Syst. Biol. 49(3):422–434, 2000

Testing Hybridization Hypotheses Based on Incongruent Gene Trees

TAO SANG[1] AND YANG ZHONG[2]

[1]*Department of Botany and Plant Pathology, Michigan State University, East Lansing, Michigan 48824, USA; E-mail: sang@pilot.msu.edu*

[2]*Institute of Biodiversity Science, School of Life Sciences, Fudan University, Shanghai 200433, China; E-mail: yangzhong@fudan.edu*

Abstract.—Hybridization is an important evolutionary mechanism in plants and has been increasingly documented in animals. Difficulty in reconstruction of reticulate evolution, however, has been a long-standing problem in phylogenetics. Consequently, hybrid speciation may play a major role in causing topological incongruence between gene trees. The incongruence, in turn, offers an opportunity to detect hybrid speciation. Here we characterized certain distinctions between hybridization and other biological processes, including lineage sorting, paralogy, and lateral gene transfer, that are responsible for topological incongruence between gene trees. Consider two incongruent gene trees with three taxa, A, B, and C, where B is a sister group of A on gene tree 1 but a sister group of C on gene tree 2. With a theoretical model based on the molecular clock, we demonstrate that time of divergence of each gene between taxa A and C is nearly equal in the case of hybridization (B is a hybrid) or lateral gene transfer, but differs significantly in the case of lineage sorting or paralogy. After developing a bootstrap test to test these alternative hypotheses, we extended the model and test to account for incongruent gene trees with numerous taxa. Computer simulation studies supported the validity of the theoretical model and bootstrap test when each gene evolved at a constant rate. The computer simulation also suggested that the model remained valid as long as the rate heterogeneity was occurring proportionally in the same taxa for both genes. Although the model could not test hypotheses of hybridization versus lateral gene transfer as the cause of incongruence, these two processes may be distinguished by comparing phylogenies of multiple unlinked genes. [Gene tree; hybridization; phylogeny; species tree; topological incongruence.]

Hybridization, especially when coupled with polyploidization, is an important evolutionary mechanism in plants (Stebbins, 1950; Grant, 1981). Masterson (1994) has suggested that ~70% of angiosperms are polyploids, implying the possibility of a tremendous amount of hybridization in the evolutionary history of flowering plants. In animals, hybrid species are increasingly being documented with the application of molecular markers (Bullini, 1994).

Although hybridization has received considerable attention from evolutionary biologists (e.g., Arnold, 1997; Rieseberg, 1997), accurate reconstruction of hybrid speciation remains challenging. The reticulate nature of hybrid speciation fails to meet the basic assumption of cladistics that speciation occurs in a bifurcated manner (Hennig, 1966). Consequently, a phylogenetic tree containing unidentified hybrids remains an inaccurate reconstruction of the phylogeny. Furthermore, including hybrids in a phylogenetic analysis may result in an increased amount of homoplasy and possibly disrupt relationships of the other taxa (McDade, 1995).

Because of theoretical and technical difficulties involved in direct reconstruction of reticulate evolution, efforts have been made to explore alternative methods and new sources of data for improving our ability to reconstruct hybrid speciation. Three approaches have been used most frequently: (1) identifying hybrids before the phylogenetic analyses, (2) detecting hybrids according to their cladistic behavior, and (3) reconstructing hybrid speciation by comparing discordant positions of hybrids between phylogenetic trees obtained from independent data sets.

With the realization that inclusion of unidentified hybrids in a cladistic analysis may distort the relationships of other taxa, it is logical to follow the approach of identifying hybridization before starting the analysis, excluding the hybrids from the analysis, and adding the hybrids back onto the tree by connecting them with the putative parents. Morphological and molecular intermediacy or combination (or both) have served as criteria for identification of hybrids and their putative parents. In particular, continued development of new molecular markers, such as allozyme, restriction fragment length polymorphism, RAPD, ISSR, and DNA sequences, has contributed enormously to the

accuracy of identification of hybrids and their parentage (Rieseberg and Ellstrand, 1993; Sang et al., 1995; Campbell et al., 1997; Wolfe et al., 1998). Difficulty arises, however, when the hybrid is too ancient to maintain recognizable morphological or molecular intermediacy/combination.

Instead of avoiding the potential problems caused by hybrids in a cladistic analysis, the second approach chooses to identify a hybrid based on its behavior on a cladogram (Funk, 1985). Similar to the first approach, it relies on the assumption that a hybrid maintains morphological or molecular intermediacy/combination between the parents. The intermediacy or combination can be reflected in cladistic characters that are intermediate or polymorphic relative to those of the parents. Consequently, inclusion of the hybrid will increase the amount of homoplasy, largely because of parallelism between the hybrid and the parents (Funk, 1985). It is expected, therefore, that homoplasy can be considerably diminished by removing the hybrid from the data matrix. A computer program, RETICLAD, has been designed to implement this approach (Rieseberg and Morefield, 1995). This approach, however, may also be problematic when dealing with ancient hybrids in which morphological or molecular intermediacy/combination has been obscured. In addition, other factors, such as ancestral polymorphism or convergent evolution, can introduce homoplasy.

Extensive uses of chloroplast DNA (cpDNA) in plant phylogenetic studies have led to the third approach, which detects hybridization by comparing phylogenetic trees. In the majority of angiosperm species, cpDNA is maternally inherited (Mogensen, 1996). In the case of biparental inheritance, the polymorphism can be fixed relatively rapidly because of the small effective population size of cpDNA, which is one-quarter of a nuclear gene (Moore, 1995). Therefore, a cpDNA tree will most likely represent a uniparental phylogeny, in most cases, the maternal genealogy. When the cpDNA phylogeny is compared with another phylogenetic tree (morphology or molecular), conflicting positions of a taxon between these trees may be viewed as evidence for the hybrid origin of this taxon. A taxon possessing the cpDNA from a morphologically distinct taxon is known as cpDNA capture (Rieseberg and Soltis, 1991).

Besides cpDNA, nuclear ribosomal DNA (nrDNA) has been used most frequently in molecular phylogenetic studies of plants. When a species has substantially discordant positions between the cpDNA and nrDNA phylogenies, there is a possibility that the species may be a hybrid (Soltis and Kuzoff, 1995; Kellogg et al., 1996). The hybrid could have inherited cpDNA form the maternal parent and fixed nrDNA sequences of the paternal parent by way of concerted evolution and thus would have different sister group relationships between the cpDNA and nrDNA trees (Wendel et al., 1995; Sang et al., 1997). Comparing incongruence between gene trees thus opens the opportunity of reconstructing ancient hybridization, an event for which morphological intermediacy and molecular additivity in the hybrid subsequently have been obscured.

However, factors other than hybridization, such as random sorting of ancestral polymorphism (lineage sorting), gene duplication/deletion (paralogy), lateral gene transfer, or erroneous phylogenetic reconstruction, can also cause topological incongruence between gene trees. For example, paralogy in nrDNA repeats could potentially lead to inaccurate phylogenetic reconstructions in some plant groups, although nrDNA evolves together through concerted evolution (Buckler et al., 1997). Understanding and dealing with incongruence among gene trees are among the most acute theoretical issues in phylogenetics and have attracted considerable attention (e.g., de Queiroz et al., 1995; Huelsenbeck et al., 1996; Doyle, 1997; Maddison, 1997; Wendel and Doyle, 1998). Determining whether topological incongruence is caused by hybridization or by other factors, therefore, represents a major challenge to the approach of detecting hybridization by gene tree comparison. Here we will focus our discussion on incongruence caused by different phylogenetic histories of data sets and leave out the issue of erroneous phylogenetic reconstruction.

Maddison (1997) developed a theoretical model to distinguish the biological processes that may potentially cause incongruence between gene trees, including lineage sorting, gene duplication/extinction, and lateral gene transfer/hybridization. He used

a cladistic approach in which the number of events required to convert a gene tree to the species tree by assuming a certain process was counted; the process that required the fewest events was considered to be the cause of the incongruence between gene trees. However, at least two practical problems are associated with this model: the algorithmic difficulty of assessing all the possible topologies for a large number of species, and the determination of the appropriate weight of each event of different processes (Maddison, 1997).

In this paper, we attempt to depict patterns of topological incongruence between gene trees. Then, we focus on characterization of certain distinctions between hybridization and other processes as causes of topological incongruence. At first, to develop theoretical models, we contrast hybridization and lineage sorting, two processes often considered to be competing hypotheses for the incongruence of gene trees at lower taxonomical levels. Then we discuss how the models can be extended for testing other hypotheses, such as paralogy and lateral gene transfer, and ultimately how the hybridization hypothesis can be tested versus the rest of the hypotheses responsible for topological incongruence between gene trees.

Models for Comparing Incongruent Gene Trees

Three-Species Model

Let us first compare incongruent gene trees with only three ingroup species. Figures 1a and 1b show two gene trees generated from sequences of gene 1 and gene 2 of the ingroup species A, B, and C and the outgroup O. A_1, B_1, C_1, and O_1 are sequences of gene 1 from species A, B, C, and O, respectively; A_2, B_2, C_2, and O_2 are sequences of gene 2 from species A, B, C, and O, respectively. The two gene trees have incongruent topologies because of the discordant positions of the genes of species B; that is, B_1 forms a sister group with C_1 on gene tree 1, whereas B_2 is a sister group of A_2 on gene tree 2 (Figs. 1a, 1b).

Based on this simple example, a basic theoretical model can be constructed for testing the likelihood of hybridization versus lineage sorting or paralogy as the cause of the incongruence between the gene trees. On both gene trees, let T_0 be the time (million years before present) when genes of the ingroup species diverged from those of the outgroup species. T_0 is also the time when the ingroup species diverged from the outgroup species on the species tree, if the outgroup is selected in such a way that its relationship with the ingroup is most likely to represent a true species relationship. On gene tree 1, let T_i be the time when gene A_1 diverged from genes B_1 and C_1 , and T_j be the time when B_1 diverged from C_1. On gene tree 2, let T_k be the time when gene C_2 diverged from genes A_2 and B_2, and T_m be the time when A_2 diverged from B_2.

If this incongruence is caused by hybridization, then the species tree, with B being the hybrid and A and C being the parental lineages, can be inferred (Fig. 1c). The hybrid species B fixed sequences of gene 1 that are similar to those of species C, and thus genes

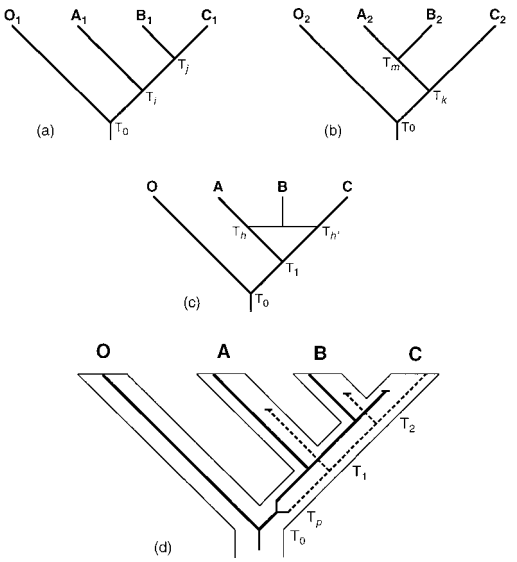

FIGURE 1. Gene trees and species trees of three-species model, where A, B, and C are ingroup species and O is an outgroup species. T_i, T_j, T_k, and T_m represent divergence times between genes A_1 and (B_1, C_1), B_1 and C_1, C_2 and (A_2, B_2), and A_2 and B_2, respectively. T_h and $T_{h'}$ represent times when the lineages that hybridized to give rise to B diverged from A and C, respectively. T_0, T_1, and T_2 represent times of speciation. T_p represents time of occurrence of polymorphic alleles. (a) Tree of gene 1. A_1, B_1, C_1, and O_1 are sequences of gene 1 of species A, B, C, and O, respectively. (b) Tree of gene 2. A_2, B_2, C_2, and O_2 are sequences of gene 2 of species A, B, C, and O, respectively. (c) Species tree inferred based on the hypothesis that B is a hybrid species. (d) Species tree (outlined by thin solid lines) inferred on the basis that gene 2 has undergone lineage sorting. Phylogeny of gene 2 is illustrated by thick solid and broken lines representing two ancestral alleles.

B_1 and C_1 form a sister group on gene tree 1. Meanwhile, the hybrid B inherited gene 2 from a parental lineage that is closely related to A, resulting in a sister group relationship between B_2 and A_2. Let T_1 be the time when species A and C diverged, and T_h and $T_{h'}$ be the times when the lineages that hybridized to give rise to B diverged from A and C, respectively. T_h will not equal $T_{h'}$ when the hybridization is ancient and the parental lineages that hybridized to produce B are extinct and diverged from A and C at different times.

Alternatively, lineage sorting of one gene may be the cause of the incongruence between the gene trees. Assume that gene tree 1 represents the species tree and gene 2 has undergone lineage sorting. The species tree and the phylogeny of gene 2 are illustrated in Figure 1d. Two polymorphic alleles of gene 2 arose in the common ancestor of species A, B, and C. Subsequently, one allele is maintained only in A and B, and the other is maintained only in C (Fig. 1d). Gene tree 2, reconstructed from these alleles, thus differs from the species tree. Let T_p be the time when the polymorphic alleles of gene 2 arose, T_1 be the time when species A diverged from species B and C, and T_2 be the time when species B and C diverged.

Under the hybridization hypothesis, $T_i = T_1$ and $T_j = T_{h'}$; $T_k = T_1$ and $T_m = T_h$ (Figs. 1a–c). Then, $T_i = T_k$. Under the lineage sorting hypothesis, $T_i = T_1$ and $T_j = T_2$; $T_k = T_p$, $T_m = T_1$ (Figs. 1a, 1b, 1d). Because $T_p > T_1$, then $T_k > T_i$. Therefore, we can test hypotheses of hybridization versus lineage sorting by testing $T_k = T_i$, or $T_k - T_i = 0$. Defining $\Delta(x, y) = T_x - T_y$, we can test:

$$\Delta(k, i) = \begin{cases} < 0, & \text{if hybridization;} \\ a(a > 0), & \text{if lineage-sorting for gene 2.} \end{cases} \quad (1)$$

In the real case, we do not know which gene has undergone lineage sorting before the test. If $\Delta(k, i) < 0$, lineage sorting may have occurred in gene 1.

To calculate $\Delta(k, i)$, the following expressions can be defined and inferred. Define $d(U_i, V_i)$ as the estimated number of nucleotide substitutions between sequences U_i and V_i of gene i of species U and V, respectively. Define r_i as the rate of nucleotide substitutions of gene i. Assuming the existence of a molecular clock, that is, a constant rate of nucleotide substitutions for both genes (Li, 1997), the substitution rates of gene 1 and gene 2 can be calculated as follows (Figs. 1a, 1b):

$$r_1 = \frac{d(O_1, A_1) + d(O_1, C_1)}{4T_0} = \frac{d(A_1, C_1)}{2T_i}, \quad (2)$$

$$r_2 = \frac{d(O_2, A_2) + d(O_2, C_2)}{4T_0} = \frac{d(A_2, C_2)}{2T_k}. \quad (3)$$

From Equations 2 and 3, we obtain

$$T_i = \frac{2d(A_1, C_1)T_0}{d(O_1, A_1) + d(O_1, C_1)}, \quad (4)$$

$$T_k = \frac{2d(A_2, C_2)T_0}{d(O_2, A_2) + d(O_2, C_2)}. \quad (5)$$

Therefore,

$$\Delta(k, i) = \left[\frac{2d(A_2, C_2)}{d(O_2, A_2) + d(O_2, C_2)} - \frac{2d(A_1, C_1)}{d(O_1, A_1) + d(O_1, C_1)} \right] T_0, \quad (6)$$

or

$$\Delta(k, i) = \Delta_0 T_0 \quad (7)$$

where

$$\Delta_0 = \frac{2d(A_2, C_2)}{d(O_2, A_2) + d(O_2, C_2)} - \frac{2d(A_1, C_1)}{d(O_1, A_1) + d(O_1, C_1)}. \quad (8)$$

Four-Species Model

The four-species model contains an additional ingroup species, D. The times of gene divergence are labeled on the two gene trees (Figs. 2a, 2b). The two genes of species C, C_1 and C_2, have discordant positions on the two gene trees, which leads to incongruence between the gene trees. Like the three-species model, the incongruence can be explained by either hybridization or lineage sorting. The species tree of hybridization (C is a hybrid between A and D) and the times of speciation and hybridization are shown in Fig. 2c. Under this hypothesis, $T_i = T_m = T_1$, $T_j = T_n = T_2$, $T_k = T_{h'}$, and $T_q = T_h$.

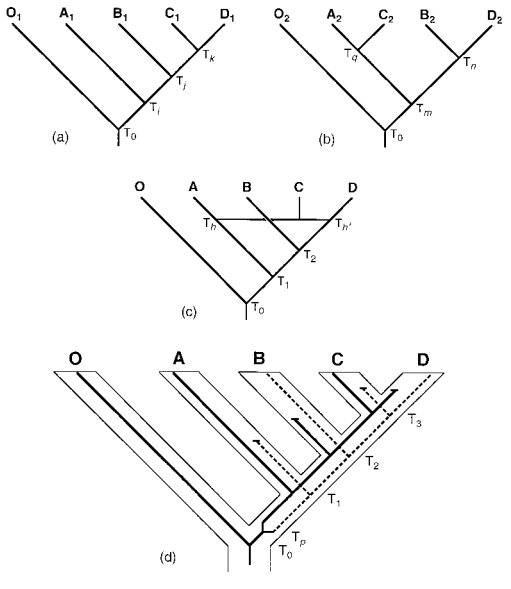

FIGURE 2. Gene trees and species trees of four-species model, where A, B, C, and D are ingroup species and O is an outgroup species. (a) Tree of gene 1. (b) Tree of gene 2. (c) Species tree inferred based on the hypothesis that species C is a hybrid. (d) Species tree inferred on the basis that gene 2 has undergone lineage sorting.

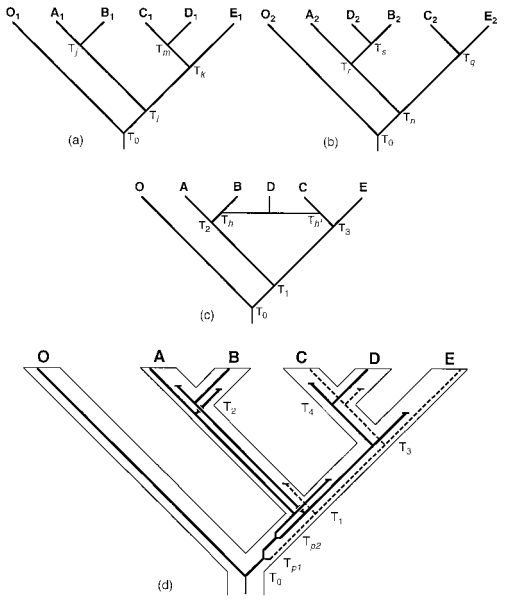

FIGURE 3. Gene trees and species trees of five-species model, where A, B, C, D, and E are ingroup species and O is an outgroup species. (a) Tree of gene 1. (b) Tree of gene 2. (c) Species tree inferred based on the hypothesis that species D is a hybrid. (d) Species tree inferred on the basis that gene 2 has undergone lineage sorting.

Alternatively, the incongruence between the gene trees is caused by lineage sorting. Assuming that gene 1 represents the species tree and gene 2 has undergone lineage sorting, the species tree and the contained phylogeny of gene 2 are illustrated in Figure 2d. Under this hypothesis, $T_i = T_q = T_1$, $T_j = T_n = T_2$, $T_k = T_3$, and $T_m = T_p$. Because $T_p > T_1$, then $T_m > T_i$.

Similar to the three-species model, these two hypotheses can be tested by testing $\Delta(m, i) = 0$, where

$$\Delta(m, i) = \ddagger \left[\frac{2d(A_2, D_2)}{d(O_2, A_2) + d(O_2, D_2)} - \frac{2d(A_1, D_1)}{d(O_1, A_1) + d(O_1, D_1)} \right] T_0. \quad (9)$$

Five-Species Model

The observed gene trees and hypothetical species trees of the five-species model are shown in Figure 3. Genes of species D, D_1 and D_2, display discordant positions on the two gene trees. Under the hybridization hypothesis, in which D is a hybrid between B and C (Fig. 3c), $T_i = T_n = T_1$, $T_j = T_r = T_2$, $T_k = T_q = T_3$, $T_m = T_{h'}$, and $T_s = T_h$. To explain the incongruence by lineage sorting of gene 2 (Fig. 3d), it is necessary to assume that polymorphic alleles of the gene arose twice, at times, T_{p1} and T_{p2}, in the common ancestor of the ingroup species. Under this hypothesis, $T_i = T_s = T_1$, $T_j = T_2$, $T_k = T_q = T_3$, $T_m = T_4$, $T_n = T_{p1}$, and $T_r = T_{p2}$. Because $T_{p1} > T_1$ and $T_{p2} > T_2$, then $T_n > T_i$ and $T_r > T_j$.

In this model, alternative hypotheses can be tested at two branching points of the gene trees: $\Delta(n, i) = 0$ and $\Delta(r, j) = 0$. The differences are calculated as follows:

$$\Delta(n, i) = \ddagger \left[\frac{2d(B_2, C_2)}{d(O_2, B_2) + d(O_2, C_2)} - \frac{2d(B_1, C_1)}{d(O_1, B_1) + d(O_1, C_1)} \right] T_0, \quad (10)$$

$$\Delta(r, j) = \ddagger \left[\frac{2d(A_2, B_2)}{d(O_2, A_2) + d(O_2, B_2)} - \frac{2d(A_1, B_1)}{d(O_1, A_1) + d(O_1, B_1)} \right] T_0. \quad (11)$$

If both differences are not significantly different from 0, the hybridization hypothesis

is favored. The lineage-sorting hypothesis is supported if both differences are significantly larger than 0.

Numerous-Species Model

If incongruence between two gene trees is caused by one species, the incongruent gene trees with any given number of species can be simplified to fit one of the three-, four-, or five-species models. In other words, these simple models can be extended to accommodate topological incongruence between gene trees with any given number of species after a simplification process, that is, converting a monophyletic or paraphyletic group into an individual species. For example, species h has different positions between two gene trees of 11 ingroup species, a, b, c, d, e, f, g, h, i, j, and k (Fig. 4). First, positions of some of the sister groups were switched without altering the topology of the gene trees. Second, the monophyletic groups, (d, e) and (a, (b, c)), are designated as single species, A and B, respectively; the paraphyletic group, (f((j,k),g)), is designated as E. With these rearrangements and conversions, the new trees fit the five-species model (Figs. 3, 4).

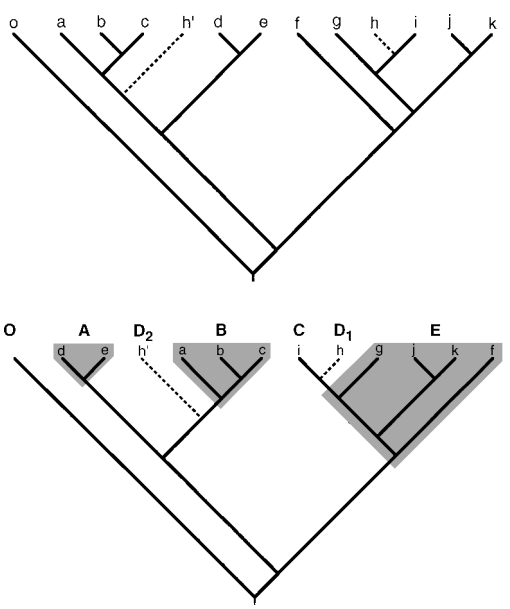

FIGURE 4. Simplification of incongruent gene trees with eleven ingroup species (a–k). (Top) Incongruence between two gene trees caused by conflicting positions of species h. Dashed branches represent different positions of h on two gene trees, whereas relationships of the remaining species are the same between the gene trees. (Bottom) Simplified gene trees that fit the five-species model.

With regard to hypothesis testing, one or more species can be chosen from a reduced group to represent this group. If two or more species are chosen, the hypotheses can be tested multiple times. Conducting the test over all possible combinations of the species may increase the amount of confidence of the hypothesis testing. Sequences that do not evolve in a clocklike fashion, however, should not be chosen to represent the group.

STATISTICAL TEST

A statistical test was designed to test hypotheses of hybridization versus lineage sorting, based on the theoretical models described above. For the three-species model, we need to test whether $\Delta(k, i)$ is significantly larger than 0 (Equation 1). According to Equation 7, testing $\Delta(k, i) = 0$ is equal to testing $\Delta_0 = 0$.

The number of nucleotide substitutions, d, can be estimated directly from sequence divergence when the sequence divergence is low. Otherwise, correction models, such as the Jukes–Cantor (1969) model or the Kimura (1980) two-parameter model, should be used to correct for possible multiple hits. Each estimated number of nucleotide substitutions has a variance that can be calculated depending on which model is used. However, calculation of the variance of Δ_0 based on these individual variances and their covariances (e.g., by using Taylor progression) remains difficult. The variance of the ratio of nucleotide substitutions tends to be too large to permit any reasonable power of a significance test (B. Gaut, pers. comm.; Z. Yang, pers. comm.).

Therefore, we chose the bootstrap method to carry out the statistical test (Efron, 1979; Felsenstein, 1985; B. Gaut, pers. comm.; Z. Yang, pers. comm.). Define $\Delta_0^{(j)}$ as a Δ_0 calculated from the jth bootstrapped data sets, where $j = 1, 2, \ldots, n$. To estimate the deviation of Δ_0 from the resampled information, the absolute values of $\Delta_0^{(j)}$ were used in the following calculation. The mean of $|\Delta_0^{(j)}|$ can be estimated as $M = \Sigma\,|\Delta_0^{(j)}|/n$. The variance of $|\Delta_0^{(j)}|$ is calculated as $V = \Sigma\,(|\Delta_0^{(j)}| - M)^2/n$. The statistics of the significance test can be calculated as $Z = [|\Delta_0| - M]/V^{1/2}$.

TABLE 1. Combinations of changes of parameters studied by computer simulation when $T_0 = 15$. In the case of the lack of a molecular clock, a correlation of rate heterogeneity is assumed between two genes; that is, the substitution rates of both genes in species A are either 1.5 times faster or 0.8 times slower than in the other species.

	With molecular clock						Without molecular clock			
							$r_1 = 1, r_2 = 1$			
	$r_1 = 1, r_2 = 1$		$r_1 = 1, r_2 = 2$		$r_1 = 2, r_2 = 1$		$r_A = 1.5$		$r_A = 0.8$	
	T_i	T_k	T_i	T_k	T_i	T_k	T_i	T_k	T_i	T_k
Hybridization	10	10	10	10	10	10	10	10	10	10
Lineage sorting	10	11	10	11	10	11	10	11	10	11
	10	12	10	12	10	12	10	12	10	12

COMPUTER SIMULATION

Computer simulation increasingly has been applied to exploration of theoretical issues in phylogenetics (e.g., Hillis et al., 1994; Huelsenbeck, 1995). A computer simulation is performed here to examine the validity of the theoretical models and the power of the statistical test under various evolutionary assumptions. Only the three-species model is examined with the computer simulation study. The implications of the simulation results should apply to the four- and five-species models, however, because of the similar nature of the test for all the models.

Two 1-kb DNA sequences, S_1 and S_2, were randomly generated and assigned to gene 1 and gene 2 of the most recent common ancestor of the outgroup and ingroup species (Figs. 1a, 1b), respectively. The G–C content of each sequence was set at 50%. Rates of nucleotide substitutions (r_i) were assumed to be 1 or 2 substitutions per 1,000 sites per million years. This corresponds to 1 (or 2) $\times 10^{-9}$ substitutions per site per year, which is biologically reasonable (Li, 1997). When $r_i = 1$, one substitution was generated randomly for a sequence during a one-million-year interval along a gene tree. All types of substitutions were treated at an equal probability. Through this process, gene sequences of the species A_1, B_1, C_1, O_1, A_2, B_2, C_2, and O_2 were generated.

The simulation experiment, with the same initial sequences of S_1 and S_2, was performed 20 times for testing hypotheses of hybridization versus lineage sorting and for testing the impact of several parameters on the model. The parameters, which were tested under the hypotheses of hybridization ($T_k - T_i = 0$) or lineage sorting ($T_k - T_i = 1, 2$), include relative substitution rates between genes 1 and 2, the time of divergence between ingroup and outgroup, and the molecular clock. With the molecular clock engaged (constant substitution rates for both genes), we tested hypotheses of hybridization and lineage sorting when the relative substitution rates of the genes were the same or when one gene evolved twice as fast as the other (Table 1). We also tested the hypotheses when substitution rates were not constant for either gene, but with the restriction that the rate heterogeneity was correlated between two genes; that is, the substitution rates of both genes increased or decreased proportionally in the same taxa (Table 1). These tests were done under the conditions of $T_0 = 15$ and $T_j = T_m = 5$ (Fig. 1; Table 1). Assigning different values to T_j and T_m would not affect testing results (data not shown).

We explored the impact of increasing the overall sequence divergence on the hypothesis testing. When the overall divergence was doubled—$T_0 = 30$ and $T_j = T_m = 10$—we tested the following combinations under condition of $r_1 = r_2 = 1$: $T_i = 20$ and $T_k = 20$; $T_i = 20$ and $T_k = 21$; $T_i = 20$ and $T_k = 22$; $T_i = 20$ and $T_k = 23$; and $T_i = 20$ and $T_k = 24$.

Using sequences generated from the computer simulation, each Δ_0 was calculated with Equation 8. Given the relatively low divergence between the sequences (up to 3%), the number of nucleotide substitutions was estimated directly from the sequence divergence. The difference between each Δ_0 and 0 was tested through 100 bootstrap replications.

RESULTS AND DISCUSSION

The computer simulation study demonstrated that the statistical test is able to support either hybridization or lineage-sorting hypotheses predicted by the theoretical

model. If each gene evolves at a constant rate, then Δ_0 equals the ratio of $(T_k - T_i)/T_0$ (see proof in Appendix). Thus, when T_0 is fixed, the model predicts that the greater the difference between T_k and T_i, the greater the Δ_0 value, that is, the more likely that lineage sorting has occurred. When T_0 was set at 15 and the hybridization hypothesis is engaged ($T_i = T_k = 10$), the Δ_0 values fluctuate around 0 (Fig. 5a), and >80% of Δ_0 values are not significantly larger than 0 ($P < 0.05$) (Fig. 6a). Therefore, the hybridization hypothesis was confirmed by the test in >80% of simulation runs.

Under the lineage-sorting hypothesis and when $T_k - T_i = 1$, almost all the Δ_0 values are >0 (Fig. 5a), and 45% ($P < 0.01$) to 85% ($P < 0.05$) of the values are significantly >0 (Fig. 6a). These results imply that when the ingroup has diverged from the outgroup for 15 million years and when the ancestral polymorphic alleles occurred one million years before the divergence of species A and C, the chance that the test is able to confirm the lineage sorting is ~50% or more. When $T_k - T_i = 2$, all the Δ_0 values are >0 (Fig. 5a), and >90% of test results are significant at $P < 0.01$ (Fig. 6a). This suggests that the chance of detecting lineage sorting increases rather rapidly as the time of origination of the ancestral polymorphism becomes more ancient relative to the time of the divergence of species A and C (Figs. 1d, 6a).

A slight fluctuation of Δ_0 during 20 runs of simulation for each combination is due to the small number of parallel substitutions and multiple hits. We tried to correct each of the sequence divergence by using the one parameter model (Jukes and Cantor, 1969), but the degree of fluctuation of Δ_0 values was not reduced significantly thereby (data not shown). This is probably a result of the calculation of the ratio of the sequence divergence, which canceled the effect of the correction. The fact that the simulation studies supported the theoretical models in the presence of multiple hits suggests that the models and statistical test can tolerate a certain number of multiple hits.

The computer simulation study demonstrated that doubling the overall divergence did not alter the test results markedly. When T_0 was increased to 30, the number of significant test results when $T_k - T_i = 2$ was comparable with that for when $T_k - T_i = 1$ and $T_0 = 15$ (Figs. 5a, 5d, 6a, 6d). Because Δ_0 remains the same when $T_k - T_i$ and T_0 increase or decrease proportionally, the model should work at various values for sequence divergence, which correspond, in most cases, to various taxonomical values. However, if the divergence of the ingroup is fixed, that is, if $T_k - T_i$ remains constant, increasing T_0 (choosing a more distantly related outgroup) will lead to a smaller Δ_0. The number of significant test values when $T_0 = 30$ and $T_k - T_i = 1$ is intermediate between those of $T_k - T_i = 0$ and $T_k - T_i = 1$ when $T_0 = 15$ (Figs. 5a, 5d, 6a, 6d), which suggests that choosing an outgroup that is more closely related to the ingroup will permit a more sensitive test for lineage sorting.

Because rates of nucleotide substitution vary widely among genes (Li, 1997), one must determine whether the theoretical model can still hold when two genes evolve at different rates. Theoretically, Δ_0 should not be affected by rate differences between the two genes, as long as the substitution rate of each gene is constant (see proof in Appendix). The simulation results supported this prediction. In the case where the substitution rate of either gene 1 or gene 2 was doubled, the test results were similar in both cases and were also similar to that obtained by assuming the equal substitution rate of the two genes (Figs. 5a–c, 6a–c).

However, the lack of constant substitution rates of one or both genes among the studied species will violate the basic assumption of the model. Using simulation studies, we explored the validity of the model when the rate heterogeneity was correlated between the two genes; that is, rates increased or decreased proportionally in both genes of the same species. Satisfying this condition should extend application of the model, given that the rate heterogeneity caused by factors such as generation time effect can be correlated between genes or even across different genomes (Wu and Li, 1985; Gaut et al., 1992). For example, grasses evolve more rapidly than palms at synonymous sites in a mitochondrial, a nuclear, and a chloroplast gene, and the rate increases in grasses are correlated among these genes (Eyre-Walker and Gaut, 1997). The results of the simulation study (Figs. 5e–f, 6e–f), in which substitution rates of both genes of species A were proportionally higher or lower than those of species C and O, suggested that the model should be valid under such a condition.

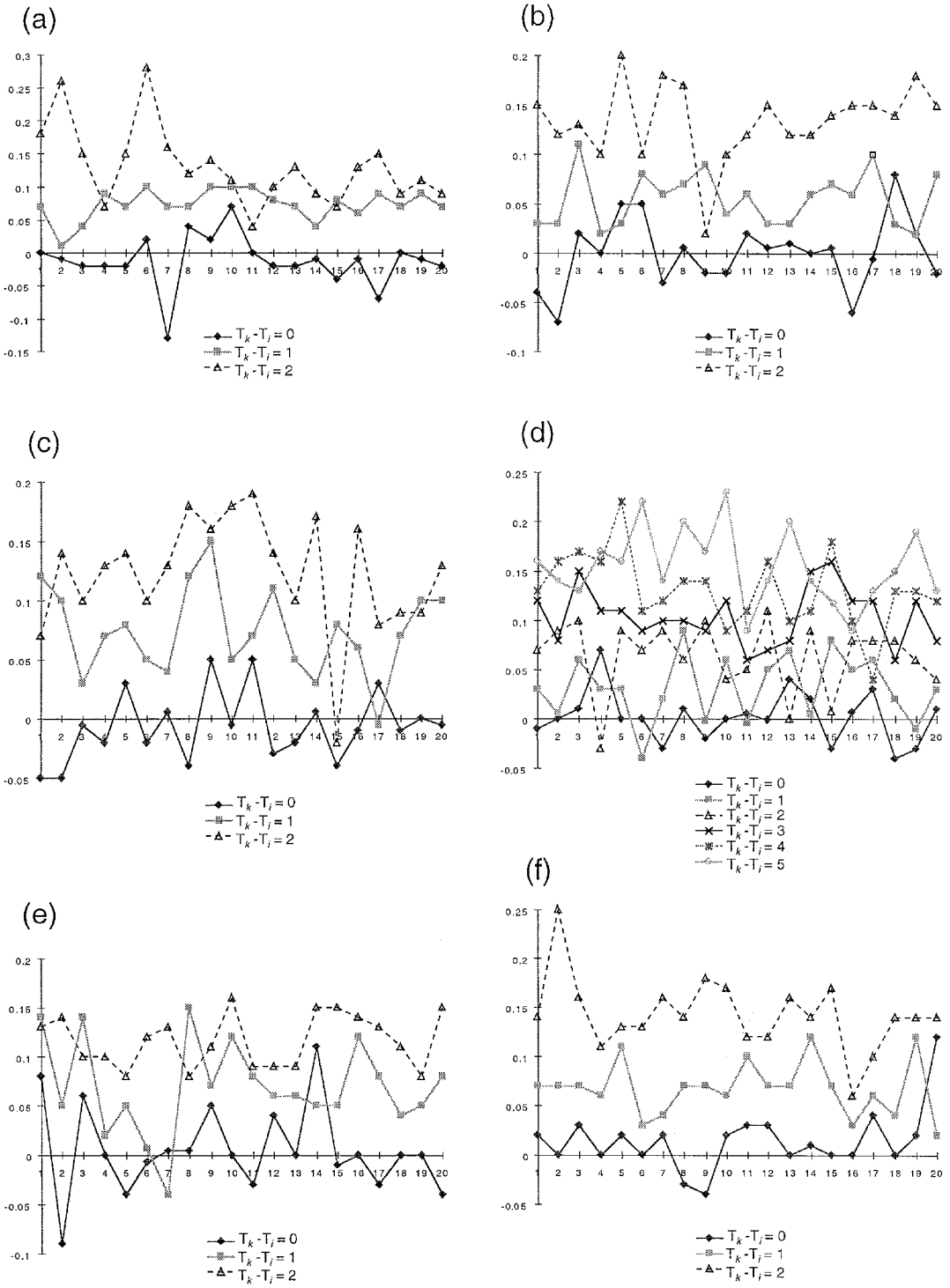

FIGURE 5. Δ_0 values generated from computer simulation. Each Δ_0 value is represented by an abscissa; Δ_0 values resulting from 20 simulation runs for a combination of a hypothesis and various parameters are connected by a certain type of line for clearer visualization (Table 1). (a) $T_0 = 15$, $r_1 = r_2 = 1$. (b) $T_0 = 15$, $r_1 = 1$, $r_2 = 2$. (c) $T_0 = 15$, $r_1 = 2$, $r_2 = 1$. (d) $T_0 = 30$, $r_1 = r_2 = 1$. (e) $T_0 = 15$, $r_1 = r_2 = 1$, $r_A = 1.5$. (f) $T_0 = 15$, $r_1 = r_2 = 1$, $r_A = 0.8$.

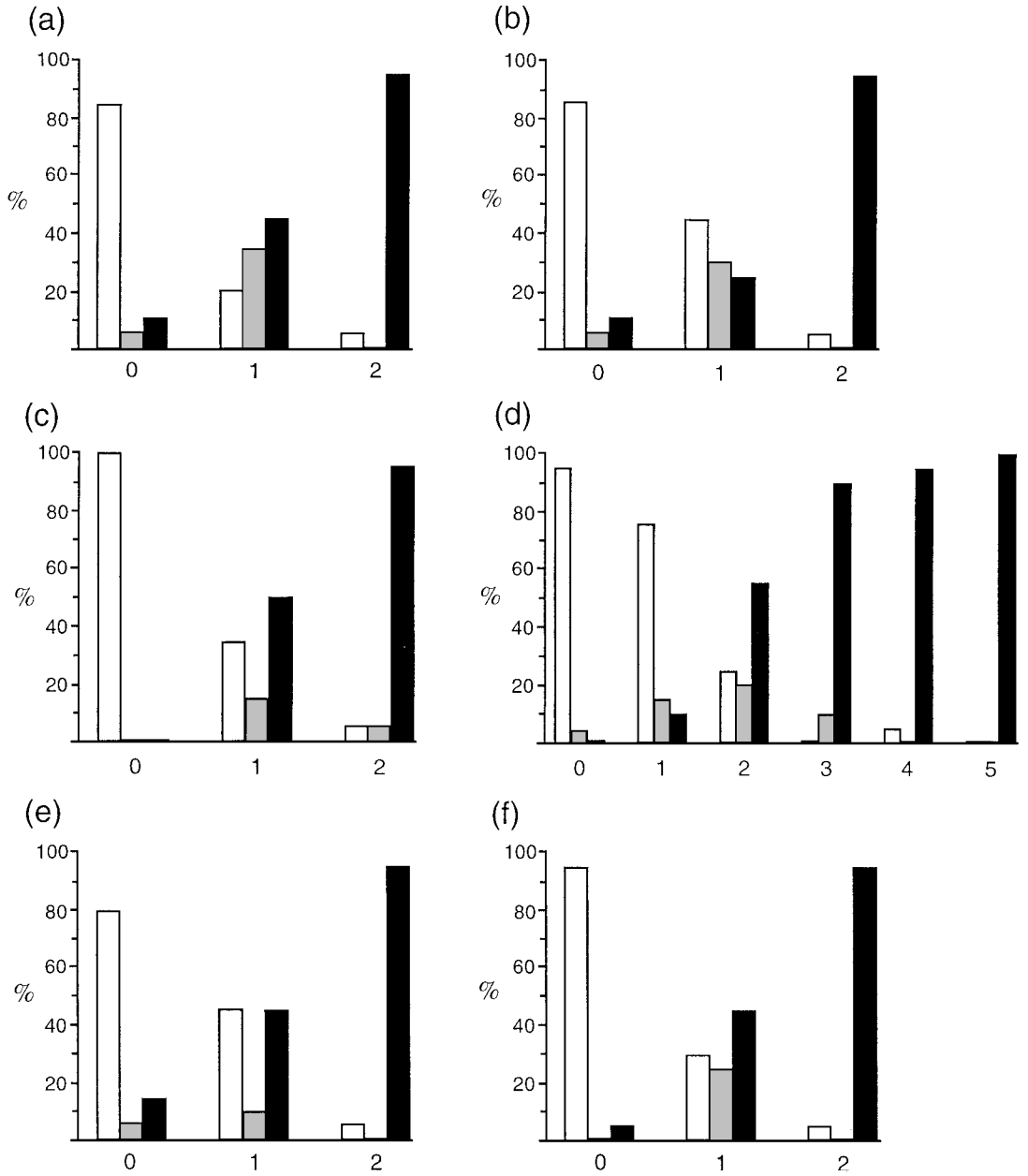

FIGURE 6. Test results of computer simulation study. (a), (b), (c), (d), (e), and (f) show test results of Δ_0 presented in Figures 5a, 5b, 5c, 5d, 5e, and 5f, respectively. Histograms illustrate percentages of probabilities of the test results of 20 simulation runs for each combination of a hypothesis and various parameters; $P > 0.05$ (open), $0.05 > P > 0.01$ (shaded), and $P < 0.01$ (solid). Numbers 0, 1, 2, 3, 4, and 5 below the histograms represent corresponding values of $(T_k - T_i)$ in Figure 5.

Because the validity of the theoretical models relies on the molecular clock for both genes, a relative rate test must be performed before the hybridization hypothesis can be tested. However, the degree of sensitivity to rate heterogeneity may differ between the relative rate test (Wu and Li, 1985) and the present bootstrap test. Problems may arise if the bootstrap test is more sensitive than the relative rate test to rate heterogeneity. In other words, the rate heterogeneity can be small enough to pass the relative rate test

TABLE 2. Comparison of probabilities obtained from relative rate (Wu and Li, 1985) under the Jukes–Cantor (1969) model and bootstrap tests based on the three-species model. For both gene trees, $T_0 = 15$ and $T_i = T_k = 10$; $r_1 = r_2 = 1$ for all the sequences except r_{A_2}, which varies from 1.2, 1.3, or 1.4 to 1.5, meaning that the substitution rate of sequence A_2 is 1.2, 1.3, 1.4, or 1.5 times faster than those of the remaining sequences.

	r_{A_2}			
	1.5	1.4	1.3	1.2
Relative rate test	0.16	0.21	0.27	0.34
Bootstrap test	<0.01	<0.01	0.04	0.20

but large enough to affect the results of hypothesis testing with the bootstrap test. To address this question, computer simulation was designed to assess the relative sensitivity of the two tests to the degree of rate heterogeneity. Under the hybridization hypothesis ($T_k - T_i = 0$), the substitution rate of gene A_2 was increased while those of the other genes were kept the same (Table 2). The probabilities obtained from both tests are compared in Table 2.

When substitution rates are variable among the species and the rate variation is not correlated between the two genes, the model is rather sensitive to rate heterogeneity. We explored the extent to which rate heterogeneity misleads the test results. Under the hybridization model ($T_k - T_i = 0$), the hybridization hypothesis was rejected ($P < 0.01$) when gene A_2 evolved >1.4 times more rapidly than the other genes (Table 2). When A_2 evolved 1.3 times more rapidly, the hybridization was still rejected ($P = 0.04$). When the substitution rate of A_2 was 1.2 times greater than that of the other genes, the hybridization hypothesis could not be rejected.

Therefore, although our model and test can tolerate a certain degree of rate heterogeneity, they appear to be quite sensitive to rate heterogeneity, at least more sensitive than the relative rate test of Wu and Li (1985). Consequently, the results of the bootstrap test may not be free of bias attributable to the rate heterogeneity, even though the molecular clock is not rejected by the relative rate test. Future work should explore further the impact of the degree of rate heterogeneity and the way of measuring it on the efficiency of the bootstrap test. For example, a maximum likelihood method represents another approach to reconstructing a phylogeny and estimating branch lengths. If sequences of two genes are determined by a likelihood ratio test to have evolved among the studied species in a clocklike fashion (Huelsenbeck and Rannala, 1997; Baldwin and Sanderson, 1998), then the branch lengths estimated by maximum likelihood can be used in the bootstrap test.

So far we have discussed lineage sorting only as an alternative explanation to hybridization in causing incongruence between gene trees. Now we extend our discussion to account for the other two biological processes that are also responsible for the departure of gene trees from the species tree: gene duplication/extinction (paralogy), and lateral gene transfer (de Queiroz et al., 1995; Doyle, 1997; Maddison, 1997; Wendel and Doyle, 1998).

Paralogy has an impact on the topology of gene trees similar to that of lineage sorting and can be tested by using the same model: $\Delta(k, i) > 0$. In these models, paralogy is analogous to lineage sorting when T_k is viewed as the time of duplication of two loci instead of two alleles of the same locus. The subsequent extinction of one allele from a species under the lineage-sorting hypothesis is analogous to extinction or incomplete sampling of one of the loci in the paralogy hypothesis. The same species tree can be inferred no matter which of the two processes caused the incongruence.

Despite the similar behavior of lineage sorting and paralogy in the models, they are different biological processes that can be distinguished in certain aspects (Maddison, 1997; Wendel and Doyle, 1998). Basically, the occurrence of lineage sorting depends on a persistence of ancestral alleles through a common ancestor of the diverged species. Therefore, lineage sorting can often be found when an ancestral branch leading to the species is short (few generations) and wide (large effective population size) (Pamilo and Nei, 1988; Hudson, 1992; Maddison, 1997). Therefore, lineage sorting is topology-dependent and is expected to occur often along the short branches of a gene tree. The occurrence of paralogy, on the other hand, depends largely on the dynamics of gene duplication and extinction during its evolution (Morton et al., 1996). Understanding molecular evolution of individual genes can help assess the likelihood of occurrence of paralogy at each locus.

Lateral gene transfer follows the same test result as hybridization, $\Delta(k, i) = 0$, for

the theoretical model. However, lateral gene transfer often happens to transposable elements (Clark et al., 1994; Syvanen, 1994). It is also very unlikely that many unlinked genes can be transferred from one species to the other in parallel. In contrast, hybridization, particularly by way of allopolyploidy, combines two entire genomes. Therefore, if $\Delta(k, i) = 0$ is the test result for multiple unlinked genes, the incongruence is much more likely to result from hybridization than from lateral gene transfer.

ACKNOWLEDGMENTS

We thank B. Gaut, F. Kong, and Z. Yang for valuable suggestions on statistical tests, and D. Ferguson, R. Olmstead, and two anonymous reviewers for valuable suggestions and comments on the manuscript.

REFERENCES

ARNOLD, M. L. 1997. Natural hybridization and evolution. Oxford Univ. Press, New York.

BALDWIN, B. G., AND M. J. SANDERSON. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae) Proc. Natl. Acad. Sci. USA 95:9402–9406.

BUCKLER, E. S., IV, A. ANTHONY, AND T. P. HOLTSFORD. 1997. The evolution of ribosomal DNA: Divergent paralogues and phylogenetic implications. Genetics 145:821–832.

BULLINI, L. 1994. Origin and evolution of animal hybrid species. Trends Ecol. Evol. 9:422–426.

CAMPBELL, C. S., M. F. WOJCIECHOWSKI, B. G. BALDWIN, L. A. ALICE, AND M. J. DONOGHUE. 1997. Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). Mol. Biol. Evol. 14:81–90.

CLARK, J. B., W. P. MADDISON, AND M. G. KIDWELL. 1994. Phylogenetic analysis supports horizontal transfer of *P* transposable elements. Mol. Biol. Evol. 11:40–50.

DE QUEIROZ, A., M. J. DONOGHUE, AND J. KIM. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26:657–681.

DOYLE, J. J. 1997. Trees within trees: Genes and species, molecules and morphology. Syst. Biol. 46:537–553.

EFRON, B. 1979. Bootstrap methods: Another look at the jacknife. Ann. Stat. 7:1–26.

EYRE-WALKER, A., AND B. S. GAUT. 1997. Correlated rates of synonymous site evolution across plant genomes. Mol. Biol. Evol. 14:455–460.

FELSENSTEIN, J. 1985. Confidence limits on phylogenetics: An approach using the bootstrap. Evolution 39:783–791.

FUNK, V. A. 1985. Phylogenetic pattern and hybridization. Ann. Mo. Bot. Gard. 72: 681–715.

GAUT, B. S., S. V. MUSE, W. D. CLARK, AND M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J. Mol. Evol. 35:292–303.

GRANT, V. 1981. Plant speciation, 2nd edition. Columbia Univ. Press, New York.

HENNIG, W. 1966. Phylogenetic systematics. Univ. Illinios Press, Urbana-Champaign.

HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenetics. Science 264:671–677.

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

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK, J. P., J. J. BULL, AND C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

HUELSENBECK, J. P., AND B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KELLOGG, E. A., R. APPELS, AND R. J. MASON-GAMER. 1996. When genes tell different stories: The diploid genera of *Triticeae* (Gramineae). Syst. Bot. 21:321–347.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Massachusetts.

MADDISON, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

MASTERSON, J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. Science 264:421–424.

MCDADE, L. A. 1995. Hybridization and phylogenetics. Pages 305–331 *in* Experimental and molecular approaches to plant biosystematics (P. C. Hoch and A. G. Stephenson, eds.), Missouri Botanical Garden, St. Louis.

MOGENSEN, H. L. 1996. The hows and whys of cytoplasmic inheritance in seed plants. Am. J. Bot. 83:383–404.

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

MORTON, B. R., B. GAUT, AND M. CLEGG. 1996. Evolution of alcohol dehydrogenase genes in the Palm and Grass families. Proc. Natl. Acad. Sci. USA 93:11735–11739.

PAMILO, P., AND M. NEI. 1988. Relationship between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

RIESEBERG, L. H. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359–389.

RIESEBERG, L. H., AND N. C. ELLSTRAND. 1993. What can molecular and morphological markers tell us about plant hybridization? Crit. Rev. Plant Sci. 12:213–241.

RIESEBERG, L. H., AND J. D. MOREFIELD. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. Pages 333–353 *in* Experimental and molecular approaches to plant biosystematics (P. C. Hoch and A. G. Stephenson, eds.), Missouri Botanical Garden, St. Louis.

RIESEBERG, L. H., AND D. E. SOLTIS. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants 5:65–84.

SANG, T., D. J. CRAWFORD, AND T. F. STUESSY. 1995. Documentation of reticulate evolution in peonies

(*Paeonia*) using sequences of internal transcribed spacer of nuclear ribosomal DNA: Implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. USA 92:6813–6817.

SANG, T., D. J. CRAWFORD, AND T. F. STUESSY. 1997. Chloroplast phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. 84:1120–1136.

SOLTIS, D. E., AND R. K. KUZOFF. 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). Evolution 49:727–742.

STEBBINS, G. L., JR. 1950. Variation and evolution in plants. Columbia Univ. Press, New York.

SYVANEN, M. 1994. Horizontal gene transfer: Evidence and possible consequences. Annu. Rev. Genet. 28:237–261.

WENDEL, J. F., AND J. J. DOYLE. 1998. Phylogenetic incongruence: Window into genome history and molecular evolution. Pages 265–296 *in* Molecular systematics of plants, II: DNA sequencing (D. Soltis, P. Soltis, and J. Doyle, eds.) Kluwer Academic Publishers, Boston.

WENDEL, J. F., A. SCHNABEL, AND T. SEELANAN. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA 92:280–284.

WOLFE, A. D., Q. Y. XIANG, AND S. R. KEPHART. 1998. Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). Proc. Natl. Acad. Sci. USA 95:5112–5115.

WU, C.-I., AND W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA 82:1741–1745.

Received 16 November 1998; accepted 27 January 1999
Associate Editor: R. Olmstead

APPENDIX

The branch length (the number of nucleotide substitutions) of the gene trees 1 and 2 of the three species model can be calculated as follows (Figs. 1a, 1b):

$$d(A_1, C_1) = 2T_i r_1,\ d(O_1, A_1) = d(O_1, C_1) = 2T_0 r_1;$$

$$d(A_2, C_2) = 2T_k r_2,\ d(O_2, A_2) = d(O_2, C_2) = 2T_0 r_2.$$

According to Equation 8,

$$\Delta_0 = \frac{4T_k r_2}{4T_0 r_2} - \frac{4T_i r_1}{4T_0 r_1}.$$

Define $r_2 = \alpha r_1$, where $\alpha > 0$ ($r_2 = r_1$ if $\alpha = 1$), then

$$\Delta_0 = \frac{4\alpha T_k r_1}{4\alpha T_0 r_1} - \frac{4T_i r_1}{4T_0 r_1} = \frac{T_k - T_i}{T_0}.$$