Advances in Applied Mathematics 102 (2019) 1–17

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Enumeration of lonely pairs of gene trees and species trees by means of antipodal cherries

Noah A. Rosenberg

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

ARTICLE INFO

Article history:
Received 28 February 2018
Received in revised form 1 September 2018
Accepted 4 September 2018
Available online xxxx

MSC:
05A15
05C05
92D15

Keywords:
Cherries
Coalescent histories
Phylogenetics

ABSTRACT

In mathematical phylogenetics, given a rooted binary leaf-labeled gene tree topology G and a rooted binary leaf-labeled species tree topology S with the same leaf labels, a coalescent history represents a possible mapping of the list of gene tree coalescences to the associated branches of the species tree on which those coalescences take place. For certain families of ordered pairs (G, S), the number of coalescent histories increases exponentially or even faster than exponentially with the number of leaves n. Other pairs have only a single coalescent history. We term a pair (G, S) *lonely* if it has only one coalescent history. Here, we characterize the set of all lonely pairs (G, S). Further, we characterize the set of pairs of rooted binary unlabeled tree shapes at least one of the labelings of which is lonely. We provide formulas for counting lonely pairs and pairs of unlabeled tree shapes with at least one lonely labeling. The lonely pairs provide a set of examples of pairs (G, S) for which the number of compact coalescent histories—which condense coalescent histories into a set of equivalence classes—is equal to the number of coalescent histories. Application of the condition that characterizes lonely pairs can also be used to reduce computation time for the enumeration of coalescent histories.

E-mail address: noahr@stanford.edu.

https://doi.org/10.1016/j.aam.2018.09.001

1. Introduction

The study of evolutionary trees makes a distinction between *species trees*, trees that describe relationships among sets of species, and *gene trees*, trees that describe relationships among genetic lineages of members of those species [3,11,12]. In considering the evolution of gene trees in relation to species trees, a variety of new types of combinatorial structures have emerged, each specifying some feature of the relationship between the branching patterns of gene trees and those of species trees [2,4,5,11,17,19,21,22].

Coalescent histories are prominent among the structures useful in the study of gene trees and species trees. For a given gene tree and species tree, a coalescent history describes an evolutionary scenario for gene lineage evolution on the branches of the species tree. More formally, viewing a rooted binary tree "backward in time" from the leaves to the root, each internal node of the tree, including the root node, represents a *coalescence*: an instance at which lineages represented by a set of leaves find common ancestors. We term a node or edge v of a tree an *ancestor* of a node or edge u if u lies on a path from v to a leaf; u is said to be a *descendant* of v. Trivially, v is an ancestor or descendant of itself. We then have the following definition.

Definition 1. Consider a rooted binary leaf-labeled tree G (the "gene tree topology") and a rooted binary leaf-labeled tree S (the "species tree topology"), labeled by the same set of mutually distinct leaf labels. A *coalescent history* f associates with each coalescence v in G an edge $f(v)$ in S, such that two properties are satisfied.

(i) For each gene tree coalescence v in G, the species tree edge $f(v)$ in S is ancestral to each leaf node of S that shares a label with a leaf that descends from v.

(ii) For each pair of gene tree coalescences u, v with the property that v is ancestral to u in G, $f(v)$ is ancestral to $f(u)$ in S.

Treating a gene tree as evolving on the branches of a species tree, coalescent histories describe permissible lists of edges of the species tree—including as a possibility an edge ancestral to its root—where the coalescences of the gene tree can take place [5,14]. From a biological perspective, the pair of constraints in the definition encodes the rules that (i) a set of gene lineages can coalesce only in a branch of the species tree that is possible for them all to reach, and (ii) ancestors can coalesce no more recently than their descendants (Fig. 1).

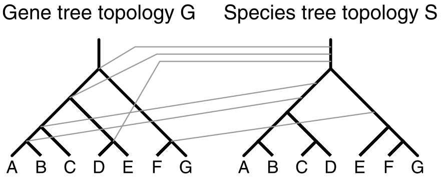
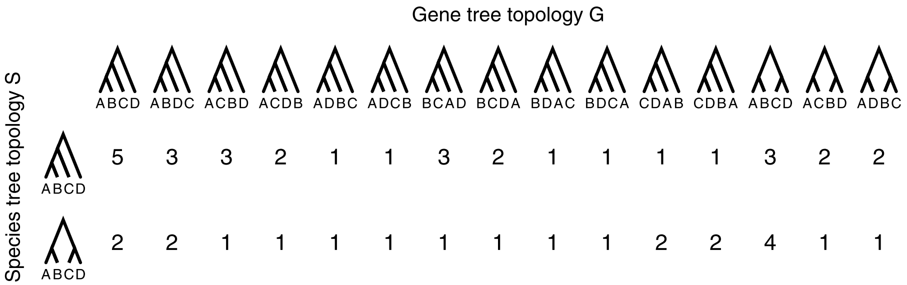


Fig. 1. A coalescent history for a labeled gene tree topology G and labeled species tree topology S. Gray lines represent the mapping of coalescences of the gene tree to edges of the species tree.



Species tree topology S	ABCD	ABDC	ACBD	ACDB	ADBC	ADCB	BCAD	BCDA	BDAC	BDCA	CDAB	CDBA	ABCD	ACBD	ADBC
ABCD	5	3	3	2	1	1	3	2	1	1	1	1	3	2	2
ABCD	2	2	1	1	1	1	1	1	1	1	2	2	4	1	1

Fig. 2. The number of coalescent histories for pairs (G, S) consisting of a 4-taxon labeled gene tree topology G and labeled species tree topology S. All 15 labeled gene tree topologies appear for representative labelings of the two unlabeled species tree shapes. Pairs with 1 coalescent history are the lonely pairs. The values are obtained by counting coalescent histories in Tables 4 and 5 of [13].

Given G and S, different coalescent histories f_1, f_2 are distinguished by having one or more coalescences v in G for which $f_1(v) \neq f_2(v)$ in S. It is natural to enumerate the coalescent histories associated with a pair (G, S), and several studies have contributed to enumerative investigations of coalescent histories [1,5,7,8,14–16,20].

Most enumerative results to date have focused on cases with many coalescent histories. Particularly for matching gene trees and species trees—that is, when G and S have the same labeled topology—gene lineages have multiple ways of traveling through the species tree to reach a common ancestor. The first studies exhibited tree families with $G = S$ for which the number of coalescent histories has exponential growth [5,14]. For matching caterpillar gene tree and species tree labeled topologies with n leaves, coalescent histories can be identified with monotonic paths that do not cross the diagonal of an $(n-1) \times (n-1)$ square lattice [1]. Such paths are enumerated by the Catalan numbers $C_{n-1} = \binom{2n-2}{n-1}/n$, which by Stirling's approximation have asymptotic growth $C_n \sim 4^n/(n^{3/2}\sqrt{\pi})$. Subsequent work has demonstrated exponential growth for other "caterpillar-like" families [8,15], and super-exponential growth in one case, that of the "lodgepole" trees [7].

For non-matching gene tree and species tree labeled topologies (G, S), with $G \neq S$, the number of coalescent histories can also be large [16,20]. For example, when S is a caterpillar labeled topology with $n \geqslant 7$ leaves, a labeled topology $G \neq S$ exists for which the pair (G, S) has more coalescent histories than (S, S) [16]. However, for fixed small species tree labeled topologies S, a salient feature of the distribution of the number of coalescent histories across labeled gene tree topologies G is the many gene trees that produce only 1 coalescent history [16, Table 1]. For a 4-taxon caterpillar species tree, 6 of 15 gene tree topologies produce only 1 coalescent history, and for a 4-taxon balanced species tree, 10 of 15 have this property (Fig. 2).

Given a rooted binary leaf-labeled gene tree topology G and species tree topology S with the same label set, we term the pair (G, S) *lonely* if and only if the number of coalescent histories for the pair is exactly 1. Here, we characterize the set of lonely pairs of labeled topologies with n leaves, exhibiting a formula for their enumeration. We also

Table 1
The number of labeled topologies with $n \geqslant 2$ leaves divided at the root into subtrees of size p and $n - p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, for small n.

	p					
n	1	2	3	4	5	Total
2	1					1
3	3					3
4	12	3				15
5	75	30				105
6	630	225	90			945
7	6615	2205	1575			10395
8	83160	26460	17640	7875		135135
9	1216215	374220	238140	198450		2027025
10	20270250	6081075	3742200	2976750	1389150	34459425

The table is computed using values of $L_{n,p}$ from eq. (2). The column on the right gives T_n from eq. (1).

characterize the set of pairs of unlabeled tree shapes (g, s) at least one of whose labelings gives a lonely pair.

2. Preliminaries

We consider all trees to be rooted and binary. Trees are leaf-labeled, except where specified. For convenience, a "tree" refers to a rooted binary leaf-labeled topology, except where specified. All trees with n leaves, or taxa, are assumed to have n distinct labels taken bijectively from a shared set of n labels. In particular, for leaf-labeled n-leaf rooted binary trees G and S representing a gene tree and species tree, respectively, this assumption corresponds to an assumption that for a fixed species tree S, we examine gene trees G that consider one gene lineage per species.

The number of labeled topologies with $n \geqslant 2$ leaves is [18, eq. 2.2]

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

The number of labeled topologies with $n \geqslant 2$ leaves that are divided at the root into subtrees of size p and $n - p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, is

$$L_{n,p} = \frac{\binom{n}{p} T_p T_{n-p}}{2^{\delta_{p,n-p}}}, \tag{2}$$

where δ is the Kronecker delta [10, section 2.1]. For small n, $L_{n,p}$ appears in Table 1.

The number of unlabeled tree shapes with $n \geqslant 2$ leaves is given by the recursion

$$t_n = \begin{cases} \sum_{p=1}^{(n-1)/2} t_p t_{n-p}, & n \text{ odd} \\ \frac{t_{n/2}(t_{n/2}+1)}{2} + \sum_{p=1}^{n/2-1} t_p t_{n-p}, & n \text{ even}, \end{cases} \tag{3}$$

Table 2
The number of unlabeled tree shapes with $n \geqslant 2$ leaves divided at the root into subtrees of size p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, for small n.

	p					
n	1	2	3	4	5	Total
2	1					1
3	1					1
4	1	1				2
5	2	1				3
6	3	2	1			6
7	6	3	2			11
8	11	6	3	3		23
9	23	11	6	6		46
10	46	23	11	12	6	98

The table is computed using values of $\ell_{n,p}$ from eq. (4). The column on the right gives t_n from eq. (3).

starting with $t_1 = 1$ [10, section 2.2]. The number of unlabeled tree shapes that have $n \geqslant 2$ leaves and that are divided at the root into subtrees of size p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, is

$$\ell_{n,p} = t_p t_{n-p} - \delta_{p,n-p} \binom{t_p}{2}. \tag{4}$$

This result follows by extracting the term of the recursive eq. 3 corresponding to the decomposition of unlabeled tree shapes at the root into subtrees of size p and $n-p$ leaves, using the Kronecker delta to combine the cases of n odd and n even into a single equation. For small n, $\ell_{n,p}$ appears in Table 2.

3. Enumerative results

Recall that we are in the setting in which G and S represent rooted binary leaf-labeled trees of size $n \geqslant 2$ leaves, each considering the same bijectively associated leaf set. We consider ordered pairs (G, S) and associated ordered pairs of unlabeled tree shapes (g, s). For convenience, we identify leaves with their labels. We also denote the two subtrees immediately descended from the root of S by S_L and S_R respectively, without loss of generality considering the "right" subtree S_R to have at least as many leaves as the "left" subtree S_L. We label the edge ancestral to the root node of S by $e_r(S)$.

3.1. Antipodal cherries

A *cherry node* of a rooted binary tree, labeled or unlabeled, or a *cherry* for short, is an internal node with exactly two descendant leaves. A key concept needed for characterizing lonely pairs (G, S) of size $n \geqslant 2$ leaves is the idea of an antipodal cherry (Fig. 3).

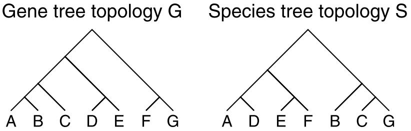


Fig. 3. Antipodal cherries. For labeled gene tree topology G, cherries (A, B) and (F, G) are antipodal cherries with respect to labeled species tree topology S. Cherry (D, E) of G is not antipodal with respect to S.

Definition 2. An *antipodal cherry* for a labeled topology G with respect to another labeled topology S is a cherry v of G in which the subtrees S_L and S_R immediately descended from the root of S each contain exactly one leaf descended from v in G.

Theorem 3. *A pair of labeled topologies* (G, S) *is lonely if and only if each cherry of* G *is antipodal with respect to* S.

Proof. Suppose each cherry v of G is antipodal with respect to S. Each internal node of G is either a cherry node or a non-cherry node. By part (i) of the definition of coalescent histories, each coalescent history f for (G, S) must map v to $e_r(S)$, as $e_r(S)$ is the only edge of S ancestral to both leaves of S that correspond to the leaves of an antipodal cherry of G. Each non-cherry internal node w of G must be ancestral to a cherry node of G; because all cherries of G are antipodal with respect to S, w must be ancestral to some antipodal cherry v^*. Hence, by part (ii) of the definition of coalescent histories, because $f(v^*) = e_r(S)$, f must also associate w with $e_r(S)$. We have therefore shown that f associates *all* internal nodes of G with $e_r(S)$, so that (G, S) has only a single coalescent history: the coalescent history associating all internal nodes of G with $e_r(S)$. Hence, (G, S) is a lonely pair.

For the converse, suppose at least one cherry v of G is not antipodal with respect to S. Then there exists a non-root edge e of S ancestral in S to both leaves descended from v in G. (G, S) then has at least two coalescent histories: one that associates v with e, and one that associates v with $e_r(S)$. □

The theorem provides a simple condition on (G, S) that indicates if (G, S) is a lonely pair: it suffices to examine the cherries of G to determine if they are all antipodal with respect to S.

3.2. Lonely-generating pairs of unlabeled tree shapes

To count the number of lonely pairs (G, S) of labeled topologies of size $n \geqslant 2$ leaves, we count the number of pairs (G, S) in which each cherry of G is antipodal with respect to S. Suppose the two subtrees of S descended from the root have sizes p and $n - p$, with $1 \leqslant p \leqslant \lfloor n/2 \rfloor$. For each cherry of G to be antipodal with respect to S, it must be possible to choose the pair of leaves of the cherry by selecting one leaf from among the p leaves of S_L and the other leaf from among the $n - p$ leaves of S_R. Thus, for (G, S) to be lonely, the number of cherries k of G must satisfy $1 \leqslant k \leqslant p$.

Definition 4. A pair of unlabeled tree shapes (g, s) is said to be *lonely-generating* if and only if (g, s) admits a labeling (G, S) such that (G, S) is lonely.

Proposition 5. *A pair of unlabeled tree shapes (g, s) is lonely-generating if and only if the number of cherries k of g is less than or equal to the number of leaves p in the smaller of the two subtrees immediately descended from the root of s.*

Proof. Suppose the number of cherries k of g is less than or equal to the number of leaves p in s_L, the smaller of the two subtrees immediately descended from the root of s. We assign labels of G and S. For each cherry v of g, identify one of the labels of its descendant leaves with a label for one of the p leaves in s_L; because $k \leqslant p$, such identifications can be simultaneously made for all cherries. Identify the label of the other descendant leaf of v with a label for one of the $n - p$ leaves in s_R. Assign the $n - 2k$ remaining labels of G and S arbitrarily. Each cherry of the labeled tree G is antipodal with respect to S. By Theorem 3, (G, S) is lonely, so that (g, s) is lonely-generating.

Now suppose $k > p$, and consider an arbitrary labeling G of g and an arbitrary labeling S of s. Then there exists at least one cherry of G both of whose leaves lie in S_R, the larger subtree of the root of S. This cherry of G is not antipodal with respect to S, so that by Theorem 3, (G, S) is not lonely, and hence (g, s) is not lonely-generating. □

To count lonely-generating pairs (g, s), we require a lemma concerning cherries.

Lemma 6. *For $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, the number of unlabeled tree shapes that have exactly k cherries is*

$$v_{n,k} = \sum_{p=1}^{\lfloor n/2 \rfloor} \Bigg[(1 - \delta_{p,n-p}) \Bigg(\sum_{i=0}^{\lfloor p/2 \rfloor} v_{p,i} v_{n-p,k-i} \Bigg) + \delta_{p,n-p} \Bigg[\sum_{i=0}^{\min\{\lfloor p/2 \rfloor, \lfloor k/2 \rfloor\}} \Bigg(v_{p,i} v_{n-p,k-i} - \delta_{i,k-i} \binom{v_{p,i}}{2} \Bigg) \Bigg] \Bigg], \tag{5}$$

where we define $v_{1,0} = v_{2,1} = 1$ and $v_{n,0} = 0$ for all $n \geqslant 2$.

Proof. We consider a decomposition of unlabeled tree shapes at the root, placing $i \geqslant 0$ cherries into the subtree of size p leaves and $k - i$ cherries into the subtree of size $n - p$ leaves. The 1-leaf unlabeled tree shape has 0 cherries, and each shape with $n \geqslant 2$ leaves has at least 1 cherry.

For odd n and for even n with $p < n/2$, the two subtrees of the root have distinct unlabeled shapes, and we tabulate $v_{p,i} v_{n-p,k-i}$ unlabeled tree shapes with exactly k cherries, i in the subtree of size p and $k - i$ in the subtree of size $n - p$.

For even n with $p = n/2$, the two subtrees of the root have the same size. To avoid double-counting, i must be bounded above by $\lfloor k/2 \rfloor$ as well as by $\lfloor p/2 \rfloor$. If $i < k/2$, then

Table 3
The number of unlabeled tree shapes with $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ cherries, for small n.

	k					
n	1	2	3	4	5	Total
2	1					1
3	1					1
4	1	1				2
5	1	2				3
6	1	4	1			6
7	1	6	4			11
8	1	9	11	2		23
9	1	12	24	9		46
10	1	16	46	32	3	98

The table is computed using values of $v_{n,k}$ from Lemma 6. The column on the right gives t_n from eq. (3).

the two subtrees continue to have distinct unlabeled shapes, and we still have $v_{p,i}v_{n-p,k-i}$ shapes for the desired quantity.

For even n with $p = n/2$ and $i = k/2$, the subtrees at the root might not have distinct unlabeled shapes. If the shapes are identical, then the number of unlabeled shapes with k cherries is $v_{n/2,k/2}$. If they are distinct, then the number of unlabeled shapes with k cherries is $\binom{v_{n/2,k/2}}{2}$. The total number of unlabeled shapes with k cherries is then equal to $v_{n/2,k/2} + \binom{v_{n/2,k/2}}{2}$ for even n with $p = n/2$ and $i = k/2$. We use the Kronecker delta to obtain one equation that combines all cases. □

For small n, values of $v_{n,k}$ appear in Table 3.

Proposition 7. *The number of lonely-generating pairs of unlabeled tree shapes* (g, s) *for* $n \geqslant 2$ *leaves is*

$$z_n = \sum_{p=1}^{\lfloor n/2 \rfloor} \ell_{n,p} \sum_{k=1}^{p} v_{n,k}. \tag{6}$$

Proof. By Proposition 5, we must count the number of pairs (g, s) for which the number of cherries k of g is less than or equal to the number of leaves p of the smaller of the two subtrees of s. Considering the results of eq. (4) and Lemma 6 for all possible (p, k) with $k \leqslant p$, we obtain the result. □

Note that by reversing the order of the summation, the quantity in Proposition 7 can be written $\sum_{k=1}^{\lfloor n/2 \rfloor} v_{n,k} \sum_{p=k}^{\lfloor n/2 \rfloor} \ell_{n,p}$. It is convenient to report the marginal sums in Proposition 7 as corollaries.

Corollary 8. *For a given unlabeled tree shape* g *with* $n \geqslant 2$ *leaves and* k *cherries,* $1 \leqslant k \leqslant \lfloor n/2 \rfloor$*, the number of unlabeled tree shapes* s *for which* (g, s) *is lonely-generating is* $y_{n,k} = \sum_{p=k}^{\lfloor n/2 \rfloor} \ell_{n,p}$.

Table 4
For a given unlabeled species tree shape s with $n \geqslant 2$ leaves divided at the root into subtrees of size p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, the number of unlabeled gene tree shapes g for which (g,s) is lonely-generating, for small n.

	p				
n	1	2	3	4	5
2	1				
3	1				
4	1	2			
5	1	3			
6	1	5	6		
7	1	7	11		
8	1	10	21	23	
9	1	13	37	46	
10	1	17	63	95	98

The table is computed using values of $x_{n,p}$ from Corollary 9.

Table 5
For a given unlabeled gene tree shape g with $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ cherries, the number of unlabeled species tree shapes s for which (g,s) is lonely-generating, for small n.

	k				
n	1	2	3	4	5
2	1				
3	1				
4	2	1			
5	3	1			
6	6	3	1		
7	11	5	2		
8	23	12	6	3	
9	46	23	12	6	
10	98	52	29	18	6

The table is computed using values of $y_{n,k}$ from Corollary 8.

Corollary 9. *For a given unlabeled tree shape s with $n \geqslant 2$ leaves that is divided at the root into subtrees of size p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, the number of unlabeled tree shapes g for which (g,s) is lonely-generating is $x_{n,p} = \sum_{k=1}^{p} v_{n,k}$.*

For small n, the quantities $x_{n,p}$ and $y_{n,k}$ appear in Tables 4 and 5, respectively.

3.3. Lonely pairs of labeled topologies

To enumerate lonely pairs of labeled topologies (G,S), we next need a pair of lemmas that count labeled topologies satisfying conditions concerning cherries.

Lemma 10. *For $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, the number of labeled topologies that have exactly k cherries is*

$$V_{n,k} = \frac{n!\,(n-2)!}{2^{2k-1}(n-2k)!\,k!\,(k-1)!}. \tag{7}$$

Table 6
The number of labeled topologies with $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ cherries, for small n.

	k					
n	1	2	3	4	5	Total
2	1					1
3	3					3
4	12	3				15
5	60	45				105
6	360	540	45			945
7	2520	6300	1575			10395
8	20160	75600	37800	1575		135135
9	181440	952560	793800	99225		2027025
10	1814400	12700800	15876000	3969000	99225	34459425

The table is computed using values of $V_{n,k}$ from Lemma 10. The column on the right gives T_n from eq. (1).

Proof. This result follows by multiplying the probability that a labeled topology chosen uniformly at random has k cherries (Theorem 6 of [23]) by the number of labeled topologies T_n (eq. (1)). □

For small n, values of $V_{n,k}$ appear in Table 6.

Lemma 11. *For $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, the number of labeled topologies that have exactly k cherries, the leaf pairings of which are specified, is*

$$W_{n,k} = \frac{2^k (n-2k)!\,k!}{n!} V_{n,k}. \tag{8}$$

Proof. Lemma 10 gives the number of labeled topologies that have exactly k cherries, or $V_{n,k}$. Each of these labeled topologies has a set of leaf pairings for the k cherries, each of which appears in the same number of labeled topologies. The number of possible sets of leaf pairings is $\binom{n}{2k} T_{k+1} = n!/[2^k (n-2k)!\,k!]$, where $\binom{n}{2k}$ is the number of ways of choosing $2k$ leaves to place in the k cherries, and T_{k+1} (eq. (1)) gives the number of perfect matchings placing $2k$ elements into k pairs [6]. We divide the number of labeled topologies with exactly k cherries, $V_{n,k}$, by the number of sets of leaf pairings for the k cherries to obtain the result. □

Lemma 12. *For a given labeled topology S with $n \geqslant 2$ leaves that is divided at the root into subtrees of size p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, the number of labeled topologies G for which (G, S) is lonely is*

$$X_{n,p} = \sum_{k=1}^{p} \frac{p!\,(n-p)!\,(n-2)!}{2^{k-1} k!\,(p-k)!\,(n-p-k)!\,(k-1)!}. \tag{9}$$

Proof. By Theorem 3, we must count labeled topologies G all of whose cherries are antipodal with respect to S. The number of cherries k satisfies $1 \leqslant k \leqslant p$, or else at least one cherry of G will have both its leaves in the larger subtree S_R of S and will not be an antipodal cherry with respect to S. Each antipodal cherry contains one of the p leaves of S_L and one of the $n-p$ leaves of S_R.

Given k with $1 \leqslant k \leqslant p$, the number of ways that k of the p leaves in S_L can be chosen for placement in cherries of G, each in a different cherry, is $\binom{p}{k}$. Once these k leaves have been selected, the number of ways that they can be paired with k leaves from S_R to produce k cherries is $(n-p)(n-p-1)\cdots(n-p-k+1) = (n-p)!/(n-p-k)!$, sequentially choosing without replacement among leaves of S_R to form the cherries.

We must then multiply the product $\binom{p}{k}(n-p)!/(n-p-k)!$, describing the number of ways of selecting the k leaf pairs for the k cherries, by $W_{n,k}$, the number of labeled topologies that contain k specific cherries, and no other cherries (Lemma 11). We obtain the result by simplifying the expression

$$\sum_{k=1}^{p} \binom{p}{k} \frac{(n-p)!}{(n-p-k)!} W_{n,k}. \quad \square$$

Lemma 13. *For a given labeled topology G with $n \geqslant 2$ leaves and k cherries, $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, the number of labeled topologies S for which (G, S) is lonely is*

$$Y_{n,k} = \sum_{p=k}^{\lfloor n/2 \rfloor} \frac{(n-2k)!\,(2n-2p-2)!\,(2p-2)!}{2^{n-k-2}(p-k)!\,(n-p-k)!\,(n-p-1)!\,(p-1)!} \left(\frac{1}{2}\right)^{\delta_{p,n-p}}. \tag{10}$$

Proof. By Theorem 3, we must count labeled topologies S with respect to which all k of the cherries of G are antipodal. The number of leaves p in the smaller subtree S_L of S must be at least k, or else at least one cherry of G will not be an antipodal cherry with respect to S.

Given p with $k \leqslant p \leqslant \lfloor n/2 \rfloor$, the number of ways of placing one leaf of each of the k cherries of G in S_L and the other leaf in S_R is $2^{k-\delta_{p,n-p}}$. The Kronecker delta reflects the fact that if $p = n-p$, then each assignment of the leaves of the cherries to subtrees of S is counted twice, once when a set of leaves from the k cherries of G is chosen for placement in S_L, and once when its complement with respect to the set of $2k$ leaves in the k cherries is chosen for placement in S_L. For each assignment of the leaves of the k cherries to S_L and S_R, the number of ways of choosing leaves of G for placement among the p leaves in subtree S_L is $\binom{n-2k}{p-k}$, as $n-2k$ leaves of S lie outside the k cherries.

We must then multiply the product $2^{k-\delta_{p,n-p}}\binom{n-2k}{p-k}$, describing the number of ways of assigning the p and $n-p$ leaves to the subtrees of S, by the number of labeled topologies with each assignment. This quantity is $T_p T_{n-p}$, where T_n gives the number of labeled topologies with a specified set of n leaves (eq. (1)). We obtain the result by simplifying the expression

$$\sum_{p=k}^{\lfloor n/2 \rfloor} 2^{k-\delta_{p,n-p}} \binom{n-2k}{p-k} T_p T_{n-p}. \quad \square$$

Table 7
For a given labeled species tree topology S with $n \geqslant 2$ leaves divided at the root into subtrees of size p and $n - p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, the number of labeled gene tree topologies G for which (G, S) is lonely, for small n.

	p				
n	1	2	3	4	5
2	1				
3	2				
4	6	10			
5	24	54			
6	120	336	450		
7	720	2400	3960		
8	5040	19440	37800	46440	
9	40320	176400	393120	567000	
10	362880	1774080	4445280	7318080	8580600

The table is computed using values of $X_{n,p}$ from Lemma 12.

Table 8
For a given labeled gene tree topology G with $n \geqslant 2$ leaves and $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ cherries, the number of labeled species tree topologies S for which (G, S) is lonely, for small n.

	k				
n	1	2	3	4	5
2	1				
3	2				
4	8	2			
5	48	12			
6	384	96	36		
7	3840	960	360		
8	46080	11520	4320	1800	
9	645120	161280	60480	25200	
10	10321920	2580480	967680	403200	176400

The table is computed using values of $Y_{n,k}$ from Lemma 13.

For small n, the quantities $X_{n,p}$ and $Y_{n,k}$ appear in Tables 7 and 8, respectively.

Theorem 14. *The number of lonely pairs of labeled topologies* (G, S) *for* $n \geqslant 2$ *leaves is*

$$Z_n = \sum_{p=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{p} \frac{(2n-2p-2)!\,(2p-2)!\,n!\,(n-2)!}{2^{n+k-3}(p-k)!\,(n-p-k)!\,(n-p-1)!\,(p-1)!\,k!\,(k-1)!} \left(\frac{1}{2}\right)^{\delta_{p,n-p}}. \tag{11}$$

Proof. For a fixed labeled topology G with $n \geqslant 2$ leaves and k cherries, $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, Lemma 13 gives the number of labeled topologies S for which (G, S) is lonely. We sum the result of Lemma 13 over all possible labeled topologies G. In particular, the number of labeled topologies G with k cherries is $V_{n,k}$ (Lemma 10). To obtain the result, we simplify the sum $\sum_{k=1}^{\lfloor n/2 \rfloor} V_{n,k} Y_{n,k}$. □

Note that we can prove Theorem 14 by using Lemma 12 instead of Lemma 13. For a fixed labeled topology S with n leaves that is divided at the root into subtrees of size p

Table 9
The number of lonely pairs for small n.

Number of leaves n	Number of unlabeled tree shapes t_n (eq. (3))	Number of labeled topologies T_n (eq. (1))	Number of lonely-generating pairs z_n (Proposition 7)	Number of lonely pairs Z_n (Theorem 14)
2	1	1	1	1
3	1	3	1	6
4	2	15	3	102
5	3	105	5	3420
6	6	945	19	191700
7	11	10395	49	16291800
8	23	135135	203	1966015800
9	46	2027025	664	321188943600
10	98	34459425	2858	68482943802000

and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, we sum the result of Lemma 12 over all possible values of p, noting that the number of labeled topologies S divided at the root into subtrees of size p and $n-p$ leaves is $L_{n,p}$ (eq. (2)). We obtain the same expression in eq. (11) by simplifying the sum $\sum_{p=1}^{\lfloor n/2 \rfloor} L_{n,p} X_{n,p}$.

In Theorem 14 as well as in Lemmas 12 and 13, some terms can be factored out of the summand; the formulas are written with all terms inside the sum to highlight that the sum can proceed in the reverse order, with $\sum_{k=1}^{\lfloor n/2 \rfloor} \sum_{p=k}^{\lfloor n/2 \rfloor}$ in place of $\sum_{p=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{p}$ in eq. (11). For small n, we report values of z_n, the number of lonely-generating unlabeled pairs, and Z_n, the number of lonely pairs, in Table 9.

4. Probabilities

Until this point, our results have been stated as enumerations. We convert them to probabilities that choices of unlabeled or labeled trees give rise to lonely-generating or lonely pairs by dividing by the sizes of associated classes of trees. These probabilities follow from results 7–9 and 12–14 (Table 10).

Fig. 4A plots the probability z_n/t_n^2 that an unlabeled pair (g, s) chosen uniformly at random is lonely-generating. The plot illustrates a decline in the probability as n increases, with fewer than 10% of pairs at $n = 40$ being lonely-generating. In other words, as n increases, it is observed that the probability decreases that the random unlabeled species tree shape s is divided at the root into subtrees with size p and $n-p$, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$, such that p is greater than or equal to the number of cherries k in an unlabeled gene tree shape g also chosen uniformly at random.

Examining the probability $y_{n,k}/t_n$ that an unlabeled pair (g, s) is lonely-generating when g is fixed and has k cherries, we observe that for fixed k, $y_{n,k}/t_n$ generally increases for increasing n (Fig. 4B). It is trivial to demonstrate that for fixed n, $y_{n,k}/t_n$ decreases monotonically from $y_{n,1}/t_n = 1$ as k increases from 1 to $\lfloor n/2 \rfloor$; as k increases, fewer terms $\ell_{n,p}$ are summed in the formula $y_{n,k} = \sum_{p=k}^{\lfloor n/2 \rfloor} \ell_{n,p}$, so that $y_{n,k}$ is smaller.

Table 10
Probabilities for lonely-generating and lonely pairs of trees.

Quantity that is fixed	Quantity being selected at random	Condition whose probability is being calculated	Number of choices satisfying the condition	Number of possible choices	Probability
–	(g, s)	(g, s) is LG	z_n (Proposition 7)	t_n^2 (eq. (3))	z_n/t_n^2 (Fig. 4A)
Unlabeled g	s	(g, s) is LG	$y_{n,k}$ (Corollary 8)	t_n (eq. (3))	$y_{n,k}/t_n$ (Fig. 4B)
Unlabeled s	g	(g, s) is LG	$x_{n,p}$ (Corollary 9)	t_n (eq. (3))	$x_{n,p}/t_n$ (Fig. 4C)
–	(G, S)	(G, S) is lonely	Z_n (Theorem 14)	T_n^2 (eq. (1))	Z_n/T_n^2 (Fig. 4D)
Labeled G	S	(G, S) is lonely	$Y_{n,k}$ (Lemma 13)	T_n (eq. (1))	$Y_{n,k}/T_n$ (Fig. 4E)
Labeled S	G	(G, S) is lonely	$X_{n,p}$ (Lemma 12)	T_n (eq. (1))	$X_{n,p}/T_n$ (Fig. 4F)

Fixed gene trees g and G are assumed to have $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ cherries. Fixed species trees s and S are assumed to be divided at the root into subtrees with p and $n-p$ leaves, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$. LG, lonely-generating.

The effects seen for $y_{n,k}/t_n$ are reversed for $x_{n,p}/t_n$, the probability that an unlabeled pair (g, s) is lonely-generating when s is fixed and divided at the root into subtrees with size p and $n-p$, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$ (Fig. 4C). The figure illustrates a decrease of $x_{n,p}/t_n$ with n for fixed p, starting from $x_{2p,p}/t_{2p} = 1$. For fixed n, $x_{n,p}/t_n$ increases monotonically to $x_{n,\lfloor n/2 \rfloor}/t_n = 1$, as p increases and more terms are incorporated into $x_{n,p} = \sum_{k=1}^{p} v_{n,k}$.

For labeled trees, the probability Z_n/T_n^2 that a labeled pair (G, S) chosen uniformly at random is lonely is seen to decrease with n, faster than the decay of z_n/t_n^2 (Fig. 4D). The computation of Z_n/T_n tabulates lonely pairs (G, S), whereas z_n/t_n tabulates pairs (g, s) that are only required to be lonely-generating; each lonely pair (G, S) represents a labeling of a lonely-generating pair (g, s), but not every labeling of a lonely-generating pair (g, s) produces a lonely pair (G, S).

The probability $Y_{n,k}/T_n$ that a labeled pair (G, S) is lonely when G is fixed and has k cherries (Fig. 4E), and the probability $X_{n,p}/T_n$ that a labeled pair (G, S) is lonely when S is fixed and is divided at the root into subtrees with size p and $n-p$, $1 \leqslant p \leqslant \lfloor n/2 \rfloor$ (Fig. 4F), are observed to have somewhat similar behavior to the corresponding unlabeled quantities $y_{n,k}/t_n$ and $x_{n,p}/t_n$, but with smaller values. Like $y_{n,k}/t_n$, $Y_{n,k}/T_n$ is seen to decrease with k for fixed n, and like $x_{n,p}/t_n$, $X_{n,p}/T_n$ is observed to decrease with n for fixed p and to increase with p for fixed n. One difference is that whereas $y_{n,k}/t_n$ is seen to increase with n for fixed k, $Y_{n,k}/T_n$ is seen to decrease with n for fixed k. This result has the interpretation that whereas the fraction of unlabeled tree shapes s whose smaller subtree at the root has size at least k increases, the fraction decreases that a labeled tree topology S both has a divide at the root with a smaller subtree of size at least k and satisfies the restriction that its labeling causes all cherries for a fixed labeled tree topology G to be antipodal with respect to S.

5. Discussion

This study has examined the features of pairs consisting of a gene tree and a species tree, characterizing the lonely-generating unlabeled pairs (Proposition 5) and the lonely labeled pairs (Theorem 3). The condition that causes loneliness is that all cherries of the

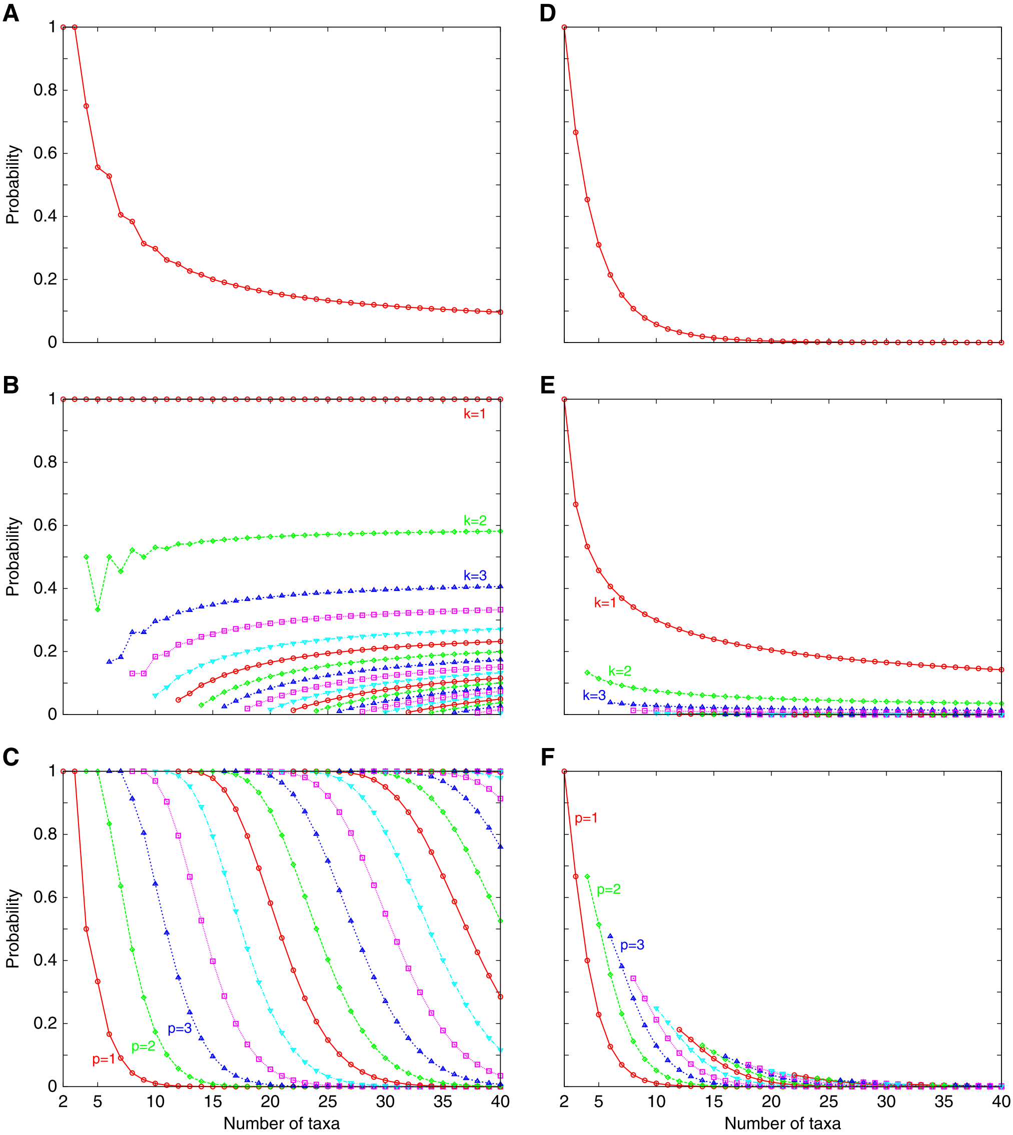


Fig. 4. Probabilities that tree pairs are lonely-generating or lonely, for $2 \leqslant n \leqslant 40$. (A) Probability z_n/t_n^2 that an unlabeled pair (g, s) chosen uniformly at random is lonely-generating. (B) Probability $y_{n,k}/t_n$ that an unlabeled pair (g, s) is lonely-generating when g is fixed with k cherries and s is chosen uniformly at random. (C) Probability $x_{n,p}/t_n$ that an unlabeled pair (g, s) is lonely-generating when s is fixed and divided at the root into subtrees with p and $n-p$ leaves and g is chosen uniformly at random. (D) Probability Z_n/T_n^2 that a labeled pair (G, S) chosen uniformly at random is lonely. (E) Probability $Y_{n,k}/T_n$ that a labeled pair (G, S) is lonely when G is fixed with k cherries and S is chosen uniformly at random. (F) Probability $X_{n,p}/T_n$ that a labeled pair (G, S) is lonely when S is fixed and divided at the root into subtrees with p and $n-p$ leaves and G is chosen uniformly at random. In (B), (C), (E), and (F), the first three curves are labeled, and subsequent curves for values of k and p up to 20 follow in sequential order.

gene tree are antipodal with respect to the species tree. Application of this condition enables an enumeration of lonely-generating pairs via a recursive formula (Proposition 7) and of lonely pairs via a summation (Theorem 14).

The key feature of a gene tree that influences its potential to produce lonely pairs is its number of cherries: because each additional cherry is less likely to be antipodal in a gene tree with a larger number of cherries, gene trees with more cherries appear in fewer lonely pairs (Lemma 13). The key feature of a species tree is the pair of sizes for the two subtrees immediately descended from the root (Lemma 12): because species trees for which the root node is more balanced can accommodate more antipodal cherries, species trees that are more balanced at the root appear in more lonely pairs.

Gene trees and species trees in lonely pairs are "distant" in the sense that the cherries of the gene tree are antipodal with respect to the species tree. However, the characterization of lonely pairs via antipodal cherries has the consequence that other ways of examining differences between gene trees and species trees need not be closely related to loneliness. For example, for the Robinson–Foulds (RF) distance, counting splits that appear in one but not the other of a pair of trees [18, p. 25], a pair with a relatively large value need not be lonely—as is seen for the 4-leaf trees $G = ((A, C), (B, D))$ and $S = (((A, B), C), D)$ in Fig. 2, with 2 coalescent histories and RF distance 2. A pair with minimal Robinson–Foulds distance can be lonely, such as for $G = (((C, D), A), B)$ and $S = (((A, B), C), D)$, with 1 coalescent history and RF distance 0.

Recall that the interest in lonely pairs arises from the importance of coalescent histories to combinatorial and probabilistic features of gene trees and species trees. While a recursive computation can give the number of coalescent histories for an arbitrary pair consisting of a gene tree and a species tree [14], in demonstrating that a simple condition characterizes the set of lonely pairs, we have found a way of obtaining the number of coalescent histories for such pairs that is simpler than the recursive computation. The condition of Theorem 3 can be checked; if it holds, then the number of coalescent histories is equal to 1, and only if it fails is the recursive computation necessary.

A second setting in which lonely pairs have appeared is in the analysis of compact coalescent histories. A compact coalescent history groups into one equivalence class all coalescent histories that for each species tree edge have the same numbers of coalescences. For several families of gene trees and species trees of increasing size, the number of compact coalescent histories grows exponentially slower with the number of leaves than the number of coalescent histories [9]. The lonely pairs, however, with only one coalescent history, also have only one compact coalescent history, illustrating that compact coalescent histories need not be less numerous than coalescent histories.

Although many enumerative studies of coalescent histories have now been performed, focusing on enumerating the coalescent histories of various families of matching and non-matching shapes [1,5,7,8,14–16,20], this analysis is the first to begin from a value for the number of coalescent histories, namely 1, and to characterize all pairs consisting of a gene tree and a species tree whose number of coalescent histories is equal to that value. It will be of interest to determine if similar results can be obtained for the pairs that produce other specified values for the number of coalescent histories.

Acknowledgments

I thank an anonymous reviewer for helpful comments. Support was provided by NIH grant R01 GM117590.

References

[1] J.H. Degnan, Gene Tree Distributions Under the Coalescent Process, PhD thesis, University of New Mexico, Albuquerque, 2005.

[2] J.H. Degnan, J.A. Rhodes, There are no caterpillars in a wicked forest, Theor. Popul. Biol. 105 (2015) 17–23.

[3] J.H. Degnan, N.A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent, Trends Ecol. Evol. 24 (2009) 332–340.

[4] J.H. Degnan, N.A. Rosenberg, T. Stadler, The probability distribution of ranked gene trees on a species tree, Math. Biosci. 235 (2012) 45–55.

[5] J.H. Degnan, L.A. Salter, Gene tree distributions under the coalescent process, Evolution 59 (2005) 24–37.

[6] P.W. Diaconis, S.P. Holmes, Matchings and phylogenetic trees, Proc. Natl. Acad. Sci. USA 95 (1998) 14600–14602.

[7] F. Disanto, N.A. Rosenberg, Coalescent histories for lodgepole species trees, J. Comput. Biol. 22 (2015) 918–929.

[8] F. Disanto, N.A. Rosenberg, Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (2016) 913–925.

[9] F. Disanto, N.A. Rosenberg, Enumeration of compact coalescent histories for matching gene trees and species trees, J. Math. Biol. (2019), https://doi.org/10.1007/s00285-018-1271-5.

[10] E.F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. in Appl. Probab. 3 (1971) 44–77.

[11] W.P. Maddison, Gene trees in species trees, Syst. Biol. 46 (1997) 523–536.

[12] P. Pamilo, M. Nei, Relationships between gene trees and species trees, Mol. Biol. Evol. 5 (1988) 568–583.

[13] N.A. Rosenberg, The probability of topological concordance of gene trees and species trees, Theor. Popul. Biol. 61 (2002) 225–247.

[14] N.A. Rosenberg, Counting coalescent histories, J. Comput. Biol. 14 (2007) 360–377.

[15] N.A. Rosenberg, Coalescent histories for caterpillar-like families, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 1253–1262.

[16] N.A. Rosenberg, J.H. Degnan, Coalescent histories for discordant gene trees and species trees, Theor. Popul. Biol. 77 (2010) 145–151.

[17] N.A. Rosenberg, R. Tao, Discordance of species trees with their most likely gene trees: the case of five taxa, Syst. Biol. 57 (2008) 131–140.

[18] M. Steel, Phylogeny: Discrete and Random Processes in Evolution, Society for Industrial and Applied Mathematics, Philadelphia, 2016.

[19] C. Than, L. Nakhleh, Species tree inference by minimizing deep coalescences, PLoS Comput. Biol. 5 (2009) e1000501.

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

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

[22] Y. Wu, An algorithm for computing the gene tree probability under the multispecies coalescent and its application in the inference of population tree, Bioinformatics 32 (2016), i225–i233.

[23] T. Wu, K.P. Choi, On joint subtree distributions under two evolutionary models, Theor. Popul. Biol. 108 (2016) 13–23.