Points of View

Syst. Biol. 53(1):81–89, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490264752

Congruence Versus Phylogenetic Accuracy: Revisiting the Incongruence Length Difference Test

ANDREW L. HIPP,[1] JOCELYN C. HALL,[2] AND KENNETH J. SYTSMA[1]

[1]*Department of Botany, University of Wisconsin, 430 Lincoln Drive, Madison, Wisconsin 53706, USA; E-mail: alhipp@wisc.edu (A.L.H.)*
[2]*Arnold Arboretum, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA*

Phylogenies inferred from independent data partitions usually differ from one another in topology despite the fact that they are drawn from the same set of organisms (Rodrigo et al., 1993). Some topological differences are due to sampling error or to the use of inappropriate phylogenetic models. These types of topological incongruence do not have their origin in genealogical discordance, i.e., differences between phylogenies underlying the respective data partitions (Baum et al., 1998). Incongruence that is not due to genealogical discordance can often be addressed by modifying the model used in phylogenetic reconstruction (Cunningham, 1997b), and combining data is an appropriate way of dealing with random topological differences that are attributable to sampling error. However, other topological differences, e.g., those arising from lineage sorting (Maddison, 1997; Avise, 2000) and hybridization (Dumolin-Lapègue et al., 1997; Rieseberg, 1997; McKinnon et al., 1999; Avise, 2000), reflect genealogical discordance between the data partitions.

Most systematists consider data partitions to be combinable if and only if they are not strongly incongruent with one another (Sytsma, 1990; Bull et al., 1993; Huelsenbeck et al., 1996; Baum et al., 1998; Johnson and Soltis, 1998; Thornton and DeSalle, 2000; Yoder et al., 2001; Barker and Lutzoni, 2002; Buckley et al., 2002). Systematists who follow this prior agreement or conditional combination approach to analyzing multiple data partitions (Bull et al., 1993; Huelsenbeck et al., 1996; Johnson and Soltis, 1998) evaluate incongruence using tests such as the incongruence length difference (ILD) test (Farris et al., 1994, 1995) or other tests of taxonomic congruence (Templeton, 1983; Kishino and Hasegawa, 1989; Larson, 1994; Shimodaira and Hasegawa, 1999) before deciding whether the partitions should be analyzed in combination. Data that exhibit strong incongruence are then analyzed separately or under assumptions that minimize incongruence (Cunningham, 1997b).

In their article "Failure of the ILD to determine data combinability for slow loris phylogeny," Yoder et al. (2001) critiqued the ILD test based on the observation that it will sometimes identify data partitions as incongruent when in fact those partitions combine to produce an accurate estimate of organismal phylogeny. They described the ILD test as a failed test of data combinability, maintaining that the presumed accuracy of trees inferred from combined data indicates the congruence of the data partitions. We have two objections to their argument (2001:421) that "the ILD [should] never be used as a test of data partition combinability." First, what Yoder et al. described as a flaw in the ILD test as applied to their data, i.e., an apparent inverse relationship between phylogenetic accuracy and data partition congruence as measured by the ILD test, turns out to be an artifact of analysis. There is in fact a bimodal relationship between congruence and accuracy: as either data partition is upweighted, homoplasy in the combined data set is swamped by homoplasy within the upweighted data partition, reducing the significance of the ILD test. At the same time, the topology of the combined analysis shifts to reflect the topology of the upweighted data partition. This phenomenon is predictable and can be accounted for in the analysis (Dowton and Austin, 2002). Second, Yoder et al.'s expectation that ILD test results should predict the phylogenetic accuracy of the combined data analysis is unreasonable. The ILD test is used to evaluate the null hypothesis that characters that make up two or more data partitions are drawn at random from a single population of characters, i.e., a population of characters that reflects a single phylogeny and a single set of evolutionary processes (Farris et al., 1995). Because accuracy of trees derived from a data set depends on many factors other than congruence among data partitions, the ILD test cannot be used to directly address questions related to phylogenetic accuracy. Genealogically discordant data can be combined to yield accurate phylogenies, whereas data that are congruent (both genealogically concordant and homogeneous in underlying evolutionary process) can be combined to yield phylogenies that do not accurately represent organismal history (Cunningham, 1997a). A damaging critique of the ILD test would have to appeal to criteria other than

phylogenetic accuracy. We demonstrate, moreover, that the ILD test supports the Templeton (1983), Kishino–Hasegawa (KH) (1989), and Shimodaira–Hasegawa (SH) (1999) tests in identifying two points of incongruence between Yoder et al.'s data partitions. We conclude that Yoder et al.'s arguments fail to demonstrate that the ILD test fails as a test of data partition congruence.

Background: Summary of the ILD Test

The ILD test (Farris et al., 1994, 1995) is one of the most commonly used statistical measures of character incongruence between phylogenetic data partitions (for reviews, see Huelsenbeck et al., 1996; Mason-Gamer and Kellogg, 1996; Cunningham, 1997a, 1997b; Johnson and Soltis, 1998; Dolphin et al., 2000; Thornton and DeSalle, 2000; Yoder et al., 2001; Dowton and Austin, 2002; Barker and Lutzoni, 2002; Darlu and Lecointre, 2002). The test is based on an expectation that data partitions that reflect different topologies or different underlying evolutionary processes will have higher overall homoplasy in combination than will data partitions that reflect a single topology and evolutionary process. Consequently, combined analysis of incongruent data sets should yield trees that are significantly longer than the sum of the tree lengths inferred from each data partition separately. The ILD test statistic, D, is the difference between tree lengths of combined data partitions and the sum of tree lengths of data partitions analyzed separately:

$$D = L_{(1+2+\cdots+N)} - (L_1 + L_2 + \cdots + L_N),$$

where L_N is the length of the most-parsimonious tree(s) found for each data partition N and $L_{(1+2+\cdots+N)}$ is the length of the most-parsimonious tree(s) for the combined data. By comparing D to a distribution generated by randomly partitioning the combined data according to the number and size of the original data partitions, the ILD test provides a P value that estimates the type I error rate, i.e., the probability of rejecting the null hypothesis that data partitions are congruent with one another when in fact the partitions are congruent with one another (Farris et al., 1994).

The ILD test as implemented in PAUP* (Swofford et al., 1998) and ARN (Farris et al., 1995) utilizes tree length (parsimony) to calculate the test statistic, but the method could in principle be implemented using likelihood or distance methods (Cunningham, 1997b). Although the conditions under which the test should return a significant result are probably not all known, recent studies have demonstrated that in addition to character incongruence caused by genealogical discordance the test is sensitive to between-partition differences in among-site rate variation (Darlu and Lecointre, 2002), overall evolutionary rates (Barker and Lutzoni, 2002; Darlu and Lecointre, 2002), levels of noise (Dolphin et al., 2000), and relative size of the data partitions being tested (Dowton and Austin, 2002). These differences appear to affect the ILD test results through their effect on the amount of phylogenetic structure in the data (Barker and Lutzoni, 2002), increasing the probability of type I errors (the error of incorrectly rejecting the correct hypothesis of congruence). The ILD test appears to be less susceptible to type II errors when sufficient numbers of informative sites are available (Darlu and Lecointre, 2002) and when data partitions are appropriately weighted relative to one another (Dowton and Austin, 2002).

ILD Test Performance in Yoder et al.'s Study

There Is No Inverse Relationship Between Phylogenetic Accuracy and ILD P Value

Yoder et al. argued that the slow lorises are monophyletic, a resolution that is supported by the morphological data but not by the molecular data. They performed the ILD test under a variety of weighting strategies and found that in most cases the test detects significant incongruence between the molecular and morphological data partitions ($P < 0.01$; the two molecular data partitions were congruent with each other). They also found that weighting schemes in which the ILD test returned a low P value (indicating incongruence) were those in which combined analysis supported the supposed correct monophyly of the Loridae, generally with bootstrap values of 85–100% and 66% in one case (Yoder et al., 2001: tables 4, 5). In contrast, combined analysis supported a paraphyletic Loridae (bootstrap values of 43–63%) under weighting schemes in which the data partitions passed the ILD test. Presuming, based on morphological data, that the Loridae are monophyletic, Yoder et al. (2001:419) considered their result as demonstrating a "complete reversal of congruence and accuracy" and therefore concluded that the ILD test is a failure.

This "reversal," however, is an artifact of not analyzing across a sufficiently broad range of relative data partition weights. To demonstrate this fact, we reanalyzed Yoder et al.'s data using bootstrap and ILD analyses following Yoder et al.'s methods and using data reassembled using IRBP and cytochrome *b* sequences deposited in GenBank (Yoder et al., 2001: table 2) and morphological data matrices presented by Yoder (1994: appendix, table 1) and Yoder et al. (2001: table 3).

As the morphological data are weighted more heavily, the increasing "equality" of the data partitions as measured in tree length makes the ILD test increasingly sensitive to heterogeneity between them (Table 1). This finding appears to contradict Farris et al.'s (1995:318) assertion that difference in number of characters between data partitions "is hardly a problem when assessing incongruence between matrices" but it is in keeping with another recent study that showed that increasing the size of one data partition over the other tends to reduce incongruence as measured by the ILD test (Dowton and Austin, 2002).

Not surprisingly, even heavier weighting of the morphological data causes ILD test P values to increase above the significance threshold while the topology of the combined analysis continues to reflect the morphology tree (Table 1). This is consistent with an expectation

TABLE 1. Analysis of Yoder et al.'s (2001) reduced (nine-taxon) data set under a range of relative morphological data partition weights. All analyses were performed as described by Yoder et al. Characters were reweighted evenly within the morphological data partition; characters within each molecular data partition were maintained at a relative weight of 1. The table indicates that there is not, as Yoder et al. described (2001:000), an "inverse relationship between congruence and accuracy." Rather, the ILD test is sensitive to different weighting strategies: when weighting strongly emphasizes either of the data partitions, the ILD test is focused on within-partition homogeneity rather than the between-partition heterogeneity. M = Loridae monophyly, the presumed accurate topology; P − denotes Loridae paraphyly.

IRBP vs. Morphology		
Morph weighting	**Bootstrap**	**ILD p-value**
0.01	P = 68	0.253
0.05	P = 69	0.262
0.1	P = 54	0.011
0.2	M = 66	0.00011
0.3	M = 86	0.00002
0.4	M = 96	0.00003
0.5	M = 99	0.00004
1	M = 100	0.00004
2	M = 100	0.002
5	M = 100	0.026
10	M = 100	0.127
15	M = 100	0.219
20	M = 100	0.4
25	M = 100	0.409
30	M = 100	0.408
35	M = 100	0.422
40	M = 100	0.426

cytB vs. morphology		
Morph weighting	**Bootstrap**	**ILD p-value**
0.01	P = 53	0.977
0.05	P = 52	0.97
0.1	P = 50	0.876
0.2	P = 47	0.55
0.3	P = 39	0.303
0.4	M = 49	0.154
0.5	M = 59	0.097
1	M = 91	0.014
2	M = 100	0.000638
5	M = 100	0.00052
10	M = 100	0.007
15	M = 100	0.066
20	M = 100	0.087
25	M = 100	0.228
30	M = 100	0.348
35	M = 100	0.376
40	M = 100	0.44

that ILD test results will increase in significance as homoplasy in the combined data increases relative to homoplasy in the data partitions analyzed separately. Increasing either data partition strengthens the homoplasy within data partitions relative to that of the combined data partitions. In other words, Yoder et al.'s data demonstrate not that there is an inverse relationship between ILD test *P* value and accuracy but that (1) the equalization of data partitions' relative weights causes the ILD test to become more sensitive to incongruence between those data partitions and (2) increased weighting of any one data partition increases the chances that the combined data tree will reflect the topology found in that data partition. The conclusion that there is an inverse relationship between phylogenetic accuracy and congruence is a consequence of overlooking the portion of the data weighting spectrum in which morphological data are most heavily weighted.

Phylogenetic Accuracy Is Not a Reliable Indicator of Data Partition Congruence

Yoder et al. argued that (2001:421)

> All of the conclusions with regard to the reliability of the ILD test rest on the assumption that the slow loris clade is real and thus phylogenetically accurate. If this assumption is false, then the ILD test could be said to have performed nearly perfectly, giving accurate results when no heterogeneity was detected and false results when it was.

This argument appears to presuppose that phylogenetic accuracy is a good indicator of data partition congruence, an assumption that would be reasonable if phylogenetic accuracy depended on the congruence of underlying data partitions. However, this is not the case.

Improving the evolutionary model used in phylogenetic inference should increase both congruence between data partitions and the accuracy of the phylogeny recovered (Cunningham, 1997b) but only if those partitions share a common and accurate organismal phylogeny in the first place. Under some conditions, however, independent data partitions for a given set of organisms are expected to misrepresent phylogenetic relationships in the same way. Long-branch attraction, for instance, can produce inaccurate phylogenies despite congruence between seemingly independent data partitions (Felsenstein, 1978; Hendy and Penny, 1989; Swofford et al., 2001). Likewise, there are conditions under which incongruent data partitions are expected to yield accurate phylogenies in combination, e.g., when a data partition that adheres to the organismal phylogeny is larger or more heavily weighted than the others (e.g., Dowton and Austin, 2002), exhibits less homoplasy at a level of inquiry relevant to the study at hand, or has a large number of accurate and informative characters at a particularly important node. If one assumes that the slow lorises truly are monophyletic, then this last scenario applies to Yoder et al.'s data.

Thus, the accuracy of a combined data tree does not depend on congruence between the separate data sets that it comprises. When an accurate topology is weighted strongly enough, the combined data tree will be accurate irrespective of incongruence. As Yoder et al. demonstrated, with sufficient weighting a combined data tree can be made to fit the topology of any data partition. In their study, the strong morphological signal for monophyly of the Loridae swamps the weaker molecular signal for paraphyly, producing a tree that matches what Yoder et al. expected of the organismal phylogeny based on morphological observations. We do not take issue with the presumed accuracy of this tree. We maintain, however, that phylogenetic accuracy cannot be used in this way to evaluate the ILD test or any other test of data partition congruence.

Using the ILD Test with Other Tests to Explore Incongruence

The decision to combine data cannot be based on the results of a single test. It has been proposed that the ILD test be used as a starting point for comparing data partitions (Mason-Gamer and Kellogg, 1996). We see this as the test's most appropriate role. Recent studies have demonstrated that the ILD test is subject to a high rate of type I error when the null hypothesis being addressed is that data partitions are genealogically concordant. This high error rate is due to the test's sensitivity to between-partition differences in noise and evolutionary rate and extremes of rate heterogeneity among sites within the data as a whole (Dolphin et al., 2000; Yoder et al., 2001; Barker and Lutzoni, 2002; Darlu and Lecointre, 2002; Dowton and Austin, 2002). These same studies, however, suggest that data partitions that pass the ILD test exhibit only minor topological incongruence (i.e., incongruence involving at most a relatively small region of the tree or a relatively small number of taxa), contain relatively few variable nucleotide positions (Darlu and Lecointre, 2002), differ strongly in number of informative characters (Dowton and Austin, 2002), or possess very high rate heterogeneity among sites ($\alpha = 0.06$, Darlu and Lecointre, 2002). In other words, these studies suggest that the ILD test is not overly susceptible to type II errors (regarding the null hypothesis of genealogical concordance between partitions) when data partitions are weighted appropriately relative to one another, are informative for the phylogenetic level under investigation, and are sufficient in number of informative base pairs, except in cases of extreme rate heterogeneity. All these conditions can be evaluated by researchers, suggesting that "passing grades" on the ILD test are meaningful and that the ILD test can serve as a conservative first test of data partition congruence.

Following an initial screen on all taxa, the ILD test can be used to help determine which taxa contribute the most to incongruence between data partitions; changes in ILD test P value for a given data set should reflect changes in congruence within that data set even when the test turns out to be overly conservative for some data sets relative to others. Yoder et al. (2001:416) restricted their study of the behavior of the ILD test to nine species, partly to "focus on incongruence related to the slow lorises." Their subsequent arguments based on the presumed monophyly of the Loridae rest in part on the assumption that they succeeded in this goal. To investigate this assumption and to explore the potential sources of incongruence in their data, we reanalyzed all eight-taxon permutations of Yoder et al.'s reduced nine-taxon data set using bootstrapping and the partition homogeneity test implemented in PAUP* 4.0b8-10 (Swofford, 1998), following methods of Yoder et al. with two exceptions: randomization seeds were not specified, and only heuristic searches (not branch-and-bound searches) were performed. An initial set of trials indicated that this divergence from their methods had no effect on our conclusions.

If the sole source of incongruence in Yoder et al.'s nine-taxon data set were the resolution of the Loridae, removal of taxa not involved in resolution of the Loridae might be expected to increase the significance of ILD test results, because removal of those taxa would focus the test further on the one source of incongruence. To the contrary, removal of either of two taxa not in the Loridae (*Lemur catta* and *Daubentonia madagascariensis*) had a substantial effect on the test result, as did removal of the contentious member of the Loridae (*Perodicticus potto*; Table 2). Based on this result, we inferred that the ILD test is picking up on not just differences in how morphological and molecular data resolve the slow lorises but also in how the two types of data resolve relationships within the clade that contains *Lemur*, *Propithecus*, and *Daubentonia* (Fig. 1).

Once the ILD test has been used to evaluate character incongruence and to identify specific taxa contributing to it, the nature of the incongruence can be more closely investigated using tests of the statistical support within each data partition for alternative topologies.

TABLE 2. Analyses of all eight-taxon sets within Yoder et al.'s (2001) reduced (nine-taxon) data set under equal weightings. All analyses were performed as described by Yoder et al. Each of the nine taxa of the authors' reduced taxon set was removed one at a time for nine tests of eight taxa each. The results indicate that the reduced-taxon set of Yoder et al. does not focus the ILD test solely on the problematic slow loris group but also on the clade that includes *Daubentonia* and *Lemur*. These two taxa differ in relative position between the molecular and morphological data partitions (see Fig. 1), and thus it is not surprising that the removal of either one of them raises the ILD test P value to the passing level. In all tests, the Loridae are resolved as monophyletic, with bootstrap values as indicated.

Taxon excluded	ILD p-value	Bootstrap support for Loris monophyly
Otolemur crassicaudatus	0.005	69%
Galago moholi	0.008	95%
Galagoides demidoff	0.046	96%
Perodicticus potto	0.375	n/a
Loris tardigradus	0.005	79%
Nycticebus coucang	0.007	73%
Lemur catta	0.26	91%
Propithecus tattersalli	0.003	88%
Daubentonia madagascariensis	0.238	93%

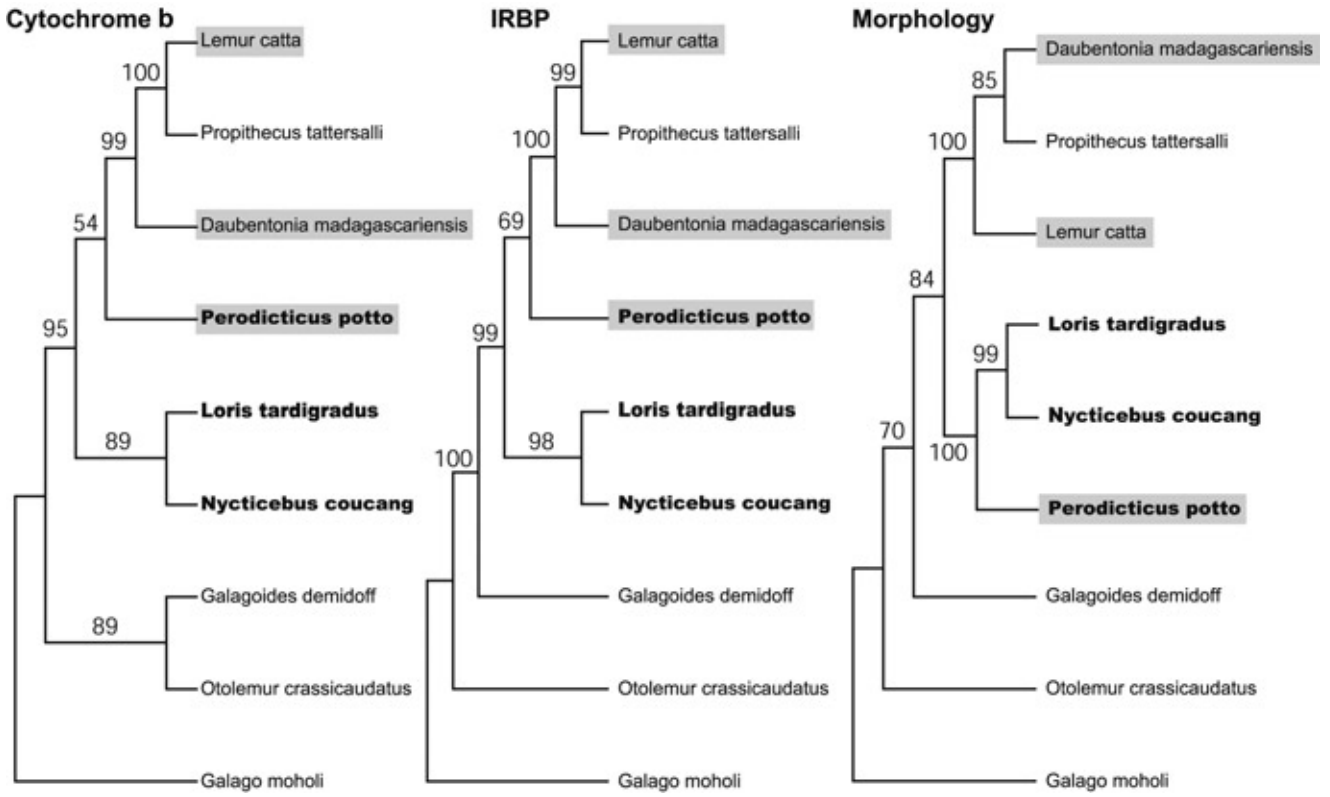

FIGURE 1. Unrooted cladograms for reduced taxon set, analyzing each of Yoder et al.'s (2001) data partitions separately. Slow lorises indicated in bold. Removal of any of the three shaded taxa reduces incongruence sufficiently that the data pass a three-way ILD test. Analyses were performed on unweighted data as described by Yoder et al. Numbers above branches are bootstrap values (also performed as described by Yoder et al.). Cytochrome *b*: length = 968, consistency index (CI) = 0.561, retention index (RI) = 0.368; IRBP: length = 116, CI = 0.809, RI = 0.841; morphology: length = 91, CI = 0.725, RI = 0.764.

The Wilcoxon signed-ranks test is most commonly used in this context (Templeton, 1983; Larson, 1994). However, because this test was designed to evaluate the support for topologies selected a priori, it is inappropriately applied in situations in which topologies are compared to a most-parsimonious tree (Shimodaira and Hasegawa, 1999; Goldman et al., 2000; Shimodaira, 2002). Under these conditions, the *P* value underestimates the confidence interval for the most-parsimonious topology (Shimodaira, 2002). A correction has been proposed for both this test, and the KH (1999) test, which is subject to the same problem (Goldman et al., 2000), and multiple-comparison methods appropriate to the evaluation of a posteriori hypotheses have been developed (Shimodaira and Hasegawa, 1999; Goldman et al., 2000; Shimodaira, 2002).

To further investigate the nature of incongruence in Yoder et al.'s data, we used the Templeton test and the parsimony-based implementation of the KH test, which facilitated the comparison of molecular data with morphological data in the absence of an implemented likelihood model for the latter (although see the model of Lewis, 2001, as implemented in MrBayes 3; Huelsenbeck and Ronquist, 2001). *P* values reported are Bonferroni corrected and one tailed, correcting for multiple comparisons and the fact that one of the topologies in each comparison is known to be optimal (Goldman et al., 2000). We compared KH and Templeton test results with the likelihood-based SH test for all cases in which molecular data were used to evaluate the differences between topologies. Confidence intervals determined using the SH test vary depending on how many tree topologies are evaluated at a time (Buckley et al., 2001; Shimodaira, 2002), and the test tends to overestimate the confidence interval around the optimal tree, especially for comparisons that involve many trees (Shimodaira, 2002). The Templeton and KH tests, however, tend to underestimate the confidence interval around optimal trees (Shimodaira and Hasegawa, 1999; Buckley et al., 2001; Shimodaira, 2002). Consequently, points of agreement between the tests should provide a robust estimate of the confidence interval about the optimal tree.

Data were divided into two partitions, molecular (IRBP + cytochrome *b*) and morphological, with all characters weighted equally. The tests were implemented using the appropriate Tree Score options in PAUP* 4.0b8-10. One-tailed Templeton and KH tests are reported with both uncorrected and Bonferroni-corrected *P* values (Buckley et al., 2001) (Tables 3, 4). Fully resolved

TABLE 3. Templeton and Kishino–Hasegawa (KH) tests on Yoder et al.'s (2001) data divided into molecular and morphological partitions. Data were divided into one molecular (IRBP and cytochrome *b*) and one morphological partition for analysis; all data were equally weighted. *P* values are all one tailed, with Templeton test values above the KH test values; *P* values in parentheses are Bonferroni corrected, calculated by multiplying the *P* values by the total number of comparisons ($n = 15$ for the 12-taxon test; $n = 6$ for the 9-taxon test). Results indicate that molecular data reject one of the two most-parsimonious morphological trees in the 12-taxon test and the only most-parsimonious morphological tree in the 9-taxon tests; morphological data reject the most-parsimonious molecular trees. Neither the morphological nor the molecular data reject the combined data trees.

12-taxon data sets

Data Set	# Trees	Constraint Topology				
		Morphology 1	Morphology 2	Molecular	Combined 1	Combined 2
Morphology	2	n.a.	n.a.	0.0021 (0.0315) 0.0012 (0.018)	0.3438 (1.000) / 0.2539 (1.000) 0.2087 (1.000) / 0.1601 (1.000)	0.4018 (1.000)/ 0.3238 (1.000) 0.3099 (1.000) / 0.2473 (1.000)
Molecular	1	0.005 (0.075) 0.0038 (0.057)	<0.0001 (0.0015) <0.0001 (0.0015)	n.a.	0.2593 (1.000) 0.21935 (1.000)	0.11475 (1.000) 0.0808 (1.000)
Combined	2	0.0733 / 0.1390 0.0587 / 0.1020	0.0001 (0.0015) / <0.0001 (0.0015) <0.0001 (0.0015) / 0.0001 (0.0015)	0.4075 (1.000) / 0.3677 (1.000) 0.3528 (1.000) / 0.6122 (1.000)	n.a.	n.a.

9-taxon data sets

Data Set	# Trees	Constraint Topology		
		Morphology	Molecular	Combined
Morphology	1	n/a	<0.0001 (0.0006) 0.001 (0.0006)	0.1875 (1.000) 0.0906 (0.5436)
Molecular	1	<0.0001 (0.0006) <0.0001 (0.0006)	n/a	0.1239 (0.7434) 0.0890 (.5340)
Combined	1	0.0005 (0.0030) 0.0003 (0.0018)	0.1744 (1.000) 0.1372 (0.8232)	n/a

TABLE 4. Templeton, Kishino–Hasegawa (KH), and Shimodaira–Hasegawa (SH) tests of four specific hypotheses in the nine-taxon data set. Data were divided into one molecular (IRBP and cytochrome *b*) and one morphological partition for analysis. *P* values are all one tailed, with Templeton test values above the KH test values; for molecular data only, SH tests are presented below KH and Templeton test results. *P* values in parentheses are Bonferroni corrected, calculated by multiplying the *P* values by the total number of comparisons ($n = 4$). The SH test was performed on a mix of seven a priori and a posteriori trees: (1) The MP molecular tree (same as the maximum likelihood tree); (2) *Lemur* and *Propithecus* constrained not to be sister to one another in molecular tree; (3) Loridae constrained to be monophyletic in molecular tree; (4, 5) the two MP morphology trees; (6) Loridae constrained not to be monophyletic in the morphological tree; and (7) *Lemur* and *Propithecus* constrained to be sister to one another in morphological tree. Constraint topologies: Loris = Loridae monophyletic; ~Loris = Loridae nonmonophyletic; (L, P) = *Lemur* and *Propithecus* sister to one another; ~(L, P) = *Lemur* and *Propithecus* not sister to one another.

Data partition	Constraint Tree			
	(L, P)	Loris	~ (L, P)	~ Loris
Morphological	0.0899 (0.3594) / 0.0899 (0.3594) 0.0899 (0.3594) / 0.0906 (0.3624)	n/a	n/a	<0.0001 (0.0004) <0.0001/<0.0001
Molecular	n/a	0.0890 (0.3558) / 0.0890 (0.3558) 0.0890 (0.3558) / 0.0890 (0.3560) 0.562	<0.0001 (0.0004) / <0.0001 (0.0004) 0.0001 (0.0004) / 0.0001 (0.0004) 0.003	n/a

most-parsimonious trees were used as constraints. SH tests were implemented in the likelihood tree scores menu, using likelihood parameters reported by Yoder et al. and simulating RELL and full optimization distributions using 1,000 bootstrap replicates (Tables 4, 5). Maximum parsimony (MP) topologies were compared with constrained MP topologies and unconstrained topologies 5–12 steps longer than the MP topologies for each data partition and for the combined data, for a total of 31 trees (Table 5); this set of trees includes the maximum likelihood ML topology for the molecular data, which is not significantly different from the MP topology (Table 5). Although this method is an imperfect means of choosing trees for simultaneous comparison in the SH test, it serves to eliminate extremely unlikely topologies from analysis (Shimodaira and Hasegawa, 1999; Buckley et al., 2001) and increases the conservativeness of the SH test (Shimodaira, 2002).

The tests show molecular data to be incompatible with morphological trees and morphological data to be incompatible with molecular trees for the most part, but neither the molecular data nor the morphological data reject the combined tree (Tables 3–5). As demonstrated by our individual taxon removals using the ILD test (Table 2), two points of incongruence are apparent between the molecular and morphological data in the nine-taxon tests: the morphological data support monophyly of the Loridae and a sister relationship between *Propithecus* and *Daubentonia*, whereas the molecular data support paraphyly of the Loridae and a sister relationship between *Propithecus* and *Lemur*.

If support for these relationships were strong for each data partition, the existence of a most-parsimonious tree that neither partition rejects would be surprising. However, test results using constraint trees that reflect the two clades of contention reveal that although the morphological data strongly reject nonmonophyly of the slow lorises, they do not reject the most-parsimonious tree in which *Propithecus* and *Lemur* are sister to one another (Table 4). Similarly, although the molecular data strongly reject the most-parsimonious trees in which *Propithecus* and *Lemur* are not sister to one another, they do not reject monophyly of the slow lorises (Tables 4, 5). Thus, each of the nine-taxon data partitions strongly supports only one of the two points of incongruence, and the combined nine-taxon tree, which is monophyletic for the Loridae and places *Propithecus* sister to *Lemur*, is not rejected by either data set. The 95% confidence interval for the molecular tree determined using the SH test rejects two trees in which *Lemur* and *Propithecus* are sister. However, this result probably should not be taken as an indication of incongruence between the morphological and molecular data because these two trees, in which *Galago*, *Galagoides*, and *Otolemur* fail to form a partition, are not among the most-parsimonious morphological trees.

The existence of a set of trees based on combined data that neither data partition can reject may be taken as evidence that Yoder et al.'s data are combinable. At the same time, the strong support in each data partition for a topology rejected by the opposing partition is intriguing and worth exploring further. Use of the ILD test in conjunction with the Templeton, KH, and SH tests appears to have effectively identified the proximal sources of incongruence in Yoder et al.'s data. In the end, the causes of this incongruence must be inferred by other means.

CONCLUSIONS

As has been demonstrated in simulation studies (Dolphin et al., 2000; Barker and Lutzoni, 2002; Darlu and Lecointre, 2002; Dowton and Austin, 2002) and as suggested in Yoder et al., significant ILD test *P* values should not be taken as a conclusive demonstration that analyzing independent data partitions in combination will produce misleading phylogenies. However, these studies do not support categorical or unqualified rejection of the ILD test. It is probably unreasonable to expect that any test of data incongruence would be capable of identifying cases in which combining data will increase phylogenetic accuracy (Huelsenbeck et al., 1996). The question of whether to combine data is complex and must be explored using a range of methods (e.g., Mason-Gamer and Kellogg, 1996; Cunningham, 1997a; Johnson and Soltis, 1998; Buckley et al., 2002). Where the

TABLE 5. Shimodaira–Hasegawa (SH), Kishino–Hasegawa (KH), and Templeton tests of the support for three key partitions in the nine-taxon data set. Tests were performed on MP molecular, morphological, and combined analyses and on trees that were up to 12 steps longer than the MP tree. Morphological data were excluded for all tests; results reflect only the confidence interval around the optimal molecular tree. *P* values are all one tailed; *P* values in parentheses are Bonferroni corrected, calculated by multiplying the *P* values by $n - 1$, where n = number of topologies ($n = 31$). Loris = Loridae monophyletic; (L, P) = *Lemur* and *Propithecus* sister to one another; (G, G, O) = *Galago*, *Galagoides*, and *Otolemur* form a partition.

			1-tailed	1-tailed	Supports partition of:		
		SH test	Templeton test	KH test	Loridae	L,P	G,G,O
Molecular trees:	MP + 7 steps	(best)	0.295 (1.000)	(best)	No	Yes	Yes
		0.930	(best)	0.414 (1.000)	No	Yes	Yes
		0.854	0.0890 (1.000)	0.142 (1.000)	No	Yes	Yes
		0.645	0.1890 (1.000)	0.116 (1.000)	No	Yes	Yes
		0.603	0.250 (1.000)	0.072 (1.000)	No	Yes	Yes
		0.585	0.316 (1.000)	0.146 (1.000)	No	Yes	Yes
		0.539	0.118 (1.000)	0.101 (1.000)	No	Yes	Yes
Combined trees:	MP + 12 steps	1.000	0.295 (1.000)	1.000 (1.000)	No	Yes	Yes
		0.930	(best)	0.414 (1.000)	No	Yes	Yes
		0.854	0.0890 (1.000)	0.142 (1.000)	Yes	Yes	Yes
		0.585	0.316 (1.000)	0.146 (1.000)	No	Yes	Yes
		0.510	0.080 (1.000)	0.066 (1.000)	Yes	Yes	Yes
		0.450	0.138 (1.000)	0.035 (1.000)	Yes	Yes	Yes
Morphological trees:	MP + 5 steps	0.854	0.089 (1.000)	0.142 (1.000)	Yes	Yes	Yes
		0.510	0.080 (1.000)	0.066 (1.000)	Yes	Yes	Yes
		0.445	<0.0001 (0.003)	0.031 (0.930)	Yes	Yes	Yes
		0.037	0.0034 (0.102)	0.002 (0.060)	Yes	Yes	No
		0.029	<0.0001 (0.003)	0.002 (0.060)	Yes	Yes	No (poly)
		0.005	<0.0001 (0.003)	0.001 (0.030)	Yes	No	Yes
		0.004	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	Yes
		0.003	0.0001 (0.003)	<0.001 (<0.001)	Yes	No	Yes
		0.003	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	Yes
		0.003	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	Yes
		0.002	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	Yes
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No (poly)
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No
		<0.0001	<0.0001 (0.003)	<0.001 (<0.001)	Yes	No	No

precise nature of incongruence between data partitions cannot be inferred, it may be most appropriate to assume Swofford's (1991:329) "admittedly non-Popperian position that an ambiguous solution that contains the truth is, in many situations, preferable to an unambiguous solution that is wrong."

ACKNOWLEDGMENTS

We thank David A. Baum for technical advice on analysis and for enriching our understanding of what may constitute combinable data and both David and the Systematics Lab Group of the UW–Madison Botany Department for comments on earlier drafts of this paper. The arguments and analysis presented here benefited significantly from comments by Cliff Cunningham, Francois Lutzoni, Chris Simon, and an anonymous reviewer.

REFERENCES

Avise, J. C. 2000. Phylogeography: The history and formation of species. Harvard University Press. Cambridge, Massachusetts.

Barker, F. K., and F. M. Lutzoni. 2002. The utility of the incongruence length difference test. Syst. Biol. 51:625–637.

Baum, D. A., R. Small, and J. F. Wendel. 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. Syst. Biol. 47:181–207.

Buckley, T. R., C. Simon, H. Shimodaira, and G. K. Chambers. 2001. Evaluating hypotheses on the origin and evolution of the

New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. Mol. Biol. Evol. 18:223–234.
Buckley, T. R., P. Arensburger, C. Simon, and G. K. Chambers. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.
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.
Cunningham, C. W. 1997a. 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.
Cunningham, C. W. 1997b. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14:733–740.
Darlu, P., and G. Lecointre. 2002. When does the incongruence length difference test fail? Mol. Biol. Evol. 19:432.
Dolphin, K., R. Belshaw, D. L. C. 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 congruence does not necessarily indicate increased phylogenetic accuracy: The behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol. 51:19–31.
Dumolin-Lapègue, S., B. Demesure, S. Fineschi, V. Le Corre, and R. J. Petit. 1997. Phylogeographic structure of white oaks throughout the European continent. Genetics 146:1475–1487.
Farris, J. D., M. Källersjö, A. G. Kluge, and C. Bult. 1994. Testing significance of incongruence. Cladistics 10:315–319.
Farris, J. D., M. Källersjö, A. G. Kluge, and C. Bult. 1995. Constructing a significance test for incongruence. Syst. Biol. 44:570–572.
Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.
Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.
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 F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Biometrics 17:754–755.
Johnson, L. A., and D. E. Soltis. 1998. Assessing congruence: Empirical examples from molecular data. Pages 297–348 *in* Molecular systematics of plants II: DNA sequencing (D. E. Soltis, P. S. Soltis, and J. J. Doyle, eds). Kluwer, Norwell, Massachusetts.
Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimates of the evolutionary tree topologies from sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.
Larson, A. 1994. The comparison of morphological and molecular data in phylogenetic systematics. Pages 371–390 *in* Molecular ecology and evolution: Approaches and applications (B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, eds.). Birkhäuser Verlag, Basel, Switzerland.
Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
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.
McKinnon, G. E., D. A. Steane, B. M. Potts, and R. E. Vaillancourt. 1999. Incongruence between chloroplast and species phylogenies in *Eucalyptus* subgenus *Monocalyptus* (Myrtaceae). Am. J. Bot. 86:1038–1046.
Rieseberg, L. H. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359–389.
Rodrigo, A. G., M. Kelly-Borges, P. R. Bergquist, and P. L. Bergquist. 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. N.Z. J. Bot. 31:257–268.
Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.
Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.
Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongrugent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, Oxford, U.K.
Swofford, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer, Sunderland, Massachusetts.
Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.
Sytsma, K. J. 1990. DNA and morphology: Inference of plant phylogeny. Trends Ecol. Evol. 5:104–110.
Templeton, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244.
Thornton, J. W., and R. DeSalle. 2000. A new method to localize and test the significance of incongruence: Detecting domain shuffling in the nuclear receptor superfamily. Syst. Biol. 49:183–201.
Yoder, A. D. 1994. Relative position of the Cheirogaleidae in strepsirhine phylogeny: A comparison of morphological and molecular methods and results. Am. J. Phys. Anthropol. 94:25–46.
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 1 July 2002; reviews returned 30 January 2003; final acceptance 4 September 2003
Associate Editor: François Lutzoni