



Difficulties in Detecting Hybridization

Mark T. Holder; Jennifer A. Anderson; Alisha K. Holloway

Systematic Biology, Vol. 50, No. 6. (Nov. - Dec., 2001), pp. 978-982.

Stable URL:

<http://links.jstor.org/sici?sici=1063-5157%28200111%2F12%2950%3A6%3C978%3ADIDH%3E2.0.CO%3B2-P>

Systematic Biology is currently published by Society of Systematic Biologists.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ssbiol.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

genome? To what extent do the particular assumptions made in the analysis (e.g., independence of sites, evolutionary model, same substitution rate across sites) affect the number and type of local maxima? Such questions provide a rich source for investigation into the efficiencies of phylogenetic reconstruction methods.

ACKNOWLEDGMENTS

I thank Dennis Pearl, Richard Olmstead, and two anonymous reviewers for helpful comments on an earlier draft of this manuscript.

REFERENCES

- BROWN, E., AND W. DAY. 1984. A computationally efficient approximation to the nearest neighbor interchange metric. *J. Classif.* 1:93–124.
- CHAN, S., H. BERNARD, S. ONG, S. CHAN, B. HOFMANN, AND H. DELIUS. 1992. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: A showcase for the molecular evolution of DNA viruses. *J. Virol.* 66:5714–5725.
- CHAN, S., H. DELIUS, A. HALPERN, AND H. BERNARD. 1995. Analysis of genomic sequences of 95 papillomavirus types: Uniting typing, phylogeny, and taxonomy. *J. Virol.* 69:3074–3083.
- CHANG, J. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Bio.* 137:51–73.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: A justification. *Evolution* 38:16–24.
- FELSENSTEIN, J. 1993. PHYLIP (Phylogenetic inference package), version 3.5p. Univ. of Washington, Seattle.
- GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182–198.
- HASEGAWA, M., H. KISHINO, AND N. SAITOU. 1991. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* 32:443–445.
- KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29:170–179.
- LEWIS, P. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* 15:277–283.
- MADDISON, D. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool.* 40:315–328.
- ONG, C., S. NEE, A. RAMBAUT, H. BERNARD, AND P. HARVEY. 1997. Elucidating the population histories and transmission dynamics of papillomaviruses using phylogenetic trees. *J. Mol. Evol.* 44:199–206.
- PAGE, R. 1993. On islands of trees and the efficacy of different methods of branch swapping in finding most-parsimonious trees. *Syst. Biol.* 42:200–210.
- PAGE, R. 2001. COMPONENT, version 2.0. The Natural History Museum, London.
- ROGERS, J. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 46:354–357.
- SALTER, L. 1999. Simulation-based estimation of phylogenetic trees. Ph.D. Dissertation, Ohio State Univ., Columbus.
- SALTER, L., AND D. PEARL. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst. Biol.* 50:7–17.
- SANDERSON, M., AND J. KIM. 2000. Parametric phylogenetics? *Syst. Biol.* 49:817–829.
- SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- YANG, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* 43:329–342.

Received 8 March 2000; accepted 9 October 2000
Associate Editor: R. Olmstead

Difficulties in Detecting Hybridization

MARK T. HOLDER,^{1,2} JENNIFER A. ANDERSON,¹ AND ALISHA K. HOLLOWAY¹

¹Section of Integrative Biology, School of Biological Sciences, University of Texas, Austin, Texas 78712, USA;

E-mail: mtholder@mail.utexas.edu, janders@mail.utexas.edu, aholloway@mail.utexas.edu

²Institute for Cellular and Molecular Biology, University of Texas, Austin, Texas 78712, USA

Sang and Zhong (2000) proposed a test to distinguish between hybridization and lineage sorting, two of the many evolutionary processes that can produce discordant gene trees. Consider two gene trees for three taxa A, B, and C (Fig. 1A,B). If B and C are sister

taxa on gene tree 1, and A and B are sister taxa on gene tree 2, then either taxon B is a hybrid species (Fig. 1C), or one of the gene trees is incorrect because of lineage sorting (Fig. 1D,E). When faced with these two discordant gene trees, Sang and Zhong

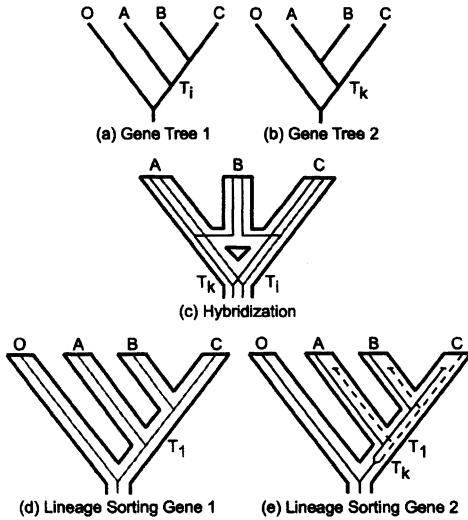


FIGURE 1. Gene trees and species trees of ingroup taxa A, B, and C, and outgroup taxon O. T_i and T_k are the divergence times between taxa A and C on gene trees 1 and 2, respectively. T_1 represents the speciation event between taxa A and C. (a) Gene tree 1. O_1 , A_1 , B_1 , and C_1 are the sequences of taxa O, A, B, and C, respectively. (b) Gene tree 2. (c) Under the hybridization hypothesis, the species tree (outlined by thick, solid lines) shows taxon B intermediate between taxa A and C. Both gene trees are drawn inside of the species tree. Gene 1 is indicated in solid gray lines, and gene 2 is in solid black lines. (d) Under lineage sorting, the species tree (outlined by thick, solid lines) and gene tree 1 (solid gray lines) are in agreement. (e) Under lineage sorting, gene tree 2 disagrees with the species tree. T_k indicates the origin of the ancestral polymorphism, and T_1 is the speciation event between taxa A/B and C. The solid black lines represent one allele of gene 2 uniting taxa A and B, and the dashed black lines indicate the other allele, which is fixed in taxon C.

(hereafter referred to as S&Z) recommended calculating a test statistic based on the divergence time of taxa A and C on each tree. If the two estimates of divergence time are equal, hybridization is favored; if the two estimates are not equal, lineage sorting is supported. This test implicitly assumes that the variance in coalescence times of genes is negligible, except in cases of lineage sorting. However, in real data, coalescence times vary among genes, regardless of whether lineage sorting is a problem. The S&Z statistic may be able to detect a difference in coalescence times between two genes, but we will show that this statistic does not reliably discriminate between hybridization and lineage sorting.

S&Z estimated the time of divergence between taxa A and C on gene tree 1 and gene tree 2 as T_i and T_k , respectively (Fig. 1A,B), and designated the true time

of the speciation event leading to A and C as T_1 . They assumed that if hybridization explains the discordant gene topologies, then T_i and T_k must both date the speciation event between taxa A and C ($T_1 = T_i = T_k$; Fig. 1C). Hybridization is therefore supported when the difference between T_i and T_k is not significantly different from zero. In contrast, if lineage sorting is responsible for the disagreement among gene trees, the coalescence time for one of the genes (gene 2, Fig. 1E) dates the origin of an ancestral polymorphism as preceding the speciation event ($T_k > T_1$). Thus, if the difference between T_i and T_k is significantly greater than zero, their model implies lineage sorting.

COALESCENCE TIMES

The assumptions that T_i and T_k should date the speciation event (T_1) in the case of hybridization and that T_i should date the speciation event when lineage sorting is involved are incorrect. T_i and T_k are better calculated by summing the time since the speciation event (T_1) and the time at which the alleles for each of the genes were distinct in the population of the last common ancestor. We will refer to the coalescence time of the alleles A and C in the population of the last common ancestor as CT_i and CT_k . Thus, T_i would actually equal the sum of T_1 and a coalescence time (CT_i), and T_k would be the sum of T_1 and a different coalescence time (CT_k). Detecting a difference between these two divergence times does not necessarily support the lineage sorting hypothesis. Indeed, one may detect a significant difference in coalescence times between the two genes ($T_i \neq T_k$) when actually a hybridization event is responsible for the pattern of discordance between the gene trees.

Under the hybridization hypothesis, the expected coalescence times for both genes (CT_i and CT_k) should be drawn from an exponential distribution (Hudson, 1982). For a gene undergoing lineage sorting, the distribution of coalescence times for the gene in the ancestral population (CT_k) should be drawn from a different distribution. If A and B are sister to each other on one gene tree because of ancestral polymorphism, then the divergence of their alleles precedes T_1 by one coalescence time. CT_k is sum of this coalescence time and the time to coalescence between the allele found in C and the last common

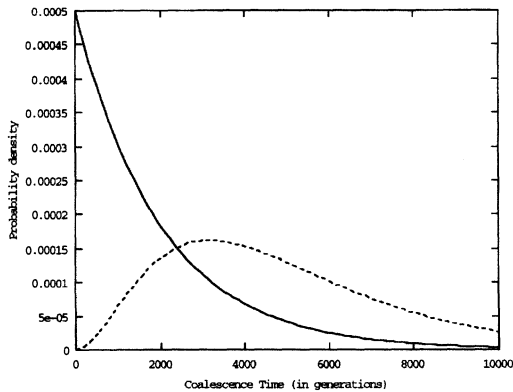


FIGURE 2. The coalescence times CT_i (solid line) and CT_k (dashed line) for a population size of 1,000, a sequence length of 1,000, and a mutation rate of 1×10^{-9} . CT_i is calculated by using the standard exponential equation for the coalescence time in a population of 1,000. CT_k is the deepest coalescence time of three alleles in a population, conditional on there being at least one parsimony-informative mutation in the genealogy.

ancestor of the alleles in A and B. Because lineage sorting will create a problem for phylogeny estimation only when a mutation occurs uniting A and B, the calculation of this second coalescence time must be altered so that the times are drawn from a distribution of coalescence times, given that there is a mutation. The expectation is that CT_k is larger under the lineage sorting hypothesis than under hybridization, as S&Z suggest, but the difference is slight compared with the large variance in coalescence times (Fig. 2).

SIMULATIONS

S&Z presented simulations to demonstrate that their statistic can detect a significant difference between T_i and T_k when these divergence times differ by 1–2 million years. Unfortunately, they failed to simulate any variance in coalescence time, and Figure 2 illustrates that the distributions for CT_i and CT_k (and therefore T_i and T_k) have high variance and are broadly overlapping. To account for this variation, we repeated the first set of simulations described by S&Z, using the same species tree (outgroup separated from A and C by 15 million years, and A separated from C by 10 million years), sequence length (1,000 bases), and mutation rate (1 change per site per 10^9 years). Unlike S&Z, we accounted for variation in coalescence times by using the

two distributions described above. Because the effective population size of a gene determines the distribution of its coalescence times, we simulated data over a range of population sizes (10^1 – 10^5 ; see Table 1). For each set of parameters we simulated 1,000 datasets under the hybridization hypothesis (all coalescence times were drawn from the standard coalescence distribution) and 1,000 datasets under the lineage sorting hypothesis (coalescence times for gene 1 were drawn from the standard coalescence time distribution; for gene 2, a mutation was added to the branch leading to taxon A, and the coalescence time was drawn from the distribution of coalescence times given that at least one mutation occurred while the alleles were in the last common ancestor). For each dataset we calculated the S&Z statistic, Δ_0 . This statistic is expected to be zero under the hybridization model but different from zero under the lineage sorting hypothesis. To judge whether a simulated dataset rejected the null hypothesis of hybridization, 500 bootstrap replicates of that dataset were used to assess the variance of Δ_0 . To replicate the approach used by S&Z, we then calculated a z-statistic by dividing Δ_0 by its standard deviation, using the bootstrap method (based on the two-tailed critical values of the z-test, $z > 1.96$ indicates that the null hypothesis is rejected at the 5% level, and $z > 2.57$ rejects the null hypothesis at the 1% level). We found that the Δ_0 statistic was significantly different from zero only about 5% of the time whether the data were simulated under hybridization or lineage sorting under a wide range of population sizes. When simulating with very large

TABLE 1. The number of simulations in which Sang and Zhong's Δ_0 statistic was significantly different from 0 (at $\alpha = 0.05$ and 0.01), out of 1,000 simulations of hybridization and lineage sorting under conditions described in the text. Simulations were performed at five different effective population sizes. The data below were simulated by assuming the first speciation event occurred 15 million years ago.

Effective pop. size	Hybridization		Lineage sorting	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
1,000	51	10	53	12
10,000	52	13	57	9
100,000	52	9	55	15
1,000,000	137	60	145	59
10,000,000	493	377	500	394

populations sizes (10^8), the null was rejected in about half of the simulation, but this was true regardless of which model the data were simulated under (see Table 1 for an example). Thus the S&Z test does not reliably distinguish between these two hypotheses under our simulation conditions. For many population sizes, the type II error rate is very high, indicating that the test has no power to reject the null hypothesis of hybridization, and in large enough populations, the genes had significantly different divergence times even when they were being drawn from the same distribution.

Our simulation results contradict those of S&Z. They rejected hybridization only 15% of the time when the data were simulated under hybridization but rejected hybridization at least 80% of the time when the data were simulated under lineage sorting. This discrepancy may be attributed to three sources. First, they did not simulate any variation in coalescence time (instead, they assumed that under lineage sorting, the coalescence time for gene 2 was 1 million years and all other coalescence times were 0). Second, they did not simulate any variation in the number of mutations along the branches of each of the gene trees. The nucleotide substitution process is typically modeled as a Poisson process; mutations are rare and are independent of mutations at other sites. In contrast, S&Z specified the number of changes on each branch, and the only variance in their simulations arose from multiple hits. In essence, they assumed an extremely strict version of a clock model of evolution, in which a mutation was added every million years. This methodology severely underestimates the variability one is likely to observe in real sequences. Finally, their method of bootstrapping to judge significance differed from ours (they briefly describe a bootstrap methodology, but we have been unable to replicate their results). Using their simulation protocol, we have been able to produce Δ_0 values similar to the values they presented, but our bootstrapping results indicate that none of these values of Δ_0 is statistically different from zero (data not shown).

One potential objection to our simulations is that we forced both genes to have the same effective population size. Because the effective population size of a gene is affected by many forces of molecular evolution, different genes might have different effective popula-

tion sizes. For example, genes under balancing selection have larger effective population sizes and should therefore have longer coalescence times. If T_k is found to be greater than T_i , then gene 2 may be under balancing selection, making it predisposed to lineage sorting. If such genes are rare and the effective population size of most genes is small, S&Z's assumption that $T_i = T_k = T_1$ under the hybridization hypothesis might be reasonable. Therefore, under the assumption that most genes coalesce rapidly and a few genes have substantially longer coalescence times, the S&Z test might yield valid results. However, S&Z present no data to support the validity of this assumption, nor do their simulations prove that their test is powerful enough to detect signal in data with realistic amounts of variability.

PROSPECTS FOR INFERRING THE HISTORY OF SPECIES FROM GENE SEQUENCES

No currently available method for inferring species trees or networks uses all of the information in the data. In light of the relationship shown in Figure 2, distinguishing between hybridization and lineage sorting, may not seem feasible. Fortunately, other sources of information are relevant to the problem. For example, although both hypotheses can produce the same gene tree topologies, the proportions of each of the topologies in a large sample of genes should differ between the two hypotheses. Under hybridization, the genes in the hybrid taxon should track one parental species or the other, but if a short internal branch is leading to lineage sorting problems, then three topologies might be expected in almost equal proportions. Even with two genes sampled, branch length information, which is ignored in the S&Z statistic, can help discriminate between hybridization and lineage sorting. For example, the gene trees shown in Figures 3A and 3B have most likely been produced by different processes from those shown in Figures 3E and 3F. The gene trees shown in Figure 3A and 3B appear to be the result of recent hybridization, whereas those in Figures 3E and 3F are easily explained by lineage sorting. The S&Z test ignores all of the branch length information pertaining to the putatively hybrid taxon and relies exclusively on the branch lengths between taxa A,

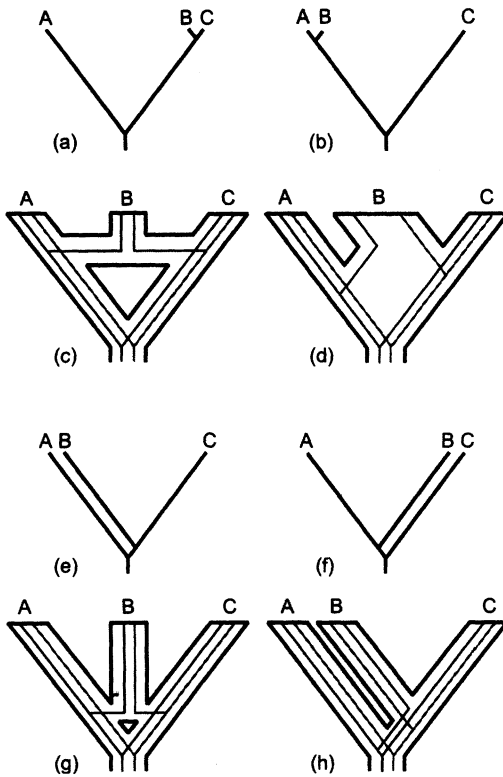


FIGURE 3. Two pairs of conflicting gene trees for three taxa A, B, and C with a depiction of how they can be reconciled with a species history featuring hybridization or lineage sorting. The gene trees in (a) and (b) can easily fit onto a species network under the hybridization hypothesis shown in (c), but their branch lengths are not compatible with the lineage sorting hypothesis shown in (d). The gene trees in (e) and (f) have the same topology as those in (a) and (b), but their branch lengths are easily reconciled with either hybridization (g) or lineage sorting (h).

C, and the outgroup. Taxon B is effectively removed from the gene trees, resulting in very similar trees that are difficult to distinguish. Clearly the examples in Figure 3 are extreme, but they suggest that branch lengths provide important information that should not be ignored.

Ideally, each hypothesis of the history of a group of organisms should be evaluated by considering all of the compatible gene

trees. A method should consider how probable the gene tree is, given the species history, and how probable the observed sequence data are, given each gene tree (see Maddison 1997, for a more thorough discussion). Hypotheses could be evaluated by maximizing their likelihood, or if a Bayesian perspective is adopted, by maximizing their posterior probabilities. Maximization of the likelihood would be computationally intensive because likelihoods would be summed over a huge number of possible gene trees. The Bayesian approach of using Markov chain Monte Carlo (MCMC) methods to approximate the integrals of complex, multidimensional functions may be the best candidate for incorporating all of the information of the likelihood function in a computationally feasible way. An efficient MCMC algorithm has not been fully worked out for the problem of gene trees within a species tree, but algorithms dealing with a related problem, horizontal transfer across parts of a tree, have already been shown to be effective (Huelsenbeck et al., 2000).

Given that maximum likelihood and Bayesian methods are currently unavailable for this problem, it is tempting to try to develop a fast, approximate statistic (such as S&Z's) to identify the cause of gene incongruence. In reality, the underlying problem may be too difficult for such an approach to be generally applicable.

REFERENCES

- HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pages 1–44 in *Oxford surveys in evolutionary biology* (D. Futuyma and J. Antonovics, eds.). Oxford Univ. Press, Oxford.
- HUELSENBECK, J. P., B. RANNALA, AND B. LARGET. 2000. A Bayesian framework for the analysis of cospeciation. *Evolution* 54:352–364.
- MADDISON, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- SANG, T., AND Y. ZHONG. 2000. Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.* 49:422–434.

Received 18 December 2000; accepted 22 February 2001
Associate Editor: R. Olmstead