

- >Martins, E.P. and T. F. Hansen. 1996. The statistical analysis of
- >interspecific data: a review and evaluation of phylogenetic comparative
- >methods. IN: Phylogenies and the Comparative Method in Animal Behavior.
- >Oxford University Press (E. Martins, ed.). Oxford University Press.

CHAPTER 2

The Statistical Analysis of Interspecific Data: A Review and Evaluation of Phylogenetic Comparative Methods

Emília P. Martins and Thomas F. Hansen

In the last few decades, the comparative method has undergone a virtual renaissance in evolutionary biology as researchers develop new ways to incorporate taxonomic and phylogenetic information into the design and analysis of interspecific data (see Brooks & McLennan 1991; Harvey & Pagel 1991; McKittrick 1993; Miles & Dunham 1993; and Maddison 1994 for recent reviews). Although phylogenies were an integral part of many classical ethological studies, their use has been largely neglected by other fields in the study of animal behavior. Recent advances in systematics and evolutionary biology have shown the critical importance of incorporating phylogenies into comparative studies in all areas of biology. From a statistical perspective, phylogenies are needed to transform comparative data so that they do not violate the assumptions of standard statistical analyses (e.g., regression, ANOVA, chi-squared tests). From a more biological perspective, phylogenies and new phylogenetic comparative methods (PCMs) allow researchers to infer the patterns and processes of character evolution from the patterns observed in data measured from extant species.

The first part of this chapter is a discussion of some general issues involved in comparative analysis. The second part reviews and critically

evaluates most of the PCMs available today, and is intended as a reference manual. We discuss the ideas, assumptions and interpretations of the methods and evaluate them from a statistical point of view. We do not give the kind of technical description necessary for use of the methods. Although there are many excellent reviews of comparative methods, there is no substitute for reading the original papers when it comes to application. In the Discussion, we explore the role and importance of statistical models in any comparative analysis.

Why phylogenetic comparative methods are used

A. To solve the statistical problem of dependent data.

One reason for incorporating phylogenetic information into a comparative study is to address the problem of statistical dependence. This is the “degrees of freedom” or “effective sample size” problem that has been mentioned frequently in the recent evolutionary and systematics literature. In statistics, the accuracy of a parameter estimate or hypothesis test depends on the number of degrees of freedom available, which in turn depends on the effective sample size of measured data. This effective sample size depends on the number of *independent* data points in the study rather than on the number of samples taken. For example, if 15 data points are measured from one animal and 15 more are measured from a second, the effective sample size is somewhat less than 30, because individual animals are often consistently different from one another (e.g., Martin & Kraemer 1987; Boake 1989), and only two individuals were measured. This “pooling fallacy” (Machlis et al. 1985) occurs when statistics are conducted as if the effective sample size were the number of data points measured (in this case, 30) and the dependence due to individual differences in behavior has been ignored. Standard errors can be greatly underestimated by this problem, such that spurious patterns may be found. Similar sorts of pooling fallacies can occur whenever important factors (e.g., individual identity, sex, preferred habitat, body size) are left out of the models used in statistical analyses. Whenever these missing variables are correlated with the variables of interest, excluding them from the analysis can lead us to judge a pattern unimportant even

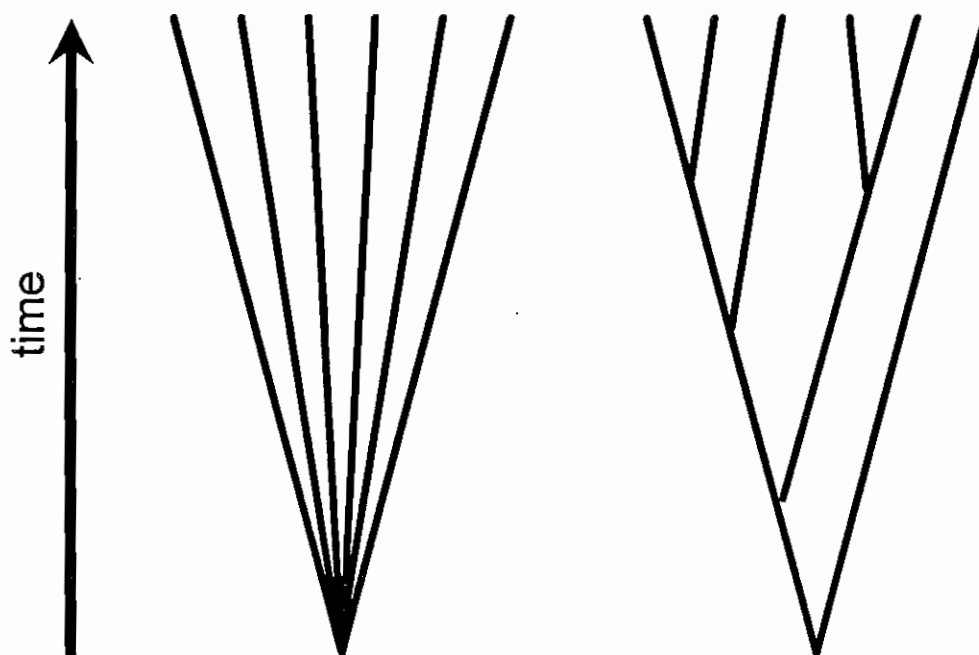


Figure 1. The phylogeny on the left depicts the situation assumed by most statistical analyses of comparative data. Species arise instantaneously from a single common ancestor in a “star” radiation. The phylogeny on the right depicts a more realistic situation as would be allowed by any one of a number of available phylogenetic comparative methods. Although there are six species in this group, the effective sample size of a study in which the data consist of measurements of each of these species will be somewhat less than six.

though it is important. This is particularly aggravated if the pattern of character variation differs among different phylogenetic or taxonomic groups and phylogeny is not taken into account.

Comparative data serve as an illustration of a particularly common and frequently ignored pooling fallacy that can occur whenever multiple data points are collected from the same genus, family, or phylogenetic clade. Closely related species are often more similar to each other than they are to distantly related species because of their shared evolutionary histories. In many studies, however, data are pooled as if there were no dependence among species due to phylogenetic history. Unless the data are transformed in some way to take phylogenetic information into account, the number of species measured in a set of interspecific data is likely to be an inflated estimate of the number of independent data points or effective sample size for the study (Fig. 1).

It has been shown both analytically and empirically that dependence of this sort in comparative studies can lead to serious statistical problems (e.g., Felsenstein 1985; Grafen 1989; Martins & Garland 1991b; Gittleman & Luh 1992, 1993; Martins in review). Accuracy of estimation is usually decreased and hypothesis tests are inadequate. For example, a Pearson correlation coefficient calculated between two traits measured in several different species is likely to exaggerate the absolute magnitude of the relationship between those two traits such that an association is found much more often than expected by chance alone when the two traits are actually independent of one another. Phylogenetic comparative methods allow incorporation of available taxonomic or phylogenetic information into the calculation of correlation coefficients and other statistics, thereby taking the degree of statistical dependence into account and greatly improving the reliability of parameter estimates and hypothesis tests.

B. To answer evolutionary questions

Dependence due to phylogenetic relationships among species in comparative data contains information about the evolutionary process. In general, we expect there to be an underlying correspondence between the phenotypes of existing species and the historical patterns of speciation. We will refer to this correspondence as *phylogenetic correlation*. Phylogenetic correlation is what allows us to reconstruct phylogenies from morphological or molecular data. When the phylogeny is known, we can also use any phylogenetic correlation that might exist in other traits (e.g., behavior) to infer the form and direction of phenotypic evolutionary change along that phylogeny. To do this adequately, we must first consider why phylogenetic correlation arises.

Phylogenetic correlation, as discussed by Hansen & Martins (in press; Martins & Hansen 1995), arises through common ancestry. Phenotypes of extant species are expected to be correlated with those of their ancestors. Because all species share some common ancestors, the phenotypes of extant species are also expected to be correlated with each other. Usually, the similarity between the phenotype of a species and that of its ancestor decreases with time as mutations appear and species phenotypes respond to natural selection in a changing environment. In general, we expect distantly related species to be less

similar to each other than are closely related species, and thus exhibit less phylogenetic correlation. Phylogenetic correlation is expected for complex behavior patterns and behavior patterns governed by complex developmental rules (e.g., communicative displays), because such traits are often slowly evolving and constrained in the changes that can occur. Similarly, behavior patterns that are of extreme importance in a wide variety of environmental conditions (e.g., antipredator strategies, parent-offspring interactions) may be resistant to evolutionary change, and may therefore retain more phylogenetic correlation.

On the other hand, many behavioral and life history traits show high levels of within species variation and are capable of rapid adaptation and smooth tracking of changes in the environment. Such traits may be more influenced by recent selective history than by ancient conditions and are therefore not likely to show phylogenetic correlation directly. This is illustrated by the success of optimality models in behavioral ecology (e. g., Krebs & Davies 1987; Parker & Maynard Smith 1990) which predict species behavior based on present day conditions (without regard to phylogeny). However, the constraint functions that are an integral part of the optimality models may themselves be subject to phylogenetic correlation, and even adaptive traits may be phylogenetically correlated due to indirect mechanisms. Phylogenetic correlation may also be caused by forces other than direct evolution when animals seek out and live in environments similar to those of their ancestors. In these cases, selective constraints are inherited along with the environment, and similarities among species due to their shared evolutionary history results. In conclusion, we expect some degree of phylogenetic correlation in most comparative data sets, and the burden of proof is on those who wish to argue for its absence.

Even when the goals of the comparative study are not explicitly evolutionary, it is important to consider the sources of phylogenetic correlation. Phylogenetic comparative methods are usually based on either statistical or evolutionary models that describe how the phylogenetic correlation of comparative data came to be. Some methods (e.g., phylogenetic autocorrelation, nested ANOVA; see below) are designed to find and incorporate whatever phylogenetic "effect" is present in the data, regardless of how it arose. Other methods (e.g., independent contrasts, microevolutionary model-based methods; see below) are based on explicit models of phenotypic evolution which

assume that the actual causes of phylogenetic correlation are known and are of a particular sort. Which sort of method is more appropriate depends on the details of the biological question of interest. Phylogenetic comparative studies can be used to answer questions such as: (a) Is a behavior pattern phylogenetically constrained? If so, by how much? (b) What was the ancestral state of the behavior? (c) How quickly did the behavior evolve? (d) In what order did a set of behavior patterns appear? Did they appear before or after some other trait? (e) Did two behavior patterns evolve in a correlated fashion, or did a behavior evolve in association with ecological factors? (f) Did a behavior pattern evolve neutrally via random genetic drift alone, or was it subjected to selection? Our ability to answer these questions will often depend on the development of a specific method for addressing that question. When the sources of phylogenetic correlation are not explicitly stated in the description of a specific method, problems of interpretation may arise (Frumhoff & Reeve 1994; Leroi et al. 1994; Martins & Hansen 1995; Nee et al. Chap. 13, in this volume).

Not taking phylogenetic information into account in standard statistical analyses is equivalent to assuming that the phylogenetic relationships among species are irrelevant and that any phylogenetic correlation is insignificant. The differences between a mouse and a dove due to one being a mammal and the other a bird may be unimportant, for example, if all species in the study arose simultaneously from a single common ancestor in a "star" radiation (Fig. 1). In this case, phylogenetic history does not lead to the sorts of dependence described above because all of the measured species are related to each other to the same degree. The assumption that phylogenetic relationships are irrelevant may also be reasonable if the response to selection is fast in comparison to the rate of speciation, such that historical constraints are quickly erased from species phenotypes. In any of these latter situations, our ability to estimate parameters of the evolutionary process and to answer the questions listed above will also be quite limited.

Getting comparative data

The first step in conducting a phylogenetic comparative study is to obtain a set of comparative or interspecific data. The richest source of comparative data is the library. With animal welfare, conservation, and

funding concerns limiting the use of animals in behavioral research, comparative or metaanalyses have become increasingly important as ways of generating hypotheses and gathering preliminary evidence. The study of animal behavior has existed for long enough now that many important data have already been collected and can be reused through a comparative approach to provide fresh insights into old problems in animal behavior. In other cases, data must be collected specifically for the comparative project from laboratory or field experiments.

Whether collecting from the field, laboratory or library, there are a number of issues researchers need to be concerned with in choosing and collecting comparative data. Several of these problems occur with any "metaanalysis" when data are combined from separate studies done by different researchers. First, data collected in different studies are likely to vary in their statistical and biological reliability due to differences in data collection methods and numbers of individuals sampled. This variability among data sets should be taken into account whenever possible, possibly by weighing the data from different studies by their reliability. Second, most experiments are designed to address only a few specific questions, and often make small, hidden assumptions that may be inconsequential to the aims of the original study but are crucial to the comparative study. Each data point in a metaanalysis must be carefully researched to ensure that the data are appropriate to answer the question at hand. Gaillard et al. (1994) provide an illustration of such problems in an excellent discussion of this issue. Finally, a sample of results gleaned from the behavioral and ecological literature is almost certainly biased because of the unfortunate tendency to think that results that do not show statistically significant patterns are not interesting and therefore not publishable. When these results are compiled in a metaanalysis, the final conclusions are likely to be an exaggerated view of the real world. This problem is particularly aggravated in metaanalyses based on studies that report results only as the outcome of statistical tests or as p values. This is only one of many reasons that we support efforts to publish raw data whenever possible and highly recommend that comparative biologists restrict their statistical analyses of combined data sets to situations in which raw data or at least parameter estimates are available.

Even when all of the data for a comparative study are collected by a single researcher, other issues need to be considered. First, when data

are collected from several species, it is not always a trivial problem to determine that the same behavior pattern is being measured in all the different species. This is the classical problem of homology, which has been discussed at length in the evolutionary systematics literature (Hall 1994; in this volume, Lauder & Reilly, Chap. 4; de Queiroz & Wimberger, Chap. 7; Irwin, Chap. 8). Ontogenetic, physiological, and morphological information about the mechanistic bases of the behavior pattern may be useful in ensuring that the behavior of interest is essentially the same pattern with variations for different species. Second, most phylogenetic comparative methods still assume that there is no variation among individuals or populations within a species and that a single number (i.e., the species mean phenotype) can represent the entire species (in this volume, see Foster & Cameron, Chap. 5 for further discussion). If there are considerable sex or age differences in the behavior, it may be important to analyze data for the different sexes or age groups separately. Often, it will also take substantial effort to find the right measures (e.g., mean, median, coefficient of variation) to describe each species effectively. Finally, as in any other behavioral study, it is important to remember that body size or any other confounding variables should, as always, be measured and taken into account in the comparative study. Phylogenetic correlation in these confounding variables should also be incorporated into the analysis.

Throughout this chapter, we refer to data measured in different species, but much of the discussion is also true whether the data are measured from different populations, genera, or even genetically related individuals within populations. Evolutionary biology suggests that all taxa are related to one another. Thus, data measured from any living organisms are likely to be dependent to some extent or another. The relative importance of this dependence to statistical or biological concerns will depend on the details of the questions being considered and the underlying microevolutionary processes. Although the evolutionary processes leading to the relationships among species will differ from those that cause relationships among different populations or individuals, there will often be a direct extension of interspecific methods that can be used at the population or individual level. When doing analyses at levels other than the species level, though, special care must be taken in the interpretation of results.

Getting phylogenetic information

Phylogenies can be obtained from a number of different sources, and means of doing so have been discussed at length (see Hillis & Moritz 1990; Felsenstein 1988 for reviews). At best, direct fossil evidence will exist showing the patterns and timing of speciation events in the clade. More commonly, phylogenetic hypotheses derived from morphological or molecular information will be available, and branch lengths on the tree will be given in units of genetic sequence divergence or the minimum number of morphological evolutionary changes required. In some cases, estimates of the time since divergence of various species on the phylogeny (branch lengths) will also be available. At worst, taxonomic information can be used to infer hierarchical topological relationships (assuming that all large taxonomic groups are monophyletic and that species within genera, genera within families, families within orders, and so on have evolved independently of one another).

All statistical analyses of interspecific data require that some phylogenetic information be either available or assumed. In most cases, it is assumed that both the phylogenetic relationships among species (i.e., a phylogenetic topology) and the process of phenotypic evolution underlying the particular character(s) that have been measured (usually described as branch lengths on the tree in units of expected variance or amount of phenotypic change) are known. For example, most standard statistical methods (without making any attempt to incorporate phylogenetic information) assume that: a) the similarities among species expected due to phylogenetic relationships can be described as a "star" phylogeny (e.g., Fig. 1a) and b) the character has evolved at the same rate and for the same length of time along each branch of this topology. Other methods (see below) allow for more flexibility of assumptions and for information from molecular phylogenies or quantitative genetic experiments to be incorporated into the analysis. The accuracy of a phylogenetic comparative study depends on the accuracy of the phylogenetic information provided, so it is well worth the effort to find a reasonable tree. ¶

Once a phylogenetic topology is obtained, branch lengths must be estimated, inferred, or assumed. In most systematics studies, branch lengths are reported in units of time. For phylogenetic comparative

methods, however, branch lengths are usually needed in units of the relative amount of change expected in the characters being analyzed (these units are often referred to as "expected variance of character change"). This expected amount of change will be a function of the evolutionary process (e.g., random genetic drift, selection) underlying the characters of interest as well as the phylogeny, and usually branch lengths must be transformed from units of time, sequence divergence, or the minimum number of evolutionary changes into these character-specific units. For example, if phenotypic evolution occurs as a gradual, "clocklike" process (e.g., if the character is evolving only by random genetic drift), branch lengths in units of expected variance of change will be directly proportional to branch lengths in units of time. Alternatively, if phenotypic evolution is thought of as a punctuational or "burstlike" process in which the character undergoes bursts of change at speciation events followed by long intervals of stasis, then branch lengths in units of expected variance of change will be proportional to the number of speciation events occurring along each branch. If a phylogeny based on molecular or morphological information is available, we might also assume that branch lengths in units of expected variance of change are proportional to branch lengths in units of phenetic similarity (usually measured as sequence divergence or number of evolutionary changes). This assumes that the characters used to reconstruct the phylogeny evolved at a rate that is linearly proportional to the rate at which the characters in the comparative study evolved.

There are a number of statistical procedures that infer phylogenetic topologies or branch lengths in units of expected variance of change using the information available in the comparative data themselves. The resulting branch lengths have intrinsic interest, as they are essentially a description of the microevolutionary process underlying phenotypic change. Because most of these procedures are specific to one type of phylogenetic comparative method or another, we will discuss them individually below. They should be used cautiously because they are generally conservative. There is only so much variability present in a set of comparative data. These procedures choose to use that variation first to determine phylogenetic relationships and only afterwards to infer adaptive evolution. With convergent evolution, for example, these procedures first try to explain the similarity between species as being due to shared evolutionary history and only afterwards as the result of a

response to similar selective pressures. The resulting estimates of evolutionary parameters may be conservative or even biased. For the same reason, phylogenetic information should be derived independently of the characters used in the comparative study whenever possible.

Often, several competing phylogenetic hypotheses and/or sets of branch lengths exist. If no independent phylogenetic information at all is available, it is also possible to generate a large set of possible phylogenies using computer simulation techniques (Losos 1995, Martins in press). Most phylogenetic comparative methods can incorporate most sorts of phylogenetic information, with the accuracy of the results depending primarily on how well the phylogenetic information actually resembles evolutionary "truth." When several possible "truths" exist, we recommend that the comparative analyses be conducted on each of the possible trees and that the variance in results be incorporated explicitly into the estimation and hypothesis testing of any resulting phylogenetic statistics (Martins in press). This provides results that are not specific to any one of the possible phylogenies. Phylogenetic methods (reviewed below) also differ in terms of their robustness to different types of phylogenies and branch length transformation procedures as well as in terms of their robustness to inaccurate phylogenetic information. Thus, accuracy of phylogenetic information and robustness to inaccurate information should be an important consideration in the choice of method.

Choosing a phylogenetic comparative method

Several types of phylogenetic comparative methods have been proposed recently and many of the questions mentioned above can now be answered given the right sort of data. Methods differ in their theoretical perspective, biological assumptions and statistical properties, the types of variables for which they were designed, the types of phylogenetic information they require, and the availability and accessibility of computer programs needed to apply them. Most of the methods have been proposed as ways of estimating the evolutionary relationship between two traits (answering, for example these questions: Has the evolution of one trait constrained evolution of a second? Did two traits evolve in a correlated fashion?), but they can also be used to infer the actual sequence and pattern of evolutionary changes acting in single

traits. A few methods can shed light on the question of evolutionary process (e.g., whether the trait has been evolving neutrally or was subjected to selection). In general, phylogenetic comparative methods involve making some assumptions about the phylogeny and process of phenotypic evolution in the clade and then fitting a statistical model or using a computer algorithm to transform the mean phenotypes of extant species into phylogenetically relevant units. For continuously varying characters, these new phylogenetic statistics can then be used in correlations, regressions, or other analyses instead of the raw species data. For categorical or "state" traits, techniques have been devised to answer specific questions (e.g., Have two traits evolved in a correlated fashion? What is the relationship between gains and losses in a single trait?). Herein, we review seven general classes of alternative techniques that can be used to analyze comparative data phylogenetically.

Before choosing a method, we highly recommend doing some exploratory data analysis. A certain familiarity with the data eases the choice of method, helps to generate hypotheses and guards against making unreasonable assumptions. In the case of comparative data, it is a good idea to get some feeling for the pattern of phylogenetic correlations before applying complex statistical procedures. Several of the methods described below include or can be used as exploratory tools. In particular we recommend *MacClade* (Maddison & Maddison 1992), a computer program, which friendly graphical interface invites you to play with your data. *MacClade* is, as mentioned by Felsenstein, "positively addictive".

A. *Inferred changes methods*

1. *The methods*

a. General description — Inferred changes methods are the most popular type of phylogenetic comparative method in use today (see Brooks & McLennan 1991; Harvey & Pagel 1991; Maddison & Maddison 1992; Miles & Dunham 1993; and Maddison 1994 for reviews). These methods begin by using prescribed algorithms (usually parsimony) to infer the character states of hypothetical ancestors on a known phylogeny and the corresponding magnitude and direction of evolutionary changes occurring along each branch. In many cases, this

inference is the final goal of the study and only qualitative conclusions are made. Alternatively, once the magnitude and direction of evolutionary changes in the trait along the tree are known, these changes can be used instead of the raw species data in explicit statistical procedures (e.g., regressions). This requires some further manipulation, as inferred evolutionary changes on a phylogeny are not statistically independent and violate the assumptions of most statistical procedures.

Note that once the character reconstruction algorithm has been applied, the data points for the study are evolutionary changes rather than species phenotypes. Thus, the effective sample size of any comparative study conducted using these methods can be no larger than the number of evolutionary changes that have occurred, regardless of the number of species measured. If only a single change has occurred in each of two behavior patterns during the entire history of a clade, trying to determine whether the two traits have evolved in a correlated fashion is similar to trying to estimate a correlation after having measured only a single data point. In these situations, any errors in the phylogeny or the data can have extreme effects. For example, Basolo (1990a, b) showed that both female swordtail (*Xiphophorus helleri*) and platy (*Xiphophorus maculatus*) fish prefer males with longer swords (an extension of the lower caudal fin), even though male *X. maculatus* fish do not normally have swords on their tails. Since a phylogenetic reconstruction suggested that swordlessness was the ancestral state for the genus *Xiphophorus*, Basolo (1990b) concluded that female preference for long swordtails evolved before the existence of the male trait and that the "preexisting bias" hypothesis of sexual selection was supported. Unfortunately, Basolo's (1990b) interpretation relies on the location of a single evolutionary change (from nonsworded fish to sworded fish; Fig. 2), and Basolo's analysis is comparable to estimating a correlation (between the existence of male swords and female preference for sword tails) with a single data point. Thus, it is not surprising that when a phylogenetic reconstruction was done on a more recent phylogeny (Meyer et al. 1994), this new phylogenetic information suggested that the common ancestor of all *Xiphophorus* species was probably sworded, thereby overturning Basolo's original conclusion. Future analyses using several sworded and unsworded species from the genus and a more quantitative and powerful method (e.g., Maddison 1990) may provide a more conclusive answer to

whether the “preexisting bias” hypothesis is a reasonable model for sexual selection.

b. Continuous traits — Although there are many ways in which the ancestral states of a continuous character might be estimated from

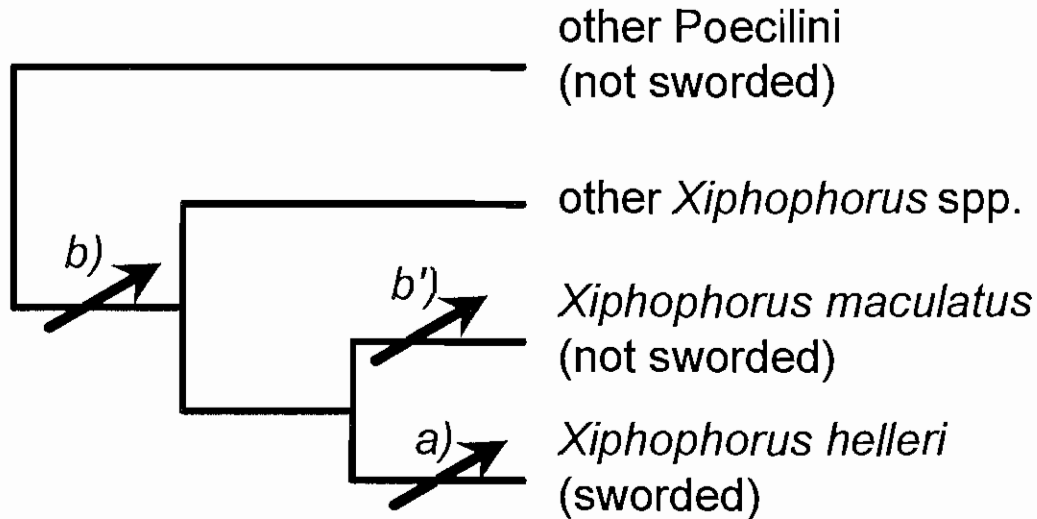


Figure 2. An example from Basolo (1990a, b) and Meyer et al. (1994) of how even minor changes in a phylogenetic hypothesis can have a serious impact on the conclusions of a study when a small number of evolutionary changes (in this case one versus two) are being considered. Using a parsimony reconstruction, Basolo (1990b) suggested that the common ancestor of all *Xiphophorus* species was swordless, with a single evolutionary change at (a) leading to the presence of swords in present-day *X. helleri*. Using empirical manipulations of sword length, Basolo (1990a, b) also showed a female preference for long swords on males in both *Xiphophorus maculatus* and *X. helleri*. Given only this information, it seems likely that the common ancestor of the two species also had a female preference for long male swords. A phylogenetic reconstruction in Basolo (1990b) concluded that there was evidence for the preexisting bias hypothesis of sexual selection because female preference for male swords seems to predate the presence of male swords in this genus. In contrast, using a more recent phylogeny, Meyer et al. (1994) found that swordtails have appeared repeatedly in the evolution of *Xiphophorus*, and that there were probably two evolutionary changes of interest: b) an evolutionary change towards swordedness before the ancestor of all *Xiphophorus*, and b') a loss of swordedness in *X. maculatus*. If this second phylogenetic reconstruction is true, Basolo's (1990b) study does not provide evidence for the preexisting bias hypothesis.

comparative data, two algorithms have been in common use. Both are parsimonylike algorithms designed to minimize the total amount of phenotypic evolution occurring along a phylogeny. "Wagner" or "linear" parsimony is based on the absolute value of the evolutionary change occurring along each branch of the tree and minimizing the sum of those changes across the entire tree (Edwards & Cavalli-Sforza 1967; Farris 1970; Swofford & Maddison 1987). In the "sum of squared-changes parsimony" or "minimum evolution" algorithm, the evolutionary change occurring along each branch of the tree is squared and the overall sum of these squared changes along the entire tree is minimized (Huey & Bennett 1987, Maddison 1991). The Wagner parsimony algorithm is simple to apply but may occasionally give multiple solutions (Swofford & Maddison 1987). The sum of squared-changes method results in a unique solution and can be implemented using iterative procedures (Huey & Bennett 1987), recursive calculation (Maddison 1991) or direct computation (McArdle & Rodrigo 1994).

The above algorithms are applied independently to each of the characters of interest. Once the ancestral phenotypes of those characters have been estimated, the evolutionary change occurring along each branch of the phylogeny can be computed. The resulting evolutionary changes can then be used in standard statistical procedures (e.g., regression, correlation, principal components analysis) to estimate possible relationships among characters or between characters and environments. As there will be $2(N-1)$ evolutionary changes available for each set of N species, it is not immediately obvious what sample size should be used for computing confidence intervals. Huey and Bennett (1987), for example, suggested that although they had inferred $2(N-1)$ evolutionary changes, it was more reasonable to assume that the effective sample size was closer to N . Martins and Garland (1991) showed how phylogenetic randomization tests can be used to estimate the true effective sample size and to conduct hypothesis tests using the sum of squared-changes method.

c. Categorical traits — Ridley's (1983) book was the first modern account of comparative methods from an inferred changes perspective. He suggested that character transitions reconstructed by parsimony methods could be regarded as independent evolutionary changes and proposed that these changes be used as independent trials in testing for adaptation. For example, the hypothesis that a specific character state

(e.g., male precopulatory behavior) is an adaptation to a specific "environment" (e.g., predictable receptivity in females) could be tested by comparing the number of times the character state had been attained in this as opposed to alternative environments. There are many algorithms available to reconstruct the ancestral states of categorical or "state" characters evolving under parsimony. As these have been reviewed frequently and recently (e.g., Maddison & Maddison 1992; Maddison 1994), we will not do so herein. Instead we concentrate on methods that address what to do once the ancestral states are obtained.

Sillén-Tullberg (1988) used Ridley's approach to study the evolution of warning coloration (aposematism) of butterfly larvae. She found that out of 23 evolutionary changes toward gregariousness, 15 to 18 occurred in areas of the phylogeny in which the taxa were aposematic. Without using statistics, she rejected the popular hypothesis that aposematism evolved by kin-selection among gregarious and unpalatable caterpillars, concluding instead that "unpalatability is an important predisposing factor for the evolution of egg clustering and larval gregariousness." Maddison (1990) criticized this approach because it did not take the distribution of aposematism on the phylogeny into account. Aposematism was much more common than crypsis in Sillén-Tullberg's (1988) data set. Thus, we expect most evolutionary changes towards gregariousness to be associated with aposematism even without any causal or functional relationship between the two characters. Maddison (1990) developed his concentrated-changes test to correct this sort of problem. Under the null hypothesis that gregariousness is uniformly distributed on the phylogeny, Maddison's method calculates the expected distribution of changes in one first variable (e.g., gregariousness) given the state of a second causal variable (e.g., coloration) and given the actual number of changes (gains and losses of gregariousness) that occurred on the phylogeny as a whole. This distribution can then be used to evaluate whether the observed number of gains in gregariousness is associated with aposematism. As an illustration, Maddison reanalyzed Sillén-Tullberg's data and found that 15 to 18 evolutionary gains in gregariousness in aposematic taxa, given 23 gains and six losses on the phylogeny as a whole, are well within the range expected from the null hypothesis of no association between the traits (Fig. 3).

Use of Maddison's method assumes that the data are either a random or a complete sample of the clade under study. As discussed in Maddison (1990), any change in that phylogeny produced by adding or subtracting taxa will affect the outcome of his method. Sillén-Tullberg (1993) responded to Maddison's reanalysis of her data by pointing out that the taxa included in her original study (Sillén-Tullberg 1988) were not a random sample of butterfly taxa. Aposematic taxa were, in fact, strongly overrepresented. Thus, the results of Maddison's reanalysis were due more to the dependence of his method on which taxa were sampled than to butterfly biology. Sillén-Tullberg (1993) proposed a variant of Maddison's method, the contingent states test, which uses a contingency table to test whether gains of gregariousness are dependent on aposematism. In this method, Sillén-Tullberg considers only those branches along which gains of gregariousness are possible (i.e. branches in which the solitary state is ancestral). Thus, the results of the contingent states test depend somewhat less on which species are sampled in the clade. Using this new method and a more representative data set, Sillén-Tullberg again found a strong relationship between aposematism and gregariousness.

If the phylogeny is known and all relevant species in the clade have been measured, Maddison's approach gives a reasonable test of relationship between two characters in that clade. If the study is being used to make inferences about a large clade from a smaller sample of measured species, then the way in which those species were sampled becomes important, and Sillén-Tullberg's approach may be preferred. Both Sillén-Tullberg's and Maddison's methods are directional or causal in their approach as they only consider evolutionary changes that can be unambiguously associated with the state of the second causal, variable (i.e., they disregard branches along which the causal variable has changed). These methods can also be used to ask more general evolutionary questions such as whether evolutionary changes in a trait are concentrated in defined parts of the phylogeny.

In his general discussion of the inferred changes approach, Ridley (1983) proposed using contingency tables to test whether two characters have evolved independently of each other or whether a single trait has evolved independently of the environment. This method uses only those branches along which at least one of the characters has changed. Grafen and Ridley (in press, and, in this volume, Ridley & Grafen, Chap. 3)

formalized this into the “independent character evolutions” or ICE test and developed a second method, the ICDE (independent character-differences evolution method), which also tests a null hypothesis of independent evolution but avoids some of the ancestral reconstructions and thereby some of the problems of the ICE.

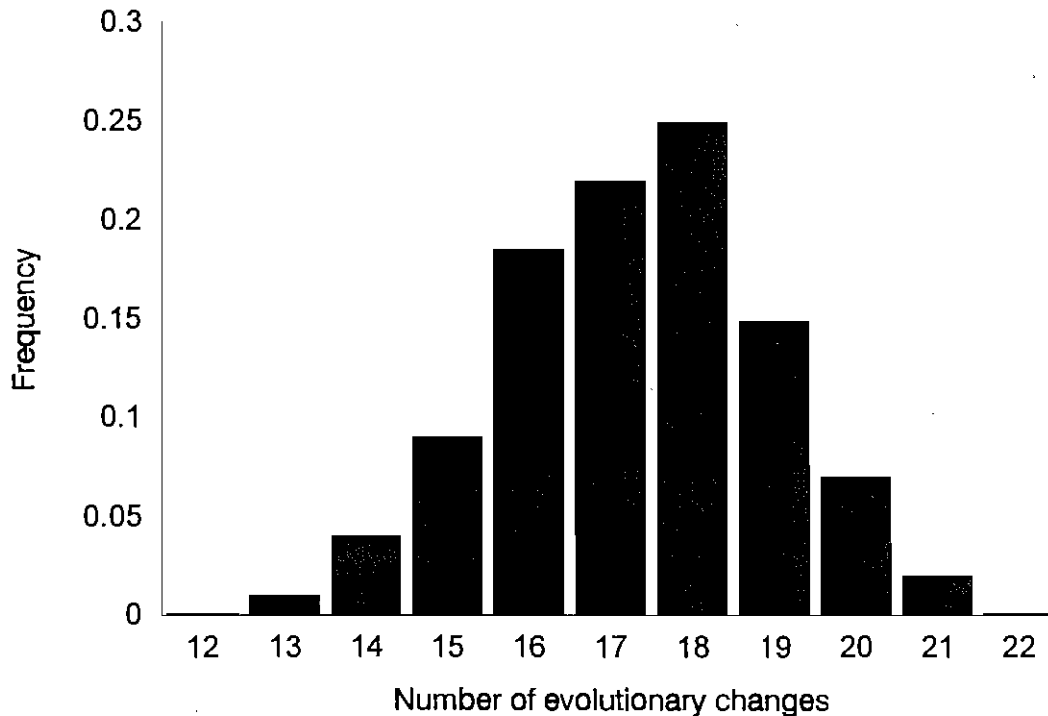


Figure 3. A frequency distribution from Maddison (1990) illustrating the use of his method to analyze data from Sillén-Tullberg (1988). Computer simulation was used to determine the number of times gregariousness in butterfly larvae is expected to evolve in taxa that also exhibit warning coloration, given 1) that there have been 23 gains and 6 losses in gregariousness during the evolution of the entire clade of 136 taxa (determined by using parsimony to reconstruct the ancestral states of gregariousness), and 2) that gains and losses are distributed randomly along the phylogeny. The histogram describes the number of computer simulation trials in which a particular number of evolutionary gains in gregariousness were observed along branches of the phylogeny in which warning coloration has evolved. Sillén-Tullberg (1988) observed 15-18 evolutionary gains in gregariousness. Results in this range occur in more than 90% of the trials and are thus very compatible with a null hypothesis of no association between gregariousness and warning coloration (but see Sillén-Tullberg 1993).

2. Assumptions

As is true for most phylogenetic methods, all of the above methods assume that within-species variation does not exist or is negligible in comparison to the level of among-species variability and that the phylogeny is known. The use of parsimony to reconstruct the ancestral states implies a few additional assumptions. Notably, parsimony implies that evolutionary changes in the character(s) are relatively rare and uniformly distributed over branches. Exactly how rare and how uniformly distributed is determined by the details of the algorithm (e.g., the sum of squared changes algorithm assumes that the distribution of evolutionary changes which gives the smallest sum of squared changes is the most likely to be true). Whether or not the parsimony assumption is reasonable is an empirical question. In particular, note that although this assumption may be reasonable in phylogeny reconstruction when the traits are chosen for their slow rate of evolution (but see Felsenstein & Sober 1986), it may often be unreasonable in comparative studies when the traits are usually chosen because they are thought to have been responding to selection. The ancestor reconstruction methods for continuous traits described above can be interpreted in terms other than parsimony. Maddison (1990) showed that the estimates of ancestral states obtained by squared-changes parsimony (with weighted branch lengths) are the same as would be obtained by a likelihood approach assuming that the traits evolved as if by Brownian motion (a common model of evolution discussed under independent contrasts below).

The inferred changes methods also assume that the ancestral states are known and not, as is almost always the case, estimated. Although evolutionary changes on different branches of the phylogeny may be independent, changes estimated using a parsimony algorithm will probably not be statistically independent. Thus, the above methods do not completely solve the problem of statistical dependence. Grafen and Ridley's (in press) ICDE methods are less vulnerable to this problem as most ancestor reconstruction in this method is either avoided or replaced with randomization. Using computer simulation, Maddison (1990) argued that his method is relatively robust to inaccurate parsimony reconstruction of ancestral states and is unlikely to give consistently misleading results if conservative significance levels are used.

Many of the above methods also assume that all branches are of equal length (i.e., that the same amount of phenotypic evolution is expected to occur along each branch of the phylogeny). This is sometimes associated with a model of punctuated evolution in which all changes occur during speciation. A model of punctuated evolution, however, is only reasonable in such analyses if all speciation events (including those leading to extinct taxa) are known and included in the phylogeny. Thus, when alternative branch lengths are available, they can and should be incorporated into the algorithm for estimating evolutionary changes. For example, the iterative sum of squared changes algorithm (e.g., Huey & Bennett 1987) proceeds by estimating the phenotype of each ancestor as the average of its three closest relatives on the phylogeny. Variable branch lengths can be incorporated by estimating ancestral phenotypes as weighted rather than simple averages, using the branch lengths as the weights (Maddison 1991; Martins & Garland 1991b). Similarly, Sanderson (1991) proposed a variant of Maddison's (1990) method which allows the probability of character change to vary on different branches.

3. Variations for dealing with unknown phylogenetic information

None.

4. Strengths and weaknesses

Although relatively few computer simulation studies have included tests of the inferred changes methods, the results have been quite encouraging. Again, for categorical characters, Maddison (1990) found that his method is relatively robust to inaccurate parsimony reconstruction of ancestral states. For continuous characters, the sum of squared changes method with appropriately weighted branch lengths (Maddison 1991, Martins & Garland 1991b) has been found to result in consistently good estimates of evolutionary relationships that are sometimes superior to those produced by other methods for continuous characters (Martins & Garland 1991b; Martins in review).

For categorical variables, the greatest weakness of the inferred changes methods is the lack of development of parameter estimation as opposed to testing of null hypotheses. With the exception of Sanderson's (1991) method, these methods are further restricted in the assumption

that all branches on the tree are of equal length. For continuous variables, the greatest weakness is the lack of explicit, analytical tools for computing standard errors and conducting hypothesis tests. The randomization procedure proposed by Martins and Garland (1991) is adequate but time-consuming and requires making a number of explicit assumptions about the microevolutionary processes guiding change in the character of interest. Finally, for both continuous and categorical variables, no extensions of the inferred changes methods have been developed to address the problem of inaccurate or unknown phylogenetic information.

5. Computational difficulties and computer programs

The popularity of inferred changes methods is almost certainly due to their relative computational simplicity and intuitive appeal. There are several user-friendly computer programs available to infer the ancestral states of characters along a known phylogeny under most of the available parsimony algorithms (e.g., Maddison & Maddison 1992; Felsenstein 1993; Swofford 1993; and Martins & Garland 1991a). We particularly recommend Maddison & Maddison's (1992) *MacClade* as an exceptional heuristic tool for any biologist interested in working with categorical comparative data. Once the ancestral states are known and evolutionary changes inferred, changes in continuous characters can be analyzed by hand, or using any standard statistical package. For categorical characters, Ridley's method can be applied by hand or using any statistical package that perform chi-squared tests; and probabilities from Maddison's (1990) method are calculated by the "concentrated changes" test of *MacClade* (Maddison & Maddison 1992). Sanderson (1991) and Sillén-Tullberg (1993) offer computer programs to calculate probabilities using their variants of Maddison's method.

B. Independent Contrasts and related methods

1. The methods

a. General description — There are several methods that can be grouped under the broad category of "independent contrasts," and which originated with the method proposed by Felsenstein (1985). The

independent contrasts method was originally devised as a way of addressing the problem of statistical dependence of comparative data within a population genetics framework. As described by Felsenstein (1985), this method assumes that the evolution of the measured trait can be described using a Brownian motion model of phenotypic evolution. Brownian motion is a stochastic process in which the evolutionary change in phenotype occurring during any interval of time is normally distributed, has variance proportional to the time interval, and is independent of the state of the phenotype at the beginning of the interval. It may be used to describe the evolution of a continuous character undergoing random genetic drift, a character responding to selection when the direction of selection is constantly shifting back and forth at random, or a character under stabilizing selection around an optimum that itself evolves according to Brownian motion (Felsenstein 1988, Hansen & Martins in press). Under such a model, although the species data are not independent, the differences or "contrasts" between certain pairs of species on the phylogeny are independent of one another (Fig. 4). If the phylogenetic relationships among species and the amount of expected change along each branch of the phylogeny are known, a simple algorithm can be used to transform the measured species data into a set of $N - 1$ standardized and independent "contrasts" (where N is the number of species that were measured).

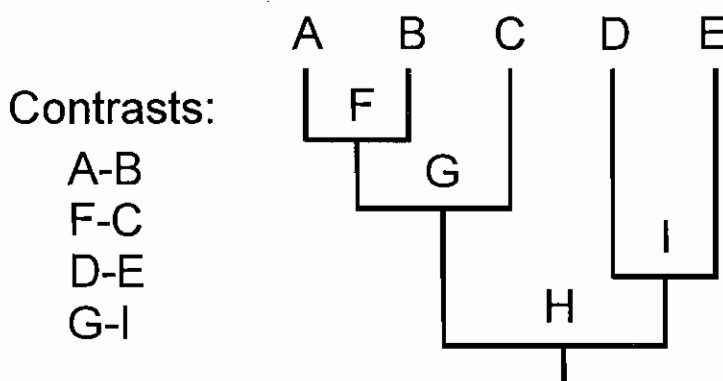


Figure 4. Illustration of Felsenstein's (1985) method of independent contrasts. Under a Brownian motion model of phenotypic evolution, although the measured species (A-E) are not statistically independent of one another, certain "contrasts" or differences between pairs of taxa are. Felsenstein's (1985) method consists of a numerical algorithm for calculating $N-1$ of these contrasts (where N is the number of measured species) and then standardizing them appropriately, such that they can be used in any standard statistics.

Under the Brownian motion assumption, contrasts are normally distributed. If the phylogeny is known, contrasts are also independent of one another. Standardization ensures that the final contrasts have the same variance (i.e., be homoscedastic) and a mean of zero. Thus, Felsenstein's contrasts are independent, normally distributed, and homoscedastic variables and fit the assumptions of most statistical tests. Once independent contrasts are formed and standardized, they can be used instead of the raw species data in any standard statistical procedure (e.g., regressions and ANOVA) as a way of conducting the analysis while taking phylogenetic information into account. Note that because the expectation of the contrasts is zero, any statistical model involving these contrasts should not include a grand mean. For example, in a regression of contrasts on other variables, the regression should be forced through the origin without estimating a nonzero y intercept (see Garland et al. 1992 for further explanation).

A number of "pairwise" contrast methods have arisen from Felsenstein's (1985) suggestion that we consider data collected from "two seals, two whales, two bats, two deer, etc." (e.g., Burt, 1989, Møller & Birkhead 1992, and in this volume, Nee et al., Chap. 13). If the pairs are formed using only closely related pairs of extant species (e.g., two bats, two whales) they are independent under less stringent assumptions than a Brownian motion (see below). On the other hand, as they use less of the information contained in the data, they lead to less precise and less powerful estimates of evolutionary parameters.

A method related to independent contrasts is the generalized least squares technique proposed by Grafen (1989, 1992). Grafen describes a set of comparative data as the sum of interspecific values of many other measured and unmeasured traits in a multiple regression format. He then uses generalized least squares and a "contrasts" approach to incorporate phylogenetic information into a multiple regression model describing the importance of each trait in predicting variation in the single response trait (e.g., using regression slopes, correlation coefficients, coefficients of determination). Grafen's method is based on statistical rather than population genetic assumptions, and requires that the error term in the multiple regression model consist of normally distributed variables with known variances and covariances. Grafen's "standard regression" calculates these variances and covariances of the error term using a known phylogeny and an assumption of character evolution (e.g.,

branch lengths in units of expected character change as in Felsenstein's method).

Felsenstein's (1985) contrasts method can be viewed as a model based on restricted maximum likelihood (REML), whereas Grafen (1989) relies on generalized least squares (GLS). When the phylogeny and branch lengths are known, Felsenstein's and Grafen's methods yield identical standardized contrasts and measures of the relationships among traits. The main practical differences between the two methods occur (a) when the phylogeny and branch lengths are uncertain (see below), (b) when categorical variables are included in the analysis (see below), and (c) in the assumptions and interpretations of the results. The results of statistics conducted on Felsenstein contrasts can be interpreted directly in historical terms. For example, a correlation between Felsenstein contrasts in two behavioral characters is an estimate of the correlation between the evolutionary changes occurring in the two traits at each generation. A regression slope describing the regression of one set of Felsenstein contrasts on another is an estimate of the amount of variation in one behavior pattern explained by the second at each generation through the evolutionary history of the clade. Grafen (1989, 1992) argues that most phylogenetic correlation is lost through evolutionary time due to the action of selection. His method partitions variation in one trait (the response or Y variable) into several components. Some of these components are other measured traits (predictor or X variables). The regression model estimates the relationship between the response and predictor variables and interprets this as due to recent adaptation (i.e., occurring at the tips of the phylogeny). The last component is a residual error term that describes the phylogenetic correlation due to evolutionary change in the response variable along the rest of the phylogeny.

Note that although one step of Felsenstein's algorithm is described as estimating "ancestral" states, the method was not intended as a means of estimating the ancestral states or magnitude of evolutionary changes occurring along the phylogeny. The "ancestral states" calculated as an intermediate step of the contrasts algorithm are the weighted averages of the two descendant species resulting from each node. This is a *local* estimate of the ancestral states (i.e., a reasonable estimate if the entire tree consists only of the ancestor and its descendants). A better estimate

of the ancestral states would be a weighted average of all the extant species phenotypes (Maddison 1991).

b. Continuous vs. categorical traits — Felsenstein's version of independent contrasts is designed for use only with continuously varying characters (e.g., body size, display rate). Grafen's (1989, 1992) version allows for the use of categorical predictor ("independent" or "X") variables described as sets of dichotomous "dummy" variables (0/1; see Draper & Smith 1981 for an excellent description of how this is done). When these are incorporated into a multiple regression, the relationship between continuous and categorical traits can be estimated. This difference between the two methods is due to the fact that Felsenstein derived his method from an assumption that the traits evolve as if by Brownian motion, whereas Grafen assumes that the error terms in his multiple regression models are normally distributed with means of zero and known variance and covariance. Although categorical variables cannot evolve by Brownian motion nor be normally distributed, the residual error terms can.

2. Assumptions

Felsenstein's (1985) original method assumes that: (a) within-species variation does not exist or is negligible, (b) phylogenetic relationships are known and can be described as a standard binary tree structure (i.e., the phylogenetic topology is known), (c) the process of phenotypic evolution for the characters being considered can be described as a Brownian motion process, and (d) the relative rate of this process along each branch of the tree (i.e., branch lengths in units of expected amount of character change) is also known (or that branch lengths in units of time are available and that the rate of phenotypic evolution is constant throughout the clade).

Grafen's (1989) regressions assume that (a) within-species variation does not exist or is negligible, (b) that the predictor or X variables are independent of each other and the error term, and (c) that the variance-covariance of the error term has been taken into account by the GLS procedure. When the variance-covariance of the error term is compatible with what would be obtained from a Brownian motion, the last two assumptions of Grafen's method are comparable to the last two assumptions of Felsenstein's method, but Grafen's method may also

apply to other situations. Grafen (and others, see below) also relax the assumptions that the phylogeny and branch lengths are known by using algorithms and estimation procedures to infer some of this information from the comparative data themselves.

The pairwise comparison methods relax the Brownian motion assumption of Felsenstein's (1985) method but still assume that each pair of species has been diverging for the same length of time and at the same rate as every other pair (such that each contrast has the same variance) and that the differences are normally distributed. If nonparametric statistics are used, these assumptions are also relaxed, but interpretations are much less clear.

3. Variations for dealing with unknown phylogenetic information

Several variants of the independent contrasts method have been proposed to address the problems of obtaining accurate phylogenies and branch lengths for the analysis. In the original description of his method, for example, Grafen (1989) proposed "the phylogenetic regression," which includes a means of reducing unresolved polytomies into binary topologies. Beginning with an initial "working" phylogeny and Ridley's (1983) "radiation principle," Grafen suggests that a single data point be "extracted" from each unresolved polytomy on the tree by setting the nodal value equal to the weighted average of the trait values of its descendants. Grafen (1989) then incorporates the error generated from this procedure into the variance-covariance structure of the generalized least squares model. Pagel (1992) presents a similar approach, which differs mostly in suggestions that some phenetic or other information be used to reduce multiple nodes initially before applying his "expansion" procedure. The "pairwise" methods can be seen as an extreme form of these topology reductions designed for situations in which the available phylogenetic information only allows the identification of closely related pairs of extant species (e.g., two bats, two whales). All of these topology reduction methods are conservative and may be useful when polytomies are due to uncertainties in the available phylogenetic information rather than to a belief that several species really did radiate simultaneously from a single point. In the latter case, Felsenstein's (1985) suggestion of resolving the polytomy into a set of bifurcations by inserting branches

of zero length is more appropriate (which species are paired with which does not matter).

Other methods consider ways of transforming branch lengths on a rough "working" phylogeny into units of expected amount of phenotypic change for particular characters in the study. For example, Grafen's (1989) phylogenetic regression also includes a statistical estimator, ρ , for inferring the relative amount of phenotypic evolution expected along each branch of the topology (i.e., branch lengths in units of expected variance of character change). He suggests that each branch length (h) from the working phylogeny (scaled such that the tree has a total length of one) be raised to a power ρ , and then that $1 - h^\rho$ be considered as a measure of the expected amount of similarity due to phylogenetic relationships (i.e., the variance of the trait value or contrast). Grafen (1989) then shows how ρ can be estimated along with the other parameters of his regression model using maximum likelihood techniques. Alternatively, using the framework of Felsenstein's method, Martins (1994) proposed an iterative least squares procedure in which branch lengths in units of time can be transformed into units of expected variance of character change using a Brownian motion model underlying each character rather than the full regression model used by Grafen. Martins' method can also transform branch lengths based on an Ornstein-Uhlenbeck model of phenotypic evolution under random genetic drift with stabilizing selection. Thus, although Martins' (1994) procedure was described originally as a means of estimating the rate of phenotypic evolution of a single character from comparative data, it is also useful as a way of testing whether the Brownian motion model underlying Felsenstein's (1985) method adequately describes the interspecific variation observed. Finally, from a more subjective approach, Garland et al. (1991) suggest that the absolute value of each standardized independent contrast be plotted versus its standard deviation (the square root of the branch lengths in units of expected variance of change) and that a qualitative, visual analysis be used to determine whether there are any general patterns which would suggest that the branch lengths are in the wrong units. They then suggest that statistical transformations of the data or branch lengths be used to correct this problem. One potential problem with this approach is that any statistical transformations of the data or branch lengths are, in effect, changing the microevolutionary assumptions underlying the

comparative analysis. Purely statistical solutions to the problem of unknown or incomplete phylogenetic information can lead to difficulties in interpreting results.

4. Strengths and weaknesses

If the assumptions and the phylogenetic information provided are correct, the independent contrasts method is guaranteed analytically to correct the problem of statistical dependence (Felsenstein 1985, Grafen 1989). Since Felsenstein's (1985) method may be derived from population genetic principles, it also has the advantage of yielding results that can be interpreted from an evolutionary perspective. Simple extensions of the method can be used to estimate related parameters such as the rate of phenotypic evolution (e.g., Garland 1992; Martins 1994). Both Felsenstein's (1985) and Grafen's (1989) methods have also been shown to provide quite good estimates of evolutionary relationships between traits, often much better estimates than provided by other methods, including inferred changes or spatial autocorrelation techniques (Martins and Garland 1991b; Grafen & Ridley in press; Martins in review). Even when phylogenetic information is incorrect, Felsenstein's contrasts method usually matches or exceeds the statistical performance of other comparative methods proposed for continuous traits.

Most of the criticisms and discussion of independent contrasts have centered on the obvious difficulties of obtaining accurate and complete information about both the phylogenetic relationships (the tree structure) and the process of phenotypic evolution (branch lengths in units of expected character change). The topology reduction algorithms and branch length transformation procedures proposed to address the problem of unavailable phylogenetic information may not always perform as well as the original method and should be used with caution (Gittleman & Luh 1992, 1993; Purvis et al. 1994; Martins in review). These methods are statistically conservative in the sense that the resulting confidence intervals are too wide, and the power of statistical tests to detect real evolutionary patterns is low. When these methods are used, any variation in the comparative data is explained first by reference to the phylogeny and will only later be considered as the possible result of correlated evolution, convergence, or other responses

to selection. Thus, relationships between traits can be obscured, and the absolute values of estimates may be downwardly biased. Again, if the Brownian motion model is appropriate and the phylogenetic information is known, most of the above variations of the basic independent contrast method yield identical results, and Felsenstein's (1985) original method or Grafen's (1989) "standard regression" should be applied. Use of the pairwise methods, topology reduction algorithms, and branch length estimation procedures should yield similar results but may be less accurate or even biased as the methods use less of the available information. If Brownian motion is not appropriate or if the estimated phylogenetic information is inaccurate, none of these variants of the independent contrasts approach is guaranteed to correct the problem of statistical dependence.

5. Computational difficulties and computer programs

For small phylogenies, Felsenstein's method can be computed by hand. For larger phylogenies, his method has been implemented in *PHYLIP* which is menu-driven, user-friendly, and available for DOS/Windows, Macintosh and UNIX platforms (Felsenstein 1993). Grafen's method can be implemented using a program that is available from the author. This program is in *GLIM*, and requires substantial knowledge of that programming language for implementation. Purvis and Rambaut's (in press) *CAIC* offers a user-friendly Macintosh program to calculate independent contrasts using Pagel's (1992) topology reduction procedure. Martins and Garland's (1991a) *CMAPI* implements several variants of Felsenstein's independent contrast method for DOS machines. Martins' (1995) *COMPARE* is available for DOS/Windows, Macintosh, and UNIX platforms; it calculates Felsenstein contrasts, estimate the rate of phenotypic evolution using Martins' (1994) method, and implement several other comparative methods.

C. Explicit model based methods for categorical traits

1. The methods

There are a number of methods which, like Felsenstein's (1985) independent contrasts, are based on explicit models of phenotypic

evolution (e.g., Harvey & Pagel 1991, p101; Janson 1992; Sanderson 1993; and Pagel 1994). These methods differ from the contrasts methods, however, in that they were explicitly designed for use with dichotomous or categorical (i.e., yes/no or "state") variables and therefore cannot be used to form "contrasts." The methods begin by building a probability model of the possible transitions between character states. Most then estimate these transition probabilities between states using maximum likelihood techniques.

These probability model based methods were designed to address a rather wide variety of evolutionary questions. Janson's (1992) method asks whether a single trait has evolved neutrally along a phylogeny or whether it was subjected to evolutionary constraints of various defined sorts. Sanderson's (1993) method considers whether there have been significantly more gains than losses in a trait along a tree (Fig. 5). Pagel's method (Pagel 1994; Harvey & Pagel 1991, p. 101) is concerned with determining whether two traits have evolved in a correlated fashion, by estimating the probability that evolutionary changes in one trait have been associated with evolutionary changes in a second. Similar probability models have been used by authors considering the use of DNA sequences to reconstruct phylogenetic trees, and questions in molecular evolution (e.g., Barry & Hartigan 1987; Ritland & Clegg 1987; Hendy 1989; Hendy & Penny 1989; Reeves 1992). Although the models proposed in these latter studies could be applied to the question of phenotypic evolution, as of yet they are primarily concerned with tree reconstruction and are beyond the scope of this chapter.

2. Assumptions

All of the methods described in this section are based on specified probability models. The methods usually require that each character exhibit only a few (usually two) states, that the probability of change between those states be constant throughout evolution in the clade, that within-species variability be negligible, and that the phylogeny be known without error. Each method also makes a number of other simplifying assumptions that differ from method to method. All except Pagel (1994) assume that the ancestral character states are known. In practice, when applying Janson's (1992) and Sanderson's (1993) methods, for example, the ancestral states are estimated from the extant

species data using some sort of parsimony algorithm, and then used as if they were known without error. In Pagel's (1994) method, the likelihood is summed over all possible assignments of ancestral states.

3. Variations for dealing with unknown phylogenetic information

None.

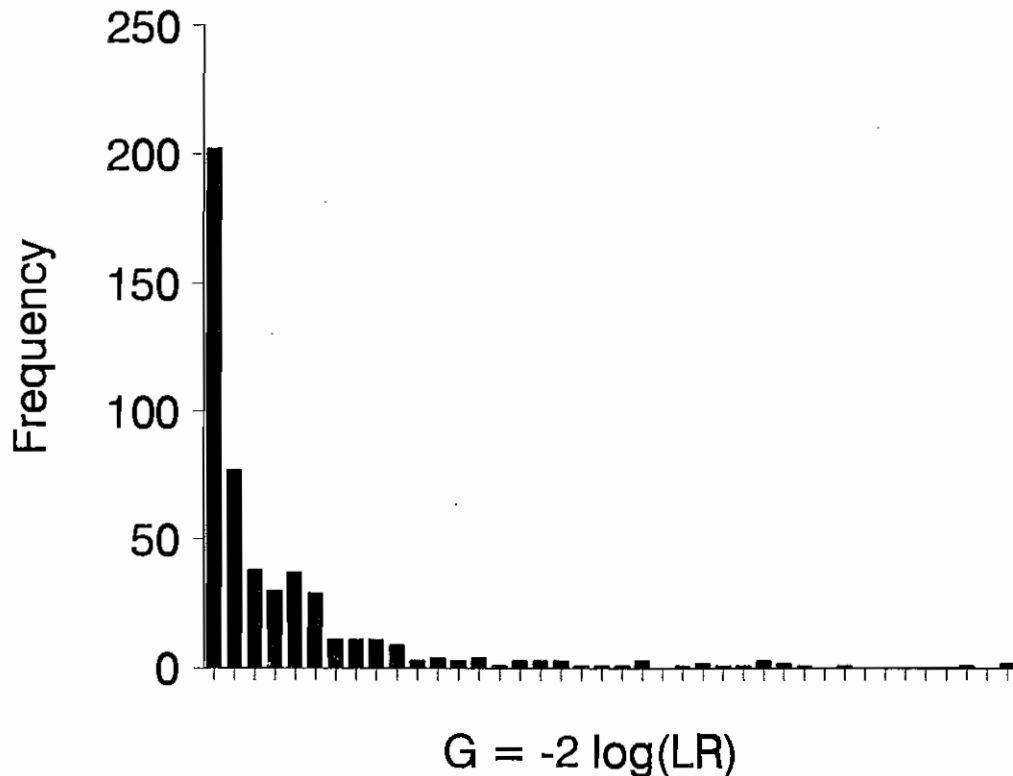


Figure 5. A frequency distribution modified from Sanderson (1993), illustrating the use of his method. Beginning with reconstructions of the ancestral states of a phenotypic character, Sanderson's (1993) method can be used (among other things) to test whether the rate at which a dichotomous (i.e., yes/no) trait is gained equals the rate at which the same character is lost along a known phylogeny. Use of the method yields a quantity, $G = -2 \log(LR)$, where LR is a likelihood ratio calculated from the measured data. This figure depicts the frequency distribution of G obtained by computer simulation for a phylogeny of 128 taxa. Such distributions can be used to conduct simple hypothesis tests about biases in evolutionary gains and losses. For example, an empirical measure of G calculated for a specific trait evolving along this phylogeny of 128 taxa is said to be significantly different from that expected under the null hypothesis of equal rates of gains and losses if it falls in the rightmost 5% of the distribution depicted in this figure.

4. Strengths and weaknesses

The main advantage of the methods discussed in this section lies in their explicit reliance on probabilistic models. As with Felsenstein's (1985) contrasts method, these probabilistic models make it much easier to relate both the assumptions of the methods and the interpretations of resulting parameter estimates to the underlying microevolutionary processes (Hansen & Martins in press). Although such comparisons have not been pursued with most of the methods reviewed in this section, the possibility of doing so remains one of the strengths of the existing methods.

As with many of the inferred changes methods, the primary weakness of most of the probability model methods is the requirement that ancestral character states be known without error. As discussed above, this can lead to inflated sample sizes and estimated parameters may reflect more about the method of reconstructing the ancestral states than about the evolutionary process being inferred. Pagel (1994) solves this problem by utilizing as likelihood the marginal distribution of the extant species which does not depend on the ancestral character states. This leaves only the state of the root to be specified

5. Computational difficulties and computer programs

Pagel's method is computationally intensive, requiring several onerous computations (especially with more than a few possible character states). Janson's and Sanderson's methods are less computer-intensive. Programs to implement these procedures are available from each author on request.

D. Spatial autocorrelation

1. The method

Cheverud et al. (1985, Cheverud & Dow 1985a) suggested the use of spatial or network autocorrelational techniques (as adapted from Cliff & Ord 1981) to incorporate phylogenetic information into comparative analyses. Their method uses a spatial autoregressive model to partition the variation in each measured species phenotype into a factor

describing the predicted phenotype of the species given the phylogeny and the measured phenotypes of all the other species in the clade (i.e., the “phylogenetic component”) and a factor unique to that species (the “residual component”). In mathematical terms, the model is $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}$ where \mathbf{y} is a vector of the observed species phenotypes, $\rho\mathbf{W}\mathbf{y}$ is the phylogenetic component, and $\boldsymbol{\varepsilon}$ is the residual component. In the phylogenetic component, \mathbf{W} is a $N \times N$ “phylogenetic connectivity matrix” where N is the number of species. The off-diagonal elements of \mathbf{W} describe the expected similarity among species due to phylogenetic relationships, whereas the diagonal elements are set to zero such that for each species the phylogenetic component is the phenotype for that species predicted by the phenotypes of all the other species not including itself. In practice, the degree of expected phenotypic similarity (i.e., each element of \mathbf{W}) is a combination of the known phylogeny and a particular hypothesis regarding how much the character is expected to change as it evolves along the branches of the phylogeny. It is loosely comparable to the “branch lengths in units of expected variance of change” required for Felsenstein’s (1985) contrasts method.

The autoregressive model is fit by estimating the autocorrelation coefficient, ρ , (which is completely unrelated to Grafen’s ρ above) using maximum likelihood techniques. This coefficient describes the amount of variation in the character explained by the particular phylogenetic hypothesis (represented by \mathbf{W}). Positive values of ρ imply that closely related species are phenotypically similar to one another in the way specified by the \mathbf{W} matrix, whereas negative values suggest that character displacement or other forces have led to closely related species being more phenotypically different than distantly related species.

As pointed out by Cheverud et al. (1985), a measure of the overall fit of the model [e.g., $r^2 = 1 - \Sigma(\boldsymbol{\varepsilon}_i^2) / \Sigma(\mathbf{y}_i^2)$, where the summations are over all species i] can be used as an estimate of the amount of variation in the character explained by the phylogenetic hypothesis (i.e., a measure of the phylogenetic correlation discussed above). A low value of r^2 implies either that the species phenotypes are not well predicted by the phenotypes of the rest of the clade or that the phylogenetic hypothesis and/or autoregressive model is incorrect. If the model describes the data reasonably well, the residual components (obtained by subtraction; $\boldsymbol{\varepsilon} = \mathbf{y} - \rho\mathbf{W}\mathbf{y}$) will be statistically independent variables on which further

statistical analyses (e.g., regression, ANOVA, multivariate statistics) can be performed. Thus, fitting this model to measured comparative data is equivalent to conducting a linear transformation or filtering of those data that takes phylogenetic information into account.

In an extension of the basic method, Gittleman and Kot (1990) developed the use of Moran's I and spatial autocorrelograms as diagnostic tools to determine whether or not the model fits comparative data well (see also Gittleman & Luh 1992, 1993, Purvis et al. 1994, and in this volume, Gittleman et al., Chap. 6). An autocorrelogram is a histogram plot of Moran's I by historical distance. Moran's I is a relative measure of the expected similarity among species phenotypes due to a specific phylogenetic connectivity matrix. High values of Moran's I indicate that the proposed phylogenetic or taxonomic hypothesis does indeed explain some of the variation present in the data. If I is small or negative, the spatial autoregressive model may not provide a reasonable fit to the comparative data, and/or incorporation of historical information may be unnecessary. An autocorrelogram can be obtained by slicing a taxonomy or phylogeny into several vertical segments and calculating Moran's I to describe the extent of phylogenetic correlation within each segment. For example, if only taxonomic information is available, we might calculate Moran's I at the species, genus, family and order levels (e.g., Fig. 6). When phylogenetic information is available, we might slice the phylogeny into vertical segments to describe the phylogenetic correlation in data from species which have a common ancestor 0–10 million years ago, 10–20 my ago, and so on. In the latter case, it is not clear what the best interval of time (or genetic distance, or number of evolutionary changes, or other unit of branch length) is most useful for calculation of Moran's I (see Purvis et al. 1994 for a suggestion). Once the autocorrelogram has been obtained, it can be used to determine whether an autoregressive model is appropriate for a particular set of data and relationship matrix (\mathbf{W}). If an autoregressive model is appropriate, we expect Moran's I to be positive at several levels (indicating that phylogenetic correlation can be detected) and to decrease with increasing taxonomic or phylogenetic distance. If particular values of Moran's I are small or negative, we might consider altering the relationship matrix (e.g., by using a cut-off value) to reduce the effects of those parts of the phylogeny which do not fit the

autoregressive model. Thus, visual inspection of a correlogram can serve as a diagnostic tool in applying the spatial autoregressive model.

The spatial autoregressive model proposed by Cheverud et al. (1985) is based on linear regression techniques and is designed primarily for the analysis of continuous characters. As illustrated by the authors, dichotomous (yes/no) characters can also be analyzed using this model, and by extension, categorical or "state" characters can be

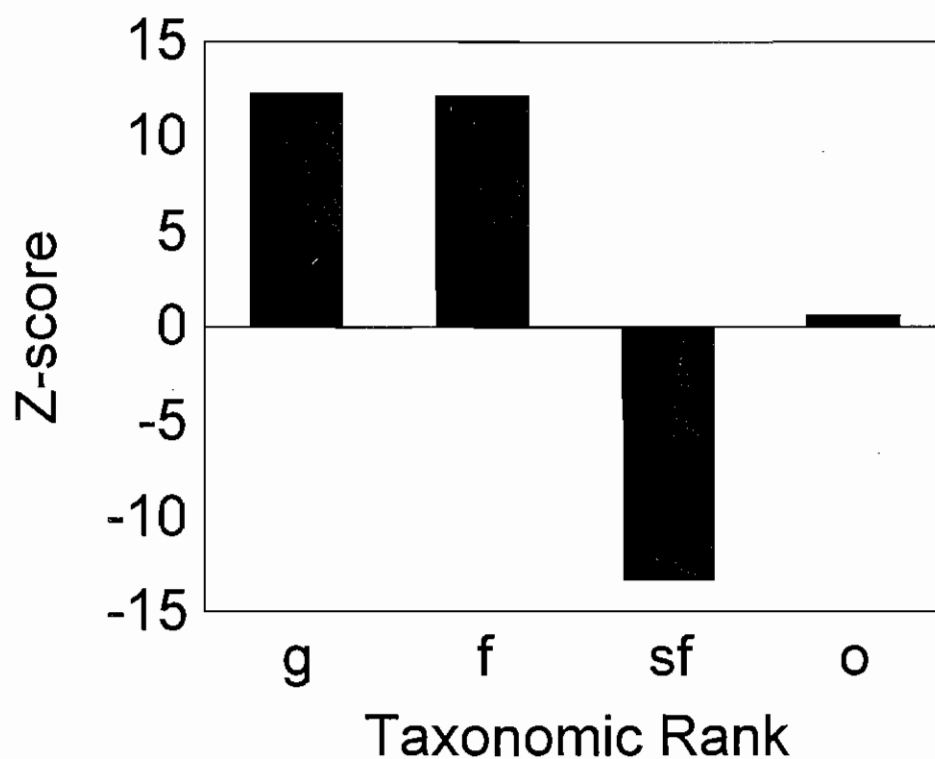


Figure 6. Phylogenetic correlogram modified from Gittleman & Kot's (1990) analysis of carnivore body weight (using Nowak & Paradiso's 1983 classification of 123 species). By varying the phylogenetic relationship matrix, normalized values of Moran's I and their respective z -scores were calculated at the genus, family, superfamily and order levels. These values describe the amount of phylogenetic correlation (i.e., a relationship between character variation and the phylogeny) present at each taxonomic rank, and can be used to suggest the best taxonomic level for further analysis, and to determine whether phylogenetic information need be taken into account for such analyses. Z -scores with absolute values greater than 1.96 are considered to be significantly different from zero. In this case, the correlogram shows that phylogenetic relationships among species would have a significant effect on statistical analyses conducted at the genus, family, or superfamily levels.

considered if they are first converted into sets of dichotomous "dummy" variables (e.g., see Draper & Smith 1981).

2. Assumptions

The biological assumptions of the spatial autoregressive method have not been explicitly stated but include at least the following partially overlapping points: (a) within-species variation is negligible or incorporated into the analysis, (b) the phylogeny is known, (c) each species phenotype can be described as a linear function of the phenotypes of all of the other species on the phylogeny (i.e., an autoregressive model), (d) a species phenotype can be represented as a simple sum of phylogenetic and a nonphylogenetic components, and (e) only the non-phylogenetic component (ϵ) is of interest in the further study of the trait.

The spatial autoregressive method is a purely statistical approach which was developed in the absence of an explicit model of phenotypic evolution. Although this gives the method substantial robustness and flexibility to variation in the microevolutionary process, it also makes evolutionary interpretations difficult. In principle, the W matrix can be specified so as to correspond to a given phylogenetic correlation pattern, but the relationship is quite complex (Martins & Hansen, 1995). In computer simulation tests of the method, the W matrix has not been set in this way to correspond to the model actually used to generate the data (such as Brownian motion). Thus, it is not surprising that computer simulation results show that although this method provides reasonable estimates of the evolutionary relationship between two continuous characters, these estimates are almost always somewhat less accurate than those provided by the inferred changes and independent contrasts methods (Martins in review). The residual components of the autoregressive model are particularly difficult to interpret in an evolutionary framework. One possibility is that the residual components represent recent adaptations to a changing environment that occurred after each species split away from its most recent common ancestor. If species can respond that quickly to selective forces, however, it is not clear why their phenotypes would retain any phylogenetic component at all or why we would want to consider possible adaptation to the

environment along the most recent branches of the phylogeny but ignore it elsewhere.

3. Variations for dealing with unknown phylogenetic information

In another extension of the spatial autoregressive model, Gittleman and Kot (1990) proposed the use of a statistical estimator that allows flexibility in the phylogenetic connectivity matrix (\mathbf{W}) when phylogenetic information is uncertain or unavailable. Elements of the \mathbf{W} matrix describe the expected degree of phenotypic similarity between species. For example, the element of \mathbf{W} corresponding to species i and j (w_{ij}) is the inverse of the expected phenotypic divergence between those two species ($w_{ij} = d_{ij}^{-1}$, usually given as the inverse of the time that the two species have been evolving independently of one another). Gittleman and Kot (1990) proposed that $d_{ij}^{-\alpha}$ (or $\exp[-\alpha d_{ij}]$) be used, instead of d_{ij}^{-1} as the elements of \mathbf{W} , and showed how α can be estimated using maximum likelihood techniques. The parameter α stretches and shrinks the phylogeny (in much the same way as Grafen's ρ parameter above) to vary the phylogenetic connectivity matrix \mathbf{W} when the phylogeny or expected amount of phenotypic change along that phylogeny are not known. Alpha also serves as a rough estimate of the rate or tempo of phenotypic evolution in a particular character, given a particular phylogeny and value of ρ .

4. Strengths and weaknesses

The spatial autoregressive method has great heuristic appeal because it offers not only a solution to the problem of statistical dependence but also an estimate of the magnitude of this problem (i.e., the amount of variation in the trait that is due to phylogenetic history). Gittleman and Kot (1990) have contributed substantially to the method by providing useful exploratory and diagnostic tools to test the assumptions of the model and a statistical estimator (α) that allows for flexibility when phylogenetic information is unknown or uncertain. Representation of the phylogeny and model of phenotypic evolution as a connectivity matrix also gives this method somewhat more flexibility than the independent contrasts approach in that phylogenetic relationships need not be represented as a binary phylogeny. This may be of substantial

importance for studies involving (a) species hybridization or horizontal gene transfer, (b) ecological factors which may not evolve by the usual microevolutionary models or along a phylogeny, and (c) population or individual level comparisons (e.g., Edwards & Kot in press; Foster & Cameron, Chap. 5 in this volume).

Possibly the most important weakness of this method is that it can have extremely poor statistical performance when the model does not fit the data well. Computer simulation results suggest that if the comparative data really have little or no phylogenetic correlation or if the sample size is too small for any phylogenetic correlation to be detected (fewer than about 40 species), the spatial autoregressive method can give far more inaccurate parameter estimates and hypothesis tests than would not taking phylogeny into account at all (Martins in review; but see Gittleman & Luh 1992, 1993). Fortunately, Gittleman and Kot's (1990) diagnostic techniques can and should be used to identify such situations and avoid these potential problems (see also Gittleman & Luh 1992, 1993; Purvis et al. 1994). Martins' (in review) results also suggest that Gittleman and Kot's α parameter should be used cautiously. When a strong correlation between two characters exists, use of the α parameter can lead to biased results (strong correlations will be diminished). Thus, we recommend that analyses always be conducted both with and without the α parameter and that differences between the results of the two analyses be interpreted cautiously.

A further difficulty is that the method is purely univariate and must be applied to each trait separately. Correlations between traits are ignored in the estimation and correction of phylogenetic effects. Thus, when the method is used to estimate the correlation between two traits, we first assume that the relationship between characters is insignificant in comparison to the effect of phylogenetic history and fit spatial autoregressive models to each of the two traits independently. We then calculate a correlation coefficient between the residuals (i.e., the nonphylogenetic components) from the two models to determine whether there is a relationship between the two traits. If a substantial relationship between the traits exists, then the original assumption that the relationship between characters is insignificant in comparison to the phylogenetic effect has been violated. A better approach would include

both traits in a more complex transformation model (e.g., as suggested by Ely & Kurland 1989).

5. Computational difficulties and computer programs

The spatial autoregressive method is computationally intensive and requires a computer to do the analyses. Programs to conduct the analyses are available from several researchers. Cheverud and Dow's (1985b) MINRHO3 is a simple BASIC program which requires calls to a separate set of commercial matrix manipulation software and fits the original autoregressive model. Luh et al. (1994) offer a Macintosh program that will fit the spatial autoregressive model to a set of comparative data with the modifications proposed by Gittleman and Kot (1990). Martins' (1995) COMPARE is available for DOS/Windows, Macintosh and UNIX platforms, and will fit the spatial autoregressive model with and without the modifications by Gittleman and Kot (1990).

E. Comparing relationship matrices

1. The methods

As in the methods described above, given an appropriate model of phenotypic evolution, variation in any set of comparative data can be described as either a branching tree structure (a phenogram) or a phylogenetic relationship matrix. Thus, there have been several studies in which researchers constructed a phenogram or distance matrix describing the patterns of similarity observed in a set of interspecific data and compared this to a known phylogeny or phylogenetic distance matrix. The comparison can be made qualitatively or statistically, and relationships between the two types of trees or matrices can be interpreted as a measure of the amount of phylogenetic correlation or constraint in the data.

If interspecific measurements of a behavior pattern are described as a phenogram, this branching tree can then be compared to other trees (e.g., a phylogenetic hypothesis based on morphological or molecular data) to determine how similar the evolution of the behavior has been to the patterns of speciation in the clade. Phenograms based on behavioral trees can also be compared to trees based on other sorts of characters to

test whether the evolution of behavior has been qualitatively different from the evolution of other traits (e.g., morphology). Such comparisons can be conducted using standard techniques such as consistency or retention indices borrowed from traditional systematics (e.g., in this volume, de Queiroz & Wimberger, Chap. 7; Irwin, Chap. 8).

Alternatively, many authors have described a set of comparative data as an $N \times N$ phylogenetic relationship matrix where N is the number of taxa, and the elements of the matrix are some measure of the similarity or differences between each pair of taxa. Given an appropriate evolutionary or statistical model of phenotypic evolution, these matrices can be developed for individual characters or groups of characters and then compared, usually using some sort of permutation test. Matrix comparisons or "matrix correlation analysis" has been particularly popular in anthropology and population genetics when the taxa being considered are populations rather than species and where comparisons are generally conducted using Mantel's test (a permutation test; see Smouse and Long 1992 for review). At the species level, Legendre et al. (in press) proposed an alternative way of using multiple regression to compare the variation in a set of interspecific data (described as a phenetic distance matrix) with a phylogenetic relationship matrix based on independent information. In this method, a correlation coefficient is used to describe the relationship between the two matrices and a permutation test is used to correct for the non-independence within each matrix. The correlation coefficient provides a measure of phylogenetic correlation.

This general approach can be used with either continuous or categorical characters. Whether a specific method can be used with categorical or continuous characters depends primarily on the algorithms used to describe interspecific variation as a branching tree structure or as a relationship matrix.

2. Assumptions

There are several assumptions implicitly buried in (a) the creation of the branching trees or matrices from real data, and (b) the permutation tests or other statistics used to compare matrices. Algorithms for constructing phenograms or relationship matrices from comparative data usually make several assumptions about the process of phenotypic evolution

underlying the character that has been measured (e.g., Martins & Hansen 1995). Permutation tests or statistics that summarize the shape of a branching tree necessarily reduce the variation present in a set of comparative data by concentrating on a few important aspects of the matrix or tree. The choice of aspects to consider implicitly assumes that only these aspects are important in the evolution of the characters.

3. Variations for dealing with unknown phylogenetic information

None.

4. Strengths and weaknesses

The basic matrix comparison approach is intuitively simple and easy to perform. As mentioned below, calculations can be conducted using a wide variety of available computer programs.

There are many different ways of comparing matrices or branching structures and each makes its own set of limiting assumptions. Little effort has been made to view these methods in an explicitly evolutionary framework, so it is quite difficult to understand the meaning and effect of the exact assumptions made by different methods. Both matrix and tree comparison methods can also be extremely vulnerable to various scaling effects and may consider two sets of comparative data that are simply scaled versions of each other to be qualitatively different phenomena. Given the difficulties regarding both assumptions and interpretation of the results, we do not recommend use of these methods without a deep understanding of the evolutionary implications of the models underlying the analyses of the particular method being used.

5. Computational difficulties and computer programs

Most computer packages offering ways of reconstructing phylogenies also offer ways of building phenetic tree-like structures or relationship matrices from sets of comparative data (e.g., Maddison & Maddison 1992; Felsenstein 1993; Swofford 1993). These programs also usually calculate summary statistics such as consistency indices that can be used to compare branching trees to each another. Permutation tests for comparing relationship matrices are becoming increasingly common in standard statistical packages (e.g., SPSS, SAS).

F. Lynch's method

1. The method

Lynch (1991) developed another linear model to incorporate degree of relatedness into comparative analyses. He drew an analogy between individuals within populations and species within a phylogeny and expanded an existing quantitative-genetic model to partition variation in the observed species mean phenotypes (\bar{z}) into a phylogeny wide mean phenotype (μ), a phylogenetically heritable component (a) and a residual component (e) due to nonadditive phylogenetic effects, environmental effects and measurement error ($\bar{z} = \mu + a + e$). Maximum likelihood techniques can be used to estimate each of these components which might then be used in further analyses to examine the evolutionary histories of individual traits or of relationships among traits. For example, ancestral values of a character might be estimated from the mean phenotype (μ) and the heritable component (a) without the use of the nonheritable, residual component (e). The full model is multivariate, such that relationships between two characters can be estimated directly from the model as the correlation between the heritable components of the two traits.

The method is based on linear regression techniques and is therefore designed for use with continuously varying characters. As with Grafen's (1989) and Cheverud et al.'s (1985) methods (described above), it may also be reasonable to consider categorical or "state" characters as sets of dichotomous "dummy" variables with this method.

2. Assumptions

Lynch's (1991) method assumes that (a) phylogenetic information is available and can be represented as a matrix (G) of expected similarities among species and (b) interspecific variation in a trait resulting from shared evolution along a phylogeny is well described by a linear model with normally distributed, additive, phylogenetically heritable components and phylogenetically uncorrelated residual components that have a mean of zero. The first assumption is similar to that used in the spatial autocorrelation methods discussed above. The two methods differ substantially in interpretation, however, as the spatial

autocorrelation methods conduct further analyses using the residual component of the model, whereas Lynch suggests that the phylogenetically heritable component should be used. Essentially, Lynch's method views character evolution and phylogenetic correlation as one and the same process in much the same way as Felsenstein's contrasts method does. The spatial autoregression is more similar to Grafen's (1989) regression method in interpretation.

3. Variations for dealing with unknown phylogenetic information

None.

4. Strengths and weaknesses

The parameters of Lynch's model are easy to interpret via the analogy with similar parameters in quantitative genetics. As with the spatial autoregressive method, Lynch's technique also has the advantage of not being limited by the binary structure of a phylogeny, and it provides an estimate of the degree of phylogenetic correlation (his "phylogenetic heritability"). Lynch's method further improves on earlier methods by incorporating measures of within-species variability into the analysis.

The method is computationally difficult and requires an iterative procedure to converge on a final result. Some further development is required before the method can be put to practical use (Lynch, pers. comm.). The method has not yet been implemented in any distributed computer programs, and has not yet been applied to even a single set of real data. It must be regarded as promising but still in a preliminary stage of development.

5. Computational difficulties and computer programs

See above.

G. Nested ANOVA

1. The method

Nested ANOVA models (e.g., Clutton-Brock & Harvey 1977, 1979, 1984, Bell 1989) were the first to be suggested in the recent comparative

method literature (see Harvey & Pagel 1991 for review). The authors of these methods suggest that the phylogenetic relationships among species can be described as a set of hierarchical factors in ANOVA models. For example, if data were measured from 10 individuals in each of the six species in Fig. 1b, we might consider analyzing these data with a nested ANOVA in which the species data are nested within three groups representing the three different taxonomic classes of animals (mammals, birds and insects). With more species, a researcher might nest species within genera, genera within families, families within orders and so on. Thus, the method provides estimates of the effects due to belonging to a certain genus, family, and so on.

When this is done, adaptation can be investigated by choosing a taxonomic level that is relatively free of phylogenetic correlation for the analysis (Clutton-Brock & Harvey, 1977). For example, if species are very influenced by the genus to which they belong, but the genus effect is not strongly influenced by the family to which the genus belongs, the analysis can be performed with genus means as data points. Stearns' (1983) "phylogenetic-subtraction" method performs the analysis on the species level but first subtracts all the estimated higher order effects from the species data. This differs fundamentally in interpretation from Clutton-Brock and Harvey's original suggestion as Stearns uses the information which is discarded in the Clutton-Brock & Harvey approach and *visa versa*. The phylogenetic subtraction method studies recent species-specific adaptation by attempting to control for phylogenetic correlation due to evolution occurring prior to the species level.

ANOVA models of this sort were designed originally for use with continuous variables. Although nested logistic regression and log-linear models (the analog of ANOVA for categorical variables) could be used in much the same way as nested ANOVA to take phylogenetic relationships into account, this application has not been developed.

2. Assumptions

Nested ANOVA methods assume that: (a) phylogenetic relationships can be described as a strictly hierarchical tree structure in which each taxonomic group within a larger group (e.g., all species within a genus, all genera within a family) have evolved independently of one another as a "star" phylogeny and (b) the process of phenotypic evolution has

been constant throughout the entire clade and that each taxonomic group has evolved in the same way (at the same rate and for the same length of time) as all other taxa on the tree. These are essentially the same assumptions as those of the nonphylogenetic approach, adding only the hierarchical structure of a taxonomy and assuming that this taxonomy directly reflects phylogenetic and evolutionary relationships. The validity of the nested ANOVA approach depends primarily on the extent to which the taxonomic clustering used in the nested ANOVA is similar to a phylogeny describing the relationships among species.

3. Variations for dealing with unknown phylogenetic information

None.

4. Strengths and weaknesses

The nested ANOVA approach has been popular, as it is relatively simple to apply using most standard statistical computer packages (e. g., SPSS, SAS). It is useful as an exploratory tool to describe the broad patterns of phylogenetic correlation. Harvey & Pagel (1991) termed this method obsolete because it neither solves the problem of statistical dependence nor proposes better estimates of evolutionary relationships. Rather than assuming that species are independent, it assumes that genera, families or orders are independent. When only hierarchical taxonomic information is available, these assumptions may seem reasonable and the method may still be useful. If any alternative phylogenetic information is available, any of the methods that allows for greater flexibility in the phylogenetic structure may be preferable.

5. Computational difficulties and computer programs

Nested ANOVA is a standard parametric statistical technique that is implemented in most commercial statistical software.

Discussion

The main goal of most statistical analyses in animal behavior is to obtain an estimate of a statistical parameter and some measure of its uncertainty. For example, we might use statistics to tell us whether the

home range sizes of dominant animals are larger or smaller than those of subordinate animals and whether that difference is greater than what would be expected by chance alone. A comparison of the mean home range size for each group of animal will give us an answer to the first question, and a standard error, a confidence interval or — less informatively — a p value from a hypothesis test can be used to answer the second. Both answers, however, will be given within the bounds of some assumed statistical model and the range of available data. For example, if the data in our study were measured from individual sparrows, we would have some reservations about any conclusions made about the behavior of lizards from those data. Similarly, any conclusions or predictions made outside the bounds of the statistical model may be subject to error.

When interspecific data and phylogenetic comparative methods are considered, the statistical model is a complex one, consisting of three parts: (a) any hypothesis of the patterns of speciation underlying the measured species (i.e., a phylogeny, often with branch lengths), (b) a model or assumption regarding how this phylogeny translates into the expected relationships among species phenotypes (e.g., parsimony, the Brownian motion assumption underlying Felsenstein's independent contrasts, the $y = \rho W y + \varepsilon$ model of the spatial autocorrelation method); and (c) a statistical model or assumption describing how the parameter we are interested in (e.g., difference in home range size) can be estimated from the measured data (e.g., a specified regression model). Thus, any conclusion based on the statistical analysis of interspecific data will depend on the accuracy and appropriateness of the assumptions made in each of these three parts, whether or not these assumptions are explicitly specified.

For example, imagine a study in which we want to know how much of the variation in mammal home range size can be predicted by differences in body size. We have measured the home range and body sizes of several different species of mammals and estimate the slope of a simple linear regression (b_1) of home range size (the response or y variable) on body size (the predictor or x variable) without making any attempt to take phylogenetic information into account. We use the statistical model: $y = b_0 + b_1 x + e$ (where b_0 gives us a y intercept and e is an error term), and make a number of assumptions. For example, if we want the regression slope to be an unbiased estimate of the "true"

relationship between these two characters, we might be assuming in the above three-part assumption, that: (a) the phylogeny of mammals is a "star" phylogeny, as in Fig. 1a, (b) both home range and body size have evolved at the same rate along each branch of that phylogeny, and (c) home range size is linearly related to body size, no within-species variation exists, and the error term in the regression model consists of independent, normally distributed variables. To estimate a standard confidence interval on the resulting regression slope (or to conduct a hypothesis test), we must also assume that the error terms in the regression model are statistically uncorrelated with each other and have the same variance. Other sets of assumptions are also possible. For example, instead of the first two of the above assumptions, we might assume that brain and body size were responding so quickly to selective pressures in these species that no evidence of shared phylogenetic histories remains in the data. In this case, the shape of the phylogeny (assumption a), and the rate of phenotypic evolution (assumption b) need not be known. The third assumption remains unchanged.

Lack of specificity or vagueness in the underlying model does not mean that a method is more robust or general. For example, Felsenstein's independent contrast method has been viewed skeptically due to its Brownian motion assumption and its requirement of phylogenetic information. All of the existing methods, however, make comparable sorts of assumptions either implicitly or explicitly, since they must all address the three parts of the comparative method model listed above. For example, as reviewed above, Grafen's (1989) phylogenetic regression assumes that the error term in his multiple regression model consists of normally distributed variables with specified variances and covariances. As it stands, Brownian motion is the only microevolutionary model grounded in a population and quantitative genetic framework that produces the variance-covariance matrix actually used in Grafen's applications. What Grafen's method gains by allowing for the possibility of other, unknown microevolutionary models, it loses in terms of the evolutionary interpretation of the results. Similarly, any specification of the connectivity matrix in the phylogenetic autocorrelation method implicitly assumes a model on the same level of detail as Brownian motion. In fact, none of the popular, existing microevolutionary models will produce the patterns of covariation among species assumed by

existing applications of the method (Martins & Hansen 1995). In our opinion, given that most of the existing methods make comparable assumptions, explicitness of those assumptions should be viewed as a strength rather than a weakness of a phylogenetic comparative method.

In fact, the specification of the above three-part assumption is a major (and usually unquantified) source of error in the statistical analysis of comparative data. The robustness of different phylogenetic methods to errors in these assumptions and the incorporation of different aspects of this complex model lie at the forefront of the development of modern phylogenetic comparative methods. Most modern comparative methods were originally designed to incorporate known phylogenetic information into the analysis of comparative data, and thereby explicitly address assumptions a and b. Several other methods have been designed to infer the needed phylogenetic information when it is not known (e.g., Grafen 1989, Gittleman & Kot 1990, Pagel 1992; reviewed above), or to incorporate uncertainty in that information into the final estimates of the uncertainties of parameter estimates (e.g., Losos 1995, Martins in press).

Computer simulation studies have shown that several of the available phylogenetic comparative methods can produce estimates of evolutionary relationships with reasonable rates of Type I error. One concern regarding these studies is that many put an unfortunate emphasis on statistical hypothesis testing over parameter estimation. The ubiquitous p values found in empirical studies can be very misleading in comparative analyses (or any behavioral study; see Yoccoz 1991 for documentation and discussion). The most common and serious of these misinterpretations is to confuse statistical significance with biological importance. For example, in determining whether body size is a good predictor of home range size, the question should not be whether a relationship between the two traits exists at all but rather whether the relationship is large enough to be of biological importance. Given any association between the two traits, no matter how weak or biologically unimportant, we only need a large enough sample size to make the estimated correlation significantly different from zero. Similarly, when the sample size is small, a statistical hypothesis test may fail to reject a null hypothesis of no relationship between two traits on some specified significance level, even though an important biological relationship exists and is, in fact, indicated by the estimate.

The best conclusion in this situation is not that "there is no evidence for a relationship" but rather that the best estimate of the parameter indicates an important pattern, although this estimate is too uncertain to be taken very seriously.

On an absolute scale, all of the phylogenetic comparative methods that have been tested (i.e., several versions of inferred changes, independent contrasts, and spatial autoregressive methods; all designed for use primarily with continuous characters) produce unbiased, reasonably accurate estimates that are usually far superior to those produced using a nonphylogenetic approach. Although some techniques will not perform well when the species are actually relatively independent of one another (i.e., as depicted in the phylogeny on the left of Fig. 1) or when convergence of unrelated species diminishes the impact of evolutionary history, diagnostic techniques can be used to identify these situations and to avoid poor method performance. Other methods perform reasonably well with any phylogeny and model of evolutionary change when these are known, and all of the tested methods are also relatively robust to inaccuracies in the available phylogenetic information.

More importantly, phylogenetic methods have opened up many new horizons in animal behavior, by allowing us to answer evolutionary questions that did not seem possible previously given traits that do not appear in the fossil record and which may not always be suitable for phylogenetic reconstruction (but see, in this volume, Irwin, Chap. 8, and de Queiroz & Wimberger, Chap. 7). Using phylogenetic methods, we can ask about the rate of behavioral change (e.g., in this volume, Gittleman et al., Chap. 6), the sequence of changes (e.g., in this volume, Crespi, Chap. 9), etc. We can also solve a troublesome statistical problem that has been the subject of much recent discussion. As new techniques for addressing the problems of unknown or uncertain phylogenetic information are developed, there seems to be very little to lose by applying a phylogenetic comparative method and much to gain. Thus, phylogenetic analyses of comparative studies will probably one day become as widely used as repeated-measures ANOVA or paired sample *t*-tests as a simple way of correcting a common statistical problem and a much more powerful tool for exploring questions in the evolution of behavior.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation to EPM (#DEB-9406964).

References

- Barry, D., & J. A. Hartigan. 1987. Statistical analysis of hominoid molecular evolution. *Stat. Sci.*, 2, 191–210.
- Basolo, A. L. 1990a. Female preference for male sword length in the green swordtail, *Xiphophorus helleri* (Pisces: Poeciliidae). *Anim. Behav.*, 40, 332–338.
- Basolo, A. L. 1990b. Female preference predates the evolution of the sword in swordtail fish. *Science*, 250, 808–810.
- Bell, G. 1989. A comparative method. *Am. Nat.*, 133, 553–571.
- Boake, C. R. B. 1989. Repeatability: its role in evolutionary studies of mating behavior. *Evol. Ecol.*, 3, 173–182.
- Brooks, D. R., & D. A. McLennan. 1991. *Phylogeny, Ecology, and Behavior*. Chicago: Chicago University Press.
- Burt, A. 1989. Comparative methods using phylogenetically independent contrasts. *Oxford Surveys in Evolutionary Biology*, 6, 33–53.
- Cheverud, J. M., & M. M. Dow. 1985a. An autocorrelation analysis of genetic variation due to lineal fission in social groups of rhesus macaques. *Am. J. Phys Anthropol.*, 67, 113–121.
- Cheverud, J. M., & M. M. Dow. 1985b. *MINRH03*. Distributed by the authors. Evanston, Illinois: Northwestern University.
- Cheverud, J. M., M. M. Dow, & W. Leutenegger. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weights among primates. *Evolution*, 39, 1335–1351.
- Cliff, A. D., & J. K. Ord. 1981. *Spatial Processes: Models and Applications*. London: Pion Press.
- Clutton-Brock, T. H., & P. H. Harvey. 1977. Primate ecology and social organization. *J. Zool. Lond.*, 183, 1–39.
- Clutton-Brock, T. H., & P. H. Harvey. 1979. Comparison and adaptation. *Proc. R. Soc. Lond. B*, 205, 547–565.
- Clutton-Brock, T. H., & P. H. Harvey. 1984. Comparative approaches to investigating adaptation. In: *Behavioral ecology: An evolutionary approach*. 2nd ed. (Ed. by J. R. Krebs & N. B. Davies), pp. 7–29. Oxford, England: Blackwell Press.
- Draper, N., & Smith. 1981. *Applied Regression Analysis*. New York: John Wiley and Sons, Inc.
- Edwards, A. F. W., & L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. In: *Phenetic and phylogenetic classification*. (Ed. by W. H. Heywood & J. McNeill), pp. 67–76. London: Systematics Association Publication No. 6.
- Edwards, S. V., & M. Kot. (in press). Comparative methods at the species level: Geographic variation in morphology and group size in Gray-crowned Babblers (*Pomatostomus temporalis*). *Evolution*.

- Ely, J., & J. A. Kurland. 1989. Spatial autocorrelation, phylogenetic constraints, and the causes of sexual dimorphism in primates. *Int. J. Primatol.*, 10, 151–171.
- Farris, J. S. 1970. Methods for computing Wagner trees. *Syst. Zool.*, 19, 83–92.
- Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.*, 126, 1–25.
- Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, 19, 445–471.
- Felsenstein, J. 1993. *PHYLIP (Phylogeny Inference Package) version 3.5*. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Felsenstein, J., & E. Sober. 1986. Parsimony and likelihood: An exchange. *Syst. Zool.*, 35, 617–626.
- Frumhoff, P. C., & H. K. Reeve. 1994. "Using phylogenies to test hypotheses of adaptation: a critique of some current proposals." *Evolution*, 48, 172–180.
- Gaillard, J.-M., D. Allainé, D. Pontier, N. G. Yoccoz, & D. E. L. Promislow. 1994. Senescence in natural populations of mammals: a reanalysis. *Evolution*, 48, 509–516.
- Garland, T. Jr. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am. Nat.*, 140, 509–519.
- Garland, T. Jr., P. H. Harvey, & A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrast. *Syst. Biol.*, 41, 18–32.
- Gittleman, J. L., & M. Kot. 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Zool.*, 39, 227–241.
- Gittleman, J. L., & H.-K. Luh. 1992. On comparing comparative methods. *Ann. Rev. Ecol. Syst.*, 23, 383–404.
- Gittleman, J. L., & H.-K. Luh. 1993. Phylogeny, evolutionary models and comparative methods: A simulation study. In: *Phylogenetics and Ecology* (Ed. by P. Eggleton & D. Vane-Wright), pp. 103–122. London: Academic Press.
- Grafen, A. 1989. The phylogenetic regression. *Phil. Trans R. Soc. B*, 326, 199–157.
- Grafen, A. 1992. The uniqueness of the phylogenetic regression. *J. theor. Biol.*, 156, 405–423.
- Grafen, A., & M. Ridley. (in press). Statistical tests for discrete cross-species data. *J. Theor. Biol.*,
- Hall, B. K. (Ed.) 1994. *Homology: the Hierarchical Basis of Comparative Biology*. San Diego, California: Academic Press.
- Hansen, T. F., & E. P. Martins. (in press). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*,
- Harvey, P. H., & M. D. Pagel. 1991. *The Comparative Method in Evolutionary Biology*. Oxford, England: Oxford University Press.
- Hendy, M. D. 1989. The relationship between simple evolutionary tree models and observable sequence data. *Syst. Zool.*, 38, 310–321.
- Hendy, M. D., & D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.*, 38, 297–309.
- Hillis, D. M., & C. Moritz. 1990. *Molecular Systematics*. Sunderland, Massachusetts: Sinauer and Associates.

- Huey, R. B., & A. F. Bennett. 1987. Phylogenetic studies of coadaptation: Preferred temperatures versus optimal performance temperatures of lizards. *Evolution*, 41, 1098–1115.
- Janson, C. H. 1992. Measuring evolutionary constraints: a Markov model for phylogenetic transitions among seed dispersal syndromes. *Evolution*, 46, 136–158.
- Krebs, J. R. & N. B. Davies. 1987. *An Introduction to Behavioural Ecology*. Sec Ed. Oxford. Blackwell Scientific Publications.
- Legendre, P., F.-J. Lapointe, & P. Casgrain. (in press). Modeling brain evolution from behavior: A permutational regression approach. *Evolution*.
- Leroi, A. M., M. R. Rose, and G. V. Lauder. 1994. "What does the comparative method reveal about adaptation?" *Am. Nat.* 143, 381–402.
- Losos, J. B. 1995. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Syst. Biol.* 43, 117–123.
- Luh, H-K., J. Gittleman, & M. Kot. 1994. *PA: Spatial autoregression computer program*. Distributed by the authors. Department of Zoology, University of Tennessee, Knoxville.
- Lynch, M. 1991. Methods for analysis of comparative data in evolutionary biology. *Evolution*, 45, 1065–1080.
- Machlis, L., P. W. D. Dodd, & J. C. Fentress. 1985. The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Z. Tierpsychol.*, 68, 201–214.
- Maddison, D. R. 1994. Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. *Annu. Rev. Entomol.*, 39, 267–292.
- 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, 539–537.
- Maddison, W. P. 1991. Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Syst. Zool.*, 40, 304–314.
- Maddison, W. P., & D. R. Maddison. 1992. *MacClade: Analysis of Phylogeny and Character Evolution*. Sunderland, Massachusetts: Sinauer and Associates.
- Martin, P., & H. C. Kraemer. 1987. Individual differences in behaviour and their statistical consequences. *Anim. Behav.*, 35, 1366–1375.
- Martins, E. P. 1994. Estimating rates of character change from comparative data. *Am. Nat.*, 144, 193–209.
- Martins, E. P. 1995. *COMPARE: statistical analysis of comparative data, version 1.0*. Distributed by the author. Department of Biology, University of Oregon, Eugene.
- Martins, E. P. (in press). Conducting phylogenetic comparative analyses when the phylogeny is not known. *Evolution*.
- Martins, E. P. (in review). Phylogenies, spatial autoregression, and the comparative method: A computer simulation test. *Evolution*.
- Martins, E. P., & T. Garland, Jr. 1991a. *CMA: Comparative Method Analysis Package*. Distributed by the author. Department of Zoology, University of Wisconsin, Madison.
- Martins, E. P., & T. Garland, Jr. 1991b. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution*, 45, 534–557.

- Martins, E. P., & T. F. Hansen. 1995. A microevolutionary link between phylogenies and comparative data. In: *New Uses for New Phylogenies*. (Ed. by P. Harvey, J. Maynard-Smith, & A. Leigh-Brown) Oxford, England: Oxford University Press.
- McArdle, B., & A. G. Rodrigo. 1994. Estimating the ancestral states of a continuous-valued character using squared-change parsimony: An analytical solution. *Syst. Zool.* 43, 573–577.
- McKittrick, M. C. 1993. Phylogenetic constraint in evolutionary theory: Has it any explanatory power? *Annu. Rev. Ecol. Syst.*, 24, 307–330.
- Meyer, A., J. M. Morrissey, & M. Schartl. 1994. Recurrent origin of a sexually selected trait in *Xiphophorus* fishes inferred from a molecular phylogeny. *Nature*, 368, 539–542.
- Miles, D. B., & A. E. Dunham. 1993. Historical perspectives in ecology and evolutionary biology: The use of phylogenetic comparative analysis. *Annu. Rev. Ecol. Syst.*, 24, 587–619.
- Møller, A. P., & T. R. Birkhead. 1992. A pairwise comparative method as illustrated by copulation frequency in birds. *Am. Nat.*, 139, 644–656.
- Nowak, R. M., & J. L. Paradiso. 1983. *Walker's Mammals of the World*. Baltimore: Johns Hopkins University Press.
- Pagel, M. D. 1992. A method for the analysis of comparative data. *J. Theor. Biol.*, 156, 431–442.
- Pagel, M. D. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B*, 255, 37–45.
- Parker, G. A., & J. Maynard Smith. 1990. Optimality theory in evolutionary biology. *Nature*, 348, 27–33.
- Purvis, A. & A. Rambaut. (in press). Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.*
- Purvis, A., J. L. Gittleman, & H-K. Luh. 1994. Truth or consequences: Effects of phylogenetic accuracy on two comparative methods. *J. Theor. Biol.*, 167, 293–300.
- Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.*, 35, 17–31.
- Ridley, M. 1983. *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating*. Oxford, England: Clarendon press.
- Ritland, K., & M. T. Clegg. 1987. Evolutionary analysis of plant DNA sequences. *Am. Nat.*, 130, S74–S100.
- Sanderson, M. J. 1991. In search of homoplastic tendencies: Statistical inference of topological patterns in homoplasy. *Evolution*, 45, 351–358.
- Sanderson, M. J. 1993. Reversibility in evolution: A maximum likelihood approach to character gain/loss bias in phylogenies. *Evolution*, 47, 236–252.
- Sillén-Tullberg, B. 1988. Evolution of gregariousness in aposematic butterfly larvae: A phylogenetic analysis. *Evolution*, 42, 293–305.
- Sillén-Tullberg, B. 1993. The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution*, 47, 1182–1191.
- Smouse, P. E., & Long, J. C. 1992. Matrix correlation analysis in anthropology and genetics. *Yearbook Phys. Anthropol.*, 35, 187–213.

- Stearns, S. C. 1983. The influence of size and phylogeny on patterns of covariation among life-history traits in mammals. *Oikos*, 41, 173–187.
- Swofford, D. L. 1993. *PAUP: A computer program for phylogenetic inference using maximum parsimony*. Washington D. C.: Smithsonian Institution.
- Swofford, D. L., & W. P. Maddison. 1987. Reconstructing ancestral states under Wagner parsimony. *Math Biosci.*, 87, 199–229.
- Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull. Ecol. Soc. Am.*, 72, 106–111.