Syst. Biol. 45(1):27–47, 1996

TESTING HYPOTHESES OF CORRELATED EVOLUTION USING PHYLOGENETICALLY INDEPENDENT CONTRASTS: SENSITIVITY TO DEVIATIONS FROM BROWNIAN MOTION

RAMÓN DÍAZ-URIARTE[1] AND THEODORE GARLAND, JR.[2]

Department of Zoology, University of Wisconsin, Madison, Wisconsin 53706-1381, USA

Abstract.—We examined the statistical performance (in terms of type I error rates) of Felsenstein's (1985, Am. Nat. 125:1–15) comparative method of phylogenetically independent contrasts for testing hypotheses about evolutionary correlations of continuous-valued characters. We simulated data along two different phylogenies, one for 15 species of plethodontid salamanders and the other for 49 species of Carnivora and ungulates. We implemented 15 different models of character evolution, 14 of which deviated from Brownian motion, which is in effect assumed by the method. The models studied included the Ornstein–Uhlenbeck process and punctuated equilibrium (change allowed in only one daughter at each bifurcation) both with and without trends and limits on how far phenotypes could evolve. As has been shown in several previous simulation studies, a nonphylogenetic Pearson correlation of species' mean values yielded inflated type I error rates under most models, including that of simple Brownian motion. Independent contrasts yielded acceptable type I error rates under Brownian motion (and in preliminary studies under slight deviations from this model), but they were inflated under most other models. This new result confirms the model dependence of independent contrasts. However, when branch lengths were checked and transformed, then type I error rates of independent contrasts were reduced. Moreover, the maximum observed type I error rates never exceeded twice the nominal P value at $\alpha = 0.05$. In comparison, the nonphylogenetic correlation tended to yield extremely inflated (and highly variable) type I error rates. These results constitute another demonstration of the general superiority of phylogenetically based statistical methods over nonphylogenetic ones, even under extreme deviations from a Brownian motion model. These results also show the necessity of checking the assumptions of statistical comparative methods and indicate that diagnostic checks and remedial measures can substantially improve the performance of the independent contrasts method. [Comparative method; computer simulation; independent contrasts; phylogeny; hypothesis testing; statistics; correlated evolution.]

If statistical approaches are to be used in analyzing comparative data, then phylogenetically based methods are essential (e.g., Harvey and Pagel, 1991; Losos, 1994; Losos and Miles, 1994; Pagel, 1994a, 1994b; Martins, 1996, in press). When methods that account for phylogenetic topology and branch lengths are not used, the de facto assumption is that the phylogeny of the species being studied is adequately represented as a star (a single hard polytomy, sensu Maddison, 1989) with equal branch lengths. For most comparative studies, a star phylogeny is far less realistic than the best available hypothesis of phylogenetic relationships (Harvey and Pagel, 1991; Pagel and Harvey, 1992; Purvis, 1995). Of the available methods that can be used with any phylogenetic tree (including both soft [reflecting uncertainty] and hard polytomies) and various branch lengths, Felsenstein's (1985) phylogenetically independent contrasts is probably the most frequently used (e.g., see Miles and Dunham, 1993; Garland and Adolph, 1994), and its statistical performance is as good as or better than that of other comparable methods (Grafen, 1989; Martins and Garland, 1991; Pagel, 1993; Purvis et al., 1994; Martins, in press).

The three main assumptions of independent contrasts are (1) a correct topology, (2) branch lengths measured in units of expected variance of character evolution, and (3) a Brownian motion (BM) model of character evolution (Felsenstein, 1985, 1988). These three assumptions allow the computation of phylogenetically independent contrasts that can be used both for statis-

[1] E-mail: rdiaz@macc.wisc.edu.
[2] E-mail: tgarland@macc.wisc.edu.

tical estimation and hypothesis testing. When these three assumptions are correct and the resulting contrasts behave adequately (e.g., for correlations, bivariate normality of contrasts; for regressions, homoscedasticity of residuals), then independent contrasts yield the nominal type I error rates (probability of rejecting the null hypothesis when it is true) for testing the significance of correlations and regressions (Grafen, 1989; Martins and Garland, 1991; Purvis et al., 1994; Martins, in press). When these assumptions are violated, however, type I error rates may deviate from nominal values (e.g., with respect to systematic errors in branch lengths; see Martins and Garland, 1991; Gittleman and Luh, 1992, 1994; Purvis et al., 1994). Inflated type I error rates seriously affect hypothesis testing: they lead to the rejection of the null hypothesis more frequently than specified by the nominal *P* value.

Independent contrasts are easily applied to hard polytomies (Purvis and Garland, 1993), although application to soft polytomies is more complicated (Grafen, 1989, 1992; Harvey and Pagel, 1991; Pagel, 1992; Pagel and Harvey, 1992: see Purvis and Garland, 1993, for summary and clarification). In the face of great topological uncertainty, computer simulation of random phylogenies (and simultaneously of the characters being studied) can be employed to create appropriate null distributions of the test statistic (Losos, 1994; Martins, 1996).

Assumptions 2 and 3 are related in the sense that modifications of branch lengths alone can be used to change the evolutionary model. For example, a BM simulation with all branch lengths set equal is effectively equivalent to a speciational model (assuming that all species [extant and extinct] are included in the analyses; e.g., Rohlf et al., 1990; Martins and Garland, 1991; Kim et al., 1993). Similarly, variation in evolutionary rates can be achieved by differentially altering branch lengths in different parts of the phylogeny. Moreover, a BM simulation in which limits to character evolution are employed is no longer BM (see Garland et al., 1993) and results in some branches (those along which limits are reached) being in error (i.e., they underestimate expected variance of character evolution).

Even if the computations of independent contrasts are based upon the three assumptions listed above, it is not obvious how adversely violations of those assumptions will affect the performance of the method (Felsenstein, 1985, 1988; Martins and Garland, 1991; Miles and Dunham, 1993; Martins, in press). For instance, although BM is a very simple and probably unrealistic model for the evolutionary process, it does not follow that independent contrasts cannot profitably be applied to real data. Brownian motion character evolution may be a sufficient condition for applying independent contrasts, but is it necessary? The key issue is the robustness of the independent contrasts method when (1) the BM model is incorrect and/or (2) the available branch lengths are not reasonable surrogates for expected variance of character change.

Herein, we use computer simulations of character evolution on specified phylogenetic trees to address the effects of deviations from the BM model on the performance of independent contrasts for testing hypotheses about correlated evolution. The alternative evolutionary models employed range from seemingly slight deviations (such as the Ornstein–Uhlenbeck model with a moderate amount of stabilizing selection) to extreme deviations (e.g., punctuated equilibrium, in which character evolution can occur in only one daughter at each cladogenic event) from BM.

In several of the models, the range of possible values for the phenotypic characters was limited. Many if not all phenotypic traits show limits in nature (e.g., Huey and Bennett, 1987; Garland et al., 1993; Bauwens et al., 1995). Limits violate the BM model because as a character approaches a limit the probability of change in one direction (towards the limit) becomes smaller than the probability of change in the other direction (cf. Garland et al., 1993). A heuristic way to visualize part of the effect of a limit is to isolate it

from the violation of the BM model per se and to envision that some of the "real" branch segments are in effect shorter than what they appear to be because the variance is constrained. If we were to use the original branch lengths to compute independent contrasts, then we would not be using branch lengths in units of expected variance of change. An added feature of using simulations with limits is that they produce a nonuniform distortion of the relationship between branch length and expected variance of change, such that the relationship may be different in different parts of the phylogeny (especially near the tips). The net result is nonuniform rates of evolution over the phylogeny as a whole.

As with any statistical methodology, independent contrasts approaches should not be applied blindly. If diagnostic checks of the assumptions are possible, then they should be used and remedial measures taken as appropriate. At least some deviations from a simple BM model may be detectable, and several ways of checking the adequacy of the available "starter" branch lengths have been suggested (Grafen, 1989, 1992; Garland et al., 1991, 1992; see also Martins, 1994). In this study, we investigated the performance of independent contrasts when the true evolutionary model is unknown, as will generally be the case. Garland et al.'s (1992) procedure (also suggested by Pagel, 1992) is similar to suggestions by Purvis and Rambaut (1995 [CAIC user's guide]) and is of general applicability: after standardized independent contrasts have been computed, check for any nonrandom pattern (e.g., linear or nonlinear relationships, differences among clades [Garland, 1992]) in scatterplots of the absolute value of standardized contrasts versus their standard deviations. Patterns in these scatterplots indicate that the branch lengths used are not adequate for standardizing the contrasts. Garland et al. (1992) suggested transforming the branch lengths (or the characters) using a family of powers of the branch lengths plus the log of the branch lengths until a pattern-free scatterplot is obtained (version 2.0 of the PDTREE program of Garland et al. [1993] allows these procedures). We tested whether these branch-length transformations actually do improve the performance of independent contrasts.

METHODS

Models of Character Evolution

To simulate the evolution of two continuous-valued characters along specified phylogenetic trees, we used three basic evolutionary models: BM, Ornstein–Uhlenbeck (OU), and punctuated equilibrium (PE). We used two different phylogenies: one for 49 species of Carnivora and ungulates (Garland et al., 1993: fig. 1) and another for 15 species of salamanders (Sessions and Larson, 1987; as used by Martins and Garland, 1991; Martins, in press).

The BM model is a good approximation of evolution by purely random genetic drift with no selection but may also be appropriate for some forms of selection, such as that caused by randomly fluctuating environmental conditions (Felsenstein, 1985, 1988). It is the simplest of the models considered herein and is implemented in the PDSIMUL program (Garland et al., 1993) by drawing a random change for each branch segment (i.e., $X_{n+1} = X_n + \Delta X$, where X_n is the value of trait X at node n, X_{n+1} is the value of the trait at node n + 1 [i.e., a descendant node], and ΔX is drawn from a normal distribution). The variance of the distribution from which the ΔX values are drawn is made proportional to the length of each branch segment, which results in a gradual model of evolution (Gradual Brownian in PDSIMUL). When the evolution of two characters (X and Y) is simulated together (as in our case), the way to introduce correlation among the changes is to draw the changes (ΔX and ΔY) from a bivariate normal distribution with the desired correlation.

The OU model is an extension of the BM model, in which character changes are affected by a force that pushes them toward some central value (Felsenstein, 1988; Garland et al., 1993; Martins, 1994). Biologically, this model can be thought of as mimicking the movements of a population's

mean phenotype that is simultaneously (1) being pushed towards a selective peak by the action of natural selection and (2) wandering back and forth on its way toward the peak (or about the peak if it is already there) because of random genetic drift; thus, selection acts as a rubber band, tending to return the population mean to the peak. Alternatively, the OU process can be used to model the movement of the adaptive peak itself (e.g., as environmental conditions change, perhaps stochastically, over time), with the population mean always being relatively close to the optimum (Felsenstein, 1988; Martins, 1994).

The PE model is fundamentally different from the other two models in that character change occurs only at speciation events and only in one of the two daughter species (see Eldredge and Gould, 1972; Gould and Eldredge, 1977; implementations by Raup and Gould, 1974; Colwell and Winkler, 1984; Kim et al., 1993; discussion by Martins and Garland, 1991). The implicit assumption of using the PE model here (see also Garland et al., 1993) is that all species in the clade, whether extant or extinct, are included in the data set; otherwise, additional opportunities for character evolution should be allowed, one for each additional speciation event that has occurred along each branch segment. The "punctuational" model of Martins and Garland (1991), Gittleman and Luh (1992, 1994), and Purvis et al. (1994) was simply a BM model on a phylogeny with all branch lengths set equal to unity (see also Huey and Bennett, 1987), i.e., change could occur in both daughters, and is better referred to as a "speciational" model (following Rohlf et al., 1990; Kim et al., 1993: Speciational Brownian in PDSIMUL).

For each of the three basic models, we simulated character evolution both without and with limits. Limits were implemented in two ways using PDSIMUL (Garland et al., 1993). The Replace algorithm checks each time a change can occur (i.e., once for each branch segment) to see if the next added value would take the trait out of bounds. If so, then a new change (simulated value) is used, rechecked to make sure the limit(s) is not exceeded, and so forth. Under the Truncate Change algorithm, if a trait attempts to "evolve" past a boundary it is forced to stop at that boundary, forfeiting the rest of its change. In neither case is the trait stuck at these limits, however, because the next change may take it back in the opposite direction. Examples of the effects of the two limit algorithms on the distribution of tip data are shown in Figure 1.

In all of the above simulations, the specified initial values and final means of the traits were the same: an arbitrary value of 100. In addition, we modeled evolutionary trends by using different initial values and final means (or adaptive peaks for OU). The net effect of this implementation is a movement of the mean value (across all lineages) of the simulated data sets as evolution progresses from the root node to the tip values (see Garland et al., 1993). This implementation was used in simulations with limits for the three models of evolution (BM, OU, PE). An example of the effect of this model on the distribution of the tip values is shown in Figure 2b; the distribution is skewed. We also modeled evolution with shifting means for BM and OU models without limits, but the results (not shown) were exactly the same as those in which the means did not move, as suggested by Felsenstein (1985) and Grafen (1989).

Fifteen models of evolution were employed (Table 1), the first nine with the same starting and ending values (100) and the last six with different starting and ending values (means shifting from 10 at the root node to an expected 190 across tip nodes). These simulations were applied to two different phylogenies, one with 49 and one with 15 species. In all cases with limits, we used values that yielded rather strong effects (e.g., a distribution of tip values that was far from normal; see Figs. 1, 2) because we wanted to be sure to obtain an effect on type I error rates if one existed.

Parameters of the Computer Simulations

The PDSIMUL program simulates bivariate evolution of continuous-valued char-

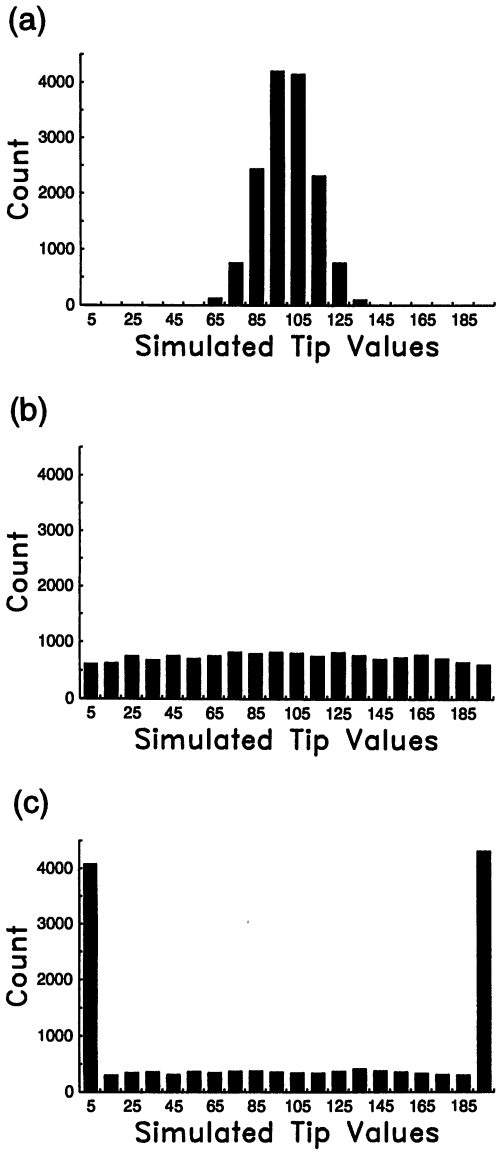

FIGURE 1. Effects of the limits algorithms on the distribution of tip data. Each histogram includes 15,000 tip values, corresponding to 1,000 simulations for the 15-species phylogeny, and represents one of six replicate simulations (Table 1). (a) BM, no limits. (b) BM, strong limits with Replace algorithm in PDSIMUL. (c) BM, strong limits with Truncate Change algorithm. For (b) and (c), upper limit for tip values was 200 and lower limit 0.00000001.

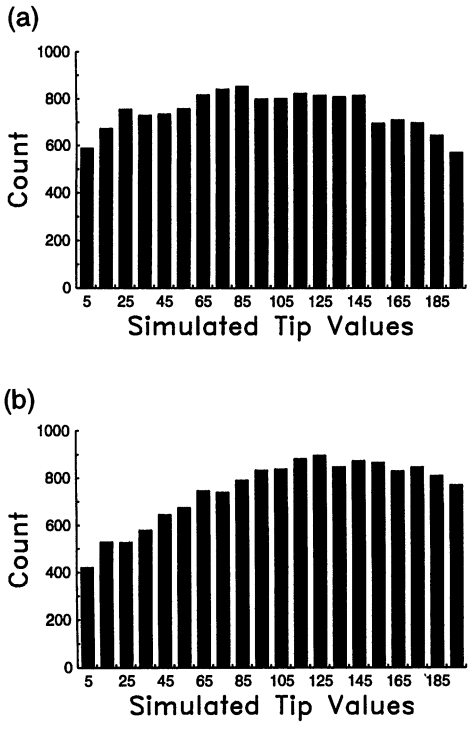

FIGURE 2. Example of effects of the movement of the mean value (mimicking an evolutionary trend) on the distribution of tip data, when limits (upper = 200, lower = 0.00000001) are present. Each histogram includes 15,000 tip values, corresponding to 1,000 simulations on the 15-species phylogeny, and represents one of six replicate simulations (Table 1). (a) OUReplace, no trend. (b) OUReplaceShifting, trend toward larger values.

acters along specified phylogenetic trees. This program allows specification of various things, including (1) the starting values at the root (basal node) of the tree (Initial Values), (2) the correlation of the distribution from which the changes are drawn (Correlation of Input Distribution = input correlation of Martins and Garland [1991]), (3) the desired variances of the distribution of the final tip values (Variances-Tip), and (4) the expected mean of the tip values (Final Means [Adaptive Peak for OU models]). Because we are concerned only with type I error rates in the present study, all input correlations were set equal to 0. In other words, we were interested in how frequently the true null hypothesis (input correlation = 0) was rejected by the

TABLE 1. Description and parameters of the 15 models of evolution used. In the PDSIMUL program (Garland et al., 1993), Variances-Tip denotes the expected variance of simulated data sets in the absence of limits. These variances are given for the 49-species phylogeny (Garland et al., 1993) and the 15-species phylogeny (Sessions and Larson, 1987; Martins and Garland, 1991). In all simulations, the trait values had an upper limit of 200 and a lower limit of 0.00000001. All simulations with means shifting had an Initial Value of 10 and an expected Final Means of 190; simulations without means shifting had initial and final values of 100. The Decay Constant for the OU model was 0.000000028 for the 49-species phylogeny and 0.00000003 for the 15-species phylogeny.

Abbreviation	Description	Variances-Tip	
		49 species	15 species
BM	Brownian motion	100	100
BMReplace	Brownian motion, limits with Replace algorithm	10,000	50,000
BMTruncate	Brownian motion, limits with Truncate algorithm	10,000	50,000
OU	Ornstein–Uhlenbeck	100	100
OUReplace	Ornstein–Uhlenbeck, limits with Replace algorithm	40,000	60,000
OUTruncate	Ornstein–Uhlenbeck, limits with Truncate algorithm	40,000	60,000
PE	punctuated equilibrium	100	100
PEReplace	punctuated equilibrium, limits with Replace algorithm	15,000	25,000
PETruncate	punctuated equilibrium, limits with Truncate algorithm	15,000	25,000
BMReplaceShifting	Brownian motion, limits with Replace algorithm, means shifting	30,000	40,000
BMTruncateShifting	Brownian motion, limits with Truncate algorithm, means shifting	30,000	40,000
OUReplaceShifting	Ornstein–Uhlenbeck, limits with Replace algorithm, means shifting	50,000	60,000
OUTruncateshifting	Ornstein–Uhlenbeck, limits with Truncate algorithm, means shifting	50,000	70,000
PEReplaceShifting	punctuated equilibrium, limits with Replace algorithm, means shifting	60,000	70,000
PETruncateShifting	puncutated equilibrium, limits with Truncate algorithm, means shifting	60,000	70,000

different methods and how this frequency compared with the nominal α level, e.g., 0.05.

For all simulations in which the initial values and expected final means were set to be the same, the value was set to 100 for both traits. For simulations in which the initial values and final means were different, the initial values were set at 10 and the final means at 190, again for both traits. In simulations with limits, the upper limit was 200 and the lower limit was 0.00000001. For the OU model of character evolution, the value of the Decay Constants in PDSIMUL (the strength of the "rubber band" or "spring") was set to 0.000000028 (49-species phylogeny) or 0.00000003 (15-species phylogeny), i.e., about twice the reciprocal of the tree height (see PDSIMUL documentation). We utilized these values because preliminary studies using values equal to the reciprocal of the tree height caused little or no inflation of type I error rates. An important area for future research would be simulations with biologically realistic values of decay constants (cf. Martins, 1994).

The desired variances across the tip values had to be varied when limits were implemented. The variances that the user specifies in PDSIMUL (Variances-Tip) are the variances that the tip distribution of the data would have on average if the simulations were not constrained by limits. For the simulations without limits, desired tip variances were set to 100 for both traits so that no tip value would be >200 or <0.00000001 (see Table 1).

If limits are imposed and simulated data reach those limits, then variances must be increased to yield variances of the tip data similar to those of data sets simulated without limits. As the user-specified variances are increased, the effect of limits becomes progressively stronger; the changes are drawn from distributions with larger

variances, making it more likely that characters will reach the limits at some point during simulated evolution proceeding from root to tips. As the limits are reached, the increase in variance for the set of actual tip values is much slower than the increase in user-specified variance until a point is reached beyond which increases in the value of the user-specified variance no longer produce any increase in the realized tip variance per se. However, the distributions of tip data deviate farther and farther from normal (e.g., see Fig. 1c), which makes it inappropriate to compare the distributions solely in terms of their variances.

For the simulations with limits, we chose very large variances (see Table 1), which resulted in strong effects of the limits. These large variances were chosen because preliminary results indicated that the stronger the limits the higher the type I error rates. Moreover, we found that weak limits, or weak OU decay constants, had little or no effect on type I error rates (results not shown). This implementation is fully equivalent to the alternative of keeping the variance fixed and making the interval between upper and lower limits narrower.

For each combination of (model of evolution) × (limit implementation), we replicated the simulated data set of $n = 1{,}000$ (i.e., each data set consisted of 1,000 simulated evolutionary processes, each involving either 15 or 49 species at the tips of the phylogeny) six times. The parameters of the six simulated data sets were identical within each model × limit, except for the seed of the pseudorandom number generator, which was chosen from a table of random numbers. We chose six replicates as a reasonable compromise between the conflicting desires for logistic ease and high statistical power when comparing the effects of different models.

Analysis of the Simulated Data Sets

The simulated data sets were analyzed with a nonphylogenetic correlation and with three versions of phylogenetically independent contrasts (Table 2). The first of these, an ordinary Pearson product–moment correlation of the 15 or 49 species values (hereinafter referred to as TIPS, following Martins and Garland, 1991), was computed using a version of CMTPANAL (Martins and Garland, 1991) modified to allow batch processing.

TABLE 2. Descriptions of the comparative methods tested and degrees of freedom (df) used in the analyses (n is number of species, either 15 or 49). The degrees of freedom were used to establish the conventional critical values for the correlation coefficients.

Method	Description	df
TIPS	nonphylogenetic Pearson correlation	$n - 2$
IC	Felsenstein's (1985) method of phylogenetically independent contrasts with no branch length transformations; FL1G of Martins and Garland (1991)	$n - 2$
ICblt	Felsenstein's (1985) method after checking for adequate branch length standardization and using branch length transformations if appropriate, as indicated by Garland et al. (1992)	$n - 4$
ICblte	same as for ICblt but checking and excluding those simulated data sets in which the branch length transformations did not achieve an adequate standardization of contrasts; data sets for which the correlation between absolute value of standardized contrasts and their standard deviations (square root of sum of branch lengths) was statistically significant for either one or both traits at $P = 0.05$ (for df $= n - 3$) were regarded as not appropriately standardized (Garland et al., 1992) and were excluded	$n - 4$

For the independent contrasts analyses, we used the new program PDERROR. This program analyzes each simulated data set and produces, among other output, the correlation coefficient (through the origin) as estimated with Felsenstein's (1985) phylogenetically independent contrasts (FL1G of Martins and Garland, 1991; also available in version 2.0 of PDTREE [Garland et al., 1993]). This ordinary application of independent contrasts is referred to as IC.

PDERROR also allows transformations of branch lengths, as proposed by Garland et al. (1992), prior to computing a correla-

tion with independent contrasts. For each trait of each simulated data set, the program examines all transformations of branch lengths obtained by raising the branch lengths to powers ranging from 0 to 2 in intervals of 0.1, plus the log (base 10) of the branch lengths (note that a power of 0 yields all branch lengths equal to unity). For each trait and for each possible branch-length transform, the program computes the Pearson product–moment correlation between the absolute value of the standardized contrasts and their standard deviations and then selects the transform that gives the smallest absolute value of the correlation coefficient (the chosen transformation can be the power of 1, which yields no transformation). This procedure is performed independently for the two traits in each simulated data set. Therefore, the correlation coefficient can be computed for two traits that have been standardized using differently transformed branch lengths, as has been done with some real data sets (e.g., Garland et al., 1992; Garland, 1994). Once branch lengths have been transformed, standardized independent contrasts are computed as usual, and the correlation coefficient is also computed; we denote this method as ICblt (IC with branch length transformations).

Use of the branch length transformations does not guarantee that an appropriate standardization of branch lengths will be achieved. Consequently, for each simulated data set, we also checked whether the transformed branch lengths actually yielded a Pearson product–moment correlation between the absolute values of standardized contrasts and their standard deviations that was not statistically significant at $P = 0.05$ for a two-tailed test with $df = n - 3$, where n = number of species (following the suggestion of Garland et al., 1992) and this correlation is not forced through the origin. We then excluded those simulations for which the correlation remained statistically significant and recomputed type I error rates; we denote this as ICblte (IC with branch length transformation checking and excluding some cases) (Table 2). This method is not really different from ICblt; it is a check of whether it is appropriate to continue the analysis for a given data set (Losos [1994] also used a checking procedure in his simulations; see also Gittleman and Luh, 1994). This checking and exclusion procedure is appropriate because with a real data set an investigator would not proceed with the analysis if assumptions of the statistical method (even after suitable branch length transformations) were clearly violated. Consequently, of the results presented here, those for ICblte may give the best indication of the "real-world" peformance of independent contrasts.

In summary, for each simulated data set we obtained four different distributions of correlation coefficients that can be used to estimate evolutionary correlations between two continuous-valued characters (see Table 2): TIPS (an ordinary nonphylogenetic Pearson correlation), IC (Felsenstein's [1985] method of phylogenetically independent contrasts), ICblt (Felsenstein's method after transforming branch lengths as suggested by Garland et al. [1991, 1992]), and ICblte (ICblt after excluding those simulations in which adequate standardization was not achieved, i.e., one of the statistical assumptions of the method was violated).

Computing and Testing Type I Error Rates

To analyze the performance of each method with regard to hypothesis testing, we calculated its type I error rate (probability of rejecting the null hypothesis when it is true) in two different ways. In both cases, the question asked was whether a given method yields type I error rates significantly different from those expected under standard normal theory when the true correlation between traits is zero (on average).

First, we compared the overall distribution of the 1,000 correlation coefficients estimated by a given method for a given set of simulated data with the theoretical distribution of Pearson correlation coefficients. We determined the number of positive correlation coefficients larger than the

critical values given by the one-tailed distribution of correlation coefficients at α = 0.010, 0.025, 0.050, 0.100, and 0.250 (Zar, 1984: table B.16) plus the number of negative correlation coefficients smaller than the critical values at those α levels. We then calculated the difference between the number of correlation coefficients exceeding critical values for successive α levels (i.e., for positive correlation coefficients for α levels between 0 and 0.01, 0.01 and 0.025, 0.025 and 0.05, 0.05 and 0.10, 0.10 and 0.25, and 0.25 and 1.00 and similarly for negative correlation coefficients). These observed differences were compared with expected differences, based on a standard Pearson's r distribution, by using a chi-square test with df = 11 (for 12 intervals). This method is very similar to that of Martins and Garland (1991), except that we compared the "unfolded" distributions (negative and positive correlation coefficients [with same absolute value] were assigned to different intervals), whereas Martins and Garland (1991) compared "folded" distributions (correlation coefficients were assigned to intervals as a function of their absolute value). Therefore, our procedure is more sensitive to nonsymmetric distributions of correlation coefficients. The unfolded distribution comparison dictates the use of critical values from the one-tailed distribution of correlation coefficients, and they are used for defining the 12 intervals.

Second, we computed the observed frequency of correlation coefficients (of 1,000 total) for a nominal α level of 0.05 (expectation is 50), which is commonly used for hypothesis testing (cf. Grafen, 1989; Purvis et al., 1994). We determined the number of correlation coefficients exceeding the critical value for α = 0.05 in a two-tailed test; thus, in these analyses we lumped together positive and negative correlation coefficients. We then used a binomial test (Conover, 1980) to obtain the P value of each observed frequency (six for each model × phylogeny combination).

This binomial test P value specifically tests for inflated type I error rates, which are of particular interest because in general nonphylogenetic statistical methods applied to comparative data lead to such inflation. We compared the observed frequency of correlation coefficients with the expected frequency (0.05) with a one-tailed test, the alternative hypothesis being that the observed frequency was larger than the nominal type I error rate. Therefore, equal deviations from the expected frequency of 0.05 were not given equal weight. Frequencies >0.05 constituted evidence in favor of the alternative hypothesis that the method yielded inflated type I error rates, but frequencies <0.05 did not count against the null hypothesis because values <0.05 do not inflate the type I error rate. For instance, for n = 1,000, an observed frequency of 0.055 would have a P value of 0.2529, whereas an observed frequency of 0.045 would have a P value of 0.7853. The values are not symmetric because of the asymmetry of the binomial distribution. We employed the exact binomial probabilities, not the normal approximation.

Absolute and Relative Performance of Methods

We evaluated absolute type I error rates of the different methods by computing combined probabilities for the six replicate chi-square tests (Table 3) and for the six replicate binomial tests (Table 4), which were based on the six replicate computer-simulated data sets (each using a different random number seed) for each model × phylogeny combination. The combined P values in Table 3 indicate whether the overall distribution of correlation coefficients obtained with the different methods is significantly different from the expected distribution; this test is sensitive to any kind of deviation from the expected distribution, including both inflation and deflation of the type I error rate. Comparison with the overall distribution is the most typical way of testing whether type I error rates of a statistical method deviate from expectations under standard normal theory (e.g., Martins and Garland, 1991). The combined P values in Table 4 test whether the P values at α = 0.05 are inflated, allowing us to address a more specific question about the type I error rates of a statistical method at

TABLE 3. Performance of alternative methods for estimating correlations of species' mean values. A chi-square test was used to compare the distribution of correlation coefficients as estimated by each method with the expected distribution of correlation coefficients for a true correlation of zero under standard normal theory. For each evolutionary model, the basic data set of 1,000 simulations was replicated six times; all parameters of the replicates were the same except for the seed of the pseudorandom number generator. Values reported are the mean and range of P values of the chi-square for the six replicate simulations (P values of 0.000 signify P values <0.0005) and the combined probability of the six independent chi-square tests. Small P values indicate poor performance (inflated type I error rates.) For each model, the first row corresponds to the 49-species phylogeny and the second row corresponds to the 15-species phylogeny.

Evolutionary model	TIPS			IC			ICblt			ICblte			
	Mean	Range	Combined	Mean	Range	Combined	Mean	Range	Combined	Mean	Range	Combined	No. simulations[a]
BM	0.000	0.000–0.000	<0.001	0.526	0.068–0.959	>0.25	0.474	0.093–0.712	>0.5	0.474	0.093–0.712	>0.5	1,000
	0.000	0.000–0.000	<0.001	0.526	0.105–0.999	>0.5	0.062	0.015–0.220	<0.001	0.068	0.015–0.233	<0.001	997
BMReplace	0.000	0.000–0.000	<0.001	0.228	0.014–0.843	<0.01	0.303	0.002–0.669	<0.025	0.304	0.002–0.673	<0.025	997
	0.720	0.412–0.986	>0.5	0.000	0.000–0.000	<0.001	0.313	0.007–0.646	<0.025	0.436	0.032–0.891	>0.25	879
BMTruncate	0.000	0.000–0.000	<0.001	0.162	0.006–0.520	<0.005	0.271	0.081–0.624	>0.1	0.273	0.081–0.624	>0.1	999
	0.151	0.000–0.562	<0.001	0.003	0.000–0.008	<0.001	0.413	0.148–0.893	>0.25	0.450	0.183–0.915	>0.25	963
OU	0.000	0.000–0.000	<0.001	0.074	0.002–0.291	<0.001	0.327	0.010–0.889	<0.05	0.335	0.009–0.836	<0.05	994
	0.001	0.000–0.001	<0.001	0.166	0.001–0.639	<0.001	0.336	0.065–0.525	>0.1	0.317	0.069–0.461	>0.1	975
OUReplace	0.005	0.000–0.017	<0.001	0.000	0.000–0.000	<0.001	0.065	0.001–0.212	<0.001	0.134	0.000–0.408	<0.001	650
	0.585	0.210–0.904	>0.5	0.000	0.000–0.000	<0.001	0.334	0.002–0.682	<0.025	0.410	0.052–0.709	>0.25	864
OUTruncate	0.000	0.000–0.000	<0.001	0.002	0.000–0.007	<0.001	0.078	0.000–0.256	<0.001	0.088	0.000–0.270	<0.001	964
	0.266	0.039–0.578	$0.1 > P > 0.05$	0.000	0.000–0.000	<0.001	0.583	0.139–0.980	>0.5	0.642	0.293–0.950	>0.5	927
PE	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.002	0.000–0.005	<0.001	0.002	0.000–0.004	<0.001	959
	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.146	0.016–0.375	<0.005	0.166	0.012–0.489	<0.005	975
PEReplace	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.035	0.000–0.201	<0.001	0.188	0.000–0.896	<0.001	540
	0.233	0.000–0.863	<0.001	0.000	0.000–0.000	<0.001	0.032	0.000–0.064	<0.001	0.318	0.000–0.686	<0.005	849
PETruncate	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.017	0.000–0.096	<0.001	0.010	0.000–0.055	<0.001	932
	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.281	0.012–0.485	$0.1 > P > 0.05$	0.305	0.003–0.613	<0.025	955
BMReplace-Shifting	0.001	0.000–0.001	<0.001	0.000	0.000–0.000	<0.001	0.032	0.009–0.078	<0.001	0.059	0.009–0.100	<0.001	842
	0.380	0.019–0.815	>0.1	0.000	0.000–0.000	<0.001	0.594	0.197–0.919	>0.5	0.673	0.304–0.913	>0.5	886
BMTruncate-Shifting	0.000	0.000–0.000	<0.001	0.093	0.000–0.294	<0.001	0.514	0.008–0.945	>0.1	0.517	0.010–0.914	>0.1	991
	0.006	0.000–0.023	<0.001	0.123	0.000–0.586	<0.001	0.453	0.029–0.944	>0.1	0.450	0.037–0.983	>0.1	972
OUReplace-Shifting	0.206	0.001–0.534	<0.005	0.000	0.000–0.000	<0.001	0.068	0.000–0.375	<0.001	0.152	0.000–0.570	<0.001	532
	0.480	0.156–0.733	>0.5	0.000	0.000–0.000	<0.001	0.247	0.073–0.576	$0.1 > P > 0.05$	0.468	0.050–0.903	>0.25	863
OUTruncate-Shifting	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.252	0.027–0.823	<0.025	0.263	0.026–0.747	<0.025	932
	0.265	0.018–0.953	<0.025	0.000	0.000–0.000	<0.001	0.531	0.228–0.756	>0.5	0.505	0.152–0.706	>0.5	932
PEReplace-Shifting	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	—[b]	—	—	191
	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	891

TABLE 3. Continued.

Evolutionary model	TIPS			IC			ICblt			ICblte			
	Mean	Range	Combined	Mean	Range	Combined	Mean	Range	Combined	Mean	Range	Combined	No. simulations[a]
PETruncate-	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	845
Shifting	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	0.000	0.000–0.000	<0.001	937
Total (n = 90)	0.014	0.000–0.534	<0.001	0.073	0.000–0.959	<0.001	0.162	0.000–0.945	<0.001	0.200	0.000–0.914	<0.001	825
	0.206	0.000–0.986	<0.001	0.055	0.000–0.999	<0.001	0.280	0.000–0.980	<0.001	0.347	0.000–0.983	<0.001	924

[a] Number of simulations that were considered adequately standardized in the analyses with the ICBlte method (this number must be ≤1,000); these numbers also apply to Table 4.
[b] Chi-square test could not be computed because of small expected frequencies.

a commonly used P value (cf. Grafen, 1989; Purvis et al., 1994).

Following Sokal and Rohlf (1981:779–781), the ln of the P value is distributed as $-\frac{1}{2}\chi^2$ with df = 2 (or $-2 \ln P$ is distributed as χ^2 with df = 2). We first computed the $-2 \ln P$ for each of the six independent chi-square or binomial tests and then added these six values; if all the null hypotheses were true, then this quantity would be distributed as a chi-square with 2 × 6 (number of independent tests) = 12 degrees of freedom.

The combination of the six independent significance tests (one for each replicate simulation) in both Tables 3 and 4 increases the statistical power for detecting deviations from the expected distribution (see Sokal and Rohlf, 1981:779–781, for more details). For example, in Table 3 the PERreplace model with the 15-species phylogeny has a mean chi-square P value of 0.233, which might seem nonsignificant; but the combination of the six independent tests reveals that the distribution of correlation coefficients was, in fact, significantly different from the standard expectation.

To assess relative performance of the methods, we compared them using the six independent P values of (1) the chi-square tests, which indicate deviations from expected for the overall distribution of correlation coefficients (Table 3) and (2) the number of correlation coefficients exceeding the critical value at $\alpha = 0.05$ (Table 4). The two different ways of comparing methods generally yielded similar results. Statistical significance of differences between methods was tested with nonparametric Wilcoxon signed-rank tests (Conover, 1980) (paired t-tests were not used because the distribution of the differences deviated significantly from normal.) These tests were performed separately for each of the 30 model × phylogeny combinations (n = 6 for each test) and for the data from all models combined for a given phylogeny (n = 90).

Statistical analyses of the data obtained from CMTPANAL and PDERROR were performed with SYSTAT version 5.0 (for obtaining binomial probabilities), SAS (for

TABLE 4. Performance of alternative methods for estimating correlations of species' mean values. Values reported are the mean number and range of correlation coefficients exceeding the critical value for $\alpha = 0.05$ for the six replicate simulations (in all cases the expected number under the null hypothesis [correct type I error rate at $P = 0.05$] is 50) and the combined probability of the corresponding six independent binomial tests. For each replicate simulation of each model, we tested whether the number of correlation coefficients was >50 (values ≥54 are significant at $P = 0.05$); therefore, the combined P value indicates whether the P value at the $\alpha = 0.05$ significance level is inflated. For each model, the first row corresponds to the 49-species phylogeny and the second row corresponds to the 15-species phylogeny.

	TIPS			IC			ICblt			ICblte[a]		
	No.			No.			No.			No.		
Evolutionary model	Mean	Range	P	Mean	Range	P	Mean	Range	P	Mean	Range	P
BM	420	395–445	<0.001	52	45–66	>0.25	45	38–62	>0.75	45	38–62	>0.75
	157	143–168	<0.001	47	39–60	>0.5	32	30–37	>0.5	32	30–37	>0.5
BMReplace	212	183–277	<0.001	62	49–73	<0.001	55	43–65	<0.05	55	43–65	<0.05
	52	48–60	>0.25	117	114–123	<0.001	58	53–71	<0.01	55	49–73	$0.1 > P > 0.05$
BMTruncate	324	273–364	<0.001	60	52–65	<0.005	49	45–53	>0.75	49	44–53	>0.75
	63	54–74	<0.001	78	75–83	<0.001	42	33–51	>0.5	42	34–50	>0.5
OU	203	196–213	<0.001	67	50–81	<0.001	58	49–66	<0.01	58	48–66	<0.01
	91	77–106	<0.001	63	56–71	<0.001	38	32–42	>0.5	38	32–43	>0.5
OUReplace	72	64–83	<0.001	106	91–120	<0.001	66	60–74	<0.001	68	57–81	<0.001
	51	42–57	>0.5	129	114–139	<0.001	57	51–68	<0.025	54	48–58	>0.25
OUTruncate	122	114–135	<0.001	79	70–96	<0.001	66	59–79	<0.001	66	60–79	<0.001
	63	52–68	<0.001	96	90–104	<0.001	51	44–55	>0.5	51	45–56	>0.5
PE	401	374–418	<0.001	162	149–178	<0.001	78	57–95	<0.001	78	56–93	<0.001
	214	198–229	<0.001	130	120–136	<0.001	67	59–75	<0.001	67	59–77	<0.001
PEReplace	96	85–106	<0.001	185	163–198	<0.001	74	67–94	<0.001	72	48–94	<0.001
	66	58–79	<0.001	151	136–167	<0.001	72	62–88	<0.001	65	50–84	<0.001
PETruncate	230	215–247	<0.001	160	149–178	<0.001	74	71–83	<0.001	74	68–85	<0.001
	108	93–130	<0.001	123	116–129	<0.001	62	56–67	<0.001	61	53–69	<0.001
BMReplaceShifting	82	75–96	<0.001	84	77–88	<0.001	66	58–81	<0.001	63	54–76	<0.001
	55	43–74	<0.025	117	109–124	<0.001	55	46–63	$0.1 > P > 0.05$	52	44–57	>0.5
BMTruncateShifting	158	152–168	<0.001	65	55–76	<0.001	55	46–69	<0.05	55	45–68	<0.05
	66	57–72	<0.001	75	59–87	<0.001	42	34–50	>0.5	42	35–50	>0.5
OUReplaceShifting	61	51–74	<0.001	109	102–118	<0.001	73	63–87	<0.001	72	60–82	<0.001
	54	48–64	>0.1	131	116–144	<0.001	61	51–75	<0.001	58	53–73	<0.025
OUTruncateShifting	92	79–108	<0.001	79	70–98	<0.001	60	54–67	<0.005	60	54–68	<0.005
	58	50–68	<0.025	93	84–105	<0.001	50	43–58	>0.5	49	42–58	>0.5
PEReplaceShifting	60	51–69	<0.001	186	174–205	<0.001	96	85–104	<0.001	86	74–107	<0.001
	71	55–87	<0.001	151	137–162	<0.001	74	63–81	<0.001	70	60–77	<0.001

TABLE 4. Continued.

Evolutionary model	TIPS			IC			ICblt			ICblte[a]		
	No.			No.			No.			No.		
	Mean	Range	P	Mean	Range	P	Mean	Range	P	Mean	Range	P
PETruncateShifting	109	101–114	<0.001	160	142–185	<0.001	75	64–89	<0.001	72	60–89	<0.001
	80	73–86	<0.001	129	117–145	<0.001	59	51–65	<0.005	58	48–63	<0.025
Total (n = 90)	176	51–445	<0.001	108	45–205	<0.001	66	38–104	<0.001	65	38–107	<0.001
	83	42–229	<0.001	109	39–167	<0.001	55	30–88	<0.001	53	30–84	<0.001

[a] Because the number of simulations for ICblte (see Table 3) was frequently <1,000, the number shown in this table is the product of the number of correlation coefficients exceeding the critical value for α = 0.05 multiplied by (1,000/no. simulations). This makes the expected number under the null hypothesis equal to 50 and allows direct comparison with the other methods.

ANOVA), and SPSS/PC+ version 5.0 (all other analyses). Statistical significance was judged at $P = 0.05$.

Table 3 refers to the overall distribution of correlation coefficients, whereas Table 4 refers specifically to the α level of 0.05. Both tables include two types of information: (1) the mean value (and range) of the relevant statistic for the six replicate simulations (same parameters but different random number seeds) and (2) the combined probability of the corresponding six independent significance tests. The first type of information allows comparison of the relative performance of each method; the second indicates the absolute performance of each method.

RESULTS

We used computer simulation to study the statistical performance (in terms of type I error rates) of four methods for estimating evolutionary correlations based on mean values for a set of species: a nonphylogenetic Pearson correlation of tip values (TIPS), independent contrasts (IC; Felsenstein, 1985), independent contrasts after transformation of branch lengths (ICblt; as suggested by Garland et al., 1992), and ICblt after excluding those simulated data sets for which branch length transformation did not yield adequate standardization of contrasts (ICblte).

A conventional Pearson correlation of species' mean values (TIPS) usually yields significantly inflated (and often greatly so) type I error rates (25 of 30 model × phylogeny combinations, Table 3; 27 of 30 model × phylogeny combinations, Table 4). Hence, ignoring phylogeny (which is equivalent to assuming the phylogeny to be a star with no hierarchical structure and equal branch lengths) would lead one to claim statistical significance too frequently and generally more frequently than for the phylogenetic methods. This result is consistent with all previous simulation studies (Grafen, 1989; Martins and Garland, 1991; Gittleman and Luh, 1994; Purvis et al., 1994; Martins, in press; see also Garland et al., 1993). All cases in which ignoring phylogeny (TIPS) does not lead to inflated

type I error rates are for simulation models involving limits with the Replace algorithm. In these cases, long branches will, in effect, tend to become shortened as characters evolve to reach limits. The effects of different algorithms for implementing limits on character evolution (several are available in PDSIMUL) warrants further study.

Except under a model of simple BM evolution, naive application of Felsenstein's (1985) method of phylogenetically independent contrasts (IC) often yields inflated type I error rates. The cases in which this inflation occurs do not correspond closely to those in which TIPS fails (see Tables 3, 4).

Checking and transforming branch lengths as described by Garland et al. (1992) improves the performance of independent contrasts (see Tables 3, 4). Nonetheless, inflated type I error rates still occur in two-thirds of the model × phylogeny combinations examined (ICblt: 19 of 30 cases, Table 3; 21 of 30 cases, Table 4). However, for some of the model × phylogeny combinations in which type I error rates are inflated, adequate standardization of contrasts (via transformation of branch lengths) could not be achieved in some or many of the 1,000 simulations (No. simulations, Table 3). If these cases were real data being analyzed, then the investigator presumably would not have relied on conventional critical values for determining significance levels. In other words, checks of the branch length diagnostics used herein and/or of the bivariate distribution of contrasts (which we did not check) would have stopped the investigator from relying on conventional critical values. Thus, the results for ICblte, which are somewhat better than that for ICblt, are probably the most reasonable indicator of the performance of ICblt (at least based on the models of evolution studied here, which may or may not be biologically realistic).

Absolute Performance of Each Method

For the nonphylogenetic TIPS method, the combined P values in Table 3 show that the overall distribution of correlation coefficients was significantly different from the expected distribution for all 15 models with the 49-species phylogeny and for 10 of 15 models with the 15-species phylogeny. For IC, with both phylogenies, the overall distribution of correlation coefficients differed significantly from expectations for every model except BM. Performance improved, however, with branch length transformations; the distribution of correlation coefficients for ICblt showed significant deviations from expectations in only 19 of 30 cases. When the simulations without appropriate standardization were excluded (ICblte), performance improved further, but the deviations from the expected distribution of correlation coefficients were still statistically significant in 17 of 29 cases (Table 3). ICblt and ICblte results under BM differed significantly from expectation with the 15-species phylogeny because of the small number of correlation coefficients in the tails of the distribution; in other words, these methods are too conservative because of the loss of degrees of freedom for unnecessary transformation of branch lengths.

Considering only $\alpha = 0.05$, Table 4 shows that for TIPS the number of correlation coefficients exceeding the critical value was significantly >50 in 27 of 30 cases. Significantly inflated type I error also occurred for IC in 28 of 30 cases. Again, however, performance of IC improved when branch lengths were transformed; significant inflation of type I error rates occurred in 22 of 30 cases for ICblt and in 20 of 30 cases for ICblte.

The performance of the methods was not completely independent of the phylogeny (see also Martins, in press); in general, all methods tended to perform better with the 15-species phylogeny. Moreover, interactions among evolutionary model, method of analysis, and phylogeny were statistically significant, as demonstrated by ANOVAs testing the effects of model of evolution, method of analysis, and phylogeny on the chi-square P values (after arcsine transformation) and on the number of correlation coefficients larger than the crit-

ical value at $\alpha = 0.05$ (after log transformation). The data were analyzed according to the following split-plot model (Snedecor and Cochran, 1989):

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} + \gamma_l + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + (\alpha\beta\gamma)_{ijl} + \delta_{ijkl},$$

where Y is the (arcsine of the) P value of the chi-square or the (log of the) number of correlation coefficients larger than the critical value at $\alpha = 0.05$; α is the effect of phylogeny, β is the effect of model of evolution, and γ is the effect of method of analysis; and ϵ and δ correspond to the "whole plot" and "subplot" errors, respectively. In both cases, all second-order interactions (model × method, model × phylogeny, method × phylogeny) and the third-order interaction (model × method × phylogeny) were significant at $P < 0.001$.

Relative Performance of the Different Methods

With independent contrasts, transformation of branch lengths always improved performance, both from the perspective of the overall distribution of correlation coefficients and considering only the 0.05 α level. Considering the overall distribution of correlation coefficients (Table 3), in all but one case the P values for ICblt were as large or larger than those for IC, meaning that the distribution of correlation coefficients obtained after transformation of branch lengths was closer to the standard expectation. For the 90 simulations combined, this difference was statistically significant (Table 3). Better results were achieved with untransformed branch lengths in only a single case: BM with the 15-species phylogeny. This difference, however, is caused by a smaller number of correlation coefficients for smaller α levels (e.g., see Table 4 for $\alpha = 0.05$). In this particular case, the type I error rate for ICblt would actually be smaller than the nominal error rate (with a corresponding decrease in power).

Considering only the number of correlation coefficients larger than the critical value at $\alpha = 0.05$, Table 4 shows that it was always closer to 50 for ICblt than for IC. This difference was statistically significant for both phylogenies (Wilcoxon test, $P < 0.05$), both for all the simulated data sets combined ($n = 90$) and for each one of the 15 evolutionary models ($n = 6$ for each).

Independent contrasts with transformed branch lengths (ICblt) also generally performed better than did the nonphylogenetic correlation (TIPS). From the perspective of the overall distribution of correlation coefficients (Table 3, chi-square tests), ICblt was usually significantly better than TIPS for both phylogenies (Wilcoxon tests), with the exceptions of BMReplace and OUReplaceShifting (TIPS significantly better) and OUReplace and PEReplace (TIPS nonsignificantly better) with the 15-species phylogeny and OUReplaceShifting (TIPS significantly better) with the 49-species phylogeny. For the α level of 0.05 (Table 4), ICblt generally performed significantly better than TIPS (Wilcoxon tests) with both phylogenies; the exceptions were PEReplaceShifting with 49 species and OUReplaceShifting and PETruncateShifting with 15 species, for which TIPS was significantly better than ICblt.

These results differ from the comparison between TIPS and IC, i.e., the naive application of independent contrasts. From the perspective of the overall distribution of correlation coefficients (Table 3), with the 15-species phylogeny, TIPS performed better than IC for all data sets combined and for every individual model except BM, OU, PE, PETruncate, BMTruncateShifting, PEReplaceShifting, and PETruncateShifting; with the 49-species phylogeny, TIPS performed significantly better than IC for OUReplaceShifting. Considering only the 0.05 α level (Table 4), with the 15-species phylogeny TIPS actually performed significantly better than IC for all data sets combined and for all individual models except BM, OU, PE, PETruncate, and BMTruncateShifting; with the 49-species phylogeny, TIPS performed significantly better than IC for the models OUReplace, PEReplace, OUReplaceShifting, PEReplaceShifting, and PETruncateShifting. In summary, independent contrasts are routinely better than TIPS (in virtually all models exam-

ined) only if the adequate standardization of branch lengths is checked.

Transformation of branch lengths did not always yield adequate standardization of contrasts (Table 3, No. simulations). However, the differences in performance before removing (ICblt) and after removing (ICblte) those cases were generally small (but see PEReplace and OUReplaceShifting models in Table 3). Nonetheless, for the overall distribution of correlation coefficients (Table 3), the differences were significant for all data sets combined (with both phylogenies) and for OUReplaceShifting with the 49-species phylogeny and PEReplace and OUReplaceShifting with the 15-species phylogeny. Considering only the α level of 0.05 (Table 4), the differences between ICblt and ICblte for all data sets combined were significant for the 15-species phylogeny and marginally significant for the 49-species phylogeny ($P = 0.074$), as well as for the individual cases of OUReplace, PEReplace, BMReplaceShifting, and PEReplaceShifting (15 species) and BMReplaceShifting (49 species). Moreover, of those cases in which TIPS performed better than ICblt, the ICblte variant actually performed significantly better than TIPS for the PETruncateShifting model ($\alpha = 0.05$; 15 species); the differences were not significant for PEReplace, OUReplaceShifting, OUReplace, and OUReplaceShifting (overall distribution: 49, 49, 15, and 15 species, respectively). For OUReplaceShifting and PEReplaceShifting ($\alpha = 0.05$; 15 and 49 species, respectively) and for BMReplace (overall distribution: 15 species), TIPS performed better than ICblte. In summary, although the differences between ICblte and ICblt were not always large, exclusion of those cases in which the phylogenetic contrasts were not properly standardized did yield better results.

Discussion

We have used Monte Carlo simulations to study the effects of violating assumptions of Felsenstein's (1985, 1988) method of phylogenetically independent contrasts for testing the statistical significance of a bivariate evolutionary correlation of continuous-valued characters. As with any such analytical method, several assumptions are inherent: the topology is correct, branch lengths are in units of expected variances of character change, evolutionary rates have been constant, and character evolution has occurred by a process that can be modeled as BM. Our focus is on the last of these assumptions because it has received little study (other than comparisons of gradual and speciational Brownian motion: Martins and Garland, 1991; Gittleman and Luh, 1992, 1994; Purvis et al., 1994) and because the independent contrasts method has been criticized specifically for its (apparent) reliance on a BM model (e.g., Miles and Dunham, 1993; Wenzel and Carpenter, 1994: but see Pagel, 1994a:44). Moreover, some workers have suggested that of the available comparative methods independent contrasts may be particularly committed to a BM model of evolution (cf. Gittleman and Luh, 1994; Purvis et al., 1994).

Obviously, BM may not adequately represent the evolution of real characters, so an understanding of the robustness of independent contrasts when BM does not apply is crucial. We know that BM character evolution is a sufficient condition for independent contrasts approaches to yield acceptable type I error rates (assuming correct phylogenetic information and constancy of evolutionary rates), but is it necessary? We have concentrated on type I error rates because this is the aspect of statistical performance that is usually of primary concern in comparative studies (e.g., Harvey and Pagel, 1991; Martins and Garland, 1991; Garland et al., 1993; Purvis et al., 1994; Martins, in press).

Some of the non-BM simulation models that we have used also effectively result in errors in branch lengths, e.g., when limits are employed with what is nominally BM. Specifically, those branches along which limits are reached will, in terms of the expected variance of evolutionary change, be shorter than their nominal value. Thus, simulating models that deviate from BM can also engender errors in branch lengths, but herein we have made no attempt to

separate the effects on type I error rates of simulation model per se from the effects of any consequent branch length errors.

Performance of Independent Contrasts When Brownian Motion Does Not Apply

Independent contrasts yield inflated type I error rates (see Tables 3, 4) under most of the 14 studied models (see Table 1) that deviate from BM. Nevertheless, not all deviations from simple BM lead to inflated type I error rates. For example, simply imposing a trend on BM (shifting the mean) did not lead to inflated type I error rates; this insensitivity was suggested by both Felsenstein (1985) and Grafen (1989). In addition, preliminary studies showed that weak limits, or weak OU decay constants, had little or no effect on type I error rates. An important area for future research will be attempting to determine how much biologically realistic deviations from BM character evolution affect the performance of independent contrasts. We have studied relatively extreme deviations (e.g., Fig. 1c), and we suspect that they may not be biologically realistic.

The models of evolution that produced the greatest inflation of type I error rates were usually those in which limits were implemented with the Replace algorithm of PDSIMUL (Garland et al., 1993), both with and without means shifting (evolutionary trends). These simulations tended to show relatively uniform distributions of tip values versus normal or bimodal distributions under no limits and Truncate Change, respectively (see Fig. 1). The punctuated equilibrium models also seemed to present special difficulties. This result is not surprising because the independent contrasts method computes the value for internal nodes as a weighted average of the two daughter nodes, whereas in PE evolution only one of the daughters changes while the other remains unchanged (see also Garland et al., 1993:282). So, we should not expect that just transforming branch lengths (e.g., setting all equal to unity) would yield correct type I error rates. This evolutionary model probably poses particular problems because the violation of assumptions is related not only to rates of evolutionary change but also to the way daughter nodes are related algebraically to the ancestor node.

Effects of Checking Branch Length Standardization

Checking and transforming branch lengths as suggested by Garland et al. (1991, 1992) almost always yields improved type I error rates with independent contrasts. The importance of branch length transformations was also implicitly emphasized by Grafen (1989). Similar conclusions have been reached in simulation studies of phylogenetic autocorrelation procedures (Gittleman and Luh, 1994; Martins, in press) in which the phylogenetic weighting matrix can be modified (Gittleman and Kot, 1990).

In our automated checking procedures used for the simulated data, branch lengths for each character were always transformed and two degrees of freedom were always subtracted for determining the critical value of the correlation coefficient between characters (even if the transformation exponent was unity, which actually leaves the branch lengths unchanged). Two degrees of freedom were subtracted because the data were used to estimate the branch length transformation with the best fit. The rationale here is that when the data themselves are used to justify a transformation (e.g., the Box-Cox transformation discussed by Sokal and Rohlf [1981]; see implementation by Reynolds and Lee [1996]), as opposed to applying a transformation chosen a priori (e.g., as in the application of a log-log transformation in many studies of bivariate allometry), then degrees of freedom should be lost for the additional parameter(s) being estimated. This argument is controversial in the statistical literature, and other acceptable approaches exist (E. V. Nordheim, pers. comm.); it is, however, more conservative than not subtracting degrees of freedom.

When a BM model does apply, however, transformations will usually be unnecessary and should lead to some loss of sta-

tistical power if one degree of freedom is subtracted for each branch length transform applied, particularly for a small number of species (Table 4, ICblt with BM for the 15-species phylogeny: 32 is considerably smaller than 50). Clearly, unnecessary branch length transformations, and the resulting loss of degrees of freedom, should be avoided. One strategy is to transform branch lengths only when the need is obvious, e.g., when the diagnostic used (Garland et al., 1992) shows a statistically significant pattern (e.g., Garland et al. [1991] and Bauwens et al. [1995] did not transform branch lengths), so that the decrease in degrees of freedom will occur only when "necessary." However, this procedure risks failing to detect cases in which transformation would have been appropriate: some violation of the assumptions might be real but not statistically significant for the number of species considered (i.e., the power of the diagnostics is not high enough). The power (and more generally the overall statistical behavior) of alternative diagnostics for detecting violations of the method warrants further study (see also Losos and Miles, 1994).

When branch lengths are checked and transformed (Table 4, ICblt and ICblte), type I error rates seem to have an upper bound, at least at $\alpha = 0.05$; in only one case was the number of correlation coefficients >100 out of 1,000, i.e., 0.10. This result differs (see Table 4) from those of independent contrasts without branch length transformations (IC) and particularly from the nonphylogenetic correlation (TIPS), for which the actual type I error rates were sometimes as much as eight times the nominal error rate. The apparent bounding of type I error rates under the ICblt procedures should be checked for other models of character evolution (e.g., character displacement) and for other phylogenies.

Performance of Independent Contrasts with Small Phylogenies and Polytomies

Only in one case did TIPS perform better than ICblte in a nontrivial sense (Table 4: OUReplaceShifting, 15 species). Although TIPS performed better than ICblte in some other instances, significantly or not, most of these differences were trivial. In some of those cases, ICblte performed well; in other cases where ICblte performed poorly, TIPS also yielded very poor results. Thus, if one checks branch lengths (as described by Garland et al., 1992) and relies on standard distributions to obtain critical values only when adequate branch length standardization is achieved, then one has little to lose by applying independent contrasts instead of TIPS (see also Purvis and Garland, 1993). Moreover, although ICblte differed in relative and absolute performance between the two phylogenies, it did not generally perform poorly with the 15-species phylogeny.

Therefore (contra Gittleman and Luh, 1992, 1994), independent contrasts, at least when applied with the recommended branch length checks, does not necessarily produce unacceptable type I error rates with small sample sizes. Martins and Garland (1991) and Martins (in press) conducted simulations under gradual and speciational BM using the same 15-species phylogeny employed herein and obtained results similar to ours. As noted elsewhere (Bauwens et al., 1995; Martins, in press), Gittleman and Luh's (1992, 1994) finding of poor performance for independent contrasts used with small sample sizes is attributable to the way polytomies were handled. Gittleman and Luh (1992, 1994) simulated both the characters and the phylogenetic trees and allowed polytomies. Following Pagel (1992), only a single contrast was computed for each polytomy. For their simulations with only 10 species, results for independent contrasts (e.g., Gittleman and Luh, 1992: fig. 4, 1994: fig. 5) showed substantial numbers of estimates of correlations at or near -1 or $+1$, which resulted in inflated type I error rates. The reason for such extreme values is that in some of their simulations with multifurcations only one or a few contrasts were computed. In the limit, when the simulated tree was completely unresolved, only a single contrast was computed, which, when analyzed by correlation through the

origin, necessarily yields a value of either −1 or +1.

Contrary to Pagel (1992), Purvis and Garland (1993) argued that computation of independent contrasts should be done following Felsenstein's (1985) original suggestion, in which case the number of contrasts computed is always $n - 1$, where n is the number of species, regardless of the number of unresolved nodes in the phylogeny. Thus, estimates of correlations for a phylogeny with 10 species would always be based on nine contrasts, never on as few as one, and estimates of −1 or +1 would be highly unlikely to occur. When the phylogeny is a star, the procedure advocated by Purvis and Garland (1993) will yield an independent contrasts estimate of a correlation that is exactly the same as a conventional nonphylogenetic Pearson correlation of tip values; this behavior is appropriate, because here the phylogeny has no hierarchical structure. Degrees of freedom for hypothesis testing can then be bounded (see Purvis and Garland, 1993; for an empirical example, see Christian and Garland, 1996). In comparison, all previous simulation studies involving independent contrasts with polytomies have assumed the minimal degrees of freedom (i.e., one per node in the phylogenetic tree) and have also extracted only one contrast per node in the phylogenetic tree for purposes of estimation (for computing the single contrast at each polytomous nodes, Gittleman and Luh [1992, 1994] and Purvis et al. [1994] followed Pagel [1992], whereas Grafen [1989] used a different procedure [see also Grafen, 1992; Pagel and Harvey, 1992]). Consequently, those studies are really tests of alternative ways of resolving polytomies and of obtaining branch lengths, rather than constituting general tests of the performance of independent contrasts as originally proposed by Felsenstein (1985).

Conclusions and Further Extensions

Testing for character correlations with phylogenetically independent contrasts and using the branch length diagnostics and transformations as suggested by Garland et al. (1992) seemed to yield bounded type I error rates, with a nominal $P = 0.05$ not exceeding an actual P of 0.10. This procedure thus seems relatively robust with respect to the assumed model of evolution and so may have some broad-sense validity (sensu Pagel and Harvey, 1992; Purvis et al., 1994). Moreover, although the branch length checks recommended by Garland et al. (1992) yielded results that were much improved over those obtained when branch lengths were not checked, even better performance might be obtained using other or additional branch length diagnostics. For example, Reynolds and Lee (1996) used the same branch length diagnostics but employed a modified Box-Cox procedure to estimate the optimal branch length transformations. A different diagnostic was suggested by Grafen (1989:146): plotting the absolute values of residuals (from the regression of standardized contrasts of Y on standardized contrasts of X) versus height of basal node for each contrast. Two other alternatives were suggested by Purvis and Rambaut (1995 [CAIC user's guide]): plotting the absolute values of standardized contrasts versus their estimated nodal values (e.g., Brandl et al., 1994:111) and plotting the absolute values of standardized contrasts versus average height of their two daughter nodes. In our simulations, the data could also have been checked for different rates of evolution in different parts of the phylogeny (see, e.g., Garland, 1992), because one of the likely effects of limits is to produce nonuniform rates of evolution. If such a pattern were found, then the investigator could transform branches in different parts of the tree in different ways (Grafen, 1989:146).

Other methods for estimating or transforming branch lengths are also available. From the perspective of multiple regression analysis, Grafen (1989, 1992) proposed using maximum likelihood techniques simultaneously to transform branch lengths (his ρ parameter) while estimating the partial regression coefficients. It is not clear how best to employ this approach for bivariate correlation or under some of the circumstances considered in this paper.

Determining the appropriate use of ρ for correlation studies in which different rates of evolution are allowed for each trait was beyond the scope of this paper but warrants further study. Martins's (1994) approach uses information on intraspecific phenotypic variation (in addition to mean values for a series of species) to estimate rates of evolutionary change in continuous-valued characters; it is applicable if evolution is known to be following either a BM or an OU model (OU also requires knowledge of the restraining force) and can guide choices about branch length transformations.

As would always be the case with real data, assumptions of the final statistical tests to be applied also must be checked, such as linearity, bivariate normality, homoscedasticity, and the presence of influential points or outliers (e.g., Hawkins, 1980; cf. Garland et al., 1993: fig. 5; Garland and Adolph, 1994: fig. 4). When distributional assumptions of parametric statistical methods cannot be met, then such nonparametric methods as sign tests can sometimes be applied (Felsenstein, 1985:13). Independent contrasts approaches are being widely used because of their relative simplicity (e.g., Garland and Adolph, 1994) and their applicability to a wide range of statistical procedures (e.g., multiple regression, analysis of covariance, principal components analysis). Nevertheless, the importance of additional checks, both of the assumptions of independent contrasts per se and of the statistical procedures in which they are employed, should not be underestimated. With respect to hypothesis testing, another alternative is to use computer simulation methods (overview by Crowley, 1992) to generate phylogenetically correct null distributions, as has been proposed elsewhere (Garland et al., 1991; Martins and Garland, 1991; Garland et al., 1993; Losos, 1994; Martins, 1996).

ACKNOWLEDGMENTS

We thank P. E. Midford for computer programing, E. V. Nordheim for many helpful discussions and statistical advice, and C. Lázaro-Perea for help in debugging PDERROR. D. Cannatella, J. Losos, P. E. Midford, E. V. Nordheim, and an anonymous reviewer made helpful comments on the manuscript. This work was supported by N.S.F. grants IBN-9157268 (PYI), DEB-9220872, and DEB-9509343 to T.G. and by a grant from the La Caixa Fellowship Program to R.D.-U.

REFERENCES

BAUWENS, D., T. GARLAND, JR., A. M. CASTILLA, AND R. VAN DAMME. 1995. Evolution of sprint speed in lacertid lizards: Morphological, physiological, and behavioral covariation. Evolution 49:848–863.

BRANDL, R., A. KRISTIN, AND B. LEISLER. 1994. Dietary niche breadth in a local community of passerine birds: An analysis using phylogenetic contrasts. Oecologia 98:109–116.

CHRISTIAN, A., AND T. GARLAND, JR. 1996. Scaling of limb proportions in monitor lizards (Squamata: Varanidae). J. Herpetol. (in press).

COLWELL, R. K., AND D. W. WINKLER. 1984. A null model for null models in biogeography. Pages 344–359 *in* Ecological communities, conceptual issues and the evidence (D. R. Strong, Jr., D. Simberloff, L. G. Abele, and A. B. Thistle, eds.). Princeton Univ. Press, Princeton, New Jersey.

CONOVER, W. J. 1980. Practical nonparametric statistics, 2nd edition. John Wiley & Sons, New York.

CROWLEY, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. Annu. Rev. Ecol. Syst. 23:405–447.

ELDREDGE, N., AND S. J. GOULD. 1972. Punctuated equilibria: An alternative to phyletic gradualism. Pages 82–115 *in* Models in paleobiology (T. J. M. Schopf, ed.). Freeman, Cooper and Co., San Francisco.

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

FELSENSTEIN, J. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. 19:445–471.

GARLAND, T., JR. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140:509–519.

GARLAND, T., JR. 1994. Phylogenetic analyses of lizard endurance capacity in relation to body size and body temperature. Pages 237–259 *in* Lizard ecology: Historical and experimental perspectives (L. J. Vitt and E. R. Pianka, eds.). Princeton Univ. Press, Princeton, New Jersey.

GARLAND, T., JR., AND S. C. ADOLPH. 1994. Why not to do two species comparative studies: Limitations on inferring adaptation. Physiol. Zool. 67:797–828.

GARLAND, T., JR., A. W. DICKERMAN, C. M. JANIS, AND J. A. JONES. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 42:265–292.

GARLAND, T., JR., P. H. HARVEY, AND A. R. IVES. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

GARLAND, T., JR., R. B. HUEY, AND A. F. BENNETT. 1991. Phylogeny and thermal physiology in lizards: A reanalysis. Evolution 45:1969–1975.

GITTLEMAN, J. L., AND M. KOT. 1990. Adaptation: Sta-

tistics and a null model for estimating phylogenetic effects. Syst. Zool. 39:227–241.

GITTLEMAN, J. L., AND H.-K. LUH. 1992. On comparing comparative methods. Annu. Rev. Ecol. Syst. 23: 383–404.

GITTLEMAN, J. L., AND H.-K. LUH. 1994. Phylogeny, evolutionary models and comparative methods: A simulation study. Linn. Soc. Symp. Ser. 17:103–122.

GOULD, S. J., AND N. ELDREDGE. 1977. Punctuated equilibria: The tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

GRAFEN, A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B 326:119–157.

GRAFEN, A. 1992. The uniqueness of the phylogenetic regression. J. Theor. Biol. 156:405–423.

HARVEY, P. H., AND M. D. PAGEL. 1991. The comparative method in evolutionary biology. Oxford Univ. Press, Oxford, England.

HAWKINS, D. M. 1980. Identification of outliers. Chapman and Hall, London.

HUEY, R. B., AND A. F. BENNETT. 1987. Phylogenetic studies of coadaptation: Preferred temperatures versus optimal performance temperatures of lizards. Evolution 41:1098–1115.

KIM, J., F. J. ROHLF, AND R. R. SOKAL. 1993. The accuracy of phylogenetic estimation using the neighbor-joining method. Evolution 47:471–486.

LOSOS, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. Syst. Biol. 43:117–123.

LOSOS, J. B., AND D. B. MILES. 1994. Adaptation, constraint, and the comparative method: Phylogenetic issues and methods. Pages 60–98 *in* Ecological morphology: Integrative organismal biology (P. C. Wainwright and S. Reilly, eds.). Univ. Chicago Press, Chicago.

MADDISON, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

MARTINS, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. Am. Nat. 144: 193–209.

MARTINS, E. P. 1996. Conducting phylogenetic comparative analyses when the phylogeny is not known. Evolution 50:12–22.

MARTINS, E. P. In press. Phylogenies, spatial autoregression, and the comparative method: A computer simulation test. Evolution.

MARTINS, E. P., AND T. GARLAND, JR. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. Evolution 45: 534–557.

MILES, D. B., AND A. E. DUNHAM. 1993. Historical perspectives in ecology and evolutionary biology: The use of phylogenetic comparative analyses. Annu. Rev. Ecol. Syst. 24:587–619.

PAGEL, M. D. 1992. A method for the analysis of comparative data. J. Theor. Biol. 156:431–442.

PAGEL, M. 1993. Seeking the evolutionary regression coefficient: An analysis of what comparative methods measure. J. Theor. Biol. 164:191–205.

PAGEL, M. D. 1994a. The adaptationist wager. Linn. Soc. Symp. Ser. 17:29–51.

PAGEL, M. D. 1994b. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B 255:37–45.

PAGEL, M. D., AND P. H. HARVEY. 1992. On solving the correct problem: Wishing does not make it so. J. Theor. Biol. 156:425–430.

PURVIS, A. 1995. A composite estimate of primate phylogeny. Philos. Trans. R. Soc. Lond. B 348:405–421.

PURVIS, A., AND T. GARLAND, JR. 1993. Polytomies in comparative analyses of continuous characters. Syst. Biol. 42:569–575.

PURVIS, A., J. L. GITTLEMAN, AND H.-K. LUH. 1994. Truth or consequences: Effects of phylogenetic accuracy on two comparative methods. J. Theor. Biol. 167:293–300.

PURVIS, A., AND A. RAMBAUT. 1995. Comparative analysis by independent contrasts (CAIC): An Apple Macintosh application for analysing comparative data. Comput. Appl. Biosci. 11:247–251.

RAUP, D. M., AND S. J. GOULD. 1974. Stochastic simulation and evolution of morphology—Towards a nomothetic paleontology. Syst. Zool. 23:305–322.

REYNOLDS, P. S., AND R. M. LEE III. 1996. Phylogenetic analysis of avian energetics: Passerines and nonpasserines are not significantly different. Am. Nat. (in press).

ROHLF, F. J., W. S. CHANG, R. R. SOKAL, AND J. KIM. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution 44: 1671–1684.

SESSIONS, S. K., AND A. LARSON. 1987. Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. Evolution 41:1239–1251.

SNEDECOR, G. W., AND W. G. COCHRAN. 1989. Statistical methods, 8th edition. Iowa State Univ. Press, Ames.

Sokal, R. R., AND F. J. ROHLF. 1981. Biometry, 2nd edition. W. H. Freeman, New York.

WENZEL, J. W., AND J. M. CARPENTER. 1994. Comparing methods: Adaptive traits and tests of adaptation. Linn. Soc. Symp. Ser. 17:79–101.

ZAR, J. H. 1984. Biostatistical analysis, 2nd edition. Prentice-Hall, Englewood Cliffs, New Jersey.

Received 19 December 1994; accepted 6 October 1995
Associate Editor: Allan Larson