Points of View

Syst. Biol. 51(3):524–527, 2002
DOI: 10.1080/10635150290069931

Type 1 Error Rates of the Parsimony Permutation Tail Probability Test

MARK WILKINSON,[1] PEDRO R. PERES-NETO,[2] PETER G. FOSTER,[1]
AND CLIVE B. MONCRIEFF[1]

[1]*Department of Zoology, The Natural History Museum, London SW7 5BD, UK; E-mail: marw@nhm.ac.uk*
[2]*Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada*

Archie (1989) and Faith and Cranston (1991) independently developed a parsimony-based randomization test for assessing the quality of a phylogenetic data matrix. Matrix randomization tests have had a mixed reception from phylogeneticists (e.g., Källersjö et al., 1992; Alroy, 1994; Carpenter et al., 1998; Wilkinson, 1998; Siddall, 2001). In general, however, these are well-founded statistical techniques (Manly, 1991) that may be well-suited to phylogenetic contexts where models or assumptions underlying parametric statistical methods are either difficult to justify or to test. In a matrix randomization test, a test statistic (typically a measure of data "quality") is calculated for the original data, and the result is contrasted against a null distribution of the test statistic determined by repeated randomization of the data. Randomization is by random permutation of the assignment of character states to taxa within each character. Essentially, each character in the dataset is independently shuffled so that congruence between characters is reduced to the extent that would be expected by chance alone. The random permutation preserves some features of the data that are known to affect measures of data quality, such as the total number of characters and taxa and the numbers of taxa with each character state within each character (Archie, 1989; Sanderson and Donoghue, 1989; Faith and Cranston, 1991). Thus the null distribution represents a distribution that one would expect from comparable phylogenetically uninformative data. The simplest parsimony-based matrix randomization tests use the length of the most-parsimonious trees (MPTs) as the test statistic, comparing this for real and randomly permuted data. A corresponding simple test statistic for the null hypothesis that the data are indistinguishable from random is the parsimony permutation tail probability or parsimony PTP (Faith and Cranston, 1991). The parsimony PTP is the proportion of data sets (real and randomly permuted) that yield MPTs as short or shorter than the MPTs for the original data.

Slowinski and Crother (1998) used 40 real data sets in an empirical evaluation of the utility of the parsimony PTP. Specifically, they compared PTPs with the fraction of clades supported by bootstrap proportions exceeding 50%. In addition, they compared PTPs with the resolution of strict component consensus trees. They reported that data sets that appear to be poorly structured, based on bootstrap analyses or because they have a poorly resolved strict component consensus, tend to have significant PTPs, and they concluded that (p. 300) "the PTP test is too liberal" and is of limited utility. Peres-Neto and Marques (2000) expressed concern at the use of one statistical test (the bootstrap) to evaluate another (parsimony PTP) and presented simulation studies that attempted to address the performance of the PTP test more directly. Their simulation studies involved performing PTP tests on randomly generated data. Because data are generated randomly, the null hypothesis is true and the number of times that the null hypothesis is rejected

correctly estimates the Type 1 error rate of the PTP test, that is, the probability of wrongly rejecting the null hypothesis when it is true. On the basis of their simulations, Peres-Neto and Marques (2000) reported unacceptably high Type 1 error rates for the parsimony PTP (e.g., >0.4 for nominal $\alpha = 0.05$). These results, if correct, would undermine the utility of this parsimony randomization test and led Peres-Neto and Marques (2000:423) to suggest, "Perhaps it is time to propose new tools for assessing character covariation in phylogenetic data." However, we discovered a mistake in the code they used to generate the "random" data, which invalidates the results of their study. Here we report more accurate Type 1 error rates for the parsimony PTP test, estimated by the simulation method of Peres-Neto and Marques (2000), for the range of parameter combinations they originally considered. Our results indicate that the parsimony PTP is a conservative test of the null hypothesis, thus underlining the potential utility of the test in phylogenetics.

Materials and Methods

Type 1 error rates were estimated by using the simulation approach of Peres-Neto and Marques (2000). For a given set of parameters, multiple data sets were generated randomly and tested, and the proportion of tests yielding results significant at $\alpha = 0.05$ was determined. The range of parameters used in the simulations followed that of Peres-Neto and Marques (2000). In the first set of simulations, 200 random binary data states were generated, with the two states (0 and 1) being equally likely for each of two values of the number of characters (40 and 80) and six values of the number of terminal taxa (increments of 5 up to 30). An all-zero outgroup was added in each random matrix, and a parsimony PTP test was performed using 1,000 permutations. The outgroup was included in tree length estimation but not in the permutation. The second set of simulations differed only in the random data consisting of four equiprobable states (0, 1, 2, 3) that were treated as unordered. The third set of simulations explored unequal frequencies of character states, using a larger sample (1,000) of randomly generated binary data sets of 40 characters for two cases, equiprobability of the states, and probabilities of 0.65 and 0.35 for states 0 and 1, respectively.

Simulations were carried out in two ways. First they were run with a corrected version of the software employed by Peres-Neto and Marques (2000), which used Hennig 86 (Farris, 1988) to perform parsimony analysis; for this, we used the exact (ie) algorithm. They were also run with independently developed software using PAUP* (Swofford, 1998) to perform the parsimony PTP tests with heuristic searches (10 random addition sequences and TBR branch swapping). All simulations were replicated by using both systems.

The addition of an all-zero outgroup to the randomly generated data sets was intended solely to emulate the original study; we do not consider this an essential part of the simulation process. Given that the "ingroup" data were randomly permuted, the unpermuted outgroup would always be random with respect to the ingroup. Simulations performed without the addition of an outgroup yield similar results (not shown).

Results

Results from the parallel tests using different software were concordant and have been combined to increase sample size. Type 1 error rates of the parsimony PTP, the proportions of tests of randomly generated data yielding significant results (PTP ≤ 0.05), for each of the three sets of simulations are shown in Table 1. The Type 1 error rates are

TABLE 1. Type 1 error rates of the parsimony PTP measured as the proportions of trials of randomly generated data sets yielding PTPs ≤ 0.05 by using (A) binary characters and equiprobable character states, (B) four-state characters and equiprobable character states, and (C) 40 binary characters and differing probabilities of states 0 and 1. (A) and (B) used 400 trials; (C) used 2,000. These results correspond to those reported in Figures 1–3 of Peres-Neto and Marques (2000).

	No. of characters		No. of characters		Probability of state 1	
Taxa	40	80	40	80	0.5	0.35
5	0.045	0.018	0.013	0.038	0.039	0.031
10	0.033	0.050	0.025	0.038	0.037	0.037
15	0.038	0.033	0.033	0.045	0.035	0.028
20	0.028	0.038	0.035	0.038	0.030	0.036
25	0.053	0.063	0.045	0.040	0.043	0.049
30	0.030	0.045	0.073	0.048	0.033	0.033

generally low, in contrast to the high values reported previously. Indeed, it is striking that the rates are lower than the expectation of 5% in the large majority of the simulations. There are no obvious differences in the performance of the test across the sampled combinations of parameters.

DISCUSSION

The results previously reported by Peres-Neto and Marques (2000) appear to support a pessimistic or even nihilistic view of the parsimony PTP test. However, that view is illusory and results from an unfortunate error in the code used to generate random data in their study. Briefly, the error made the probability of assigning a particular character state contingent on the state previously generated. Character data were generated for each "species" in turn creating sequences of states that were often more similar or more different between species than one would expect by pure chance. The effect of the error would be expected to increase with numbers of species, which explains the finding that rejection of the null hypotheses became easier as the number of species increased (Peres-Neto and Marques, 2000:Figs. 1 and 2). This also explains the reported increase in Type 1 error rates with an increase in the proportion of one of the character states, which increased the chance of generating similar sequences of character states.

In direct contrast to the original results, our estimates indicate that the parsimony PTP is a relatively conservative test statistic. Over a range of numbers of taxa, characters, character states, and relative proportions of character states the Type 1 error rate is mostly <5%. In only 3 of the 36 simulations did it exceed this error rate, and in each case only marginally so. In none of the 12 simulations using 2,000 trials did the error rate exceed 5%, thus suggesting that the three outliers are attributable to sampling error.

Our results demonstrate that the parsimony PTP test cannot be considered too liberal because of any unacceptably high Type 1 error rate. The present results also make more intuitive sense than those reported in the original study. Indeed, the error in the original study was discovered because the current authors were unable to conceive of any obvious mechanism that would account for the reported results. We have no reason to expect truly random data to generate anything greater than the 5% Type 1 error rate when the statistical test is based on random permutation. This is because what constitutes the "real" data is simply a random choice from the set of all its possible permutations.

An interesting aspect of our results that demands explanation is the relatively low error rates. The reason for this deviation from expectations is explained by the discontinuous nature of the distribution of the test statistic. Because parsimony tree lengths are not continuous, there is no need for a clear break between the shortest 5% of the tree lengths and the longest 95% of the tree lengths. Rather, some tree lengths may be clustered on the threshold such that<5% of the tree lengths will be shorter than the threshold. In such cases, the observed tree length would need to be among these (<5%) shortest tree lengths to be significant, the Type 1 error rate will be <5%, and the test will be conservative.

To assess this possibility we developed a revised version of the test that expressly accounts for this source of error. A test on discrete data is not fundamentally distinct from a test based on continuous data grouped into discrete bundles. Imagine that members of the bundle clustered on the threshold all have "true" test value (based on an underlying continuous distribution) that range uniformly from the value associated with a "conservative" test to that associated with a "liberal" test (i.e., including and excluding the entire bundle). The midpoint of this range now corresponds to the threshold, so that one half of the members of the bundle are considered to be among the 5% of values in the tail of the overall distribution. When we applied this method of calculating the test statistic, the previous "conservative" nature of the test results disappeared entirely, and the test statistics nominally set at $\alpha = 0.05$ clustered very close to the nominal value. Indeed, the whole distribution of the test-statistic approximated very well to the expected uniform distribution associated with P-values. Note that we would expect a matrix randomization test using test statistics better approximated by a continuous distribution (such as

log-likelihoods) to be almost unbiased (i.e., not conservative).

Matrix randomization tests such as the parsimony PTP seem to have two uses or interpretations. In the first, the emphasis is on the failure to reject the null hypothesis as a justification for discounting phylogenetic relationships based on the data. Passing the test is seen as a minimum requirement of data if one is to invest any confidence at all in the phylogenetic relationships inferred from it (e.g., Alroy, 1994; Wilkinson, 1997, 1998). In the second, the emphasis is on passing the test as justification for placing confidence in the results of what phylogenetic inferences are based on the data (e.g., Lee, 2001). The first approach is the more conservative and, we believe, more reasonable. It is not known how much phylogenetic signal is required of data for them to pass the parsimony PTP test or whether this level of signal is sufficient for accurate phylogenies to be expected. Type 2 error rates remain unexplored. However, Slowinski and Crother's (1998) comparisons with bootstrapping suggest that passing the parsimony PTP cannot generally be assumed to guarantee well-supported phylogenetic hypotheses. Certainly, phylogenetic signals may not be uniformly distributed across a data matrix, and the fact that a given data matrix passes the test does not entail that subsets of it would similarly pass the test (Faith and Cranston, 1991; Fu and Murphy, 1999). In addition, many data sets with nonphylogenetic (but nonrandom) structure are likely to pass the test (Källersjö et al., 1992; Alroy, 1994). Thus we cannot reasonably infer that the data passing a PTP test support well-founded inferences or even are "phylogenetically well structured" (e.g., Lee, 2001). Other approaches should be used to investigate the strength of support for relationships inferred from data that have passed the parsimony PTP test. From a conservative perspective, where avoiding ill-founded hypotheses of relationships is deemed most important, the possibility that the parsimony PTP may be slightly conservative is not a problem. Given that the use of the PTP is to protect us from poorly founded inferences, the low Type 1 error rate simply means that a greater degree of protection than the nominal 5% is being provided, although this might also imply that the test has lower power. If the conservativeness of the test is considered problematic, then the revised version of the test we have described can be used.

ACKNOWLEDGMENTS

Thanks are due to Joe Thorley for enlightening discussion of the importance of tree length distributions being discontinuous and to Mike Sanderson and an anonymous reviewer for suggesting improvements to the manuscript. P. R. P.-N. was funded by a CNP-q fellowship.

REFERENCES

ALROY, J. 1994. Four permutation tests for the presence of phylogenetic structure. Syst. Biol. 43:430–437.

ARCHIE, J. W. 1989. A randomisation test for phylogenetic information in systematic data. Syst. Zool. 38:239–252.

CARPENTER, J. M., P. A. GOLOBOFF, AND J. S. FARRIS. 1998. PTP is meaningless, T-PTP is contradictory: A reply to Trueman. Cladistics 14:105–116.

FAITH, D. P., AND P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics 7:1–28.

FARRIS, J. S. 1988. Hennig86, version 1.5. Distributed by the author, Port Jefferson Station, NewYork.

FU, J. Z., AND R. W. MURPHY. 1999. Discriminating and locating character covariance: An application of permutation tail probability (PTP) analysis. Syst. Biol. 48:380–395.

KÄLLERSJÖ, M., J. S. FARRIS, A. G. KLUGE, AND C. BULT. 1992. Skewness and permutation. Cladistics 8:275–287.

LEE, M. S. Y. 2001. Molecules, morphology and the monophyly of diapsid reptiles. Contrib. Zool. 70:1–22.

MANLY, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.

PERES-NETO, P. R., AND F. MARQUES. 2000. When are random data not random, or is the PTP test useful? Cladistics 16:420–424.

SANDERSON, M. J., AND M. J. DONOGHUE. 1989. Patterns of variation in levels of homoplasy. Evolution 43:1781–1795.

SIDDALL, M. E. 2001. Computer-intensive randomization in systematics. Cladistics 17:S35–S52.

SLOWINSKI, J. B., AND B. I. CROTHER. 1998. Is the PTP test useful? Cladistics 14:297–302.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (and other methods. Version 4.0. Sinauer Associates, Sunderland, Massachusetts.

WILKINSON, M. 1997. On phylogenetic relationships within *Dendrotriton* (Amphibia: Caudata: Plethodontidae): Is there sufficient evidence? Herpetol. J. 7:55–65.

WILKINSON, M. 1998. Split support and split conflict randomization tests in phylogenetic inference. Syst. Biol. 47:673–695.

First submitted 19 July 2001; revision submitted 17 December 2001; final acceptance 17 December 2002

Associate Editor: Mike Sanderson