Correcting the Problem of False Incongruence Due to Noise Imbalance in the Incongruence Length Difference (ILD) Test

DONALD L. J. QUICKE,1,2,3 OWEN R. JONES,1 AND DANIELLA R. EPSTEIN1

1 Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire, SL5 7PY, UK; E-mail: d.quicke@imperial.ac.uk (D.L.J.Q.)
2 NERC Centre for Population Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire, SL5 7PY, UK
3 Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK

Abstract.—The incongruence length difference (ILD) test is prone to suggesting significant conflict between character partitions when these differ only in the amount of undirected homoplasy (noise). This has been shown to be due to nonlinearity in the relationship between tree length and noise. Here we show that by standardizing either tree length or 1-retention index on a 0-to-1 scale, and then taking the arcsine of the value, the resulting value is linearly related to noise except at extremely high noise levels. We then investigate the effect of substituting these values instead of raw tree metrics in a modified ILD test (here called arcsine-ILD) for two types of noise. We show that, using the modified metric instead of the raw length, the results of ILD tests agreed better with desirable properties. [Congruence; homoplasy; incongruence; noise; partition homogeneity test.]

The incongruence length difference (ILD) test, also sometimes referred to as the partition homogeneity test, was invented to test whether two or more data partitions were supporting statistically different phylogenetic hypotheses (Farris et al., 1994, 1995). It has been widely used, often prior to deciding whether or not to combine the character partitions for simultaneous analysis. Recently, modifications to the ILD test have been proposed. For example, the relative incongruence length difference (RILD) test has been developed that allows identification of the part(s) of the tree where incongruence is maximal (Aagesen et al., 2005). Clearly being able to test whether data partitions are statistically incongruent or not is desirable, not only from the perspective of deciding whether or not partitions should be combined for analysis but also because real incongruence points to other biological phenomena.

However, since its introduction there have been a series of papers pointing out that the ILD test may give misleading results, principally suggesting significant incongruence in cases where it doesn’t appear to exist. Indeed, Yoder et al. (2001) even stated that, “...we recommend that the ILD never be used as a test of data partition combinability,” and more recently Barker and Lutzoni (2002), following a range of simulation studies, conclude that “The precise utility and appropriate uses of the ILD test remain to be established.” The first seeds of doubt about the validity of the test’s statistic start with Cunningham (1997) who suggested that invariant characters could artificially increase apparent incongruence and this was subsequently explored by Lee (2001) who extended this point to include also phylogenetically uninformative characters and showed that the probability of obtaining a significant P-value from the test increased with the proportion of uninformative characters and, in agreement with Cunningham’s inference, provided evidence that this phenomenon was most problematic when there was a substantial difference in the number of uninformative characters between the partitions. In any case, this issue is simply resolved by excluding noninformative characters prior to carrying out the test. In an exploration of the effect of different nucleotide substitution models, Dowton and Austin (2002) noted bimodality in the outcome of the ILD in that either increasing or decreasing the contribution of one of a pair of data partitions leads to increased congruence. Again these results indicate potential problems are greatest if the data partitions vary greatly in size, a phenomenon that was also mentioned by Darlu and Lecointre (2002). In addition, Darlu and Lecointre (2006) make the point that when data sets are simulated under different evolutionary conditions (e.g., constant versus variable evolutionary rates, homogeneous versus heterogenous rates of substitution), the probability of a type 2 error (falsely rejecting the true hypothesis of congruence) was greater than 5% and tends to increase with the number of sites.

Baker and DeSalle (1997) noted that there was no clear pattern in the various incongruent relationships that the ILD test revealed and suggested that one reason might be that different genes evolve at different rates because their problematic genes clearly showed evidence of saturation, whereas their nonproblematic ones did not. These studies point to potential problems with the ILD test, including sensitivity to noise (satisfaction in genes) and imbalance in signal strength (uninformative character imbalance), but mostly they fail to identify the mechanisms involved and therefore were unable to offer any solutions. Dolphin et al. (2000) investigated the effect of noise on test results using simulated 16-taxon data partitions with a single character supporting each node and showed that with initially congruent data partitions, the test started to suggest significant incongruence as the proportion of noisy characters in one of the partitions increased. The test, therefore, is not simply detecting positively incongruent signals but rather it is detecting differences in the nature of the data. This was in fact anticipated in the sense that the test was initially expected 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., from a population of characters that reflects a single
phylogeny and a single set of evolutionary processes (Farris et al., 1995)" (Hipp et al., 2004). The conflict investigated by Dolphin et al. (2000) was where the two data partitions were drawn from populations of characters that differed only in noisiness rather than one having a different conflicting phylogenetic signal. They found that the ILD test could indicate significant conflict when random noise was unequally distributed between the partitions. Such noise might be expected, for example, with third codon positions as they approach saturation, and thus these could show apparently significant ILD conflict when compared with less homoplastic second codon positions, even though they are in the same gene and consequently share the same phylogenetic history. Such incongruence between codon positions of the same gene is illustrated by Vidal and Lecointre (1998), who found more incongruence among codon positions within a given gene than there was between genes.

Dolphin et al. (2000) showed that the incongruence resulting from unequally distributed noise was the result of the nonlinear relationship between the amount of noise in a data set and tree length—the ILD test depending on this relationship being linear. Dolphin et al. (2000) suggested a solution to the problem of noise imbalance between data partitions. They proposed that ILD tests could also be carried out between each partition and randomized versions of them (where the character states within each character were permuted). However, this would be impractical in many cases because the ILD test P-values generated from these studies would often be extremely small and thus unrealistically large numbers of ILD replicates would be required to estimate them with sufficient accuracy.

Here we present a modification of the ILD test, which we term “arcsine-ILD,” which linearizes the relationship between tree length and noise and thus largely overcomes the tendency of the ILD test to reject the null hypothesis that the data partitions are random samples of the same data set due solely to any unequal distribution of noise.

The argumentation for the modification of tree length and noise is as follows. When there are noise levels that completely saturate the phylogenetic signal (i.e., the data are almost 100% noisy), the slope of the relationship between noise and length will be zero because, when all but one character is random (i.e., pure noise), the one remaining true character effectively will be random with respect to each of the others. Therefore, shuffling the states between taxa within this remaining true character are equally likely to increase or decrease tree length. When noise levels are smaller, the slope of the same relationship is asymptotic to a nonzero value. This is because randomizing some small proportion of characters or data cells in a matrix will lead, on average, to a given number of extra steps. Randomization of either another character or another cell in the matrix will, given a moderately large data set, usually be almost independent (i.e., it will affect a different subset of taxa) and therefore the effect on tree length will almost be the same as modifying the first character.

The arcsine-ILD is based on the above end constraints of the relationship between tree length and noise described above, which suggest that the curve describing the length-noise relationship will be a close approximation to a nonrectangular hyperbola, and although not strictly hyperbolic (because at the extremes of noise, the length of the most parsimonious tree actually reaches the asymptote), the departure from the ideal is small and, as illustrated below, can probably be ignored. An appropriate linearizing transformation of this form of curve is obtained by scaling length on a 0-to-1 scale and taking the arcsine (i.e., sin⁻¹). Using simulations we demonstrate that this is indeed the case and that by using the arcsine-transformed lengths instead of raw lengths in the ILD test, the results are more in tune with the desired properties of the test both when data are replaced by noise or noise is added to one otherwise unperturbed partition.

**METHODS**

We used a variety of tree sizes to test these methods, ranging from 8-taxon trees to 256-taxon trees. In each case, using PAUP* (Swofford, 2001), we generated 15 pairs of trees using the equiprobable model, which dictates that every possible tree has the same probability. Each tree was represented as a matrix such that a single unambiguous binary character supported each node in the tree; thus, a 32-taxon tree had 29 characters.

**Adding Noise**

In order to investigate the effect of noise on the length, retention index (RI), and upon the outcome of the ILD tests we took each matrix (i.e., tree) and generated increasingly noisy versions of them. We added noise using two methods. The first method, which we term “replacement noise,” involved randomly selecting a single character within the matrix and randomly shuffling (permuting) the binary states within that character in such a way as to ensure that the number of cells with each state remained the same as in the unshuffled character. We added increasing amounts of noise by randomly selecting and shuffling states in additional characters. Using this protocol we could generate matrices with several levels of noise, ranging from 0% to 100% noisy. Permuting whole characters in this way is a mild and precisely quantifiable way of introducing noise, compared with, for example, randomly picking single cells or pairs of cells, since the latter results in multiple characters being affected even when the total number of cells permuted in the matrix is just the same as the number of taxa.

The second method, which we term “addition noise,” involved adding noisy characters to the matrix (rather than replacing them). We selected characters at random, and shuffled the binary states within the characters to create a noisy character. We then added the resulting noisy character to the original matrix, thereby diluting the original signal. We added noise so that the original signal, which with no noise made up 100% of the matrix, was diluted incrementally by the addition of noisy characters until it made up only 0.5% of the matrix. Thus, we
ended up with a range of data sets that were between 0% and 99.5% noisy. This process results in very large data sets and, therefore, we restricted this method of adding noise to the 32-taxon trees.

Estimating Tree Length and Retention Indices

We estimated tree lengths and retention indices (RIs) for each tree by carrying out maximum parsimony heuristic tree searches in PAUP* with 20 random additions holding no more than 10 trees at any one time. We used 1-RI rather than RI in the analyses in order to force the index to increase with increasing homoplasy.

Then we calculated a standardized length and standardized 1-RI value for each tree so that the values ranged from 0 to 1. To standardize tree length (or 1-RI) on this scale, it was first necessary to estimate the asymptotic tree length (or 1-RI) with maximum noise for a particular tree size. In order to do this we created 200 matrices with the states within all characters shuffled (i.e., 100% noise [or 99.5% noise with addition noise]), and subjected each tree to parsimony analysis. We then calculated the standardized length as (observed length – minimum possible length)/(mean length of trees from maximally noisy data – minimum possible length). The calculations for 1-RI were analogous.

ILD Tests

We carried out ILD tests on pairs of matrices (or data partitions). The matrices (trees) under comparison were either exact duplicates (i.e., they were completely congruent) or were different (i.e., they were highly incongruent). We first compared non-noisy partitions, and then we repeatedly implemented the ILD test to compare the non-noisy partition with each of the 20 replicates for each noise level. We did this for both congruent and incongruent matrices. The ILD test was deemed to show a significant difference between partitions if more than 5% of the ILD replicates were longer (or had a greater 1-RI) than the original unshuffled matrix. We repeated these ILD tests using the arcsine-transformed values for length and 1-RI. Analogous to working with raw lengths, as in the standard ILD test (REF), the arcsined values were weighted by the proportion of characters in the relevant partition. For replacement noise, this effects no difference as both partitions were of equal size, but with addition noise, the partition sizes differed. Thus, for example, with 99.5% noise the second partition for a 32-taxon tree had 5800 characters, whereas the first unpermuted partition had 29 characters. The arcsines of the standardized length for each partition were multiplied by 5800/5829 and 29/5829, respectively. The P-value of the ILD test was estimated from the proportion of ILD replicates where the summed metric (length or 1-RI or the weighted arcsined values for these) was larger than the metrics calculated from the original unpermuted matrices.

Effect of Noise Imbalance

Shape of the relationships.—We explored the relationship between the amount of noise added to one of the trees and (i) the length; (ii) the arcsine-transformed standardized length; (iii) the 1-RI; and (iv) the arcsine-transformed standardized 1-RI. We estimated means and standard errors of the mean under each of the “noisiness” conditions for each response variable (standardized length and standardized 1-RI and their arcsine-transformed counterparts). We evaluated whether the data were best described by a linear model or whether there was significant curvature in the relationships by first fitting linear regression models through the data then testing whether the model could be improved by the addition of a quadratic term (which would indicate significant non-linearity). We tested the significance of the addition of the quadratic term by comparing the nested models using the anova() procedure in R (R Development Core Team, 2006).

Significance of ILD tests.—Then, for 32-taxon trees, for each of the four response variables (length, weighted arc sine standardized transformed length, 1-RI, and weighted arc sine-transformed standardized 1-RI), we graphically explored the relationship between the amount of noise added to one of the matrix pairs and (i) the proportion of cases where the ILD test indicated that the partitions (trees) were significantly different and (ii) the mean P-values that were calculated by the ILD tests.

Generality of the Method with Large Trees

In order to evaluate the generality of this approach with larger trees, we investigated the effect of the arcsine standardized transformation on the metrics for trees of sizes ranging from 8 to 256 taxa. In this case we were only interested in whether the arcsine standardized transformation of the response variables linearized their relationship with increasing replacement noise. For each simulation of the effect of adding noise, we incrementally added between 0% and 100% noise to the partition.

The permuted nexus files and PAUP* (Swofford, 2001) command lines used for these analyses were created, and output log files were parsed, using R programs (available from the senior author on request). The statistical analyses were also carried out in R (version 2.3.1, R Development Core Team, 2006; on a Macintosh computer).

RESULTS

Tree Length versus Noise

With increasing numbers of whole characters permuted, the standardized tree length showed the predicted pattern, initially climbing rapidly and nearly linearly as more characters were permuted, but leveling off as the number of permuted characters approached the total number of characters (Fig. 1; left panel, open symbols). This pattern was consistent for all sizes of trees tested here.
FIGURE 1. The relationship between proportion noise and standardized length (left) and standardized 1-RI (right) for trees ranging in size from 8 to 256 taxa, with replacement noise added as described Methods. Filled symbols represent the arcsine-transformed data, whereas open symbols represent the untransformed data. The points have been jittered on the x-axis. The black line shows the theoretical relationship between noise and the arcsine-transformed data and runs from a noise proportion of 0.0 to 1.0 with a slope of arcsine (1). Although the results from one tree of each size are presented here, this analysis have been carried out with other trees and qualitatively identical results were found.
The relationship between arcsine-transformed standardized length and replacement noise appears essentially linear (Fig. 1; left panel, filled symbols). Again, this pattern was consistent for the range of tree sizes considered here. Linearity towards the right of the graph is somewhat sensitive to the estimate of the mean asymptotic metric (length or 1-RI) of fully permuted data, and therefore it is desirable to carry out at least 100, and ideally more, randomizations to generate this estimate (we carried out 200). Further, with these simulations there is a problem with the arcsine calculation at very high noise levels because some randomizations will give lengths greater than the estimated mean and therefore standardized lengths greater than 1, for which nonimaginary arcsines do not exist, and this results in the bend in the arcsine-transformed standardized length relationship when most of the characters are randomized. This was particularly evident for the smaller tree sizes.

The linearization of standardized 1-RI was qualitatively very similar to the linearization of standardized length (Fig. 1; right panel).

**ILD Test Results**

**Congruent partitions with replacement noise.**—For pairs of congruent partitions, the proportion of unmodified ILD tests that are significant increases with the amount of noise imbalance (Fig. 2: left panel, open circles). Initially, when both pairs of matrices are pure signal, the ILD test never indicates significance. However, as one of the pair is rendered increasingly noisy, the proportion of spuriously significant ILD tests increases. The rate of increase is high for raw length data and, with one of the matrices rendered 50% noisy, ~50% of the ILD tests are found (incorrectly) to be in conflict. However, if the ILD tests are performed using arcsine-transformed standardized length (rather than raw length), the rate of increase in spurious significance is dramatically reduced. For example, with a pair of matrices that includes an ~80% noisy partition, using raw data, 100% of the ILD tests indicate a spurious significant difference between partitions. However, if arcsine-transformed data are used, less than 10% of the tests are spuriously significant. The mean $P$-values of the tests follow an analogous pattern with the $P$-values decreasing more rapidly with increasing noise imbalance for raw length data than for arcsine-transformed data (Fig. 2: grey symbols). The results for 1-RI show qualitatively similar patterns (Fig. 3).

**Figure 2.** Results of ILD test and arcsine-ILD test using tree length for pairs of data partitions that support either congruent (left panel) or incongruent (right panel) random trees under different replacement noise conditions. Noise was added to the second partition by shuffling randomly selected characters in the second partition. The error bars represent the range for the $P$-value.
Congruent partitions with addition noise.—With increasing addition noise, initially congruent partitions remain congruent until the noisy partition contains more than 90% noisy characters, whereupon some test results were first found to show significant conflict (Fig. 4; left panel, filled circles). Arcsine transformation of the standardized length had little effect, though curiously, at the highest noise level a larger proportion of tests were found to be significant. The P-values for these tests (Fig. 4; grey symbols) mirror these results.

The ILD test results obtained using 1-RI were qualitatively similar, although a small proportion of the tests...
indicated significant differences at intermediate noise levels (Fig. 5; left panel). Only at very high noise levels did the results from the raw metric differ from the transformed metric, with more than 80% of tests suggesting significant conflict (Fig. 5; open circles) compared to 20% of tests suggesting significant conflict using the transformed metric (Fig. 5; closed circles). Again, the P-values mirrored these results (Fig. 5; grey symbols).

Incongruent partitions with replacement noise.—As expected, for pairs of incongruent partitions, the proportion of significant ILD tests tends to decrease with increasing noise imbalance (Fig. 2: right panel). Initially, with no noise imbalance, all of the ILD tests correctly indicate significant incongruence between partitions. However, as the amount of noise imbalance increases, the proportion of significant ILD tests does not decrease, and 100% of tests indicate significant incongruence when all characters have been permuted (Fig. 2; right panel, open circles).

When the ILD tests are carried out using arcsine-transformed standardized length, increasing noise results in a decrease in the proportion of significant test results (Fig. 2; right panel, filled circles).

As above, the P-values associated with these ILD tests reflect the same pattern (Fig. 2; right panel, grey symbols). The results of the ILD tests using 1-RI (Fig. 3; right panel) were virtually identical.

Incongruent partitions with addition noise.—The results of the ILD tests obtained using raw lengths showed that increasing noise starts to lead to a loss of detected incongruence above 50% noise. With an almost completely noisy second partition, close to 0% of tests were significant, even though the original 29 incongruent characters are still present in the partition (Fig. 4; right panel, open circles). Tests based upon weighted and arcsine-transformed standardized length provided qualitatively similar results; only the rate of decline in the proportion of significant tests was lower (Fig. 4; right panel, closed circles) and even with 99.5% noise only about 50% of the tests were nonsignificant.

The results using 1-RI generally showed greater robustness to the addition of noise with significant incongruence being detected until partition two contained more than 70% noisy characters (Fig. 5; right panel, open circles). However, arcsine-transforming the data resulted in a lower proportion of incongruent test results at the three highest noise levels (although only marginally at ~80% noise; Fig. 5; right panel, filled circles).

**DISCUSSION**

Desirable properties of using arcsine-transformed metrics when implementing the ILD test were apparent with both addition and replacement noise under most circumstances. These two types of noise led to qualitatively different outcomes in that the response to addition noise, because it preserves the initial signal in the partition, was more robust to increasing noise than was the response to replacement noise. With replacement noise, spurious incongruence was detected by the ILD test with moderate amounts (~50%) of initially congruent signal characters replaced by noise (Fig. 2; left panel). However, with addition noise, significant incongruence was not apparent until the second partition was made extremely noisy. Therefore, the detection of apparently significant conflict by the ILD test as a result of noise is likely to
be a greater problem when signal is weak rather than depending on the amount of noise per se.

With incongruent partitions, significant conflict is detected by the ILD test using raw length when one partition comprised pure signal and the other is completely noisy, as in the case of replacement noise (Fig. 2; right panel). Thus, the unmodified ILD test does reveal significant differences when the data are not drawn from the same distribution (Farris et al., 1994), but this does not necessarily indicate that the signals are different: it could be that only the noise in the two partitions differ. This appears to be because the relationship between noise and tree length is nonlinear and, therefore, when the partitions differ in the amount of noise the standard ILD test metric (sums of lengths) is not additive (Dolphin et al., 2000).

Our results show that by arcsine-transforming the standardized length (or 1-RI), this new metric shows a far more linear relationship with noise (Fig. 1). We then show that the results obtained using this metric in the ILD test overcome most of the detection of apparent incongruence, even when one partition is completely randomly

When the initial signal in the second partition is preserved, as in addition noise, the behavior of the ILD test is less at odds with the desired outcome in that initially congruent partitions are not wrongly detected as being incongruent until the amount of added noise is quite extreme (Fig. 4; left panel), though with initially incongruent partitions, the addition of noise leads to the loss of detected incongruence at relatively moderate levels of noise (Fig. 4; right panel).

By substituting the arcsine of standardized and weighted tree metric instead of the raw tree metric as the value used in the calculations in the ILD test, little difference is apparent when the partitions are initially congruent, while a substantial improvement is made in the detection of incongruence in the light of moderate to high levels of added noise (Fig. 4). Implementing the new test requires only that a reasonable estimate is made of the mean value of the tree metrics from each partition with all characters permuted, so that the metrics can be standardized prior to taking the arcsine.

We therefore propose that our modification of the ILD test (arcsine-ILD) provides at least a partial solution to some of the anomalous behavior of the test that has previously been the subject of contention. The R code required to implement, in conjunction with PAUP*, the arcsine-ILD test, is available from http://systematicbiology.org or on request from the senior author and will be published elsewhere.

ACKNOWLEDGEMENTS

We wish to thank Andy Purvis and David Orme for useful discussions.

REFERENCES


First submitted 11 March 2006; final acceptance 6 March 2007
Associate Editor: Frank Anderson