Supplementary Material

Gene tree discordance, phylogenetic inference and the multispecies coalescent

James H. Degnan^{1,2} and Noah A. Rosenberg^{1,3,4}

^{1}Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
^{2}Current address: Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
^{3}Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, USA
^{4}Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA
Corresponding authors: Degnan, J.H. (j.degnan@math.canterbury.ac.nz); Rosenberg, N.A. (rnoah@umich.edu).

Box S1. Probabilities for gene genealogies and gene trees

Although ‘gene tree’ is sometimes used to indicate only the branching pattern for a set of coalescences, the branch lengths might also be of interest. We use ‘gene genealogy’ to indicate the gene tree topology together with its branch lengths (equivalently, coalescent times). Because gene tree topologies are discrete entities, with only a finite number of topologies possible for any given number of taxa, each topology has a positive probability. Gene genealogies, however, include branch lengths measured with real numbers and are, therefore, described using a probability density. The joint density of the coalescent times and gene tree topology given a fixed species tree is described in Ref. [30]. Because densities for gene genealogies include branch length information, they are well-suited for species tree inference and for determining species tree likelihoods [6,7]. The density is also useful for estimating ancestral population sizes and divergence times. Discrete probabilities of gene trees given the species tree have also been derived [25], and are related to gene genealogy densities in that integrating over coalescent times in the density results in the discrete topology probabilities. Topology probabilities are useful for predicting gene tree discordance and the existence of AGTs, for designing simulations [56] and for inferring species trees using approximate methods [15,63].

The distribution for the gene genealogy with coalescent times can be obtained for a particular gene tree-species tree combination by considering the waiting time to the next coalescence within each population (branch of the species tree) to be exponentially distributed, where the rate depends on the number of lineages in the population. Each coalescence and each failure of two or more lineages to coalesce in a population contribute to the density. If there are i lineages at a given point in time in a population with effective population size N_e chromosomes, the waiting time in coalescent time units (going backwards in time) to the ‘next’ coalescence is exponential with rate $\binom{i}{2}/N_e$ (mutation units can be used by replacing N_e by $\theta/2$). The overall density is obtained by multiplying the contributions to the density from each population. The form of the density depends on the populations in which coalescences occur. This makes it difficult to integrate over coalescent times to obtain the probability of a gene tree topology.

Gene tree topology probabilities can be computed directly without using the gene genealogy density in Ref. [30] by summing over different cases, each of which represents a 'coalescent history.' The probability of the entire history is obtained by multiplying probabilities of events across all branches of the species tree. Summing over coalescent histories yields the probability for an entire gene tree topology given the species tree:

$$\begin{aligned}&\Pr(\textit{gene tree} \mid \textit{species tree})\\ &= \sum_{\textit{histories}} \Pr(\textit{history})\\ &= \sum_{\textit{histories}} \prod_{\textit{branches}} \Pr(\textit{coalescences on branch})\end{aligned}$$

The probability of a given set of coalescences on a branch can be computed using the function $g_{ij}(t)$, which gives the probability that i lineages coalesce to j lineages within t coalescent time units [S1]. The $g_{ij}(t)$ terms must be multiplied by the probability that the $i-j$ coalescences occurring in a population are compatible with the gene tree topology. For example, in Figure SIb, no coalescences occur in populations 1 and 2, so these branches contribute $g_{22}(t_1)$ and $g_{22}(t_2)$ to the probability. Two coalescences occur on branch 3; because there are four lineages entering the branch, and two exiting it (moving from the present to the past), this branch contributes $(2/18)g_{42}(t_3)$ to the probability. The coefficient 2/18 arises from the assumption that all coalescences are equally likely; there are $\binom{4}{2}\binom{3}{2}=18$ ways that two coalescences could occur, and only two of these 18 (either B and C followed by A and D, or vice versa) are compatible with the gene tree topology. After the two coalescences in population 3, there are three gene lineages in population 4 — (BC), (AD), and E — and the probability is 1/3 that the first coalescence in population 4 joins E and (BC). Combining these probabilities, the probability of this coalescent history is the expression in Figure SIb.

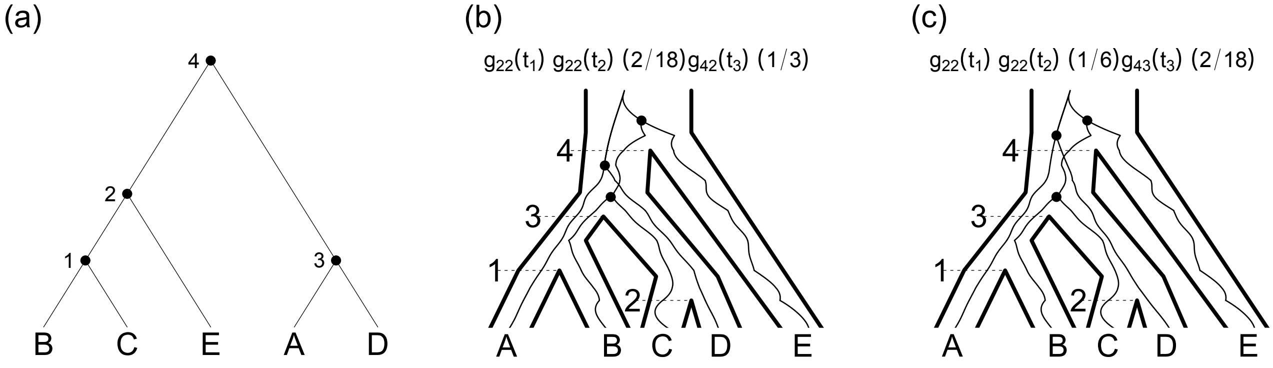
The number of coalescent histories for a given combination of a gene tree and species tree depends on the tree topologies. Gene trees that match the species tree typically (but not always) have more coalescent histories. When the gene tree matches the species tree, trees that are more balanced typically have fewer coalescent histories, and therefore it takes less time to compute their probabilities. When the gene tree and species tree match and they have the same pectinate n–taxon topology, the number of coalescent histories is the Catalan number, $(2n-1)!/[n!(n-1)!]$ [25,77]; more generally, recursive methods for counting and enumerating coalescent histories appear in Refs [77,78].

In addition to being useful for computing gene tree probabilities, coalescent histories can be used in coalescent hidden Markov models (HMMs) that investigate parameters such as ancestral population sizes and divergence times [39]. States of the Markov process are coalescent histories, and the HMM describes transition probabilities for different sites along the genome to change coalescent histories via recombination. Assuming the species tree topology is known, the number of states in this type of HMM is $\sum_k H_k$, where *k* ranges over all possible rooted binary gene tree topologies and H_k is the number of coalescent histories for the gene tree-species tree pair.

Online Supplemental References

1 Tavaré, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* 26, 119-164

Figure SI. Example gene tree in a species tree. Solid circles represent coalescences. **(a)** depicts an example gene tree, and **(b)** and **(c)** each show a different coalescent history, (3,4,3) in (b) and (3,4,4) in (c). The coalescent history in (b) is (3,4,3) because nodes 1 and 3 on the gene tree correspond to coalescences in branch 3 of the species tree. The second element of the coalescent history is 4 in both (b) and (c) because E coalesces with B and C in branch 4 on the species tree. Probabilities of coalescent histories are also shown, where t_i is the length in coalescent time units of branch i . Two other coalescent histories are possible for this combination of gene tree and species tree: (4,4,3) and (4,4,4).


(a)
4
2
1
3
B C E A D
(b)
$g_{22}(t_1)$ $g_{22}(t_2)$ $(2/18)g_{42}(t_3)$ $(1/3)$
4
3
1
2
A B C D E
(c)
$g_{22}(t_1)$ $g_{22}(t_2)$ $(1/6)g_{43}(t_3)$ $(2/18)$
4
3
1
2
A B C D E

Online Supplementary Material Box 1, Figure I.