Evolution, 45(3), 1991, pp. 534–557

PHYLOGENETIC ANALYSES OF THE CORRELATED EVOLUTION OF CONTINUOUS CHARACTERS: A SIMULATION STUDY

EMÍLIA P. MARTINS AND THEODORE GARLAND, JR.
Department of Zoology, University of Wisconsin, Madison, WI 53706 USA

Abstract.—We use computer simulation to compare the statistical properties of several methods that have been proposed for estimating the evolutionary correlation between two continuous traits, and define alternative evolutionary correlations that may be of interest. We focus on Felsenstein's (1985) method and some variations of it and on several "minimum evolution" methods (of which the procedure of Huey and Bennett [1987] is a special case), as compared with a nonphylogenetic correlation. The last, a simple correlation of trait values across the tips of a phylogeny, virtually always yields inflated Type I error rates, relatively low power, and relatively poor estimates of evolutionary correlations. We therefore cannot recommend its use. In contrast, Felsenstein's (1985) method yields acceptable significance tests, high power, and good estimates of what we term the input correlation and the standardized realized evolutionary correlation, given complete phylogenetic information and knowledge of the rate and mode of character change (e.g., gradual and proportional to time ["Brownian motion"] or punctuational, with change only at speciation events). Inaccurate branch length information may affect any method adversely, but only rarely does it cause Felsenstein's (1985) method to perform worse than do the others tested. Other proposed methods generally yield inflated Type I error rates and have lower power. However, certain minimum evolution methods (although not the specific procedure used by Huey and Bennett [1987]) often provide more accurate estimates of what we term the unstandardized realized evolutionary correlation, and their use is recommended when estimation of this correlation is desired. We also demonstrate how correct Type I error rates can be obtained for any method by reference to an empirical null distribution derived from computer simulations, and provide practical suggestions on choosing an analytical method, based both on the evolutionary correlation of interest and on the availability of branch lengths and knowledge of the model of evolutionary change appropriate for the characters being analyzed. Computer programs that implement the various methods and that will simulate (correlated) character evolution along a known phylogeny are available from the authors on request. These programs can be used to test the effectiveness of any new methods that might be proposed, and to check the generality of our conclusions with regard to other phylogenies.

Key words.—Coadaptation, coevolution, comparative method, computer simulation, gradualism, phylogenetic analysis, phylogenetic inertia, phylogeny, punctuated equilibrium, systematics.

Received September 12, 1989. Accepted August 3, 1990.

Much of evolutionary biology is inherently historical in nature. Although the fossil record provides the only direct evidence of past changes, attempts to reconstruct the evolutionary history of features of organisms based primarily or exclusively on neontological data are common. This is particularly true for some types of traits, such as behavior or physiology, for which paleontological evidence is rarely available. If a single trait is studied, a common goal is to infer its state in an ancestral form, generally based on some sort of parsimony argument (e.g., Farris, 1970; Ruben and Bennett, 1980; Larson, 1984; Campbell et al., 1985; Swofford and Maddison, 1987; Donoghue, 1989; Maddison, 1990). Alternatively, the distribution of a trait may be examined in relation to ecological or environmental characteristics (Huey, 1987). Relationships identified by such comparisons may be used to formulate hypotheses about adaptation or to test preexisting hypotheses (Harvey and Mace, 1982; Ridley, 1983; Greene, 1986; Lauder, 1986; Huey, 1987; Krebs and Davies, 1987 Ch. 2; Coddington, 1988; Gittleman, 1988; Hailman, 1988; Bell, 1989; Donoghue, 1989; Harvey and Pagel, 1991).

Another common question in comparative studies is whether traits have evolved in a correlated fashion (e.g., Huey and Bennett, 1987; Sessions and Larson, 1987; Losos, 1990). Typically, species are compared and patterns are defined by the existence of statistically significant relationships between traits. However, because organisms descend in a hierarchical fashion from common ancestors, data for different species are not independent and standard statistical techniques are inappropriate for compara-

tive analyses (e.g., Felsenstein, 1985). A number of quantitative methods for inferring relationships between traits, while taking phylogeny into account, have therefore been proposed (reviews in Ridley, 1983; Felsenstein, 1988; Pagel and Harvey, 1988*a*; Maddison, 1990; Harvey and Pagel, 1991).

Our purpose herein is to use computer simulation to compare the assumptions, statistical properties, and practical applicability of several methods for testing whether two continuous traits have evolved in a correlated fashion. Practical considerations do not allow us to consider all available methods; rather, we have chosen some that are becoming well known and that may be seen as exemplars of three larger categories of approaches. We did not consider methods designed for categorical characters (e.g., Ridley, 1983; Maddison, 1990) or which address the rather different problem of separating the effects of phylogenetic "inertia" or "constraints" from specific adaptation (cf. Ballinger, 1983; Cheverud et al., 1985; Bell, 1989), although at least two of these methods (Cheverud et al., 1985; Lynch, 1991) may also be used to test for correlated evolution (cf. Gittleman and Kot, 1990). In all cases, we assumed the true phylogeny for the species being compared was known to the level of an entirely dichotomous topology. Thus, we did not consider problems introduced by errors in the available topology or by unresolved polytomies ("unrecognized phylogeny" sensu Grafen, 1989). We assumed also that the available phylogeny was based on characters independent of those being studied for evolutionary correlation (Felsenstein, 1985, 1988).

The first general category of comparative methods we consider consists of those that are explicitly nonphylogenetic. As an example, we consider a simple correlation across the tips of a phylogeny (nested ANOVA approaches are a special case of this, as discussed below). Although this procedure is, in principle, known to be statistically unacceptable for hypothesis testing, how poorly it actually performs has been quantified only by Grafen (1989).

The second category of approaches is derived from Felsenstein's (1985) method of standardized independent contrasts. Felsenstein's (1985) method has been shown analytically to yield acceptable significance tests, but requires complete knowledge of both phylogenetic topology and branch lengths (in units of expected variance of character change). Moreover, how well it estimates the evolutionary relationship between traits has not been considered. We wished, therefore, to verify its performance through computer simulation, to determine how susceptible it is to incomplete or inaccurate knowledge of branch lengths (or, equivalently, rates of character change), to test its ability to estimate different evolutionary correlations (see below), and to test some variants of this method.

The third category of approaches is "minimum evolution" methods, similar to that applied by Huey and Bennett (1987). These methods were originally proposed for explicit reconstruction of ancestral states and hence might be expected to provide better estimates of correlations among the evolutionary changes occurring in two traits. To test this possibility, we define three types of evolutionary correlation that may be desired, and test the abilities of all methods to estimate these. This distinction is important, because most authors have not been explicit as to what, exactly, is being estimated by a particular method, and because the different evolutionary correlations vary in statistical and biological relevance. Acceptable means of hypothesis testing have not been proposed for minimum evolution methods. We therefore develop a procedure for obtaining appropriate significance tests through the creation of computer-simulated null distributions. This procedure can be used for hypothesis testing with any method, and therefore allows us to choose among methods for reasons other than hypothesis testing (e.g., statistical estimation). Overall, we provide a framework within which any new comparative methods may be tested.

METHODS

The Methods Compared.—We tested three categories of approaches in this study. The "**TIPS**" method (Table 1) is simply a Pearson product–moment correlation between the raw values of two traits for a series of species. TIPS is the traditional "equilibri-

TABLE 1. Short descriptions of the comparative methods tested and abbreviations used in the text.

Abbreviation	Assumed model of evolutionary change	Description
TIPS	Not specified	The traditional nonphylogenetic "equilibrium" approach, a simple Pearson product–moment correlation of species values across the tips of a phylogeny
FL1G	Gradual	Felsenstein's (1985) method of standardized contrasts
FL2G	Gradual	Felsenstein's (1985) method, but not standardizing contrasts
FL1P	Punctuational	Felsenstein's (1985) method of standardized contrasts
FL2P	Punctuational	Felsenstein's (1985) method, but not standardizing contrasts
ME1G	Gradual	Minimum evolution method using all inferred changes along the phylogeny
ME2G	Gradual	Minimum evolution method using only inferred changes that lead to tips
ME1P	Punctuational	Minimum evolution method using all inferred changes along the phylogeny
ME2P	Punctuational	Minimum evolution method using only inferred changes that lead to tips (method most similar to that used by Huey and Bennett, 1987)

um" analysis (Lauder, 1981; Huey, 1987), and has been termed the "nonphylogenetic approach" by Felsenstein (1988) and the "naive species regression" by Grafen (1989). TIPS has been criticized primarily because trait values measured in different species are not independent of each other, and thus should not be used with most standard statistical tests (e.g., Felsenstein, 1985). The TIPS method would be adequate for comparative studies if (1) phylogenetic inertia were entirely absent, in which case characters would respond instantaneously to natural selection in the current environment (Felsenstein, 1985 p. 6) or (2) the species studied derived from a "star" phylogeny as depicted in Felsenstein's (1985) Figure 2. Otherwise, the problem of nonindependent data points translates statistically into a question concerning the appropriate degrees of freedom to be used in tests of significance (Felsenstein, 1985). Hierarchical phylogenetic relationships between species effectively decrease the available degrees of freedom by some unknown quantity.

Ignorance concerning appropriate degrees of freedom can be dealt with in various ways. Crook (1965) suggested comparing means for higher taxonomic units to achieve greater independence of points. Clutton-Brock and Harvey (1977, 1984; Harvey and Mace, 1982; see also Bell, 1989) suggested a similar approach, but using nested analysis of variance to determine nonarbitrarily which taxonomic level should be used for averaging. These methods require phylogenetic information in that they implicitly assume that taxa represent monophyletic groups of comparable age and hence comparable expected amounts of within-taxon divergence. Although such methods may reduce the problem of statistical nonindependence, they "are only a partial solution" (Clutton-Brock and Harvey, 1977 p. 8). Just as species within a genus may not be independent, so may genera within a family not be independent. Nested ANOVA approaches thus represent a special case of TIPS, in which some attempt to decrease the effects of nonindependence has been made by prior taxonomic averaging. They are arbitrary to the extent that all taxonomic levels are arbitrary, and they are inefficient as they discard information from taxonomic levels below that at which averaging is performed. Harvey and Pagel (1991) conclude that nested ANOVA approaches are now obsolete, and we do not consider them separately.

The second category of approaches we tested solves the problem of nonindependence by computing statistics that describe the available data, but that are independent of each other. For example, Felsenstein's (1985, 1988; see also Burt, 1989) technique computes statistically independent contrasts for each trait. Appropriately standardized, these contrasts can be viewed as having been drawn from a normal probability distribution with mean of zero and unit variance; thus, exact significance tests are available. Standardization requires that branch lengths be available in units of expected variance of evolutionary change for the characters of interest. If the rate of evolution is assumed to be constant, then these

branch lengths can be estimated as absolute or relative time. If rates of evolution are *not* constant over the entire tree, standardization is still possible, but details of the rate changes must be known.

We tested both Felsenstein's (1985) original method and a variant under two models of evolutionary change. In method "**FL1G**" (Table 1), change was assumed to occur gradually such that the expected variance of change in each trait was simply proportional to time. Standardization was thus accomplished by dividing each contrast by the square root of the sum of the branch lengths (in units of relative time). (Felsenstein's s_x^2 and s_y^2, which indicate the relative rates at which two traits evolve, are not used for computation of correlation coefficients, and may be ignored.) In "**FL1P**" (Table 1), change was assumed to be "punctuational," (or "speciational," Rohlf et al., 1990), occurring only at speciation events, such that the expected variance of change in each trait was proportional to the number of speciation events on the phylogeny. (A further implicit assumption is that *all* speciation events within the clade, whether leading to measured species or not, are known and counted.) Under this model, appropriate standardization of contrasts was accomplished by setting all branch lengths equal (the specific value used does not matter) and dividing each contrast as before by the square root of the sum of the branch lengths (i.e., the square root of the sum of the number of speciation events occurring along a branch).

The desirability of standardizing contrasts in the way proposed by Felsenstein (1985) is unclear (cf. Felsenstein, 1988 p. 465; Bell, 1988 pp. 554, 565; Harvey and Pagel, 1991). For example, under a gradual model of change, this method of standardization implies that evolutionary changes occurring over short periods of time are weighted equally to those occurring over longer periods of time. Alternatively, changes occurring over long periods of time can be given greater weight simply by computing a correlation between nonstandardized contrasts ("FL2"), although standard significance tests would not be appropriate. For both FL1 and FL2, nodal values used to compute contrasts were estimated as the weighted average of the values for the two descendents of that node (as indicated by Felsenstein, 1985) with weights being proportional to branch lengths in units of expected variance of change. Again, we tested two models of evolutionary change, and designated the correlation from the nonstandardized gradual version of Felsenstein's (1985) method "**FL2G**" and the correlation from the nonstandardized punctuational version "**FL2P**" (Table 1). Once contrasts were obtained for each of these four versions, we calculated correlation coefficients by the following formula:

$$r = \frac{\sum [(\text{Contrast Trait A}) \times (\text{Contrast Trait B})]}{\left[\sum (\text{Contrast Trait A})^2 \times \sum (\text{Contrast Trait B})^2\right]^{0.5}}$$

Although intuitively appealing, branch lengths in units of time do not necessarily provide the best estimate of expected variance of character change, unless evolution is gradual and clock-like. As an alternative, Grafen (1989) developed the "phylogenetic regression," which can be used with incomplete phylogenetic information (ignorance of branch lengths and/or unresolved polytomies). This method employs maximum likelihood to obtain estimates of "branch lengths" in units of expected variance of change. These branch lengths are then used to standardize contrasts (equivalent to those of Felsenstein [1985]), on which statistical analyses can be performed. As another alternative, Harvey and Pagel (1991) suggest calculating contrasts as described by Felsenstein (1985), and then using residual analysis and, if necessary, weighted regres-

sion as a remedial measure to determine whether contrasts actually require standardization for significance testing. Thus, Felsenstein's, Grafen's, and Harvey and Pagel's procedures all offer significance tests based on explicit statistical assumptions. As differences among these three procedures will be due primarily to the accuracy of the available phylogenetic information (including rates of character change), we do not test either Grafen's or Harvey and Pagel's methods separately. A useful extension of our comparisons would be to determine how much statistical performance declines if one (unnecessarily) applies Grafen's (1989) or Harvey and Pagel's (1991) procedures even when complete and accurate phylogenetic information *is* available.

A third way to deal with the degrees of freedom problem in hypothesis testing is by using a sort of randomization test that takes phylogeny into account. This requires that a topology be available and that one be willing to specify a model of evolutionary change (e.g., gradual or punctuational). The usual null hypothesis in comparative studies is that of no relationship between traits. Thus, the evolution of traits showing no correlation may be simulated many times along a known phylogeny, until multiple sets of species values are obtained. Analysis of these data by *any* method will result in the creation of an empirical null distribution of correlation coefficients for that method on that phylogeny. Hypothesis testing of the results of analysis of real data may then be conducted by comparison with this null distribution rather than against standard distributions of critical values to obtain reasonable Type I error rates despite problems of nonindependence. It is important to note that *all* proposed methods of phylogenetic analysis make implicit assumptions concerning patterns and rates of evolutionary change of the traits being considered. Creation of computer-simulated null distributions requires that these assumptions be made explicit.

This third procedure for hypothesis testing is also useful as it allows us to consider the third category of methods, which does not deal explicitly with the problem of nonindependence, but which may be preferred for other reasons (e.g., statistical estimation). For example, some authors developed methods in which nodal values on a phylogeny are inferred, changes along branch segments are computed, and those changes are used to compute a correlation (e.g., Farris, 1970; Huey and Bennett, 1987; Swofford and Maddison, 1987). As nodal values are inferred from the values of their descendents and hypothetical ancestors, these methods are also subject to the problem of nonindependence. The most common method of inferring nodal values is to use a parsimony or "minimum evolution" algorithm. (A problem with all parsimony reconstructions is that they may underestimate the frequency of parallel change [Felsenstein, 1983, 1985]). This algorithm may minimize the sum of changes ("Wagner parsimony": Farris, 1970; Swofford and Maddison, 1987) or the sum of squared changes (Huey and Bennett, 1987; Maddison, 1991). These alternatives may yield different results, but we suspect such differences will depend primarily on the phylogeny in question rather than on intrinsic differences between the algorithms. One difficulty with implementing Farris' (1970) algorithm is that it may yield multiple solutions (Swofford and Maddison, 1987). For the sake of simplicity, and to avoid complications due to multiple solutions, we chose to test only minimum evolution methods based on an algorithm that minimizes the sum of squared changes.

We first used an iterative "means algorithm" that estimates nodal values on a phylogeny by setting each node equal to the mean of the nearest three nodes or tips (Huey and Bennett, 1987; see Maddison, 1991, for a direct algorithm for estimating nodal values). Means were computed in two different ways. The first involved weighting by branch lengths in units of time, which is appropriate under a gradual model of character change ("**ME__G**"); the second assumed all branch lengths were equal, which is appropriate for a punctuational model of change ("**ME__P**") (Table 1). Once nodal values were computed, inferred changes between connected points on the phylogeny were obtained by simple subtraction. Differences were calculated between all connected points on the phylogeny ("**ME1__**") or only between nodes and tips ("**ME2__**"), as was done by Huey and Bennett (1987). Finally, a standard Pearson product–moment cor-

relation coefficient between these inferred changes for the two traits was computed (see Appendix A for an example).

Huey and Bennett (1987) actually applied a special case of the general minimum evolution method we have outlined. They performed the iterative means algorithm on generic averages (rather than on species values), assuming all branch lengths were equal (as in our "punctuational" model of evolutionary change, MEP), and without allowing change on the branch leading to the "outgroup." They then computed changes between nodes and tips only (as in our ME2) due to uncertainty as to proper degrees of freedom and in order to avoid using more data points than original tips (R. B. Huey, pers. comm.). Finally, they computed a least-squares linear regression between the changes for the two traits, and significance tested with $N - 2$ degrees of freedom, with N equal to the number of changes between nodes and tips (excluding the "outgroup"). Actual Type I error rates under this procedure are unknown. Except for the generic averaging, Huey and Bennett's (1987) procedure is most similar to the version we have termed ME2P.

Alternative Evolutionary Correlations.— One can view the evolutionary changes in two traits occurring at each generation as random variables drawn from a bivariate probability distribution of possible changes. This probability distribution is determined by prevailing selection pressures, by genetic and environmental correlations (and their interactions), and by random genetic drift, mutation, gene flow, and any other factors affecting gene frequencies within a population. The parameters of the probability distribution describe a specific pattern of evolutionary change. For example, a bivariate probability distribution with means of zero and a correlation of 0.5 would indicate no net change, on average, in either character, but an overall trend for positive correlation between changes in the two characters. The relationship between the evolutionary changes in two traits might thus be described by the correlation of the bivariate probability distribution from which changes are drawn, which we term the *input correlation* (Fig. 1).

The actual pattern of correlated changes that occurs in a particular group of organisms from generation to generation may differ from the parameters of the input distribution. Thus, for some purposes, a statistic that estimates the correlation of actual evolutionary changes in the traits at each generation might be preferred. This *correlation across generations* (Fig. 1), is a realization of finite sampling from the input distribution. In the laboratory, barnyard, or garden, one might actually observe such changes. Similarly, the fossil record sometimes may approach the resolution necessary to quantify changes in two traits as they occurred over multiple generations (e.g., Williamson, 1981). The correlation across generations could not be computed from our simulations, simply because changes did not occur every generation. Although the correlation between the actual changes drawn during any single simulation was available, this is not very useful because the error variance of this statistic depends on the frequency of sampling, which was determined for practical rather than biological reasons (see below).

A similar but more accessible type of evolutionary correlation is based on changes that occur between speciation events (cf. Huey and Bennett, 1987). Given an independently derived phylogeny, one can estimate values for each trait of interest at the nodes (hypothetical ancestors), for example by using a minimum evolution (parsimony) algorithm, as discussed in the previous section. Simple subtraction between trait values for points on the phylogeny then yields the inferred changes in each character. The correlation of these changes we term the *realized evolutionary correlation* (Fig. 1). The realized evolutionary correlation was calculated from the simulated data as the correlation between internode differences in each trait, with each difference being either standardized (*SREC*) or not standardized (*UREC*). Standardization was accomplished by dividing the difference between each set of points on the phylogeny by the square root of the branch length (in units of expected variance of change) along which the difference occurred. This standardization differentially weights changes occurring along different branches, as discussed above for FL1 and FL2.

As shown in Figure 1, the correlation across generations, as a finite sample from

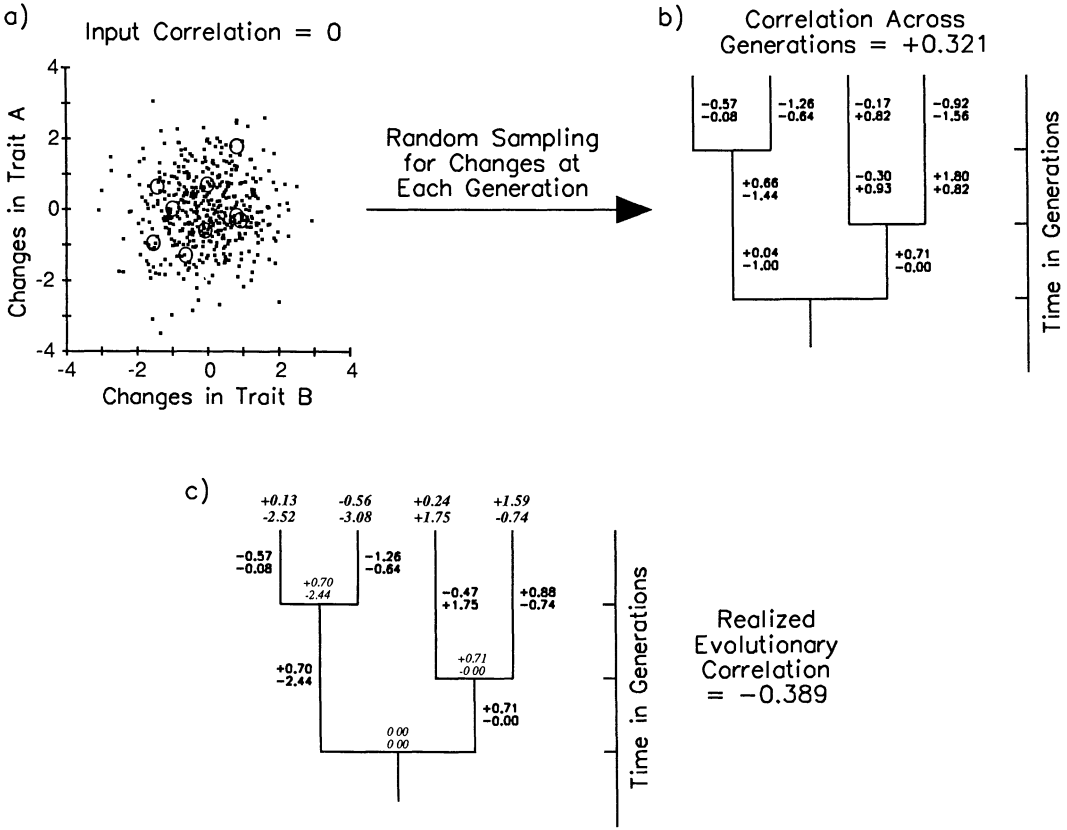


FIG. 1. Alternative correlations which may be used to describe the evolutionary relationships between traits. (a) The bivariate (normal) distribution from which changes are drawn each generation. The correlation of this distribution is termed the input correlation. Five hundred points are depicted, nine of which (open circles) were randomly sampled to produce the correlation across generations, depicted in (b). In this example, these nine randomly selected points yield a correlation of +0.321 between changes in trait A (top) and changes in trait B (bottom). (c) The changes that produce the correlation across generations can be grouped into changes that occur between nodes (hypothetical ancestors, italicized) and from nodes to tips of the phylogeny. The (unstandardized) realized evolutionary correlation is computed from these changes, and equals −0.389 in this example.

the probability distribution of possible changes, may serve as an estimate of the input correlation. In turn, the realized evolutionary correlation results from grouping the changes used to obtain the correlation across generations, and may serve as an estimate of it. The relationship between these three statistics might be viewed as a sort of random effects model, in which the realized evolutionary correlation is a function of (1) the input correlation, (2) the sampling realization of the input correlation, which constitutes the correlation across generations in a particular trial (actual evolution or simulation), and (3) a random error term.

Just as nonindependence due to phylogenetic relationships causes problems for significance testing, so too may it lead to inaccuracies of estimation. If, for example, some clades are more speciose than others, and if phylogenetic inertia is strong for the traits of interest, then estimates of the correlation between these traits may be biased by those clades containing large numbers of species (as noted by Harvey and Mace, 1982 p. 346). Differential probabilities of speciation and/or extinction that are correlated with the traits being studied might also lead to biases. Which evolutionary correlation TIPS best estimates has not been discussed. Felsenstein's (1985) method is explicitly designed to estimate the input correlation, whereas minimum evolution methods seem intended to estimate realized

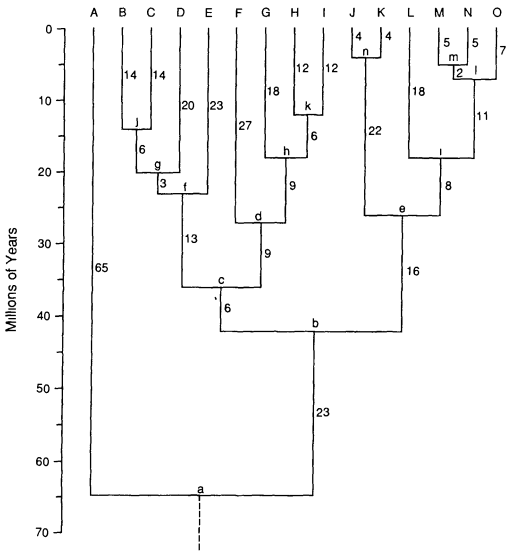


FIG. 2. Phylogeny for 15 species of plethodontid salamanders, from Sessions and Larson's (1987) Figure 3. Vertical axis represents time since divergence; horizontal axis is arbitrary.

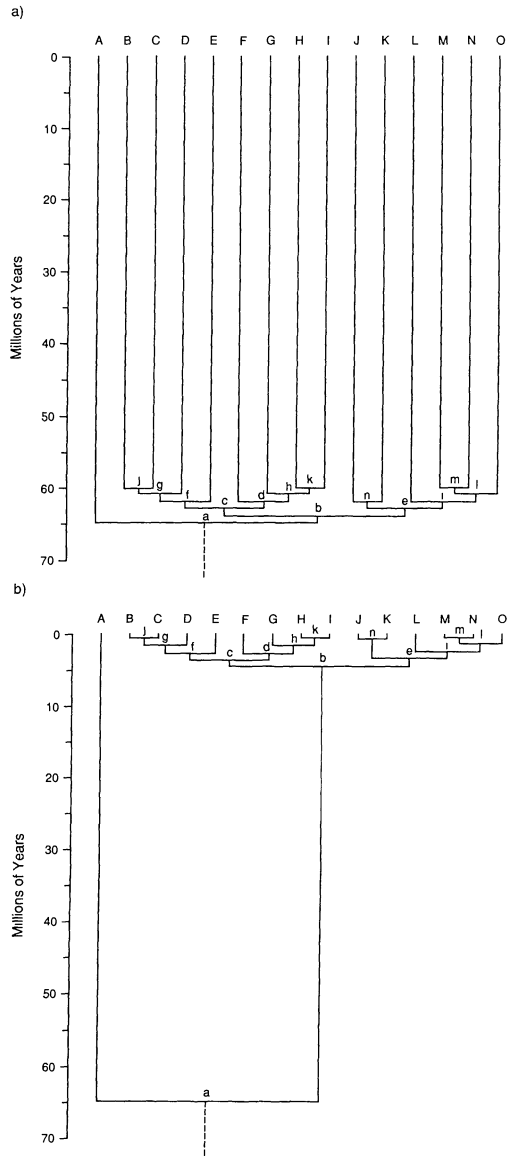


FIG. 3. Phylogeny (from Fig. 2), with branch lengths modified (a) to increase independence among contemporary tip species and (b) to decrease independence among contemporary tip species.

evolutionary correlations (cf. Huey and Bennett, 1987).

Computer Simulation of Data.—Each of the methods described above was used to analyze data created by computer simulation. Simulated evolution was begun at the base of a known phylogeny, and random changes were added to the previous value at each step until trait values were obtained for each species on the phylogeny. Simulations were repeated 1,000 times for each combination of phylogeny and model of character change. Analysis of the simulated species values by each of the methods thus resulted in distributions of 1,000 correlation coefficients for each method.

We felt it appropriate to begin our studies with a single phylogeny—one actually used in a comparative study. Sessions and Larson (1987) used Felsenstein's (1985) method in a study of developmental correlates of genome size in plethodontid salamanders (see also Burt, 1989). We thus used their phylogeny for 15 species (based on morphological and biochemical data), as depicted in their Figure 3 and in our Figure 2. We also modified this phylogeny to represent extreme situations. First, we created what is almost a "star" phylogeny (Fig. 3a), with all divergence occurring very early in the radiation (we arbitrarily made all internode branch lengths equal to one million years). Second, we created the opposite of a star phylogeny, with almost all divergence occurring very late in the radiation (Fig. 3b), thus aggravating the problem of nonindependence among species. Simulating data along one of these two phylogenies and then

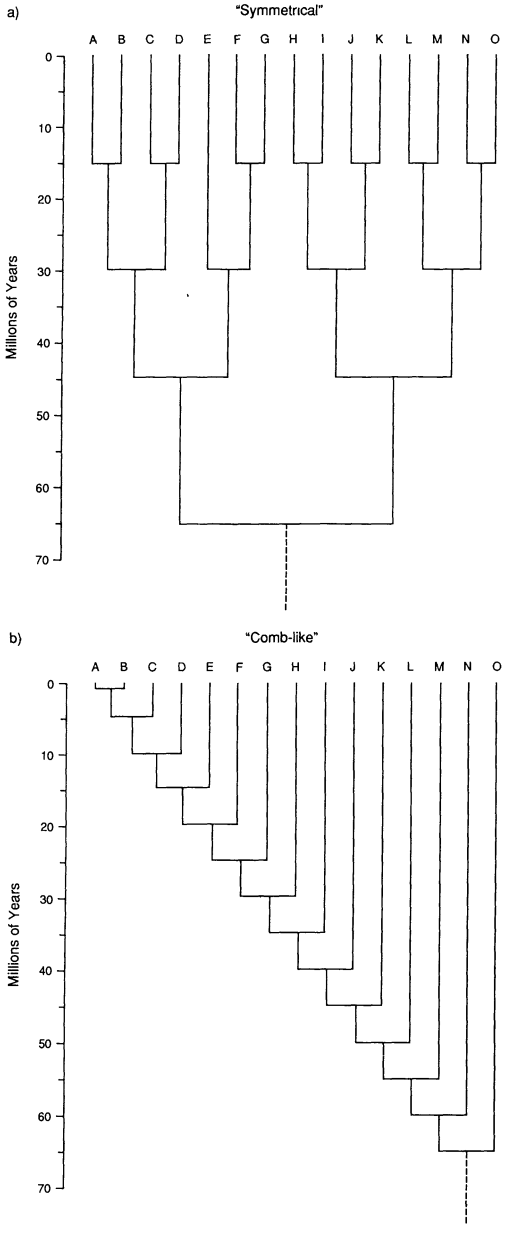


FIG. 4. Phylogenies with topologies extremely different from those of Figures 2 or 3.

analyzing these data *as if they applied to the other phylogeny* allowed us to test the robustness of methods when using extremely inaccurate branch length information. Finally, we ran additional simulations on the two different topologies depicted in Figure 4, one with rather symmetrical branching (Fig. 4a) and the other with branching in a somewhat comb-like pattern (Fig. 4b).

The frequency at which random character changes were allowed to occur along the phylogeny was specified by the model of evolutionary change, either gradual or punctuational. Gradual change potentially occurs in each trait each generation [in which case, time may be used as an estimate of expected variance of change in both Felsenstein's (1985) and minimum evolution analyses]. Equivalently, we noted the lowest common denominator of all branch lengths, converted all branch lengths to multiples of this common denominator, then allowed each character to change this number of times along a given branch.

An extreme punctuational model, in contrast, requires that all evolutionary change occurs during or soon after speciation events (e.g., Eldredge and Gould, 1972; Gould and Eldredge, 1977), such that expected variance of change is proportional to the number of speciation events rather than to time. Purely punctuational change poses great conceptual problems for any comparative method. First, if change occurs only at speciation events, then one would need to know of *every* past speciation event that had occurred anywhere within the clade being studied, regardless of whether the species were extinct or extant and regardless of whether they were actually included in the analysis. This knowledge would be necessary for proper standardization of contrasts in Felsenstein's (1985) method and for appropriate computation of nodes in minimum evolution methods. Second, some formulations of punctuated equilibrium allow change in only one of the two daughter lineages following a bifurcation (e.g., simulations by Raup and Gould, 1974; Colwell and Winkler, 1984). If evolution were known to occur strictly in this fashion, it would lead to difficult computational uncertainties owing to lack of knowledge as to which of a pair of daughter species experienced character change. We thus chose to assume that a single change occurred in both daughters following each bifurcation, as has been depicted by Schopf (1981) and Douglas and Avise (1982). Maynard Smith (1983 p. 21) points out that there is no compelling genetic reason to think it necessary that *no*

change would occur in the daughter species representing the large "mother" population.

A separate issue is the form of the underlying distribution from which changes are drawn. Felsenstein (1985, 1988) argued that if evolutionary change occurred as by Brownian motion (involving many small, independent changes), as in a gradual model of change, then the total change along any branch will be normally distributed with mean of zero and variance proportional to the number of small, independent changes. As explained above, our simulations potentially use only a single change for the shortest branch on a tree. To ensure that the total change for each branch was normally distributed, we drew individual changes in each trait from a bivariate normal probability distribution.

For punctuational change, we know of no detailed discussions as to the appropriate distribution from which to draw changes. Other authors (e.g., Raup and Gould, 1974; Colwell and Winkler, 1984) offered no justification for the use of any given distribution in punctuational simulations. Colwell and Winkler (1984 p. 352) noted, however, that "a normal random distribution" might represent "a biologically more realistic rule for character change." One justification for using a normal distribution is that if punctuational change following bifurcations is due to bursts of genetic drift at multiple loci, then the distribution of such changes is expected to approach normality (Slatkin and Lande, 1976; J. Felsenstein, pers. comm.). We therefore used a bivariate normal distribution of changes for punctuational as well as for gradual simulations.

In both cases, this distribution was created by first drawing pseudorandom numbers from a uniform distribution, using Borland's Turbo Pascal 4.0 random number generator. These numbers were transformed using a Box–Mueller algorithm, and then combined into a bivariate normal distribution with specified means, variances, and covariance.

For heuristic purposes we performed simulations that yielded means and variances for sets of simulated tips which were comparable to those of real data (genome size: $\bar{x} = 33.6$, $s^2 = 98.4$; and regenerative growth rate: $\bar{x} = 5.4$, $s^2 = 2.8$) obtained from Sessions and Larson's (1987) Table 2 for each of the 15 species on the phylogeny of Figure 2. (For calculating correlations, means and variances of a simulated data set are irrelevant, but would be important for calculation of slopes of linear regressions [cf. Pagel and Harvey, 1988*b*].) A method of moments estimator (developed by E. V. Nordheim) was used to obtain the appropriate variances (as explained in our package of computer programs). To obtain tip values with the desired means, starting trait values were set equal to the means of Sessions and Larson's (1987) data, and means of the bivariate normal probability distribution were set equal to zero. Thus, on average, the number of lineages showing net increase and net decrease in a given trait was equal, and, overall, a clade was expected to show no net directional change.

Maintaining constant means and variances for the bivariate distribution of character changes yields stochastically constant rates of evolution in gradual simulations. This is consistent with assumptions required for Felsenstein's (1985) method to yield accurate significance tests when simple time is used for branch lengths. For punctuational simulations, constant means and variances yield character change that varies in rate (i.e., change per unit time), since changes are drawn not in relation to time but in proportion to number of speciation events.

Analysis.—To compare methods with regard to hypothesis testing, Type I error (the probability of rejecting the null hypothesis when it is in fact true—also termed "empirical size" or simply "size") was calculated as compared to Pearson's r distribution with $N - 2 = 13$ *df* (two-tailed test). For TIPS, $N - 2$ is the usual degrees of freedom. For both variations of Felsenstein's (1985) procedure (FL1 and FL2), a total of $N - 1$ independent contrasts are produced (where N = number of tip values). One *df* is lost in computing the correlation coefficient, leaving a total of $(N - 1) - 1 = N - 2$ *df*. For ME1, $2N - 2$ inferred changes are used to compute a correlation coefficient, whereas for ME2, only N changes are used (from most recent nodes to tips). As discussed above, several *df* are, in effect, lost by both minimum evolution methods, due to the

nonindependence of inferred changes, and the correct degrees of freedom for significance testing are unknown. One possibility would be to use $N - 2$ *df*, where N is the original number of tip values. Thus, we compared all methods to a standard distribution of critical values for a Pearson's r with $N - 2$ *df*.

For each method, we determined the number of correlation coefficients exceeding the critical value given by the standard distribution for Pearson's r (from Zar's, 1984 Table B.16) at $\alpha = 0.005$, 0.010, 0.020, 0.050, 0.100, 0.200, 0.500, and 1.000. We then calculated the difference between the number of correlation coefficients exceeding successive α levels (i.e., between $\alpha = 0.005$ and 0.010, between $\alpha = 0.010$ and 0.020, etc.). These observed differences were compared to expected differences, based on a standard Pearson's r distribution, by using a χ^2 goodness-of-fit test (a total of seven intervals was used, so results were compared to a χ^2 distribution with 6 *df*). Type I error rates for each method were calculated in this way for all simulations in which the null hypothesis was no correlation (i.e., the input correlation was zero).

Power (the ability to detect nonzero relationships when they exist, defined formally as one minus the Type II error rate) was determined for each method by significance testing against its own simulated null distribution. The power of each method (for a two-tailed test with $\alpha = 0.05$) was calculated for each of four alternative hypotheses (input correlation = 0.25, 0.50, 0.75, 0.90). Cochran's Q test was used to compare the power of different methods at each alternative hypothesis, while blocking by simulation ($N = 1{,}000$).

To compare estimation, the results for each method were compared to the input correlation and to the standardized and nonstandardized forms of the realized evolutionary correlation (Fig. 1). The mean deviation (an index of the bias) and the mean deviation squared (MDS, an index of the mean squared error) of each method were calculated for all three evolutionary correlations. Ninety-five percent confidence intervals and nonparametric sign tests were used to test whether mean deviations for each method differed significantly from zero and were distributed symmetrically about zero, respectively. Friedman's test was used to compare the mean deviation and MDS of the different methods. (Friedman's test is a two-factor nonparametric test. For our analyses, one factor was the method, the other factor was the simulated data set, with 1,000 levels.)

As a final index of estimation, we computed coefficients of determination (r^2) between the distribution of correlation coefficients obtained for each method and the two distributions of realized evolutionary correlations. This allowed us to compare methods on the basis of predictive ability of these two forms of correlations. (Note that coefficients of determination are measures of *linear* association only, and are not sensitive to bias; in practice, however, all of the methods we tested proved to be unbiased.) For an approximate statistical test of whether methods differed significantly in predictive ability, we arbitrarily split the simulated data for each input correlation into 10 groups of 100 sets. For each of the 10 groups, we computed r^2 between the results for each method and either the *UREC* or the *SREC*, yielding 90 r^2s for each evolutionary correlation being estimated. We then conducted 2-way ANOVAs without replication, with method (1–9; see Table 1) and "trial" (10 groups of 100 simulations) as factors, to test for differences among the methods while blocking for "trial" effects.

RESULTS

Simulations of Gradual Character Change

For two traits evolving gradually along the phylogeny of Figure 2, only Felsenstein's (1985) method (FL1G) provided acceptable Type I error rates (Table 2; Fig. 5). All other methods tended to overestimate the significance of the observed correlation (Table 2). Power was always highest for FL1G and lowest for TIPS (Table 3).

Biases (as indicated by mean deviation) were always less than 0.025 in magnitude (Tables 2 and 4). Thus, *all* of the methods may, for all practical purposes, be considered unbiased estimators of the input correlation and of the realized evolutionary correlations. (Bias increased and became significant as the input correlation increased

TABLE 2. Statistical comparison of methods for estimating evolutionary correlations. The evolution of two independent traits was simulated 1,000 times along the phylogeny in Figure 3 (from Sessions and Larson, 1987), under a gradual model of change (changes drawn from a bivariate normal distribution of random numbers with means and covariance equal to zero). Means of simulated values across tips of phylogeny were 33.6 and 5.4; mean variances across tips were 99.5 and 2.7. Simulated tip values were analyzed by each method to obtain a distribution of 1,000 correlation coefficients for each method. Two forms of realized evolutionary correlation (see text for explanation) also were obtained from the simulation (*UREC* and *SREC*). Type I error was calculated for several different *P* values (null hypothesis was a two-tailed *r* distribution with $N - 2 = 13$ *df*; critical value at $\alpha = 0.05$ is 0.514, for example) and intervals of Type I error were compared to an expected distribution with a χ^2 goodness-of-fit test ($df = 6$, $\alpha = 0.05$, critical value = 12.6). Mean deviation = (method − *REC*) × 1,000. Mean deviation squared = (method − *REC*)^2 × 1,000.

							Compared to							
							Input correlation		Unstandardized realized evolutionary correlation			Standardized realized evolutionary correlation		
	Percentiles				$\alpha = 0.05$ Type I error	χ^2 Type I error	Mean dev. × 10^3	MDS × 10^3	Mean dev. × 10^3	MDS × 10^3	r^2 (%)	Mean dev. × 10^3	MDS × 10^3	r^2 (%)
Method	2.5	97.5	2.5–97.5	Variance										
TIPS	−0.636	0.694	1.330	0.122	0.159	177.15*	3.53	122.2	2.10	80.6	34.8	2.34	87.3	28.5
FL1G	−0.513	0.507	1.020	0.072	0.046	8.52	−2.47	72.1	−3.91	45.0	42.9	−3.67	36.4	49.5
FL2G	−0.569	0.572	1.141	0.088	0.074	23.84*	−0.57	88.1	−2.00	38.1	57.0	−1.76	52.5	40.4
FL1P	−0.563	0.559	1.122	0.087	0.078	38.01*	−1.95	87.2	−3.38	37.4	57.4	−3.14	51.8	40.6
FL2P	−0.570	0.576	1.146	0.092	0.089	45.07*	−1.51	92.0	−2.94	39.5	57.1	−2.71	56.2	38.9
ME1G	−0.582	0.594	1.176	0.094	0.091	40.63*	−1.08	93.4	−2.51	38.7	58.6	−2.28	58.1	37.9
ME2G	−0.639	0.646	1.285	0.113	0.139	132.67*	2.35	113.2	0.92	55.5	50.9	1.16	79.5	29.8
ME1P	−0.567	0.549	1.116	0.086	0.074	27.96*	−2.26	85.5	−3.70	38.2	55.8	−3.46	50.3	41.2
ME2P	−0.617	0.589	1.206	0.098	0.097	49.12*	−2.29	97.9	−3.72	55.0	44.7	−3.49	64.1	34.6
Friedman's test χ^2							2.08[1]	246.3[2]	2.08[1]	404.0[2]		2.08[1]	515.9[2]	
UREC	−0.463	0.464	0.927	0.057										
SREC	−0.400	0.377	0.777	0.037										

* $P < 0.05$.
[1] 95% confidence intervals for all mean deviations in these analyses included zero.
[2] MDS differs significantly among methods.

TABLE 3. Power of methods for estimating the input correlation of a distribution from which changes are drawn. The power of each method (for a two-tailed test, $\alpha = 0.05$) was calculated for each of four alternative hypotheses (input correlation = 0.25, 0.50, 0.75, 0.90) by comparing the distribution of results for each method with the appropriate simulated null distribution (input correlation = 0, from Table 2). Power is the proportion of correlation coefficients in the alternative distribution that exceeds the critical value (upper or lower 2.5 percentile) provided by the null distribution of that method.

Method	Input correlation of alternative distribution			
	0.25	0.50	0.75	0.90
TIPS	0.082	0.258	0.682	0.972
FL1G	0.161	0.533	0.941	1.000
FL2G	0.130	0.403	0.873	0.997
FL1P	0.143	0.418	0.893	0.999
FL2P	0.134	0.392	0.864	0.997
ME1G	0.117	0.358	0.846	0.996
ME2G	0.097	0.281	0.747	0.983
ME1P	0.155	0.440	0.902	0.999
ME2P	0.135	0.381	0.849	0.993
Cochran's Q	131.05*	641.49*	848.25*	124.04*

* Powers differ significantly among methods ($P < 0.05$).

in magnitude. However, this is to be expected due to the asymmetry of the distribution of a correlation coefficient.)

Methods differed significantly in terms of both mean deviation squared (MDS) and coefficients of determination (r^2) [Friedman's tests for MDS, Tables 2 and 4; two-way ANOVAs for r^2: $P < 0.001$ for both the unstandardized realized evolutionary correlation (*UREC*) and the standardized realized evolutionary correlation (*SREC*) for all input correlations]. TIPS always yielded the worst estimate (highest MDS and lowest r^2) of all three types of correlations, whereas FL1G was the best predictor (lowest MDS and highest r^2) of the input correlation and of the *SREC*. With one exception, ME1G always provided the highest r^2 with the *UREC*, but the results for MDS were inconsistent (Tables 2 and 4).

Simulations of Punctuational Character Change

Results for significance testing were quite different for simulations of punctuational change along the phylogeny of Figure 2 with an input correlation of zero. FL2G, FL1P, FL2P, ME1G, and ME1P *all* yielded acceptable Type I error rates, whereas TIPS, FL1G, ME2G, and ME2P yielded excessively high levels (Table 5). As in the simulations of gradual change, Cochran's Q test demonstrated significant differences in power among the methods at all input correlations except 0.25 ($P < 0.0001$; results not shown). At input correlations of 0.50 and 0.75, TIPS showed the lowest power, ME2G and ME2P showed somewhat higher power, and the other methods showed still higher power. At an input correlation of 0.90, all methods showed approximately equal power, with the exception that TIPS was lower than the rest.

Results for statistical estimation were similar to those obtained with simulations of gradual change. Mean deviations were small (<0.030), and all methods may be considered unbiased estimators of the three evolutionary correlations. TIPS again provided the worst estimate (the greatest MDS and the lowest r^2) in estimating all three types of correlations. FL1P and ME1P, which correctly assume change is punctuational, always gave the best estimates (the lowest MDS and the highest r^2) of both the input correlation and the *UREC*. More surprisingly, FL1G, which *incorrectly* assumes that change has been gradual, consistently provided the best estimate of the *SREC* (Table 5 for input correlation of zero; other results not shown).

Simulations on Phylogenies with Extreme Branch Lengths

Simulations of the gradual evolution of traits along the two phylogenies depicted in Figure 3 serve to illustrate the effectiveness of methods under different extremes of phy-

logenetic nonindependence (power calculations were not completed for these simulations). (Note that under a punctuational model, all three versions of the plethodontid phylogeny [Figs. 2 and 3] are equivalent.)

As expected, for data simulated along what is almost a binary "star" phylogeny (Fig. 3a), Type I error rates were acceptable for TIPS and for *all* of the methods that assume gradual evolution (FL1G, FL2G, ME1G, and ME2G) (Table 6). In contrast, all four of the methods that incorrectly assume punctuational change (FL1P, FL2P, ME1P, and ME2P) exhibited excessively high Type I error rates. Similarly, although values for both mean deviation and MDS for all methods were quite small (MD < 0.030; MDS ranging from 0.006 to 0.108; cf. Tables 2 and 4), they were lower for TIPS and for those methods that assume gradual change than for those methods that incorrectly assume punctuational change. r^2 values were relatively large with the *UREC* (ranging from 83 to 92% for TIPS and for all of the methods that assume gradual change, but from 61 to 81% for all of the methods that assume punctuational change), but somewhat smaller with the *SREC* (ranging from 34 to 49% for methods assuming gradual change, with methods assuming punctuational change again doing somewhat worse).

On the phylogeny that aggravates the problem of nonindependence (Fig. 3b), only FL1G provided acceptable Type I errors (Table 6). Not surprisingly, all of the other methods yielded excessively high Type I error rates (Table 6), as they do not adequately correct for nonindependence of data points. Again, all methods provided unbiased estimators of all three evolutionary correlations. In this case, however, high MDS (ranging from 0.035 to 0.291) and low r^2 values (ranging from 7 to 42%) were obtained for all methods estimating all three evolutionary correlations, with the exception of FL1G estimating the input correlation and the *SREC*. Thus, increasing phylogenetic nonindependence of species values led to the decreased ability of *all* methods except FL1G to estimate any of the three evolutionary correlations. Differences among the methods were significant (Friedman's tests for MDS, two-way ANOVAs for r^2; $P < 0.05$ in all cases; results not shown),

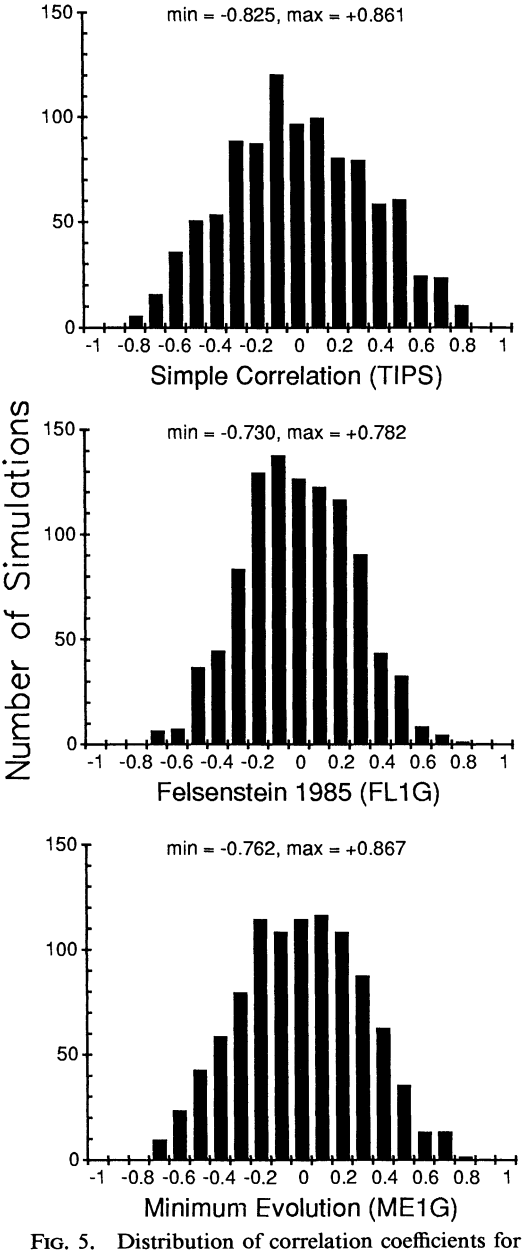


FIG. 5. Distribution of correlation coefficients for simulations under a gradual model of change, with an input correlation of zero, and done on the phylogeny depicted in Figure 2. (a) Simple Pearson product–moment correlation across 15 tip values ("TIPS"). (b) Correlation based on 14 standardized independent contrasts (Felsenstein, 1985; "FL1G"). (c) Pearson product–moment correlation based on 28 inferred changes between nodes and between nodes and tips ("ME1G"). With $N - 2 = 13$ *df* (where N = the number of tips), correlations >0.514 in magnitude would be judged significant at $P < 0.05$ for a 2-tailed test. Only FL1G yields acceptable Type I error rates; TIPS and ME1G yield excessively high Type I error rates (see Table 2).

TABLE 4. Bias (indicated by mean deviation), mean squared error ([mean deviation]^2), and predictive value (r^2) of various methods for estimating three types of evolutionary correlations. The evolution of two traits was simulated 1,000 times along the phylogeny in Figure 2, under a gradual model of change. Input correlation of the bivariate normal distribution of changes varied between 0.25 and 0.90.

		Input correlation											
		0.25			0.50			0.75			0.90		
Method		Input	*UREC*	*SREC*	Input	*UREC*	*SREC*	Input	*UREC*	*SREC*	Input	*UREC*	*SREC*
TIPS	Dev. × 1000	−7.9	−0.1	−11.9	−22.7[1,2]	−8.7[2]	−16.7[1,2]	−19.8[1,2]	−8.7	−11.6	−7.7[1,2]	−2.3[2]	−4.2[2]
	MDS × 1000	117.3	77.2	85.6	82.3	59.1	65.1	32.4	22.0	24.5	6.8	5.4	5.7
	% r^2		34.7	27.1		29.1	21.0		31.7	24.0		22.2	17.0
FL1G	Dev. × 1000	−8.8	−0.9	−12.8[1]	−8.9[2]	5.1	−3.0	−11.2[1,2]	−0.1	−3.0	−5.2[1,2]	0.1	−1.7[2]
	MDS × 1000	64.4	42.4	33.6	42.2	32.8	23.7	16.0	10.0	8.2	2.9	2.3	1.7
	% r^2		40.8	48.0		33.0	43.8		41.6	48.5		33.2	40.9
FL2G	Dev. × 1000	−10.2	−2.4	−14.2	−17.0[1,2]	−3.0	−11.0	−16.6[1,2]	−5.5[2]	−8.4[1]	−6.3[1,2]	−1.0[2]	−2.8[2]
	MDS × 1000	84.8	39.5	53.4	55.6	28.9	36.4	21.8	10.0	13.9	4.3	2.6	3.1
	% r^2		53.7	37.2		48.6	34.4		53.5	35.9		41.6	28.4
FL1P	Dev. × 1000	−13.4	−5.6	−17.4[1]	−14.8[1,2]	−0.8	−8.8	−16.5[1,2]	−5.4[2]	−8.4[1,2]	−6.6[1,2]	−1.3[2]	−3.1[2]
	MDS × 1000	82.4	36.9	50.9	52.6	26.5	33.5	21.5	9.8	13.4	4.2	2.3	2.9
	% r^2		55.4	38.4		50.4	36.2		53.7	37.0		44.9	30.7
FL2P	Dev. × 1000	−13.7	−5.8	−17.7	−16.1[1,2]	−2.1	−10.2	−17.2[1,2]	−6.1[2]	−9.0[1,2]	−7.0[1,2]	−1.7[2]	−3.5[2]
	MDS × 1000	87.2	39.0	56.0	56.4	27.9	37.4	22.9	10.7	14.9	4.6	2.6	3.3
	% r^2		55.3	35.9		50.7	33.7		52.9	34.5		43.3	28.0
ME1G	Dev. × 1000	−11.7	−3.8	−15.7[1]	−17.6[1]	−3.6[2]	−11.6	−18.6[1,2]	−7.5[1]	−10.4[1,2]	−7.3[1,2]	−2.0[2]	−3.8[1,2]
	MDS × 1000	86.7	38.0	56.0	57.5	28.0	38.2	23.5	10.5	15.4	4.7	2.6	3.4
	% r^2		56.2	35.6		51.5	33.6		55.0	34.0		44.7	27.7
ME2G	Dev. × 1000	−14.9	−7.1	−18.9[1]	−22.0[1,2]	−8.0[2]	−16.0[1]	−24.4[1,2]	−13.4[1,2]	−16.3[1]	−10.9[1,2]	−5.6[1,2]	−7.4[1,2]
	MDS × 1000	100.4	51.2	72.0	69.7	38.2	49.6	30.7	16.1	22.1	6.4	3.9	4.9
	% r^2		49.0	28.6		45.0	28.8		47.7	27.6		39.0	22.5
ME1P	Dev. × 1000	−14.5	−6.6	−18.5[1]	−13.5[2]	0.5	−7.6	−15.4[1,2]	−4.4[2]	−7.3[1,2]	−6.4[1,2]	−1.1[2]	−2.9[2]
	MDS × 1000	81.4	37.4	49.6	51.2	26.6	32.1	20.9	9.9	12.8	4.1	2.3	2.8
	% r^2		54.4	39.3		49.2	37.2		52.2	38.2		43.7	31.7
ME2P	Dev. × 1000	−17.7	−9.8	−21.6	−14.0	0.0	−8.0	−18.4[1,2]	−7.4[2]	−10.3[1,2]	−8.4[1,2]	−3.1[2]	−4.9[1,2]
	MDS × 1000	89.9	51.4	58.9	56.6	36.3	37.5	24.4	14.3	16.2	5.0	3.2	3.5
	% r^2		43.8	34.8		38.3	33.7		41.1	32.9		36.0	28.3
Friedman's test χ^2	Dev. × 1000	15.6[3]	15.6[3]	15.6[3]	7.2	7.1	7.1	18.1[3]	18.1[3]	18.1[3]	49.2[3]	49.1[3]	49.1[3]
	MDS × 1000	175.7[3]	311.5[3]	447.5[3]	22.5[3]	366.8[3]	440.3[3]	18.9[3]	362.1[3]	526.8[3]	49.1[3]	301.4[3]	481.3[3]

[1] 95% CI do not include 0.
[2] Sign test, distributions asymmetrical about 0.
[3] Methods differ significantly.

TABLE 5. Statistical comparison of methods for estimating evolutionary correlations. The evolution of two independent traits was simulated 1,000 times along the phylogeny in Figure 2, under a punctuational model of change (changes drawn from a bivariate normal distribution of random numbers with means and covariance equal to zero). Means of simulated values across tips of phylogeny were 33.6 and 5.3; mean variances across tips were 98.6 and 2.7. See Table 2 heading for further explanation.

							Compared to							
							Input correlation		Unstandardized realized evolutionary correlation			Standardized realized evolutionary correlation		
	Percentiles				$\alpha = 0.05$ Type I	χ^2 Type I	Mean dev.	MDS	Mean dev.	MDS	r^2	Mean dev.	MDS	r^2
Method	2.5	97.5	2.5–97.5	Variance	error	error	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	(%)	$\times 10^3$	$\times 10^3$	(%)
TIPS	−0.640	0.663	1.303	0.127	0.158	234.06*	−5.24	126.7	1.44	93.9	26.1	0.60	113.1	16.9
FL1G	−0.582	0.579	1.161	0.086	0.075	29.91*	−2.63	85.7	4.05	52.5	39.1	3.21	55.8	38.3
FL2G	−0.512	0.551	1.063	0.077	0.061	4.65	0.24	76.5	6.92	41.4	46.1	6.08	61.1	29.6
FL1P	−0.498	0.522	1.020	0.072	0.050	3.93	0.41	72.1	7.09	36.0	50.3	6.25	58.3	30.0
FL2P	−0.501	0.537	1.038	0.074	0.055	3.21	0.19	73.6	6.87	37.6	49.1	6.03	61.8	27.7
ME1G	−0.501	0.534	1.035	0.074	0.058	1.22	1.37	74.0	8.05	37.9	48.9	7.21	59.7	29.7
ME2G	−0.559	0.581	1.140	0.089	0.088	26.92*	1.72	88.7	8.40	53.4	40.0	7.56	75.4	24.1
ME1P	−0.503	0.524	1.027	0.072	0.052	5.47	0.43	72.3	7.11	35.9	50.4	6.27	58.1	30.2
ME2P	−0.576	0.578	1.154	0.090	0.084	131.64*	−0.73	90.1	5.95	54.0	40.2	5.11	75.8	24.4
Friedman's test χ^2							6.42[1]	219.0[2]	6.44[1]	440.0[2]		6.42[1]	252.1[2]	
UREC	−0.388	0.396	0.784	0.039										
SREC	−0.447	0.461	0.908	0.056										

* $P < 0.05$.
[1] 95% confidence intervals for all mean deviations in these analyses included zero.
[2] MDS differs significantly among methods.

TABLE 6. Comparison of Type I error rates for simulations on extreme phylogenies. The evolution of two independent traits was simulated 1,000 times along various phylogenies with either gradual or punctuational change. Values are Type I error rates at $\alpha = 0.05$ (expected = 0.050) and χ^2 and significance for overall deviation of Type I error rates from expected (see text).

	Model of change/simulation topology/analysis topology															
	Gradual/Fig. 3a/ same		Gradual/Fig. 3b/ same		Gradual/Fig. 3a/ Fig. 3b		Gradual/Fig. 3b/ Fig. 3a		Gradual/Fig. 4a/ same		Gradual/Fig. 4b/ same		Punctuational/ Fig. 4a/same		Punctuational/ Fig. 4b/same	
Method	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2	$\alpha = 0.05$	χ^2
TIPS	0.047	2.1	0.422	853.8*	0.047	2.1	0.422	853.8*	0.203	385.2*	0.094	62.8*	0.239	482.1*	0.154	172.5*
FL1G	0.048	1.8	0.054	3.0	0.110	97.7*	0.417	943.7*	0.063	13.0*	0.044	7.1	0.138	179.5*	0.050	2.8
FL2G	0.056	4.2	0.428	986.8*	0.061	11.2	0.424	863.1*	0.061	11.2	0.069	9.2	0.095	48.8*	0.049	5.6
FL1P	0.079	32.2*	0.399	783.8*	0.079	32.2*	0.399	783.8*	0.062	13.1*	0.097	60.5*	0.055	2.7	0.048	5.6
FL2P	0.069	18.0*	0.421	827.1*	0.069	19.2*	0.421	827.1*	0.065	5.7	0.102	65.5*	0.055	2.9	0.051	6.5
ME1G	0.051	3.0	0.460	919.9*	0.071	24.6*	0.422	885.8*	0.064	14.7*	0.095	43.5*	0.062	5.7	0.050	7.6
ME2G	0.051	3.0	0.390	801.6*	0.099	73.9*	0.422	885.8*	0.121	116.2*	0.109	60.4*	0.065	6.3	0.083	49.1*
ME1P	0.078	37.1*	0.371	808.7*	0.078	37.1*	0.371	800.1*	0.062	14.7*	0.103	59.8*	0.061	6.4	0.048	5.9
ME2P	0.118	127.8*	0.314	827.3*	0.118	127.8*	0.314	827.3*	0.115	117.1*	0.139	160.0*	0.084	23.1*	0.090	57.0*

* $P < 0.05$.

and were due almost entirely to the effect of FL1G, which provided much better estimates of the input correlation and the *SREC* (lower MDS and higher r^2), but a much worse estimate of the *UREC* (higher MDS and lower r^2), than did any of the other methods (results not shown).

Robustness with Inaccurate Branch Lengths

To test the robustness of methods given extremely inaccurate information as to branch lengths, data simulated under a gradual model of change along the nearly independent phylogeny (Fig. 3a) were analyzed as if evolution had occurred along the extremely nonindependent phylogeny (Fig. 3b), and vice versa. The same sets of simulated data as were used in the previous section (*Simulations on Phylogenies with Extreme Branch Lengths*) were used in these analyses to allow for comparison of results based solely on misinformation concerning branch lengths.

Simulated Along Binary Star Phylogeny, Analyzed as If Nonindependent Phylogeny.—Not surprisingly, only TIPS and FL2G provided acceptable Type I errors (Table 6). (TIPS is unaffected by inaccurate branch lengths, whereas FL2G uses branch lengths only in calculations of nodes, and not for standardization of the contrasts.) Although mean deviations were again small, estimation of all three evolutionary correlations (as measured by both MDS and r^2) by all methods was worse than when using correct branch lengths in calculations. FL1G showed the largest increase in MDS and the largest decrease in r^2, but these changes only served to bring it within the range of the other methods. TIPS yielded the highest predictive power for both the *UREC* (r^2 = 91%) and the *SREC* (48%). Differences among the other methods were quite small, with r^2 ranging from 62% (ME2P) to 83% (FL2G) for the *UREC,* and from 34% (ME2P) to 45% (FL2G) for the *SREC.*

Simulated Along Nonindependent Phylogeny, Analyzed as If Binary Star Phylogeny.—None of the methods provided acceptable Type I error rates (Table 6). However, all methods once again provided unbiased estimators of the input correla-

tion, the *UREC,* and the *SREC.* MDS again increased (ranging from 0.169 to 0.272) in comparison with analyses performed with accurate branch length information, with FL1G showing the largest increases (for estimation of the input correlation, 0.073 versus 0.258; for estimation of the *SREC,* 0.035 versus 0.234). FL1G also showed the largest changes in r^2 values with both the *UREC* and the *SREC,* decreasing from 51 to 41% with the *SREC* and actually *increasing* from 7 to 41% with the *UREC.* r^2 values with the *SREC* were quite low in all cases (<14%), whereas r^2 values with the *UREC* (with the exception of FL1G) changed little from those obtained for data that were analyzed with accurate information (range = 40–43%).

Simulations on Altered Topologies

Gradual Model of Change.—For the comb-like phylogeny of Figure 4b, both FL1G and FL2G yielded correct Type I error rates (Table 6). For gradual evolution along the symmetrical phylogeny of Figure 4a only FL2G and FL2P yielded acceptable tests of significance (Table 6), but FL1G, FL1P, ME1G, and ME1P all provided marginally acceptable Type I error rates (χ^2 = 13.0, 13.1, 14.7, and 14.7, respectively, versus critical value of 12.6). Thus, considering all results for gradual simulations, only FL1G seems to offer *consistently* accurate significance tests. Other methods, except ME2, sometimes may yield correct Type I error rates, as compared with critical values for a standard Pearson's r with $N - 2$ *df,* but they cannot be relied upon to do so (Tables 2 and 6).

Punctuational Model of Change.—Surprisingly, FL1P, FL2P, ME1G, and ME1P *all* yielded acceptable Type I error rates for simulations of punctuational change on the two phylogenies of Figure 4 (as they did for simulations of punctuational change on the phylogeny of Fig. 2). ME2G yielded acceptable Type I error rates only on the comb-like phylogeny of Figure 4b, whereas FL1G and FL2G (which incorrectly assume change has been gradual) only yielded correct Type I error rates on the symmetrical phylogeny of Figure 4a. TIPS and ME2P again yielded excessively high Type I error rates on both phylogenies (Table 6).

Discussion

Not attempting to take phylogeny into account is statistically unacceptable. As predicted by a number of authors, standard statistical analyses that ignore phylogeny entirely (as in TIPS) yield inflated Type I error rates, low power, and relatively inaccurate estimates of evolutionary correlations (Fig. 1). Any of the other methods we have compared (Table 1), all of which make some attempt to correct for phylogenetic effects, perform better than does TIPS (unless the actual phylogeny is close to a "star" [e.g., Fig. 3a], in which case the methods perform equally well). Even when extremely inaccurate information concerning model of change (gradual vs. punctuational) or relative branch lengths is used, other methods perform no worse than does TIPS. (The single exception occurs when the actual phylogeny is close to a star, but phylogenetic information is extremely inaccurate and claims that species are very nonindependent of each other. In practice, having branch lengths with such an *extreme* degree of inaccuracy seems quite unlikely.)

A second conclusion from our simulations is that ME2, the minimum evolution method that uses only the changes between most recent nodes and tips (and which is most similar to that used by Huey and Bennett [1987]), never performs better and often performs considerably worse than does ME1, which uses changes between inferred nodes as well as changes between nodes and tips. Thus, if one of these minimum evolution methods is to be used, it should be ME1, not ME2. We conclude that neither TIPS nor ME2 should be seriously considered for analyzing comparative data.

Significance Testing.—We demonstrate that *any* of the available methods may be used to obtain accurate Type I error rates, if hypothesis testing is conducted against an empirical null distribution created through computer simulation along the phylogeny of interest and under an appropriate model of change. This may be desirable if a method is preferred for other reasons (e.g., estimation) or if complete phylogenetic information is not available (or considered unreliable). Otherwise, Felsenstein's (1985) method of standardized contrasts (FL1) is

the only method tested that *consistently* provided acceptable Type I error rates, given an accurate phylogeny and model of character change. Under a punctuational model of evolution, FL2P, ME1G, and ME1P also yielded correct Type I error rates (Tables 5 and 6), even when topology was varied drastically (cf. Figs. 3 and 5).

Power.—For data simulated under a gradual model of change, FL1 clearly had higher power than did any other method (Table 3). TIPS and ME2G (neither of which is recommended) had the lowest power; other methods were intermediate. Interestingly, ME1P, which (incorrectly) assumes a punctuational model of change, showed higher power (lower Type II error) than did ME1G.

For data simulated under a punctuational model of change, TIPS again had the lowest power, both versions of ME2 had intermediate power, and all other methods had similar, higher power. There seemed to be little difference in power between those methods that assume gradual change (incorrectly) and those that assume punctuational change.

Statistical Estimation.—We compared methods in terms of their abilities to estimate the input correlation and two types of realized evolutionary correlation (*UREC* and *SREC*; see Methods and Fig. 1). Importantly, all of the methods we compared yield what may be considered unbiased estimates of all three statistics (mean deviation < 0.03 in all cases). Methods varied a great deal, however, in terms of both mean deviation squared (for all three evolutionary correlations) and coefficients of determination (with the two forms of *REC*s). For simulations along the phylogeny of Figure 2, TIPS consistently provided the worst estimate of all three evolutionary correlations at all input correlations and for simulations under both gradual and punctuational models of change (Tables 2, 4, 5).

Felsenstein's (1985) method was intended to estimate the input correlation and, not surprisingly, FL1 consistently showed the lowest MDS in estimating it. In punctuational simulations, ME1P did as well as FL1P (Table 5). FL1 was highly subject to inaccuracy of information concerning both model of change and relative branch lengths (Table 6). However, inaccurate information (in terms either of model of change or of branch lengths) only made FL1 perform as well or as poorly as did the other methods in estimating the input correlation. One thus has little to lose by choosing FL1 to estimate the input correlation. FL1G also provided the best estimate of the standardized realized evolutionary correlation (*SREC*), under all conditions, regardless of the model of change.

The unstandardized realized evolutionary correlation (*UREC*), however, was best estimated by ME1 in both gradual and punctuational simulations. Thus, ME1 seems best suited for the type of evolutionary reconstructions desired by Huey and Bennett (1987). FL1 yielded an equally good estimate of the *UREC* in simulations under a punctuational model of change (FL1P), but was a poor estimator of the *UREC* in simulations under a gradual model of change (FL1G).

The absolute predictive ability of the various methods for either form of realized evolutionary correlation varied depending on the phylogeny and on the model of evolutionary change. Coefficients of determination with the *SREC* were generally lower than with the *UREC* for all methods except FL1G. In general, r^2s always were less than 60% (Tables 2, 4, 5; other results not shown). Only for data simulated gradually along a binary "star" phylogeny (Fig. 4a) did predictive ability become quite high for the *UREC* (r^2 ranged from 76 to 92%, except $r^2 = 62\%$ for ME2P), although r^2s for the *SREC* still ranged from only 41 to 49% (except $r^2 = 34\%$ for ME2P). The foregoing would not seem to represent very good predictive ability in any absolute sense, although r^2 would probably tend to be higher with larger sample sizes and/or phenotypic data showing a greater range.

Unanswered Questions.—We certainly have not investigated all factors that may lead to variation among methods in statistical performance. Our simulations envision evolution as a nondirectional process (a type of "random walk" [or diffusion, in continuous time]) of consistently gradual (stochastically constant average rate) or punctuational change, with a constant relationship between two traits. It would be

possible, however, to vary the model of evolutionary process in a number of ways that might affect the performance of different methods.

A bivariate normal probability distribution with means of zero would be characteristic of evolutionary change occurring as by Brownian motion, which "corresponds well to what we expect if genetic drift is the mechanism of character change" (Felsenstein, 1988 p. 464). As genetic drift is not the only mechanism of character change, it would be of obvious interest to explore the consequences of choosing changes from a log-normal or other probability distribution, rather than from a normal. Variable rates of evolution could be modeled by allowing the variance of the distribution of changes for either or both traits to change during the course of simulations, either randomly or systematically, similar to changing the "step variance" of a random walk (Bookstein, 1987, 1988).

The evolutionary relationship between characters may change over time (Harvey and Mace, 1982; Felsenstein, 1985 p. 14; Huey, 1987; Bell, 1989; but see Pagel and Harvey, 1988*b* regarding potential statistical artifacts). This could be due to changing genetic correlations or to changing patterns of selection, for example. Thus, altering the covariance (input correlation of Fig. 1) of the bivariate distribution of changes within a single simulation would be of considerable interest. It also would be possible to allow the phylogeny itself to vary randomly from simulation to simulation (cf. Raup and Gould, 1974; Fiala and Sokal, 1985). We did not do this simply to avoid confounding effects of variation in the phylogeny per se (topology, distribution of branch lengths) with those due to model of evolutionary change, magnitude of input correlation, etc.

Finally, the statistical and biological underpinnings of the alternative evolutionary correlations (Fig. 1) deserve further study. Various authors clearly seem interested in different correlations, but the distinctions we note have not previously been explicated. Perhaps the input correlation will be most useful for attempting to uncover general biological "laws" that apply to large groups of organisms; for example, attempting to infer what happened in all mammals from studies of a single family. Realized evolutionary correlations, on the other hand, may be most appropriate for those who are interested in the evolution of traits in a restricted study group, such as a particular family of mammals, but do not necessarily wish to generalize to all mammals. Another possibility, as suggested by J. M. Cheverud (pers. comm.), is that the input correlation is closer to microevolutionary processes, whereas the realized evolutionary correlations better estimate macroevolutionary patterns. In any case, we have shown that analytical methods do in fact differ significantly with regard to how well they estimate the input correlation and the two forms of realized evolutionary correlation, one of which (the *SREC*) gives greater weight to changes occurring over short time spans.

Recommendations.—Choice of a method in a comparative study will depend on the availability of phylogenetic information and on the question of primary interest. Phylogenetic information, including branch lengths of various sorts, is becoming available at an increasing rate. So too are alternative comparative methods. We have not compared the statistical properties of *all* available methods, although several of the alternatives are quite similar to those we have compared (see Harvey and Pagel [1991] for a thorough review of currently available methods). All of the programs used herein are available from the authors on request, and may be used to compare new methods on any phylogeny. We emphasize that our results are based on simulations run on a limited number of phylogenies. We therefore recommend that future comparative studies might prudently include at least limited simulations in order to compare methods on the relevant phylogeny.

1. No phylogenetic information available. Although phylogenies may be fundamental to comparative biology (Felsenstein, 1985), they are simply unavailable for many groups of organisms. One might, however, be willing to construct a topology based solely or largely on the available taxonomy, with, for example, one node on an unresolved polytomy for each genus, family, etc., as suggested by Burt (1989), Grafen (1989), and others. This is a reasonable first step, *assuming that taxonomic groups are mono-*

phyletic and that groups at a given taxonomic level are of comparable age. From this point, both Grafen (1989) and Harvey and Pagel (1991) offer modifications of Felsenstein's (1985) method that can be applied with partially unresolved topologies. Alternatively, given a completely resolved topology, choice of analytical method will depend on whether information concerning branch lengths is available.

2. Phylogenetic topology but no branch lengths available. It is important to realize that all of the methods we have considered (except TIPS) require knowledge of branch lengths, in units of expected variance of change. For example, the minimum evolution methods we have outlined require branch lengths for computing nodes and for simulation of appropriate null distributions for significance testing.

One way to "avoid" the need for branch lengths in computations is to assume they are all equal, which is equivalent to assuming that character change has been punctuational (e.g., Huey and Bennett, 1987) *and* that all speciation events are known and have been counted. Given these assumptions, then either Felsenstein's (1985) method (FL1P), FL2P, or a minimum evolution method (e.g., ME1P) can be applied. All of these methods gave acceptable Type I error rates for all of our punctuational simulations (Tables 5 and 6), although this should be verified for phylogenies with other than 15 species. In terms of both power and estimation (for all three evolutionary correlations), FL1P, FL2P, and ME1P are very similar (Table 5; other results not shown).

Several alternatives for estimating branch lengths are available. At the very least, one might use a taxonomically based estimate of branch lengths similar to that suggested by Cheverud et al's. (1985) Figure 1, although this procedure is arbitrary (cf. Gittleman and Kot, 1990). Time, which may be estimated from molecular clock and/or paleontological information, is an appropriate estimate of expected variance of change *if* character change is known to be gradual (Brownian motion). Another possibility would be to use overall rates of DNA sequence change to estimate branch lengths. Such information might be obtained from actual sequence data or from single copy nuclear DNA–DNA hybridization studies. Although branch lengths could be estimated with algorithms that make the restrictive assumption of equal rates, using a pairwise tree-construction algorithm that allows rates to differ among branches would be preferable (cf. Springer and Krajewski, 1989). Thus, if a topology is available from other information (e.g., a cladistic analysis of morphological characters), one might use information on DNA sequence divergence to estimate branch lengths only.

It is also possible to use the characters under study to infer branch lengths. Grafen (1989) offers a maximum likelihood technique for doing so, which he presents in the context of standardizing Felsenstein's (1985) independent contrasts. Alternatively, Harvey and Pagel (1991) suggest not standardizing contrasts but instead using residual analysis and remedial measures such as weighted regression, if necessary. An important area for future research will be developing robust techniques for estimating branch lengths in units of expected variance of change. Also of use would be developing techniques for inferring, solely from neontological data, whether the characters of interest have evolved in a gradual or in a punctuational fashion (cf. Douglas and Avise, 1982; Burt, 1989; Lemen and Freeman, 1989; Mindell et al., 1989).

3. Phylogenetic topology and branch lengths available. If the correct phylogenetic topology and branch lengths are available, then Felsenstein's (1985) method (FL1G) should be applied to obtain the most acceptable significance testing, highest power, and best estimates of either the input correlation or the standardized realized evolutionary correlation. For the best estimate of the unstandardized realized evolutionary correlation, however, ME1G and FL2G should be applied, but these must be significance tested against simulated null distributions. This requires specifying the appropriate model of character change, but this is known, in effect, if branch lengths (in expected variance of change) are available.

4. Two characters strongly correlated with a third. One final suggestion concerns the analysis of two traits that are themselves

strongly correlated with a third. Obvious examples include such characters as brain size or metabolic rate, which scale allometrically with body mass. A common approach is to regress both characters on body mass, compute residuals, then look for correlations between these residuals (e.g., Harvey and Mace, 1982; Garland and Huey, 1987; Garland et al., 1988). Such residuals may be analyzed by techniques that take phylogeny into account (e.g., Bell, 1989). However, *phylogeny should be taken into account during creation of the residuals.* One could, for example, use Felsenstein's method (FL1) to compute standardized independent contrasts for metabolic rate, brain size, and body size. Alternatively, a minimum evolution method (ME1) might be used to compute inferred changes for these traits along a phylogeny (e.g., Losos, 1990). These independent contrasts or inferred changes could then be used to compute separate regressions of metabolic rate on body size and of brain size on body size. Residuals from these two regressions should be free of the confounding effects of both body size *and* phylogeny, and could be tested for correlation using standard procedures (in the case of FL1) or by reference to computer-simulated null distributions (in the case of ME1).

Acknowledgments

We would like to thank the following people for help: J. Felsenstein for frequent advice, motivation, and the encouragement to get us started on this study; E. V. Nordheim for numerous helpful discussions, the method of movements variance estimator, statistical advice, and comments on the manuscript; A. Grafen, P. H. Harvey, W. P. Maddison, and M. D. Pagel for copies of unpublished manuscripts; J. R. Baylis, J. B. Losos, and W. P. Maddison for insightful comments on earlier versions; J. M. Cheverud, J. Felsenstein, J. L. Gittleman, M. Lynch, and an anonymous reviewer for comments on the submitted version of the manuscript. This project was supported by funds from the Wisconsin Alumni Research Foundation (administered by the Graduate School, University of Wisconsin–Madison) to T.G. and by a National Science Foundation Graduate Fellowship to E.P.M.

Literature Cited

Ballinger, R. E. 1983. Life-history variations, pp. 241–260. *In* R. B. Huey, E. R. Pianka, and T. W. Schoener (eds.), Lizard Ecology: Studies of a Model Organism. Harvard, Cambridge, MA.

Bell, G. 1989. A comparative method. Am. Nat. 133:553–571.

Bookstein, F. L. 1987. Random walk and the existence of evolutionary rates. Paleobiology 13:446–464.

———. 1988. Random walk and the biometrics of morphological characters. Evol. Biol. 23:369–398.

Burt, A. 1989. Comparative methods using phylogenetically independent contrasts. Oxford Surv. Evol. Biol. 6:33–53.

Campbell, J. W., D. D. Smith, Jr., and J. E. Vorhaben. 1985. Avian and mammalian mitochondrial ammonia-detoxifying systems in tortoise liver. Science 228:349–351.

Cheverud, J. M., M. M. Dow, and W. Leutenegger. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: Sexual dimorphism in body weight among primates. Evolution 39:1335–1351.

Clutton-Brock, T. H., and P. H. Harvey. 1977. Primate ecology and social organization. J. Zool., London 183:1–39.

———. 1984. Comparative approaches to investigating adaptation, pp. 7–29. *In* J. R. Krebs and N. B. Davies (eds.), Behavioral Ecology: An Evolutionary Approach, 2nd ed. Blackwell, Oxford, U.K.

Coddington, J. A. 1988. Cladistic tests of adaptational hypotheses. Cladistics 4:3–22.

Colwell, R. K., and D. W. Winkler. 1984. A null model for null models in biogeography, pp. 344–359. *In* D. R. Strong, Jr., D. Simberloff, L. G. Abele, and A. B. Thistle (eds.), Ecological Communities, Conceptual Issues and the Evidence. Princeton Univ. Press, Princeton, NJ.

Crook, J. H. 1965. The adaptive significance of avian social organization. Symp. Zool. Soc. London 14:181–218.

Donoghue, M. J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. Evolution 43:1137–1156.

Douglas, M. E., and J. C. Avise. 1982. Speciation rates and morphological divergence in fishes: Tests of gradual versus rectangular modes of evolutionary change. Evolution 36:224–232.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: An alternative to phyletic gradualism, pp. 82–115. *In* T. J. M. Schopf (ed.), Models in Paleobiology. Freeman, Cooper, San Francisco, CA.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

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

———. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. 19:445–471.

Fiala, K. L., and R. R. Sokal. 1985. Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. Evolution 39: 609–622.

Garland, T., Jr., and R. B. Huey. 1987. Testing

symmorphosis: Does structure match functional requirements? Evolution 41:1404–1409.

GARLAND, T., JR., F. GEISER, AND R. V. BAUDINETTE. 1988. Comparative locomotor performance of marsupial and placental mammals. J. Zool., London 215:505–522.

GITTLEMAN, J. L. 1988. The comparative approach in ethology: Aims and limitations. Perspect. Ethol. 8:55–83.

GITTLEMAN, J. L., AND M. KOT. 1990. Adaptation: Statistics and a null model for estimating phylogenetic effects. Syst. Zool. 39:227–241.

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. Phil. Trans. R. Soc. London Ser. B 326:119–157.

GREENE, H. 1986. Diet and arboreality in the Emerald Monitor, *Varanus prasinus,* with comments on the study of adaptation. Fieldiana Zool., new series, no. 31:1–12.

HAILMAN, J. P. 1988. Operationalism, optimality, and optimism: Suitabilities versus adaptations of organisms, pp. 85–116. *In* M.-W. Ho and S. W. Fox (eds.), Evolutionary Processes and Metaphors. Wiley, Chichester, U.K.

HARVEY, P. H., AND G. M. MACE. 1982. Comparisons between taxa and adaptive trends: Problems of methodology, pp. 343–361. *In* King's College Sociobiology Group (eds.), Current Problems in Sociobiology. Cambridge Univ. Press, Cambridge, U.K.

HARVEY, P. H., AND M. D. PAGEL. 1991. The Comparative Method in Evolutionary Biology. Oxford Univ. Press, Oxford, U.K.

HUEY, R. B. 1987. Phylogeny, history, and the comparative method, pp. 76–98. *In* M. E. Feder, A. F. Bennett, W. Burggren, and R. B. Huey (eds.), New Directions in Ecological Physiology. Cambridge Univ. Press, N.Y.

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

KREBS, J. R., AND N. B. DAVIES. 1987. An Introduction to Behavioural Ecology, 2nd ed. Sinauer Associates, Sunderland, MA.

LARSON, A. 1984. Neontological inferences of evolutionary pattern and process in the salamander family Plethodontidae. Evol. Biol. 17:119–217.

LAUDER, G. V. 1981. Form and function: Structural analysis in evolutionary morphology. Paleobiology 7:430–442.

———. 1986. Homology, analogy, and the evolution of behavior, pp. 9–40. *In* M. H. Nitecki and J. A. Kitchell (eds.), Evolution of Animal Behavior. Oxford Univ. Press, N.Y.

LEMEN, C. A., AND P. W. FREEMAN. 1989. Testing macroevolutionary hypotheses with cladistic analysis: Evidence against rectangular evolution. Evolution 43:1538–1554.

LOSOS, J. B. 1990. Concordant evolution of locomotor behaviour, display rate, and morphology in *Anolis* lizards. Anim. Behav. 39:879–890.

LYNCH, M. 1991. Methods for the analysis of comparative data in evolutionary ecology. Evolution. *In press.*

MADDISON, W. P. 1990. A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539–557.

———. 1991. Squared-change parsimony reconstructions of ancestral states for continuous-valued characts on a phylogenetic tree. Syst. Zool. *In press.*

MAYNARD SMITH, J. 1983. The genetics of stasis and punctuation. Annu. Rev. Genet. 17:11–25.

MINDELL, D. P., J. W. SITES, JR., AND D. GAUR. 1989. Speciational evolution: A phylogenetic test with allozymes in *Sceloporus* (Reptilia). Cladistics 5:49–62.

PAGEL, M. D., AND P. H. HARVEY. 1988*a*. Recent developments in the analysis of comparative data. Quart. Rev. Biol. 63:413–440.

———. 1988*b*. The taxon-level problem in the evolution of mammalian brain size: Facts and artifacts. Am. Nat. 132:344–359.

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

RIDLEY, M. 1983. The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating. Clarendon, Oxford, U.K.

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.

RUBEN, J. A., AND A. F. BENNETT. 1980. The vertebrate pattern of activity metabolism: Its antiquity and possible relation to vertebrate origins. Nature (London) 286:886–888.

SCHOPF, T. J. M. 1981. Punctuated equilibrium and evolutionary stasis. Paleobiology 7:156–166.

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.

SLATKIN, M., AND R. LANDE. 1976. Niche width in a fluctuating environment—density independent model. Am. Nat. 110:31–55.

SPRINGER, M. S., AND C. KRAJEWSKI. 1989. Additive distances, rate variation, and the perfect-fit theorem. Syst. Zool. 38:371–375.

SWOFFORD, D. L., AND W. P. MADDISON. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87:199–229.

WILLIAMSON, P. G. 1981. Paleontological documentation of speciation in Cenozoic molluscs from Turkana basin. Nature (London) 293:437–443.

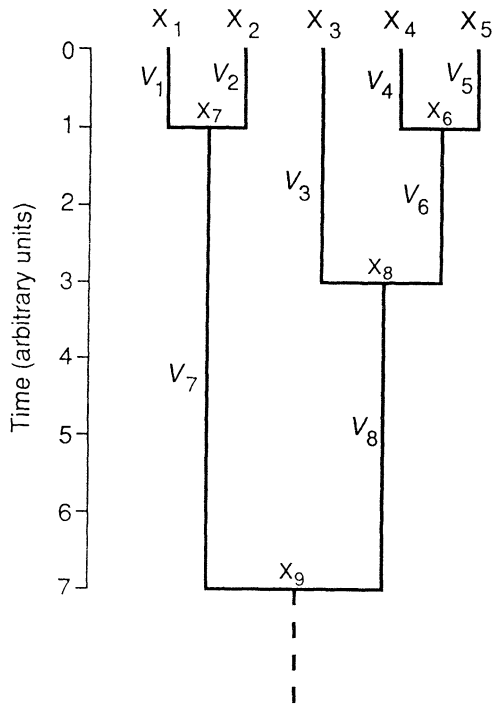
ZAR, J. H. 1984. Biostatistical Analysis, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Corresponding Editor: M. R. Lynch

Appendix A

Character A: 15 9 18 35 28

Character B: 1 2 3 4 5



For purposes of illustration, we have calculated correlation coefficients using the topology, branch lengths, and tip data depicted above. We have also calculated correlations for the phylogeny and data on genome size and growth rate given in Sessions and Larson's (1987) Table 2 for 15 species.

Method	Correlation coefficients	
	Sample	Sessions and Larson
TIPS	0.789	−0.365
FL1G	−0.016	−0.486
FL2G	0.584	−0.510
FL1P	0.408	−0.579
FL2P	0.532	−0.554
ME1G	0.498	−0.584
ME2G	−0.277	−0.760
ME1P	0.388	−0.581
ME2P	−0.140	−0.696