Theoretical Population Biology **61**, 225–247 (2002)
doi:10.1006/tpbi.2001.1568, available online at http://www.idealibrary.com on IDEAL®

TPB

The Probability of Topological Concordance of Gene Trees and Species Trees

Noah A. Rosenberg

Department of Biological Sciences, Stanford University, Stanford, California 94305; and Program in Molecular and Computational Biology, 1042 W. 36th Place, DRB 155, University of Southern California, Los Angeles, California 90089
E-mail: noahr@usc.edu

Received September 1, 2001

The concordance of gene trees and species trees is reconsidered in detail, allowing for samples of arbitrary size to be taken from the species. A sense of concordance for gene tree and species tree topologies is clarified, such that if the "collapsed gene tree" produced by a gene tree has the same topology as the species tree, the gene tree is said to be *topologically concordant* with the species tree. The term *speciodendric* is introduced to refer to genes whose trees are topologically concordant with species trees. For a given three-species topology, probabilities of each of the three possible collapsed gene tree topologies are given, as are probabilities of monophyletic concordance and concordance in the sense of N. Takahata (1989), *Genetics* 122, 957–966. Increasing the sample size is found to increase the probability of topological concordance, but a limit exists on how much the topological concordance probability can be increased. Suggested sample sizes beyond which this probability can be increased only minimally are given. The results are discussed in terms of implications for molecular studies of phylogenetics and speciation.

Key Words: **population divergence; coalescence; consistency; phylogeny; reciprocal monophyly; human evolution; trichotomy; speciodendricity.**

1. INTRODUCTION

It has long been known that the genealogical history of orthologous genomic regions of several species need not be identical to the history of the species themselves (e.g., Hudson, 1983; Nei, 1986; Neigel and Avise, 1986; Doyle, 1992; Ruvolo, 1994; Maddison, 1997; Nichols, 2001; Nordborg, 2001). Two main phenomena can explain this apparent anomaly. First, ancient coalescence of lineages can occur in an order that differs from the branching order of species. Second, if genes are exchanged between two species that are not sister species, subsequent to their divergence from a common ancestor, gene trees may place those two species together in a clade. This grouping will disagree with the species tree. Depending on the taxa under consideration, gene exchange may result from horizontal gene transfer or from hybridization.

In practice, other causes can explain disagreements between gene genealogies and species tree topologies. The assumption that a genomic region is orthologous across all species studied may be erroneous. Alternatively, if insufficient genetic information is used, a gene tree may be incorrectly inferred, potentially leading to discordance with the species tree.

Understanding the relationship between gene trees and species trees is useful for deducing properties of specific genes (e.g., Ting *et al.*, 2000) and for inference of species phylogenies from discordant gene trees (e.g., Ruvolo, 1997; Satta *et al.*, 2000; Chen and Li, 2001). The fraction of genes whose trees agree with a species tree can also be used to estimate population sizes of ancestral species (e.g., Chen and Li, 2001; Takahata and Satta, 2002).

Applications that use gene trees to study individual genes, species trees, or ancestral population sizes require

0040-5809/02 $35.00

the probability of concordance of gene trees and species trees under a species divergence model. Thus, this probability has been a frequent source of discussion (Hudson, 1983; Nei, 1986; Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Hudson, 1992; Moore, 1995). Assume that three species have equal and constant haploid population sizes (all equal to N) and equal generation times. If one lineage is sampled from each species and if the species tree topology and gene genealogy are known exactly, the concordance probability $P(T_2)$ for the gene tree and species tree is (Hudson, 1983; Nei, 1986)

$$P(T_2) = 1 - \tfrac{2}{3} e^{-T_2}. \quad (1)$$

In (1), T_2 is the quotient of the number of generations that elapsed between the more ancient divergence and the more recent divergence, and the haploid population size N. Equation (1) follows from the fact that the waiting time to the coalescence of two lineages is exponentially distributed with mean 1, in units of N generations (e.g., Tajima, 1983). The probability that the gene tree topology is determined by a coalescence that occurs between the two divergence points is $1 - e^{-T_2}$. In this circumstance, the gene tree is congruent to the species tree. If the gene tree topology is determined by a coalescence prior to the more ancient divergence, an event that has probability e^{-T_2}, three ancestral lineages are present. Then the probability that the most recent coalescence joins the two lineages ancestral to the pair of sister species equals $1/3$. Thus, the concordance probability is $1 - e^{-T_2} + (1/3)\, e^{-T_2}$. The probability of either discordant tree topology is $\frac{1}{2}\,[1 - P(T_2)]$, or $(1/3)\, e^{-T_2}$. Similar reasoning gives corresponding concordance probabilities in cases of four and five species (Pamilo and Nei, 1988).

A natural extension to this work is to determine the effect of increasing the sample sizes above one lineage per species. However, with multiple lineages per species, the meaning of "concordance" of gene trees and species trees is unclear. Using different definitions of concordance, Pamilo and Nei (1988) and Takahata (1989) reached different conclusions about the effect of sample size on concordance probability. Both senses of concordance were somewhat problematic. With Pamilo and Nei's (1988) definition, concordance was difficult to assess analytically, and some possible gene trees were classified so that they were not concordant with *any* of the possible species tree topologies (other problems with the definition were discussed by Takahata, 1989). Although this failure to classify all gene trees as concordant with some species tree can be resolved (Takahata, 1989), it is hard for the definition to accommodate more than three species. Because Pamilo and Nei's sense of concordance depends on a distance measurement between pairs of species (based on mean coalescence times of two lineages, one from each species), the gene tree topology is decided from a pairwise distance matrix. With three species, this decision is straightforward; with more species, however, results may depend on which algorithm for constructing the topology from the matrix is used (for example, UPGMA or neighbor-joining).

Takahata's (1989) definition, though more mathematically tractable and more easily generalizable, had the flaw (recognized by Takahata) that for samples of size one, it did not recover the intuitive definition of concordance used by previous authors (e.g., Hudson, 1983; Pamilo and Nei, 1988), namely that of gene trees and species trees having the same topology. As with Pamilo and Nei's (1988) definition, under Takahata's (1989) definition, gene trees could also be constructed that were not concordant with any species tree topology.

In this article, I reconsider the concordance probability using a precise definition of the *topological concordance* of gene trees and species trees. As described in Section 2, for samples of size 1 taken from each species, this definition coincides with the intuitive sense of agreement between gene trees and species trees. For larger sample sizes, the definition is closely related to Takahata's (1989) use of "consistency." Using the new definition and a three-species divergence model, in Section 3 I calculate the probability that given a gene, a sample of arbitrary size, and a species phylogeny, the gene tree is topologically concordant with the species tree. I also give the probability that gene trees and species trees are *monophyletically concordant*, that is, topologically concordant in such a way that all three species are monophyletic. Using simulations in Section 4, I discuss the effects of divergence times and sample sizes on the topological concordance probability. Implications for studies of phylogeny and speciation are described in Section 5.

This article differs from Takahata's (1989) approach, in that the new definition of concordance enables computation of the likelihoods of all three collapsed gene tree topologies given the species tree topology. Additionally, the present method allows large-sample limiting concordance probabilities (*speciodendricity* probabilities) to be computed fairly easily. Also, when samples differ in size across species, the probabilities of genotype data conditioned on alternate topologies are not equal, and they cannot be calculated from half of one minus the concordance probability. Adjusted likelihoods, incorporating this fact, are given here.

The main question addressed is: "conditioned on the species tree topology and assuming no gene exchange between species, what is the probability that a tree of orthologous genes is topologically concordant with a species tree?" Because I am concerned only with the

relationship of the genealogical shape of gene trees to species trees, several important issues are not considered. First, I assume that full knowledge of gene trees is available. In practice, however, gene trees are inferred from the DNA sequences of copies of a gene in different individuals. Error in reconstructed gene trees can be introduced by stochastic differences in the number of mutational changes that have happened along different lineages, by heterogeneity in mutation rates across sites, by failure to account for intragenic recombination, by problems with heuristics that underlie phylogenetic inference algorithms, or by genotyping errors. Some of these issues have been studied by Saitou and Nei (1986).

2. TERMINOLOGY

The terms "species" and "population" are imperfect for the concept needed here, namely that of organisms that are grouped with a common label and that are treated as having descended from the bifurcation (or multifurcation) of a similar ancestral group. Each group maintains the same label for the entire period between its origin and its bifurcation into two new groups (if such an event occurs). For lack of a better term, I refer to such groups as "species." It is to be understood that these groups can be different species in the traditional "biological species concept" sense, or different populations within a traditional species.

Because the coalescent approach treats time as increasing backwards from the present, I adopt the same convention. However, I still use "before" to mean "more ancient," and I employ "later" and "after" to mean "more recent." The directionality of other words that refer to time should be clear from the context.

2.1. Congruence and Topological Concordance

Many terms have been used to codify the concept of a gene tree and species tree having the same topology. Gene trees and species trees have been referred to as being "in agreement," "concordant," "congruent," "consistent," "identical," and "isomorphic," and gene trees as "matching" or "tracking" the species tree. For the purposes of this article, supposing that one lineage is sampled from each of several species, the gene tree and species tree are said to be *congruent* if and only if they have the same topology (Figs. 1i and 1ii).

If more than one lineage is sampled from any of the species, then the gene tree has more tips than the species tree and the two cannot have the same topology. Thus, the words "congruent," "identical" and "isomorphic" are inappropriate when sample sizes are greater than one. In this situation, I refer to a gene tree and species tree as being *topologically concordant* if and only if the *collapsed gene tree* is congruent to the species tree. To construct the collapsed gene tree from a gene tree, proceed backwards in time until a coalescence of lineages occurs between two

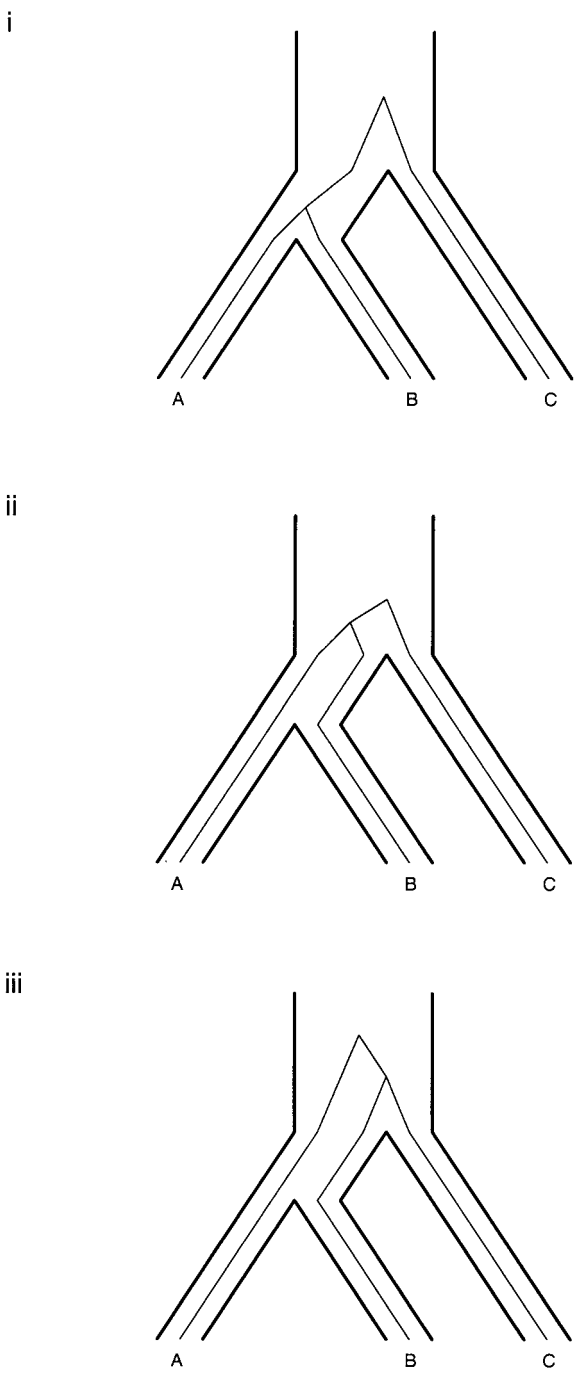

FIG. 1. *Congruence* of gene trees and species trees. *A*, *B*, and *C* are present-day species. (i) Gene tree that is both congruent and Takahata-congruent to the species tree. (ii) Gene tree that is congruent but not Takahata-congruent to the species tree. (iii) Gene tree that is neither congruent nor Takahata-congruent to the species tree.

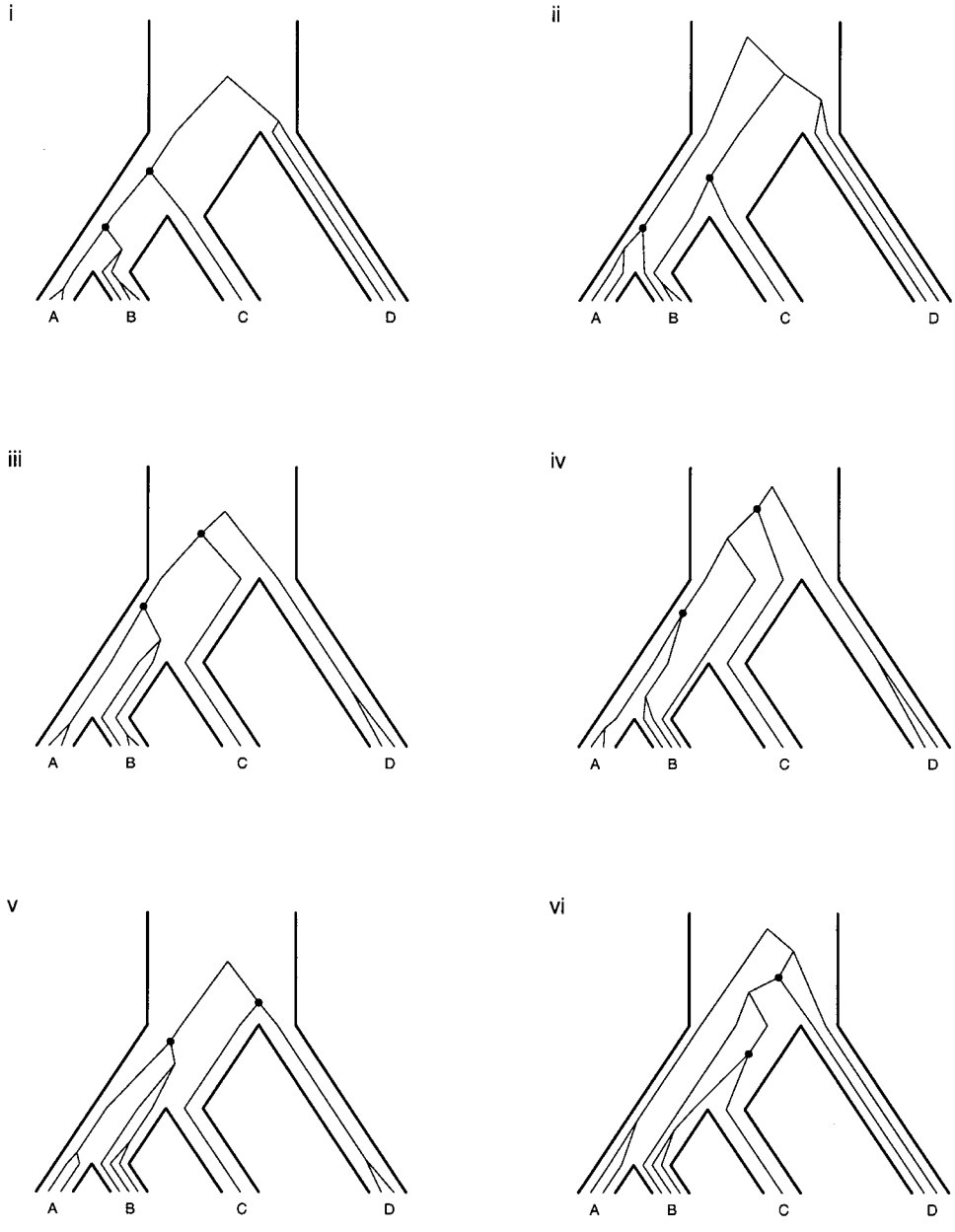

FIG. 2. *Concordance* of gene trees and species trees. *A*, *B*, *C*, and *D* are present-day species. Circles indicate interspecific coalescences that are used in determining the collapsed gene tree. The collapsed gene trees for (i), (ii), (iii), and (iv) all have topology (((*AB*) *C*) *D*). For (v) the collapsed gene tree has topology ((*AB*)(*CD*)), and for (vi), the topology is (((*BC*) *D*) *A*). The six trees represent the six classes partitioned by the terms *topological concordance*, *Takahata-concordance*, *monophyletic concordance*, and *all species monophyletic*. (i) All four terms apply. (ii) Topological concordance, Takahata-concordance. (iii) Topological concordance, monophyletic concordance, all species monophyletic. (iv) Topological concordance. (v) All species monophyletic. (vi) None of the terms applies.

species. Group the two species involved in this coalescence into a clade. Continue backwards in time until another coalescence occurs between two clades (where "clade" is understood to subsume "species" as a special case). If both clades involved in this coalescence have already experienced inter-clade coalescences, ignore the event. If one or neither of the clades has already had inter-clade coalescences, group these two clades into a larger clade. Proceed backwards in time until all species have been involved in inter-clade coalescences. Examples of collapsed gene trees are described in Fig. 2. Note that this definition of topological concordance between gene trees

and species trees recovers the definition of congruence when only one lineage is sampled from each species, because the collapsed gene tree will be identical to the gene tree itself. Also, *topological concordance probability* describes the quantity P^* that was briefly discussed by Takahata (1989).

For the case of three species, the new definition of "topologically concordant" is similar to "consistent," as given by Takahata (1989). With three species, Takahata (1989) defined the gene tree and species tree to be "consistent" if the most recent interspecific coalescence occurred between the pair of sister species in the phylogeny, and if this event took place later than the first bifurcation of the ancestral group to all three species (Figs. 2i and 2ii). The difference between the definition of topological concordance here and Takahata's (1989) use of "consistency" is that in the present formulation, if the most recent interspecific coalescence happened prior to the first bifurcation of the ancestral group, the gene tree and species tree would still be topologically concordant if this event took place between the sister species (Figs. 2iii and 2iv). In this same situation, Takahata's (1989) definition would label them "inconsistent." In many circumstances, the probability that the most recent coalescence occurred before the original bifurcation is negligibly small, so that Takahata's (1989) definition is often a reasonable approximation to the one here, as will be seen below.

I use "concordant" because "consistent" has many different meanings in phylogenetic contexts. Henceforth I distinguish between the new definition and that of Takahata using "topologically concordant" (or simply "concordant") and "Takahata-concordant" (with samples of size one, "congruent" and "Takahata-congruent"). Stated precisely, a gene tree taken from any number of species is *Takahata-concordant* with the species tree if and only if (a) the collapsed gene tree is congruent to the species tree, and (b) the collapsed gene tree contains no coalescences prior to the most ancient species divergence. A gene tree is *topologically concordant* with the species tree if and only if (a) holds. In the case of one lineage per species, it is acceptable to use "topologically concordant" and "congruent" interchangeably.

It is useful to define another form of concordance, similar to Neigel and Avise's (1986) "phylogenetic status I" and Mountain and Cavalli-Sforza's (1997) "consistency." A gene tree and species tree are defined to be *monophyletically concordant* or *M-concordant* if and only if (a) the gene tree and species tree are topologically concordant, and (b) for each species, all lineages sampled from that species form a monophyletic group. For the case of two species, examples of M-concordance are given in Figs. 2i and 2iii. If only two species are considered, M-concordance is equivalent to "reciprocal monophyly" (e.g., Moritz, 1994).

In summary, we have the following relationships between the concepts, producing the six classes of genealogies shown in Figs. 2i–2vi:

1. Monophyletic concordance implies topological concordance.
2. Takahata-concordance implies topological concordance.
3. Monophyletic concordance implies all species are monophyletic.
4. Topological concordance and all species monophyletic imply monophyletic concordance.

2.2. Speciodendric Genes

"Orthology" was defined by Fitch (1970) to include genes whose homology was the result of speciation and subsequent descent, with no duplication. According to Fitch (1970, p. 113), for orthologous genes, "the history of the gene reflects the history of the species." Although meanings of this term have since diversified (Ouzounis, 1999), recent usage of "orthologous" has focused on the first part of Fitch's idea: genes that have diverged via speciation as opposed to duplication. No term has come to have the meaning of the second part: genes whose trees reflect the species tree. To fill this gap in terminology, I propose the term *speciodendric*. Stated precisely, a gene is *speciodendric* with respect to a given set of species if the gene tree constructed from *all copies of the gene in all of the species in the set* is topologically concordant with the species tree. It is understood that only genes that are homologous (*sensu* Fitch, 2000) across a set of species can have this property.

Note that Fitch's (2000) re-definition of "orthology" contains a misleading statement about gene trees made from orthologous genes (p. 228). It is not true that all orthologous genes are speciodendric: consider Fig. 3ii, in which genes are orthologous, but not speciodendric with respect to the species shown. It is also not true that all speciodendric genes are orthologous. To see this, consider Fig. 3iii, in which xenologous genes, or genes for which transfers across species are a part of their histories, *are* speciodendric for the three species (of course, xenology need not imply speciodendricity: in Fig. 3iv, xenologous genes are not speciodendric). It is even possible for paralogous genes, those whose histories reflect duplication, to be speciodendric, and vice versa (Fig. 3v).

3. THEORY

In this section, I compute the probability that a gene tree and species tree are topologically concordant in the case of three species. This probability depends on sample sizes and on demographic histories of the three species.

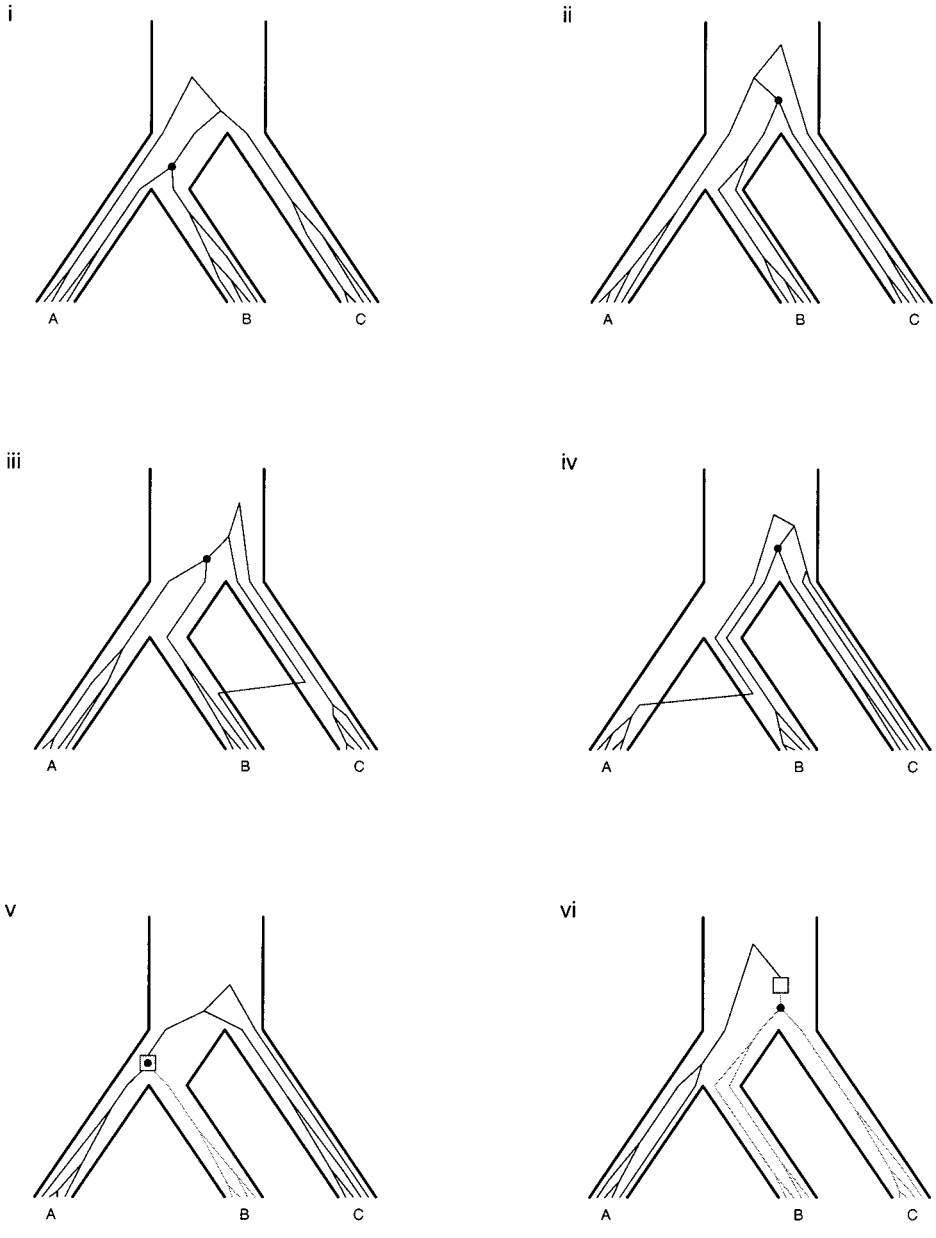

FIG 3. Gene trees for speciodendric and non-speciodendric genes. Circles indicate interspecific coalescences that are used in determining the collapsed gene tree. Squares indicate gene duplications. After a duplication event, "new" copies of genes are drawn more lightly than the old copies. For ease of representation, each species is treated as having only four lineages. (i) Speciodendric orthologous genes. (ii) Non-speciodendric orthologous genes. (iii) Speciodendric xenologous genes. (iv) Non-speciodendric xenologous genes. (v) Speciodendric paralogous genes. (vi) Non-speciodendric paralogous genes.

Suppose that an ancestral group of organisms separated into two descendant clades $t_3 + t_2$ generations in the past. One of the clades separated further into two groups (A and B) t_3 generations in the past. It is simplest to assume that each modern and ancestral species has had constant haploid population size N during its entire existence, so that time can be easily scaled in coalescent units (here, t generations equals $T = \frac{t}{N}$ coalescent time units). It would be straightforward to assume that the size of an ancestral species is the sum of the sizes of its descendants: then the scaling of time would be different before and after the divergence of the ancestor. Before

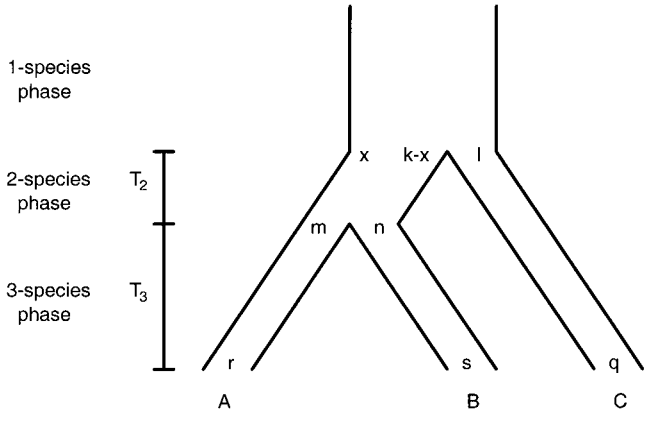

FIG. 4. Three-species divergence model. The quantities r, s, and q are numbers of sampled lineages. T_3 and T_2 are lengths of time periods in coalescent units. The remaining variables, m, n, l, x, and $k-x$, all represent numbers of ancestral lineages. Note that the variable x, which represents the number of ancestral lineages to species A present at time T_3+T_2, is only sensible when no interspecific coalescences occur during the two-species phase.

the divergence, t generations would equal $\frac{t}{2N}$ coalescent time units and after the divergence, t generations would equal $\frac{t}{N}$ coalescenttime units. This complication, as well as deterministic fluctuations in the number of individuals in each species, or different population sizes or mating systems across species, could be accommodated by deducing results in coalescent units and rescaling to units of generations (see Nordborg, 2001).

The history of the three species is divided into "phases," in each of which the demographic properties of the three species are constant for the duration of the phase. As soon as a species divergence is reached, a new phase is entered. The model shown in Fig. 4 diagrams the "three-species phase" the "two-species phase," and the "one-species phase" or "ancestral phase." Looking backwards in time, as soon as the first interspecific coalescence occurs, the collapsed gene tree is determined. If this event occurs between lineages ancestral to species A and B, the gene tree is topologically concordant with the species tree.

3.1. Takahata-Concordance Probability

In the present, r, s, and q lineages are sampled from species A, B, and C, respectively. Let $g_{ij}(T)$ be the probability that i lineages derive from j lineages that existed T coalescent time units in the past (e.g., Tavaré, 1984, Eq. (6.1)),

$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!\,(k-j)!\, i_{(k)}}, \quad (2)$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ for $k \geqslant 1$ with $a_{(0)} = 1$; and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geqslant 1$ with $a_{[0]} = 1$. $g_{ij}(T) = 0$, except when $1 \leqslant j \leqslant i$. Note that $g_{ij}(0) = \delta_{ij}$, where δ_{ij} is Kronecker's delta. Also, of course $\sum_{j=1}^{i} g_{ij}(T) = 1$ for all T. For example, (2) yields

$$g_{11}(T) = 1 \quad \begin{matrix} g_{21}(T) = 1-e^{-T} \\ g_{22}(T) = e^{-T} \end{matrix} \quad \begin{matrix} g_{31}(T) = 1-\frac{3}{2}e^{-T}+\frac{1}{2}e^{-3T} \\ g_{32}(T) = \frac{3}{2}e^{-T}-\frac{3}{2}e^{-3T} \\ g_{33}(T) = e^{-3T}. \end{matrix}$$

The probability that species A and B are respectively represented by m and n ancestral lineages at time T_3 is $g_{rm}(T_3)\, g_{sn}(T_3)$. The probability that the $m+n$ lineages in the ancestral species at time T_3 coalesce to k lineages at time T_3+T_2 is then $g_{m+n,k}(T_2)$. During the process of coalescence of these $m+n$ lineages to k lineages in the two-species phase, denote the probability that an interspecific coalescence occurs between a lineage of species A and a lineage of species B by $F_k^{A,B}(m, n, 0)$.

More generally, suppose that in an ancestral species, a, b, and c lineages represent descendant species A, B, and C, respectively, and that coalescences take place until the total number of lineages is k. Then let $F_k^{A,B}(a, b, c)$ be the probability that at least one interspecific coalescence occurs during this process, *and that the most recent interspecific coalescence joins a lineage from species A and a lineage from species B*. Similarly, $F_k^{A,C}(a, b, c)$ and $F_k^{B,C}(a, b, c)$ are the probabilities that an interspecific coalescence occurs and that the most recent interspecific coalescence is between lineages from species A and C, and lineages from species B and C, respectively. Values of $F_k^{A,B}(a, b, c)$ can be computed as in the Appendix; some are given in Table I. $F_k^{A,B}(m, n, 0)$ is equal to Takahata's (1989, Eq. (11) and Table 1) H_{jk}, where j in his notation corresponds to $m+n$ here.

The probability of Takahata-concordance is (equivalently to Eq. (14) of Takahata, 1989)

$$P_T(r, s, q, T_3, T_2) = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} g_{rm}(T_3)\, g_{sn}(T_3) \times g_{m+n,k}(T_2)\, F_k^{A,B}(m, n, 0). \quad (3)$$

Intuitively, (3) is the conditional probability of Takahata-concordance given configurations of lineages throughout the history of the three species, summed over possible configurations.

TABLE I

Values of $F_k^{A,B}(a, b, c)$

(a, b)	$c=1$	$c=2$	$c=3$	$c=4$	$c=1$	$c=2$	$c=3$	$c=4$
	$k=1$ (same values as for $k=2$)				$k=3$			
(1,1)	0.333	0.222	0.167	0.133	0	0.167	0.150	0.127
(2,1)	0.389	0.261	0.197	0.158	0.333	0.250	0.193	0.157
(2,2)	0.478	0.333	0.257	0.209	0.467	0.331	0.256	0.209
(3,1)	0.417	0.280	0.211	0.169	0.400	0.277	0.210	0.169
(3,2)	0.523	0.372	0.289	0.237	0.520	0.371	0.289	0.237
(3,3)	0.578	0.422	0.333	0.276	0.577	0.422	0.333	0.276
(4,1)	0.433	0.291	0.219	0.176	0.427	0.290	0.219	0.176
(4,2)	0.551	0.395	0.309	0.255	0.550	0.395	0.309	0.255
(4,3)	0.611	0.454	0.362	0.302	0.611	0.453	0.362	0.302
(4,4)	0.648	0.491	0.397	0.333	0.648	0.491	0.397	0.333
	$k=4$				$k=5$			
(1,1)	0	0	0.100	0.107	0	0	0	0.067
(2,1)	0	0.200	0.180	0.152	0	0	0.133	0.137
(2,2)	0.400	0.320	0.253	0.208	0	0.267	0.241	0.204
(3,1)	0.300	0.260	0.206	0.167	0	0.200	0.190	0.163
(3,2)	0.500	0.368	0.288	0.237	0.400	0.352	0.285	0.236
(3,3)	0.571	0.421	0.333	0.276	0.543	0.416	0.332	0.276
(4,1)	0.387	0.283	0.217	0.175	0.267	0.260	0.212	0.173
(4,2)	0.542	0.394	0.309	0.254	0.508	0.389	0.308	0.254
(4,3)	0.609	0.453	0.362	0.302	0.599	0.452	0.362	0.302
(4,4)	0.647	0.491	0.397	0.333	0.644	0.490	0.397	0.333

Note. For a, b, and c lineages representing species A, B, and C, respectively, in coalescing to k total lineages, $F_k^{A,B}(a, b, c)$ is the probability that the most recent interspecific coalescence joins a lineage of A and a lineage of B.

3.2. Topological Concordance Probability

To compute the probability of *topological concordance*, a term must be added to the Takahata-concordance probability for the probability of all of the following: (a) no interspecific coalescences happen in the two-species phase; (b) the most recent interspecific coalescence happens in the one-species phase; and (c) this coalescence joins ancestral lineages of species A and B. Assuming that m and n lineages from species A and B are present at time T_3, and that these lineages have k total ancestors at time T_3+T_2, the probability that *no* interspecific coalescences happen in the two-species phase is $1-F_k^{A,B}(m, n, 0)$. All the coalescences are intraspecific, and at time T_3+T_2, there are, say, X_1 and X_2 lineages ancestral to species A and species B, respectively. Because k total lineages are present at time T_3+T_2, $X_1+X_2=k$. Also, each species is represented by at least one lineage, so $1 \leqslant X_1, X_2 \leqslant k-1$.

In order to determine probabilities of events in the one-species phase, we will need to consider all possible values of X_1 and X_2. Thus, $Pr(X_1=x, X_2=k-x \mid X_1+X_2=k)$ is needed. This probability, henceforth denoted $W_{(m,n),(x,k-x)}(T_2)$, depends on the numbers of ancestral lineages to species A and B at time T_3 (m and n), the number of lineages at time T_3+T_2 (k), and the duration of the two-species phase (T_2). Using Bayes's theorem, we have

$$\begin{aligned} W_{(m,n),(x,k-x)}(T_2) &= Pr(X_1=x, X_2=k-x \mid X_1+X_2=k) \\ &= Pr(X_1=x, X_2=k-x)/Pr(X_1+X_2=k) \\ &= Pr(X_1=x)\,Pr(X_2=k-x)/Pr(X_1+X_2=k) \\ &= \frac{g_{mx}(T_2)\,g_{n,k-x}(T_2)}{\sum_{i=1}^{k-1} g_{mi}(T_2)\,g_{n,k-i}(T_2)}. \end{aligned} \tag{4}$$

Simultaneous to the entry of lineages from A and B into the one-species phase, lineages ancestral to species C also enter the one-species phase. The probability that species C is represented by l ancestral lineages at time T_3+T_2 is $g_{ql}(T_3+T_2)$.

The last quantity needed for the calculation is the probability $F_1^{A,B}(a, b, c)$ that for a, b, and c lineages from species A, B, and C present at the ancestral divergence, the most recent interspecific coalescence occurs between lineages ancestral to species A and B. This probability is necessary because if the most recent interspecific coalescence involves a lineage from species C, the collapsed gene tree will be discordant with the species tree.

Combining the various components, the topological concordance probability is

$$\begin{aligned} P_C(r, s, q, T_3, T_2) = \sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n}\Bigg[& g_{rm}(T_3)\,g_{sn}(T_3)\,g_{m+n,k}(T_2) \\ & \times\Bigg[F_k^{A,B}(m, n, 0)+[1-F_k^{A,B}(m, n, 0)] \\ & \times\sum_{x=1}^{k-1}\Bigg[W_{(m,n),(x,k-x)}(T_2) \\ & \times\sum_{l=1}^{q}[g_{ql}(T_3+T_2)\,F_1^{A,B}(x, k-x, l)]\Bigg]\Bigg]\Bigg]. \end{aligned} \tag{5}$$

For samples of size 1 in each species, (5) recovers the formula given in (1), while the Takahata-concordance probability in (3) gives $1-e^{-T_2}$ (as was noted by Takahata, 1989).

3.3. Speciodendricity Probability

As a special case, when sample sizes equal the total number of copies of the gene in the respective species, (5) gives the probability that the gene is speciodendric. It will be seen in Section 4.2 that the topological

concordance probability often converges rapidly as sample size increases. Because population sizes tend to be large, so that the probability of speciodendricity is close to the limit given by the abstract case of an infinite sample size, a useful approximation to the speciodendricity probability is obtained by substituting ∞ into (5) in place of r, s, and q. Thus, the probability of speciodendricity is approximately

$$P_S(T_3, T_2) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{m+n} \Bigg[g_m(T_3)\, g_n(T_3)\, g_{m+n,k}(T_2) \times \Bigg[F_k^{A,B}(m,n,0) + [1 - F_k^{A,B}(m,n,0)] \times \sum_{x=1}^{k-1} \Bigg[W_{(m,n),(x,k-x)}(T_2) \times \sum_{l=1}^{\infty} [g_l(T_3+T_2)\, F_1^{A,B}(x,k-x,l)] \Bigg] \Bigg] \Bigg], \tag{6}$$

where $g_j(T)$ is the large-sample limiting probability that at time T, a sample has j ancestral lineages (Tavaré, 1984, Eqs. (6.3) and (6.4)).

3.4. Probabilities of the Alternate Topologies

If sample sizes differ between species A and B, then the probabilities of alternate topologies $((AC)\,B)$ and $((BC)\,A)$ (this notation is the same as in, for example, Pamilo and Nei, 1988) cannot simply be obtained by halving the probability of discordance. The probabilities of these topologies are analogous to (5), except that each topology can only be obtained if the most recent interspecific coalescence occurs in the ancestral phase. Thus, the expressions are simpler than the corresponding expression for topology $((AB)\,C)$ given in (5):

$$Q_{((AC)\,B)}(r,s,q,T_3,T_2) = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \Bigg[g_{rm}(T_3)\, g_{sn}(T_3)\, g_{m+n,k}(T_2) \times [1 - F_k^{A,B}(m,n,0)] \sum_{x=1}^{k-1} \Bigg[W_{(m,n),(x,k-x)}(T_2) \times \sum_{l=1}^{q} [g_{ql}(T_3+T_2)\, F_1^{A,C}(x,k-x,l)] \Bigg] \Bigg] \tag{7}$$

$$Q_{((BC)\,A)}(r,s,q,T_3,T_2) = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \Bigg[g_{rm}(T_3)\, g_{sn}(T_3)\, g_{m+n,k}(T_2) \times [1 - F_k^{A,B}(m,n,0)] \sum_{x=1}^{k-1} \Bigg[W_{(m,n),(x,k-x)}(T_2) \times \sum_{l=1}^{q} [g_{ql}(T_3+T_2)\, F_1^{B,C}(x,k-x,l)] \Bigg] \Bigg]. \tag{8}$$

By adding P_C (Eq. (5)), $Q_{((AC)\,B)}$, and $Q_{((BC)\,A)}$, it is easily verified that the sum of the probabilities of the three collapsed gene tree topologies equals 1.

For a trifurcation (that is, $T_2 = 0$), (5) simplifies considerably. It no longer makes sense to describe a topological concordance probability. Rather, the interpretation here is that the probability that the collapsed gene tree has topology $((AB)\,C)$ is

$$P_{tri}(r,s,q,T_3,T_2) = \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{l=1}^{q} g_{rm}(T_3)\, g_{sn}(T_3)\, g_{ql}(T_3) \times F_1^{A,B}(m,n,l). \tag{9}$$

Similar equations result for the probabilities of the other two topologies, $((AC)\,B)$ and $((BC)\,A)$. Of course, if all sample sizes are equal, each topology has probability 1/3.

3.5. Monophyletic Concordance Probability

Unlike the topological concordance probability, the *monophyletic concordance* probability is not simply equal to 1 if only two species are considered. For the present, consider species A and B only. Because both species must be monophyletic for the gene tree and species tree to be M-concordant, the only interspecific coalescence must join a lineage ancestral to all lineages of species A and a lineage ancestral to all lineages of species B. In other words, the m and n lineages ancestral to species A and B at the divergence time must coalesce to two lineages without experiencing any interspecific coalescences. The probability that this happens is $1 - F_2^{A,B}(m,n,0)$. Therefore, for two species, the M-concordance probability is

$$P_{M2}(r,s,T_3) = \sum_{m=1}^{r} \sum_{n=1}^{s} g_{rm}(T_3)\, g_{sn}(T_3) \times [1 - F_2^{A,B}(m,n,0)]. \tag{10}$$

Special cases of (10) were obtained by Tajima (1983) and Takahata and Nei (1985). Using the values of $g_{ij}(T)$ from (2), it is straightforward to show that (10) reduces to Tajima's (1983, Fig. 5a) solution for $r=s=2$, or $(1-\frac{2}{3}e^{-T_3})^2$. Equation (10) recovers the formula that describes simulations of M-concordance performed by Neigel and Avise (1986).

For the three species A, B, and C, the calculation of M-concordance probability can be separated into two parts, based on the value of k, the total number of lineages ancestral to species A and B at time T_3+T_2.

If $k=1$, then monophyletic concordance occurs if and only if (a) only one interspecific coalescence happens during the two-species phase, and this coalescence is the most ancient coalescence in that phase, and (b) the l ancestral lineages to species C in the ancestral phase coalesce to one lineage without coalescing with the one lineage ancestral to species A and B. The probability of (a) is $1-F_2^{A,B}(m,n,0)$. The probability of (b) equals $1-F_2^{A,C}(1,0,l)$, or equivalently, $1-F_2^{B,C}(0,1,l)$. This probability equals $\frac{2}{l(l+1)}$ (Property 5 in the Appendix).

If $k>1$, then M-concordance requires all of the following: (a) no interspecific coalescences occur during the two-species phase; (b) no interspecific coalescences occur during the one-species phase until all lineages have coalesced to three lineages; and (c) the last three lineages coalesce in the order given by the species tree. The probability of (a) is $1-F_k^{A,B}(m,n,0)$; (b) has probability $1-F_3^{A,B}(x,k-x,l)-F_3^{A,C}(x,k-x,l)-F_3^{B,C}(x,k-x,l)$; and the probability of (c) is simply $1/3$. Thus, the probability of M-concordance is

$$\begin{aligned}P_{M3}(r,s,q,T_3,T_2)&=\sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n}\sum_{l=1}^{q}\Bigg[g_{rm}(T_3)\,g_{sn}(T_3)\,g_{m+n,k}(T_2)\\&\times g_{ql}(T_3+T_2)\Bigg[\delta_{k,1}[1-F_2^{A,B}(m,n,0)]\frac{2}{l(l+1)}\\&+(1-\delta_{k,1})[1-F_k^{A,B}(m,n,0)]\\&\times\sum_{x=1}^{k-1}\Bigg[W_{(m,n),(x,k-x)}(T_2)[1-F_3^{A,B}(x,k-x,l)\\&-F_3^{A,C}(x,k-x,l)-F_3^{B,C}(x,k-x,l)]\frac{1}{3}\Bigg]\Bigg]\Bigg].\end{aligned}\tag{11}$$

The probability that all species are monophyletic (but not necessarily that M-concordance is achieved) is obtained from (11) simply by removing the factor of $1/3$:

$$\begin{aligned}P_{monophyly}(r,s,q,T_3,T_2)&=\sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n}\sum_{l=1}^{q}\Bigg[g_{rm}(T_3)\,g_{sn}(T_3)\,g_{m+n,k}(T_2)\\&\times g_{ql}(T_3+T_2)\Bigg[\delta_{k,1}[1-F_2^{A,B}(m,n,0)]\frac{2}{l(l+1)}\\&+(1-\delta_{k,1})[1-F_k^{A,B}(m,n,0)]\\&\times\sum_{x=1}^{k-1}[W_{(m,n),(x,k-x)}(T_2)\,[1-F_3^{A,B}(x,k-x,l)\\&-F_3^{A,C}(x,k-x,l)-F_3^{B,C}(x,k-x,l)]]\Bigg]\Bigg].\end{aligned}\tag{12}$$

In case of a trifurcation ($T_2=0$), monophyly is obtained if no interspecific coalescences occur until three lineages remain. Thus, the probability of monophyly in this case is

$$\begin{aligned}P_{Mtri}(r,s,q,T_3,0)&=\sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{l=1}^{q}g_{rm}(T_3)\,g_{sn}(T_3)\,g_{ql}(T_3)\\&\times[1-F_3^{A,B}(m,n,l)\\&-F_3^{A,C}(m,n,l)-F_3^{B,C}(m,n,l)].\end{aligned}\tag{13}$$

3.6. Comparison of Three Types of Concordance

A comparison of analytically calculated probabilities of different types of concordance is given in Table II. As can be observed from the table, topological concordance probability can be substantially larger than Takahata-concordance probability when T_2 is small, and when either the numbers of sampled lineages are small or T_3 is large. Under these conditions, it is unlikely that an interspecific coalescence occurs during the two-species phase, so that the most recent interspecific coalescence frequently takes place in the one-species phase. If the coalescence occurs in accord with the species tree topology, this occurrence can be counted towards the topological concordance probability but not towards the Takahata-concordance probability. The discrepancy between the two probabilities is smaller at large values of the internodal time T_2. If T_3 and T_2 are held constant, this discrepancy decreases with increasing sample size. Analytically computed values in Table II agree with closely with Takahata's (1989, Table 3) work, in which Takahata-concordance and topological concordance probabilities were simulated at many of the same values of (r,s,q,T_3,T_2) shown in Table II.

TABLE II

Probabilities of Concordance in the Three-Species Divergence Model

(r, s, q)	T_3	T_2	Takahata-concordance probability	Topological concordance probability	Monophyly probability (all three species)	Monophyletic concordance probability (all three species)	Monophyletic concordance probability (species A and B)
(1,1,1)	Any value	0	0	0.333	1	0.333	1
		0.05	0.049	0.366	1	0.366	
		0.5	0.393	0.596	1	0.596	
		5	0.993	0.996	1	0.996	
(2,2,1)	0.05	0	0	0.469	0.048	0.016	0.134
		0.05	0.169	0.551	0.056	0.019	
		0.5	0.762	0.855	0.109	0.049	
		5	0.998	0.9990	0.134	0.133	
	0.5	0	0	0.413	0.247	0.082	0.355
		0.05	0.118	0.476	0.260	0.092	
		0.5	0.628	0.763	0.329	0.175	
		5	0.997	0.998	0.355	0.353	
	5	0	0	0.334	0.989	0.330	0.991
		0.05	0.049	0.367	0.989	0.362	
		0.5	0.396	0.598	0.991	0.590	
		5	0.993	0.996	0.991	0.987	
(2,2,2)	0.05	0	0	0.333	0.011	0.004	0.134
		0.05	0.169	0.445	0.015	0.005	
		0.5	0.762	0.837	0.059	0.028	
		5	0.998	0.9990	0.133	0.132	
	0.5	0	0	0.333	0.124	0.041	0.355
		0.05	0.118	0.410	0.137	0.049	
		0.5	0.628	0.746	0.234	0.127	
		5	0.997	0.998	0.354	0.352	
	5	0	0	0.333	0.983	0.328	0.991
		0.05	0.049	0.366	0.984	0.360	
		0.5	0.396	0.597	0.987	0.589	
		5	0.993	0.996	0.991	0.987	
(5,5,1)	0.05	0	0	0.674	0.0002	0.00007	0.002
		0.05	0.604	0.859	0.0003	0.0001	
		0.5	0.989	0.994	0.001	0.0005	
		5	0.99996	0.99997	0.002	0.002	
	0.5	0	0	0.519	0.031	0.010	0.082
		0.05	0.261	0.630	0.035	0.012	
		0.5	0.846	0.907	0.066	0.030	
		5	0.9990	0.9994	0.082	0.081	
	5	0	0	0.335	0.978	0.326	0.982
		0.05	0.050	0.368	0.978	0.358	
		0.5	0.399	0.599	0.981	0.584	
		5	0.993	0.996	0.982	0.978	
(5,5,5)	0.05	0	0	0.333	0.000002	0.0000005	0.002
		0.05	0.604	0.734	0.000005	0.000002	
		0.5	0.989	0.992	0.0003	0.0001	
		5	0.99996	0.99997	0.002	0.002	
	0.5	0	0	0.333	0.006	0.002	0.082
		0.05	0.261	0.502	0.007	0.003	
		0.5	0.846	0.892	0.030	0.014	
		5	0.9990	0.999	0.081	0.081	
	5	0	0	0.333	0.967	0.322	0.982
		0.05	0.050	0.367	0.968	0.354	
		0.5	0.399	0.599	0.975	0.580	
		5	0.993	0.996	0.982	0.978	

Note. Notation is defined in Fig. 4. Note that "monophyletic concordance" and "both species monophyletic" are equivalent for two species. Takahata-concordance probability depends only on r, s, T_3, and T_2; monophyletic concordance probability for species A and B depends only on r, s, and T_3. Takahata-concordance, topological concordance, three-species monophyly, three-species monophyletic concordance, and two-species monophyletic concordance probabilities are computed using (3), (5), (12), (11), and (10), respectively.

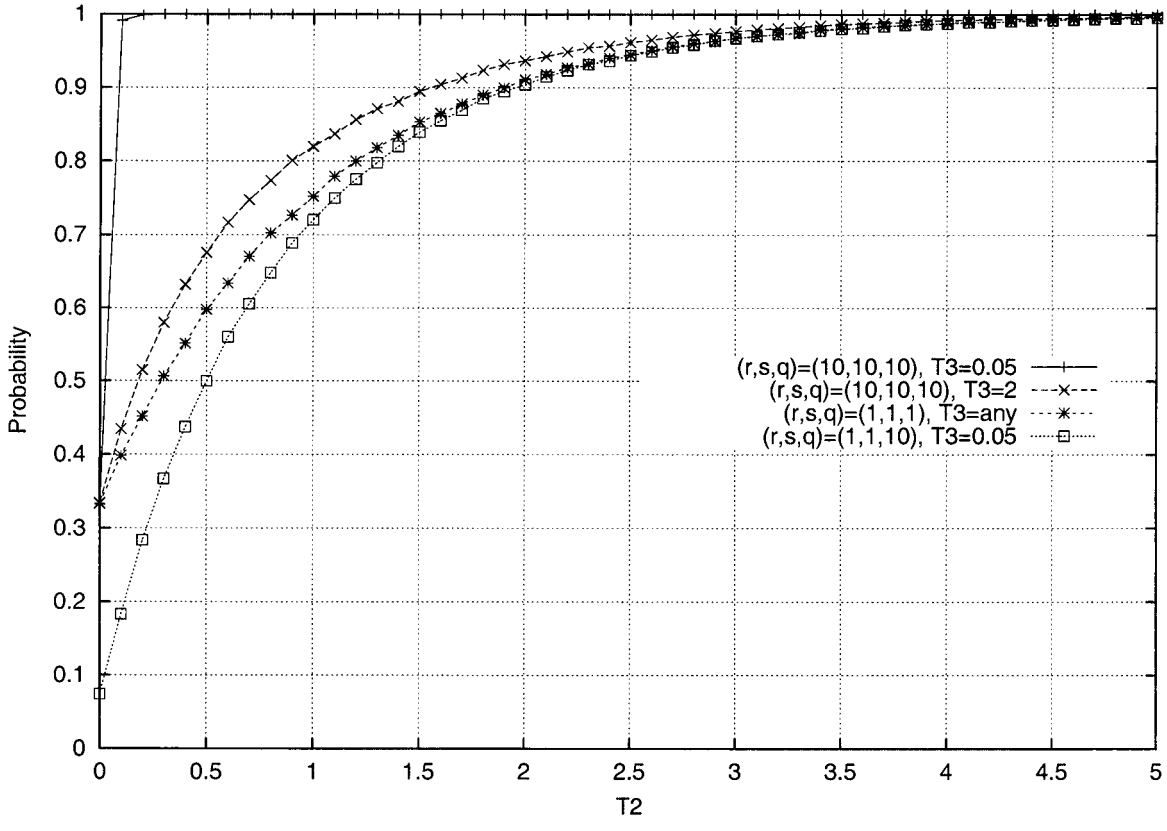

FIG. 5. Probability of topological concordance as a function of T_2, the time between divergences in the three-species model. Each point is based on 100,000 simulated instances of the model. Sample sizes in species A, B, and C were r, s, and q respectively, and the time since the divergence of sister species was T_3.

Like Takahata-concordance, M-concordance is a stricter condition than topological concordance. Thus, analytically computed probabilities of three-species M-concordance are at most equal to corresponding probabilities of topological concordance (Table II). For samples of size one, M-concordance and topological concordance have the same meaning. Unlike the Takahata-concordance and topological concordance probabilities, the three-species M-concordance probabilities *decrease* with increasing sample size; thus, values of (r, s, q, T_3, T_2) can be chosen for which the M-concordance probability is either greater or less than the Takahata-concordance probability. The decrease in M-concordance probability with sample size, which occurs rapidly at small T_3, results from the fact that the number of lineages in the two-species phase increases with sample size. Consequently, the chances of an interspecific coalescence more recently than when two ancestral lineages are reached are increased.

As T_2 increases, the three-species M-concordance probability increases because monophyly is not prevented by interspecific coalescences that involve species C. For large T_2, the three-species M-concordance probability increases towards the probability that all three species are monophyletic, which in turn increases towards the two-species M-concordance probability. For small T_2, because the probability of events during the two-species phase is small, the collapsed gene tree is usually determined in the one-species phase, so that the probability of M-concordance is about $1/3$ of the probability of monophyly. The assertion that if all species are monophyletic then gene trees and species trees are likely to be topologically concordant (e.g., Takahata and Slatkin, 1990) does not hold for small T_2.

For small T_3, both two-species and three-species M-concordance probabilities are small, especially with large sample size. As T_3 increases, the probability of monophyly of each species increases, so that M-concordance is determined by whether or not the single ancestral lineages for each of the species produce a topology congruent to the species tree topology. Thus, for large T_3, the M-concordance probability is approximated by (1).

The remainder of this article focuses on the topological concordance probability. Of the senses discussed, only

the topological definition of concordance allows gene trees to be partitioned into the three topological classes. As we will see in Section 5, this property is useful for phylogenetic applications.

4. SIMULATIONS

4.1. Procedure

To explore the topological concordance probability in large samples, I used a standard coalescent simulation (e.g., Hudson, 1990). Although for small samples, the exact formula (5) is easy to compute, for large samples, recursive computations of $F_k^{A,B}$ can be more time-consuming than simulations. The simulations were analogous to those of Takahata (1989).

In each species, an exponential random variable of mean $\frac{2N}{j(j-1)}$ was simulated (where j was the number of sampled lineages in the species and N was the total number of individuals in the species) for the time of the most recent coalescence. If this time was more recent than any species divergence, two random numbers were chosen to decide which two of the j lineages coalesced, and the extant number of lineages was updated by subtracting one. This process of coalescence was continued until the divergence time of species A and B. If a divergence occurred between species that had j_1 and j_2 lineages at the divergence time, the simulation proceeded in the ancestral species using $j_1 + j_2$ lineages. I continued the simulation as above, taking into account each species divergence, until the most recent interspecific coalescence. If the most recent interspecific coalescence joined lineages from the sister species, the simulation was counted as having produced topological concordance.

In cases tested, when 100,000 simulations were performed with each set of parameter values, the topological concordance probability determined by simulation was usually within 0.001 of the analytically computed value (not shown). At parameter values for which Takahata (1989) performed similar simulations, the simulations here gave nearly identical results.

4.2. Properties of the Topological Concordance Probability

The probability of topological concordance is a rather complicated function of r, s, q, T_3, and T_2 (Eq. (5)). The key determinants of the topological concordance probability are the numbers of ancestral lineages to the samples of species A and B at time T_3, and the amount of time that these ancestral lineages have to coalesce (that is, T_2). The behavior of the topological concordance probability can be determined by considering several cases.

Large Values of T_2

When T_2 is sufficiently large, lineages from species A and B almost always have time to coalesce during the two-species phase. This is true regardless of the values of r, s, q, and T_3 (Fig. 5).

To understand this behavior, suppose that only one lineage from each of the sister species A and B enters the two-species phase at time T_3. The probability that these two lineages coalesce in this phase is $g_{21}(T_2)$, or $1-e^{-T_2}$. For large values of T_2, this probability is close to 1, and topological concordance is nearly always obtained (Fig. 5). Increasing the sample sizes r and s can only increase the topological concordance probability: if more lineages are present during the two-species phase, the chances of an interspecific coalescence during that phase are greater. Increasing T_3 causes lineages to coalesce within species, so that few lineages are represented in the two-species phase. Thus, increasing this parameter counteracts increases in sample sizes. In any case, however, at least one lineage will be represented from each species in the two-species phase, so that the topological concordance probability is at least $1-e^{-T_2}$.

Lastly, the sample size of species C has little effect at large values of T_2, because with large T_2, the collapsed gene tree topology is usually determined in the two-species phase. Thus, if T_2 is believed to be large, a gene tree inferred for any values of r, s, q, and T_3 is likely to reflect the species tree. Increasing the sample size is not necessary in this case.

Small Values of T_2 and Large Values of T_3

If T_2 is small, other parameters can significantly affect the topological concordance probability (Fig. 5). At large values of T_3, regardless of sample size, only one lineage is likely represented from each of the sister species during the two-species phase. This is attributable to the fact that the waiting time until coalescence of a large number of lineages has a mean of 2 coalescent units (e.g., Nordborg, 2001). Thus, if $T_3 \gg 2$, a large sample from a species in the present usually reflects a sample of only one ancestral lineage T_3 time units in the past. In this situation, increasing the sample size cannot increase the topological concordance probability (Fig. 6).

Because this part of the parameter space behaves as if samples of size 1 have been taken from all species, the topological concordance probability can be computed

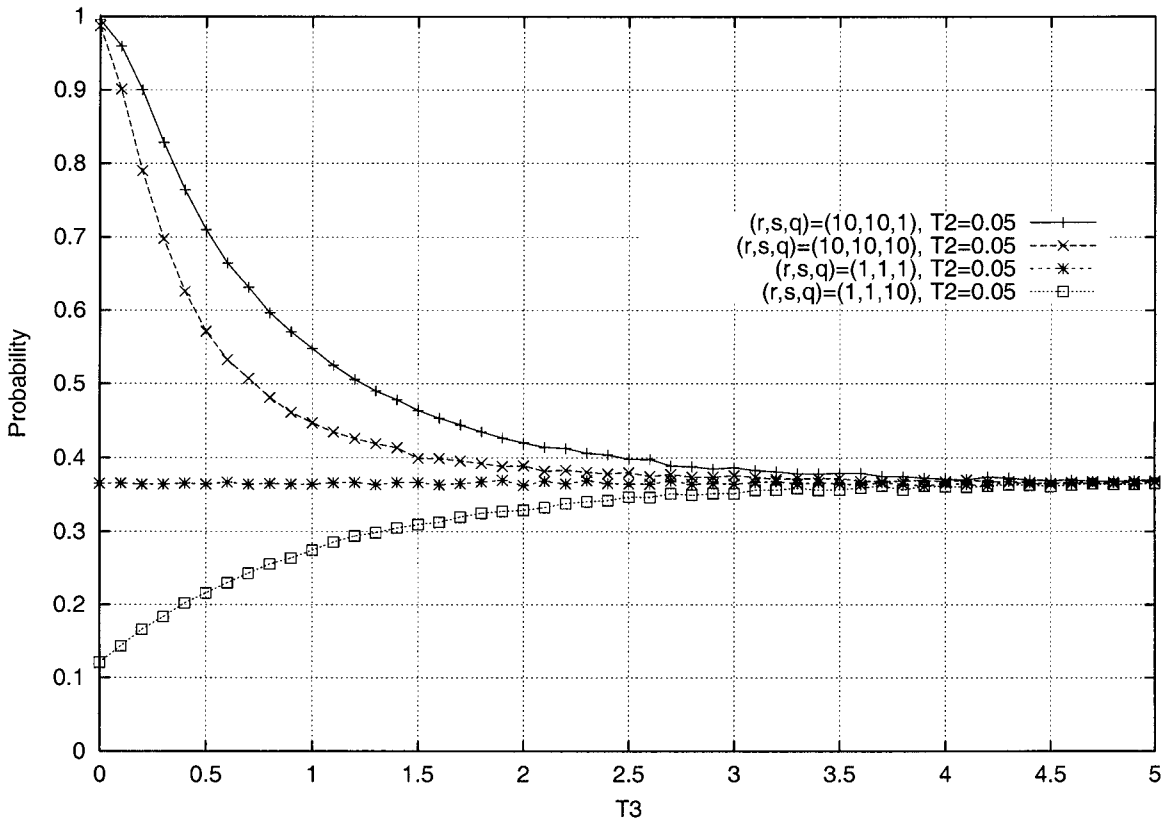

FIG. 6. Probability of topological concordance as a function of T_3, the time since the most recent divergence in the three-species model. Each point is based on 100,000 simulated instances of the model. Sample sizes in species A, B, and C were r, s, and q respectively, and the time between the sister-species divergence and the ancestral divergence was T_2.

from (1). In this circumstance, it is very likely that a large fraction of genes will produce topologies incompatible with the species tree. Here, considering topologies from many genes is more useful than increasing sample sizes within species.

Small Values of T_2 and Small Values of T_3

Limiting Behavior. If both T_2 and T_3 are small, then the sample sizes r, s, and q can significantly affect the topological concordance probability (Fig. 7). It is clear that increasing r and s while holding q constant can increase the topological concordance probability, and that increasing q while holding r and s constant can decrease it. In contrast to what might be expected, however, increasing only r or only r and s without bound does not lead to a topological concordance probability of 1; similarly, increasing q without bound does not lead to an eventual topological concordance probability of 0. At fixed nonzero T_3 and T_2, the large-sample limit of the topological concordance probability is a value that lies strictly between 0 and 1.

These observations are a consequence of two facts. First, as mentioned earlier, the topological concordance probability depends largely on the numbers of ancestral lineages of species A and B that are present in the two-species phase. As these numbers increase and the length of this phase increases, it becomes more likely that an interspecific coalescence will happen during the phase.

Second, regardless of the sample sizes used in the present, the numbers of ancestral lineages (m and n in the notation of Fig. 4) cannot be increased beyond a certain limit. For a very large sample size, most coalescences take place quickly, and few ancestral lineages are represented. Increasing the sample size in the present increases the sample size in the past by a comparatively small amount.

Formally, Tavaré (1984, Eq. (6.5)) showed that if r lineages are sampled from species A, and if m is the random number of ancestral lineages present at time T_3 in

the past, then the probability that m is at least c (where $1 \leqslant c \leqslant r$) satisfies

$$e^{-c(c-1)T_3/2} \leqslant Pr[m \geqslant c] \leqslant \min\left\{1, \frac{(2c-1)!}{(c!)^2} e^{-c(c-1)T_3/2}\right\}. \tag{14}$$

Note that the bounds on $Pr[m \geqslant c]$ are uniform; that is, they do not depend on the number of sampled lineages, r. In other words, no matter how many lineages r are sampled, the probability that at least c ancestral lineages are represented at time T_3 cannot be increased above $\frac{(2c-1)!}{(c!)^2} e^{-c(c-1)T_3/2}$, or if the Stirling approximation for large c is used, $(4^c/[2c\sqrt{\pi c}])\, e^{-c(c-1)T_3/2}$.

Maximal Useful Sample Sizes. Because the distribution of the number of ancestral lineages cannot be increased above the upper bounds in (14) through use of a large sample size, it is useful to determine sample sizes larger than which topological concordance probability is only trivially affected.

To compute these sample sizes, I assume that the number of ancestral lineages at time T_3 directly impacts the topological concordance probability. Thus, I assume that if the number of ancestral lineages at time T_3 cannot be increased by an increase in sample size, then the topological concordance probability cannot be substantially increased either. The idea of the computation is to choose r large enough so that by an appropriate criterion of deviation, the distribution of A_{T_3} given $A_0 = r$ deviates from the large sample limiting distribution of A_{T_3} given $A_0 = \infty$ by less than a prespecified tolerance (where A_T is the random number of ancestral lineages to the sample at time T coalescent units in the past).

To measure the deviation of the cumulative distribution of the number of lineages at time T_3 given a sample size of r, that is, $Pr(A_{T_3} \leqslant a \mid A_0 = r)$ or $G_r(a)$, from the large-sample limiting distribution $G_\infty(a)$, it is ideal to use the total variation norm. That is, for a given ε, a maximal sample size might be the minimal r such that the following criterion holds for all values of a:

$$|G_r(a) - G_\infty(a)| < \varepsilon. \tag{15}$$

However, for ease of computation, it is convenient to measure deviations using a less cumbersome criterion. In general, T_3 is not known precisely, and the sample size chosen in a study will need to reflect this uncertainty. Thus, the reason for identifying bounds is more to develop practical rules than for rigorous mathematical precision.

For these distributions, the absolute difference between the mean of G_r and the mean of G_∞ is appropriate. Although it is in general unwise to use this norm to measure convergence (for example, consider a family of normal distributions with mean zero in which the jth distribution has variance $1 + 1/j$), the properties of the distributions G_r have been well studied (e.g., Griffiths, 1984; Tavaré, 1984), and the distributions do not exhibit behavior that would make the mean difference a problematic criterion. Conveniently, the means of the distributions are fairly easy to calculate (Griffiths, 1981; Tavaré, 1984).

Let $\varepsilon > 0$. Define the recommended sample size by R, the minimal r that satisfies

$$|E[A_{T_3} \mid A_0 = r] - E[A_{T_3} \mid A_0 = \infty]| < \varepsilon, \tag{16}$$

where the expected number of ancestral lineages at time T_3 is given by

$$E[A_{T_3} \mid A_0 = r] = \sum_{k=1}^{r} e^{-k(k-1)T_3/2} \frac{(2k-1)\, r_{[k]}}{r_{(k)}} \tag{17}$$

for finite r (Griffiths, 1981; Tavaré, 1984, Eq. (6.7)) and by

$$E[A_{T_3} \mid A_0 = \infty] = \sum_{k=1}^{\infty} e^{-k(k-1)T_3/2}(2k-1) \tag{18}$$

in the limiting case (Griffiths, 1981). Recall that $r_{(k)}$ and $r_{[k]}$ are defined in Section 3.1.

Values of R computed from (16)–(18) are shown in Table III, along with deviations of large-sample limiting topological concordance probabilities from those computed at the recommended sample sizes. As is clear from Table III and from Fig. 7, unless the sister species diverged recently (small T_3), only a small sample size is needed in order to obtain a topological concordance probability close to the large-sample limit.

4.3. Properties of the Speciodendricity Probability

The large-sample limiting topological concordance probability, or speciodendricity probability (Eq. (6)), is shown in Fig. 8. Along $T_2 = 0$, the speciodendricity probability is $1/3$. For large values of T_3 the speciodendricity probability approaches $1 - (2/3)\, e^{-T_2}$, and for large values of T_2 this probability is nearly 1.

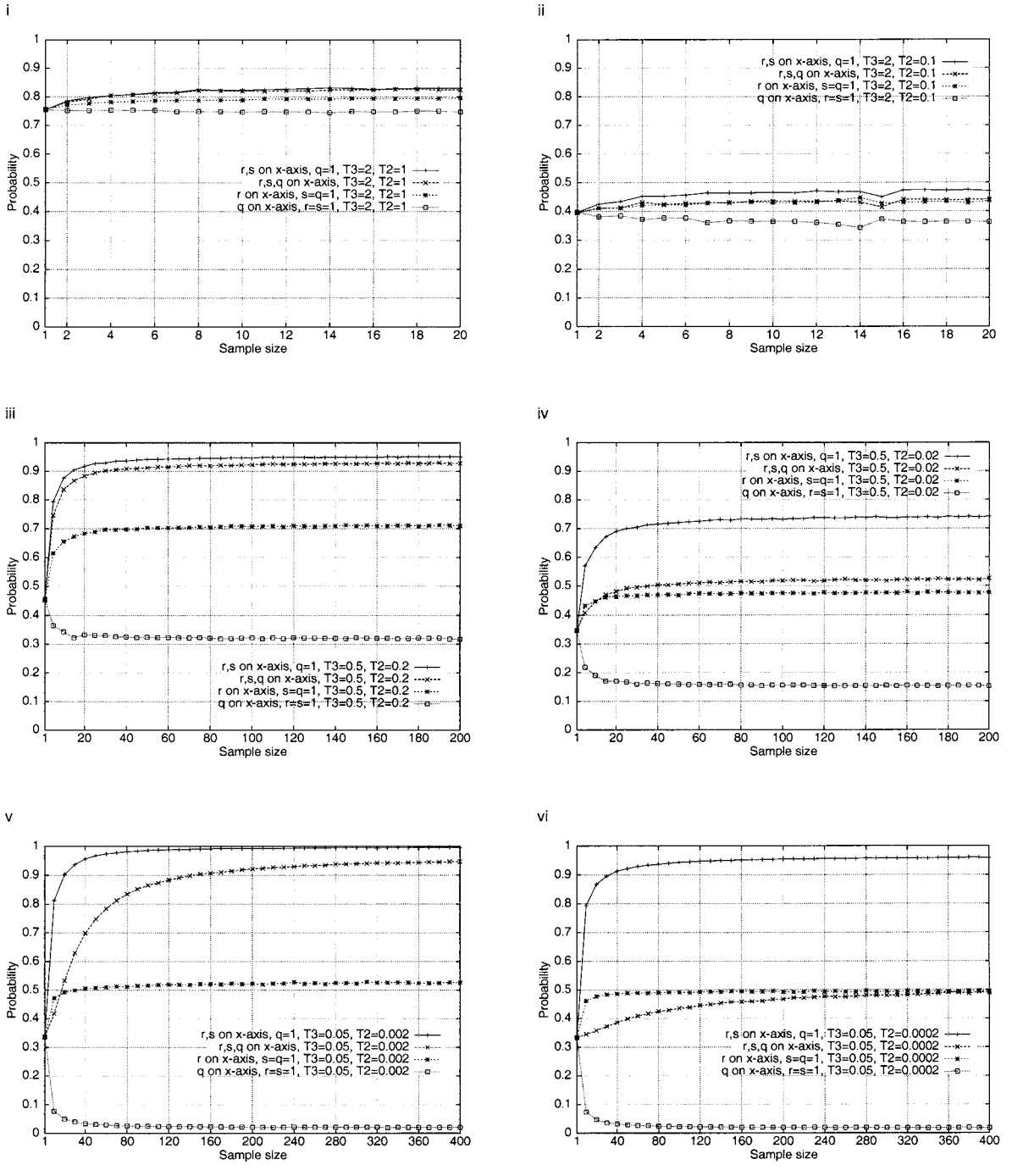

FIG. 7. Probability of topological concordance as a function of sample sizes. The independent variable differs across the four curves in each graph. For example, in the top curve, both r and s vary according to the values on the x-axis, while q is constant at 1. Each point is based on 100,000 simulated instances of the model.

TABLE III

Minimal Sample Size for Which the Mean Number of Ancestral Lineages Is Close to the Large-Sample Limit

T_3	Large-sample limiting mean number of ancestral lineages at time T_3	$\log_2(\varepsilon)$	Lower bound R	Mean number of ancestral lineages at time T_3 with a sample size of R
5	1.020	−3	1	1
2	1.418	−1	1	1
		−2	3	1.204
		−3	6	1.294
1	2.370	1	1	1
		0	3	1.577
		−1	6	1.879
		−2	14	2.126
		−3	30	2.248
0.5	4.351	2	1	1
		1	5	2.557
		0	12	3.352
		−1	28	3.853
		−2	60	4.102
		−3	124	4.226
0.2	10.340	4	1	1
		3	3	2.503
		2	16	6.479
		1	41	8.379
		0	90	9.341
		−1	190	9.840
		−2	390	10.090
		−3	790	10.215
0.1	20.337	4	6	4.803
		3	31	12.476
		2	81	16.375
		1	180	18.337
		0	380	19.337
		−1	780	19.837
		−2	1580	20.087
		−3	3180	20.212
0.05	40.335	4	61	24.474
		3	161	32.373
		2	361	36.345
		1	760	38.335
		0	1560	39.335
		−1	3160	39.835
		−2	6360	40.085
		−3	12759	40.210

Note. See Section 4.2 for descriptions of ε and R.

As discussed above in the context of sample sizes, the most complex behavior is in the region where both T_3 and T_2 are small. For small T_2, the speciodendricity probability decays quickly as T_3 increases. For small T_3, slight

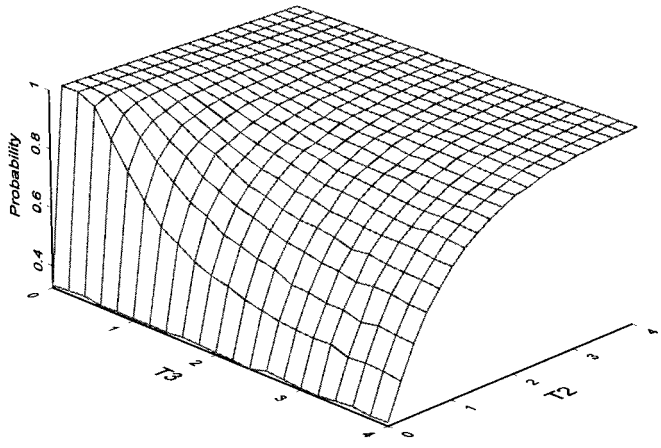

FIG. 8. Probability of speciodendricity as a function of T_3 and T_2. Each point is based on 100,000 simulated coalescent trees. The speciodendricity probability was approximated by using a large sample size for the simulations, namely 400 lineages in each species. Because the smallest nonzero value of T_3 used was 0.2, and because sample sizes of 400 give topological concordance probabilities very close to the large-sample limit if $T_3 \geqslant 0.2$ (see Table III), the graph shown is a very good approximation of the speciodendricity probability (except for $0 < T_3 < 0.2$ and T_2 small, where no points are plotted).

increases in T_2 are sufficient to increase the probability of speciodendricity near 1. Along $T_3 = 0$, P_S equals 1, with the exception that at $T_3 = T_2 = 0$, $P_S = 1/3$: $(0, 0)$ is a point of discontinuity of P_S. This result is easily explained: it is intuitive that $P_S(0, 0) = 1/3$. With $T_3 = 0$ and $T_2 > 0$, however, if the sample sizes are *infinite*, then interspecific coalescences in the two-species phase are guaranteed.

5. DISCUSSION

In this article, I have computed the likelihood function of an observed gene tree conditioned on a proposed species tree topology (together with its branch lengths). This calculation enables the use of likelihood-based inference of three-species phylogenies based on sample sizes larger than 1.

In agreement with Takahata (1989), I have found that conditions on T_3 and T_2 exist under which sample sizes can increase the topological concordance probability. If T_3 and T_2 are both small, topological concordance probability can be increased (to a point) by enlarging samples. The increase in sample size needed for achieving a desired topological concordance probability depends on T_3 and T_2, and maximal useful sample sizes can be obtained in Table III. If T_2 is large, topological concordance is nearly

guaranteed. If T_2 is small and T_3 is large, topological concordance is not likely; this result is little affected by sample size.

These results are relevant to a variety of problems.

5.1. Inference of Species Trees

Maddison (1997) suggested that species trees could be inferred from gene trees by searching for species trees that maximize the likelihood function

$$\prod_{l=1}^{L} \sum_{G \in \mathcal{G}} Pr(D \mid G)\, Pr(G \mid S), \tag{19}$$

where L is the number of unlinked loci genotyped, D is a collection of individual multilocus genotypes, $\mathcal{G}$ is the set of possible gene trees, and S is a proposed species tree (including branch lengths). A model of sequence evolution gives the function $Pr(D \mid G)$ and a model of the shape of gene genealogies gives $Pr(G \mid S)$. By allowing for large sample sizes, the present work has expanded the set of situations for which $Pr(G \mid S)$ can be calculated (at least, if G includes only the genealogical shape). Additional advances in the likelihood computation—such as for more species and for different demographic models—together with sequence evolution models and Markov chain Monte Carlo techniques for maximizing (19) will expand the class of situations for which tests of phylogenetic hypotheses can be performed using gene trees.

From a more philosophical perspective, the incorporation of samples larger than one into likelihood calculations enables within-species variation to be accommodated in molecular phylogenetic inference. Most phylogenetic algorithms are typological in nature rather than populational (see Dobzhansky, 1967) and they do not easily accommodate within-species polymorphisms or polymorphisms shared across species (Wiens, 1999). To infer relationships of closely related and recently diverged groups, only an approach that takes into account the range of variation within each group should be fully satisfying. Incorporating larger samples, that is, within-species variation, into phylogenetic likelihood computations is both more realistic given the extent of within-species variability, and, as shown by Takahata (1989) and here, it is often more likely to produce topologies concordant with the species tree.

5.2. Sample Sizes for Human Evolution Problems

The effect of sample size is that increasing the sample size increases the topological concordance probability, but that increasing the size of large samples is only minimally helpful. This result also holds if many lineages are taken from one or two of three species, but only one lineage is taken from the remaining species. The results given in Section 4.2 enable speculation on maximal useful sample sizes for various problems of interest.

A loose estimate places T_3 between 1.6 and 93.3 for humans and chimpanzees (Rosenberg and Feldman, 2002). At the low end of this range, several lineages are sufficient to reach the limiting number of ancestral lineages, and at the high end of the range, it is unnecessary to use more than one lineage. Thus, depending on T_3, it may be possible to improve upon previous attempts to resolve relationships of humans, chimpanzees, and gorillas, by adding as many as 5–10 lineages from each species. If geographic structure is taken into account, this suggested sample size will increase. It seems, however, that the divergence is sufficiently ancient that increasing the number of genes is more useful than increasing the sample sizes.

For the divergence of humans and Neanderthals, T_3 is likely between 0.5 and 10 (Rosenberg and Feldman, 2002). Again, at the high end of the range, samples of size one will reach the limiting concordance probability. At the low end, it is valuable to examine as many as, say, 20–40 Neanderthal sequences in order to study their relationship to modern humans (see Table III). As in other cases, it is useful to look at gene trees taken from many genes.

For pairs of modern human groups, values of T_3 may be as small as 0.05 (Rosenberg and Feldman, 2002). To achieve maximal accuracy, sample sizes as large as 80–200 from each group are needed (see Table III and Fig. 7). Of course, geographic structure will increase the required sample sizes further, and as above, many genes will need to be studied.

5.3. Estimation of T_3, T_2, or N

If the species phylogeny is "known," equations given here allow estimation of T_3 and T_2 from a set of individual genotypes at multiple independent loci. Suppose that the species phylogeny is $((AB)\,C)$ and that r, s, and q lineages are sampled from species A, B, and C, respectively. If L genes are typed and if trees for x, y, and z genes support topologies $((AB)\,C)$, $((AC)\,B)$, and $((BC)\,A)$, respectively, then the likelihood of the data is

$$Lik(T_3, T_2) \propto P_C(r, s, q, T_3, T_2)^x\, Q_{((AC)\,B)}(r, s, q, T_3, T_2)^y \times Q_{((BC)\,A)}(r, s, q, T_3, T_2)^z. \tag{20}$$

Maximizing (20) should provide estimates of both T_3 and T_2. In practice, it is hoped that it will be possible to separate the impacts of T_3 and T_2. However, in some situations, independent knowledge may exist about one of these two variables, in which case (20) can enable estimation of the other. Another application of (20) is the extension of multiple-locus likelihood-ratio tests of species phylogenies (Wu, 1991; Hudson, 1992; Ruvolo, 1997) to include sample sizes larger than one.

It is noted that by employing the probabilities of the alternate topologies, the precision of an estimate should be improved beyond that obtained by dividing gene trees into those that are concordant and those that are discordant with the species tree. Because the derivatives of P_C, $Q_{((AC)B)}$ and $Q_{((BC)A)}$ are unwieldy, the likelihood is best maximized numerically.

Equation (20) also provides the potential to estimate ancestral population sizes, as has been done with samples of size 1 (e.g., Chen and Li, 2001). If T_2 and T_3 can be estimated, and if independent information is available about t_2 and t_3 (measured in generations), then $t_2/\hat{T}_2$ estimates the population size for the most recent common ancestral species for A and B.

5.4. Speciation Genes

All calculations to date on concordance of gene trees and species trees, including those presented here, have treated random genes, that is, genes whose functions are assumed to have played no role in causing ancestral species to diverge. It stands to reason, however, that genes in which mutations or changes in expression contributed to species divergence will produce concordant trees much more often than will random genes: speciation genes are more likely to be speciodendric. This prediction has been confirmed for *Odysseus*, a gene thought to have been involved in the divergence of *Drosophila* species (Ting *et al.*, 2000).

To study this phenomenon, the reasoning employed in Section 5.1 may be inverted: instead of using gene trees to infer species phylogenies, species phylogenies can be used to make inferences about specific genes. For example, a species phylogeny can be assumed to be known. If a gene is found to be speciodendric with respect to that set of species, and if the probability of speciodendricity is low (as computed by Eq. (6) for a set of three species), then it might be inferred that the gene was causally linked to the divergences of the species under consideration (or genetically linked to such a "speciation gene"). This observation suggests a genomic approach: with complete genome data for a set of closely related species, genes can be tested for speciodendricity. Genes found to be speciodendric can be targeted for further study of their potential roles in speciation.

5.5. Extensions: Four or More Species

If four or more species are considered (Fig. 9), even with samples of size 1, alternate gene tree topologies do not have the same probabilities (Tables IV and V). For samples of size 1 from each species, the four-species probabilities of topological concordance agree with those found by Pamilo and Nei (1988); for larger sample sizes the probabilities can be obtained using calculations similar to those in Section 3.

In general, to deduce the probability that a gene tree is topologically concordant with a given species tree of any size, it is best to proceed backwards in time, conditioning on all possible lineage configurations and for each lineage configuration, computing the probability that if

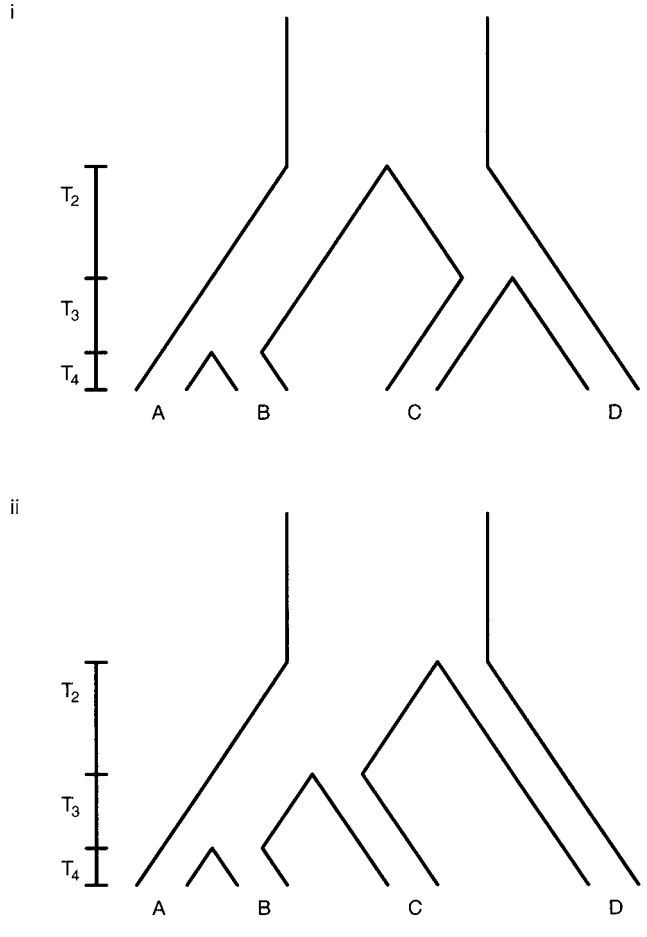

FIG. 9. The two bifurcating tree topologies that can be generated by four species. (i) Balanced tree topology. (ii) Unbalanced tree topology.

TABLE IV

Probabilities of the 15 Gene Tree Topologies When the Species Tree Topology Is ((*AB*)(*CD*))

Gene tree topology	Probability	Probability at $T_3 = T_2 = 1$
$((AB)(CD))$	$g_{21}(T_3+T_2)\, g_{21}(T_2)$ $+ g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.6867
$((AC)(BD))$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.0055
$((AD)(BC))$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{2}{6}\, \frac{1}{3}$	0.0055
$(((AB)\,C)\,D)$	$g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.1088
$(((AB)\,D)\,C)$	$g_{21}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.1088
$(((AC)\,B)\,D)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AC)\,D)\,B)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AD)\,B)\,C)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((AD)\,C)\,B)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BC)\,A)\,D)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BC)\,D)\,A)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BD)\,A)\,C)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((BD)\,C)\,A)$	$g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0028
$(((CD)\,A)\,B)$	$g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0313
$(((CD)\,B)\,A)$	$g_{22}(T_3+T_2)\, g_{21}(T_2)\, \frac{1}{3}$ $+ g_{22}(T_3+T_2)\, g_{22}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0313

Note. Notation is as in Fig. 9i. Values of $g_{ij}(T)$ are given by (2).

TABLE V

Probabilities of the 15 Gene Tree Topologies When the Species Tree Topology Is (((*AB*) *C*) *D*)

Gene tree topology	Probability	Probability at $T_3 = T_2 = 1$
$((AB)(CD))$	$g_{21}(T_3)\, g_{22}(T_2)\, \frac{1}{3} + g_{22}(T_3)$ $\times [g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{2}{6}\, \frac{1}{3}]$	0.0991
$((AC)(BD))$	$g_{22}(T_3)[g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{2}{6}\, \frac{1}{3}]$	0.0215
$((AD)(BC))$	$g_{22}(T_3)[g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{2}{6}\, \frac{1}{3}]$	0.0215
$(((AB)\,C)\,D)$	$g_{21}(T_3)[g_{21}(T_2) + g_{22}(T_2)\, \frac{1}{3}] + g_{22}(T_3)$ $\times [g_{31}(T_2)\, \frac{1}{3} + g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.5556
$(((AB)\,D)\,C)$	$g_{21}(T_3)\, g_{22}(T_2)\, \frac{1}{3} + g_{22}(T_3)$ $\times [g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.0980
$(((AC)\,B)\,D)$	$g_{22}(T_3)[g_{31}(T_2)\, \frac{1}{3} + g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.0785
$(((AC)\,D)\,B)$	$g_{22}(T_3)[g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.0205
$(((AD)\,B)\,C)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010
$(((AD)\,C)\,B)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010
$(((BC)\,A)\,D)$	$g_{22}(T_3)[g_{31}(T_2)\, \frac{1}{3} + g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.0785
$(((BC)\,D)\,A)$	$g_{22}(T_3)[g_{32}(T_2)\, \frac{1}{3}\, \frac{1}{3} + g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}]$	0.0205
$(((BD)\,A)\,C)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010
$(((BD)\,C)\,A)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010
$(((CD)\,A)\,B)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010
$(((CD)\,B)\,A)$	$g_{22}(T_3)\, g_{33}(T_2)\, \frac{1}{6}\, \frac{1}{3}$	0.0010

Note. Notation is as in Fig. 9ii. Values of $g_{ij}(T)$ are given by (2).

an interspecific coalescence occurs during a given interval, then the coalescence *violates* the proposed topology. The probability of topological concordance is then equal to one minus the probability that the proposed topology is not obtained.

If the true species tree is balanced, the probability that a gene tree is topologically concordant with the species tree is larger than if it is unbalanced. Informally, the more symmetry the tree topology has, the greater the number of sequences of coalescences that produce gene trees topologically concordant with the species tree. For example, with samples of size 1 and eight species, only 1 out of 1,587,600 random sequences of coalescences can produce a topology concordant with the species tree $(((((((AB)\,C)\,D)\,E)\,F)\,G)\,H)$. In contrast, the species tree $(((AB)(CD))((EF)(GH)))$ can be achieved in any of 80 different sequences.

This effect, that balanced species tree topologies are more likely to have concordant gene trees, increases in magnitude with the number of species. Contrast the case of eight species with the fact that for four species $(((AB)\,C)\,D)$ is achieved in 1 of 18 sequences, and $((AB)(CD))$ is achieved in only 2 of 18. The claim that balanced trees more often have speciodendric genes, however, is true only if branch lengths are very short. With short branch lengths, interspecific coalescences only take place in the ancestral phase, during which the order of coalescence is random and the above reasoning holds.

5.6. Extensions: Expanding the Model

I have made simplifying assumptions about equality and stability of population sizes and absence of population structure. A concordance probability calculation in a model with many demes per species is given by Wakeley (2000), and Takahata and Slatkin (1990) numerically computed M-concordance probability in a two-species migration model. Future work might expand these calculations to include larger sample sizes or other forms of population structure. In general, the effect of geographic structure within species is to decrease the topological concordance probability (e.g., Fig. 6 in Wakeley, 2000).

Another limitation of this work is that I ignore gene exchange between species, mistaken orthology, and mutational stochasticity. Including these factors will allow for the main determinants of discordance to be considered simultaneously. By studying a general model that includes all factors, it may be possible to determine relative probabilities for each of the different causes of an observed discordance. It may also be possible to use the

observed amount of concordance to estimate gene exchange rates and other parameters.

APPENDIX

$F_k^{A,B}(a, b, c)$ is the probability that in coalescing from a, b, and c lineages from species A, B, and C, respectively, to k total lineages, (i) an interspecific coalescence occurs and (ii) the most recent interspecific coalescence links lineages of species A and species B. Similarly, we have $F_k^{A,C}$ or $F_k^{B,C}$, if the most recent interspecific coalescence joins species A and species C, or species B and species C, respectively.

Define $F_k^{A,B}(a, b, c) = 0$ if any of $\{a, b, c\}$ is negative, or if $a+b+c \leqslant k$. We restrict attention to $a, b, c \geqslant 0$ and $k \geqslant 1$. The expression $F_k^{A,B}(a, b, c)$ only describes a sensible quantity if $a+b+c \geqslant k$. The following recurrence relation holds:

$$F_k^{A,B}(a,b,c) = \frac{ab}{\binom{a+b+c}{2}} + F_k^{A,B}(a-1,b,c)\frac{\binom{a}{2}}{\binom{a+b+c}{2}} + F_k^{A,B}(a,b-1,c)\frac{\binom{b}{2}}{\binom{a+b+c}{2}} + F_k^{A,B}(a,b,c-1)\frac{\binom{c}{2}}{\binom{a+b+c}{2}}. \quad (21)$$

The first term arises from the fact that with probability $ab/\binom{a+b+c}{2}$, the most recent coalescence occurs interspecifically and joins species A and B. Each species contributes a term in case the most recent coalescence is *intra*specific in that species. The recursion terminates with base cases that have $a+b+c=k$ and $F_k^{A,B}(a, b, c) = 0$.

Values of $F_k^{A,B}(a, b, c)$ are shown in Table I. Results for $F_k^{A,C}$ and $F_k^{B,C}$ can be obtained using Property 2 below. The probability of no interspecific coalescences during the collapsing of a, b, and c lineages to k lineages equals $1-[F_k^{A,B}(a, b, c)+F_k^{A,C}(a, b, c)+F_k^{B,C}(a, b, c)]$.

Many properties of $F_k^{A,B}$ are easily verified and justifications of some properties are given below.

Property 1. $F_k^{A,B}(a, b, c) = F_k^{A,B}(b, a, c)$.

This property allows the assumption of $a \geqslant b$, without loss of generality, as in Table I.

Property 2. $F_k^{A,B}(a,b,c) = F_k^{A,C}(a,c,b) = F_k^{B,C}(b,c,a)$.

Using Property 1, these three quantities also equal $F_k^{A,B}(b, a, c)$, $F_k^{A,C}(c, a, b)$, and $F_k^{B,C}(c, b, a)$.

Property 3. $F_k^{A,B}(0, b, c) = 0$.

This follows from repeated application of (21) to $F_k^{A,B}(0, b, c)$ until base cases are reached.

Property 4. $F_k^{A,B}(a, b, 0) = H_{jk}$, with H_{jk} as in Takahata (1989, Eq. (11)), and $j = a+b$.

Coalescences involving species C were not considered by Takahata (1989), so $F_k^{A,B}(a, b, c)$ generalizes Takahata's H_{jk}. See Takahata (1989, Table 1) for values of $F_k^{A,B}(a, b, 0)$.

Property 5. If $b \geqslant 2$, then $F_2^{A,B}(1, b, 0) = 1 - \frac{2}{b(b+1)}$.

For no interspecific coalescences to occur in the collapsing of the configuration $(1, b, 0)$ to 2 lineages, all coalescences must be intraspecific within species B. For $b \geqslant 2$, the probability of this occurrence is

$$1 - F_2^{A,B}(1,b,0) = \frac{\binom{b}{2}}{\binom{b+1}{2}} \times \frac{\binom{b-1}{2}}{\binom{b}{2}} \times \frac{\binom{b-2}{2}}{\binom{b-1}{2}} \times \cdots \times \frac{\binom{2}{2}}{\binom{3}{2}} = \frac{2}{b(b+1)}. \quad (22)$$

Property 6. $F_k^{A,B}(a, b, c)+F_k^{A,C}(a, b, c)+F_k^{B,C}(a, b, c) \leqslant 1$, with equality if $a+b+c-k > (a-1)\,\chi(a)+(b-1)\,\chi(b)+(c-1)(c)$, where $\chi(x) = 1$ if $x \geqslant 2$ and $\chi(x) = 0$ otherwise.

$F_k^{A,B}(a, b, c)+F_k^{A,C}(a, b, c)+F_k^{B,C}(a, b, c)$ is the probability that *some* type of interspecific coalescence occurs as the configuration (a, b, c) collapses to k lineages. Equality occurs if an interspecific coalescence *must* take place. Such a coalescence is guaranteed if the total number of coalescences, or $a+b+c-k$, exceeds the maximal possible number of *intra*specific coalescences, or $(a-1)\,\chi(a)+(b-1)\,\chi(b)+(c-1)\,\chi(c)$.

As a direct consequence of Property 6, if at least two of $\{a, b, c\}$ are positive, then $F_1^{A,B}(a, b, c)+F_1^{A,C}(a, b, c)+F_1^{B,C}(a, b, c) = 1$. This corollary together with Property 2 yields $F_1^{A,B}(a, a, a) = 1/3$ for $a \geqslant 1$.

Property 7. $F_k^{A,B}(a, b, c)$ is a nonincreasing function of k.

From the state (a, b, c) with $a+b+c \geqslant k$, suppose a sequence of coalescences occurs so that the most recent interspecific coalescence in the sequence occurs between species A and B and so that the total number of lineages is left at $k+1$ or greater. This sequence contributes to both $F_k^{A,B}$ and $F_{k+1}^{A,B}$. Any sequence for which the most recent interspecific coalescence leaves the number of lineages at $k+1$ or greater, and that occurs between species A and C or between B and C, will not contribute to either $F_k^{A,B}$ or to $F_{k+1}^{A,B}$. However, sequences for which no interspecific coalescence occurs in reaching $k+1$ lineages have *some chance* of having a coalescence that joins A and B as the number of lineages hits k. Thus, the chance of the most recent interspecific coalescence occurring between A and B in declining to k lineages is *at least* the chance of the same event occurring in declining to $k+1$ lineages.

Property 8. If $c > 0$, then $F_1^{A,B}(a, b, c) = F_2^{A,B}(a, b, c)$.

By Property 7, $F_1^{A,B}(a, b, c) \geqslant F_2^{A,B}(a, b, c)$. If $a = 0$ or $b = 0$, the result follows from Property 3. Otherwise, the only way for $F_1^{A,B}(a, b, c)$ to be strictly larger than $F_2^{A,B}(a, b, c)$ is if the final coalescence in a sequence links species A and B and if it is the only interspecific coalescence in the sequence. However, this is not possible: to reach a single lineage with $\{a, b, c\}$ all positive, at least two interspecific coalescences are needed.

ACKNOWLEDGMENTS

I thank Marc Feldman, Aaron Hirsh, Joanna Mountain, Magnus Nordborg, Dmitri Petrov, Dylan Schwilk, Simon Tavaré, John Wakeley and an anonymous reviewer for comments that have improved the manuscript. This work was supported by a Program in Mathematics and Molecular Biology Fellowship through the Burroughs–Wellcome Fund and by NIH Grant GM28428 to Marc Feldman.

REFERENCES

Chen, F.-C., and 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.

Dobzhansky, T. 1967. On types, genotypes, and the genetic diversity in populations, *in* "Genetic Diversity and Human Behavior" (J. N. Spuhler, Ed.), pp. 1–18, Aldine, Chicago.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy, *Syst. Bot.* **17**, 144–163.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins, *Syst. Zool.* **19**, 99–113.

Fitch, W. M. 2000. Homology: A personal view on some of the problems, *Trends Genet.* **16**, 227–231.

Griffiths, R. C. 1981. Transient distribution of the number of segregating sites in a neutral infinite-sites model with no recombination, *J. Appl. Probab.* **18**, 42–51.

Griffiths, R. C. 1984. Asymptotic line-of-descent distributions, *J. Math. Biol.* **21**, 67–75.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data, *Evolution* **37**, 203–217.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *Oxford Surv. Evol. Biol.* **7**, 1–44.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles, *Genetics* **131**, 509–512.

Maddison, W. P. 1997. Gene trees in species trees, *Syst. Biol.* **46**, 523–536.

Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees, *Evolution* **49**, 718–726.

Moritz, C. 1994. Defining "evolutionary significant units" for conservation, *Trends Ecol. Evol.* **9**, 373–375.

Mountain, J. L., and Cavalli-Sforza, L. L. 1997. Multilocus genotypes, a tree of individuals, and human evolutionary history, *Am. J. Hum. Genet.* **61**, 705–718.

Nei, M. 1986. Stochastic errors in DNA evolution and molecular phylogeny, *in* "Evolutionary Perspectives and the New Genetics" (H. Gershowitz, D. L. Rucknagel, and, R. E. Tashian, Eds.), pp. 133–147, A. R. Liss, New York.

Neigel, J. E., and Avise, J. C. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation, *in* "Evolutionary Processes and Theory" (S. Karlin and E. Nevo, Eds.), pp. 515–534, Academic Press, New York.

Nichols, R. 2001. Gene trees and species trees are not the same, *Trends Ecol. Evol.* **16**, 358–364.

Nordborg, M. 2001. Coalescent theory, *in* "Handbook of Statistical Genetics" (D. J. Balding, C. Cannings, and, M. Bishop, Eds.), Chap. 7, pp. 179–212, Wiley, Chichester, UK .

Ouzounis, C. 1999. Orthology: Another terminology muddle, *Trends Genet.* **15**, 445.

Pamilo, P., and Nei, M. 1988. Relationships between gene trees and species trees, *Mol. Biol. Evol.* **5**, 568–583.

Rosenberg, N. A., and Feldman, M. W. 2002. The relationship between coalescence times and population divergence times, *in* "Modern Developments in Theoretical Population Genetics" (M. Slatkin and M. Veuille, Eds.), Chap. 9, pp. 130–164, Oxford Univ. Press, Oxford.

Ruvolo, M. 1994. Molecular evolutionary processes and conflicting gene trees: The Hominoid case, *Am. J. Phys. Anthropol.* **94**, 89–113.

Ruvolo, M. 1997. Molecular phylogeny of the Hominoids: Inferences from multiple independent DNA sequence data sets, *Mol. Biol. Evol.* **14**, 248–265.

Saitou, N., and Nei, M. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human–chimpanzee–gorilla divergence, *J. Mol. Evol.* **24**, 189–204.

Satta, Y., Klein, J., and Takahata, N. 2000. DNA archives and our nearest relative: The trichotomy problem revisited, *Mol. Phylogenet. Evol.* **14**, 259–275.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**, 437–460.

Takahata, N. 1989. Gene genealogy in three related populations:

Consistency probability between gene and population trees, *Genetics* **122**, 957–966.

Takahata, N., and Nei, M. 1985. Gene genealogy and variance of interpopulational nucleotide differences, *Genetics* **110**, 325–344.

Takahata, N., and Satta, Y. 2002. Pre-speciation coalescence and the effective size of ancestral populations, *in* "Modern Developments in Theoretical Population Genetics" (M. Slatkin and M. Veuille, Eds.), Chap. 5, pp. 52–71, Oxford Univ. Press, Oxford.

Takahata, N., and Slatkin, M. 1990. Genealogy of neutral genes in two partially isolated populations, *Theor. Popul. Biol.* **38**, 331–350.

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

Ting, C.-T., Tsaur, S.-C., and Wu, C.-I. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus, Proc. Natl. Acad. Sci. USA* **97**, 5316–5316.

Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species, *Evolution* **54**, 1092–1101.

Wiens, J. J. 1999. Polymorphism in systematics and comparative biology, *Annu. Rev. Ecol. Syst.* **30**, 329–362.

Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms, *Genetics* **127**, 429–435.