doi:10.1111/j.1558-5646.2010.00973.x

POOR STATISTICAL PERFORMANCE OF THE MANTEL TEST IN PHYLOGENETIC COMPARATIVE ANALYSES

Luke J. Harmon[1,2] and Richard E. Glor[3]

[1]*Department of Biological Sciences, University of Idaho, Moscow, Idaho, 83844*

[2]*E-mail: lukeh@uidaho.edu*

[3]*Department of Biology, RC Box 270211, University of Rochester, Rochester, New York, 14627*

Received May 21, 2008
Accepted January 21, 2010

The Mantel test, based on comparisons of distance matrices, is commonly employed in comparative biology, but its statistical properties in this context are unknown. Here, we evaluate the performance of the Mantel test for two applications in comparative biology: testing for phylogenetic signal, and testing for an evolutionary correlation between two characters. We find that the Mantel test has poor performance compared to alternative methods, including low power and, under some circumstances, inflated type-I error. We identify a remedy for the inflated type-I error of three-way Mantel tests using phylogenetic permutations; however, this test still has considerably lower power than independent contrasts. We recommend that use of the Mantel test should be restricted to cases in which data can only be expressed as pairwise distances among taxa.

KEYWORDS: **Comparative methods, independent contrasts, phylogenetic signal, statistical power, type-I error.**

A wide variety of methods are now available for testing comparative hypotheses. One method with a long history of use and a broad range of potential applications is the Mantel test (Mantel 1967). Here, we consider the application of this test to two fundamental questions in comparative biology: (1) Does a trait exhibit "phylogenetic signal" (i.e., are related taxa more similar than expected by random chance; Cubo et al. 2005; Davis 2005; Alexander et al. 2006; Böhning-Gaese et al. 2006; Clabaut et al. 2007)?, and (2) Are two or more characters correlated after controlling for phylogenetic relatedness (Guill et al. 2003; Ossi and Kamilar 2006; Sanders et al. 2006; Cofre et al. 2007)? In both cases, the Mantel test is one of several methods available to comparative biologists. For this reason, it is important to assess its statistical performance relative to these alternatives. We use a series of simulations to show that the Mantel test is generally inferior to two alternatives, Blomberg's K-statistic for tests of phylogenetic signal and phylogenetic independent contrasts (PIC) for tests of character correlations.

The standard Mantel test asks whether two matrices of pairwise distance data are correlated (Mantel 1967). To implement this test, one first computes a test statistic (z) expressing the degree of correlation between the original data matrices. One must then determine whether this test statistic permits rejection of the null hypothesis that the observed correlation is no greater than expected by chance. Because the elements in a distance matrix are not independent from one another, the required null distribution must be generated by randomly reshuffling one of the original data matrices. To maintain the properties of these distance matrices, rows and columns are shuffled together; for example, if rows one and four are swapped, columns one and four are also exchanged. Normally, rows and columns are reshuffled such that each permutation is equally likely; this assumes that the original observations are independent from each other. The z values obtained from each permutation represent the null distribution. If the two original matrices are significantly correlated, we expect their z value to be significantly more extreme than the distribution of z values

© 2010 The Author(s). Journal compilation © 2010 The Society for the Study of Evolution.

obtained by matrix permutation. This basic version of the Mantel test is used to assess phylogenetic signal by asking whether a matrix of pairwise differences between trait values is correlated with a matrix of phylogenetic distances (typically obtained by calculating patristic distances, the sum of branch lengths separating pairs of species, across a molecular phylogeny; Cubo et al. 2005; Davis 2005; Alexander et al. 2006; Böhning-Gaese et al. 2006; Clabaut et al. 2007). From a phylogenetic perspective, simple permutations of the phylogenetic distance matrix are analogous to reshuffling taxon labels on a fixed topology. If phylogenetic signal is strong, we expect a strong correlation between matrices, indicating that those species separated by relatively shorter phylogenetic distances also tend to be those that exhibit the least character divergence.

Tests of character correlation rely on a variant of the standard Mantel test known as the partial Mantel test (Smouse et al. 1986). This test involves three matrices and asks whether two of these (representing pairwise distances for two characters of interest) are significantly correlated while holding a third (phylogenetic distances) constant. The original version of this method worked by finding residuals from linear regressions of each of the first two matrices from the third, and carrying out a Mantel test on these residual distance matrices (Smouse et al. 1986). Legendre (2000), however, showed that this permutation procedure had poor statistical properties, and suggested permuting the residuals of a null model as long as sample size is greater than 10.

In spite of its popularity, the Mantel test sometimes suffers from low power (i.e., high probability of a type-II error, or not detecting an effect when it is present) and high type-I error (i.e., erroneous rejection of a true null hypothesis) relative to alternative methods (Lapointe and Legendre 1995; Oberrath and Bohning-Gaese 2001; Nunn et al. 2006). Although both of these problems are well established by previous criticisms of the Mantel test, there are reasons to believe that their impact will be particularly profound in the case of phylogenetic applications (Lapointe and Legendre 1995). The behavior of the Mantel test is due to two aspects of the test. First, data are converted to pairwise distances, so that single values in the original data can have a cascading effect on matrix values (see Dutilleul et al. 2000). Second, matrix rows and columns are permuted as if all datapoints were independent, an assumption that is violated in many cases (see Raufaste and Rousset 2001).

Although relatively low power appears to be a general feature of the Mantel test (Legendre 2000), this problem may be unavoidable when analyzing the type of data the Mantel test was originally designed to deal with—namely, data that can only be expressed as pairwise distances (e.g., geographic distances between populations). In many phylogenetic comparative analyses, however, data that need not be expressed as pairwise distances are often converted into such measures for the purpose of conducting Mantel tests (e.g., Alexander et al. 2006; Böhning-Gaese et al. 2006; Ossi and Kamilar 2006; Cofre et al. 2007). We are specifically interested in testing the hypothesis that this practice results in reduced power relative to alternative methods that deal with the data more directly. We test this hypothesis by comparing the performance of the Mantel test to alternative methods of testing phylogenetic signal and character correlation that rely on the use of independent contrasts (ICs).

Relatively high type-I error in the partial Mantel test is another frequently discussed problem (Oden and Sokal 1992; Lapointe and Legendre 1995; Raufaste and Rousset 2001; but see Castellano and Balletto 2002; Rousset 2002; Nunn et al. 2006). In the case of phylogenetic comparative analyses, this error may stem from autocorrelation of matrix elements due to underlying phylogenetic structure (Oden and Sokal 1992; Lapointe and Legendre 1995). To test the hypothesis that phylogenetic nonindependence leads to type-I error in partial Mantel tests, we compare this method to the method for phylogenetically informed permutation introduced by Lapointe and Garland (2001).

To test both of the hypotheses discussed above, we assess the power and type-I error of the Mantel test relative to alternative methods for testing phylogenetic signal and character correlation by conducting simulations in the R framework for statistical computing (R Core Development Team 2009); source code for novel calculations are provided as Supporting information.

Methods and Results

TESTING PHYLOGENETIC SIGNAL

A number of alternatives to the Mantel test are widely used to measure phylogenetic signal (the statistical nonindependence among species trait values due to their phylogenetic relatedness; for discussion of the use of this term, see Revell et al. 2008). We focus on methods using Blomberg et al.'s (2003) K value because they are among the best studied. The K test compares the distribution of ICs, which represent phylogenetically structured comparisons among sets of related species (Felsenstein 1985), to that expected under a Brownian motion model of trait evolution. To test for phylogenetic signal, one must first calculate the K-statistic from the original data, which is the variance of ICs divided by its expectation under a Brownian motion model. The value of K is then compared to a null distribution obtained by randomly assigning traits to the tips in the phylogenetic tree. Unlike the Mantel test, data are not converted into pairwise distances to calculate K.

To compare the power of the Mantel test to that of the K statistic, we used a three-step analysis. First, we grew phylogenetic trees of varying sizes ($n = 5$, 10, 15, 20, 25, 30, 35, 40, 45, and 50 tips) under a pure-birth model with a birth rate of 0.1 using the R package GEIGER (Harmon et al. 2008). Phylogenies grown

to a fixed size end in a speciation event, resulting in two zero-length tip branches that cause problems in comparative analyses and are not representative of real trees. To correct this problem, we grew phylogenies to the desired size, then drew an additional waiting time based on the number of taxa in the tree and the birth rate, and added this amount to each tip branch in the tree. Second, we simulated character evolution across each phylogeny under three models: a nonphylogenetic model in which trait values were drawn independently from a normal distribution with unit variance; a Brownian motion model with a rate parameter $\sigma^2 = 1.0$; and an Ornstein–Uhlenbeck (OU) model, which is similar to Brownian motion but also has a tendency for characters to change toward some particular optimum (see Hansen 1997; Butler and King 2004). The OU model has one parameter, α, that governs the strength of this pull toward the optimum; stronger α means that trait values will be nearer the optimum and show less phylogenetic signal (Blomberg et al. 2003). Importantly, traits evolving under an OU model will share some degree of similarity with closely related species, but that similarity will decay more quickly than in a Brownian motion model as they evolve away from their common ancestor. Third, we tested for phylogenetic signal in each simulated dataset with the standard Mantel test and K.

To obtain the two distance matrices required by the Mantel test, we used squared Euclidean distances between species for the evolved trait and patristic distances between species on the phylogeny. We squared Euclidean distances because these values are expected to increase linearly with time under a Brownian motion model. We assessed the significance of correlations between these matrices via 1000 permutations using standard Mantel permutations implemented in R (scripts provided as Supporting information). We calculated values of Blomberg et al.'s K statistic and assessed their significance using permutations with the R package Picante (Kembel et al. 2009).

To assess type-I error in tests of phylogenetic signal for Mantel tests and the K statistic, we used simulations with no phylogenetic signal (i.e., by drawing independent random numbers from a normal distribution with variance = 1.0). Here, a significant result indicates type-I error. In this case, the Mantel test has appropriately low levels of type-I error (<5% at $\alpha = 0.05$; Fig. 1C). We also calculated power as the proportion of the time that actual phylogenetic signal in the simulations was detected. Results from our Brownian motion simulations show that Mantel tests have substantially less power to detect phylogenetic signal than the K statistic over a range of phylogenetic tree sizes (Fig. 1A). When data are simulated under an Ornstein–Uhlenbeck model, the K statistic again has higher power to detect phylogenetic signal than the Mantel test (Fig. 1B).

TESTING CHARACTER CORRELATION

We addressed the relative performance of the Mantel test in analyses of character correlation in a phylogenetic context by comparing this method to ICs, the longest standing and most widely used method for testing correlations among characters (Felsenstein 1985). Importantly, results from ICs are identical to results using phylogenetic generalized least squares (PGLS) under the assumption of a Brownian motion model of evolution (Grafen 1989).

We carry out two versions of the partial Mantel test for phylogenetic signal. For both, we calculate the Mantel test statistic (z) and compare it to a null distribution based on permutations. The two versions of the Mantel test we employ differ in how they permute rows and columns: (1) randomly, as in standard (S) Mantel tests, or (2) or using the method for phylogenetic permutation (PP) described by Lapointe and Garland (2001). The latter exchanges elements in the matrix according to the amount of branch length separating them in the phylogeny; closely related species are more likely to exchange places than more distantly related species. The method requires setting one parameter, k (note that this is distinct from Blomberg et al's [2003] K statistic), to determine the weighting given to these permutation probabilities. k can range from 1 to ∞. As k gets larger, the model converges to equally likely permutations. We use $k = 1$ in all tests described here. Other values of k give similar results, although the PP test

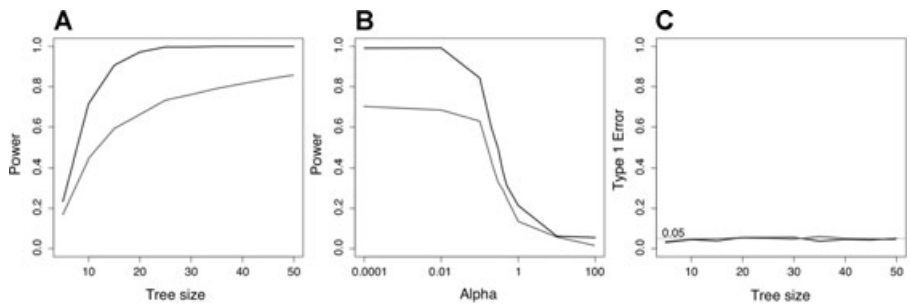

Figure 1. Statistical power (A & B) (1 – type II error) and type-I error (C) for tests of phylogenetic signal using the standard Mantel test (dotted line) and Blomberg et al.'s K statistic (solid line). In each case, 1000 simulated datasets were analyzed for a range of tree sizes (in A and C) or for trees with 25 taxa (B). Data were simulated on the tree under Brownian motion model (A), an Ornstein–Uhlenbeck model for a range of constraint parameter (α) values (B), or in the absence of phylogenetic signal (C).

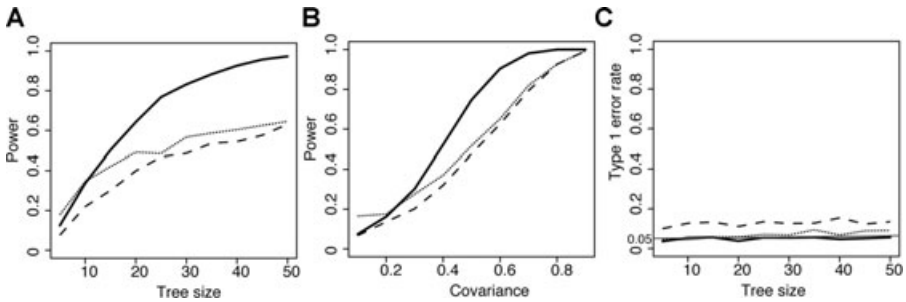

Figure 2. Statistical power (A & B) and type-I error (C) for phylogenetic tests of character correlation using the standard partial Mantel test (dashed line), PP Mantel test (dotted line), and independent contrasts (solid line). In each case, 1000 simulated datasets were analyzed for a range of tree sizes with an expected covariance of 0.5 (in A and C) or for trees with 25 taxa and a range of covariance values (B).

under large values of k converges to the standard Mantel test. To our knowledge, this is the first use of this permutation algorithm for a Mantel test. We have implemented the PP algorithm in a new R function (see Supporting information).

To compare the power and type-I error of these alternative methods, we generated simulations under a range of character covariances ($\sigma_{xy} = 0$–0.9) and tree sizes ($n = 5$, 10, 15, 20, 25, 30, 35, 40, 45, and 50 tips). First, we grew trees of various sizes under a pure-birth process, using the same method as above. Second, we simulated two characters on each tree using a multivariate Brownian motion model (see Revell and Harmon 2008). We created three sets of simulated data: (1) a set of trees of different sizes with no trait covariance ($\sigma_{xy} = 0$) to estimate type-I error rates; (2) a set of trees of different sizes with a moderate level of trait covariance ($\sigma_{xy} = 0.5$) to estimate the power of each method across a range of sample sizes; and (3) a set of trees with 25 taxa across a range of trait covariance values to estimate each method's power to detect trait covariances of different strengths. We calculated ICs using the R package ape (Paradis et al. 2004), with Mantel tests implemented as above.

Over a range of tree sizes, standard partial Mantel tests of character correlation exhibit high type-I error (~20%) relative to both the PP Mantel and IC tests, which behaved appropriately (~5% at $\alpha = 0.05$; Fig. 2C). Both types of Mantel tests suffer from low power relative to IC across a range of tree sizes and character covariances (Fig. 2A,B).

Discussion

The Mantel test is inferior to alternative methods for testing phylogenetic signal and character correlations. In the case of phylogenetic signal, the Mantel test has appropriate type-I error but suffer from markedly lower power than Blomberg et al.'s (2003) K statistic. These results lead us to suggest that simple matrix permutations methods should only be used to test phylogenetic signal when absolutely necessary (i.e., when data cannot be expressed in any form other than pairwise distances). We do not consider another method for permuting matrices, the double permutation developed by Lapointe and Legendre (1995), which acts to randomize both taxon labels and tree topology. Although this method may warrant additional attention, it is still based on the permutation of distance matrices, and our preliminary analyses (not shown) suggest that it does not improve power relative to the K statistic.

The outlook for continued application of the standard partial Mantel test to questions about character correlation in a phylogenetic context is more grim. When applied to such questions, the standard partial Mantel test suffers from both low power and high type-I error. Fortunately, the problem with type-I error appears to result from autocorrelation among species and can be corrected by conducting phylogenetically informed permutations (i.e., the PP version of the Mantel test). These results suggest that the standard partial Mantel test should never be used to test character correlations, but that PP Mantel tests offer a reasonable alternative when data can only be expressed as pairwise distances.

Although the standard Mantel test is an invaluable statistical tool with a wide range of applications in ecology and evolutionary biology (Smouse et al. 1986; Hutchison and Templeton 1999; Nicotra et al. 1999) its use should be avoided when testing two important questions in comparative biology. Mantel tests convert raw data into matrices of pairwise differences among species, an inefficient process that results in low power relative to other methods for testing phylogenetic signal and character correlation. The practice of converting species data into matrices of pairwise distances for the purpose of conducting Mantel tests of any kind should end; matrix-based analyses such as the Mantel test should only be used when the data can only be expressed in the form of pairwise distances. Even then, the standard partial Mantel tests of character correlation remain highly problematic due to their elevated type-I error. If permutation of distances matrices must be used to address questions in phylogenetic comparative biology, we propose using Lapointe and Garland's (2001) PP method for conducting phylogenetic-informed permutations that correct for the data's intrinsic autocorrelation.

ACKNOWLEDGMENTS

We thank J. B. Losos, R. B. Langerhans, J. Sullivan, D. Posada, and three anonymous reviewers for helpful suggestions during the preparation of this manuscript, and S. Kembel for providing R code for analyses.

LITERATURE CITED

Alexander, H. J., J. S. Taylor, S. S. T. Wu, and F. Breden. 2006. Parallel evolution and vicariance in the guppy (*Poecilia reticulata*) over multiple spatial and temporal scales. Evolution 60:2352–2369.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Böhning-Gaese, K., T. Caprano, K. van Ewijk, and M. Veith. 2006. Range size: disentangling current traits and phylogenetic and biogeographic factors. Am. Nat. 167:555–567.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Castellano, S., and E. Balletto. 2002. Is the partial Mantel test inadequate? Evolution 56:1871–1873.

Clabaut, C., P. M. E. Bunje, W. Salzburger, and A. Meyer. 2007. Geometric morphometric analyses provide evidence for the adaptive character of the Tanganyikan cichlid fish radiations. Evolution 61:560–578.

Cofre, H. L., K. Bohning-Gaese, and P. A. Marquet. 2007. Rarity in Chilean forest birds: which ecological and life-history traits matter? Diversity Distrib. 13:203–212.

Cubo, J., F. Ponton, M. Laurin, E. de Margerie, and J. Castanet. 2005. Phylogenetic signal in bone microstructure of sauropsids. Syst. Biol. 54:562–574.

Davis, E. B. 2005. Comparison of climate space and phylogeny of *Marmota* (Mammalia : Rodentia) indicates a connection between evolutionary history and climate preference. Proc. R. Soc. Lond. B. 272:519–526.

Dutilleul, P., J. D. Stockwell, D. Frigon, and P. Legendre. 2000. The Mantel Test versus Pearson's correlation analysis: assessment of the differences for biological and environmental studies. J. Ag. Biol. Env. Stat. 5:131–150.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Grafen, A. 1989. The phylogenetic regression. Phil. Trans. R. Soc B. 326:119–157.

Guill, J. M., D. C. Heins, and C. S. Hood. 2003. The effect of phylogeny on interspecific body shape variation in darters (Pisces : Percidae). Syst. Biol. 52:488–500.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Harmon, L. J., J. Weir, C. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131.

Hutchison, D. W., and A. R. Templeton. 1999. Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. Evolution 53:1898–1914.

Kembel, S. W., D. D. Ackerly, S. P. Blomberg, P. D. Cowan, M. R. Helmus, H. Morlon, and C. O. Webb. 2009. picante: R tools for integrating phylogenies and ecology. R package version 0.7-1. http://picante.r-forge.r-project.org.

Lapointe, F. J., and P. Legendre. 1995. Comparison tests for dendrograms: a comparative evaluation. J. Classif. 12:265–282.

Lapointe, F. J., and T. Garland. 2001. A generalized permutation model for the analysis of cross-species data. J. Classif. 18:109–127.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. J. Stat. Comput. Sim. 67:37–73.

Mantel, N. A. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.

Nicotra, A. B., R. L. Chazdon, and S. V. B. Iriarte. 1999. Spatial heterogeneity of light and woody seedling regeneration in tropical wet forests. Ecology 80:1908–1926.

Nunn, C. L., M. B. Mulder, and S. Langley. 2006. Comparative methods for studying cultural trait evolution: a simulation study. Cross-Cult. Res. 40:177–209.

Oberrath, R., and K. Böhning-Gaese. 2001. The Signed Mantel Test to cope with autocorrelation in comparative analyses. Journal of Applied Statistics 28:725–736.

Oden, N. L., and R. R. Sokal. 1992. An Investigation of Three-Matrix Permutation Tests. J. Classification 9:275–290.

Ossi, K., and J. M. Kamilar. 2006. Environmental and phylogenetic correlates of *Eulemur* behavior and ecology (Primates : Lemuridae). Behav. Ecol. Sociobiol. 61:53–64.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

R Core Development Team. 2009. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raufaste, N., and F. Rousset. 2001. Are partial mantel tests adequate? Evolution 55:1703–1705.

Revell, L. J., and L. J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res. 10:311–321.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Rousset, F. 2002. Partial Mantel tests: reply to Castellano and Balletto. Evolution 56:1874–1875.

Sanders, K. L., A. Malhotra, and R. S. Thorpe. 2006. Evidence for a Mullerian mimetic radiation in Asian pitvipers. Proc. R. Soc. Lond. B 273:1135–1141.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. 35:627–632.

Associate Editor: D. Posada

Supporting Information

The following supporting information is available for this article:

R script for carrying out Mantel tests under the phylogenetic permutation (PP) algorithm.

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.