Syst. Biol. 58(1):35–54, 2009

DOI:10.1093/sysbio/syp008

Properties of Consensus Methods for Inferring Species Trees from Gene Trees

JAMES H. DEGNAN[1,4,*], MICHAEL DEGIORGIO[2], DAVID BRYANT[3], AND NOAH A. ROSENBERG[1,2]

[1]*Department of Human Genetics, 1241 East Catherine Street, University of Michigan, Ann Arbor, MI 48109-0618, USA;*
[2]*Center for Computational Medicine and Biology, 2017 Palmer Commons, 100 Washtenaw Avenue, University of Michigan, Ann Arbor, MI 48109-2218, USA;*
[3]*Department of Mathematics, University of Auckland, Private Bag 29019, Auckland, New Zealand;*
[4]*Present address: Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand;*
**Correspondence to be sent to: Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand; E-mail: J.Degnan@math.canterbury.ac.nz.*

Abstract.—Consensus methods provide a useful strategy for summarizing information from a collection of gene trees. An important application of consensus methods is to combine gene trees to estimate a species tree. To investigate the theoretical properties of consensus trees that would be obtained from large numbers of loci evolving according to a basic evolutionary model, we construct consensus trees from rooted gene trees that occur in proportion to gene-tree probabilities derived from coalescent theory. We consider majority-rule, rooted triple (R^*), and greedy consensus trees obtained from known, rooted gene trees, both in the asymptotic case as numbers of gene trees approach infinity and for finite numbers of genes. Our results show that for some combinations of species-tree branch lengths, increasing the number of independent loci can make the rooted majority-rule consensus tree more likely to be at least partially unresolved. However, the probability that the R^* consensus tree has the species-tree topology approaches 1 as the number of gene trees approaches ∞. Although the greedy consensus algorithm can be the quickest to converge on the correct species-tree topology when increasing the number of gene trees, it can also be positively misleading. The majority-rule consensus tree is not a misleading estimator of the species-tree topology, and the R^* consensus tree is a statistically consistent estimator of the species-tree topology. Our results therefore suggest a method for using multiple loci to infer the species-tree topology, even when it is discordant with the most likely gene tree. [Anomalous gene tree; coalescence; discordance; lineage sorting; phylogenetics; statistical consistency.]

The goal of many phylogenetic and phylogeographic studies is not the estimation of the individual gene trees but rather the estimation of the species-level phylogeny or population history (Felsenstein 1988; Nei and Kumar 2000). Among methods that have been used to estimate species trees from data on multiple loci, a popular approach has been to make use of sequences concatenated across the loci. In essence, this approach assumes that all loci have the same gene tree, whose estimate is also used as the estimated species tree. Because gene trees vary both locally and across broad regions of organismal genomes (Chen and Li 2001; Pollard et al. 2006; Hobolth et al. 2007), sequence data from multiple genes are expected to be the result of heterogeneous processes. Multilocus data can be regarded as mixtures generated from different branch lengths and mutation rates on gene trees as well as from different gene-tree topologies that may arise from sources such as incomplete lineage sorting or hybridization.

As a result of these various sources of heterogeneity, concatenation can perform poorly when sequences are analyzed as if they come from a single model. Inferences may be inconsistent (Kolaczkowski and Thornton 2004), or the mixture generating the sequences might not be identifiable even when sites are generated from the same topology (Matsen and Steel 2007). Similarly, when sites are generated from different topologies but under the same mutation model, inferences from concatenated data can be misleading (Mossel and Vigoda 2005; Edwards et al. 2007; Kubatko and Degnan 2007). It is therefore useful to examine the behavior of other approaches in situations with a high level of gene-tree discordance.

One strategy for estimating species trees that does not assume that all loci reflect the same underlying gene tree is consensus, in which trees are obtained from individual loci and are then summarized in the form of a single tree. The consensus idea underlies such methods as concordance trees (Ané et al. 2007; Baum 2007), in which the support for each potential clade is estimated from multiple loci, and the most strongly supported clades are placed in a single "concordance tree," adjusting for a prior distribution on the number of distinct gene trees that a sample is predicted to contain. In the absence of processes such as horizontal gene transfer and recombination within loci, the concordance tree can be considered an estimate of the species tree. The concordance factor approach allows statistical dependence between loci but does not assume any particular mechanism for gene-tree conflict.

Relatively little is known about the theoretical properties of consensus algorithms applied to trees from multiple loci. We consider a simplified setting in which gene-tree discordance of rooted gene trees is due solely to incomplete lineage sorting. We examine 3 consensus algorithms applied to loci that are independent given a fixed species tree. In particular, we ask the question: as the number of gene trees considered from different loci increases, what is the probability that the consensus tree matches the species-tree topology?

We focus on majority-rule, rooted triple (R^*), and greedy consensus trees. A survey of these and other consensus methods can be found in Bryant (2003). We note that because the coalescent approach that we models probabilities of rooted trees, we consider consensus algorithms applied to rooted trees only, although

in practice, majority-rule and greedy consensus are often applied to unrooted trees. For rooted trees, the majority-rule consensus tree consists of those clades that occur more than 50% of the time in the collection of trees. (For simplicity, we always use 50% as the cutoff when referring to majority-rule consensus, although any greater proportion could be used instead.)

The R^* consensus tree is constructed in 2 steps. The first step is to construct a list of "uniquely favored" rooted triples. A rooted triple (AB)C on 3 taxa is said to be uniquely favored if it appears in more trees than either of the other 2 rooted triples, (AC)B or (BC)A, on the same set of 3 taxa. Note that in the case of ties, there might not be a uniquely favored triple for some sets of taxa.

In the second step, the most resolved tree that contains only uniquely favored triples is constructed, for example, using the algorithm of Bryant and Berry (2001, Corollary 2.2). To illustrate, consider the rules for clades in the R^* tree in the case of 4 taxa:

1. Clades of sizes 1 and 4 are included automatically.
2. The set {XY} is a clade exactly when (XY)Z and (XY)W are uniquely favored.
3. The set {XYZ} is a clade exactly when (XY)W, (XZ)W, and (YZ)W are uniquely favored.

For a tree with n taxa, these rules can be generalized. Let S be the set of all n taxa. Let A be a subset of S where A has at least 2 and at most $n-1$ elements, and let $S \setminus A$ be the set of all taxa in S not in A.

1′. Clades of sizes 1 and n are included automatically.
2′. A is a clade exactly when for each pair of taxa, $A_i, A_j \in A$ with $i \neq j$, for every taxon $Z \in S \setminus A$, $(A_iA_j)Z$ is uniquely favored.

These 2 rules determine the list of clades in the R^* consensus tree. Once the clades are determined, the tree can be constructed from the list of clades.

Greedy consensus trees are constructed by sequentially adding one clade at a time, the most frequently occurring clade that is compatible with clades already included in the greedy consensus tree (breaking ties randomly). Greedy consensus trees are also sometimes called "majority rule extended" (Felsenstein 1995), and the greedy consensus algorithm is implemented in PHYLIP (Felsenstein 1995) and PAUP* (Swofford 2003). Primary concordance trees (Baum 2007) can use the greedy consensus approach, where the concordance factor of a clade (or a split for unrooted trees) is the proportion of trees which exhibit the clade (split). We stress, however, that we only consider consensus algorithms applied to rooted trees.

For a given set of input trees, the greedy and R^* consensus trees are always resolutions of the majority-rule tree (Bryant 2003), meaning that every clade on the majority-rule consensus tree is also on the greedy and R^* consensus trees; however, there might be clades on the greedy or R^* consensus trees that are proper subsets of unresolved clades on the majority-rule consensus tree.

As an example to illustrate the 3 consensus methods, suppose the input gene trees are (((AB)C)D), (((AD)C)B), (((BC)A)D), and (((CD)A)B). First we determine the R^* tree. For the taxa A, B, and C, (AC)B occurs twice, whereas (AB)C and (BC)A each occur once. Thus, (AC)B is uniquely favored. The rooted triple (AC)D is uniquely favored for the taxa A, C, and D. Thus, {AC} is a clade by rule 2. However, for the taxa A, B, and D, (AB)D and (AD)B each occur twice. Similarly (BC)D and (CD)B each occur twice. Thus, (AC)B and (AC)D are the only uniquely favored triples. No other groups satisfy rule 2 or 3; hence there are no other 2- or 3-taxon clades. Thus, the R^* consensus tree is the partially unresolved tree ((AC)BD). Furthermore, because no clade occurs more than 50% of the time, the majority-rule tree for this set of input trees is a star tree. For the greedy algorithm, the 3-taxon clades {ABC} and {ACD} each occur in 50% of the trees and the 2-taxon clades {AB}, {AD}, {BC}, and {CD} each occur in 25% of the trees. The greedy algorithm therefore first selects 1 of the 2 observed 3-taxon clades at random (because they are tied for being most probable) and then randomly selects 1 of the 2 remaining compatible 2-taxon clades (because they are also tied). The result is that the greedy consensus algorithm returns each of the original input trees with probability 1/4.

The previous example illustrates how the R^* method performs when there are ties for the most frequently occurring triple. However, incompatible triples can also arise when each set of 3 taxa has a uniquely favored triple. Such conflicts also result in unresolved trees. Suppose the input trees are (((AB)C)D), (((AB)D)C), and ((AD)(BC)). Each of the rooted triples (AB)C, (AB)D, (AD)C, and (BC)D occurs in 2 of the 3 input trees, and therefore each is uniquely favored. These triples are incompatible because any resolved tree with the triples (AB)C and (BC)D also has (AC)D (Ranwez et al. 2007), but the input trees have (AD)C as a uniquely favored triple. The R^* consensus tree is ((AB)CD) and is unresolved with respect to the taxa A, C, and D because {AB} is the only 2-taxon clade which satisfies rule 2 and no 3-taxon clade satisfies rule 3. For these input trees, the majority-rule consensus tree also is ((AB)CD) because only the {AB} clade occurs more than 50% of the time. The greedy consensus tree is either (((AB)C)D) or (((AB)D)C), each with 50% probability.

The 3 consensus methods considered in this paper exhibit different behaviors as the number of genes increases. We find that when gene-tree discordance is due to incomplete lineage sorting, adding genes can increase the probability that the majority-rule consensus tree is unresolved. However, this unresolved tree is compatible with the species tree in the sense that one of its resolutions has the species-tree topology. We call sets of species-tree branch length vectors leading to this lack of resolution "unresolved zones." Also, as the number of independent, known gene trees increases, the R^* tree is more likely to be fully resolved and to match the species tree, although the R^* algorithm can be slow to converge to the correct tree. Greedy consensus trees, which are always resolved, often converge to the species-tree

topology more quickly; however, they can be misleading in the sense that adding more genes can make the greedy consensus tree less likely to match the species tree. We use the term "too-greedy zone" to denote the set of species-tree branch length vectors for which greedy consensus trees constructed from infinitely many loci disagree with the species tree. This concept is analogous to the "anomaly zone" (Degnan and Rosenberg 2006), the set of species-tree branch length vectors for which the most probable gene tree does not match the species tree. For 4-taxon asymmetric species trees, the too-greedy zone is a subset of the anomaly zone.

To illustrate the properties of the 3 consensus algorithms, we begin with 4-taxon examples of consensus trees when the number of loci approaches infinity. This section is followed by more general derivations for 4-taxon trees of the unresolved zones for majority-rule consensus trees and the too-greedy zone for greedy consensus trees. We then consider the same consensus methods with finitely many loci sampled, applying them to examples with 3 and 4 taxa. The main theoretical results of the paper (Theorems 1–4) apply to rooted, bifurcating species trees with any number of taxa and give different results for the limiting behavior of the 3 consensus methods used, assuming that gene trees are generated under the multispecies coalescent model. Proofs of the theorems are given in Appendices 1–3.

Assumptions and Definitions

We use the term "multispecies coalescent" for the model in which coalescent processes occur in each branch of a species tree and for which all possible coalescent events within a branch are equally likely. This is the model that has previously been used to calculate probabilities of gene trees given species trees (Tajima 1983; Pamilo and Nei 1988; Takahata 1989; Rosenberg 2002; Degnan and Salter 2005). The model assumes that population sizes are constant within species-tree branches (although not necessarily across branches) and that populations are panmictic. It also assumes that the genes from the different species are orthologous, that natural selection is not acting on the genes of interest, and that horizontal gene transfer and recombination do not occur within the genes of interest.

We use "gene tree" to refer to a gene-tree topology and "species tree" to refer to a species-tree topology with internal branch lengths specified. Because 2 or more lineages in a population are needed for a coalescence to occur, lengths of external branches of the species tree (branches leading to the tips) do not affect probabilities of gene-tree topologies when only one lineage is considered per species. Branch lengths on species trees are measured in "coalescent units", the number of generations divided by the effective population size (twice the effective population size for diploids; Hein et al. 2005).

Nodes on gene trees correspond to coalescent events. For example, if a node on a gene tree is the root of the subtree ((AB)C), then this node corresponds to the coalescent event that joins the lineage ancestral to (AB) with the lineage ancestral to C, where (AB) itself represents the coalesced lineage combining the lineages from taxa A and B. Clades with only 2 taxa (on either species or gene trees) are called "cherries." We use the same letter (such as A, B, etc.) to refer both to a taxon and to the gene lineage sampled from that taxon.

We use the notation (AB)C for the 3-taxon statement (rooted triple) that the most recent common ancestor (MRCA) of gene lineages A and B is not an ancestor of lineage C. This notation is similar to the notation for a 3-taxon tree but does not have the outer set of parentheses. If a given species tree (with topology and internal branch lengths specified) is σ, then $P_\sigma[\cdot]$ indicates probabilities of events for gene lineages when σ is the species tree. For example, P_σ[(AB)C] and P_σ[((AB)C)] are used to indicate the probabilities of the rooted triple (AB)C and the gene tree ((AB)C), respectively. The expression P_σ[{ABC}] is used to denote the probability that {ABC} is a clade on the gene tree.

Asymptotic Consensus Trees

Consensus trees are used to summarize a set of trees defined on the same set of taxa. A consensus algorithm takes the trees as input, so that the method of producing the input trees is not part of the consensus algorithm. Typically, the trees summarized might be estimated trees such as those that are obtained from separate genes, different models, or different bootstrap resamples. In all these cases, the consensus tree is a function of some data set and is therefore a statistic (Casella and Berger 1990).

Using gene-tree probability distributions, we can also compute the consensus tree that would be returned in the limit as the number of input gene trees approaches infinity. This calculation assumes that input gene trees are correctly estimated, independent, and generated by the multispecies coalescent model. In this setting, the proportion of occurrences for a gene-tree topology asymptotically approaches its probability under the multispecies coalescent model as the sample size (the number of independent loci) approaches infinity.

Consensus trees obtained from these asymptotic proportions are not functions of data and are therefore not statistics. Instead, they are properties solely of gene-tree probability distributions. These distributions in turn are functions of the species tree, which we can consider to be a parameter for a gene-tree distribution (Degnan and Salter 2005). Intuitively, we can also think of a consensus tree computed from gene-tree probabilities under the multispecies coalescent model as the consensus tree that would be obtained from an infinite number of independent, correctly inferred gene trees.

We define an "asymptotic consensus tree" for a species tree to be the tree topology that would be obtained if a consensus algorithm had considered gene trees in proportion to their probabilities (under the multispecies coalescent model). Under the multispecies coalescent model, every gene-tree topology has positive probability given

any species tree, and therefore every gene tree is included in the consensus algorithm. Consequently, methods such as Adams and strict consensus (Bryant 2003; Felsenstein 2004)—which preserve information shared by all input trees—result in star trees when probabilities under the multispecies coalescent model are used. Similarly, methods in which a single input tree can "veto" clades on the consensus tree (e.g., combinable component consensus; Bremer 1990; Felsenstein 2004) would necessarily result in star trees asymptotically. We therefore focus on 3 consensus algorithms that do not require strict agreement. Other consensus methods that can tolerate a high level of noise in the input trees, but which we have not investigated, such as matrix representation with parsimony (Baum 1992; Ragan 1992) could also be of interest in this setting.

The majority-rule asymptotic consensus tree (MACT) can be determined by listing the probability of monophyly for each subset of taxa. If a subset of taxa appears on the list with probability greater than 1/2, then that group is contained in the MACT. This is the same method traditionally used to determine majority-rule consensus trees, but here we use theoretical probabilities rather than observed proportions.

Similarly, the R^* asymptotic consensus tree (RACT) for n taxa can be determined by calculating the probability of each of the 3 possible rooted triples for each of the $\binom{n}{3}$ subsets of 3 taxa. The RACT then consists of those rooted triples that have the highest probability for each subset of 3 taxa. For any 3 taxa and any strictly bifurcating species tree, the rooted triple corresponding to the species tree is always the most probable (see Proposition 5 in Appendix 1)—that is, there are no ties. The set of rooted triples for all $\binom{n}{3}$ subsets of 3 taxa uniquely identifies the species tree (Steel 1992, Proposition 4); thus, the RACT is always fully resolved under the multispecies coalescent model.

The greedy asymptotic consensus tree (GACT) for n taxa can be obtained by ranking probabilities of the $2^n - n - 1$ clades with 2 or more taxa. The most probable clade is incorporated into the consensus tree, and then the list of clade probabilities is updated by removing any clades incompatible with those already in the tree. This process is repeated until the tree is fully resolved, randomly picking clades in the case of ties.

The 3 types of asymptotic consensus trees—MACT, RACT, and GACT—are purely mathematical functions of gene-tree probabilities. They are therefore properties of species trees. Consensus trees constructed from finitely many loci under different consensus algorithms are random variables and are increasingly likely to match their asymptotic counterparts as the number of loci approaches infinity.

Examples

Examples that illustrate the construction of asymptotic consensus trees for the 3 methods in this paper are shown in Table 1, which lists probabilities of each gene tree for 4 taxa, for several sets of branch lengths on the species tree in Figure 1*a*. Also listed are probabilities for 2- and 3-taxon clades and probabilities for the 12 rooted triples. For 4 taxa, there are 6 possible cherries and 4 possible 3-taxon monophyletic groups. Note that because some cherries are not mutually exclusive, the sum of probabilities over all cherries is more than 1. Also, because it is possible for a tree to not have any 3-taxon monophyletic groups, the sum of the probabilities for subsets of 3 taxa is less than 1.

For each of the examples in Table 1, majority-rule consensus returns 1 of the 4 trees illustrated in Figure 2*a*. Greedy consensus returns the matching tree for all examples in the table, except when $(x, y) = (0.05, 0.05)$, for which it returns ((AB)(CD)). This topology is also the most probable gene tree for those branch lengths. R^* consensus is the only consensus method considered which returns the matching tree for all branch lengths used. As we will see in Theorem 2, this result for R^* consensus is not limited to the examples chosen but applies to any branch lengths and any binary species tree.

As an example from the table, we see that if the species tree has topology (((AB)C)D) and has $x = 0.6$ and $y = 0.4$, then the clades {AB} and {ABC} both occur with probability greater than 1/2 and {CD} occurs with probability less than 1/2. Thus, the MACT for this species tree has the topology (((AB)C)D) because this is the only 4-taxon topology which has exactly the monophyletic groups {AB} and {ABC}. Both probabilities are only slightly larger than 1/2, however, so in a small sample of correctly inferred trees, it is possible that either {AB} or {ABC} would occur less than 50% of the time or that {CD} would occur more than 50% of the time. In these cases, the majority-rule consensus tree would be unresolved or would otherwise not match the species tree.

For the greedy consensus algorithm, we would first select the {AB} clade to be in the tree (because it is the most probable clade other than {ABCD}) and we would then eliminate all clades except {CD}, {ABC}, and {ABD} from consideration because these other clades are incompatible with {AB}. From among the 3 remaining clades, {ABC} is the most probable—hence, the GACT has clades {AB} and {ABC}, which means that (((AB)C)D) is the GACT. For the R^* consensus algorithm, the most probable rooted triples for each set of 3 taxa are (AB)C, (AB)D, (AC)D, and (BC)D. Because (((AB)C)D) is the only tree for taxa A, B, C, and D that is compatible with these rooted triples, R^* also returns the matching tree.

Choosing the branch lengths to be $(x, y) = (0.4, 0.6)$ (Table 1, second branch length column) illustrates that the behavior of MACTs is sensitive to the order of the branch lengths. Switching the lengths for x and y can change whether the MACT is fully resolved. For this tree, most gene trees (about 62%) are expected to have an {AB} clade, so this clade is very likely to be in the majority-rule consensus tree for a large enough number of gene trees; however, less than 46% of trees are expected to have {ABC} in a monophyletic group, so the MACT does not have {ABC} as a clade. Because no other group is monophyletic with probability greater than 1/2,

TABLE 1. Probabilities of 4-taxon gene-tree topologies, clades, and rooted triples for the species tree (((AB)C)D), with various sets of branch lengths. A clade (rooted triple) probability is the sum of probabilities of gene-tree topologies which have the clade (rooted triple). Branch lengths are as in the model species tree in Figure 1*a*

		Branch lengths (x, y)					
Gene-tree topology	Probability	(0.6, 0.4)	(0.4, 0.6)	(0.8, 0.3)	(0.3, 0.3)	(0.1, 0.1)	(0.05, 0.05)
1. (((AB)C)D)	p_1	0.316	0.319	0.321	0.212	0.104	0.079
2. (((AB)D)C)	p_2	0.109	0.144	0.087	0.122	0.091	0.075
3. (((AC)B)D)	p_3	0.107	0.069	0.140	0.081	0.066	0.061
4. (((AC)D)B)	p_4	0.049	0.043	0.048	0.058	0.062	0.060
5. (((AD)B)C)	p_5	0.006	0.009	0.004	0.017	0.037	0.045
6. (((AD)C)B)	p_6	0.006	0.009	0.004	0.017	0.037	0.045
7. (((BC)A)D)	p_7	0.107	0.069	0.140	0.081	0.066	0.061
8. (((BC)D)A)	p_8	0.049	0.043	0.048	0.058	0.062	0.060
9. (((BD)A)C)	p_9	0.006	0.009	0.004	0.017	0.037	0.045
10. (((BD)C)A)	p_{10}	0.006	0.009	0.004	0.017	0.037	0.045
11. (((CD)A)B)	p_{11}	0.006	0.009	0.004	0.017	0.037	0.045
12. (((CD)B)A)	p_{12}	0.006	0.009	0.004	0.017	0.037	0.045
13. ((AB)(CD))	p_{13}	0.115	0.153	0.094	0.139	0.128	0.121
14. ((AC)(BD))	p_{14}	0.055	0.052	0.052	0.075	0.099	0.105
15. ((AD)(BC))	p_{15}	0.055	0.052	0.052	0.075	0.099	0.105
Clade							
{AB}	$p_1 + p_2 + p_{13}$	0.541[a]	0.616[a]	0.499	0.473	0.322	0.275
{AC}	$p_3 + p_4 + p_{14}$	0.211	0.165	0.239	0.213	0.227	0.226
{AD}	$p_5 + p_6 + p_{15}$	0.067	0.071	0.059	0.108	0.174	0.196
{BC}	$p_7 + p_8 + p_{15}$	0.211	0.165	0.239	0.213	0.227	0.226
{BD}	$p_9 + p_{10} + p_{14}$	0.067	0.071	0.059	0.108	0.174	0.196
{CD}	$p_{11} + p_{12} + p_{13}$	0.128	0.171	0.098	0.172	0.202	0.212
{ABC}	$p_1 + p_3 + p_7$	0.530[a]	0.458	0.601[a]	0.373	0.236	0.201
{ABD}	$p_2 + p_5 + p_9$	0.121	0.162	0.094	0.155	0.165	0.166
{ACD}	$p_4 + p_6 + p_{11}$	0.061	0.061	0.055	0.091	0.136	0.151
{BCD}	$p_8 + p_{10} + p_{12}$	0.061	0.061	0.055	0.091	0.136	0.151
Rooted triple							
(AB)C	$p_1 + p_2 + p_5 + p_9 + p_{13}$	0.553	0.634	0.506	0.506	0.397	0.366
(AC)B	$p_3 + p_4 + p_6 + p_{11} + p_{14}$	0.223	0.183	0.247	0.247	0.302	0.317
(BC)A	$p_7 + p_8 + p_{10} + p_{12} + p_{15}$	0.223	0.183	0.247	0.247	0.302	0.317
(AB)D	$p_1 + p_2 + p_3 + p_7 + p_{13}$	0.755	0.755	0.778	0.634	0.454	0.397
(AD)B	$p_4 + p_5 + p_6 + p_{11} + p_{15}$	0.123	0.123	0.111	0.183	0.273	0.302
(BD)A	$p_8 + p_9 + p_{10} + p_{12} + p_{14}$	0.123	0.123	0.111	0.183	0.273	0.302
(AC)D	$p_1 + p_3 + p_4 + p_7 + p_{14}$	0.634	0.553	0.700	0.506	0.397	0.366
(AD)C	$p_2 + p_5 + p_6 + p_9 + p_{15}$	0.183	0.223	0.150	0.247	0.302	0.317
(CD)A	$p_8 + p_{10} + p_{11} + p_{12} + p_{13}$	0.183	0.150	0.247	0.223	0.302	0.317
(BC)D	$p_1 + p_3 + p_7 + p_8 + p_{15}$	0.634	0.553	0.700	0.506	0.397	0.366
(BD)C	$p_2 + p_5 + p_9 + p_{10} + p_{14}$	0.183	0.223	0.150	0.247	0.302	0.317
(CD)B	$p_4 + p_6 + p_{11} + p_{12} + p_{13}$	0.183	0.223	0.150	0.247	0.302	0.317

[a]Clade has probability greater than 1/2 and would therefore be represented in the MACT.

this MACT is not fully resolved and is ((AB)CD). Note that the lack of resolution is a theoretical limitation of majority-rule consensus and occurs even though the species tree and gene trees are fully resolved (there are no "hard" polytomies). Because asymptotic consensus trees use infinitely many resolved input trees, the lack of resolution is also not due to insufficient information—in other words, the lack of resolution cannot be overcome by adding more loci (there are no "soft" polytomies).

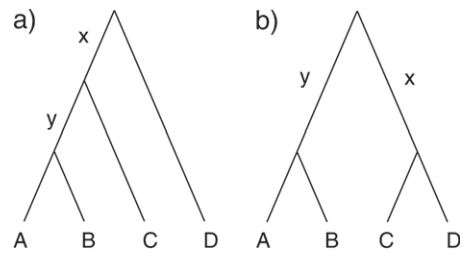

FIGURE 1. Four-taxon species trees with internal branch lengths x and y, measured in coalescent units.

When the branch lengths are $(x, y) = (0.8, 0.3)$ (Table 1, third branch length column), majority-rule consensus returns the partially resolved tree ((ABC)D). For the branch lengths $(x, y) = (0.3, 0.3)$, $(0.1, 0.1)$, $(0.05, 0.05)$ (columns 4 through 6), because no monophyletic subset of taxa has probability greater than 1/2, the MACTs for these species trees are star phylogenies. When the branch lengths are $(x, y) = (0.1, 0.1)$ and $(x, y) = (0.05, 0.05)$, ((AB)(CD)) is the most probable gene tree, although it does not match the species tree. Gene trees that are more probable than the gene tree matching the species tree are called "anomalous gene trees" (Degnan and Rosenberg 2006). When $(x, y) = (0.3, 0.3)$, no anomalous gene trees occur, illustrating that unresolved majority-rule consensus trees can arise even when there are no anomalous gene trees. When $(x, y) = (0.05, 0.05)$, the most probable clade is {AB}, which has probability 0.275, so it is included in the greedy consensus tree. The second most probable clade compatible with {AB}, however, is {CD}, which has probability 0.212, and thus the greedy consensus tree is ((AB)(CD)), which does not match the species tree. However, when $(x, y) = (0.1, 0.1)$,

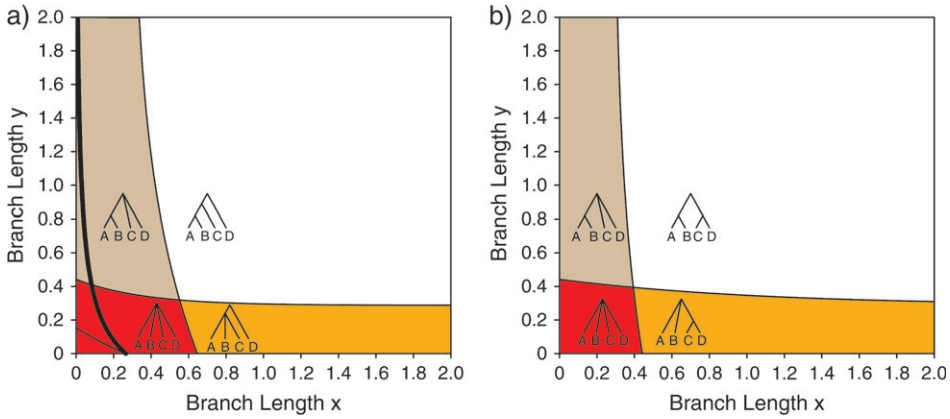

FIGURE 2. Unresolved zones for 4-taxon species trees. The shaded regions are different areas of the unresolved zones leading to different unresolved majority-rule consensus trees. Shaded regions represent values of x and y for which one of the inequalities (1–4) is violated. (*a*) The species tree is (((AB)C)D). A star tree is the limiting consensus tree for the red region, where conditions (1) and (2) both fail. The orange region corresponds to the tree with the {ABC} clade unresolved, which is where condition (1) fails. In the tan area to the left of the steeper of the 2 curves, inequality (2) is violated. For comparison, the anomaly zone is also plotted as the area under the heavy, dark curve. The anomaly zone cuts across 2 regions of the unresolved zone, and the area under the line starting from $(x, y) = (0, 0.154)$ which creates the approximately triangular region is the part of the anomaly zone with 3 anomalous gene trees. (*b*) The species tree is ((AB)(CD)). The unresolved zone in this case is similar in size to that of (*a*), but there is no anomaly zone for this species tree.

greedy consensus returns the matching tree (((AB)C)D), even though this tree is less probable than ((AB)(CD)). These examples show that different types of outcomes occur for majority-rule and greedy consensus trees, depending on the properties of the species tree. We now describe asymptotic consensus trees for more general sets of branch lengths for 3 and 4 taxa.

Majority-Rule Consensus

Three taxa.—For the case of 3-taxon trees, the MACT is resolved if the probability of the matching tree is greater than 1/2. The well-known probability of congruence for a gene tree given a 3-taxon species tree, $1 - (2/3)e^{-T}$ (e.g., Nei 1987) , where T is the length of the one internal branch, is greater than 1/2 if $T > \log(4/3) \approx 0.28768$. For smaller values of T, increasing the number of independent gene trees increases the probability that the trees do not produce a resolved majority-rule consensus tree, even though the matching gene tree is more likely than any other gene tree.

Four taxa.—For 4-taxon trees, the branch lengths needed for a clade to be in the MACT can be obtained by setting the probability of the clade to be greater than 1/2 and solving for branch length y in terms of branch length x. Clade probabilities are functions of gene-tree probabilities and are listed in Table 1. The model 4-taxon species trees are shown in Figure 1.

Details for deriving conditions for clades to be in the MACT are given in Appendices 4 and 5. First we consider the species tree with topology (((AB)C)D). Following Figure 1, let y be the length of the branch (in coalescent units) ancestral to A and B, but not C, and let x be the length of the other internal branch. Then {ABC} is a clade in the MACT if and only if

$$x > \log(4/3) \quad \text{and} \quad y > \log\left(\frac{2\,e^{2x} - 1}{3\,e^{3x} - 4\,e^{2x}}\right) \tag{1}$$

and {AB} is a clade in the MACT if and only if

$$y > \log\left(\frac{12\,e^{3x} + 2}{9\,e^{3x}}\right). \tag{2}$$

These 2 conditions partition the space of branch lengths into regions corresponding to the 4 possible MACTs for this species tree (Fig. 2*a*), where $x = \log(4/3) \approx 0.28768$ is a vertical asymptote and $y = \log(4/3)$ is a horizontal asymptote. The MACT is

(((AB)C)D)	if (1) and (2) both hold,
((ABC)D)	if (1) holds and (2) fails,
((AB)CD)	if (1) fails and (2) holds,
(ABCD)	if (1) and (2) both fail.

Similarly, if the species tree is ((AB)(CD)), with y denoting the length of the branch ancestral to (AB) and x denoting the length of the other internal branch, then {AB} is a clade in the MACT if and only if

$$y > \log\left(\frac{12\,e^{x} + 2}{9\,e^{x}}\right) \tag{3}$$

and {CD} is a clade in the MACT if and only if

$$x > \log(4/3) \quad \text{and} \quad y > \log\left(\frac{2}{9\,e^{x} - 12}\right). \tag{4}$$

These 2 conditions partition the branch length space into different regions for each of the possible MACTs (Fig. 2*b*), and $x = \log(4/3)$ and $y = \log(4/3)$ are vertical and horizontal asymptotes, respectively. The MACT is

((AB)(CD)) if (3) and (4) both hold,

((AB)CD) if (3) holds and (4) fails,

(AB(CD)) if (3) fails and (4) holds,

(ABCD) if (3) and (4) both fail.

Because inequalities (1–4) characterize all possible MACTs for 4 taxa, it follows that 4-taxon MACTs are never misleading in the sense that a 4-taxon MACT never has a clade that is not a clade in the species tree. Due to lack of resolution, however, the MACT may fail to have clades that are present in the species tree. Although we have obtained this result by explicit computation for the 4-taxon case, we will show that the result holds for larger trees (see Theorem 1 in Asymptotic Consensus Trees: General Theorems).

The plots in Figure 2 depict regions of parameter space in which MACTs are not fully resolved (and therefore do not fully recover the species tree). For the asymmetric 4-taxon tree, the anomaly zone from Degnan and Rosenberg (2006) is also depicted within the unresolved zone, which is considerably larger than the anomaly zone. For example, when we set $x = y$ for the 4-taxon asymmetric tree, the largest value of x that is still in the anomaly zone is approximately 0.1568 (Degnan and Rosenberg 2006); but for majority-rule consensus, when $x = y$, $x = y \approx 0.345$ is the largest value for which the MACT is fully unresolved and $x = y \approx 0.507$ is the largest value for which the MACT is partially unresolved, equaling ((AB)CD). For the symmetric 4-taxon tree, the line $x = y$ passes through the intersection of the 2 curves in Figure 2*b* and $x = y = 0.394$ is the largest value that results in a star consensus tree. This is somewhat surprising because these values result in the partially resolved tree ((AB)CD) for the asymmetric species tree, and the asymmetric species tree is typically more difficult to infer. For the asymmetric 4-taxon species tree, the anomaly zone is a subset of the zone in which the MACT is unresolved. For the symmetric species tree, the MACT can be unresolved but there is no anomaly zone. For 4 taxa, it is always true that if a species tree has an anomalous gene tree, then it does not have a fully resolved MACT.

R Consensus*

Three taxa.—In the case of 3 taxa, we note that the greedy and R^* algorithms are equivalent when there are infinitely many loci. (For finitely many loci, greedy and R^* consensus are not equivalent because they handle ties differently, with the R^* consensus tree sometimes being unresolved.) For both algorithms, the most frequently occurring clade determines a 3-taxon statement, and in the asymptotic case, there is a uniquely occurring most frequent tree. This tree has probability $1-(2/3)e^{-T} > 1/3$ (where T is the internal branch length), and the other 2 trees each have probability $(1/3)e^{-T} < 1/3$. Thus, for the 3-taxon case, as the number of loci approaches ∞, the probability that the matching gene tree is the most frequent approaches 1.

Four taxa.—We defer consideration of the case of 4 taxa to Asymptotic Consensus Trees: General Theorems.

Greedy Consensus

Three taxa.—Under the multispecies coalescent model, when there are 3 taxa, greedy consensus applied to gene trees is asymptotically guaranteed to result in the species tree as the number of gene trees increases. If the species tree has topology ((AB)C) and the one internal branch has length T, then a random gene tree has clade {AB} with probability $1 - (2/3)e^{-T} > 1/3$, whereas {AC} and {BC} each occur with probability less than 1/3. Thus, {AB} is always the most probable cherry for this topology and the GACT always matches the species-tree topology.

Four taxa.—For the 4-taxon symmetric species tree and for any choice of branch lengths, the GACT has the same topology as the species tree (Appendix 6). However, if the species tree is (((AB)C)D), then the GACT can be the symmetric tree ((AB)(CD)).

To find the set of branch lengths for which the GACT fails to match the asymmetric species-tree topology, let x and y be the lengths of the deeper and more recent internal branches, respectively, for the tree (((AB)C)D) (see Fig. 1*a*). For this species tree, the region where the GACT is ((AB)(CD)), the "too-greedy" zone, consists of those values of x and y for which the clade {CD} is more probable than the clade {ABC} (Appendix 7). The set of values of x and y for which $P(\{CD\}) > P(\{ABC\})$ is characterized by

$$y < \log\left[\frac{3\,e^{2x} - 2}{18(e^{3x} - e^{2x})}\right]. \qquad (5)$$

The right-hand side of this inequality is strictly less than the boundary of the anomaly zone for the species tree (((AB)C)D) (Degnan and Rosenberg 2006, Equation (4)); thus, for this species tree, the too-greedy zone is a subset of the anomaly zone (Fig. 3).

FINITE NUMBERS OF LOCI

Theory

An asymptotic consensus tree occurs in the limit as the number of loci approaches infinity. What happens with a finite number of loci? We can examine the behavior of consensus algorithms from a theoretical point of view by considering all possible finite samples of gene trees. The probability of a particular consensus tree is the sum of the probabilities of the samples of gene trees that result in that consensus tree. These probabilities can

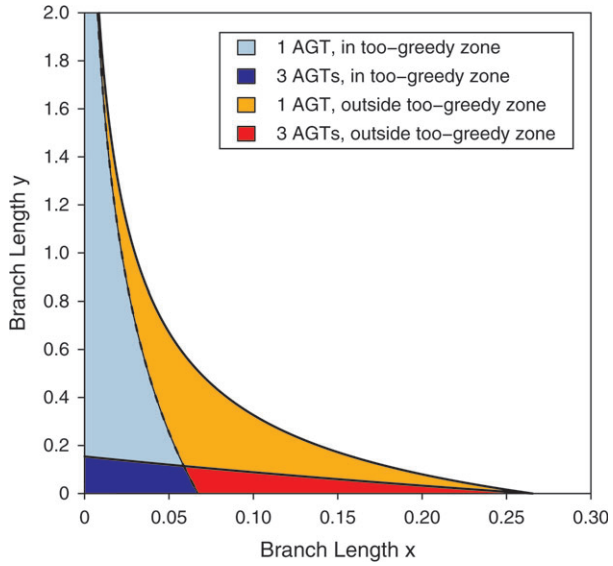

FIGURE 3. The too-greedy zone. The upper curve is the boundary of the anomaly zone for the species tree (((AB)C)D). For points below this curve, there is either one anomalous gene tree (AGT) or 3 AGTs. The 2 blue regions to the left of the curve that extends from roughly $(x, y) = (0.067, 0.0)$ to $(0.0078, 2.0)$ constitute the too-greedy zone, where the GACT is ((AB)(CD)).

be determined by noting that a sample of independent loci has a multinomial distribution, where the categories are the gene-tree topologies, and the probabilities are given by the multispecies coalescent model (Degnan and Salter 2005). Using this approach, at least for small numbers of taxa and loci, probabilities that consensus algorithms return a particular topology can be determined exactly without using simulation. Details for the method are given in Appendix 8.

Examples

Three taxa.—We illustrate the case of finite loci using 3 (Fig. 4) and 4 taxa (Figs 5 and 6). With 3 taxa, there is only one internal branch length, and this length determines all gene-tree probabilities, with the probability that the gene tree matches the species tree being $1-(2/3)e^{-T}$, where T is the length of the internal branch. We use ((AB)C) as the species tree, with branch lengths of 0.5, $\log(4/3) \approx 0.288$, and 0.1, corresponding to matching probabilities of 0.596, 0.5, and 0.397, respectively.

For the branch length of 0.5, most loci (almost 60%) are likely to have the matching topology; thus, given enough loci, all 3 methods (majority rule, R^*, and greedy) are expected to have a high probability of returning the matching tree. The greedy consensus algorithm has the highest probability of returning the matching tree for all sample sizes examined (up to 50 loci). The R^* method has the second-best performance, although by 50 loci, the greedy and R^* algorithms have roughly equivalent performance. When the branch length is chosen such that the probability of matching is 0.5 (Fig. 4*b*, with the 2 nonmatching trees each having probability 0.25), majority-rule consensus returns the correct tree at most 50% of the time (less for even sample sizes). This is not surprising because by design ((AB)C) has probability 1/2 and therefore is not likely to occur more than 50% of the time in a multinomial sample. For this case, as well as for the branch length of 0.1 (Fig. 4*c*), greedy consensus has the best performance and R^* slowly approaches greedy as the number of loci increases (and therefore the probability of ties decreases). Also, for the branch length of 0.1, no gene tree has probability greater than 50%, and therefore majority-rule consensus is increasingly likely to return a star tree as the number of loci increases.

Four taxa.—Figure 5 shows the behavior of the 3 consensus methods as the number of loci increases when the species tree is (((AB)C)D), and Figure 6 shows the corresponding results when the species tree is ((AB)(CD)). The 2 figures are similar, although the methods generally perform better with the symmetric species tree.

Figure 5*a* suggests that large numbers of loci might be needed before one majority-rule consensus tree becomes the most probable. Figures 5*b*,*c* and 6*b*,*c* show that majority-rule consensus can converge fairly quickly to a star phylogeny even though the probability of a star phylogeny decreases under R^* consensus.

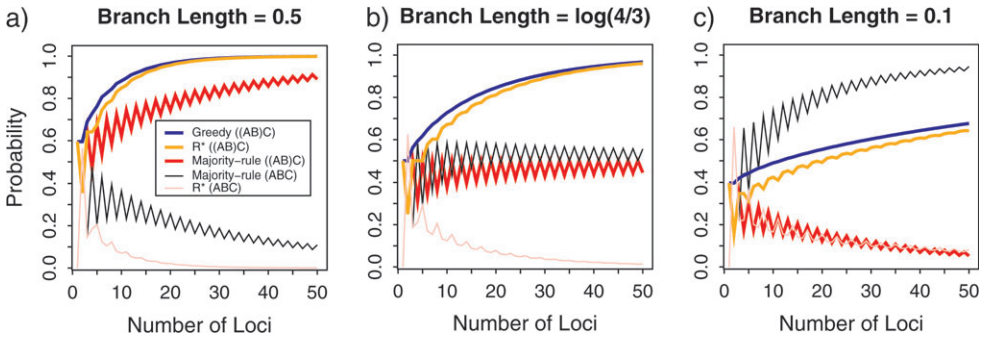

FIGURE 4. Species tree ((AB)C)—Probabilities of consensus trees from finite numbers of known gene trees. Each plot shows the probability that each of the 3 consensus methods will return either the species-tree topology ((AB)C) or a star tree (R^* and majority rule only). The legend in (*a*) also applies to each of the 3 plots.

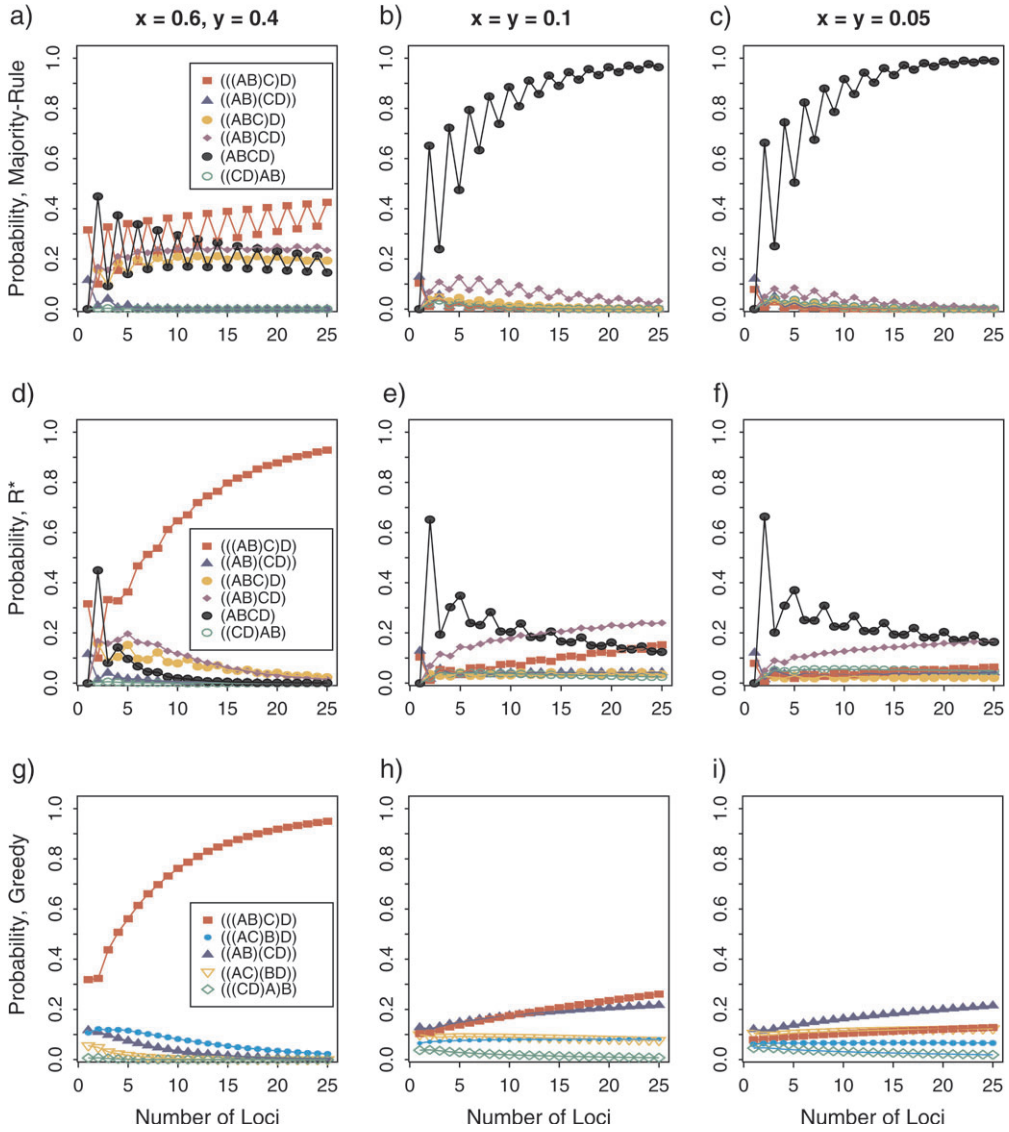

FIGURE 5. Species tree (((AB)C)D)—Probabilities of consensus trees from finite numbers of known gene trees. One consensus algorithm is used for each row of plots, and one set of branch lengths is used for each column. For the majority-rule and R^* algorithms, there are 26 possible 4-taxon consensus trees, including 15 fully resolved trees and 11 trees not fully resolved. The graphs only show some of the more frequently occurring consensus trees; consequently probabilities do not sum to 1. The legends in the left-hand column apply to the 3 plots in their corresponding rows.

For majority-rule trees, there is also an effect of having an odd or even sample size, where even sample sizes tend to give higher probabilities to unresolved trees. This occurs because even sample sizes increase the opportunity for ties for 2 or more incompatible clades, in which neither clade can be in the majority. This has the somewhat surprising consequence that a consensus tree can be less likely to match the species tree in a sample of $2n$ loci than in a sample of $2n-1$ (or even $2n-3$ or $2n-5$) loci. In being more likely to return an unresolved tree, however, majority-rule consensus is also less likely to produce a resolved tree that does not match the species tree. For the symmetric species-tree topology with branch lengths of $x = 0.6$ and $y = 0.4$, if the sample size is odd, then the majority-rule consensus tree is more likely to be the species-tree topology ((AB)(CD)) than any other topology, but for even sample sizes up to 25 loci, the unresolved tree ((CD)AB) is roughly tied in probability with ((AB)(CD)) (Fig. 6*a*). This result is consistent with Figure 2*b*, in which the point $(x, y) = (0.6, 0.4)$ is close to the boundary between the regions for ((AB)(CD)) and ((CD)AB). However, if the number of loci is sufficiently large, then majority-rule consensus is expected to return the resolved tree ((AB)(CD)) that matches the species tree because the point $(x, y) = (0.6, 0.4)$ is slightly outside the MACT unresolved zone (cf. inequalities (3) and (4)).

As the number of loci increases, the finite-sample R^* trees (Figs 5*d*,*e*,*f* and 6*d*,*e*,*f*) show increasing probability of matching the species-tree topology, including

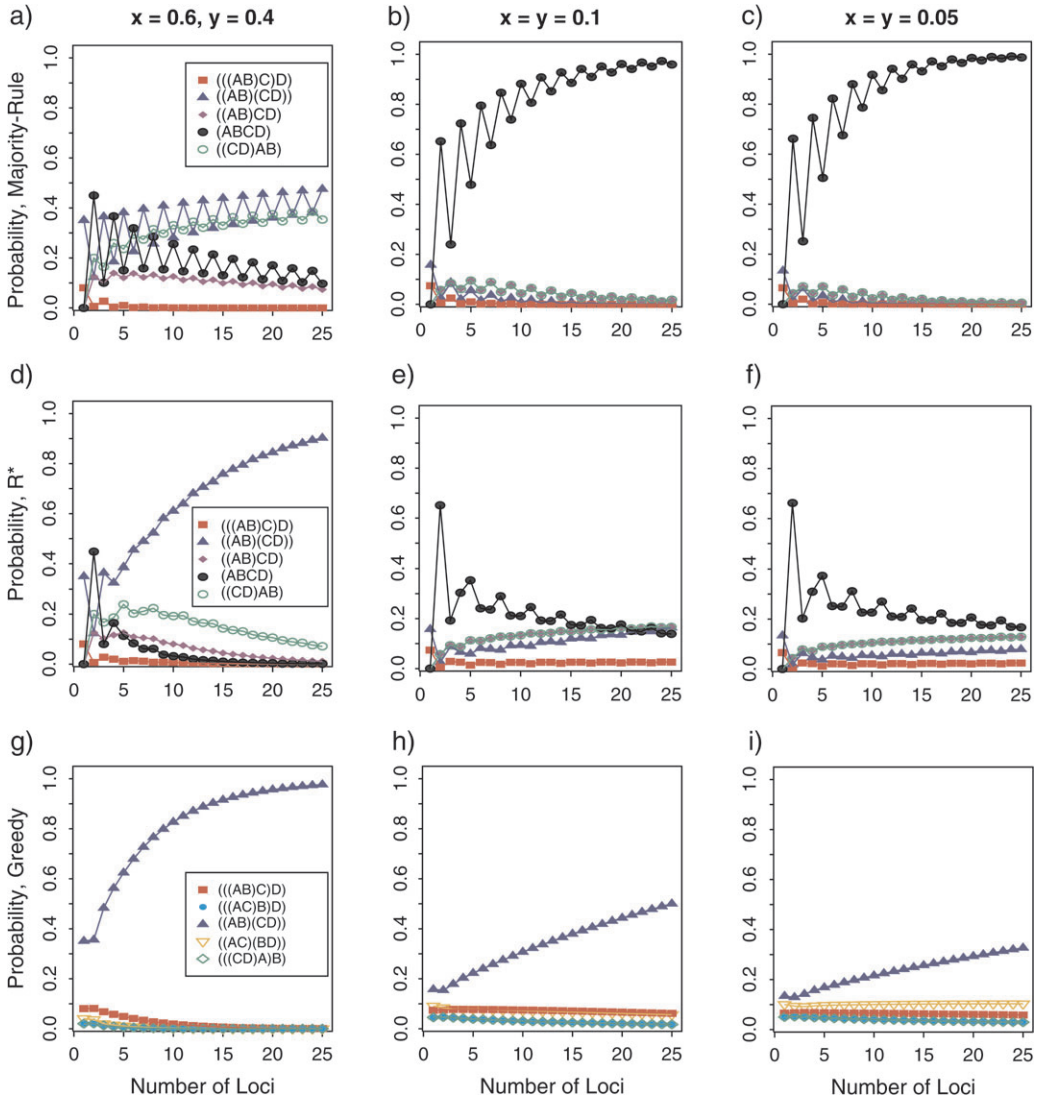

FIGURE 6. Species tree ((AB)(CD))—Probabilities of consensus trees from finite numbers of known gene trees. One consensus algorithm is used for each row of plots, and one set of branch lengths is used for each column. For the majority-rule and R^* algorithms, there are 26 possible 4-taxon consensus trees, including 15 fully resolved trees and 11 trees not fully resolved. The graphs only show some of the more frequently occurring consensus trees; consequently probabilities do not sum to 1. The legends in the left-hand column apply to the 3 plots in their corresponding rows.

for branch lengths that are in the anomaly zone, $(x, y) = (0.1, 0.1)$, and the too-greedy zone, $(x, y) = (0.05, 0.05)$. As we will see, this result agrees with our theoretical expectations of R^* consensus trees (Theorem 3); however, the increase in probability is very slow. For example, when $(x, y) = (0.1, 0.1)$ and the species tree is asymmetric (Fig. 5*e*), the 2 trees most likely to be returned are (ABCD) and ((AB)CD) until there are 23 loci, at which point the matching topology (((AB)C)D) changes from being the third to the second most probable topology. The star tree (ABCD) is the most likely tree to be inferred for 11 and fewer loci, and the trend is that the probability that the R^* consensus tree is the star tree decreases as the number of loci increases. The tree ((AB)CD), however, is still increasing in probability at 25 loci; thus, large numbers of loci might be needed for the matching tree to be the most likely tree to be returned by R^* consensus.

Greedy consensus trees show more smoothly increasing probabilities of returning the matching tree for branch lengths outside the too-greedy zone (Figs 5*g*,*h* and 6*g*,*h*,*i*). When the species tree is (((AB)C)D) and $(x, y) = (0.1, 0.1)$ (Fig. 5*h*), the gene tree ((AB)(CD)) is more probable than the matching tree, and greedy consensus is slightly more likely to return this nonmatching tree for small samples. The matching tree becomes the most probable greedy consensus tree with 11 or more loci; however, for this species tree with the extreme branch lengths of $(x, y) = (0.05, 0.05)$, increasing the number of loci is more likely to produce the nonmatching greedy

consensus tree ((AB)(CD)) (Fig. 5*i*). These results are consistent with our expectations based on the location of the too-greedy zone (Fig. 3).

Asymptotic Consensus Trees: General Theorems

Thus far we have considered in detail the properties of consensus methods with 3 and 4 taxa. It is desirable to understand how the results we have observed with 3 and 4 taxa generalize to larger numbers of taxa. Thus, we now provide theorems about MACTs, RACTs, and GACTs for rooted, binary species trees with arbitrarily many taxa. These theorems generalize the results obtained earlier regarding 3- and 4-taxon asymptotic consensus trees under the multispecies coalescent model. The proofs of the theorems are found in Appendices 1 (R^*), 2 (greedy), and 3 (majority rule).

Majority-Rule Consensus

We show that all clades on the MACT are also on the species tree and that for any species-tree topology, it is possible that the MACT is not fully resolved.

Theorem 1. *For all species-tree topologies with $n \geq 3$ taxa, (i) for all branch lengths, the MACT does not have any clades not on the species tree, and (ii) there exist branch lengths for which the MACT is not fully resolved.*

Theorem 1(i) is a consequence of the fact that the RACT always matches the species tree (Theorem 2). From the calculations in Asymptotic Consensus Trees, subsection "majority-rule consensus," Theorem 1(i) and (ii) follow for the 3- and 4-taxon cases. For larger trees, Theorem 1(ii) follows from the inconsistency of greedy consensus (Theorem 4) and from the fact that greedy consensus trees are resolutions of majority-rule trees.

R^* Consensus

We show that R^* consensus trees are consistent estimators of species-tree topologies for any number of taxa. This consistency occurs because for any set of 3 taxa, the rooted triple in the species tree is the highest probability rooted triple in the gene-tree distribution.

Theorem 2. *For a species tree σ, the RACT has the same topology as σ.*

Theorem 2 describes the RACT, which is a mathematical function of gene-tree probabilities and therefore of a species tree with branch lengths. When an R^* consensus tree is computed from a finite number of loci, however, it has some probability of not matching the species tree. For an estimator of a parameter to be statistically consistent, for any point in the parameter space, the probability that the estimator gets arbitrarily close to the parameter must approach 1 as the sample size approaches ∞. Theorem 3 states that, regardless of the species-tree topology and branch lengths, the probability that the R^* consensus tree constructed from a finite number of loci matches the species tree approaches 1 as the number of loci approaches ∞.

Theorem 3. *R^* consensus is statistically consistent.*

Greedy Consensus

The result that greedy consensus can be misleading in the 4-taxon asymmetric case generalizes to any species-tree topology with more than 4 taxa. Intuitively, by making some branches long and some short (so that coalescent events occur with probability arbitrarily close to 1 or 0), trees with 5 or more taxa can be made to behave similarly to the 4-taxon asymmetric case. The strategy of the proof therefore uses an argument similar to that used in proving Lemma 5 in Degnan and Rosenberg (2006).

Theorem 4. *For 3-taxon species-tree topologies and for 4-taxon symmetric species-tree topologies, the GACT matches the species tree; for the asymmetric topology with $n = 4$ taxa and for every species-tree topology with $n \geq 5$ taxa, there exist branch lengths such that the GACT does not match the species tree.*

Theorems 1–4 help to explain the behavior observed for majority-rule, R^*, and greedy consensus in our detailed analysis of 4 taxa under the multispecies coalescent model. In particular, the observations that MACTs can be unresolved, RACTs match the species tree, and GACTs do not necessarily match the species tree, all generalize to arbitrary numbers of taxa.

Discussion

Use of coalescent probabilities makes it possible to predict which trees are likely to be constructed using consensus of gene trees from many independent loci. We have obtained results under the multispecies coalescent model for 3 types of asymptotic consensus trees: majority rule, R^*, and greedy. Theorems 1, 2, and 4, respectively, demonstrate that with an infinite number of loci, MACTs might be unresolved, RACTs always match the species tree, and GACTs might be nonmatching. These results have implications for a common goal of phylogenetics: the inference of species trees.

Estimating Species Trees

Although concatenation of sequences is perhaps the most widely used method of estimating species trees, several alternatives to concatenation currently exist for inferring species trees. These include minimizing deep coalescence (Maddison and Knowles 2006), finding the joint posterior of the species tree and gene trees from the coalescent model in a Bayesian framework (Liu and Pearl 2007), using the most ancient speciation times compatible with the set of inferred coalescent times on a set of gene trees (called the "maximum tree" by Liu and Pearl [2007] or "GLASS tree" by Mossel and Roch [2009]),

and using probabilities of gene-tree topologies to approximate the species-tree likelihood (Carstens and Knowles 2007; Carling and Brumfield 2008). These methods are designed to estimate species trees when gene-tree conflict results from incomplete lineage sorting, and they do not assume that sequence data are generated under a single gene-tree topology.

Theorem 3 suggests a statistically consistent method for building species-tree topologies from gene-tree topologies (assuming known gene trees). This method involves inferring all rooted triples of the gene-tree topologies and then determining the clades of the estimated species tree (rules 1′ and 2′ above) to build up the tree from the $\binom{n}{3}$ rooted triples (Bryant and Berry 2001). A closely related method that estimates species trees from rooted triples is described by Ewing et al. (2008). In their method, quartets of species are estimated at each locus, where each quartet has the same out-group in addition to a set of 3 in-group taxa. These quartets with an out-group correspond to rooted triples when the out-group is removed. Quartet puzzling (Strimmer and von Haeseler 1996) is then used to build the species-tree estimate. Similarly to the justification of R^* consensus, this quartet approach to rooted triple consensus is also motivated by the idea that the most probable 3-taxon statement matches the species tree. Because quartet puzzling builds the tree heuristically, however, we expect R^* to be more conservative for smaller numbers of loci and to be more likely to return a partially unresolved tree.

Although the R^* consensus algorithm does not estimate species-tree branch lengths, rooted triples could also be used to estimate internal branch lengths on the species tree by using $P_\sigma[(\mathrm{AB})\mathrm{C}] = 1 - (2/3)\mathrm{e}^{-T}$ (Nei 1987), where T is the length separating the MRCA of A, B, and C from the MRCA of A and B. The topologies of the observed gene trees can be used to obtain maximum likelihood estimates of T. This idea has been applied to the human–chimpanzee–gorilla phylogeny (e.g., Wu 1991; Chen and Li 2001) to infer the time separating the gorilla divergence from the human–chimpanzee divergence. Wakeley (2008) gives an example in which the one internal branch length is estimated in coalescent units using 28 gene-tree topologies for 3 in-group taxa of Australian grass finches analyzed by Jennings and Edwards (2005). This approach could be extended to trees with larger numbers of taxa. The frequency of each rooted triple in the observed set of gene trees could be used to estimate species divergence times, from which the species tree (including internal branch lengths) could be constructed. Alternatively, given a species-tree topology, the set of branch lengths most compatible with the observed rooted triples could be determined using a criterion such as maximum likelihood or least squares.

Using majority rule to estimate species trees from finitely many loci is not expected to result in many false clades, but for some sets of branch lengths it is likely to result in a tree that is at least partially unresolved. It is thus expected to provide a conservative estimate of the species tree, with little power to resolve some clades for some sets of branch lengths. We note that R^* consensus appears to be more likely than majority rule to correctly recover resolved clades, both asymptotically as well as for finite numbers of loci (Figs 5 and 6).

Greedy consensus can be misleading in the sense that it can be increasingly likely to return a nonmatching, fully resolved tree as the number of loci grows. However, when there are 4 taxa, this inconsistency only occurs for a relatively small portion of the parameter space (Fig. 3), and outside this region, greedy consensus typically converges to the species-tree topology more quickly than does either majority-rule or R^* consensus.

Mutation and Recombination

In this paper, we have not considered the roles of mutation and recombination and the resulting uncertainty that occurs when gene trees are inferred from sequence data. When gene trees are estimated and the underlying species tree has short branches, some gene trees are expected to not be fully resolved due to insufficient divergence among sequences. Also, due to the inherent stochasticity in sequence evolution, some gene trees are likely to be incorrectly inferred. For finite numbers of genes, these factors would tend to increase the probability that majority-rule consensus trees would have some lack of resolution, whether or not the true MACT was fully resolved. If the MACT is a star tree, we speculate that accounting for mutation would cause convergence to a star tree to occur more quickly as the number of loci is increased. If the MACT does have some resolved clades, then uncertainty in the gene trees would be expected to increase the number of loci needed to have a high probability that an estimated majority-rule consensus tree is the same as the MACT. We expect similar effects for R^* and greedy consensus trees; ultimately, the effects of mutation on consensus trees could be assessed by simulating sequence data for independent gene trees evolving on the species tree.

When gene trees are estimated from sequence data, they are estimated with some degree of error. Because different gene-tree estimates may not have the same level of certainty, it may be desirable to give gene trees different weights before inputting them into a consensus algorithm or to only use gene trees with high support (e.g., Ebersberger et al. [2007] analyze a multilocus data set using all genes and reanalyze using only genes for which the inferred gene tree had a high posterior probability). Using trees inferred from a Bayesian analysis, for example, gene trees could be weighted by their posterior probability, with each locus contributing one unit of weight potentially distributed over several gene-tree topologies. This approach is used, for example, in estimating concordance trees (Ané et al. 2007).

Recombination within genes can cause segments within a gene to have different tree topologies, thus creating a problem similar to the one arising when concatenating genes that were generated under different topologies. If recombination within genes is fairly infrequent, then aligned sequences can be tested for

recombination (e.g., see Wiuf et al. 2001), so that only nonrecombining genes are used. A more sophisticated approach is to concatenate aligned genes and then to break up the alignments into "recombination blocks" using a hidden Markov model that treats coalescent histories of gene trees in species trees as states in the Markov chain. This approach was used by Hobolth et al. (2007) to analyze genomic data for the human–chimpanzee–gorilla tree.

Methods of combining gene trees to infer species trees often assume that gene trees are independent given the species tree. However, this is only strictly true if genes are unlinked, meaning that there is a high probability of recombination between genes. When genes are tightly linked, they may share the same evolutionary history on a short timescale. In practice, this means that the number of independent gene trees can be smaller than the number of genes shared by a collection of taxa, and estimation using multiple genes should use genes sufficiently far apart that they can be considered independent. Slatkin and Pollack (2006) studied the case of 3 taxa and found that 2 gene trees are approximately independent (given the species tree) when the 2 genes are not in linkage disequilibrium. For *Drosophila*, an example for which incomplete lineage sorting is thought to be pervasive (Pollard et al. 2006), Slatkin and Pollack found that gene trees are approximately independent when loci are, on average, ∼8 kb apart. Thus, it may be feasible to find several hundreds or thousands of genes in real genomes that can be considered independent. Asymptotic properties of multilocus estimators of species tree might therefore provide a basis for understanding analyses of such large quantities of data.

Conclusions

Our results show that when there is sufficient gene-tree discordance due to incomplete lineage sorting, majority-rule consensus trees can have a high probability of being at least partially unresolved, and for some sets of branch lengths, the probability of being unresolved can approach 1 as the number of genes increases indefinitely. However, the MACT is never resolved incorrectly; that is, it never has a clade not in the species tree. We therefore describe the MACT as not misleading; however, it is not consistent because for an estimator to be statistically consistent for a parameter (e.g., a fully resolved species tree), that estimator must produce estimates arbitrarily close to the parameter with probability approaching 1 as the sample size increases.

The fact that under the multispecies coalescent model R^* trees are asymptotically guaranteed to be fully resolved and to match the species-tree topology means that the R^* procedure is not only not misleading but is also a statistically consistent estimator of the species-tree topology. This is remarkable considering that R^* trees are based only minimally on a model of species tree–gene tree relationships. The only feature of the multispecies coalescent model used in proving the consistency of the R^* method is the fact that in this model, 3-taxon relationships that occur in the species tree are also expected to occur in the gene-tree distribution. Thus, although R^* consensus trees are consistent without explicitly incorporating gene-tree probabilities into the R^* algorithm for constructing trees, it will be important to examine how robust the R^* consensus algorithm is to violations of assumptions in the coalescent, such as the absence of population structure along ancient internal edges.

Finally, greedy consensus trees can be increasingly likely (as the number of gene trees increases) to have a topology that differs from that of the species tree. Thus, greedy consensus trees can be misleading if used as estimators of species trees. However, for 4 taxa, the region of parameter space in which greedy consensus fails to return the true tree—the too-greedy zone—is relatively small, smaller than the anomaly zone; hence, greedy consensus offers some robustness to gene-tree discordance that may cause other methods to fail to recover the species tree. In addition, the greedy consensus method outperformed our other methods for branch lengths outside the too-greedy zone. However, there may be a trade-off between consistency and speed of convergence, with greedy consensus being the quicker to converge yet statistically inconsistent and with R^* consensus being slow to converge yet statistically consistent. To test these consensus methods in practice will require examining their performance in the presence of mutation so that gene trees are estimated with uncertainty rather than treated as known.

Funding

National Science Foundation (DEB-0716904); Burroughs Wellcome Fund; Alfred P. Sloan Foundation; National Institutes of Health (T32 GM070449); NZ Marsden Fund to D.B.

Acknowledgments

We thank C. Ané, J. Cotton, F. Matsen, E. Allman, and an anonymous reviewer for comments.

References

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24: 412–426.

Baum B.R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining trees. Taxon. 41:3–10.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon. 56:417–426.

Bremer K. 1990. Combinable component consensus. Cladistics. 6: 369–372.

Bryant D. 2003. A classification of consensus methods for phylogenies. In: Janowitz M., Lapointe F.-J., McMorris F.R., Mirkin B., Roberts F.S., editors. BioConsensus. Providence (RI): Center for Discrete Mathematics and Theoretical Computer Science, American Mathematical Society. p. 163–183.

Bryant D., Berry V. 2001. A structured family of clustering and tree construction methods. Adv. Appl. Math. 27:705–732.

Carling M.D., Brumfield R.T. 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. Genetics. 178:363–377.

Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56:400–411.
Casella G., Berger R.L. 1990. Statistical inference. Belmont (CA): Duxbury Press.
Chen F.-C., Li W.-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68:444–456.
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., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution. 59:24–37.
Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., von Haeseler A. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2277.
Edwards S.V., Liu L., Pearl D.K. 2007. High resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104:5936–5941.
Ewing G.B., Ebersberger I., Schmidt H.A., von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8:118.
Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22:521–565.
Felsenstein J. 1995. Phylogenetic Inference Package (PHYLIP). Version 3.5. Seattle (WA): University of Washington.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Hein J., Schierup M.H., Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press.
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:e7.
Jennings W.B., Edwards S.V. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution. 59:2033–2047.
Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.
Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56: 17–24.
Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Matsen F.A., Steel M. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56:767–775.
Mossel E., Roch S. 2009. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE Comput. Biol. Bioinform. doi: 10.1109/TCBB.2008.66.
Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science. 309:2207–2209.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Nei M., Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.
Ragan M.A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.
Ranwez V., Berry V., Criscuolo A., Fabre P.-H., Guillemot S., Scornavacca C., Douzery E.J.P. 2007. PhySIC: a veto supertree method with desirable properties. Syst. Biol. 56:798–817.
Rosenberg N.A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.
Ross S. 1998. A first course in probability. 5th ed. Upper Saddle River (NJ): Prentice-Hall.
Slatkin M., Pollack J.L. 2006. The concordance of gene trees and species trees at two linked loci. Genetics. 172:1979–1984.
Steel M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J. Classification. 9:91–116.
Strimmer K., von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.
Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.
Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics. 122:957–966.
Tavaré S. 1984. Line-of-descent and genealogical process, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.
Wakeley J. 2008. Coalescent theory. Greenwood Village (CO): Roberts and Company.
Wiuf C., Christensen T., Hein J. 2001. A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. 18: 1929–1939.
Wu C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics. 127:429–435.

First Submitted 17 April 2008; reviews returned 7 July 2008; final acceptance 22 October 2008
Associate Editor: Cécile Ané

APPENDIX 1

R, Proofs of Theorems 2 and 3*

For terminology in the proof of Proposition 5, we say that (AB) is a lineage "containing" A and B. We additionally say that 2 taxa "join" or "are joined" on a branch b if the lineages (i.e., clades) containing those taxa coalesce on branch b. For example, if (AB) and C coalesce on branch 3, then A and C "join" on branch 3.

Proposition 5. Let σ be the species tree where S is the set of taxa on σ. For any A, B, C $\in S$, if σ has the grouping (AB)C, then P_σ[(AB)C] > P_σ[(AC)B].

Proof. Let $\mathcal{J}$ be the set of branches of σ on which A and B can join (i.e., either the lineages A and B or separate lineages containing A and B can coalesce in $\mathcal{J}$) but on which A and C cannot join. Note that $\mathcal{J}$ is nonempty and that any branch in $\mathcal{J}$ is an ancestor of species A and B and not an ancestor of species C. Let $\mathcal{K}$ be the set of branches on which gene lineages A and C can join. Any branch in $\mathcal{K}$ is an ancestor of species A and C. Because (AB)C is a rooted triple in σ, any ancestor of species A and C is also an ancestor of species B. Thus, for any branch $k \in \mathcal{K}$, if none of the lineages A, B, and C have joined, then they are free to do so on k. The probability that A and B join on a branch in $\mathcal{J}$ is positive. If A and B do not join in $\mathcal{J}$, then the probabilities that A and B, A and C, and B and C are the "first" (going backwards in time) 2 of A, B, and C to join in $\mathcal{K}$ are equal because all pairs of lineages in a population are equally likely to coalesce. Thus, P_σ[(AB)C] > P_σ[(AC)B]. □

Proof of Theorem 2. By Proposition 5, any rooted triple in the species tree has higher probability in the gene-tree

distribution than the other 2 rooted triples for the same set of 3 taxa. Thus, the set of rooted triples from which the R^* tree is constructed is exactly the set of $\binom{n}{3}$ rooted triples in the species tree, where n is the number of taxa. From Steel (1992), a tree topology is uniquely specified by its set of rooted triples, from which it follows that the only tree topology containing the $\binom{n}{3}$ triples is the topology of the species tree itself. □

The proof of Theorem 3 uses a generalized version of Bonferroni's inequality, according to which if there are k events each with probability $p = 1 - q$, then the probability that they all occur is greater than or equal to $1 - kq$ (Ross 1998, p. 63).

Proof of Theorem 3. It suffices to show that for any $\varepsilon > 0$, there exists k such that if there are at least k independent gene trees, the probability is greater than $1 - \varepsilon$ that all rooted triples in the species tree are also the most frequently occurring rooted triples for all sets of 3 taxa in the collection of gene trees. Let σ be the species tree with taxon set S. For n taxa, there are $\binom{n}{3}$ sets of 3 taxa in S. Let A, B, and C be 3 distinct taxa in S. Without loss of generality, assume that (AB)C is the jth rooted triple on σ. From Proposition 5, $P_\sigma[(AB)C] > P_\sigma[(AC)B] = P_\sigma[(BC)A]$, where the equality holds by symmetry. Thus, $P_\sigma[(AB)C] = 1/3 + \delta$ and $P_\sigma[(AC)B] = 1/3 - \delta/2$ for some $\delta > 0$. We use $\widehat{P}$ to denote sample proportions of rooted triples. For any $\varepsilon > 0$, because sample proportions converge in probability to their parametric values (by the weak law of large numbers) as the sample size tends to ∞, we can choose the number of loci k_j to be large enough that with probability greater than $1-\varepsilon/\binom{n}{3}$, $\widehat{P}_\sigma[(AB)C] > 1/3$, $\widehat{P}_\sigma[(AC)B] < 1/3$, and $\widehat{P}_\sigma[(BC)A] < 1/3$. Letting $k = \max_{j|j\in\{1,2,\ldots,\binom{n}{3}\}} k_j$, for each set of 3 taxa the probability that its most common rooted triple in the gene-tree distribution matches the rooted triple in the species tree is greater than $1 - \varepsilon/\binom{n}{3}$. By Bonferroni's inequality, the probability that all the $\binom{n}{3}$ rooted triples in the R^* tree are rooted triples in the species tree is therefore greater than $1 - \varepsilon$. □

APPENDIX 2

Greedy, Proof of Theorem 4

Lemma 6. *The 4-taxon asymmetric species-tree topology (((AB)C)D) has a set of branch lengths which makes the asymptotic greedy consensus tree fail to match the species tree.*

This set is explicitly derived in Appendix 7 and is given in inequality (5) and Figure 3.

Lemma 7. *For every bifurcating species tree with $n \geq 5$ taxa and every $k \geq 1$ with $2^{k+1} < n$, there is a node with c terminal descendants, where $2^k < c < 2^{k+1} + 1$.*

Proof. Take any $k \geq 1$. If $2^{k+1} + 1 \leq n$, then the root has $n \geq 2^{k+1} + 1$ terminal descendants. Let $\mathcal{N}_0$ denote the root node, and choosing between the 2 nodes immediately descended from the root, let $\mathcal{N}_1$ denote the internal node with the larger number of terminal descendants (choosing arbitrarily in case of a tie). Similarly, let $\mathcal{N}_2$ be the internal node (if it exists) immediately descended from $\mathcal{N}_1$ with the larger number of terminal descendants. Continue this process until a node $\mathcal{N}_m$ ($m \geq 0$) is reached which has at least $2^{k+1}+1$ terminal descendants but neither of whose immediate descendant nodes has more than 2^{k+1} terminal descendants. Call $\mathcal{N}_m$ the "minimal node." It follows that at least one of the immediate descendant nodes of the minimal node has more than 2^k terminal descendants (because otherwise the minimal node would have at most $2(2^k) < 2^{k+1} + 1$ descendants). Thus, at least one immediate descendant of the minimal node has c terminal descendants with $2^k < c < 2^{k+1} + 1$. □

Lemma 8. *If for some $k \geq 2$, all species-tree topologies with n taxa, $n \in \{2^k + 1, \ldots, 2^{k+1}\}$, have a nonempty too-greedy zone, then all species-tree topologies with $n > 2^{k+1}$ (and thus $n \geq 2^k + 1$) taxa have a nonempty too-greedy zone.*

Proof. Assume there exists $k \geq 2$ such that all species-tree topologies with $n \in \{2^k + 1, \ldots, 2^{k+1}\}$ taxa have a nonempty too-greedy zone, that is, there exist branch lengths for which the GACT does not match the species-tree topology. By Lemma 7, each species tree σ with more than 2^{k+1} ($k \geq 1$) taxa S has some node $\mathcal{N}$ with c terminal descendants, where $c \in \{2^k + 1, \ldots, 2^{k+1}\}$. Let $\sigma_\mathcal{N}$ denote the species tree rooted at $\mathcal{N}$, and let $S_\mathcal{N}$ denote the set of taxa labeling the tips of $\sigma_\mathcal{N}$. By assumption, the topology of $\sigma_\mathcal{N}$ has a nonempty too-greedy zone.

Let B be the number of branches of σ that are outside $\sigma_\mathcal{N}$. Make the lengths of all branches outside $\sigma_\mathcal{N}$ long enough that for each branch b not in $\sigma_\mathcal{N}$, the probability that all lineages on b coalesce is greater than $1 - \varepsilon/B$, where ε is chosen so that $1 - \varepsilon > 1/2$ and $1 - \varepsilon$ is greater than the probability of each clade within $\sigma_\mathcal{N}$ (i.e., each clade which is a proper subset of $S_\mathcal{N}$). Because the greedy consensus tree is a resolution of the majority-rule consensus tree, all clades of σ which include taxa outside $S_\mathcal{N}$, and the clade consisting of all taxa in $S_\mathcal{N}$, are included in the GACT. When ranking clade probabilities as is required for the algorithm for constructing the GACT, these clades are added before the clades whose sets of taxa are proper subsets of $S_\mathcal{N}$. Thus, eventually the list of candidate clades consists only of proper subsets of $S_\mathcal{N}$. When clades are accepted from this list, by assumption we accept at least one clade to be in the GACT which is not on σ. Thus, there exist branch lengths on σ for which the GACT does not match the species tree. □

Lemma 9. *For any species-tree topology with 5, 6, 7, or 8 taxa, there exists a set of branch lengths for which the GACT does not match the species tree.*

Proof. This is shown by reduction to the 4-taxon asymmetric case. For each species-tree topology with 5, 6, 7, or 8 taxa, some branches can be made long and some can be made short so as to produce the same inconsistencies

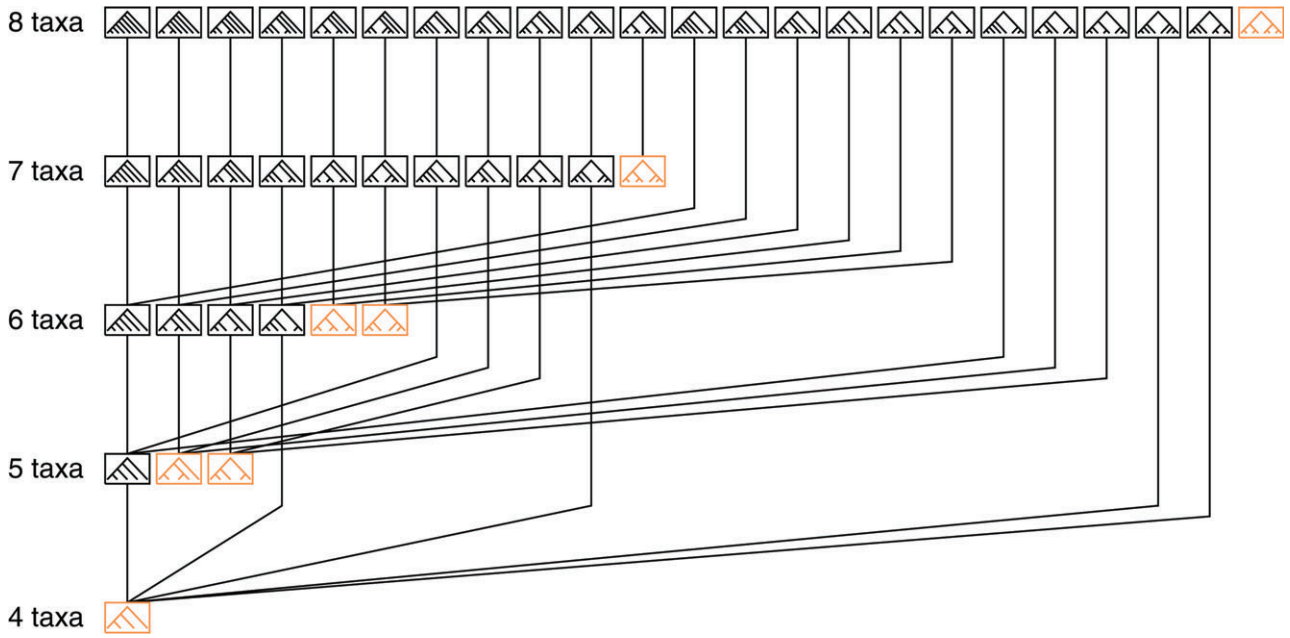

FIGURE B1. Reduction of topologies used in the proof of Lemma 9. If 2 trees are connected by an edge, then the topology with the smaller number of leaves is a left subtree of the larger tree.

as those seen in the 4-taxon case. Most cases are shown in Figure B1. Here a topology with n taxa is connected by an edge to a topology with fewer than n taxa if the smaller topology is the left subtree—from the node which is the immediate left descendant of the root—of the larger topology. In this case, for any $\varepsilon > 0$, all branches on the larger topology not in the left subtree can be made arbitrarily long. Thus, all lineages available to coalesce on long branches do coalesce with probability greater than $1 - \varepsilon$. Remaining clades then have the same order of probabilities as on the left subtree and thus are accepted by the greedy algorithm in the same order as on the left subtree.

If the greedy consensus algorithm returns a nonmatching tree for the smaller tree, then it also does so for the larger tree because the ranking of the remaining clades by frequencies is eventually the same (once the high-probability clades have already been added on the larger tree). This process of reducing trees can be repeated until one of the trees colored orange (which have no edges connecting to a smaller tree) is reached.

It then remains to be shown that the GACT does not match the species tree for the remaining orange trees from Figure B1. This is already shown explicitly for the 4-taxon case (Lemma 6). The other trees can again be reduced to the 4-taxon case by choosing certain edges to be long and others short, as shown in Figure B2. By choosing the long, orange branches to have large branch lengths, the probability that all available lineages coalesce on a branch can be made greater than $1 - \varepsilon/(2m)$, where m is the number of long branches on a tree. This makes the probability that all available lineages on long branches coalesce greater than $1-\varepsilon/2$. Because only counter examples are needed to show that the greedy consensus algorithm can return a nonmatching tree, it suffices to note that branches can be chosen to be short enough using inequality (5) or Figure 3 for the 4-taxon asymmetric tree to make the greedy consensus algorithm fail to return the tree matching the species tree with probability greater than $1 - \varepsilon/2$. Making the black internal branches sufficiently short, the probability exceeds $1 - \varepsilon$ that the entire tree returned by the greedy consensus algorithm fails to match the species-tree topology. □

Proof of Theorem 4. The result for 3 taxa follows from the fact that the matching gene tree has the highest

FIGURE B2. Reduction of the remaining trees from Figure B1 to the 4-taxon asymmetric case, for the proof of Lemma 9. Branches in orange are made long enough that all lineages on these branches coalesce with probability arbitrarily close to 1.

probability of the 3 possible gene trees. The 4-taxon asymmetric case is covered in Lemma 6. The 4-taxon symmetric case is treated in Appendix 6 by showing that for all branch lengths, (AB) and (CD) are the 2 most probable clades. We have shown that all cases with $n = 5, 6, 7$, or 8 taxa have too-greedy zones (Lemma 9). From Lemma 8, this verifies by induction that all cases with $n \geq 5$ taxa have such zones. □

APPENDIX 3

Majority Rule, Proof of Theorem 1

Proof of Theorem 1(i). This result follows from Theorem 2 and Theorem 2.14 of Bryant (2003), according to which every clade in the majority-rule consensus tree is in the R^* consensus tree. Because the MACT and RACT are the majority-rule and R^* consensus trees applied to coalescent gene-tree probabilities, every clade in the MACT must appear in the RACT. Because in the limit of infinitely many gene trees, the R^* tree is fully resolved, it follows that if the MACT has one or more multifurcations, then the R^* tree is one of the possible resolutions of the MACT. Because the RACT has the same topology as the species tree (Theorem 2), the MACT either has the species-tree topology or one of its resolutions has the same topology as the species tree. □

Proof of Theorem 1(ii). The GACT and MACT are each examples of greedy and majority-rule consensus trees, respectively. It follows that if the MACT is fully resolved, then it is the same as the GACT because greedy consensus trees are resolutions of majority-rule consensus trees (Bryant 2003). However, by Theorem 4, for any species-tree topology with $n \geq 5$ taxa, there exist branch lengths for which the GACT has a clade not on the species tree and therefore cannot be equivalent to the MACT (by Theorem 1(i)). Consequently, a sufficient condition for the MACT to be unresolved is for the GACT to not match the species tree. Because exact conditions for the MACT to not be fully resolved were obtained earlier for smaller trees (the internal branch length being no greater than $\log(4/3)$ for 3-taxon trees and one of inequalities (1–4) being required to fail for 4-taxon trees), the result follows for any species tree with $n \geq 3$ taxa. □

APPENDIX 4

Majority-Rule Unresolved Zones, Species Tree (((AB)C)D)

In this appendix, we derive conditions for which the MACT is unresolved for the 4-taxon species tree (((AB)C)D). This is done by finding branch lengths for which there exist clades with probability greater than 1/2. First, the following result about cherries, which is analogous to Proposition 5, is useful.

Proposition 10. Let σ be the species tree where S is the set of taxa on σ. Then for any A, B, C $\in S$, if {AB} is a cherry on σ, then $P_\sigma[\{AB\}] > P_\sigma[\{AC\}]$.

The proof is omitted because it is very similar to the proof of Proposition 5.

Remark 11. If {AB} is a cherry on the species tree σ, then for any taxon C, $P_\sigma[\{AC\}] = P_\sigma[\{BC\}] < 1/3$.

The equality holds by symmetry; the inequality follows from Proposition 10.

To find branch lengths for the species tree (((AB)C)D) where the MACT is resolved, consider the probabilities of clades {ABC} and {AB}. Table 1 lists the probability that A, B, and C are monophyletic as $p_1 + p_3 + p_7$, where p_i is the probability of gene tree i in the same table, because for gene trees 1, 3, and 7 (and only these gene trees), these 3 taxa are monophyletic. Table D1 can be used to compute probabilities of gene trees, clades, or rooted triples for 4-taxon trees as linear combinations of products of the terms $g_{ij}(T)$, which denote the probability that i lineages coalesce into j lineages within T coalescent time units, where $i \geq j \geq 1$ and $T > 0$. For $i = 2, 3$, the $g_{ij}(t)$ functions are (Tavaré 1984)

$$g_{21}(T) = 1 - e^{-T}, \quad g_{31}(T) = 1 - \frac{3}{2}e^{-T} + \frac{1}{2}e^{-3T},$$
$$g_{22}(T) = e^{-T}, \quad g_{32}(T) = \frac{3}{2}e^{-T} - \frac{3}{2}e^{-3T},$$
$$g_{33}(T) = e^{-3T}. \tag{D.1}$$

For example, from Table D1, for the species tree (((AB)C)D), the probability of clade {CD} is $\frac{1}{3}g_{21}(y)g_{22}(x) + \frac{1}{9}g_{22}(y)g_{32}(x) + \frac{4}{18}g_{22}(y)g_{33}(x)$.

The probability of clade {ABC} is

$$P_\sigma[\{ABC\}] = p_1 + p_3 + p_7 = 1 - \frac{2}{3}e^{-x} - \frac{1}{3}e^{-(x+y)} + \frac{1}{6}e^{-(3x+y)}. \tag{D.2}$$

Setting $P_\sigma[\{ABC\}] > 1/2$, we obtain a condition for which the MACT has the clade {ABC}. No other 3-taxon clade can be on the MACT because each of the other 3-taxon clades is incompatible with and less probable than {ABC}, and therefore each has probability less than 1/2. This claim can be verified by checking probabilities of 3-taxon clades from Table D1 and comparing coefficients of the $g_{ij}(T)$ terms. Three-taxon clades for the species tree (((AB)C)D) have probabilities

$$P_\sigma(\{ABC\}) = g_{21}(y)g_{21}(x) + \frac{1}{3}g_{21}(y)g_{22}(x) + g_{22}(y)g_{31}(x) + \frac{3}{9}g_{22}(y)g_{32}(x) + \frac{3}{18}g_{22}(y)g_{33}(x),$$
$$P_\sigma(\{ABD\}) = \frac{1}{3}g_{21}(y)g_{22}(x) + \frac{1}{9}g_{22}(y)g_{32}(x) + \frac{3}{18}g_{22}(y)g_{33}(x),$$
$$P_\sigma(\{ACD\}) = P_\sigma(\{BCD\}) = \frac{1}{9}g_{22}(y)g_{32}(x) + \frac{3}{18}g_{22}(y)g_{33}(x).$$

TABLE D1. Probabilities of 4-taxon gene-tree topologies, clades, and rooted triples as functions of terms $g_{ij}(T)$. The branch lengths x and y are as in Figure 1*a*. The probabilities of clades (rooted triples) are obtained by adding the probabilities of gene-tree topologies which have the clade (rooted triple, see Table 1). For each entry in the table, the left and right numbers are the coefficients of the $g_{ij}(T)$ (Equation (D.1)) terms for the species trees (((AB)C)D) and ((AB)(CD)), respectively

Gene-tree topology	$g_{21}(y)g_{21}(x)$	$\frac{1}{3}g_{21}(y)g_{22}(x)$	$\frac{1}{3}g_{22}(y)g_{21}(x)$	$\frac{1}{18}g_{22}(y)g_{22}(x)$	$\frac{1}{3}g_{22}(y)g_{31}(x)$	$\frac{1}{9}g_{22}(y)g_{32}(x)$	$\frac{1}{18}g_{22}(y)g_{33}(x)$
1. (((AB)C)D)	1,0	1,1	0,0	0,1	1,0	1,0	1,0
2. (((AB)D)C)	0,0	1,1	0,0	0,1	0,0	1,0	1,0
3. (((AC)B)D)	0,0	0,0	0,0	0,1	1,0	1,0	1,0
4. (((AC)D)B)	0,0	0,0	0,0	0,1	0,0	1,0	1,0
5. (((AD)B)C)	0,0	0,0	0,0	0,1	0,0	0,0	1,0
6. (((AD)C)B)	0,0	0,0	0,0	0,1	0,0	0,0	1,0
7. (((BC)A)D)	0,0	0,0	0,0	0,1	1,0	1,0	1,0
8. (((BC)D)A)	0,0	0,0	0,0	0,1	0,0	1,0	1,0
9. (((BD)A)C)	0,0	0,0	0,0	0,1	0,0	0,0	1,0
10. (((BD)C)A)	0,0	0,0	0,0	0,1	0,0	0,0	1,0
11. (((CD)A)B)	0,0	0,0	0,1	0,1	0,0	0,0	1,0
12. (((CD)B)A)	0,0	0,0	0,1	0,1	0,0	0,0	1,0
13. ((AB)(CD))	0,1	1,1	0,1	0,2	0,0	1,0	2,0
14. ((AC)(BD))	0,0	0,0	0,0	0,2	0,0	1,0	2,0
15. ((AD)(BC))	0,0	0,0	0,0	0,2	0,0	1,0	2,0
Clade							
{AB}	1,1	3,3	0,1	0,4	1,0	3,0	4,0
{AC}	0,0	0,0	0,0	0,4	1,0	3,0	4,0
{AD}	0,0	0,0	0,0	0,4	0,0	1,0	4,0
{BC}	0,0	0,0	0,0	0,4	1,0	3,0	4,0
{BD}	0,0	0,0	0,0	0,4	0,0	1,0	4,0
{CD}	0,1	1,1	0,3	0,4	0,0	1,0	4,0
{ABC}	1,0	1,1	0,0	0,3	3,0	3,0	3,0
{ABD}	0,0	1,1	0,0	0,3	0,0	1,0	3,0
{ACD}	0,0	0,0	0,1	0,3	0,0	1,0	3,0
{BCD}	0,0	0,0	0,1	0,3	0,0	1,0	3,0
Rooted triple							
(AB)C	1,1	3,3	0,1	0,6	1,0	3,0	6,0
(AC)B	0,0	0,0	0,1	0,6	1,0	3,0	6,0
(BC)A	0,0	0,0	0,1	0,6	1,0	3,0	6,0
(AB)D	1,1	3,3	0,1	0,6	3,0	5,0	6,0
(AD)B	0,0	0,0	0,1	0,6	0,0	2,0	6,0
(BD)A	0,0	0,0	0,1	0,6	0,0	2,0	6,0
(AC)D	1,0	1,1	0,0	0,6	3,0	5,0	6,0
(AD)C	0,0	1,1	0,0	0,6	0,0	2,0	6,0
(CD)A	0,1	1,1	0,3	0,6	0,0	2,0	6,0
(BC)D	1,0	1,1	0,0	0,6	3,0	5,0	6,0
(BD)C	0,0	1,1	0,0	0,6	0,0	2,0	6,0
(CD)B	0,1	1,1	0,3	0,6	0,0	2,0	6,0

The grouping {AB} is monophyletic with probability greater than 1/2 if $p_1 + p_2 + p_{13} > 1/2$. Again using Table D1 and Equation (D.1), this occurs when

$$P_\sigma[\{\text{AB}\}] = 1 - \frac{2}{3}e^{-y} - \frac{1}{9}e^{-(3x+y)} \qquad \text{(D.3)}$$

is greater than one-half. Solving for y yields inequality (2).

For the species tree (((AB)C)D), the 4 trees shown in Figure 2*a* are the only consensus trees possible regardless of the set of branch lengths. Proposition 10 guarantees that all cherries incompatible with {AB} (which includes all 2-taxon clades other than {AB} and {CD}) are less probable than {AB}. Therefore, these cherries each have probabilities lower than 1/2 and thus cannot be on the MACT. To show that {CD} cannot occur on the MACT for this species tree, it must be shown that this clade has probability less than 1/2.

The probability that {CD} is monophyletic is

$$\begin{aligned} p_{11} + p_{12} + p_{13} &= \frac{1}{3}e^{-x} - \frac{1}{6}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)} \\ &< \frac{1}{3} + \frac{1}{18}e^{-(3x+y)} < \frac{1}{3} + \frac{1}{18} < \frac{1}{2}. \end{aligned} \qquad \text{(D.4)}$$

APPENDIX 5

Majority-Rule Unresolved Zones, Species Tree ((AB)(CD))

Similar calculations as in Appendix 4 can be performed when the species tree is ((AB)(CD)). For this tree, 3-taxon groups cannot have probability greater than 1/3. For example, the probability for monophyly of {ABC} is (from Table D1 and Equation (D.1))

$$\frac{1}{3}e^{-x} - \frac{5}{18}e^{-(x+y)} < \frac{1}{3}e^{-x} < \frac{1}{3}. \qquad \text{(E.1)}$$

Thus, the MACT for a symmetric 4-taxon species tree cannot have a 3-taxon clade.

All cherries other than {AB} and {CD} are incompatible with these 2 cherries (which occur on this species tree), and from Remark 11, any 2-taxon clades other than {AB} and {CD} have probability less than 1/2 and cannot occur on the MACT. The 2 clades that can occur on the MACT have probabilities

$$P_\sigma(\{\text{AB}\}) = 1 - \frac{2}{3}e^{-y} - \frac{1}{9}e^{-(x+y)} \quad \text{and} \tag{E.2}$$

$$P_\sigma(\{\text{CD}\}) = 1 - \frac{2}{3}e^{-x} - \frac{1}{9}e^{-(x+y)}. \tag{E.3}$$

Setting these functions to be greater than 1/2 yields inequalities (3) and (4).

Here the probability that {AB} is a clade cannot be greater than 1/2 for $y \leq \log(4/3)$ and the probability of clade {CD} cannot be greater than 1/2 for $x \leq \log(4/3)$. These values form asymptotes for the unresolved zone for the symmetric species tree (Fig. 2*b*).

APPENDIX 6

The Too-Greedy Zone, Species Tree ((AB)(CD))

We now show that if the species tree has topology ((AB)(CD)), then the GACT matches the species tree. First note that for this species tree, {AB} and {CD} are always each more probable than any 3-taxon clade. This can be verified by comparing coefficients of the g_{ij} terms in the clade probabilities from Table D1 and by noting that $g_{ij}(T) > 0$ for $T > 0$:

$$\begin{aligned}
P_\sigma(\{\text{AB}\}) &= g_{21}(y)g_{21}(x) + \frac{3}{3}g_{21}(y)g_{22}(x) \\
&\quad + \frac{1}{3}g_{22}(y)g_{21}(x) + \frac{4}{18}g_{22}(y)g_{22}(x), \\
P_\sigma(\{\text{CD}\}) &= g_{21}(y)g_{21}(x) + \frac{1}{3}g_{21}(y)g_{22}(x) \\
&\quad + \frac{3}{3}g_{22}(y)g_{21}(x) + \frac{4}{18}g_{22}(y)g_{22}(x), \\
P_\sigma(\{\text{ABC}\}) &= \frac{1}{3}g_{21}(y)g_{22}(x) + \frac{3}{18}g_{22}(y)g_{22}(x), \\
P_\sigma(\{\text{ABD}\}) &= P_\sigma(\{\text{ABC}\}), \\
P_\sigma(\{\text{ACD}\}) &= \frac{1}{3}g_{22}(y)g_{21}(x) + \frac{3}{18}g_{22}(y)g_{22}(x), \\
P_\sigma(\{\text{BCD}\}) &= P_\sigma(\{\text{ACD}\}).
\end{aligned}$$

Also, from Proposition 10, {AB} and {CD} are the 2 most probable cherries. Thus, the first clade chosen in the greedy algorithm (other than {ABCD}) is either {AB} or {CD} because any other clade would be less probable than one of these 2. If {AB} is most probable, then the remaining compatible clades are {CD}, {ABC}, and {ABD}. Because {CD} is always more probable than {ABC} and {ABD}, {CD} would be chosen after {AB}. Similarly, if {CD} is chosen first, then {AB} is chosen second. Thus, the GACT is always ((AB)(CD)) for this species tree.

APPENDIX 7

The Too-Greedy Zone, Species Tree (((AB)C)D)

In this appendix, we show that when the species tree has topology (((AB)C)D), finding the branch lengths for the too-greedy zone is equivalent to determining the set of branch lengths for which {CD} is more probable than {ABC}.

For the species tree (((AB)C)D) with any set of branch lengths, {ABC} is the most probable 3-taxon clade and {AB} is the most probable 2-taxon clade. These facts can be verified by comparing clade probabilities in Table D1.

In general, {AB} is not more probable than {ABC}, however, because the branch ancestral to A and B but not C might be very short and the branch ancestral to A, B, and C, but not D, might be very long. In the latter case, {ABC} has probability near 1 and {AB} has probability near 1/3.

To show that when the species tree has topology (((AB)C)D), the GACT is always nonmatching if and only if $P[\{\text{CD}\}] > P[\{\text{ABC}\}]$, we consider cases where {ABC} is either (i) more probable than {AB}, (ii–iv) less probable than {AB}, or (v) equally probable as {AB}. In (ii–iv), we also consider whether {CD} is (ii) less probable than {ABC}, (iii) more probable than {ABC}, or (iv) equally probable as {ABC}. Because these cases exhaust all possibilities and because greedy consensus always returns a nonmatching tree in case (iii) and returns a nonmatching tree with probability 1/2 in case (iv), we get the desired result.

(i) $P[\{\text{ABC}\}] > P[\{\text{AB}\}]$. Here {ABC} is the most probable clade other than {ABCD} and is therefore included in the GACT. The remaining compatible clades are {AB}, {AC}, and {BC}. By comparing clade probabilities in Table 2, or by using Proposition 10, we observe that {AB} is the most probable of these 3 clades. Thus, the GACT is (((AB)C)D).

(ii) $P[\{\text{CD}\}] < P[\{\text{ABC}\}] < P[\{\text{AB}\}]$. In this case, {AB} is the most probable clade (other than {ABCD}) and is therefore in the GACT. The remaining compatible clades are {CD}, {ABC}, and {ABD}. Because $P[\{\text{ABD}\}] < P[\{\text{ABC}\}]$ (Table 2), {ABD} cannot be in the GACT; thus the GACT is (((AB)C)D).

(iii) $P[\{\text{ABC}\}] < P[\{\text{CD}\}] < P[\{\text{AB}\}]$. In this case, the GACT is ((AB)(CD)), so $P[\{\text{ABC}\}] < P[\{\text{CD}\}]$ is a sufficient condition for the GACT to be ((AB)(CD)).

(iv) $P[\{\text{ABC}\}] = P[\{\text{CD}\}] < P[\{\text{AB}\}]$. This equality only holds when inequality (5) is an equality, which occurs for points on the boundary of the too-greedy

zone. In this case, the GACT is ((AB)(CD)) or (((AB)C)D), each with probability 1/2.

(v) Finally, if $P[\{ABC\}] = P[\{AB\}]$, then the GACT is (((AB)C)D) because in this case {ABC} and {AB} are the 2 most probable clades.

Having considered all cases, $P[\{ABC\}] < P[\{CD\}]$ if and only if ((AB)(CD)) is the GACT and $P[\{ABC\}] = P[\{CD\}]$ if and only if ((AB)(CD)) is the GACT with probability 1/2. The probabilities of {ABC} and {CD} are given in Equations (D.2) and (D.4), respectively, in Appendix 4. Setting $P(\{CD\}) > P(\{ABC\})$ and solving for y, we obtain inequality (5).

APPENDIX 8

Probabilities of Consensus Trees for Finite Numbers of Loci

To compute the probability of a consensus tree given a finite sample of ℓ gene trees, let ℓ_i be the number of times gene tree i is observed, where i ranges from 1 to k, $\sum_{i=1}^{k} \ell_i = \ell$, and k is the number of possible gene-tree topologies. Let $c(\ell_1, \ldots, \ell_k)$ denote the consensus tree resulting from a particular sample for a particular consensus method c. The probability that a sample results in the consensus tree having topology $\mathcal{T}$ is therefore

$$\sum_{\substack{\ell_1,\ldots,\ell_k \geq 0 \\ \ell_1+\cdots+\ell_k=\ell}} \frac{\ell!}{\ell_1 ! \cdots \ell_k!} p_1^{\ell_1} \cdots p_k^{\ell_k} \, I(c(\ell_1, \ldots, \ell_k) = \mathcal{T}), \tag{H.1}$$

where I is an indicator that the consensus tree has topology $\mathcal{T}$, p_i, $i = 1, \ldots, k$, is the probability that a random gene tree has the ith topology, and the sum is over all nonnegative integer solutions to $\ell_1 + \cdots + \ell_k = \ell$. There are $\binom{\ell+k-1}{k-1}$ terms in the sum (Ross 1998, p. 13), where $k = (2n - 3)!!$ and there are n taxa (Felsenstein 2004). For 4 taxa and 25 loci, the sum has approximately 1.51×10^{10} terms.

Equation (H.1) provides a basis for evaluating probabilities of majority-rule or R^* consensus trees; however, to compute the probabilities of finite-sample greedy consensus trees, probabilities of resolutions of ties must also be taken into account. This can be done by summing over all possible tiebreaks and treating each possible tiebreak as equally likely, rather than randomly breaking ties. The probability of the greedy consensus tree having topology $\mathcal{T}$ can therefore be written as

$$\sum_{\substack{\ell_1,\ldots,\ell_k \geq 0 \\ \ell_1+\cdots+\ell_k=\ell}} \frac{\ell!}{\ell_1 ! \cdots \ell_k!} p_1^{\ell_1} \cdots p_k^{\ell_k} \Bigg[\sum_{b_1 \in B_1} \cdots \sum_{b_r \in B_r(b_1,\ldots,b_{r-1})} \prod_{j=1}^{r} \Pr(b_j) I(c(\ell_1, \ldots, \ell_k, b_1, \ldots, b_r) = \mathcal{T}) \Bigg], \tag{H.2}$$

where B_j, $j = 1, \ldots, r$, denotes the set of possible tiebreaks in the jth round, b_j denotes one way (out of $|B_j|$ possible ways, where $|B_j|$ is the number of elements in B_j) of breaking up a set of tied clade frequencies in the jth round (out of r rounds) of choosing clades for the greedy consensus tree, and $\Pr(b_j) = 1/|B_j|$ is the probability of a particular tiebreak. In general, the set B_j is a function of the choices $b_1, \ldots, b_{j-1}$ in preceding rounds of tiebreaks because the possible tiebreaks in a given round may depend on how previous ties were broken. Thus, the function c in Equation (H.2) has been given additional arguments (compared with Equation (H.1)) so that the consensus tree is a function of both the gene-tree frequencies and the tiebreaks. For n-taxon trees, there are $n - 2$ rounds of tiebreaks, assuming each case when no tiebreaks are necessary (i.e., there is one clade on the list which is uniquely most frequent) is treated as a trivial tiebreak with $|B_j| = 1$.