene tree and  $j$  is an ancestor of  $i$ , then  $i$  must coalesce more recently than  $j$  or on the same branch of the species tree as  $j$ . Again, these restrictions can be represented as a matrix  $\mathbf{R} = (r_{ij})$  where  $r_{ij} = 1$  if and only if  $i$  is an ancestor of  $j$  on the gene tree; otherwise,  $r_{ij} = 0$ . A coalescent history  $\mathbf{h} = (h_1, \dots, h_i, \dots, h_j, \dots, h_{n-2})$ , with  $1 \leq i < j \leq n-2$ , is not permissible if  $h_j < h_i$  and  $j$

is ancestral to  $i$ . Note that numbering the branches by a post-order traversal is crucial here.

Similarly, let  $\mathbf{S} = (s_{ij})$  be the matrix for the species tree where  $s_{ij} = 1$  if node  $i$  is ancestral to node  $j$ ; otherwise,  $s_{ij} = 0$ . For the example above,

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad (4a)$$

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}. \quad (4b)$$

As an example of how the  $\mathbf{R}$  matrix is used, consider the proposed but invalid history  $\mathbf{h} = (2,5,6,5,6)$ . Although this history does not violate the restrictions of the  $\mathbf{M}$  matrix, it is not a valid history because on the gene tree, clade 4 is an ancestor of clade 3 ( $r_{43} = 1$ ), and therefore clade 3 of the gene tree must coalesce on the same branch or a descendent branch from where clade 4 coalesces on the species tree. Because  $\mathbf{h}$  has  $h_4 < h_3$ , the proposed history is invalid.

To determine the set of valid coalescent histories, all vectors  $\mathbf{h} = (h_1, h_2, \dots, h_{n-2})$  that satisfy  $0 \leq h \leq (n-1)^{n-2} - 1$  can be enumerated, and those that violate restrictions given by either the  $\mathbf{M}$  or  $\mathbf{R}$  matrices can be discarded. In summary,  $\mathbf{h} \in H_{\psi}(g)$  if and only if: (1) for all  $k = 1, \dots, n-2$ , if  $h_k \leq n-2$ , then  $m_{h_k k} = 1$ ; and (2) for all  $i, j \in \{1, 2, \dots, n-2\}$ , if  $h_j < h_i$ , then  $r_{ji} = 0$ . Computational savings similar to the case of the  $\mathbf{M}$  matrix can be achieved by skipping over proposed histories with the pattern  $\mathbf{h} = (h_1, \dots, h_i, \dots, h_j, \dots, h_{n-2})$  where  $h_j < h_i$  and  $r_{ji} = 1$ . For the seven-taxon example gene tree and species tree, the valid coalescent histories are listed in Table 1. (Note that by taking advantage of the postorder traversal labeling and the fact that these matrices are lower triangular, the  $\mathbf{R}$  and  $\mathbf{S}$  matrices can be represented as  $(n-2)$ -dimensional vectors,  $\mathbf{r}$  and  $\mathbf{s}$ , rather than  $(n-2) \times (n-2)$  matrices. Let  $r_k = 0$  if the  $k$ th row of the  $\mathbf{R}$  matrix has only zeros; otherwise, let  $r_k$  be the minimum value of  $j$  that satisfies  $r_{kj} = 1$ . The definition of  $s_k$  is analogous.)

#### THE PROBABILITY OF A COALESCENT HISTORY

Because a fixed, valid coalescent history  $\mathbf{h} = (h_1, \dots, h_{n-2})$  is assumed in this section, the notation that indicates dependence on  $\mathbf{h}$  can usually be dropped.

##### Calculating $p_{u_b v_b}(\lambda_b)$

To calculate  $p_{u_b v_b}(\lambda_b)$ , one can first determine the subscripts  $u_b$  and  $v_b$  and then apply equation (1). The subscript  $u_b$  is the number of lineages present at the beginning of branch  $b$  of the species tree, that is, the number of lineages on a branch before any coalescent events have occurred. This is equivalent to the number of taxa present in clade  $b$  of the species tree minus the number of coalescent events on branches de-

scended from branch  $b$  of the species tree. Using the postorder traversal labeling, node 1 necessarily has two descendants, and for  $b > 1$ , the number of taxa  $\delta_b$  in clade  $b$  of the species tree is

$$\delta_b = 2 + \sum_{y=1}^{b-1} s_{by}. \quad (5)$$

For  $b > 1$ ,

$$u_b = \delta_b - \sum_{y=1}^{b-1} \sum_{k=1}^{n-2} I(h_k = y) s_{by}, \quad (6)$$

where  $I(h_k = y) = 1$  if clade  $k$  of the gene tree coalesces on branch  $y$  of the species tree; otherwise,  $I(h_k = y) = 0$ . For  $b = 1$ ,  $u_b$  is necessarily 2.

The subscript  $v_b$  is the number of lineages still remaining on the branch after all coalescences have occurred, which is  $u_b$  minus the number of coalescent events on branch  $b$ :

$$v_b = u_b - \sum_{k=1}^{n-2} I(h_k = b). \quad (7)$$

Because the species tree branch lengths  $\lambda_b$  are known,  $p_{u_b v_b}(\lambda_b)$  can be determined from equations (1), (6), and (7) for any particular coalescent history,  $\mathbf{h}$ .

##### Calculating $w_b$ and $d_b$

If a branch has at least one coalescent event, then there is a possibility that the events on that branch are inconsistent with the gene tree topology. The coefficients  $w_b$  and  $d_b$  reflect the probability that the coalescences within a branch are consistent with the gene tree. Here  $w_b$  is the number of correct orderings for the coalescent events within a branch, and  $d_b$  is the number of possible orderings (including incorrect orderings). Assuming that all possible orderings of lineages are equally likely,  $w_b/d_b$  is the probability that the coalescent events on branch  $b$  are consistent with the gene tree.

If  $u_b = v_b$ , then there are no coalescences on branch  $b$ , and  $w_b = d_b = 1$ . Also, if  $u_b = 2$  and  $v_b = 1$ , then there is one coalescent event on that branch, but only one coalescent event is possible (the two lineages coalesced), so there is no possibility that the coalescent event occurs in the wrong order, and again  $w_b = d_b = 1$ .

When there are more than two lineages at the beginning of a branch and there is at least one coalescent event, so that  $u_b > 2$  and  $u_b > v_b$ , then the probability that the coalescences are consistent with the gene tree is strictly less than one. If  $u_b = v_b + 1$  (i.e., if there is exactly one coalescent event), then there is only one possible ordering, and again  $w_b = 1$ . However, if  $u_b > v_b + 1$ , there may be more than one way for the coalescences to occur. As an example, for the seven-taxon case given above and the history (2,5,5,6,6), because branch 5 on the species tree has lineages D, E, F, and G coalescing into two lineages, (DF) and (EG), either D and F coalesced first, meaning closer to the tips of the tree, or E and G coalesced first. Because there are two correct orderings for these two coalescent events, the numerator of the coefficient of  $p_{42}(\lambda_5)$  is  $w_5 = 2$ .

In general, if the number of coalescent events on branch  $b$  is  $c_b$  and if there are no restrictions on the ordering of the

coalescences, then  $w_b = c_b!$ . (Note that  $c_b$  is just  $u_b - v_b$ .) To check whether there are restrictions, one must check whether any of the clades coalescing is an ancestor of another clade. For example, ((DF)(EG)) is an ancestor of (EG).

These restrictions are provided by the  $\mathbf{R}$  matrix. Let  $b_1, b_2, \dots, b_{c_b}$  be the clades coalescing on branch  $b$  of the species tree. Recall that a clade  $b_j$  can coalesce before clade  $b_i$  only if  $r_{b_j b_i} = 0$ . If there have not been any coalescences on branches descended from branch  $b$  of the species tree, then the number of ways,  $w_b$ , for the coalescences on branch  $b$  to occur can be obtained directly from the appropriate rows of  $\mathbf{R}$ . In particular,

$$w_b = c_b! \prod_{y=1}^{c_b} \frac{1}{1 + \sum_{z=1}^{n-2} r_{b,yz}}. \quad (8)$$

This formula is explained in Appendix 1.

If there have been coalescences more recent than branch  $b$  on the species tree, however, then an updated version of the  $\mathbf{R}$  matrix is needed. Here the  $\mathbf{R}$  matrix is only used to count the number of restrictions for the ordering of coalescent events within a branch, so what is needed is a matrix  $\mathbf{R}^{(\mathbf{h})} = (r_{ij}^{(\mathbf{h})})$  for which  $r_{ij}^{(\mathbf{h})} = 1$  if branch  $i$  of the gene tree is ancestral to branch  $j$  of the gene tree, and clade  $j$  did not coalesce more recently than  $i$  on the species tree; otherwise,  $r_{ij}^{(\mathbf{h})} = 0$ . Using this updated matrix (which must be recomputed for each coalescent history), the above formula still holds:

$$w_b = c_b! \prod_{y=1}^{c_b} \frac{1}{1 + \sum_{z=1}^{n-2} r_{b,yz}^{(\mathbf{h})}}. \quad (9)$$

(Updating  $\mathbf{R}$  with  $\mathbf{R}^{(\mathbf{h})}$  solves the dependence problem between the branches alluded to earlier because the events on a particular branch  $b$  are independent of the events on other branches given the information contained in  $\mathbf{R}^{(\mathbf{h})}$ ,  $u_b$ , and  $v_b$  regarding events on branches descended from  $b$ .) Letting  $c_{n-1}$  be the number of coalescent events more ancient than the root and  $w_{n-1} := w$  be the number of correct orderings of coalescent events more ancient than the root, equation (9) still holds for  $b = n - 1$ .

The number of instantiations of a coalescent history is the product of the  $w_b$  terms taken over all the branches:

$$\prod_{b=1}^{n-1} w_b = \prod_{b=1}^{n-1} \prod_{y=1}^{c_b} \frac{1}{1 + \sum_{z=1}^{n-2} r_{b,yz}^{(\mathbf{h})}}. \quad (10)$$

Here products from one to zero are considered to be one. See Graham et al. (1994, pp. 62, 106, 501–502) for discussions about conventions regarding empty products and sums. Because the ordering of coalescent events is independent between branches, each instantiation of a coalescent history is equally likely.

The denominators,  $d_b$ , are more straightforward. If there is at least one coalescence on branch  $b$ , then consider any one of the orderings of coalescences counted by  $w_b$ . If there are  $c_b$  coalescent events on branch  $b$  and  $u_b$  lineages at the beginning of the branch, then there are  $\binom{u_b}{2}$  possible coalescences as the first coalescent event. In the seven-taxon tree

example, consider the case that branch 5 of the species tree has D and F coalescing into (DF) first, followed by E and G coalescing into (EG). Of the four lineages, D, E, F, and G, the probability that two chosen at random are D and F is  $1/\binom{4}{2}$ . Given that D and F have coalesced, there are three lineages, (DF), E, and G, that must coalesce into two. The probability that E and G coalesce before (DF) coalesces with either E or G is  $1/\binom{3}{2}$ . The denominator of the coefficient of  $p_{42}(\lambda_5)$  is therefore  $d_5 = \binom{4}{2} \times \binom{3}{2}$ .

If the coalescent events on a particular branch are labeled  $k_1, k_2, \dots, k_{c_b}$ , then the probability that  $k_1$  occurs first is  $1/\binom{u_b}{2}$ ; the probability that  $k_2$  happens second, given that  $k_1$  occurred first, is  $1/\binom{u_b-1}{2}$ ; and the probability that the coalescent event  $k_{c_b}$  happened last, given that events  $k_1, k_2, \dots, k_{c_b-1}$  have already occurred, is  $1/\binom{u_b-c_b+1}{2}$ . The probability that a set of coalescent events on a branch occur in a particular order is therefore (for  $c_b > 0$ ):

$$\frac{1}{d_b} = \prod_{y=0}^{c_b-1} \frac{1}{\binom{u_b - y}{2}}. \quad (11)$$

The quantity  $d_b$  is essentially the same as  $I_{n,k}$  in Rosenberg (2003).

For coalescent events prior to the root,  $1/d$  is the probability of any particular ordering. Let  $1/d_{n-1} := 1/d$ . Then letting  $u_{n-1}$  be the number of lineages prior to the root (before any coalescences prior to the root), equation (11) still holds with  $b = n - 1$ . The term  $u_{n-1}$  is  $1 + \sum_{k=1}^{n-2} I(h_k = n - 1)$ . In the example seven-taxon gene tree and species tree, the 10 valid coalescent histories are shown graphically on the species tree in Figure 3. Table 1 lists these histories and their probabilities. For trees with long branches, most of the probability of the gene tree is determined by histories that have coalescences as close to the tips as possible; trees with shorter branches have more of the probability on histories with deep coalescences. As an example, the first history listed in Table 1, which has the fewest deep coalescences, accounts for more than two-thirds of the probability when the species tree has branch lengths of 1.0, but less than one-half of the probability when the species tree has branch lengths of 0.5.

#### THE PROBABILITY MASS FUNCTION $P_{\psi,\lambda}(G = g)$

Substituting equations (6), (7), (9), and (11) into (2) yields the full formula for the probability of a gene tree for a given species tree. To simplify the formula, let  $p_{u_{n-1}(\mathbf{h})v_{n-1}(\mathbf{h})}(\lambda_{n-1}) := 1$ . (Because there is no branch ancestral to the root,  $\lambda_{n-1}$  is, strictly speaking, meaningless, or  $\lambda_{n-1}$  can be considered infinite.) Then

$$\begin{aligned} P_{\psi,\lambda}(G = g) &= \sum_{\mathbf{h} \in H_\psi(g)} \prod_{b=1}^{n-1} \frac{c_b(\mathbf{h})! \prod_{y=1}^{c_b(\mathbf{h})} \left[ 1 + \sum_{z=1}^{n-2} r_{b,yz}^{(\mathbf{h})} \right]^{-1}}{\prod_{y=0}^{c_b(\mathbf{h})-1} \binom{u_b(\mathbf{h}) - y}{2}} p_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b) \\ &= \sum_{\mathbf{h} \in H_\psi(g)} \prod_{b=1}^{n-1} \frac{P_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b)}{\prod_{y=1}^{c_b(\mathbf{h})} \left\{ \frac{1}{y} \left[ 1 + \sum_{z=1}^{n-2} r_{b,yz}^{(\mathbf{h})} \right] \binom{u_b(\mathbf{h}) - y + 1}{2} \right\}}. \end{aligned} \quad (12)$$

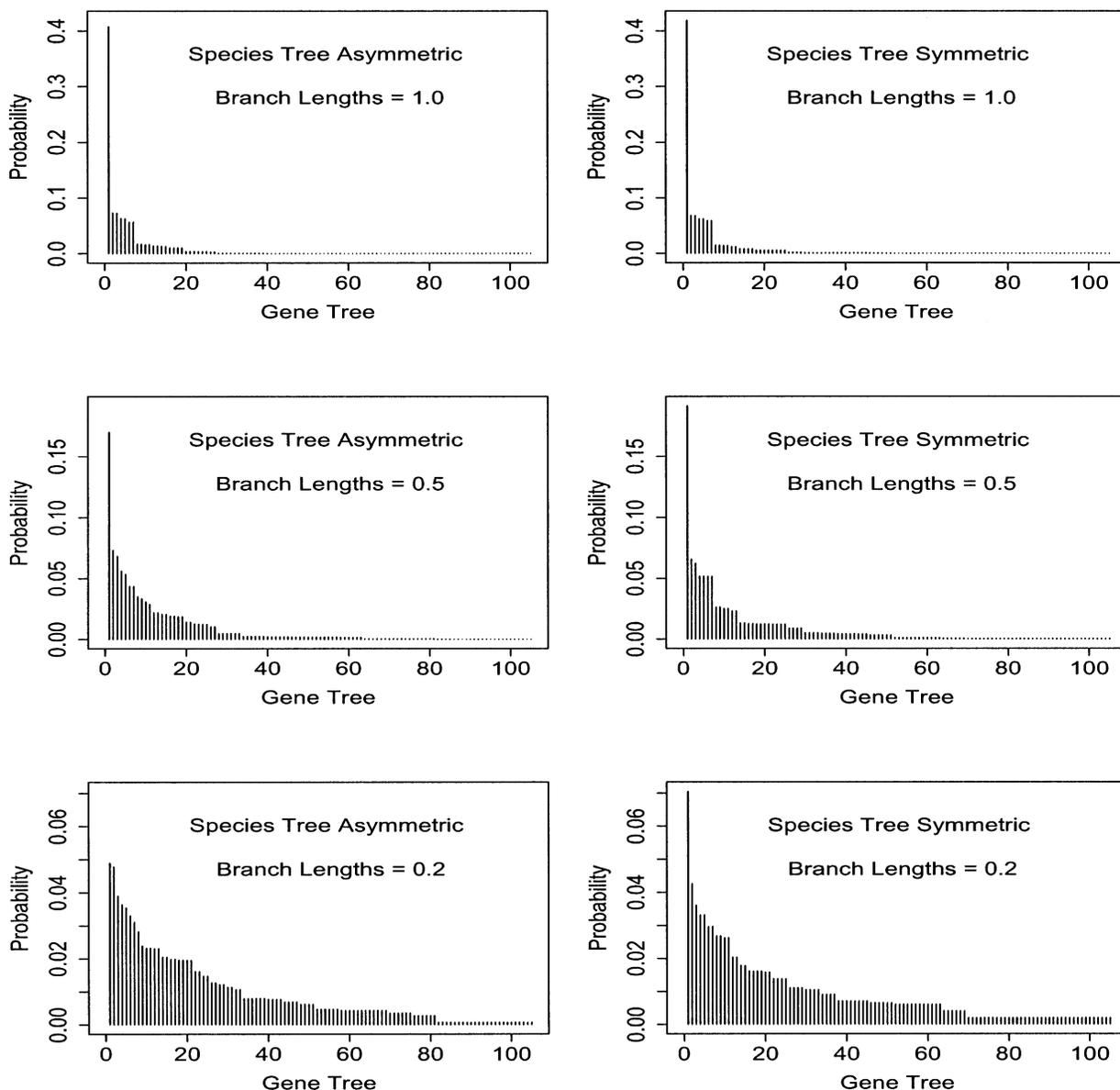


FIG. 4. Gene tree distributions for five-taxon species trees. Each plot shows the exact probability for all 105 gene trees in the distribution. For each plot, the gene trees are sorted by their probabilities. (The ordering of gene trees is not necessarily the same in the different plots.) The spikes in the distributions are for the gene trees with the same topology as the species tree.

Gene tree distributions for five-taxon trees (Figs. 4, 5) show the effects of branch lengths and symmetry in the species tree. A tree is said to be maximally symmetric when for each node  $b$  of the tree, if  $b$  has a total of  $2\delta_b$  descendants, then each of its two children has  $\delta_b$  descendants; and if  $b$  has  $2\delta_b + 1$  descendants, then one child has  $\delta_b$  descendants, and the other child has  $\delta_b + 1$  descendants. The seven-taxon species tree shown in Figure 1a is an example of a maximally symmetric tree. A tree is said to be maximally asymmetric if for each node of the tree, at least one of the children of that node is a tip.

Figure 4 depicts the effect of decreasing all branch lengths simultaneously on the tree. For all of these distributions, although the gene tree with the highest probability has the same topology as the species tree, this probability is below

0.5, and diminishes rapidly as branch lengths decrease. In general, longer branches result in distributions that are concentrated on a smaller set of trees. To illustrate this pattern, Table 2 shows the number of gene trees necessary to capture 90% of the gene tree distribution for different branch lengths and different numbers of taxa.

Figure 5 shows that having one short branch anywhere on the tree results in several trees that are nearly tied for the highest probability, regardless of the shape of the species tree. Gene trees with more symmetry can have higher probabilities because there are fewer restrictions on the orderings of coalescences. For Figures 5c and 5e, the gene tree with the same topology as the species tree has the second highest probability. The panels on the left side use  $((((AB)C)D)E)$  as the species tree, and for Figure 5c, the gene tree with this

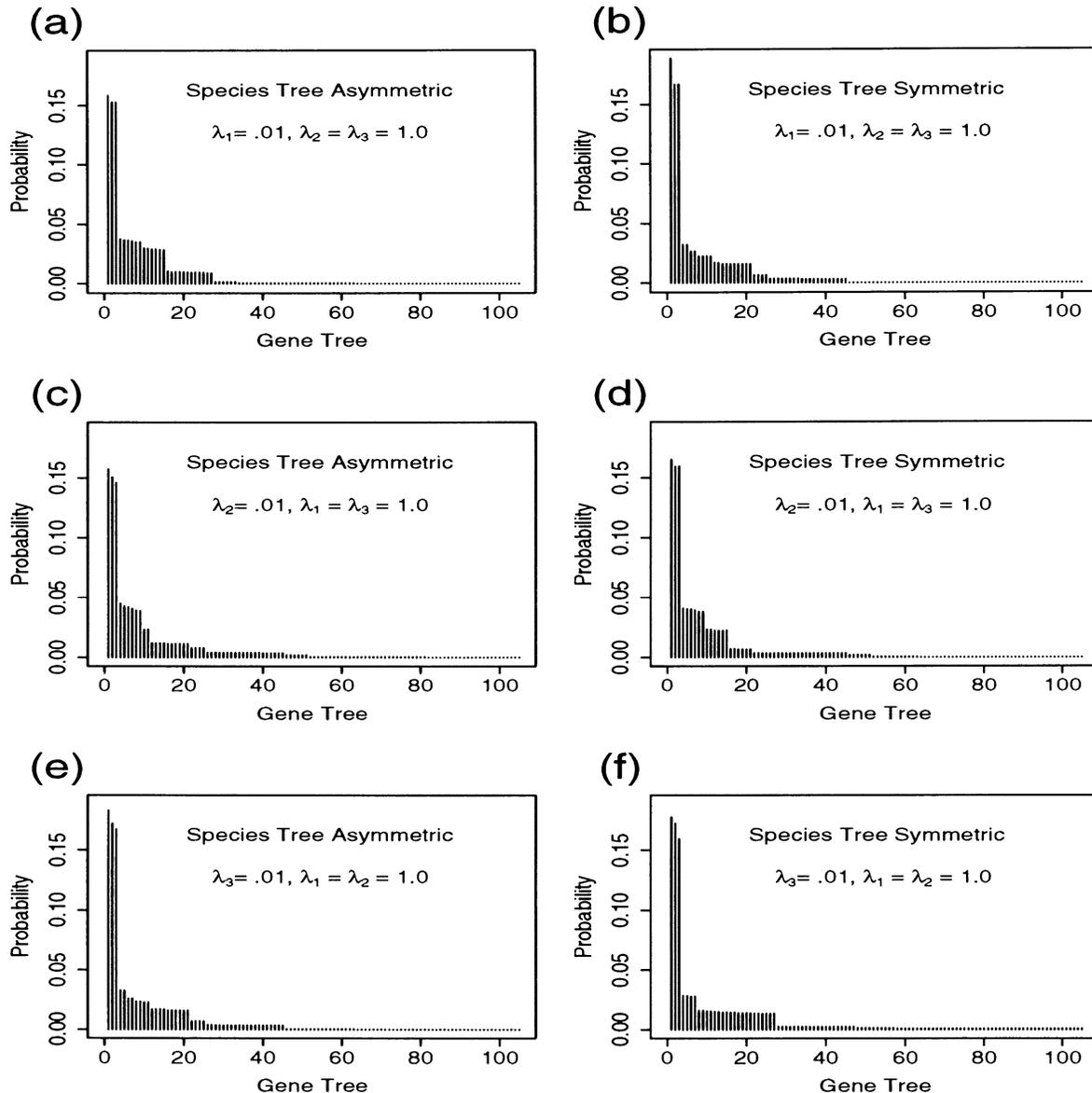


FIG. 5. Gene tree distributions for five-taxon trees with one short branch. As in Figure 4, gene trees are sorted by their probabilities. In (c) and (e), the gene tree with the species tree topology has the second highest probability.

topology has probability 0.1505, while the gene tree with topology  $((AB)(CD))E$  has probability 0.1572. Similarly, for Figure 5e, the gene tree with the species tree topology has probability 0.1723, while the gene tree with topology  $((AB)C)(DE)$  has probability 0.1837. For the other four panels in this figure, the highest probability gene tree is the one that matches the species tree.

#### THE NUMBER OF COALESCENT HISTORIES

The number of valid coalescent histories that have to be considered for a particular gene tree depends on the size of the trees, the degree of similarity between the species and gene trees, and tree shape (the degree of symmetry in the trees). Gene trees that are similar to species trees have the largest numbers of histories. Those that differ radically from the species tree have very few valid histories (often only one),

even for trees with a large number of taxa. Asymmetric species trees tend to generate a larger number of valid coalescent histories than symmetric species trees, even for gene trees that are not topologically equivalent to the species tree (Fig. 6). Table 3 shows the number of coalescent histories that must be evaluated to compute the probability that the gene tree is topologically equivalent to the species tree. This is given both for trees that are maximally asymmetric and for trees that are maximally symmetric.

When the gene and species trees have the same topology and the trees are maximally asymmetric, the number of histories can be shown to be the Catalan number  $C(n-1)$ , where  $n$  is the number of taxa, and  $C(n) = \binom{2n}{n}/(n+1) = (2n)!/[n!(n+1)!]$  (Graham et al. 1994). Using Stirling's approximation (Feller 1968), the number  $C(n)$  can be shown to be asymptotically equivalent to  $4^n/(\sqrt{\pi n^{3/2}})$ . Thus, the num-

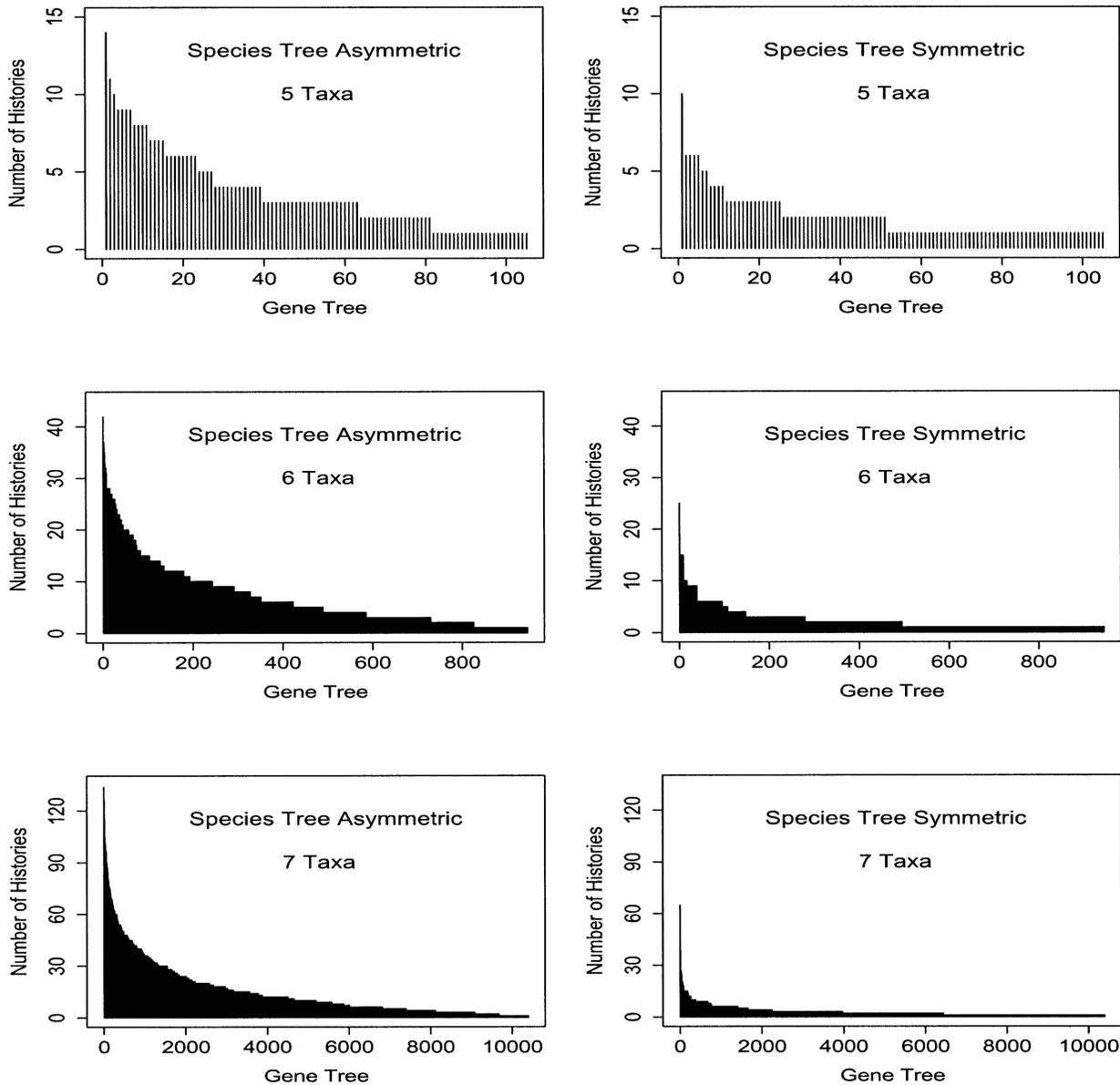


FIG. 6. The number of coalescent histories for each gene tree in the gene tree distribution. For each plot, the gene trees are sorted by the number of valid coalescent histories.

ber of histories grows more slowly in the number of taxa than the number of tree topologies,  $(2n - 3)!!$  (Felsenstein 2004). For an arbitrary sequence of tree topologies,  $\psi_3, \psi_4, \dots, \psi_n, \dots$ , where  $\psi_n$  has  $n$  taxa, larger trees do not necessarily have more coalescent histories than smaller trees. For example, the 10-taxon symmetric tree has fewer histories than the nine-taxon asymmetric tree (Table 3). However, if a sequence of tree topologies is generated by splitting tips of successive trees, then the number of histories at least doubles with each taxon added (Appendix 2). A consequence is that for maximally symmetric trees, a conservative lower bound for the number of histories, and therefore the number of terms needed to compute the exact probability, is  $2^{n-2}$ . This makes it difficult in practice to compute the probabilities of gene trees given species trees when there is a large number

of taxa, even with an algorithm available. A very conservative upper bound on the number of histories can be obtained by considering the number of proposed histories that do not violate the conditions of the  $\mathbf{M}$  matrix given above in the section on enumerating coalescent histories. For any particular gene and species tree (not necessarily with the same topology), this upper bound is

$$\prod_{y=1}^{n-2} \left( 1 + \sum_{j=1}^{n-2} m_{ij} \right). \quad (13)$$

For the seven-taxon example trees, equation (13) yields 16 as the upper bound for the number of histories.

TABLE 2. The minimum number of gene trees needed to capture 90% of the gene tree distribution as a function of the type of symmetry of the species tree (a, maximally asymmetric; s, maximally symmetric), the number of taxa ( $n$ ), and branch lengths. In the first three branch length columns, all branches have the indicated length. The fourth and fifth columns have all branches with length 1.0 except the indicated branch. Note that the minimum number of gene trees listed grows more slowly than the number of tree topologies based on the number of taxa (see Table 3).

Symmetry	$n$	Branch lengths				
		1.0	0.5	0.2	$\lambda_1 = 0.01$	$\lambda_{n-2} = 0.01$
a	4	4	7	10	7	9
a	5	13	27	58	19	21
a	6	33	118	345	51	61
a	7	96	512	2239	140	155
s	4	4	10	12	10	10
s	5	15	35	62	21	26
s	6	38	144	441	63	87
s	7	140	869	3452	207	363

APPLICATIONS

*Probability of Topological Equivalence of Gene Trees and Species Trees*

Because the complete distribution of gene trees for a given species tree is available, the probability that the gene tree has the same topology as the species tree can be computed directly. Figure 7 shows the probability that the gene tree is topologically equivalent to the species tree when branch lengths vary continuously from 0.01 to 5.00 (assuming all branches have the same length) for different numbers of taxa. This figure can also be used to determine the branch lengths that would be necessary to have any desired probability that the gene tree and species tree are topologically equivalent. Note that even for moderately long branches, the probability of topological equivalence quickly decreases with the number of taxa.

Pamilo and Nei (1988) give a conservative upper bound for this probability,

$$P_A = \prod_{i=1}^{n-2} \left(1 - \frac{2}{3}e^{-\lambda_i}\right). \tag{14}$$

From equation (12), the probability of any three-taxon gene tree matching its species tree is  $1 - \frac{2}{3}e^{-\lambda_i}$ , and the bound is based on decomposing an  $n$ -taxon species tree into  $n - 2$  three-taxon trees, one for each internal branch, and treating these trees as independent. Here each three-taxon tree consists of an internal branch, its two descendent branches, and its sister branch. For example, in the seven-taxon tree example, the three-taxon tree corresponding to branch 5 has the branches 2, 3, and 4, and could be represented as (2,(3,4)).

The closeness of this bound to the exact probability can be evaluated for different tree shapes and sizes as well as branch lengths using equation (12). Because the assumption of independence is more nearly met, as Pamilo and Nei (1988) note, when the branch lengths are larger, the bound is tighter for trees with longer branches. The bound is also tighter for trees that are more nearly symmetric (Fig. 8), because for asymmetric trees lineages are more constrained in their order of coalescence and are therefore less independent. Although

TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

Taxa	Number of histories		Number of topologies
	Asymmetric trees	Symmetric trees	
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	$6.190 \times 10^{15}$
20	1,767,263,190	100,360,324	$8.201 \times 10^{21}$

the bound is fairly close when the branch lengths are moderately large, as the number of taxa increases and branch lengths are held constant, the ratio of the bound to the exact probability increases (Fig. 8). This indicates that the bound is not asymptotically approaching the exact probability.

Notice that  $P_{\psi,\lambda}(G = \psi)$  and  $P_A$  only refer to the probability that a random gene tree has the same topology as the fixed species tree. For a given observed gene tree, the coalescent model does not provide a method for determining the probability that the species tree has the same topology as the gene tree. Because the coalescent model treats the species tree as a parameter, one could adopt a Bayesian point of view to assign probabilities to species trees given gene trees. This would require assigning a prior distribution on the space of species trees, where the space would include branch lengths as well as topologies.

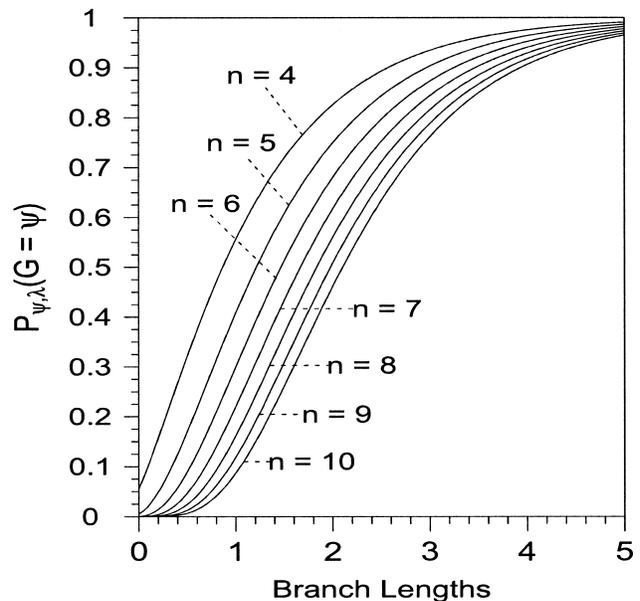


FIG. 7. The exact probability of topological equivalence between species and gene trees as a function of branch lengths and number of taxa. Probabilities were computed for branch lengths between 0.01 and 5.00 in increments of 0.01. Only asymmetric trees were used for this example. Symmetric trees show a very similar pattern (results not shown).

*Inferring the Species Tree*

Another application for the method derived here, suggested by Maddison (1997), is inferring the species tree given data for multiple genes in a maximum likelihood framework. Let  $\sigma = (\psi, \lambda)$  be the species tree with topology  $\psi$  and vector of branch lengths  $\lambda$ , and let  $\text{GTD}(\sigma)$  denote the gene tree distribution with parameter  $\sigma$ . Then consider a random sample of  $q$  gene trees,  $G_1, G_2, \dots, G_q$ , which are independent and identically distributed (i.i.d.) with distribution  $\text{GTD}(\sigma)$  and with associated datasets  $D_1, D_2, \dots, D_q$ , each of which is an alignment of  $n$  sequences from the same set of taxa. If the  $i$ th gene tree topology has vector of branch lengths  $\mathbf{t}_i$ , the density value for the branch lengths given the gene tree topology and ancestral population sizes  $\Theta$ ,  $f_{\Theta}(\mathbf{t}_i | G_i)$ , can be obtained from Yang (2002). Assuming that the datasets are independent, the likelihood of a species tree given the data is

$$l = l(\sigma | D_1, \dots, D_q) \quad (15a)$$

$$= P_{\sigma}(D_1, \dots, D_q) \quad (15b)$$

$$= \prod_{x=1}^q P_{\sigma}(D_x) \quad (15c)$$

$$= \prod_{x=1}^q \sum_{g_x, \mathbf{t}_x} P_{\sigma}(D_x | G_x = g_x, \mathbf{t}_x) P_{\sigma}(G_x = g_x, \mathbf{t}_x) \quad (15d)$$

$$= \prod_{x=1}^q \sum_{g_x, \mathbf{t}_x} P(D_x | G_x = g_x, \mathbf{t}_x) P_{\sigma}(G_x = g_x, \mathbf{t}_x) \quad (15e)$$

$$= \prod_{x=1}^q \sum_{g_x, \mathbf{t}_x} l(g_x, \mathbf{t}_x | D_x) f_{\Theta}(\mathbf{t}_x | G_x) P_{\sigma}(G_x = g_x). \quad (15f)$$

Thus

$$l = \prod_{x=1}^q \sum_{g_x, \mathbf{t}_x} l(g_x, \mathbf{t}_x | D_x) f_{\Theta}(\mathbf{t}_x | G_x) \times \sum_{\mathbf{h} \in H_{\psi}(g_x)} \prod_{b=1}^{n-1} \frac{P_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b)}{\prod_{y=1}^{c_b(\mathbf{h})} \left\{ \frac{1}{y} \left[ 1 + \sum_{z=1}^{n-2} r_{b_y(\mathbf{h})z}^{(\mathbf{h})} \right] \binom{u_b(\mathbf{h}) - y + 1}{2} \right\}}, \quad (15g)$$

where  $l(g_x, \mathbf{t}_x | D_x)$  is the likelihood of the  $x$ th gene tree given the  $x$ th dataset. This is the standard likelihood used in phylogenetic inference and depends on the model of evolution as well as the branch lengths of the gene trees. Maximizing the likelihood requires maximization over the space of gene trees (and perhaps model parameters), including all possible sets of branch lengths. In this case, the sum  $\sum_{g_x, \mathbf{t}_x}$  should be interpreted as an integral over this space. Note that the step from (15d) to (15e) assumes that the distribution of the  $x$ th dataset depends on the species tree only through the  $x$ th gene tree. Rannala and Yang (2003) derive the joint density for the branch lengths and topology of the gene tree, and use this to implement a Bayesian MCMC algorithm to find posterior densities for species divergence times and ancestral population sizes.

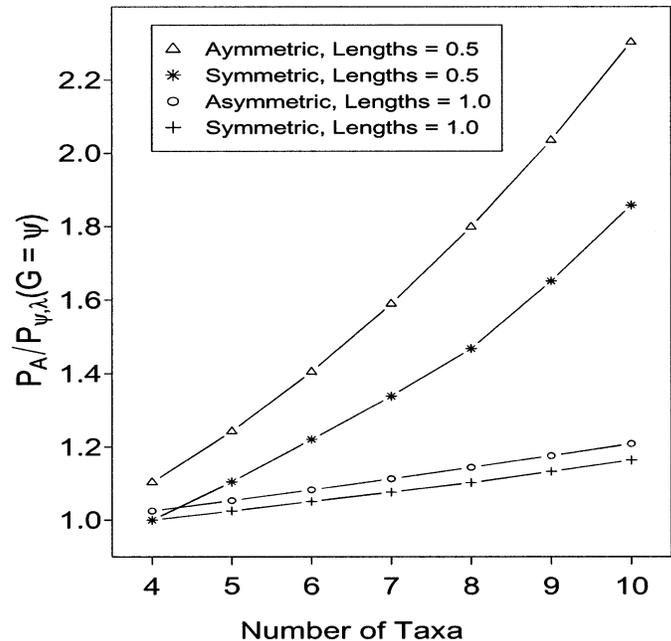


FIG. 8. The ratio of the upper bound given by equation (14) to the exact probability.

DISCUSSION

*Generalizations*

The method presented here can be extended in several ways to allow for more complex models. For example, the model used in this paper assumes population sizes within branches are constant. A more complicated model could be obtained by allowing the effective population size to vary for different branches or by allowing the effective population size to be a function of time (Kingman 1982; Felsenstein 2004). This would modify the  $p_{uv}(T)$  terms from equation (1) without changing the general algorithm.

Another assumption of the method used for this paper is that on a given branch, if there are more than two lineages present, every possible pair of lineages has an equal probability of coalescing first on that branch. This corresponds to the Yule model (Yule 1924; Aldous 2001; Rosenberg 2003), and could be modified so that the probability that two lineages coalesce in a given amount of time is a function of the similarity of the two lineages (e.g., according to an evolutionary distance measure). This would require modifying the  $w_b(\mathbf{h})/d_b(\mathbf{h})$  and  $w(\mathbf{h})/d(\mathbf{h})$  terms in equation (2), but would not alter the enumeration of histories or the calculation of the  $p_{uv}(T)$  terms.

*Tree Shape and Branch Lengths*

Tree shape (of both the gene tree and species tree) and branch lengths (of the species tree) both clearly have an effect on gene tree distributions. Effects of branch lengths on the probability of topological equivalence have been discussed earlier (Tajima 1983; Takahata and Nei 1985; Pamilo and Nei 1988), but the tree shape, in particular the degree of symmetry in both the species tree and gene tree, is also important, especially if the species tree has short branches. Ro-

senberg (2002) notes that the probability that the gene and species trees have the same topology is higher for symmetric trees. Gene tree distributions are generally flatter when branch lengths of the species tree are small, but this effect is more pronounced for asymmetric than for symmetric trees (Fig. 4).

Gene trees with a high degree of symmetry have fewer restrictions on the order of coalescent events on branches deep within the species tree. A counterintuitive consequence is that when the Yule model is used, the gene tree with the highest probability does not necessarily have the same topology as the species tree. This can occur when the gene tree has more symmetry than the species tree and when there are short branches. This phenomenon is illustrated in Figure 5 as a result of one extremely short branch, but it can also occur if all branches are sufficiently and uniformly short. For example, if the species tree has topology  $\psi = (((AB)C)D)$  and all branch lengths are 0.1, then  $P_{\psi,\lambda}(G = \psi) = 0.1037$ , but  $P_{\psi,\lambda}(G = ((AB)(CD))) = 0.1279$ . Because long branches correspond to small effective population sizes, the probability that the gene and species trees have the same probability increases dramatically when population sizes are small.

#### *Bayesian Extensions*

Although gene trees as well as species trees are generally unknown and need to be estimated, gene trees are usually easier to estimate than species trees. When a species tree is desired but only a gene tree is available, a natural question to ask is whether the species tree has a high probability of having the same topology as the gene tree. Similarly, if several gene trees are available, but they conflict, then it is natural to ask about the distribution of possible species trees. As mentioned above, however, under the coalescent model the species tree is treated as a parameter and not a random variable. There is a strange asymmetry in thinking of gene trees as random and species trees as fixed. One could argue that evolution is a stochastic process that generates both species trees and gene trees and that both kinds of trees should therefore be considered random.

Questions regarding the distribution of the species tree can be answered in a Bayesian framework where a prior distribution is specified that includes information regarding both the species tree topologies and their branch lengths. If one is willing to assume a prior distribution for species trees, other interesting questions can be answered. For example, one could obtain the unconditional probability of a gene tree by averaging (integrating) over the space of species trees. Similarly, this approach could be used to find expected differences between two random gene trees. In addition, species trees could be inferred using Bayesian methods instead of maximum likelihood by finding the species tree with the highest posterior probability. More importantly, because this approach would result in a posterior distribution on the set of possible species trees, it could be used to find a set of trees that captures most of the posterior distribution.

#### *Computational Considerations*

The number of coalescent histories is also the number of terms in the outer sum in equation (12). The amount of time

needed to compute gene tree probabilities therefore depends on this number, which is a function not only of the number of taxa but also of the amount of topological agreement between the gene and species trees, and the degree of symmetry in the trees. For the case of the gene and species trees having the same topology, the computation of the gene tree probabilities for the 16-, 18-, and 20-taxon maximally asymmetric trees took approximately 2.1, 34.9, and 581.8 min, respectively (as implemented in COAL running on LINUX on a Dell [Round Rock, TX] Precision Workstation 530 with a 1.7 GHz Xeon processor [Intel, Santa Clara, CA]). For maximally symmetric trees, the same numbers of taxa took 0.1, 1.5, and 18.6 min, respectively. For trees with much more than 20 taxa, and especially for applications that require evaluating many trees such as species tree inference, methods can be used to approximate (12), such as importance sampling on the set of coalescent histories (Robert and Casella 1999). Note also that the vector representation of coalescent histories allows the set of histories to be partitioned so that parallel computing techniques can be used to evaluate equation (12).

#### *Conclusions*

This article presents the gene tree distribution for any fixed, bifurcating species tree under the coalescent process as well as a method for computing this distribution directly. The ability to compute gene tree distributions for larger trees than previously possible will allow the natural variability of gene trees to be better understood in the absence of nonneutral evolution, horizontal transfer, hybridization, gene duplication, and other evolutionary forces. The automated computation of gene tree distributions also makes possible applications such as inferring species trees using either maximum likelihood or Bayesian approaches.

Although the method presented here allows only one sampled gene per species, this should be seen as a step toward the more general problem of determining the probability of a gene tree when intraspecific sampling is also considered.

All calculations in this paper were made using the program COAL, which is available at <http://www.coaltree.net> or by request from the authors.

#### ACKNOWLEDGMENTS

We thank J. Hey, H. Nelson, M. S. Petronis, T. F. Turner, and two anonymous reviewers for discussion and comments on an earlier draft. The idea for Figure 7 was suggested by T. F. Turner. JHD and LAS were supported by National Science Foundation grant DMS 0104290 to LAS.

#### LITERATURE CITED

- Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.* 16:23–34.
- Feller, W. 1968. An introduction to probability theory and its application. Vol. I. 3rd ed. Wiley, New York.
- Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.
- Graham, R. L., D. E. Knuth, and O. Patshnik. 1994. *Concrete mathematics*. 2nd ed. Addison-Wesley, Boston, MA.
- Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.* 36:396–405.
- Kingman, J. F. C. 1982. Exchangeability and the evolution of large

- populations. Pp. 97–112 in G. Koch and F. Spizzichino, eds. Exchangeability in probability and statistics. North-Holland, Amsterdam.
- Knowles, L. L., and W. P. Maddison. 2002. Statistical phylogeography. *Mol. Ecol.* 11:2623–2635.
- Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5(5):568–583.
- Poe, S., and A. L. Chubb. 2004. Birds in a bush: five genes indicate explosive radiation of avian orders. *Evolution* 58:404–415.
- Rannala, B., and Z. Yang. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 164:1645–1656.
- Robert, C., and G. Casella. 1999. Monte Carlo statistical methods. Springer, New York.
- Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
- Rosen, K. H. 1999. Discrete mathematics and its applications. 4th ed. WCB/McGraw-Hill, Boston, MA.
- Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* 61:225–247.
- . 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57:1465–1477.
- Sang, T., and Y. Zhong. 2000. Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.* 49:422–434.
- Syvanen, M. 1994. Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.* 28:237–261.
- Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
- Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–966.
- Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344.
- Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26:119–164.
- Watterson, G. A. 1984. Lines of descent and the coalescent. *Theor. Popul. Biol.* 26:77–92.
- Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127:429–435.
- Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–1823.
- Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F. R. S. *Philos. Trans. R. Soc. Lond. B* 213:21–87.

Corresponding Editor: J. Hey

#### APPENDIX 1

The term  $w_b$  is the number of ways of arranging  $c$  objects,  $b_1, b_2, \dots, b_c$ , where there are restrictions on the ordering of the  $c$  objects. These restrictions can be described by noting, for each object  $b_j$ , the objects that must precede  $b_j$ . In general, if  $b_j$  is constrained to appear after  $m$  objects, the number of arrangements of the  $c$  objects satisfying this constraint is  $c!/(1 + m)$ . (This can be seen by symmetry; given  $1 + m$  objects, there is the same number of sequences with any of the  $1 + m$  objects occurring last.) The constraint that  $b_j$  appears after the objects  $b_{j_1}, \dots, b_{j_m}$  is independent of the possible constraints of the objects  $b_{j_1}, \dots, b_{j_m}$ .

In general, if  $\mathbf{a} = (a_1, \dots, a_c)$ , where  $a_j$  is the number of objects that must precede the  $j$ th object, then the restrictions of the  $j$ th object reduce the number of arrangements by a factor of  $1 + a_j$ . The total number of arrangements is therefore

$$c! \prod_{j=1}^c \frac{1}{1 + a_j}. \quad (\text{A1})$$

Note that equation (A1) is the same as (8), which is used for computing  $w_b$ , where  $a_j$  is the sum of the  $b_j$ th row of the  $\mathbf{R}$  matrix, and  $b_j$  is the  $j$ th clade that coalesces on the branch.

#### APPENDIX 2

Consider a sequence of tree topologies  $\psi_3, \psi_4, \dots, \psi_n, \dots$ , where  $\psi_n$  has  $n$  taxa labeled  $A_1, A_2, \dots, A_n$ , and the  $(n + 1)$ st topology is obtained from the  $n$ th topology by replacing tip  $A_i$  with clade  $(A_i A'_i)$  for some  $i \leq n$ . Let the clades of  $\psi_n$  be labeled  $1, \dots, n - 2$ . Let the corresponding clades of  $\psi_{n+1}$  have the same labels as  $\psi_n$ , and let  $(A_i A'_i)$  be labeled zero. Let coalescent histories for the  $(n + 1)$ -taxon tree be represented by  $y = (y_0, y_1, \dots, y_{n-2})$ ; that is, let the indexing start at zero instead of one. We still have the interpretation that  $y_k = b$  means that clade  $k$  coalesces on branch  $b$ , but now we allow  $k = 0$  and  $b = 0$ . If the node immediately ancestral to  $A_i$  on the  $n$ -taxon tree is  $b_i$ , then for any history  $(h_1, h_2, \dots, h_{n-2})$  of the  $n$ -taxon tree,  $(0, h_1, h_2, \dots, h_{n-2})$  and  $(b_i, h_1, h_2, \dots, h_{n-2})$  are valid histories on the  $(n + 1)$ -taxon tree. Therefore, any such sequence of topologies has the number of histories at least doubling for each taxon added.