Theoretical Population Biology 77 (2010) 145–151

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Coalescent histories for discordant gene trees and species trees

Noah A. Rosenberg[a,b,c,*], James H. Degnan[a,d]

[a] Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, United States
[b] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, United States
[c] Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, United States
[d] Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

ARTICLE INFO

Article history:
Received 30 January 2009
Available online 11 January 2010

Keywords:
Coalescent history
Gene tree
Species tree

ABSTRACT

Given a gene tree and a species tree, a coalescent history is a list of the branches of the species tree on which coalescences in the gene tree take place. Each pair consisting of a gene tree topology and a species tree topology has some number of possible coalescent histories. Here we show that, for each $n \geq 7$, there exist a species tree topology S and a gene tree topology $G \neq S$, both with n leaves, for which the number of coalescent histories exceeds the corresponding number of coalescent histories when the species tree topology is S and the gene tree topology is also S. This result has the interpretation that the gene tree topology G discordant with the species tree topology S can be produced by the evolutionary process in more ways than can the gene tree topology that matches the species tree topology, providing further insight into the surprising combinatorial properties of gene trees that arise from their joint consideration with species trees.

1. Introduction

For a labeled species tree topology S and a labeled gene tree topology G, both with n leaves, a *coalescent history* is a list of edges of S on which the coalescences in G occur (Degnan and Salter, 2005). For each choice of S and G, many coalescent histories might exist, with each coalescent history supplying a distinct set of edges that can describe the locations of the coalescence events in G (Fig. 1).

Coalescent histories provide a combinatorial set of objects useful in gene tree probability computations. Under a standard probabilistic model describing the descent of genealogical lineages on a given species tree — the "multispecies coalescent" (Degnan and Rosenberg, 2009) — the probability that a random gene tree has a given topology can be written as a sum of terms, each representing the joint probability of the topology and a specific coalescent history (Degnan and Salter, 2005). The state space of probabilistic models of gene tree topologies at sequences of locations along a genome — as in recent genomic studies of humans, chimpanzees, and gorillas (Hobolth et al., 2007; Dutheil et al., 2009) — can also be described using coalescent histories (Degnan and Rosenberg, 2009).

Following Than et al. (2007), we provide a formal definition of a coalescent history. For a rooted tree topology T with n leaves labeled by set X, let $E(T)$ denote the set of internal edges of T, numbered in a postorder traversal. The number for a node is identified with the number for its immediate ancestral edge (Fig. 1), so that the number for a descendant edge is smaller than the number for its immediate ancestral edge (and hence it is smaller than the numbers for all its ancestral edges). We define a partial order $\leq_T$ such that, for two distinct edges e_1 and e_2, $e_1 <_T e_2$ if and only if edge e_2 is ancestral to edge e_1 in T. For each internal edge e of T, let c_e^T denote the set of labels in X of all leaves descended from e; this set is a "cluster" in T that is identified with edge e. Let C_T denote the set of clusters in T, $C_T = \{c_e^T : e \in E(T)\}$.

Definition 1.1. For a labeled gene tree topology G and a labeled species tree topology S with the same set of leaf labels, a *coalescent history* is a mapping $\alpha : C_G \to E(S)$ such that (1) for each $Y \in C_G$, $Y \subset c^S_{\alpha(Y)}$, and (2) for each $e_1, e_2 \in E(G)$, if $e_1 <_G e_2$, then $\alpha(e_1) \leq_S \alpha(e_2)$.

These conditions formalize the rule that the coalescence of the cluster Y in the gene tree topology must occur at least as deep in the species tree as the most recent common ancestor (MRCA) of Y in the species tree topology, and the rule that a cluster in the gene tree topology cannot find its MRCA on an edge in the species tree topology deeper than an edge on which one of its "superclusters" finds its MRCA. The number of mappings $\alpha : C_G \to E(S)$ is $(n-1)^{n-1}$; however, most of these mappings do not satisfy criteria (1) and (2), and therefore do not represent coalescent histories.

[*] Corresponding author at: Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, United States.
E-mail address: rnoah@umich.edu (N.A. Rosenberg).

0040-5809/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2009.12.004

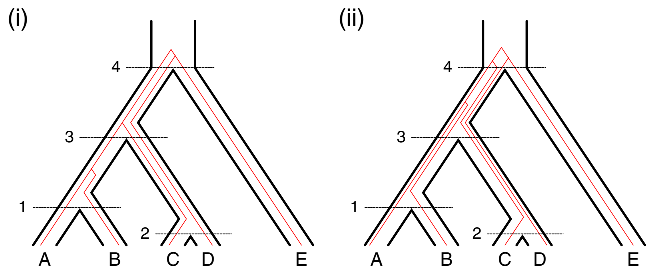


Fig. 1. Two coalescent histories for a gene tree topology and species tree topology with five leaves. Internal nodes of the species tree are numbered according to a postorder traversal, and the edge above an internal node is given the number for the node. In both (i) and (ii), the species tree, represented by thick lines, has labeled topology (((A, B), (C, D)), E), and the gene tree, represented by thin lines, has labeled topology (((A, B), C), (D, E)). However, the two diagrams represent two distinct coalescent histories. In (i), coalescences (A, B), ((A, B), C), (D, E), and (((A, B), C), (D, E)) occur on edges 1, 3, 4, and 4, respectively, whereas in (ii), they occur on edges 3, 4, 4, and 4, respectively. The two coalescent histories shown represent two of the five coalescent histories possible for gene tree topology (((A, B), C), (D, E)) and species tree topology (((A, B), (C, D)), E), the other three having the coalescences of (A, B), ((A, B), C), (D, E), and (((A, B), C), (D, E)) on edges 1, 4, 4, and 4, on edges 3, 3, 4, and 4, and on edges 4, 4, 4, and 4, respectively.

Given a gene tree topology and a species tree topology, the number of coalescent histories associated with the pair of topologies can be counted using a recursive formula (Rosenberg, 2007; Than et al., 2007). Rosenberg (2007) obtained closed-form expressions for the number of coalescent histories in various special cases with the property that the gene tree topology and species tree topology were identical. Here we investigate the number of coalescent histories when the gene tree topology and species tree topology are not necessarily identical. We show that, for each $n \geq 7$, n-leaf species tree topologies exist for which the number of coalescent histories for a nonmatching gene tree topology ($G \neq S$) can exceed the number of coalescent histories for the matching gene tree topology ($G = S$). This result is obtained by exhaustive consideration of all species tree topologies and gene tree topologies with $n \leq 9$ leaves, together with an inductive proof for $n \geq 9$ leaves.

Our main result complements the main theorem of Degnan and Rosenberg (2006). Previously, we showed that, for each species tree topology with $n \geq 5$ leaves, and for the asymmetric species tree topology with $n = 4$ leaves, branch lengths exist for the species tree such that the species tree topology disagrees with the gene tree topology it is most likely to produce under the multispecies coalescent. Informally, we show here that, for each $n \geq 7$ leaves, there exists an n-leaf species tree topology for which a discordant gene tree topology has more ways of evolving than the matching gene tree topology.

2. Definitions and background

We follow the terminology of Rosenberg (2007) and highlight a few key concepts, considering only rooted binary trees. A *caterpillar* tree is a tree in which each internal node has at least one leaf as one of its immediate descendants. A *pseudocaterpillar* tree (Rosenberg, 2007) is a tree with at least five leaves in which each internal node, with one exception, has at least one leaf as one of its immediate descendants, and in which the one internal node that does not have this property is ancestral to exactly four leaves (Fig. 2).

We restrict our attention to gene tree topologies and species tree topologies with equally many leaves (that is, we assume that only one gene lineage is examined per species). If a gene tree and species tree have the same labeled topology, we describe the topologies as *identical* and refer to the gene tree topology as *matching* the species tree topology; otherwise the gene tree topology is *nonmatching* or *discordant*. Unless otherwise stated, it is implicit that gene tree topologies and species tree topologies are labeled

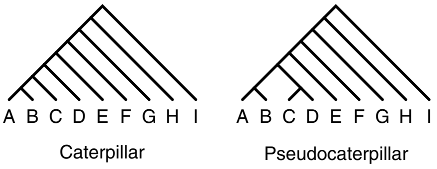


Fig. 2. Caterpillar and pseudocaterpillar tree topologies with $n = 9$ leaves.

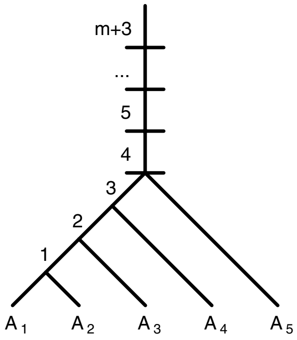


Fig. 3. An m-extended species tree topology, for which the edge above the root is artificially divided into m edges. The numbers denote labels for the edges. If the gene tree topology is $(((A_1, A_2), (A_3, A_4)), A_5)$, then an m-extended coalescent history involves a coalescence of $((A_1, A_2), (A_3, A_4))$ and A_5 on an edge k from 4 to $m + 3$, a coalescence of (A_1, A_2) and (A_3, A_4) on an edge ℓ from 3 to k, a coalescence of (A_1, A_2) on some edge from 1 to ℓ, and a coalescence of (A_3, A_4) on some edge from 3 to ℓ. If $m = 1$, then the edge above the root is not subdivided, and m-extended coalescent histories reduce simply to coalescent histories.

rather than unlabeled. Although a given coalescent history is associated with both a gene tree topology and a species tree topology, we sometimes treat the species tree topology as fixed and refer to coalescent histories as being possessed only by the gene tree topology.

An *m-extended coalescent history* for a (labeled) gene tree topology and (labeled) species tree topology is a coalescent history for the gene tree topology and species tree topology when the edge above the root of the species tree topology is subdivided into m components (Fig. 3). We denote the number of m-extended coalescent histories for a gene tree topology G and a species tree topology S, obtained using Theorem 4.3 of Rosenberg (2007), by $B_{G,S,m}$. Although we are primarily interested in the case of $m = 1$, in performing computations of $B_{G,S,1}$ it is convenient to first consider general values of m. Setting $m = 1$ reduces m-extended coalescent histories to coalescent histories; $B_{G,S,1}$ then represents the number of coalescent histories for gene tree topology G and species tree topology S.

3. Small trees

We begin by exhaustively considering the number of coalescent histories for small trees with $n \leq 9$ leaves. The number of cases that must be examined is the product of the number of labeled gene tree topologies and the number of *unlabeled* species tree topologies. Although the enumeration of coalescent histories requires a labeled species tree topology, without loss of generality, we need only consider one labeling of each unlabeled species tree topology. For convenience, when this arbitrary labeling of the species tree topology is not important for describing a result, we abbreviate the arbitrarily labeled species tree topology by its underlying unlabeled topology.

Rosenberg (2007) reported the number of coalescent histories for each case with $n \leq 9$ leaves in the setting in which the gene tree topology matched the species tree topology. To obtain the number of coalescent histories for all remaining cases with $n \leq 9$ leaves, we used the COUNTCOAL routine of the PhyloNet package

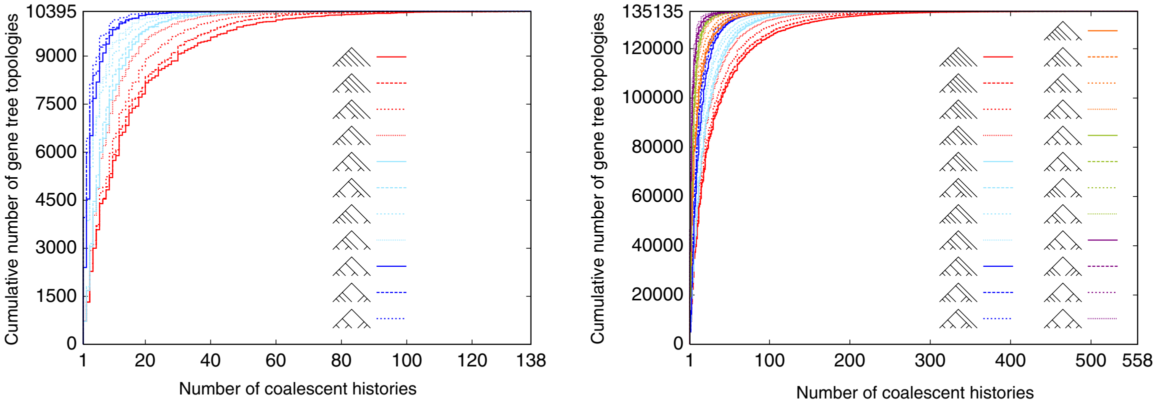


Fig. 4. Cumulative number of gene tree topologies with at most h coalescent histories, as a function of the number of coalescent histories h. Left — species tree topologies with $n = 7$ leaves; right — species tree topologies with $n = 8$ leaves.

(Than et al., 2008), which implements the efficient recursive counting algorithm described by Than et al. (2007). This approach exactly recovers the values reported previously in Tables 1–4 of Rosenberg (2007) for cases with identical gene tree and species tree topology. We have also verified many of the results obtained from PhyloNet by using the COAL package, which implements an earlier nonrecursive algorithm for counting coalescent histories (Degnan and Salter, 2005). Although our formal Definition 1.1 of a coalescent history differs slightly from the corresponding Definition 2 of Than et al. (2007), correcting a small error in condition 2 of the definition of Than et al., the underlying concept is identical, and we and Than et al. count the same sets of objects.

For each species tree topology with $n \leq 6$, Table 1 shows the distribution of the number of coalescent histories, considering all gene tree topologies with n leaves. Some of the features of this distribution can be understood by examining the properties of "history classes" (Rosenberg and Tao, 2008), each of which describes a set of gene tree topologies that have exactly the same list of coalescent histories (in the sense that the clusters of any gene tree topology G_1 in a history class can be identified with the clusters of any other gene tree topology G_2 in the history class so that the list of coalescent histories for G_1 is obtained from the list of coalescent histories for G_2 simply by substituting the clusters of G_1 for the corresponding clusters of G_2). For a fixed species tree topology, permutations of certain subsets of the gene tree topology leaf labels generate topologies in the same history class, so that the number of gene tree topologies in a history class is often a count of permutations. Because such numbers of permutations are often factorials, the number of topologies in a history class often has many divisors. For a given species tree topology, the number of gene tree topologies with exactly h coalescent histories is the sum across history classes producing h coalescent histories of the numbers of gene tree topologies in these various history classes. As each of these numbers often has many divisors, some of which are shared, it often holds that the number of gene tree topologies with h coalescent histories — the entry for h coalescent histories in Table 1 — also has many divisors.

Table 1 also illustrates that, for each species tree topology, most gene tree topologies have relatively few coalescent histories. This result can be observed even more dramatically in Fig. 4, which for species tree topologies with $n = 7$ and $n = 8$ leaves shows the cumulative number of gene tree topologies with at most h coalescent histories (for each h). For a given species tree topology, consider the median number of coalescent histories across gene tree topologies as a fraction of the maximal number of coalescent histories, and label this ratio by r. The values of r are 1/2 for the three-leaf species tree topology, 2/5 and 1/4 for the four-leaf species tree topologies, and 3/14, 3/13, and 2/10 for the five-leaf species tree topologies. The minimal and maximal values of r across six-leaf species tree topologies are 2/28 and 5/42, illustrating an increased proportion of gene tree topologies with a small number of coalescent histories. For $n = 7$ leaves, the minimal and maximal r decline further to 2/70 and 9/132, respectively; for $n = 8$ they are 2/196 and 18/429, and for $n = 9$ they are 3/588 and 30/1784.

A comparison of Table 1 with Table 1 of Rosenberg (2007) indicates that, for $n \leq 6$ leaves, if the species tree topology is fixed, then the number of coalescent histories is greatest for the matching gene tree topology. From complete enumerations of coalescent histories for $n = 7$, $n = 8$, and $n = 9$ leaves, however, we found that examples exist in which a nonmatching gene tree topology has more coalescent histories than the matching gene tree topology (Table 2). The single instance of this phenomenon for $n = 7$ leaves appears in Fig. 5, as does the single instance with $n = 7$ in which a nonmatching gene tree topology has the same number of coalescent histories as the matching gene tree topology. Both of these examples involve the caterpillar species tree topology.

Similar examples with $n = 8$ leaves are much more numerous (Table 2), and they occur for 7 of the 23 (unlabeled) species tree topologies (Table 3). In particular, the caterpillar and pseudocaterpillar species tree topologies with $n = 8$ leaves have many nonmatching gene tree topologies with more coalescent histories than the matching gene tree topology. With a pseudocaterpillar species tree topology, three nonmatching gene tree topologies each have 558 coalescent histories, considerably exceeding the 462 coalescent histories for the matching pseudocaterpillar gene tree topology and producing a tie for the greatest number of coalescent histories among all cases with $n = 8$ leaves.

Finally, for $n = 9$, examples in which nonmatching gene tree topologies have more coalescent histories than the matching gene tree topology occur for 22 of 46 (unlabeled) species tree topologies (Table 4). Noticeable structure is visible in this collection of 22 topologies. For 15 of the topologies, leaf I is immediately descended from the root; the remaining 7 topologies can be obtained by substituting the cherry (H,I) for leaf H in the species tree topologies with $n = 8$ shown in Table 3. The species tree topology with the largest number of nonmatching gene tree topologies that have more coalescent histories than the matching gene tree topology continues to be the caterpillar, with 865 such nonmatching topologies, a much greater number than the corresponding value of 42 topologies for the caterpillar with $n = 8$ leaves. The ratio of the largest number of coalescent histories for a nonmatching gene tree topology to the largest number of coalescent histories for a matching gene tree topology also increases, from 558/462 for $n = 8$ to 2511/1663 for $n = 9$.

Table 1
Frequency distribution of the number of coalescent histories among the $(2n-3)!/[2^{n-2}(n-2)!]$ gene tree topologies with n leaves, for species tree topologies with $n=3$, $n=4$, $n=5$, and $n=6$. For each species tree topology shown, the matching topology is the unique gene tree topology with the largest number of coalescent histories, and the "1" that appears in the bottom nonzero line corresponds to the matching gene tree topology.

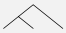
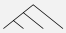
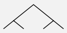
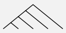
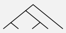
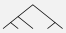
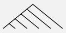
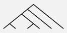
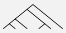
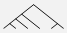
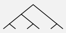
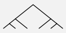
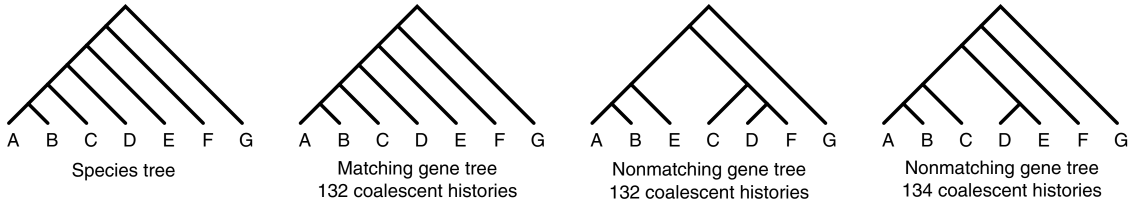
Number of coalescent histories	Number of gene tree topologies											
												
1	2	6	10	24	24	54	120	120	120	336	336	450
2	1	4	4	18	24	26	96	96	144	186	240	216
3		4	0	24	28	14	144	168	180	168	188	132
4		0	1	12	12	4	96	96	132	84	60	40
5		1		4	10	2	66	48	90	20	44	12
6				8	0	4	72	96	66	64	28	56
7				4	4	0	24	28	52	14	8	0
8				4	0	0	36	40	20	26	12	0
9				4	2	0	48	76	48	14	10	22
10				1	0	1	50	32	20	6	10	8
11				1	0		14	16	8	2	0	0
12				0	0		44	32	18	8	0	0
13				0	1		8	4	8	0	2	0
14				1			24	24	10	6	4	0
15							20	20	6	0	0	8
16							8	14	0	4	0	0
17							2	0	4	0	0	0
18							6	8	6	4	2	0
19							10	10	4	0	0	0
20							12	0	2	1	0	0
21							4	0	0	0	0	0
22							4	0	0	1	0	0
23							6	8	0	0	0	0
24							2	0	3	0	0	0
25							2	0	0	0	0	1
26							7	0	1	0	1	
27							4	0	2	0		
28							8	4	0	1		
29							0	2	0			
30							0	0	0			
31							1	0	0			
32							2	0	0			
33							1	0	0			
34							0	0	0			
35							1	0	0			
36							0	0	0			
37							2	2	1			
38							0	0				
39							0	0				
40							0	0				
41							0	0				
42							1	1				

Table 2
Properties of coalescent histories for $n \leq 9$ leaves.

Number of leaves	Number of unlabeled tree topologies (U)	Number of labeled tree topologies (L)	Number of pairs of topologies with a nonmatching gene tree ($UL-U$)	Number of pairs with more coalescent histories than the matching case	Number of unlabeled species tree topologies represented in pairs with more coalescent histories than the matching case	Maximal number of coalescent histories for a matching gene tree	Maximal number of coalescent histories for a nonmatching gene tree
3	1	3	2	0	0	2	1
4	2	15	28	0	0	5	3
5	3	105	312	0	0	14	11
6	6	945	5,664	0	0	42	37
7	11	10,395	114,334	1	1	138	137
8	23	135,135	3,108,082	71	7	462	558
9	46	2,027,025	93,243,104	1437	22	1663	2511



Species tree

Matching gene tree
132 coalescent histories

Nonmatching gene tree
132 coalescent histories

Nonmatching gene tree
134 coalescent histories

Fig. 5. Examples with $n = 7$ leaves in which a nonmatching gene tree topology has at least as many coalescent histories as the matching gene tree topology.

Table 3
Nonmatching gene tree topologies with numbers of coalescent histories greater than or equal to that of the matching gene tree topology, for $n = 8$ leaves.

Species tree topology	Number of coalescent histories for the matching gene tree topology	Number of nonmatching gene tree topologies with more coalescent histories than the matching topology	Number of nonmatching gene tree topologies with equally many coalescent histories as the matching topology	Maximal number of coalescent histories for a nonmatching gene tree topology	Gene tree topology that achieves the maximal number of coalescent histories
(((((((A, B), C), D), E), F), G), H)	429	42	0	549	(((((A, B), E), G), ((C, D), F)), H)
((((((A, B), (C, D)), E), F), G), H)	462	18	0	558	(((((A, B), E), G), ((C, D), F)), H) ((((A, B), F), (((C, D), E), G)), H) ((((A, B), (C, D)), ((E, F), G)), H)
((((((A, B), C), (D, E)), F), G), H)	453	5	1	506	(((((A, B), C), G), ((D, E), F)), H)
((((((A, B), C), D), (E, F)), G), H)	416	2	0	428	((((A, B), (C, D)), ((E, F), G)), H)
(((((A, B), (C, D)), (E, F)), G), H)	442	1	0	465	((((A, B), (C, D)), ((E, F), G)), H)
((((((A, B), C), D), E), (F, G)), H)	354	2	0	365	(((((A, B), C), (D, E)), (F, G)), H)
((((((A, B), C), D), E), F), (G, H))	264	1	1	268	(((((A, B), C), (D, E)), F), (G, H))

Table 4
Nonmatching gene tree topologies with numbers of coalescent histories greater than or equal to that of the matching gene tree topology, for $n = 9$ leaves.

Species tree topology	Number of coalescent histories for the matching gene tree topology	Number of nonmatching gene tree topologies with more coalescent histories than the matching topology	Number of nonmatching gene tree topologies with equally many coalescent histories as the matching topology	Maximal number of coalescent histories for a nonmatching gene tree topology	Gene tree topology that achieves the maximal number of coalescent histories
((((((((A, B), C), D), E), F), G), H), I)	1430	865	3	2454	(((((A, B), E), G), (((C, D), F), H)), I)
(((((((A, B), (C, D)), E), F), G), H), I)	1573	233	0	2511	(((((A, B), (C, D)), H), ((E, F), G)), I)
(((((((A, B), C), (D, E)), F), G), H), I)	1584	74	0	2238	(((((A, B), C), G), (((D, E), F), H)), I)
(((((((A, B), C), D), (E, F)), G), H), I)	1511	56	0	1970	(((((A, B), (C, D)), H), ((E, F), G)), I)
((((((A, B), (C, D)), (E, F)), G), H), I)	1663	16	0	2106	(((((A, B), (C, D)), H), ((E, F), G)), I)
((((((A, B), C), ((D, E), F)), G), H), I)	1518	7	0	1784	(((((A, B), C), G), (((D, E), F), H)), I) (((((A, B), C), H), (((D, E), F), G)), I)
(((((((A, B), C), D), E), (F, G)), H), I)	1368	41	0	1726	(((((A, B), C), (D, E)), ((F, G), H)), I)
((((((A, B), (C, D)), E), (F, G)), H), I)	1488	17	0	1726	((((A, B), ((C, D), E)), ((F, G), H)), I) (((((A, B), E), (C, D)), ((F, G), H)), I)
((((((A, B), C), (D, E)), (F, G)), H), I)	1478	3	0	1630	(((((A, B), C), (D, E)), ((F, G), H)), I)
((((((A, B), C), D), (E, (F, G))), H), I)	1316	2	1	1384	((((A, B), (C, D)), ((E, (F, G)), H)), I)
(((((A, B), (C, D)), (E, (F, G))), H), I)	1408	2	0	1503	((((A, B), (C, D)), ((E, (F, G)), H)), I)
(((((((A, B), C), D), E), F), (G, H)), I)	1155	32	2	1452	(((((A, B), E), ((C, D), F)), (G, H)), I)
((((((A, B), (C, D)), E), F), (G, H)), I)	1248	13	0	1495	(((((A, B), (C, D)), (E, F)), (G, H)), I)
((((((A, B), C), (D, E)), F), (G, H)), I)	1229	3	0	1345	(((((A, B), C), ((D, E), F)), (G, H)), I)
((((((A, B), C), D), E), (F, (G, H))), I)	1020	2	0	1058	(((((A, B), C), (D, E)), (F, (G, H))), I)
(((((((A, B), C), D), E), F), G), (H, I))	858	42	0	1098	(((((A, B), E), G), ((C, D), F)), (H, I))
((((((A, B), (C, D)), E), F), G), (H, I))	924	18	0	1116	(((((A, B), E), G), ((C, D), F)), (H, I)) ((((A, B), F), (((C, D), E), G)), (H, I)) ((((A, B), (C, D)), ((E, F), G)), (H, I))
((((((A, B), C), (D, E)), F), G), (H, I))	906	5	1	1012	(((((A, B), C), G), ((D, E), F)), (H, I))
((((((A, B), C), D), (E, F)), G), (H, I))	832	2	0	856	((((A, B), (C, D)), ((E, F), G)), (H, I))
(((((A, B), (C, D)), (E, F)), G), (H, I))	884	1	0	930	((((A, B), (C, D)), ((E, F), G)), (H, I))
((((((A, B), C), D), E), (F, G)), (H, I))	708	2	0	730	(((((A, B), C), (D, E)), (F, G)), (H, I))
((((((A, B), C), D), E), F), (G, (H, I)))	660	1	1	670	(((((A, B), C), (D, E)), F), (G, (H, I)))

This collection of results on the number of coalescent histories for $n \leq 9$ leaves illustrates that, despite an increase in the proportion of gene tree topologies with relatively few coalescent histories, an increase occurs with n in the amount by which nonmatching gene tree topologies can exceed matching gene tree topologies in their numbers of coalescent histories, and an increase occurs in the number of cases in which nonmatching gene tree topologies produce more coalescent histories than the matching gene tree topology. The next section shows that, for $n \geq 7$, such cases of a nonmatching gene tree topology exceeding the matching gene tree topology in number of coalescent histories always exist.

4. Larger trees

The following theorem states that, for each $n \geq 7$, there exists a species tree topology for which some nonmatching gene tree topology produces more coalescent histories than the matching gene tree topology.

Theorem 4.1. *If and only if $n \geq 7$, there exist a species tree topology S and a gene topology $G \neq S$ with n leaves such that $B_{G,S,1} > B_{S,S,1}$.*

We consider a family of examples involving caterpillar species tree topologies and pseudocaterpillar gene tree topologies. We label by Z_n an n-leaf caterpillar topology for $n \geq 1$,

$$Z_n = ((\ldots((((A_1, A_2), A_3), A_4), A_5), \ldots), A_n).$$

For $n \geq 5$, we label an n-leaf pseudocaterpillar topology by Y_n,

$$Y_n = ((\ldots(((A_1, A_2), (A_3, A_4)), A_5), \ldots), A_n).$$

Suppose for $n \geq 5$ that a species tree has topology Z_n. The motivation for this choice in our set of examples is that the pseudocaterpillar gene tree topology Y_n can be produced by each of two different sequences of coalescence events (either gene lineages A_1 and A_2 coalesce first, or lineages A_3 and A_4 coalesce first), whereas a caterpillar gene tree topology Z_n can only be produced by one sequence of coalescences. Because it can arise in either of two sequences, gene tree topology Y_n potentially has more coalescent histories than gene tree topology Z_n. We will see that, for sufficiently large n, the effect of the two possible sequences for gene tree topology Y_n eventually outweighs the increase in the number of coalescent histories for gene tree topology Z_n caused by its identity to the species tree topology. The transition point occurs at $n = 9$ leaves, where nonmatching gene tree topology Y_n has 1441 coalescent histories and matching gene tree topology Z_n has 1430 coalescent histories.

Lemma 4.2. *If $n \geq 9$ and $m \geq 1$, then $B_{Y_n,Z_n,m} > B_{Z_n,Z_n,m}$.*

Proof. We first count m-extended coalescent histories $B_{Y_n,Z_n,m}$ when $n = 5$. Label the edges of species tree topology Z_n in increasing order from most recent to most ancient (Fig. 3). To obtain gene tree topology Y_n when the species tree topology is Z_n, the final coalescence joining $((A_1, A_2), (A_3, A_4))$ to A_5 can occur on any of the m edges above the root of Z_n. If this coalescence occurs on edge k $(4 \leq k \leq m + 3)$, then the coalescence of (A_1, A_2) with (A_3, A_4) can occur on any edge ℓ with $3 \leq \ell \leq k$. The coalescence of A_1 and A_2 can then occur on any of the ℓ edges from 1 to ℓ, and the coalescence of A_3 and A_4 can occur on any of the $\ell - 2$ edges from 3 to ℓ. Therefore,

$$\begin{aligned} B_{Y_5,Z_5,m} &= \sum_{k=4}^{m+3} \sum_{\ell=3}^{k} \ell(\ell - 2) \\ &= \frac{1}{12} m(m^3 + 12m^2 + 47m + 72). \end{aligned} \tag{1}$$

Using the recursion for the number of m-extended coalescent histories (Rosenberg, 2007, Theorem 4.3), together with the observation that, for all m, $B_{Z_1,Z_1,m}$ trivially equals 1 because no coalescences take place in a 1-leaf gene tree, we have

$$\begin{aligned} B_{Y_{n+1},Z_{n+1},m} &= \sum_{k=1}^{m} B_{Y_n,Z_n,k+1} B_{Z_1,Z_1,k+1} \\ &= \sum_{k=1}^{m} B_{Y_n,Z_n,k+1}. \end{aligned} \tag{2}$$

Applying this equation four times, $B_{Y_9,Z_9,m}$ can be written in terms of $B_{Y_5,Z_5,m}$. Using Eq. (1), we obtain

$$B_{Y_6,Z_6,m} = \frac{1}{120} m(2m^4 + 45m^3 + 380m^2 + 1515m + 2498) \tag{3}$$

$$\begin{aligned} B_{Y_7,Z_7,m} &= \frac{1}{720} m(m + 7) \\ &\quad \times (2m^4 + 58m^3 + 634m^2 + 3302m + 7164) \end{aligned} \tag{4}$$

$$\begin{aligned} B_{Y_8,Z_8,m} &= \frac{1}{5040} m(m + 8)(m + 9) \\ &\quad \times (2m^4 + 71m^3 + 952m^2 + 6109m + 16386) \end{aligned} \tag{5}$$

$$\begin{aligned} B_{Y_9,Z_9,m} &= \frac{1}{40320} m(m + 9)(m + 10)(m + 11) \\ &\quad \times (2m^4 + 84m^3 + 1334m^2 + 10164m + 32432). \end{aligned} \tag{6}$$

Theorem 3.4 of Rosenberg (2007) provides the number of m-extended coalescent histories in the case in which the gene tree and species tree have the same caterpillar topology:

$$\begin{aligned} B_{Z_9,Z_9,m} &= \frac{1}{40320} m(m + 9)(m + 10)(m + 11)(m + 12) \\ &\quad \times (m + 13)(m + 14)(m + 15). \end{aligned}$$

We then have

$$\begin{aligned} B_{Y_9,Z_9,m} - B_{Z_9,Z_9,m} &= \frac{1}{40320} m(m + 9)(m + 10)(m + 11) \\ &\quad \times (m^4 + 30m^3 + 243m^2 + 390m - 328), \end{aligned}$$

from which it follows that $B_{Y_9,Z_9,m} > B_{Z_9,Z_9,m}$ for $m \geq 1$.

We now show by induction on n that $B_{Y_n,Z_n,m} > B_{Z_n,Z_n,m}$ for $n \geq 9$ and $m \geq 1$. Using the recursion for the number of m-extended coalescent histories (Rosenberg, 2007, Theorem 4.3) and the inductive hypothesis,

$$\begin{aligned} B_{Y_{n+1},Z_{n+1},m} &= \sum_{k=1}^{m} B_{Y_n,Z_n,k+1} \\ &> \sum_{k=1}^{m} B_{Z_n,Z_n,k+1} \\ &= B_{Z_{n+1},Z_{n+1},m}. \quad \square \end{aligned}$$

Proof of Theorem 4.1. Given the results in Section 3 for $n \leq 8$ leaves, it suffices to produce for each $n \geq 9$ an example in which a nonmatching gene tree topology leads to more coalescent histories than the matching gene tree topology. Taking $m = 1$ in Lemma 4.2, the case in which the species tree topology is Z_n and the nonmatching gene tree topology is Y_n is such an example for each $n \geq 9$. □

We can examine the relationship between the number of coalescent histories $B_{Y_n,Z_n,1}$ for the nonmatching pseudocaterpillar gene tree topology and the number of coalescent histories $B_{Z_n,Z_n,1}$ for the matching caterpillar (Table 5). $B_{Z_n,Z_n,1}$ has been shown to equal the Catalan number $C_{n-1} = (2n-2)!/[n!(n-1)!]$ (Degnan, 2005; Rosenberg, 2007). Using a similar approach to that employed by Rosenberg (2007) in obtaining $B_{Y_n,Y_n,m}$, it can be shown that, for $n \geq 6$,

$$\begin{aligned} B_{Y_n,Z_n,m} &= \frac{mF(m, n)}{(n-1)(n-2)(n-3)(n-4)(n-5)} \\ &\quad \times \binom{m + 2n - 7}{n - 6}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} F(m, n) &= 2m^4 + (13n - 33)m^3 + (32n^2 - 162n + 200)m^2 \\ &\quad + (38n^3 - 288n^2 + 705n - 555)m \\ &\quad + (19n^4 - 192n^3 + 705n^2 - 1110n + 626). \end{aligned}$$

The proof proceeds by induction on n, starting from Eq. (3) as the base case. The inductive hypothesis is that Eq. (7) holds for a given $n \geq 6$ and all m. Inserting the inductive hypothesis into Eq. (2), we evaluate $B_{Y_{n+1},Z_{n+1},m}$. Note that

Table 5
Numbers of coalescent histories for the pseudocaterpillar gene tree topology Y_n and the caterpillar gene tree topology Z_n when the species tree topology is Z_n.

n	Number of coalescent histories	
	Pseudocaterpillar	Caterpillar
5	11	14
6	37	42
7	124	132
8	420	429
9	1,441	1,430
10	5,005	4,862
11	17,576	16,796
12	62,322	58,786
13	222,870	208,012
14	802,978	742,900
15	2,912,168	2,674,440

$$\frac{mF(m,n)}{(n-1)(n-2)(n-3)(n-4)(n-5)}\binom{m+2n-7}{n-6}$$
$$=2\binom{m+2n-2}{n-1}-7\binom{m+2n-3}{n-2}$$
$$+8\binom{m+2n-4}{n-3}-2\binom{m+2n-5}{n-4}$$
$$-5\binom{m+2n-6}{n-5}+2\binom{m+2n-7}{n-6}. \qquad (8)$$

With this decomposition, we can apply Lemma 3.6 of Rosenberg (2007) to each of the six terms in Eq. (8) to evaluate the sum $\sum_{k=1}^{m} B_{Y_n,Z_n,k+1}$, assuming the inductive hypothesis. Simplifying, we obtain

$$B_{Y_{n+1},Z_{n+1},m}=\frac{mF(m,n+1)}{n(n-1)(n-2)(n-3)(n-4)}\binom{m+2n-5}{n-5},$$

which completes the proof.

By choosing $m=1$ in Eq. (7), it then follows that

$$B_{Y_n,Z_n,1}=\frac{(19n-40)(n-3)}{4(2n-3)(2n-5)}C_{n-1}. \qquad (9)$$

Thus, recalling that $B_{Z_n,Z_n,1}=C_{n-1}$, for large n, the pseudocaterpillar gene tree topology Y_n exceeds the matching caterpillar gene tree topology Z_n in coalescent histories by approximately a factor of 19/16.

5. Discussion

We have shown that, as the number of leaves increases, it becomes possible that a nonmatching gene tree topology has more coalescent histories than the matching gene tree topology. If the number of coalescent histories is viewed as the number of ways that a gene tree topology can evolve on a species tree topology, then this result means that the number of ways that a nonmatching gene tree topology can evolve sometimes exceeds the number of ways that a matching topology can evolve. Considering this result in the context of our earlier work concerning the excess in *probabilities* of nonmatching gene tree topologies compared to the probability of the matching gene tree topology (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008), we see that extreme discordance of gene trees and species trees might arise partly from inherent properties of the combinatorial structure of trees, and it need not be due specifically to the particular probability model that we have previously investigated. This potential of extreme discordance to occur in general models supports the importance of taking such discordance into account in the design and evaluation of phylogenetic algorithms.

Our analysis leaves many questions unanswered about the combinatorics of coalescent histories. While we have examined the distributions of the number of coalescent histories for small species trees, we have only provided partial explanations of the features of these distributions. Although loose bounds on the number of coalescent histories for identical gene trees and species trees are presented by Rosenberg (2007), we have not provided an upper bound for the maximal number of coalescent histories for discordant trees. Also, while Lemma 4.2 demonstrates that, for caterpillar species tree topologies with $n \geq 9$ leaves, a certain nonmatching pseudocaterpillar gene tree topology has more coalescent histories than the matching caterpillar gene tree topology, we have not characterized the species trees for which there exists a nonmatching gene tree topology with more coalescent histories than the matching topology, nor have we characterized these nonmatching gene tree topologies.

More generally, it would be desirable to characterize the determinants of the number of coalescent histories for a gene tree topology and species tree topology, or to otherwise obtain strongly predictive summary statistics. Using a set of examples with 20 leaves, Than et al. (2007) computed the number of coalescent histories in scenarios in which the gene tree topology and species tree topology differed by some number of subtree-prune-and-regraft (SPR) rearrangements, finding that the number of coalescent histories tended to decline with an increasing number of rearrangements and with an increasing number of tree edges between the "prune" and "regraft" locations (the "event diameter"). However, neither the number of SPR events nor the SPR diameter produces a monotonic decrease in the number of coalescent histories, as is evident from the observation that matching gene tree topologies, which are not separated from species tree topologies by any SPR events, can have fewer coalescent histories than nonmatching gene tree topologies, which have nonzero SPR distance to the species tree topology. Thus, other summary statistics for obtaining further results on the properties of coalescent histories will be needed, both to aid in understanding the computational complexity of probability computations that rely on coalescent histories and, ultimately, to aid in improving algorithms for performing these computations.

Acknowledgments

We thank L. Nakhleh for assistance with PhyloNet and two reviewers for helpful comments. This work was supported by NSF grant DEB-0716904, by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, and by an Alfred P. Sloan Research Fellowship.

References

Degnan, J.H., 2005. Gene tree distributions under the coalescent process. Ph.D. Thesis. University of New Mexico, Albuquerque.
Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, 762–768.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.
Dutheil, J.Y., Ganapathy, G., Hobolth, A., Mailund, T., Uyenoyama, M.K., Schierup, M.H., 2009. Ancestral population genomics: The coalescent hidden Markov model approach. Genetics 183, 259–274.
Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3, 294–304.
Rosenberg, N.A., 2007. Counting coalescent histories. J. Comput. Biol. 14, 360–377.
Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Syst. Biol. 57, 131–140.
Than, C., Ruths, D., Innan, H., Nakhleh, L., 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14, 517–535.
Than, C., Ruths, D., Nakhleh, L., 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9, 322.