

- MOORE, W. S. 1995. Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees. *Evolution* 49:718–726.
- NAYLOR, G. J. P., T. M. COLLINS, and W. M. BROWN. 1995. Hydrophobicity and phylogeny. *Nature* 373:565–566.
- PATON, T., O. HADDRATH, and A. J. BAKER. 2002. Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds. *Proc. R. Soc. Lond. B* 269:839–846.
- POE, S., and D. L. SWOFFORD. 1999. Taxon sampling revisited. *Nature* 398:299–300.
- PRAGER, E. M., and A. C. WILSON. 1976. Congruency of phylogenies derived from different proteins. *J. Mol. Evol.* 9:45–57.
- PRAGER, E. M., A. C. WILSON, D. T. OSUGA, and R. E. FEENEY. 1976. Evolution of flightless land birds on southern continents: Transferrin comparison shows monophyletic origin of ratites. *J. Mol. Evol.* 8:283–294.
- SHIMODAIRA, H., and M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16:1114–1116.
- SIBLEY, C. G., and J. E. AHLQUIST. 1990. Phylogeny and classification of birds: A study in molecular evolution. Yale Univ. Press, New Haven, Connecticut.
- STAPEL, S. O., J. A. M. LEUNISSEN, M. VERSTEEG, J. WATTEL, and W. W. DE JONG. 1984. Ratites as oldest offshoot of avian stem—Evidence from α -crystallin A sequences. *Nature* 311:257–259.
- SWOFFORD, D. L. 1999. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer, Sunderland, Massachusetts.
- SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 in *Molecular systematics*, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.
- VAN TUINEN, M., C. G. SIBLEY, and S. B. HEDGES. 2000. The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. *Mol. Biol. Evol.* 17:451–457.
- VOELKER, G., and S. V. EDWARDS. 1998. Can weighting improve bushy trees? Models of cytochrome *b* evolution and the molecular systematics of pipits and wag-tails (Aves: Motacillidae). *Syst. Biol.* 47:589–603.
- WETMORE, A. 1960. A classification for the birds of the world. *Smithson. Misc. Collect.* 139:1–37.
- YANG, Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105–111.

First submitted 28 November 2001; reviews returned
13 March 2002; final acceptance 15 April 2002
Associate Editor: Karl Kjer

Syst. Biol. 51(4):625–637, 2002
DOI: 10.1080/10635150290102302

The Utility of the Incongruence Length Difference Test

F. KEITH BARKER¹ AND FRANÇOIS M. LUTZONI²

¹Department of Ornithology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA; E-mail: fbarker@amnh.org

²Department of Biology, Duke University, Box 90338, Durham, North Carolina 27708, USA

Conditional combination of phylogenetic data requires definition of explicit criteria for combinability (Bull et al., 1993). In this context, combinability refers to the methodological validity of combining multiple sources of phylogenetic data, given the underlying assumptions (explicit or otherwise) of the analysis. Combinability has been evaluated by the effect of data set combination on phylogenetic accuracy: Combinable data sets increase accuracy (Bull et al., 1993; Cunningham, 1997b). When inferential methods are statistically consistent, this convergent property is guaranteed by statistical homogeneity of the data sets to be combined: Increasing sample size increases precision. In a phylogenetic context, data homogeneity can be defined as the sharing of a single history (topological pat-

tern of ancestor–descendant relationships among terminals) and uniform probabilities of change among character states (e.g., branch lengths and relative frequencies of character state transformation). Data sets sampling the same phylogenetic history, but with drastically different evolutionary dynamics, could yield biased estimates when combined and analyzed using a model and parameters with a poor fit to at least one of the partitions. For molecular data, these requirements are explicit in the calculation of conditional probabilities based on the maximum-likelihood criterion, where the overall likelihood is the product of individual site likelihoods, under the assumption that site patterns are independent and identically distributed (Felsenstein, 1981). However, likelihood methods allow this

requirement to be relaxed in various ways, such as by allowing sites to vary in rate (Yang, 1993) or relative probabilities of character-state transformations (Yang, 1996). Given that homogeneity of these transformation probability parameters can be relaxed, the most basic requirement of combinability is topological congruence (e.g., Mason-Gamer and Kellogg, 1996).

Whereas tests of congruence (Huelsenbeck and Bull, 1996) and process homogeneity (e.g., Sullivan, 1996; Yang, 1996) are self-evident (if computationally demanding) within a maximum-likelihood framework, the same has not been true for parsimony. This lack, in conjunction with debate regarding the effects of combining data with differing evolutionary rates in parsimony analysis, has contributed to significant controversy over how data homogeneity, topological congruence, and combinability should be assessed and interpreted (Bull et al., 1993; Kluge and Wolf, 1993; Chippindale and Wiens, 1994; Lutzoni and Vilgalys, 1995; Brower et al., 1996; Mason-Gamer and Kellogg, 1996; Nixon and Carpenter, 1996; Cunningham, 1997a, 1997b; DeSalle and Brower, 1997; Lutzoni, 1997). One of the procedures that has been developed in a parsimony context is the incongruence length difference (ILD) test (Farris et al., 1995a, 1995b). The test is based on the ILD index of Mücke and Johnson (1976), which measures the proportion of inferred homoplasy attributable to the combination of individual data sets or partitions, which may each require conflicting minimal-length topologies. The index can be defined as $(i_T - i_W)/i_T$, where i_T is the total number of homoplastic character changes required under parsimony on the shortest tree for two or more data sets analyzed simultaneously, and i_W is the sum of homoplastic changes required for each data set on its own minimum length tree (or trees). The ILD test compares the value of this index with a null distribution generated by random permutation of characters among partitions (in practice, only the sum of the tree lengths from separate analyses is calculated and compared with its permuted null distribution). The ILD test was intended to detect the presence of strongly supported character conflict ("hard" incongruence) among individual

data sets within a combined analysis (Farris et al., 1995a, 1995b). However, the test has gained wide usage in parsimony analyses both as a test of topological congruence (e.g., testing the clonality of fungi; Koufopanou et al., 1997; Geiser et al., 1998; Carbone et al., 1999) and more generally as a test of combinability (Cunningham, 1997a, 1997b; Swofford, 1998).

Although the test is widely used, a number of authors have noted peculiarities in its behavior that have called into question its validity as a criterion for congruence and combinability. In one of the few studies of the effect of combining data sets with varying significance with the ILD test, Cunningham (1997a) concluded that the ILD test performed best in predicting when data should be combined, compared with the tests of Templeton (1983) and Rodrigo et al. (1993). This conclusion was based on an analysis of the effects of adding together individual partitions in estimation of a phylogenetic hypothesis strongly supported by all the available data (proxy for a "known" phylogeny). However, Cunningham (1997a) suggested that a critical α value of somewhere between 0.01 and 0.001 was a more appropriate criterion for rejection of combinability than the generally accepted 0.05 level, suggesting an excessive type I error rate for the ILD test as a measure of combinability (see also Sullivan, 1996). Graham et al. (1998) obtained significant ILD values when testing for incongruence between sequence data from the chloroplast genome and morphological data in the angiosperm family Pontederiaceae but interpreted this conflict as the result of high levels of homoplasy in the morphological data. To support this contention, they performed the ILD test using their molecular data and random data sets with four equiprobable states, generated using the "Fill random" option in MacClade (Maddison and Maddison, 1999). Despite low structure retention in the 50% bootstrap majority rule consensus for these "random" data sets, all 20 of their replicates were incongruent with the molecular data at $\alpha \leq 0.01$. These results suggested that the ILD test might have an excessively high type I error rate as a test of congruence. Specifically, they indicated that this effect might be caused by disparity in levels of homoplasy among data sets. These results and others (e.g., Cunningham, 1997b; Stanger-Hall and

Cunningham, 1998; Yoder et al., 2001) have suggested that the test might be biased or inaccurate as a measure of both combinability and congruence.

Recently, Dolphin et al. (2000) performed a series of data set manipulations that showed conclusively that significance values of the ILD test are related to disparity in levels of homoplasy between two or more data sets. In their investigation, they permuted character states among taxa for increasingly large proportions of perfectly consistent binary data sets. These perturbed data sets were evaluated for congruence with unmanipulated data using the ILD procedure. As the proportion of permuted characters increased, the significance of the ILD likewise increased, although no well-supported structure should have been retained in the permuted data. Dolphin et al. concluded that differing levels of homoplasy between the two data sets per se caused the significant result. Their figure 3 shows the underestimation of homoplasy characteristic of the parsimony method (Archie and Felsenstein, 1993), which underlies the significant ILD values. Darlu and Lecointre (2002) generalized this result to molecular data simulated under a variety of evolutionary conditions. They found significant conflict between data sets evolved on a single tree but with contrasting lineage-specific rates of evolution and patterns of among-site rate variation (simulated using a Γ distribution of rates; Yang, 1993).

These results suggest that the ILD test is not a valid measure of "hard" (well supported) incongruence. However, it remains to be seen whether the ILD test is an appropriate measure of data set *combinability*. The ILD test could be a good measure of combinability without being an appropriate test of congruence (a necessary but not sufficient condition of combinability). Specifically, the test may combine phylogenetic congruence and uniformity of character transformation probabilities inextricably, as in a test of homogeneity (regarding which, nota bene the current naming of the ILD implementation in PAUP* is the partition homogeneity test; Swofford, 1998). To evaluate the utility of the ILD test, it must be examined not only with regard to the evolutionary conditions that yield significance but with an exploration of the consequences of data set combination as a function of the test's significance.

Here, we explore the interrelated properties of congruence, homogeneity, and combinability in the context of the ILD test. We explain briefly the underlying statistical difficulty with the ILD test as a measure of congruence and discuss its potential utility as a test of homogeneity among data partitions. The ILD test appears to be an inappropriate measure of congruence and homogeneity under reasonable simulated conditions of molecular evolution. We also assessed the utility of the ILD test as a criterion for data set combinability, as estimated by its predictive value with regard to the effect of data set combination on phylogenetic accuracy.

METHODS

DNA Sequence Data Simulations

Pairs of identical-size data sets for evaluation via the ILD test were generated stochastically according to established models of DNA sequence evolution. Contrasts between data sets in a number of factors (e.g., sample size, average substitution rates, substitution models, and lineage-specific and site-specific rate heterogeneity) are common in molecular data (e.g., Reed and Sperling, 1999; Wilgenbusch and de Quieroz, 2000), and some of these factors are known to affect significance of the ILD (Darlu and Lecointre, 2002). In this study, the only difference between pairs of data sets tested was in evolutionary rate. However, each rate comparison was repeated under a number of evolutionary models (Table 1). All data sets were generated using Seq-Gen 1.1 (Rambaut and Grassly, 1997), which allows a multiplier to be applied to all branches of an input phylogeny (option -s). All unique pairwise comparisons were made between data sets with multipliers of 1 (the base model tree and branch lengths), 5, 10, and 50 (a total of 10 rate comparisons, i.e., 1:1, 1:5, ..., 50:50). These 10 unique rate comparisons were repeated under a variety of simulated conditions of DNA sequence evolution, determined by three main axes: (1) tree shape, (2) base frequency skewness, and (3) transition/transversion bias (see Table 1). Two symmetric (perfectly balanced) base model trees were chosen for testing, one with equal branch lengths set at 0.077 (in expected number of changes per site) and the second with the five internal branches set at 0.012

TABLE 1. Conditions of DNA sequence simulations. Conditions indicated by each row were implemented with the base rates (unscaled branch lengths indicated under Tree shape) and with branch lengths scaled by multipliers of 5, 10, and 50. Within each model, 100 replicates (with 1,000 characters evolved at each rate) of all unique pairwise rate comparisons (1:1, 1:5, . . . , 50:50) were evaluated via the ILD procedure (Farris et al., 1995a, 1995b), yielding "error" estimates for a total of 80 simulated conditions.

Tree shape ^a	Base frequencies	Transition/transversion ratio	Model ^b
Even	A = C = G = T (even)	0.5 (unbiased)	JC69
	A = C = G = T	5.0 (biased)	K2P
	A = T = 0.125, C = G = 0.375 (skewed)	0.5	F81
	A = T = 0.125, C = G = 0.375	5.0	HKY85
Short internal	A = C = G = T	0.5	JC69
	A = C = G = T	5.0	K2P
	A = T = 0.125, C = G = 0.375	0.5	F81
	A = T = 0.125, C = G = 0.375	5.0	HKY85

^aEven = (1: 0, (2: 0.076923, ((3: 0.076923, 4: 0.076923): 0.076923, ((5: 0.076923, 6: 0.076923): 0.076923, (7: 0.076923, 8: 0.076923): 0.076923): 0.076923): 0.076923); short internal = (1: 0, (2: 0.117647, ((3: 0.117647, 4: 0.117647): 0.117647, ((5: 0.117647, 6: 0.117647): 0.117647, (7: 0.117647, 8: 0.117647): 0.117647): 0.117647): 0.117647).

^bJC69—Jukes and Cantor, 1969; K2P—Kimura, 1980; F81—Felsenstein, 1981; HKY85—Hasegawa et al., 1985.

and the eight external branches set at 0.118 (see Table 1). The total tree length in both cases was 1.000 (values rounded). Branch lengths generated with the base rate and with multipliers of 5 and 10 represent reasonable levels of comparison frequently encountered in problems of phylogeny estimation using DNA sequence data. The multiplier of 10 yielded data with phylogenetic signal significantly degraded by multiple substitutions (pers. obs.), and a multiplier of 50 yielded data sets that were nearly randomized. The base model of sequence evolution used was that of Jukes and Cantor (1969; JC69), which has a single rate for all nucleotide substitutions and equal representation of all four bases. Two evolutionary parameters were varied from this base model to assess their impact on significance values of the ILD test. The first of these was base frequency skewness, which was imposed by setting base frequencies for G and C to 37.5% and those for A and T to 12.5% (corresponding to the model of Felsenstein, 1981; F81). The second factor was the proportion of transitions to transversions, which was set to 5 to mimic the observed skewness in some data sets (e.g., mitochondrial DNA; corresponding to the model of Kimura, 1980; K2P). These departures were also imposed simultaneously (the model of Hasegawa et al., 1985; HKY85). All 10 rate comparisons were made for each of the four substitution models (JC69, K2P, F81, and HKY85) using both of the model trees (even and short internal), yielding a total of 80 comparisons.

Statistical Evaluation of the ILD

For each of the 80 simulated comparisons, 100 replicate data sets of 1,000 characters per partition were generated (one partition for each of the two rates being compared under a given substitution model and tree). Each of these replicate data sets was analyzed using the ILD procedure as implemented in PAUP* 4.0b8 (Swofford, 1998) using the branch-and-bound search algorithm with 100 permutation replicates to generate the null distribution. The fraction of ILD null replicates greater than the initial value (the "significance" value of the ILD) was recorded for each simulation replicate.

A more extensive analysis of the data sets simulated under the HKY85 model of sequence evolution was conducted. In addition to the ILD significance values, the inferred most-parsimonious tree (or trees, found via the branch-and-bound algorithm) for each of the two partitions separately and the two partitions analyzed simultaneously were recorded. Congruence between these trees and the generating tree was quantified by the normalized consensus fork index (nCFI; Colless, 1980), which has its maximum at complete congruence with the generating tree and its minimum when the inferred tree shares no nodes with the generating tree. Changes in the accuracy of phylogenetic estimation with data combination (Δ nCFI) were estimated by subtracting the nCFI of the low-rate data set (which invariably provided a better estimate of phylogeny under the simulation conditions used here) from the nCFI of

the combined analysis (for single-rate comparisons, the choice of single data set nCFI was arbitrary). Thus, if data combination increased accuracy over the best single data set in terms of the number of correctly inferred nodes, Δ nCFI was positive, and vice versa. Significance of the ILD (using ln-transformed P values; Cunningham, 1997b) was evaluated for its value as a predictor of Δ nCFI via least-squares regression (StatView 5.0.1, SAS Institute). The effect of combination was also quantified discretely as negative (Δ nCFI < 0) versus neutral/positive (Δ nCFI \geq 0), and the significance of the ILD for these cases ($\alpha \leq 0.05$) was noted. These data were subjected to a χ^2 contingency table analysis to determine whether significance of the ILD successfully predicted negative effects of data set combination.

RESULTS

Significance Values of the ILD as a Function of Model Conditions and Rate Comparison

The results of ILD evaluations of simulated DNA sequence data are summarized in Figure 1. Comparisons of data sets evolved at identical rates yielded very few significant values, even for extremely high rates (data sets evolved with a multiplier of 50 were essentially randomized; uncorrected p distances among sequences were ≈ 0.75 , the random expectation with even base frequencies). Comparisons of data sets with contrasting rates of nucleotide substitution, especially the 1:10, 1:50, 5:10, 5:50, and 10:50 comparisons, demonstrated significance values for the ILD test markedly in excess of 0.05. Rate comparisons based on the model tree with equal branch lengths showed rather abrupt transitions between failure to detect significant differences and complete rejection of the null hypothesis (e.g., rate proportions of 1:10 versus 1:50), especially for comparisons on this tree that included a transition/transversion bias (K2P and HKY85 models).

Increases in external branch lengths at the cost of decreasing internal branch lengths yielded reduced significance levels of the ILD test, in cases where the test yielded significant values with equal branch lengths (Figs. 1A, 1B). However, where significance values were low with equal branch lengths (1:1, 1:5, 5:5, 10:10), the percentage of sig-

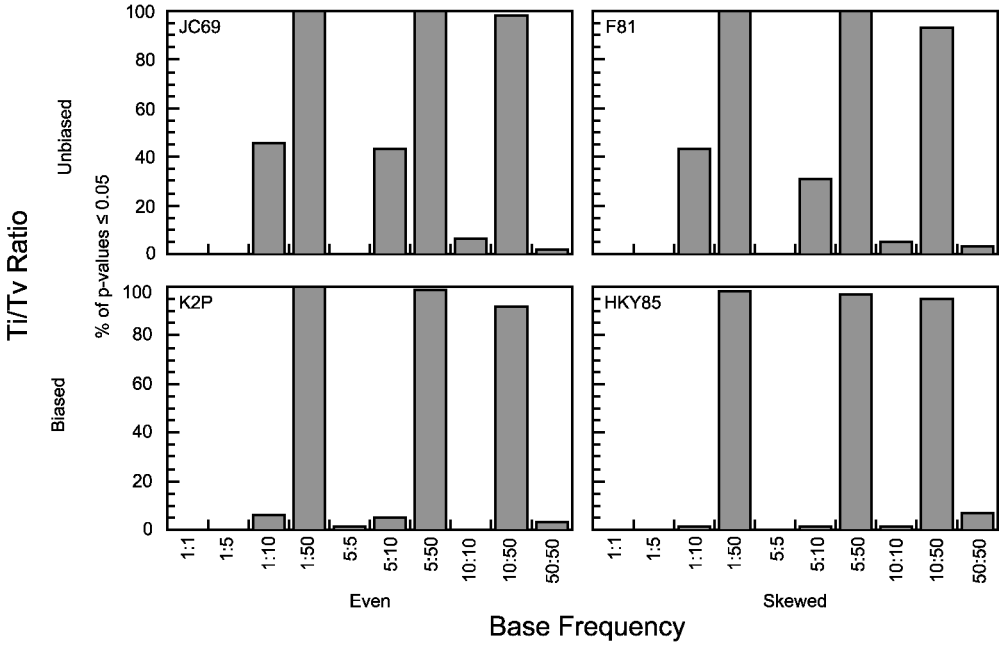
nificant comparisons tended to increase on the unequal branch length tree. Imposition of a transition/transversion bias generally reduced the significance levels of the ILD. When combined with skewed branch lengths, imposition of a transition bias had a reduced effect, although in general it still decreased the proportion of significant values. Overall, skewed base frequency had little effect on observed significance values, the most noticeable being the comparison for the unequal branch length tree with transition/transversion bias, where addition of base frequency skewness (K2P versus HKY85) appeared to have a slight reducing effect on significance levels.

ILD as a Predictor of Phylogenetic Accuracy with Data Combination

Somewhat surprisingly given the highly significant ILD values found for many of these simulated data sets (Fig. 1), phylogenetic accuracy of individual data sets and data sets in combination was extremely high (generally $\geq 90\%$ of replicates recovered the generating tree, except rate 50 data and extreme rate comparisons under the HKY85 model). Levels of accuracy for most models of sequence evolution (JC69, F81, and K2P) were high enough that there was little variation available for analysis of the ILD test as a predictor of accuracy. For this reason, we focused on analysis of the HKY85 model data (Fig. 2). Even for these data, levels of accuracy were very high for the even-branch-length trees, except for data sets including only characters evolving at the base rates with a multiplier of 50 (Fig. 2A). With 2,000 characters, even the rate 50 data yielded the correct tree in 2 of 100 replicates, indicating that not quite all phylogenetic signal was eliminated.

Phylogenetic accuracy of individual and combined data sets was severely compromised for characters evolved on the short-internal-branches tree (Fig. 2B). Even the base rate data failed to recover the correct tree in 14.5% of the replicates with 1,000 characters, although doubling the data set size increased accuracy to 100% (see 1:1 combined data set, Fig. 2B). The generating tree was never recovered from rate 50 data with this tree shape. In general, increasing substitution rates decreased accuracy for individual data sets and for

A) All branches equal length



B) Internal branches short

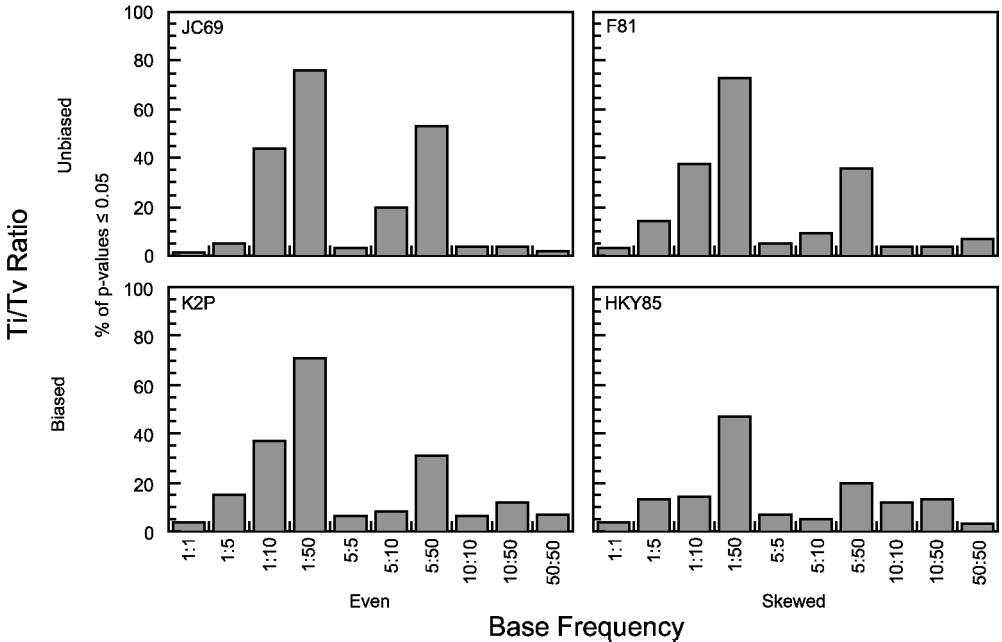
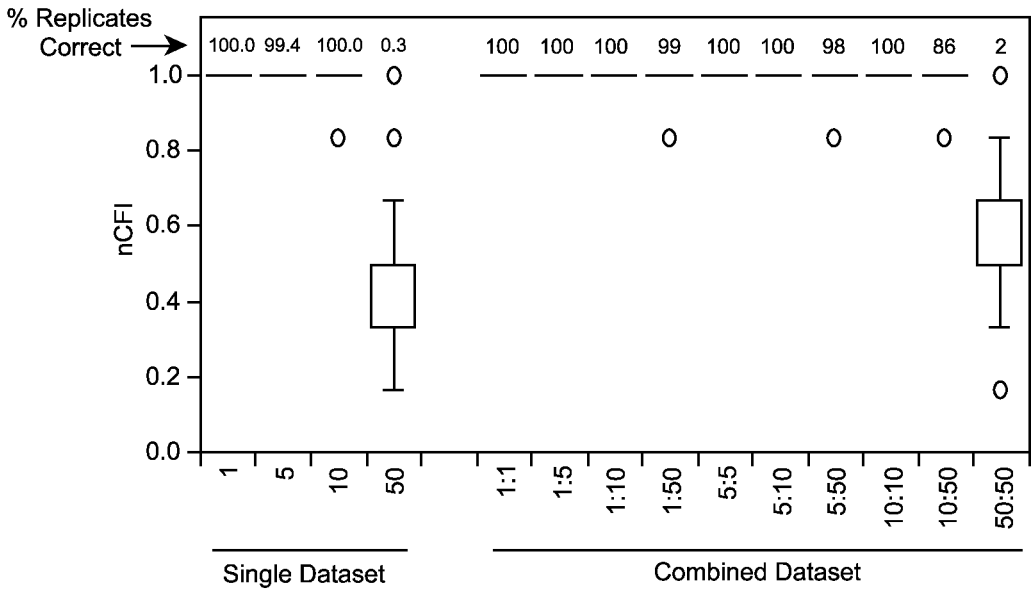


FIGURE 1. Percentage of replicate simulated DNA sequence data set comparisons for which the Ti/Tv Ratio test returned $P \leq 0.05$ (significance). (A) Results of simulations on trees with equal branch lengths. (B) Results of simulations on trees with short internal branches. See Table 1 for parameters used in each simulation. The ratios under each bar indicate the rate comparisons being reported (e.g., 5:10 indicates data sets simulated on the base tree with branch length multipliers of 5 and 10, respectively, were being prepared).

A) All branches equal length



B) Internal branches short

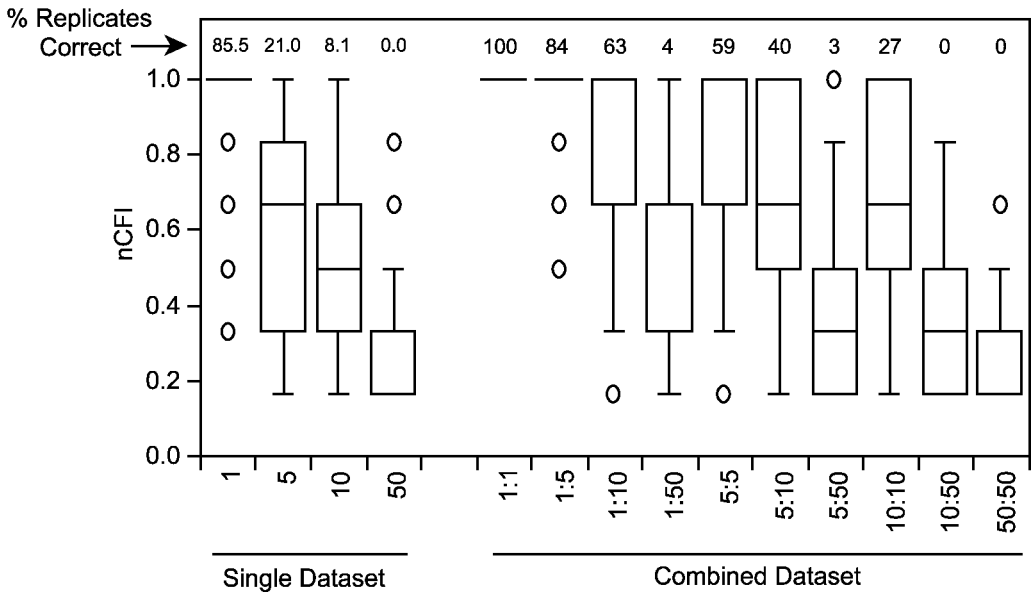


FIGURE 2. Phylogenetic accuracy of data sets evolved under the HKY85 model of sequence evolution with (A) equal branch lengths or (B) short internal branches (see Table 1). Boxplots indicate the distribution of accuracy of parsimony trees compared with the tree used to generate the data sets (as measured by the normalized consensus fork index, nCFI; Colless, 1980). The center horizontal line of each box indicates the median accuracy of a given data set composition, the lower and upper boundaries of each box indicate the 25th and 75th percentiles of data set accuracy, the whiskers outside each box extend to the 10th and 90th percentiles, and individual values in the lowest and highest 10% of the distribution are plotted as circles (one or more of these features may coincide if the corresponding percentiles overlap; e.g., data sets with 100% accuracy are represented by a single horizontal line). Accuracy of single data sets (1,000 bases, left) was estimated from 1000 replicate data sets, and that of combined data sets (2,000 bases, right) was estimated from 100 replicates. The values at the top of the box plots indicate the percentage of these replicates that recovered the tree used to generate the data sets.

combined single-rate data sets. Combining high-rate data sets with low-rate data sets generally decreased phylogenetic accuracy relative to averages for the lower rate data alone. The only exceptions to this trend were the combination of rate 1 and rate 5 data, which only slightly decreased accuracy of inference over rate 1 data alone, and the combination of rate 5 and rate 10 data, which significantly improved accuracy over rate 5 data (Fig. 2B).

Variance around this general pattern of reduced accuracy with combination of high- and low-rate data was examined for evidence of the predictive value of the ILD. Specifically, we asked whether significance of the ILD was a good predictor of the effect of combining two data sets on the accuracy of the combined estimate. We measured this effect relative to the better of the two separate estimates (invariably the lower rate data set under the conditions simulated here). For the HKY85 model of evolution, on the short-internal-branch tree, the ILD P value was a significant predictor of the effect of data combination on relative accuracy, as evaluated by simple regression (Fig. 3). Thus, decreasing significance for the ILD (higher ILD P values) is related to overall increases in phylogenetic

accuracy of the combined data estimate. We also examined this trend qualitatively using contingency table analysis. This analysis indicated that data set combination for replicates with significant ILD values resulted in reduced accuracy at a much higher frequency than did combination for replicates with nonsignificant values across a range of significance levels (Table 2). Although ILD significance was a significant predictor of the effect of data set combination on accuracy, the amount of variation in this effect explained by the ILD P value was extremely small (coefficient of determination, $r^2 = 0.11$).

Accuracy	ILD significance					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
	S	NS	S	NS	S	NS
Decreased	29	275	75	229	114	190
Increased	13	683	63	633	106	590
χ^2	30.9		43.4		61.1	

accuracy of the combined data estimate. We also examined this trend qualitatively using contingency table analysis. This analysis indicated that data set combination for replicates with significant ILD values resulted in reduced accuracy at a much higher frequency than did combination for replicates with nonsignificant values across a range of significance levels (Table 2). Although ILD significance was a significant predictor of the effect of data set combination on accuracy, the amount of variation in this effect explained by the ILD P value was extremely small (coefficient of determination, $r^2 = 0.11$).

DISCUSSION

ILD Significance as an Indicator of Topological Congruence

In agreement with previous results (Cunningham, 1997a; Graham et al., 1998; Dolphin et al., 2000; Yoder et al., 2001; Darlu and Lecointre, 2002), the simulations presented here strongly support the contention that the ILD procedure is, under certain conditions, biased as a test of congruence, that is, in terms of shared phylogenetic history. The proportion of individual ILD significance values $\leq \alpha$ ($\alpha = 0.05$) in our simulations indicates the type I error rate of the ILD as a test of topological congruence at that value of α (the probability of rejecting congruence given that congruence is true, which is the case because the data sets were generated from the same tree). In most cases simulated here, this proportion far exceeded the target value of

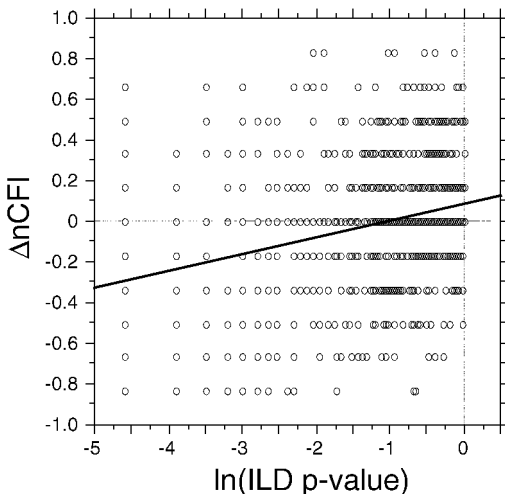


FIGURE 3. Least-squares regression of the effect of data set combination on phylogenetic accuracy ($\Delta nCFI$) as predicted by significance levels of the ILD (\ln -transformed ILD P values; includes data generated using the HKY85 model with the short-internal-branch-length tree). The equation of the regression line is $\Delta nCFI = 0.081 + 0.082[\ln(ILD P \text{ value})]$; $r^2 = 0.11$, $P_{\text{regression}} < 0.01$.

0.05. All of the factors varied, including transition/transversion ratio bias, base composition, and internal:external branch length proportions, appear to have their main effects by influencing the amount of structure in the more structured data set. The higher the contrast in degree of structure between the two data sets, the stronger the effect on significance levels of the ILD. Imposing a transition bias or base composition skewness or reducing the length of the internal branches on the model tree have the effect of reducing the level of contrast in strength of phylogenetic signal (as quantified by the consistency and retention indices) between individual data sets, especially in the most extreme cases (e.g., rate < 50 versus rate 50 comparisons).

The results presented here expand upon previous conclusions regarding the behavior of the ILD drawn from studies of randomized data. Dolphin et al. (2000) previously argued that significance of the ILD in comparisons of randomized and structured data was due to the nonlinear relationship between increasing homoplasy levels and parsimony estimates of tree length. This consistent underestimation of character change inherent to parsimony procedures is well documented and has fueled continuing debates over the appropriateness of various measures of homoplasy and phylogenetic signal (reviewed by Archie, 1996). Specifically, Archie and Felsenstein (1993) noted that the length of shortest trees for random data is usually substantially lower than that of random trees. In the context of the ILD, combined analysis of random and structured data will result in higher estimates of character change for the randomized characters than would be obtained on minimum-length trees generated using those characters alone. If the structured characters dominate in producing trees for the null replicates of the test, the mode of the null distribution will be shifted up a number of steps depending on the degree to which parsimony underestimates amounts of character change for the randomized characters alone. Consequently, comparison of the initial summed tree lengths (with changes in the randomized data significantly underestimated in the separate analysis) with the null will yield a conclusion of significance. The current results (as well as other simulation data; Darlu and Lecomte, 2002) indicate that this effect can be significant for data

that contain phylogenetic information (non-randomized data) but that exhibit varying levels of homoplasy and structure because of varying rates and patterns of evolution. Data that share a single history can, because of differences in evolutionary dynamics, exhibit significant incongruence as measured by the ILD test.

This difficulty with the application of the ILD test as a criterion of topological congruence (a measure of shared phylogenetic history) could prompt at least two responses, that is, retention of the ILD as a criterion under certain conditions or in some modified form, or rejection of the ILD as a criterion (and possibly its resurrection in some other role). Regarding the former option, delineation of conditions under which the test might be biased offers one potential remedy. A full definition of the parameter space within which the ILD test might be a statistically valid test of congruence is beyond the scope of this study. A number of factors other than evolutionary rates and patterns may affect the performance of the test, such as resolving power, sample size, and the number of character states available (Lutzoni, 1997). We have performed preliminary tests of the effects of resolving power (testing data sets against jackknifed subsets) and sample size (using independent data sets of differing sizes). Neither factor per se appears to be significant; however, both should have an impact to the degree that they affect the probability of recovering the generating tree in null model replicates when disparity in levels of homoplasy exists (predicting increased bias of the test in comparisons of large, relatively structured data sets with small, relatively unstructured data sets). Additionally, disparity in the number of available character states between data sets evolving at similar rates will result in different levels of homoplasy at sufficiently high rates of change.

Regarding the conditions tested in our study, examination of homoplasy indices for the data sets used in these simulations provides some useful information. Figure 4 summarizes consistency index (CI) and retention index (RI) values for the simulated DNA data sets. CI values are conspicuously high, even for essentially randomized data (rate 50), because of the small number of taxa in each data set. Graphically, the RI values appear more useful in discriminating among

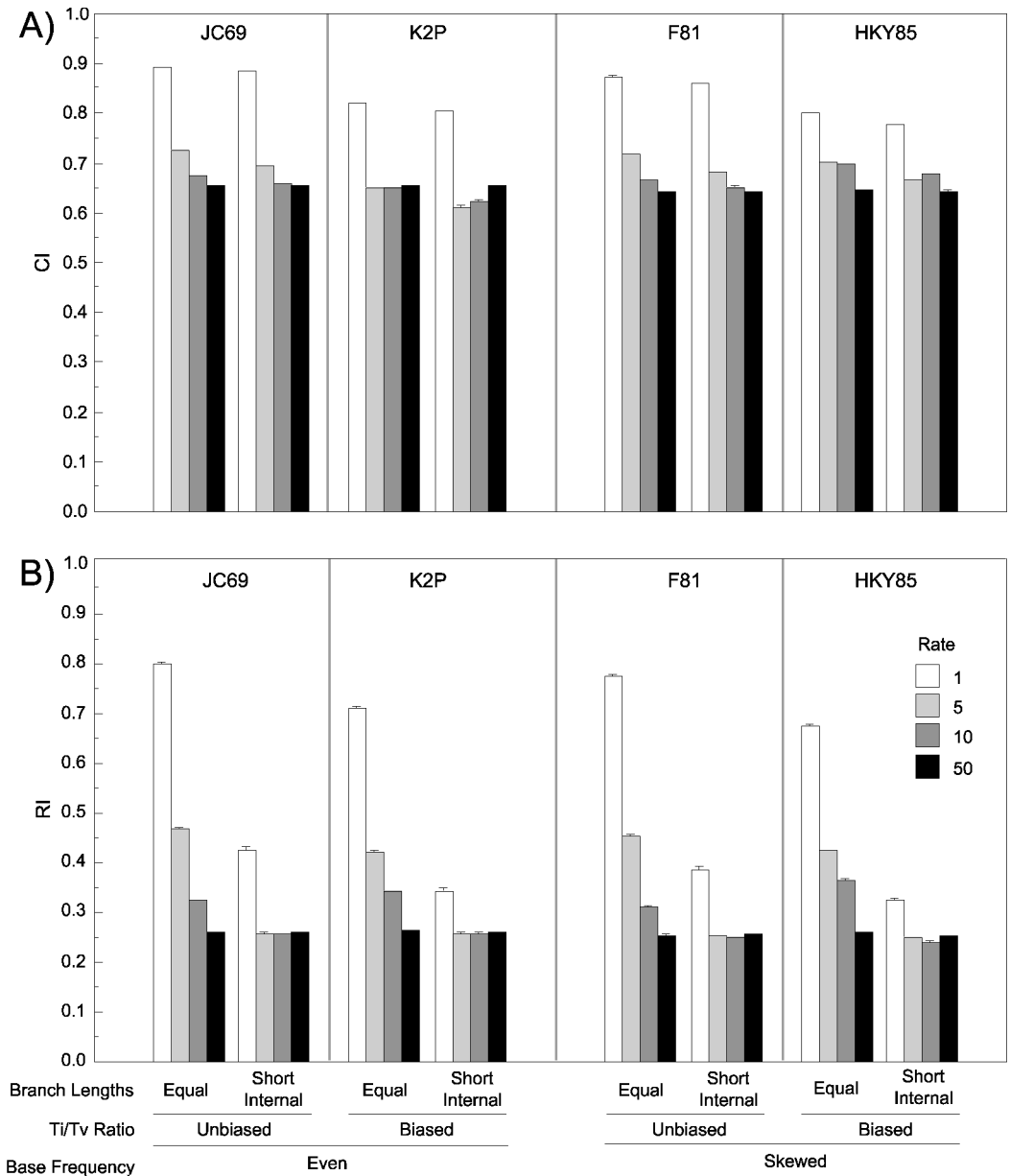


FIGURE 4. (A) Ensemble consistency index (CI) and (B) retention index (RI) for the simulated DNA sequence data sets. Error bars indicate the 95% confidence interval of each value. Reported values are for individual simulated data sets from associated shortest trees as found by the branch-and-bound search algorithm of PAUP* (Swofford, 1998).

data sets, although the overall pattern is essentially the same as that displayed by the CI. For those cases where the ILD test absolutely rejects the null hypothesis (all comparisons with rate 50), the RI of 0.25 indicates the essentially random nature of the

high-rate data set. Basically, comparison of RI = 0.25 DNA data with any more structured data (RI > 0.25) yielded significant ILD values.

In contrast, when the model tree is less easily estimated (i.e., short internal branch

lengths), RI values for the lower rate data (specifically 5 and 10) drop to levels similar to those for the rate 50 data (RI = 0.25, Fig. 4B), and the ILD test yields less significant values. An intermediate case is that of the 5:10 rate comparison under a JC69 model (Fig. 4B). In this case, RI values are approximately 0.45 for the rate 5 data and 0.30 for the rate 10 data; comparison of these data sets indicates significant (40% type I error rate, Fig. 1A) bias in the ILD on the null hypothesis of congruence. Although there is a general trend in contrasting RI values being associated with significant values of the ILD, even very small contrasts in RI may still be associated with significance. For example, the 5:50 rate comparison under the K2P model with unequal branch lengths yields significant values of the test in 30% (Fig. 1B) of the simulated replicates, but examination of data set RI values indicates that they are essentially identical (RI \approx 0.25, Fig. 4B). For this reason, it may be more appropriate to seek alternatives to the ILD as a criterion of congruence or to create modifications of the ILD that would make it a valid criterion of congruence.

ILD as an Indicator of Homogeneity

Although one option for dealing with the behavior of the ILD test would be to abandon it as a criterion of topological congruence (i.e., shared phylogenetic history), this begs the question of what exactly the ILD is measuring. One candidate interpretation of the ILD test is as a measure of homogeneity among data partitions. With this interpretation, the proportion of ILD significance values $\leq \alpha$ in our simulations is an indication of the test's statistical power to detect heterogeneity (β ; probability of rejecting homogeneity given that homogeneity is false). Under this interpretation, the ILD test seems to fare poorly. In Figure 1B, for the HKY85 model of evolution with the short-internal-branches tree and the most extreme rate comparison (1:50), only \sim 50% of ILD replicates reject homogeneity at the $\alpha = 0.05$ level.

To place this value in an appropriate context, we used a maximum-likelihood approach (Yang, 1996) to detect among-partition rate heterogeneity in a combined data set generated under the same model of evolution (HKY85 with the short-internal-branch tree) but with the smallest relative dif-

ference in rates (5:10) and the highest ILD P value of all replicates generated under this model ($P = 1.00$). A likelihood ratio test comparing the fit of a single-rate model with that of the two-rate model to these data under the HKY85 model of evolution (using the generating tree) was highly significant ($-2 \ln \ell = 268.96$, $df = 1$, $P < 0.001$; calculated using PAML 3.0c; Yang, 2000). Even under the simplest model of DNA substitution (JC69; a poor fit to these data), this rate difference was easily detectable ($-2 \ln \ell = 107.97$, $df = 1$, $P < 0.001$). Thus, in a case with the smallest contrast in rates between partitions simulated here and the largest number of factors that might obscure this contrast (base composition skewness and transition bias), the maximum-likelihood method was easily able to detect the difference. The ILD test consistently indicated heterogeneity in only the cases of greatest contrast (e.g., comparisons with rate 50 data). Thus, if the ILD test were a measure of rate homogeneity, it is an extremely inefficient one relative to other methods currently available for analysis of molecular data. In addition, the results of Darlu and Lecointre (2002) indicate that the test has little power to detect other types of heterogeneity, such as differences in lineage-specific and site-specific rate heterogeneity.

ILD as a Criterion of Combinability

Although data set homogeneity guarantees increasing phylogenetic accuracy with data set combination when analytical methods are statistically consistent, combining heterogeneous data can also increase accuracy, even if the analysis does not explicitly incorporate that heterogeneity. For example, Figure 2B indicates that the combination of rate 5 and rate 10 data significantly increased the average accuracy of phylogenetic estimation using parsimony. Others have argued that varying levels of homoplasy in different data sets might contribute to an overall robust signal (Barrett et al., 1991; Nixon and Carpenter, 1996; Vidal and Lecointre, 1998; Wenzel and Siddall, 1999). Although it may be advantageous to combine heterogeneous data, ideally some criterion should be used to indicate whether or not data combination is desirable in individual cases.

To evaluate the ILD as a criterion for combinability, we analyzed changes in

phylogenetic accuracy accompanying data set combination as a function of ILD P values. Increasing ILD P values (i.e., decreasing significance levels) were correlated with improvements in phylogenetic accuracy with data set combination (Table 2). However, the relationship was extremely weak (Fig. 3), and the amount of variance in improvement explained was generally small ($\sim 10\%$). We also examined this question from a less stringent point of view, asking whether or not significant ILD values were more likely to be associated with decreases in accuracy (and conversely whether nonsignificant ILD values were generally associated with, at worst, neutral effects of data set combination). Our χ^2 analysis of the HKY85 data on the short-internal-branches tree indicate that this trend exists and is significant (Table 2). However, nearly 30% of combined data sets with nonsignificant ILD values still had reduced accuracy relative to the better of the two separate analyses, and nearly half of the combined data sets with significant ILD values showed increased accuracy. In sum, the ILD appears to be a relatively poor indicator of data set combinability with the criterion of phylogenetic accuracy and should not be used for this purpose even when using low critical α values between 0.01 and 0.001 (see Sullivan, 1996; Cunningham, 1997a).

CONCLUSIONS

We have briefly reviewed the three related concepts of topological congruence, homogeneity among data partitions, and combinability specifically with regard to the utility of the ILD test in decisions regarding phylogenetic data analysis. Our simulation study supports previous studies in rejecting the ILD test as a unbiased measure of phylogenetic congruence (Graham et al., 1998; Dolphin et al., 2000; Darlu and Lecointre, 2002). The observed bias occurs under a biologically realistic range of parameters and cannot be easily predicted from observed levels of homoplasy. Our results further indicate that the ILD test has relatively little statistical power to detect substitution rate heterogeneity, especially relative to available alternative methods. Although significance values of the ILD broadly predict the effect of data set combination on phylogenetic accuracy, a great deal of variation

in this effect is left unexplained and decisions regarding data combination based on the ILD would be misleading in a large proportion of cases. Beyond the realm of combinability testing per se, the ILD has been used as a criterion for model choice in combined data analysis (e.g., Giribet et al., 2001), but recent results suggest that even this use may be problematic (Dowton and Austin, 2002). The precise utility and appropriate uses of the ILD test remain to be established.

ACKNOWLEDGMENTS

For comments on various versions of the manuscript, we thank Cliff Cunningham, Chris Simon, and two anonymous reviewers. This research was partially supported by a grant from the National Science Foundation, USA, Systematic Biology (DEB-9615542) to F.M.L.

REFERENCES

- ARCHIE, J. W. 1996. Measures of homoplasy. Pages 153–188 in *Homoplasy: The recurrence of similarity in evolution* (M. J. Sanderson and L. Hufford, eds.). Academic Press, San Diego, CA.
- ARCHIE, J. W., AND J. FELSENSTEIN. 1993. The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theor. Popul. Biol.* 43:52–79.
- BARRETT, M., M. J. DONOGHUE, AND E. SOBER. 1991. Against consensus. *Syst. Zool.* 40:486–493.
- BROWER, A. V. Z., R. DESALLE, AND A. VOGLER. 1996. Gene trees, species trees, and systematics: A cladistic perspective. *Annu. Rev. Ecol. Syst.* 27:423–450.
- BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD, AND P. J. WADDELL. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* 42:384–397.
- CARBONE, I., J. B. ANDERSON, AND L. M. KOHN. 1999. Patterns of descent in clonal lineages and their multilocus fingerprints are resolved with combined gene genealogies. *Evolution* 53:11–21.
- CHIPPINDALE, P. T., AND J. J. WIENS. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol.* 43:278–287.
- COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. *Syst. Zool.* 29:288–299.
- CUNNINGHAM, C. W. 1997a. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* 14:733–740.
- CUNNINGHAM, C. W. 1997b. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst. Biol.* 46:464–478.
- DARLU, P., AND G. LECOINTRE. 2002. When does the incongruence length difference test fail? *Mol. Biol. Evol.* 19:432–437.

- DESALLE, R., AND A. V. Z. BROWER. 1997. Process partitions, congruence, and the independence of characters: Inferring relationships among closely related Hawaiian *Drosophila* from multiple gene regions. *Syst. Biol.* 46:751-764.
- DOLPHIN, K., R. BELSHAW, C. D. L. ORME, AND D. L. J. QUICKE. 2000. Noise and incongruence: Interpreting results of the incongruence length difference test. *Mol. Phylogenet. Evol.* 17:401-406.
- DOWTON, M., AND A. D. AUSTIN. 2002. Increased incongruence does not necessarily indicate increased phylogenetic accuracy—The behavior of the ILD test in mixed-model analyses. *Syst. Biol.* 51:19-31.
- FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE, AND C. BULT. 1995a. Constructing a significance test for incongruence. *Syst. Biol.* 44:570-572.
- FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE, AND C. BULT. 1995b. Testing significance of incongruence. *Cladistics* 10:315-319.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376.
- GEISER, D. M., J. I. PITT, AND J. W. TAYLOR. 1998. Cryptic speciation and recombination in the aflatoxin producing fungus *Aspergillus flavus*. *Proc. Natl. Acad. Sci. USA* 95:388-393.
- GIRIBET, G., G. D. EDGEcombe, AND W. C. WHEELER. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* 413:157-161.
- GRAHAM, S. W., J. R. KOHN, B. R. MORTON, J. E. ECKENWALDER, AND S. C. H. BARRETT. 1998. Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. *Syst. Biol.* 47:545-567.
- HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174.
- HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst. Biol.* 45:92-98.
- 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.
- 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.
- KLUGE, A. G., AND A. J. WOLF. 1993. Cladistics: What's in a word? *Cladistics* 9:183-199.
- KOUFOPANOU, V., A. BURT, AND J. W. TAYLOR. 1997. Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. *Proc. Natl. Acad. Sci. USA* 94:5478-5482.
- LUTZONI, F. M. 1997. Phylogeny of lichen- and non-lichen-forming omphalinoid mushrooms and the utility of testing for combinability among multiple data sets. *Syst. Biol.* 46:373-406.
- LUTZONI, F. M., AND R. VILGALYS. 1995. Integration of morphological and molecular data sets in estimating fungal phylogenies. *Can. J. Bot.* 73:S649-S659.
- MADDISON, W. P., AND D. R. MADDISON. 1999. MacClade: Analysis of phylogeny and character evolution, version 3.08. Sinauer, Sunderland, Massachusetts.
- MASON-GAMER, R. J., AND E. A. KELLOGG. 1996. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Syst. Biol.* 45:524-545.
- MICKEVICH, M. F., AND M. S. JOHNSON. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Syst. Zool.* 25:260-270.
- NIXON, K. C., AND J. M. CARPENTER. 1996. On simultaneous analysis. *Cladistics* 12:221-241.
- RAMBAUT, A., AND N. C. GRASSLY. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13:235-238.
- REED, R. D., AND F. A. H. SPERLING. 1999. Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus *Papilio*. *Mol. Biol. Evol.* 16:286-297.
- RODRIGO, A. G., M. KELLY-BORGES, P. R. BERGQUIST, AND P. L. BERGQUIST. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *N.Z. J. Bot.* 31:257-268.
- STANGER-HALL, I. K., AND C. W. CUNNINGHAM. 1998. Support for a monophyletic Lemuriformes: Overcoming incongruence between data partitions. *Mol. Biol. Evol.* 15:1572-1577.
- SULLIVAN, J. 1996. Combining data with different distributions of among-site variation. *Syst. Biol.* 45:375-380.
- SWOFFORD, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer, Sunderland, Massachusetts.
- TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes. *Evolution* 37:221-244.
- VIDAL, N., AND G. LECOINTRE. 1998. Weighting and congruence: A case study based on three mitochondrial genes in vipers. *Mol. Phylogenet. Evol.* 9:366-374.
- WENZEL, J. W., AND M. E. SIDDALL. 1999. Noise. *Cladistics* 15:51-64.
- WILGENBUSCH, J., AND K. DE QUIEROZ. 2000. Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. *Syst. Biol.* 49:592-612.
- YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396-1401.
- YANG, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42:587-596.
- YANG, Z. 2000. Phylogenetic analysis by maximum likelihood (PAML), version 3.0c. Univ. College, London.
- YODER, A. D., J. A. IRWIN, AND B. A. PAYSEUR. 2001. Failure of the ILD to determine data combinability for slow loris phylogeny. *Syst. Biol.* 50:408-424.

First submitted 25 September 2000; reviews returned
23 January 2001; final acceptance 9 April 2002
Associate Editor: Richard Olmstead