

Invited Perspectives in Physiological Zoology

Why Not to Do Two-Species Comparative Studies: Limitations on Inferring Adaptation

Theodore Garland, Jr.¹

Stephen C. Adolph²

¹Department of Zoology, 430 Lincoln Drive, University of Wisconsin, Madison, Wisconsin 53706; ²Department of Biology, Harvey Mudd College, 301 E. Twelfth Street, Claremont, California 91711

Accepted 3/9/94

One thing cannot be evaluated unless it can be compared with another. This is, of course, why degrees of freedom in statistics are the number of observations minus one. [BRADSHAW 1987a, p. 71]

Adaptation can only be measured and indeed discussed on a comparative basis. . . . Adaptation is entirely a comparative concept. [BRADSHAW 1987a, p. 71]

Interspecific comparison is a common approach in physiological ecology, comparative physiology, and biochemistry, and in such related fields as functional morphology and ethology (Hochachka and Somero 1984; Feder et al. 1987; Brooks and McLennan 1991; Harvey and Pagel 1991). In their most basic form, comparisons are used simply to identify which characteristics differ among species. Sometimes the goal is to identify alternative physiological or biomechanical mechanisms (multiple solutions; see, e.g., Bartholomew 1987) that have achieved a similar functional endpoint (e.g., longer legs vs. faster muscles, either of which may cause higher maximal sprinting abilities), or perhaps to identify new "models" in which to study particular phenomena (see, e.g., Faraci, Kilgore, and Fedde 1984; Kellogg and Shaffer 1993).

Interspecific comparisons are also used frequently to elucidate the endpoint and/or the process of evolutionary adaptation, that is, genetic changes in response to natural selection (reviews in Harvey and Pagel 1991; Miles and Dunham 1993; Losos and Miles 1994). Specifically, interspecific correlations between some aspect(s) of the phenotype (e.g., low rates of evaporative water loss) and some aspect(s) of the environment (e.g., heat and aridity) are taken to indicate that past and/or present natural selection acting

on the character(s) of interest has played some role in causing its diversification and/or maintenance. Although most workers accept that character-environment correlations provide at least some evidence pertaining to adaptation, the reader should note that how best to study adaptation—and even how to define it—is a very controversial subject in evolutionary biology (reviews in Brooks and McLennan 1991; Reeve and Sherman 1993; Bennett 1994; Leroi, Rose, and Lauder 1994).

Claims about adaptation are commonly made from comparative studies involving only two species (or only two populations of a single species). Our main purpose here is to alert practitioners to several logical and statistical problems associated with using two-species comparisons for studying adaptation and to outline some alternative approaches. Multispecies comparisons are one such alternative. However, data from multiple species may not be independent or identically distributed in the statistical sense, which violates assumptions of ordinary statistical methods (Harvey and Pagel 1991). We therefore also discuss one phylogenetically based statistical method that can be employed for valid hypothesis testing with comparative data, Felsenstein's (1985) method of phylogenetically independent contrasts. We discuss briefly how such methods can be employed, even with incomplete phylogenetic information, and also how data for multiple populations within species can enhance comparative analyses. Phylogenetically based analyses come in a variety of flavors, and our penultimate section discusses some differences in perspectives regarding *statistical* hypothesis testing in a phylogenetic context. We conclude by pointing out that many of our criticisms of two-species comparisons apply also to comparisons aimed at discovering mechanisms underlying physiological differences between species.

How Common Are Two-Species Comparative Studies?

We reviewed the last 5 yr of *Physiological Zoology* (vols. 62–66) to ascertain the commonness of two-species comparative studies. We have published on such comparisons before and have offered adaptive interpretations (see, e.g., Sinervo and Adolph 1989), as have many of our colleagues and mentors. Thus, the present survey is intended not to criticize what has come before but to support our claim that two-species comparisons are common. We tabulated original data articles, omitting review articles, methods articles, analyses of literature data, and invited perspectives. Of 316 data articles, 229 (72%) dealt with a single species, 49 (16%) dealt with two species, and 38 (12%) dealt with three or more species. We judged that 18 of the two-species articles (37%) drew inferences about the adaptive significance of

differences (or, rarely, the absence of differences) between species. Adaptive interpretations were also common among the multispecies studies. The word “adaptation” is also used frequently in titles of articles, even when the authors do not really discuss their results in this context: such usage seems to indicate a presumption of adaptation.

To make these numbers more tangible, we highlight two recent examples. Hinsley et al. (1993) studied two closely related species of sandgrouse, one of which (*Pterocles alchata*) “also occurs in hotter and more arid regions. . . . It might therefore be expected that, under hot conditions, the pin-tailed sandgrouse would be the better thermoregulator. . . . We tested this hypothesis by comparing the metabolism, evaporative heat loss, and temperature control of these two species” (Hinsley et al. 1993, p. 21). Their prediction was partly supported by the data. In a similar way, Quinlan and Hadley (1993) studied respiration and water loss in two species of grasshoppers inhabiting different climatic zones. They suggested (p. 636) that “higher metabolic rate may be an adaptation to the reduced growing season available to *Taeniopoda*.”

In both of the foregoing cases, adaptation may well be responsible for the physiological differences observed between species. However, drawing this conclusion relies, in part, on the assumption that these species would be physiologically identical in the absence of adaptation to their respective climates. In the following sections, we consider some statistical and evolutionary concepts that call into question this assumption. We also note that some other recent articles display a cautious attitude toward drawing adaptive inferences on discovering differences between two species or between two populations. For example, although Beaupre, Dunham, and Overall (1993) found possible physiological differences between two populations, they did not attempt to attribute the differences to local adaptation. Instead, they discussed some of the perils of assuming physiological homogeneity among populations when physiological information is used to construct energy budgets and other ecological extrapolations from physiological data.

Limitations of Two-Species Comparative Studies

Physiologists are well versed in *experimental* studies. A typical example might involve purchasing 20 same-sex and same-age laboratory rats from a commercial supplier, that is, from a single interbreeding population. Ten rats would be assigned randomly to each of two groups, a control group and an experimental group. The experimental group would then receive some sort of treatment, perhaps exercise training, for several weeks. All

other conditions would be maintained identically for both groups (e.g., food, water, temperature, photoperiod, with both groups housed individually in cages randomly interspersed in the same room). At the end of the experimental treatment, the two groups would be measured for some dependent variable of interest, perhaps basal metabolic rate. A simple t -test or, possibly, an ANCOVA, with body mass as a covariate, could be used to compare the groups. The t or F statistic for the group effect would be compared to a critical value determined by reference to a standard statistical table (see, e.g., table 12 or 16 in Rohlf and Sokal 1981). With the Type I error rate, α , controlled at 0.05 (i.e., with the P value set at 0.05), one would expect to find a statistically "significant" difference owing simply to chance effects about one time in 20. That is, regardless of any experimental treatment that was applied, the two groups might differ by chance (e.g., owing to the original assignment to groups) with a probability of 0.05. This 5% possibility of falsely claiming a treatment effect when none really exists is traditionally considered as acceptable by the scientific community.

Because of differences in training, physiologists may be less well versed in the subtleties and complexities of methodologies now employed by evolutionary biologists (see also Bennett and Huey 1990; Burggren and Bemis 1990; Huey and Bennett 1990; Burggren 1991; Garland and Carter 1994). Species differ as the result of evolutionary processes, and the implications of these processes must be considered. Thus, analytical methods used by evolutionary biologists often acknowledge evolutionary processes in an explicit fashion (see, e.g., Endler 1986; Feder et al. 1987; Huey and Bennett 1987, 1990; Patton and Brylski 1987; Cohan and Hoffman 1989; Otte and Endler 1989; Bennett and Huey 1990; Burggren and Bemis 1990; Losos 1990, 1994; Baum and Larson 1991; Brooks and McLennan 1991; Harvey and Pagel 1991; James 1991; Martins and Garland 1991; Lynch 1992; Maddison and Maddison 1992; Kellogg and Shaffer 1993; Malhotra and Thorpe 1993; Garland and Carter 1994; Leroi et al. 1994; Losos and Miles 1994).

A typical *comparative* study undertaken by a physiologist might involve measuring, under controlled laboratory conditions, one or more physiological traits on each of 10 individuals from each of two species. Of course, unless all individuals were raised in the laboratory under identical conditions, at least from birth and preferably from conception or even as members of a second laboratory-reared generation, we cannot be certain that any differences we may find are the result of genetic differences between the species (Patton and Brylski 1987; Garland and Adolph 1991; Adolph and Porter 1993). We shall ignore this limitation in the present discussion.

A far more serious limitation of two-species comparisons, with respect to inferring adaptation, is as follows. Individuals from two species cannot be

considered the equivalent of individuals from two groups whose members were drawn randomly from a uniform population, such as in the rat example described above. For the latter, the appropriate null hypothesis is no difference with respect to any measure of the phenotype. But, for two species, the appropriate null hypothesis is more likely to be that a *difference* exists for any phenotypic trait.

Our point is in many ways similar to that of Hurlbert's (1984) discussion of "pseudoreplication" in ecological field experiments. For example: "Certainly, in any field situation, we *know* that two replicate plots or ponds in the same treatment are not identical. It may be of interest to know roughly *how* different they are, but a significance test of the difference is irrelevant" (Hurlbert 1984, p. 205).

As with experimental plots, differences between species are almost certain to exist for several reasons. First, although the relationship between speciation per se (i.e., the evolution of reproductive isolation) and phenotypic evolution is poorly understood, the process of speciation itself may result in genetic differentiation, which typically affects various phenotypic traits (Otte and Endler 1989). Second, the two species will (by definition) have experienced little or no genetic exchange since the time of their evolutionary divergence (the cladogenic event), and so, at a minimum, they will have diverged somewhat because of random genetic drift alone. Third, they will likely have experienced different environmental conditions (broadly defined) and so will have experienced different selective pressures and consequent adaptation. This third probable cause of species differences is, of course, what motivates the study of adaptation by comparing species. Note that random genetic drift can be opposed by uniform selection pressures that might occur in each species. However, perfectly identical selection pressures for different species seem extremely unlikely and, even given uniform selection pressures, the same genetic and phenotypic response is not guaranteed (see, e.g., Robertson 1980; Bartholomew 1987; Cohan and Hoffman 1989; Hill and Caballero 1992).

The point of the previous paragraph is that, for any two species (closely related or not), comparison of any phenotypic trait is *likely* to reveal a statistically significant difference, at least given a reasonably large sample size. Thus, the appropriate null hypothesis for comparing two species is something closer to a difference rather than no difference. The typical alternative hypothesis in species comparisons focusing on adaptation is a positive correlation between the presumed selective force (e.g., altitude, temperature) and the trait that is expected to vary adaptively (e.g., hemoglobin level, critical thermal maximum). Thus, a one-tailed test can be applied (e.g., the species living at the higher altitude should have higher hemoglobin levels).

Consequently, if our appropriate null hypothesis is that a difference will exist between the two species, then we have a 50% chance of accepting our one-tailed alternative hypothesis because of chance alone. In other words, if we accept that *any* two species will differ, then 50% of the time the difference will be in the direction predicted by our alternative hypothesis. A Type I error rate of 0.50 is an order of magnitude greater than the conventionally accepted $\alpha = 0.05$.

We can offer two other, equally startling perspectives on the (im)prudence of attempting to infer evolutionary adaptation from a two-species comparison. First, as discussed above, species comparisons rely on the demonstration of a correlation between a feature of the environment (or, in a broader sense, a presumed selective regime; sensu Baum and Larson 1991) and a phenotypic trait presumed to affect performance with respect to the environmental factor. Testing the statistical significance of a correlation or, equivalently, of a regression requires a minimum of three data points, because degrees of freedom are $N - 2$ (Sokal and Rohlf 1981). When the mean phenotypes of two species are related to the mean values of their environments, a statistically significant association cannot possibly be demonstrated, because degrees of freedom are $N - 2 = 0$! This conundrum is evaded when a *t*-test, ANOVA, or ANCOVA is used to compare the species with the use of data for each measured individual, and then any difference is just verbally related to the known or presumed (and often not measured) environmental differences between species. This evasion obviously does not validate an adaptive interpretation.

Second, attempting to infer adaptation from a two-species comparison inevitably involves the confounding of independent variables. The independent variables are (1) the environmental factor (presumed selective regime) and (2) species membership. Again, we can illustrate this problem by analogy with an experimental study. Imagine 20 rats drawn at random from the same population, with 10 assigned randomly to a control group and 10 to a treatment group. The appropriate null hypothesis is no difference between these two groups for any aspect of the phenotype that we might wish to measure. But now imagine that the 20 rats, although coming from a single population, represented 10 males and 10 females and that each sex was assigned entirely to either the control or the experimental group. Alternatively, the control and treatment groups might be housed in two different rooms, or even on two different shelves in the same room. Any of these three scenarios would perfectly confound the independent variable of interest—the experimental treatment to be applied—with another factor that might affect the dependent variable to be studied. Thus, after the experimental treatment has been applied, any difference between control and

experimental groups could, in principle, be caused either by the treatment or by the other factor. No scientist would ever make such an obvious mistake, yet comparing two species is exactly analogous: variation in the environmental factor and variation in species membership are perfectly confounded.

A final limitation of two-species comparisons is that, given a difference between species, they do not allow inference as to which species has the derived character state; at least three species, one serving as an "out-group," are required to suggest the direction of past evolutionary change within the "in-group" (e.g., Huey and Bennett 1987; Baum and Larson 1991; Brooks and McLennan 1991; Chevalier 1991). In summary, drawing inferences about adaptation solely on the basis of data from a two-species comparative study is extremely tenuous, both for statistical and logical reasons.

Enhancing Two-Species Comparisons through a Multivariate Approach

One way to enhance the value of two-species comparisons is to make a priori predictions about each of several independent traits (e.g., each trait will be higher in the species living at high altitude). Assume, as we have argued, that any two species are likely to differ for almost any aspect of the phenotype, and, hence, the probability of finding a difference in the predicted direction for any single trait approaches 0.5 (with symmetry assumed in the distribution of possible differences). With this assumed, then the probability that each of several traits will differ in the predicted direction is 0.5^K , where K denotes the number of *independent* characters measured. Thus, if two species were compared, and the a priori prediction had been that each of five characters would differ *in a specific direction*, then the probability of obtaining by chance alone the predicted differences for all five characters would be $(0.5)^5 = 0.03125$ (this is essentially a sign test). Bennett, Huey, and John-Alder (1984) measured several traits predicted a priori to show differences between two lizard species differing in speed and stamina. They did not, however, attempt the statistical test we have proposed, and with good reason; several of the traits studied clearly could not be considered as independent: for example, heart mass, hematocrit, and stamina; and muscle contractile properties and sprint speed.

In general, demonstrating that the traits of interest are and have been evolutionarily and statistically independent could be extremely problematic (cf. Leroi et al. 1994). It would require, among other things, demonstrating that the genetic correlations between each and every pair of traits are and have been zero (cf. Dohm and Garland 1993). Moreover, *any* evidence

deriving from only two species borders on the anecdotal. Generalizing from a two-species comparison would also be much more tenuous than generalizing from a multispecies study, especially given that the members of the two-species comparison likely will not have been chosen randomly for study.

Multispecies Comparative Studies as an Alternative

The statistical problems plaguing two-species comparisons of one trait can be overcome by expanding the comparative database. Doing so allows a switch to tests based on parametric or nonparametric correlation or regression. To correlate a trait with an environmental feature, three is the minimum number of species' means to which a formal statistical test can be applied, which yields 1 df. This is true whether one is doing a one-tailed or a two-tailed test. As noted above, one-tailed tests are perhaps the more common in physiological ecology, because practitioners typically have an *a priori* expectation as to what would be adaptive (i.e., what past natural selection would have favored, such as higher hemoglobin levels at higher altitudes). One-tailed tests also have higher power to detect significant relationships.

What about the Type I error rates of correlations involving mean values for only three species (or populations)? Are they also wildly inflated, as we have argued for two-species comparisons? No. If it is assumed that we are analyzing species' means as data points and that a statistical method allowing for nonindependence because of phylogenetic relationships is used, then Type I error rates will actually reflect the nominal tabular values (e.g., $\alpha = 0.05$).

We have just raised the issue of the appropriate statistical methods for analyzing multispecies, comparative data sets. The fundamental problem, with respect to hypothesis testing, is that species' mean values cannot be assumed to represent biologically or statistically independent data points (Felsenstein 1985; Harvey and Pagel 1991). Nonindependence arises because organisms descend in a hierarchical fashion from common ancestors and inherit many features from them. Thus, two species sharing a recent common ancestor will also probably share a number of features inherited from that ancestor, as compared with some other contemporary species to which they are less closely related (fig. 1).

Several procedures for dealing with the problem of statistical nonindependence of species' mean values have been proposed (reviews in Harvey and Pagel 1991; Miles and Dunham 1993; Losos and Miles 1994). Some have been developed to the point that their use is now fairly routine in comparative biology. Of these, Felsenstein's (1985) method of phy-

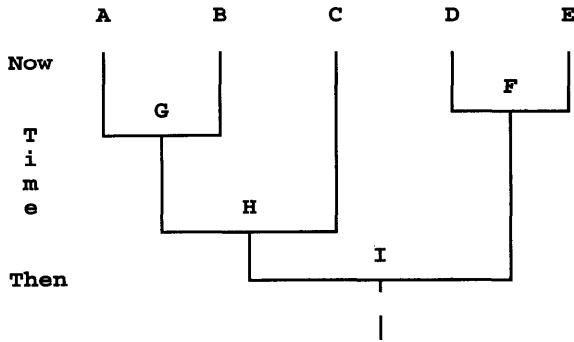


Fig. 1. Extant species A and B share a recent common ancestor (G) and, hence, are phylogenetically more closely related than either is to species C, D, or E. In addition, species A and B would probably share many features that they would have inherited from their common ancestor, G. Similar arguments apply to species D and E. In general, therefore, phenotypic mean values for species A, B, C, D, and E cannot be assumed to represent five independent data points in the statistical sense, and a correlation involving them could not properly be tested for significance with the nominal $N - 2 = 3$ df. Rather, the appropriate degrees of freedom available for hypothesis testing will, in effect, be something less than 3. The precise degree of nonindependence depends on the lengths of the branches in units of expected variance of change for the characters being studied, which may or may not be adequately represented by divergence times (see Felsenstein 1985; Grafen 1989; Harvey and Pagel 1991; Martins and Garland 1991; Garland, Harvey and Ives 1992).

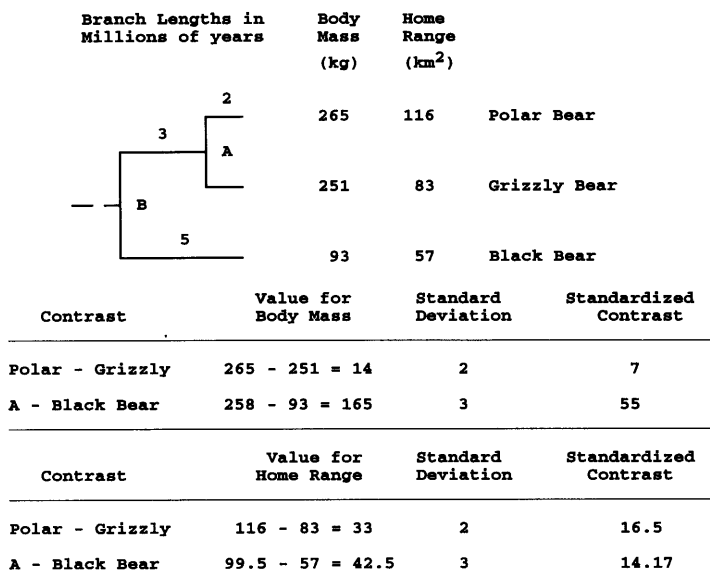
logenetically independent contrasts is the best understood and the best justified on statistical grounds (Grafen 1989; Losos 1990; Garland, Huey, and Bennett 1991; Martins and Garland 1991; Garland et al. 1992; Pagel 1992, 1993; Purvis and Garland 1993). Although originally developed to test for correlated evolution of continuously distributed characters, independent contrasts approaches are applicable to a wide range of statistical/evolutionary questions, which include rates of evolution and ANOVAs and ANCOVAs (Garland 1992; Garland et al. 1993; Martins 1993). In the next section, we offer a brief description of the phylogenetically independent contrasts approach and illustrate its application with an example. At least two articles published recently in *Physiological Zoology* have used this method (Promislow 1991; Sparti 1992). Computer programs for the PC that perform independent contrasts analyses are available from the senior author (Martins and Garland 1991; PDTREE program of Garland et al. 1993; see also Losos 1994).

Analyzing Comparative Data by Phylogenetically Independent Contrasts

The method of phylogenetically independent contrasts is based on the following logic (Felsenstein 1985). Although mean values for a series of hierarchically related species (see, e.g., fig. 1) may not be statistically independent, owing to inheritance from ancestors, these values can be transformed to be independent (at least with respect to inheritance from ancestors) by use of knowledge of the species' relationships to compute a series of trait differences (independent contrasts) between sister species or nodes. In the example shown in figure 2, the first contrast compares polar and grizzly bears, and the second compares the value at the node immediately ancestral to them (labeled "A") to the black bear. Because it is always the closest available relatives that are compared, the method of phylogenetically independent contrasts seems to satisfy the desire to compare "closely related" species, as has been suggested by many workers (see, e.g., Huey and Bennett 1987, 1990; Bennett and Huey 1990).

Once contrasts (differences between the phenotypes of sister species and/or nodes) are computed, branch lengths can then be used to weight all of the contrasts equally, which produces *standardized* independent contrasts that can be used in ordinary parametric statistical tests. Felsenstein's (1985) method is explicitly statistical and is based on a stochastic model of evolutionary change, that of Brownian motion. In this context, Brownian motion means simply that the probability of change for a given character is equally likely to be up or down at any point in the phylogeny. To implement the method, estimates of branch lengths in units of expected variance of change must be available for each character. As in any statistical analysis, assumptions of the method being used must always be checked, such as adequacy of branch lengths for standardizing contrasts, homogeneity of variance of standardized contrasts, and adequate behavior of residuals in a regression (see Grafen 1989; Harvey and Pagel 1991; Garland 1992; Garland et al. 1992; Pagel 1992).

Figure 3 illustrates both a phylogenetic and a nonphylogenetic analysis of morphology and running speed data for 14 species of *Anolis* lizards (taken from Losos 1990). The three left panels of figure 3 present the traditional nonphylogenetic analysis. The three right panels of figure 3 present independent contrasts applied (1) to estimate allometric scaling relationships of maximal sprint running speed and of hindleg length, (2) to remove the effects of body size from each trait by the computation of residuals from their respective allometric equations, and (3) to test whether residual independent contrasts in hindleg length predict variation in residual inde-



Pearson product-moment correlation of 3 tip values = 0.868

Correlation through the origin of 2 standardized contrasts = 0.742

Fig. 2. Computation of phylogenetically independent contrasts. A hypothesized phylogenetic tree for three species of bears is illustrated, with branch lengths in units of millions of years, as estimated from fossil information, and species' mean values for two characters (from Garland et al. 1993). Independent contrasts are computed as the phenotypic differences between sister species or nodes. With a Brownian motion model of character evolution assumed, these contrasts are independent in the statistical sense. The contrasts can be brought to common variance by dividing them by the square root of the sum of their branch lengths, which is their "standard deviation" (the branch length leading to node A is lengthened from three to four to account for the fact that the value at node A is estimated as opposed to measured data; hence, it should be devalued [see Felsenstein 1985]). Given a correct topology and branch lengths in units of expected variance of evolutionary change for the characters being analyzed, the standardized contrasts are independent and identically distributed, can be used in conventional parametric statistical tests, such as correlations and multiple regressions, and can provide estimates of evolutionary relationships, such as allometric scaling exponents (see, e.g., fig. 3). In this example, the correlation of standardized independent contrasts (which must always be computed through the origin) is somewhat lower than that for the original three tip data points. Such a difference occurs when the characters being analyzed tend to show phylogenetic resemblance, that is, sister taxa tend to be similar; here, the polar bear and grizzly bear are similar in body mass as compared to the black bear.

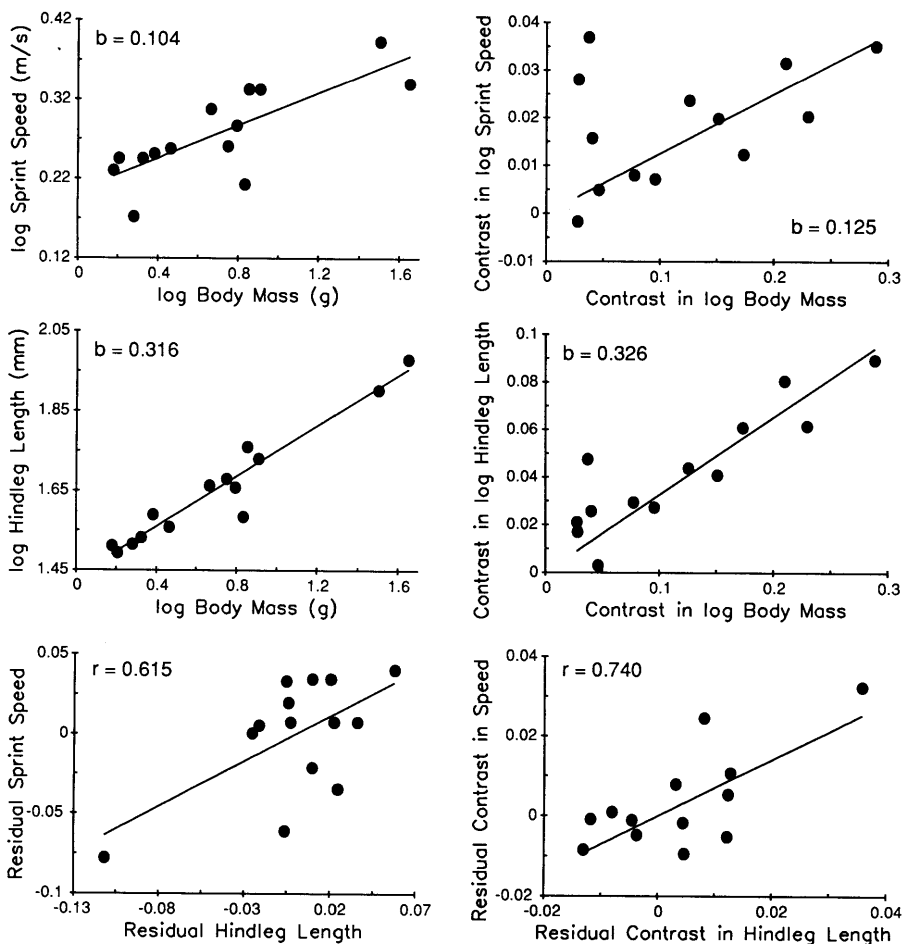
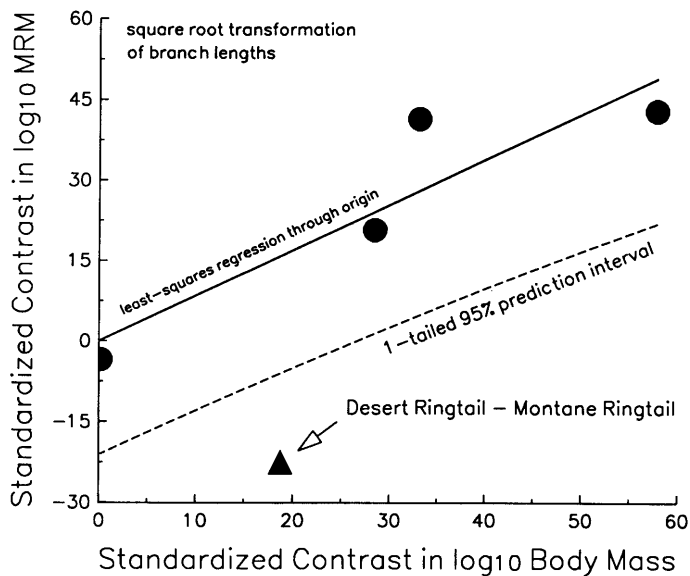
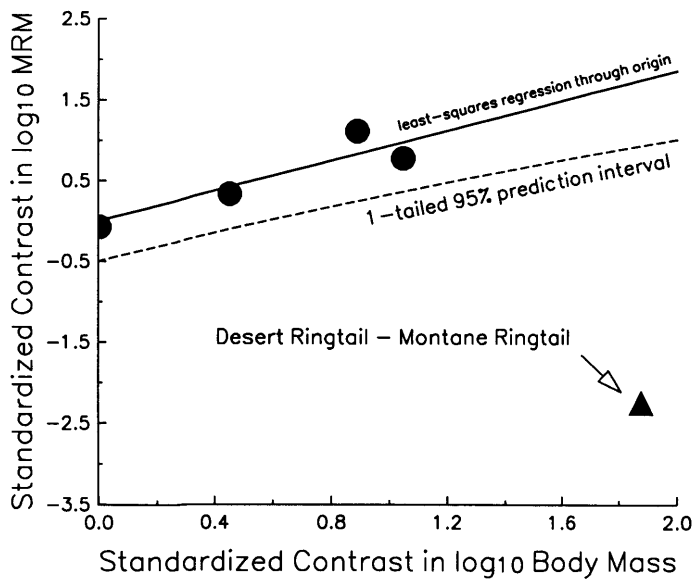
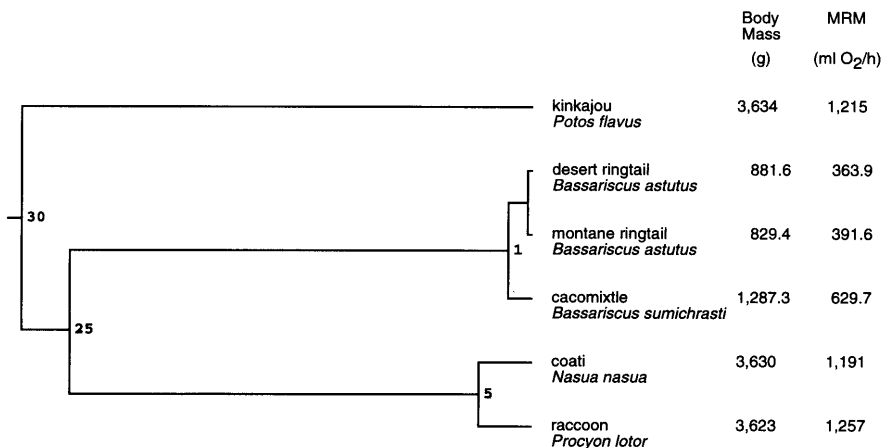


Fig. 3. Comparison of least squares linear regression and correlation applied to species' mean values (left three panels) and by means of these values transformed by the method of phylogenetically independent contrasts (right panels). Data on body mass, hindleg length, and maximal sprint running speed over 0.25 m on a photocell-timed racetrack for 14 species of *Anolis* lizards are from Losos (1990). The top left and center left panels depict, respectively, the log-log regressions of speed and of hindleg length on body mass, that is, conventional allometric plots and regression lines. The lower left panel depicts a positive relationship between the residuals of speed and the residuals of hindleg length. The top right and center right panels depict regressions of standardized independent contrasts in speed and in hindleg length regressed on body mass (all variables were log transformed prior to the computing of contrasts). The lower right panel depicts a significant positive relationship between the residuals of the independent contrasts of speed and hindleg length.

pendent contrasts in speed (for a similar example, see Garland and Janis 1993). In both cases, ordinary least squares linear regression analysis is employed, but with independent contrasts all regression lines are constrained to pass through the origin (see Garland et al. 1992). Because the 14 species' mean values probably are not statistically independent, *P* values from the conventional nonphylogenetic analyses cannot be trusted. On the other hand, *P* values from the independent contrasts analyses can generally be determined by reference to conventional statistical tables, with the assumption that the phylogeny used in computing contrasts (taken from Losos 1990) is accurate and that the branch lengths adequately standardize the independent contrasts (tests as described in Garland et al. 1992 indicate that the branch lengths presented in fig. 1 of Losos 1990 are adequate). In addition, independent contrasts estimates of the allometric scaling relationships can be shown to be superior to those from the conventional log-log plots (Martins and Garland 1991; Pagel 1993). In this example, nonphylogenetic and independent contrasts analyses yield similar estimates of the slopes and of the correlation between residuals, although the latter estimates are a little higher in all three cases. Thus, phylogenetically based statistical analyses do not always make relationships between characters appear less strong.

Another use of phylogenetically independent contrasts is illustrated in figure 4. Chevalier (1991) studied metabolism, thermoregulation, and evaporative water loss in single populations of five species and in two separate populations of a sixth species of procyonid mammals. A main question of interest to Chevalier was whether a desert population of ringtails had a lower than expected minimal resting metabolic rate in the thermal neutral zone (MRM). This raises two questions: (1) what is the "expected" MRM, and (2) how does one conduct a formal statistical test that allows for phylogenetic relationships? A traditional approach might have been to use conventional statistics to fit a 95% confidence interval to a log-log regression of MRM on body mass for a large series of mammalian species and to compare the focal species to this prediction. Concerned about comparing "apples and oranges," however (see next section), Chevalier instead measured several closely related species, which included a population of ringtails from a montane habitat.

One way to analyze his data phylogenetically is to ask whether the standardized independent contrast in MRM between the desert and montane ringtail populations is unusual as compared with the set of contrasts derived from the rest of the phylogenetic tree shown in figure 4. This can be done by regressing contrasts in MRM on contrasts in body mass, while omitting the one contrast of interest, and then fitting a 95% prediction interval for a



new observation (cf. Garland et al. 1993). We can then see whether the one contrast of interest falls outside of this interval, which it does (a one-tailed test is appropriate here because of the a priori alternative hypothesis that the desert population of ringtails would have a lower MRM). Note that the desert population is slightly larger in body mass than is the montane population, so the former is expected to also have a higher MRM; thus, a positive standardized contrast in body mass should be associated with a positive contrast in MRM, as is indicated by the solid regression line. Instead, the MRM of desert ringtails is lower, so the standardized contrast falls below the expected value, is actually negative, and is even below the one-tailed prediction interval.

Standardized independent contrasts represent estimated *rates* of evolution (Garland 1992), so we conclude that the divergence in MRM between these two ringtail populations occurred at a rate higher than usual for this clade (cf. Garland et al. 1993). One possible explanation for rapid evolutionary change in MRM (relative to the divergence in body mass) would be that past natural selection had led to a reduced MRM in the desert population; this is a common adaptive hypothesis (see, e.g., Hulbert and Dawson 1974; Nagy 1987; MacMillen and Garland 1989; Chevalier 1991). Because independent contrasts are nondirectional (Losos 1990; Harvey and Pagel 1991), however, another possibility is that the montane ringtail population has evolved a high MRM. In other words, figure 4 indicates that MRM evolved at a rate higher than expected (given the change in body mass) as these

*Fig. 4. Use of phylogenetically independent contrasts to test comparative hypotheses about deviation of a single species (of the ringtail *Bassariscus astutus*) from the pattern seen in related species. The top panel shows data (from Chevalier 1991) for body mass (g) and MRM (mL O₂/h), which are listed to the right of a cladogram for these six taxa (from Decker and Wozencraft's [1991] cladistic analysis of 129 morphological characters; branch lengths were estimated from fossil and biogeographic information). Numbers at the nodes of the cladogram indicate estimated divergence times in millions of years before present; the two ringtail populations probably began diverging about 10,000 yr ago, at the end of the last ice age. The middle panel plots standardized independent contrasts in MRM vs. body mass. The bottom panel plots standardized independent contrasts in MRM vs. body mass, with square-root transformed branch lengths. This transformation of branch lengths (cf. Garland et al. 1992) makes the contrast of interest less extreme, but it is still an outlier. (All values have been multiplied by 10,000.)*

desert and montane ringtail populations diverged, but it does not indicate whether the change occurred along the lineage leading to the desert population, along the lineage leading to the montane population, or by some combination of both. In the present case, however, squared-change parsimony reconstructions of ancestral values at the nodes, and independent biogeographic and paleontological evidence, suggest that most of the change in MRM actually did occur in the lineage leading to the desert ringtail populations (Chevalier 1991).

Statistical Power of Phylogenetic Comparative Methods

What constitutes an adequate number of species for correlating a trait with an environmental feature or for correlating two phenotypic traits? This can be viewed as a straightforward question of statistical power, or it can be viewed more subjectively. Some workers, for example, might find *any* study based on a small number of species to be unconvincing, whatever its results (see also section above, Enhancing Two-Species Comparisons through a Multivariate Approach).

Statistical power is defined as the probability that the null hypothesis will be rejected (e.g., no correlation between a trait and an environmental feature) when it is in fact false; in other words, the ability to detect a relationship when one exists. In statistical jargon, power is defined as one minus the Type II error rate, where the Type II error rate (denoted as β) is the probability of accepting the null hypothesis when it is false. It is ideal, according to many statisticians, that both the Type I error rate and the Type II error rate should be near 0.05; thus, power should be near 95% (e.g., see Peterman 1990). As has been pointed out many times in various forums, high statistical power typically does not receive the attention it should as compared with the strong emphasis placed on holding Type I error rates at $\alpha = 0.05$ (see, e.g., Peterman 1990; Thompson and Neill 1993; and references therein). Many experimental studies are conducted with Type I error rates held at 5% but with powers far less than 95%.

What is the power of phylogenetically based comparative studies? This general question has no general answer, because power depends on sample size, the strength of the relationship being studied, and the statistical test employed. When the statistical test also uses phylogenetic information, as in the method of phylogenetically independent contrasts described in the previous section, then this information too will affect power. An incorrect topology would cause the wrong species to be compared, and this could lead to a relationship's appearing either falsely strong or falsely weak. An-

other ever-present problem in phylogenetically based comparative analyses is that of errors in branch lengths. Systematic (i.e., nonrandom) errors in branch lengths can usually be detected and transformed away (Grafen 1989; Garland 1992; Garland et al. 1992). More or less random branch length errors should reduce statistical power but probably will not affect the accuracy of estimates of relationships (e.g., allometric slopes), although their effects have not been formally studied.

With the possibility of errors in the topology and branch lengths being used for analysis ignored, what is the power of the independent contrasts method for detecting an evolutionary correlation? It is, perhaps, surprising that it is exactly the same as for an ordinary Pearson product-moment correlation coefficient applied to nonphylogenetic data, irrespective of the details of the phylogeny. Figure 5 shows the power of using independent contrasts for detecting correlations of different magnitude and with different numbers of species. For studying allometric relationships, with correlations that often exceed 0.9 (see, e.g., fig. 3), a sample size of about seven species would be adequate to achieve a power of 95% for a one-tailed test. With an expected correlation of 0.75, about 13 species would be adequate. When correlations are 0.5 or lower, however, 30 or more species would be required. Given that branch lengths available for analyses will not be without error, these power figures should be viewed as upper bounds.

Although the phylogenetic relationships of species, with the assumption that they are correctly known, should not affect statistical power when independent contrasts analyses are used, the choice of species to be compared can affect power. For example, ecological physiologists often study species inhabiting extreme environments, or species known to display extreme values for some physiological trait, because they are more likely to show evidence of adaptation or to highlight some physiological principle (references in Feder et al. 1987; Burggren and Bemis 1990; Burggren 1991; Garland and Adolph 1991; Garland and Carter 1994). Including "extreme" species in a comparative study should generally increase the power to detect relationships, because it will, in effect, increase the range of the independent variable (e.g., the environmental factor, which is typically taken to indicate the presumed selective pressure).

If the extreme species is only distantly related to other species in the comparison, then one risks comparing "apples and oranges" (Huey and Bennett 1990, pp. 49–50). Physiologists, for example, have often compared almost any species with the "norm" defined by humans, domesticated Norway rats, or leopard frogs. We do not see such comparisons as very useful except for heuristic purposes. As noted by Carey (1993), even generalizations

about "the frog" are unlikely to prove very useful for predicting physiological characteristics of other amphibians. Some two-species comparisons have involved "control" species that are only distantly related to the species of interest. This is true of a comparison of a burrowing owl and a bobwhite, in which the authors did note that comparison with a nonburrowing owl of similar size would have been preferable (Boggs and Kilgore 1983), and of a comparison of a muskrat and a guinea pig that aimed to study diving adaptations in the muskrat heart (McKean et al. 1986). The latter example also suffers from the possibility that the control species might show adaptation to high altitude (also see Burggren 1991, pp. 6–7). In a similar way, comparisons of domesticated or laboratory organisms with others must be

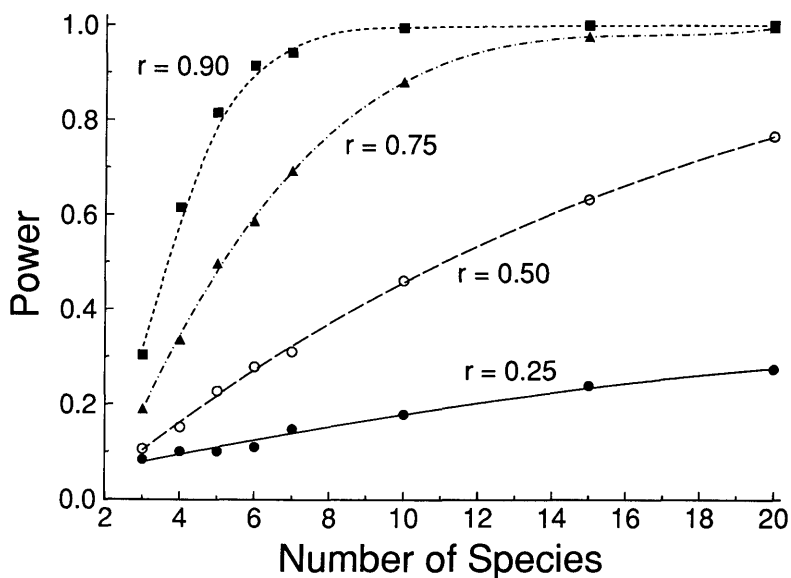


Fig. 5. Statistical power of phylogenetically independent contrasts (Felsenstein 1985) for detecting correlations of different magnitude and with different numbers of species, with the assumption that there are no errors in phylogenetic information (see text). Values are the proportion of 1,000 simulations with specified correlation (r) that exceeded the conventional critical value (from Zar's [1984] table B.16; one-tailed test, $\alpha = 0.05$). These values were obtained by means of the PDSIMUL program of Garland et al. (1993) and the CMFLANAL program of Martins and Garland (1991), although they could also have been obtained by means of standard formulas for the power of Pearson product-moment correlation coefficients applied to nonphylogenetic data. Lines are for illustrative purposes and represent nonlinear regression equations.

mindful that the former will likely show adaptations to their artificial environments.

In any case, for a preliminary comparative study, it may be useful to include species (and populations; see next section) of close, intermediate, *and* distant relationships, and to check for possible extreme contrasts (e.g., statistical outliers) that may heavily influence an overall relationship and possibly signal a point in phylogeny at which a relationship changed (cf. fig. 4; Huey 1987; Burggren 1991; figs. 4c and 6b of Garland and Janis 1993). Researchers should also avoid confounding the environmental factors presumed to lead to adaptation with phylogeny, as when all members of one ecological or behavioral category to be compared belong to one clade, whereas all members of the other category belong to a separate clade (see, e.g., Garland et al. 1993). Species within clades may not be statistically independent, and in this case become analogous to Hurlbert's (1984) pseudoreplicates.

Populations as Subjects of Comparative Studies

What happens within species can . . . be just as much part of comparative [physiology] as what is found between them. Yet, on the whole, the former has been more the concern of people interested in evolution, although they may call themselves ecological geneticists. The great value of comparisons based on different populations of a species is that any differences are far more readily related to the environments in which the populations occur, and are much less confounded by changes in characteristics acquired at some time past and of little relation to the present environment in which the species occur. . . . The differences found within species may not be as extreme as those found between species, which could be a disadvantage. But the specificity of their relationship to the environment occupied by the populations concerned is a great advantage.[BRADSHAW 1987b, pp. 14–15]

Populations within a species often show physiological differentiation (Garland and Adolph 1991; Lynch 1992), so they can serve as the subject of comparative studies just as can species, with advantages and disadvantages such as those noted in the foregoing quotation (see also Patton and Brylski 1987; James 1991; Adolph and Porter 1993; Malhotra and Thorpe 1993; *Behavior Genetics* 22:1–42). Indeed, a number of the single-species studies in *Physiological Zoology* compare two or more populations (see, e.g., Beaupre et al. 1993), and adaptive interpretations of population differences are common in many journals (see, e.g., Marken Lichtenbelt and Albers 1993). Two-population comparisons are, however, subject to the same lim-

itations discussed above for two-species comparisons. If three or more populations are studied, then it must be remembered that populations within species may often show some hierarchical genealogical structure. If so, then statistical methods incorporating this phylogenetic information are appropriate, although with some caveats (see Garland et al. 1992; Purvis and Garland 1993). One issue that needs more consideration is the appropriate null hypothesis (i.e., no difference vs. some difference) for physiological differentiation among populations within a species. Gene flow among populations, which is inherent to the definition of a species, will tend to counteract differentiation caused either by local adaptation or by genetic drift. When interpopulation gene flow is high, we might, therefore, expect little genetic or phenotypic (e.g., physiological) differentiation among populations. If gene flow is nearly ubiquitous among populations, then their reconstructed phylogenetic relationships will probably resemble a "star" with little or no phylogenetic structure (see, e.g., fig. 2 of Felsenstein 1985; Martins and Garland 1991; fig. 2 of Garland and Carter 1994). It is, therefore, worth noting that the independent contrasts approach gives the correct answer about character correlation, even if the phylogeny is a star or nearly so (see Martins and Garland 1991; Purvis and Garland 1993).

A particularly useful design for a multispecies comparative study would be to also sample two populations from each study species; this will allow inference about both microevolutionary and macroevolutionary processes (Garland et al. 1992). Note that studying only two populations from within any one species avoids any possible uncertainty about population phylogeny within that species (i.e., two populations can only be related as a simple bifurcation).

How to Deal with a Paucity of Phylogenetic Information

Many comparative biologists, having accepted the need to use phylogenetically based statistical methods, will be frustrated by a lack of information for their study organisms. Sometimes taxonomic information can be used to indicate a crude topology, but this must be done with extreme caution because most existing taxonomies are not cladistic even in intent. When available phylogenetic information is sparse or very weak, heuristic computer simulation approaches are possible (Losos 1994). Another approach is to compute independent contrasts only between those pairs or subsets of species for which one is quite confident about relationships (as mentioned by Felsenstein [1985]; see also Pagel [1992]). For example, in figure 1, we might be certain that species A and B are sisters and that species D and E are

sisters, but uncertain about the relationships of A-B, D-E, and C. We could be ultraconservative and compute only two contrasts, those between A and B and between D and E, but this would be throwing away information for species C and, hence, should lower statistical power. Alternatively, we could treat the relationships of A-B, D-E, and C as a “soft” polytomy, which would reflect our ignorance, and adjust degrees of freedom accordingly (see Purvis and Garland 1993). Possible sources of branch length information and examples are discussed elsewhere (Grafen 1989; Losos 1990; Martins and Garland 1991; Garland 1992; Garland et al. 1992; Pagel 1992, 1994; Sparti 1992; Garland et al. 1993; Martins 1993; Miles and Dunham 1993).

Alternative Views on Hypothesis Testing and Inference: Cladistics contra Statistics

We have argued that demonstrating an interspecific correlation between phenotype and environment requires adequate sample size and analyses incorporating phylogenetic information. The latter point leads us to briefly discuss some differences in logical framework found within the systematics community. These differences are difficult to describe in brief, except with some oversimplification. Nevertheless, we attempt to do so because we believe they have led to some confusion among those who wish to apply phylogenetically based methods, most of which are derived more or less directly from systematic biology, in their analyses of comparative data.

Some workers view reconstructing phylogenies as a problem of statistical inference, and statistical methods for inferring phylogenetic relationships rely on specified models of evolution (brief review in Felsenstein 1988). Others favor parsimony methods for deducing phylogenetic relationships, and the application of parsimony can be justified on purely methodological grounds without reference to any particular model of evolution (e.g., see Platnick and Funk [1983], esp. the article by Farris [1983]). Workers who typically deal with traits scored as discrete values (e.g., a simple presence-absence trait) tend to fall in the latter camp, whereas those who typically deal with traits scored on a continuous scale often take a more statistical perspective. Various exceptions exist; for example, many workers dealing with discrete molecular sequence data favor statistical approaches (see the journal *Molecular Biology and Evolution*). In any case, the foregoing dichotomy has, to some extent, been carried over into the world of contemporary “comparative methods” (compare Brooks and McLennan 1991 with chaps. 4 and 5 of Harvey and Pagel 1991). The following discussion is intended to help clarify some issues that may hinder communication between

the two perspectives, with special emphasis on the independent contrasts approach discussed above.

All, including the present authors, would agree that the identification of *multiple*, independent examples of convergent evolution in response to similar putative selective regimes provides some evidence of adaptation (Miles and Dunham 1993; Leroi et al. 1994). The greater the number of *independent* examples of convergent evolution, the stronger the evidence. When traits and environments scored as, for example, 0 or 1, to indicate presence-absence or high-low, are dealt with, a maximum-parsimony reconstruction of evolution along a cladogram will often indicate that only a few of the branches witnessed change. If only those branches inferred to have changed are used as evidence, then the apparent number of independent instances of possibly coincident change will often be small (see examples in Brooks and McLennan 1991), and some statistical methods for detecting evolutionary associations between dichotomous characters consider only those branches along which change is inferred to have occurred (reviews in Harvey and Pagel [1991]; Maddison and Maddison [1992]; Pagel [1994] has proposed a more general method for detecting correlated evolution of discrete characters). In the limit, a parsimony reconstruction of character and environment evolution might indicate only a single, possibly coincident, change in each (cf. Baum and Larson 1991). Inferring anything from this evidence, which suggests a single historical event, could be criticized for many of the same reasons that we have criticized the drawing of inferences from comparisons of two species (see also Leroi et al. 1994).

With continuous-valued characters treated by independent contrasts, the situation is quite different. Rarely will two species be scored as having *exactly* the same mean value for the character(s) or environmental feature(s) of interest. Thus, most contrasts will have a nonzero value, and each is viewed as providing *independent* evidence about the relationship. Even if a contrast does compute as zero for the character and/or the environmental feature, it is still included in assessing the comparative relationship, and, thus, still provides independent evidence—and a degree of freedom for hypothesis testing (see Garland et al. 1992). It is typical that a regression through the origin is fitted to test the statistical significance of the relationship (see, e.g., fig. 3). Another, perhaps more heuristic, way to view this is to imagine *each* independent contrast with a line connecting it to the origin. The direction (away from the origin) and magnitude of each such line defines a vector indicating the form of the comparative relationship *within that particular two-species subclade* of the overall phylogenetic tree. The statistical significance of the overall relationship is judged, in effect, by summing all of these statistically independent vectors and by determining the direction of

the net vector and whether its magnitude differs significantly from zero. Contrasts indicating no change in either character will be represented by a vector of zero length and no direction, that is, a point exactly at the origin, and, hence, will not affect the estimate of the relationship between the two traits (remember that all correlations or lines fitted to independent contrasts are constrained to pass through the origin [Garland et al. 1992]). Nevertheless, all $N - 1$ contrasts, regardless of their value, are taken as providing independent pieces of evidence about the relationship being studied (cf. Garland 1992).

Some workers favor coding continuously distributed characters as a limited number of discrete values (see Brooks and McLennan [1991, pp. 155–156, 364–366] for some illustrative examples). Reasons for such treatment may be philosophical and/or practical, for example, to allow their use in those parsimony (character optimization) algorithms that require discrete coding of characters. (Others might simply refuse to use continuously distributed characters, or even discrete characters showing polymorphism within species, in any kind of historical phylogenetic analysis.) Perhaps because of our backgrounds in physiological ecology and quantitative genetics, such categorization of what are inherently quantitative data (e.g., data for a trait that shows continuous variation within species, that is polygenic, and that shows essentially continuous variation among species) has always seemed a little perverse (see also Felsenstein 1988). Genetic drift alone should ensure that no two species or populations ever (except instantaneously) possess *exactly* the same mean for any given character. Thus, we would generally favor using the best available empirical estimate of each species' (or population's) mean value, rather than artificially lumping them into a smaller number of categories.

Optimization algorithms for tracing the evolution of continuous-valued characters are available (see Huey 1987; Huey and Bennett 1987; Huey 1989; Losos 1990; Harvey and Pagel 1991; Martins and Garland 1991; Maddison and Maddison 1992) and have even been implemented in phylogenetically based statistical methods for hypothesis testing (see, e.g., the squared-change parsimony in Martins and Garland 1991). Such algorithms can provide estimates for character values of hypothetical ancestors (interior nodes on a phylogenetic tree), and the squared-change parsimony algorithm is justified under a Brownian motion model of evolution. The performance of the squared-change parsimony algorithm when Brownian motion does not hold has not been studied in any formal way, although informal "sensitivity analyses" have been presented (Huey and Bennett 1987; Chevalier 1991). Felsenstein's (1985) independent contrasts method also uses estimates of interior nodes as an intermediate part of the computations, but

these are *not* intended as optimal reconstructions in the sense of parsimony criteria. The one exception occurs at the basal node (root) of a phylogeny, where the estimate from independent contrasts is identical to that from squared-change parsimony (see Garland et al. 1993).

Another important difference in perspective relates to the appropriateness of formal *statistical* hypothesis testing with comparative data. Felsenstein's (1985) method for continuously distributed characters is explicitly statistical and is based on a stochastic model of evolutionary change; "the objective is to find out how characters have evolved" (Felsenstein 1988, p. 124), for example, whether they have evolved in a correlated fashion (see also Martins and Garland 1991; Pagel 1993). In contrast, the parsimony procedures typically applied to reconstruct the evolution of discrete characters are *not* based on clear assumptions about an underlying evolutionary model (e.g., see Farris 1983; Maddison and Maddison 1992). Rather, many proponents justify their use as methodological rules for producing the "best" (in the sense of being most parsimonious) summaries of character distributions (e.g., cladograms). Although parsimony procedures are preferred by many workers (e.g., see Platnick and Funk 1983; Brooks and McLennan 1991), they do not readily lend themselves to statistical hypothesis testing (see, e.g., discussions in Harvey and Pagel 1991; Maddison and Maddison 1992; Pagel 1994). We would argue that, because parsimony reconstructions are only estimates (in the colloquial sense) of what happened, the evidence they may provide for or against an evolutionary association would benefit from a statistical framework (see also Felsenstein 1988). Whether this will become possible with parsimony methods per se is unclear, but alternatives for statistical hypothesis testing with discrete-valued characters are now becoming available (Maddison and Maddison 1992; Pagel 1994 and references therein).

In summary, we believe that statistical hypothesis testing can add rigor to scientific inference generally and to comparative studies of the type discussed herein and in two recent books (Brooks and McLennan 1991; Harvey and Pagel 1991). We are not dismayed that phylogenetically based statistical methods rely on assumptions, as do all methods of inference, statistical or not. The null "model" of evolution assumed by a phylogenetically based statistical method is rightly seen as including the topology, branch lengths, *and* model of character change per se (e.g., Brownian motion occurring independently in each of two characters or in a character and an environmental feature). All of these components are used in generating the null distribution for hypothesis testing. If, for example, an observed pattern of character-environment covariation is judged to be highly improbable, given the assumed model, then we logically reject one or more assumptions of

the null model. Usually, we would be most interested in whether to reject the assumption of *independent* character-environment change while accepting all other assumptions. (Note that natural selection leading to adaptation is but one factor that may cause correlated character-environment change [Martins and Garland 1991; Leroi et al. 1994].) If those other assumptions are clearly stated, they can often be tested. If they cannot be tested, then at least two options are available.

First, as with any statistical method, one can study the sensitivity of a phylogenetically based statistical method to departures from its underlying assumptions. Preliminary studies are available concerning the sensitivity of independent contrasts to errors in branch lengths (Grafen 1989; Martins and Garland 1991; Garland et al. 1992; Martins 1993) and to errors in topology (Grafen 1989; Purvis and Garland 1993; Losos 1994). It is hoped that some phylogenetically based statistical methods will prove to be robust with respect to a wide range of biologically reasonable violations of their assumptions. Second, hypothesis testing can be accomplished by computer simulation methods that, while relying on specified models of character change, are not limited to assuming any *particular* model (Martins and Garland 1991; Garland et al. 1993; see also Pagel 1994). If a range of evolutionary models all lead to the same conclusions regarding empirical data, then those conclusions are strengthened (e.g., see Garland et al. 1993; Martins 1993; Losos 1994). It would also be possible to develop comparative methods that formally consider uncertainty in the phylogenetic information while testing hypotheses about correlated character evolution (Felsenstein 1985, p. 13; Pagel 1994), rather than relying on a single specified phylogeny (see also Losos 1994).

Concluding Remarks

Most biologists would acknowledge that comparative studies have contributed much to the identification of putative evolutionary adaptations and to understanding their physiological bases. With respect to the former, comparative studies function by revealing correlations between phenotype and environment. Although comparisons involving only two species (or only two populations) are common, we have argued on both statistical and evolutionary grounds that they are, by themselves, inadequate for inferring adaptation (see also Gans [1989, pp. 634–635] on “the folly of ‘the two species comparison’”). Multispecies comparisons that attempt to correlate phenotypic with environmental variation eliminate the statistical and logical problems associated with comparing just two entities. Alternatively, one

might conduct a meta-analysis (cf. Gurevitch et al. 1992) of existing two-species comparisons pertaining to a particular hypothesized trait-environment correlation (e.g., low resting metabolic rates in species inhabiting arid habitats).

Our article should not be taken as a blanket endorsement of using multispecies comparative studies alone for inferring adaptation, because we certainly acknowledge the cautionary messages of Leroi et al. (1994). Many other types of studies can and should be employed when the goal is to elucidate evolutionary adaptation (see Reeve and Sherman 1993; Leroi et al. 1994). One example involves direct measurements of natural selection acting in extant populations (Lande and Arnold 1983; Endler 1986). This approach has recently been applied to measures of locomotor performance in reptiles, and it could be enhanced by experimental manipulation of the phenotype of individuals (references in Bennett and Huey 1990; Bennett 1994; Garland and Carter 1994; Garland and Losos 1994).

We suspect that many instances of claimed adaptive differences between two species really are so, partly because of the tendency for researchers to choose species that occur in very different (extreme) environments (see also Garland and Adolph 1991, p. 215). Studies comparing pairs of species from similar or only moderately different environments seem underrepresented in the physiological literature, presumably because of the anticipated lower likelihood of discovering a major difference. Thus, the prevalence of among-species adaptive differences cannot presently be judged, because the database is biased. Multispecies comparisons help to dilute the effect of biases in the selection of species and environments for comparative studies, and also allow for the possibility that the rank correlation between physiology and environment may not be perfect.

Multispecies comparative data must be analyzed with methods that make explicit use of phylogenetic information (Brooks and McLennan 1991; Harvey and Pagel 1991; Miles and Dunham 1993). If one is to employ a statistical framework for hypothesis testing and/or estimation (e.g., of allometric relationships; fig. 3), then Felsenstein's (1985) method of independent contrasts is probably the best currently available starting point. The alternative to employing a phylogenetically based statistical method is to employ conventional statistical methods, but doing so is tantamount to assuming no hierarchical relationships among species (i.e., that the phylogeny is a star) *and* that the characters have evolved by Brownian motion at equal rates in all lineages. At least the former of these assumptions is known to be false for most collections of species that might be studied. Thus, even incomplete phylogenetic information can be used to advantage (see, e.g., Purvis and Garland 1993; Losos 1994). Of course, as new and improved

hypotheses about phylogenetic relationships become available, reanalyses are in order and conclusions may change (see, e.g., Huey and Bennett 1987; Garland et al. 1991; references in Miles and Dunham 1993).

The independent contrasts approach was originally intended for analyzing phenotypic characteristics that differ genetically among species or populations (e.g., an estimate of the mean body mass for a series of species). The most common use of independent contrasts is in testing for correlated evolution of phenotypic characters (Losos 1990 [see our fig. 3], 1994; Garland et al. 1991; Martins and Garland 1991; Promislow 1991; Sparti 1992; Garland and Janis 1993; Martins 1993; Pagel 1993; Purvis and Garland 1993). Whether it is appropriate to treat environmental features (e.g., mean annual temperature) in the same fashion is not entirely clear, but Garland et al. (1992, pp. 29–30) discuss some of the rationale for doing so, with caveats.

What about using two-species (or two-population) comparisons for inferring physiological mechanisms underlying differences between species (see, e.g., Bennett et al. 1984; Canady, Kroodsmas, and Nottebohm 1984; Bailey, Sephton, and Driedzic 1991; Caldow and Furness 1993)? Suppose, for example, that two species are known to differ in maximum sprint speed, and we hypothesize (leading to a one-tailed test) that the faster species has muscles that can contract more rapidly (see, e.g., Abu-Ghalyun et al. 1988). Our same general point applies: species are *likely* to differ in any aspect of the phenotype we might choose to measure, which includes muscle contractile speed. Moreover, comparative studies supply only correlational data. Correlation does not demonstrate causation, although a known or hypothesized mechanism that could account for an observed correlation can strengthen the overall argument (e.g., faster-contracting muscles allow higher limb-cycling frequencies, which in turn allow higher sprint speeds), and the more precisely defined the model of causation the stronger the argument (see also Leroi et al. 1994). (Examining patterns of correlation among individuals within each of two species could also bolster or weaken arguments about a possible mechanistic link [references in Bennett 1994; Garland and Carter 1994; Garland and Losos 1994].)

A still more convincing case can be made by showing that all other possible causes of the physiological difference (e.g., leg length, locomotor muscle mass) do not differ between the two species. Rarely, however, is our understanding of the basis of variation in complex physiological traits so complete as to allow identification of *all* possible causes (e.g., on locomotor performance abilities, see Bennett et al. 1984; Abu-Ghalyun et al. 1988; Caldow and Furness 1993; Jones and Lindstedt 1993; Garland and Losos 1994). (In a similar way, environments are complex

and so differ in many ways other than those we might specify as possible causes for adaptive differences [Gans 1989, pp. 634–635; Leroi et al. 1994].) An extensive discussion of what one can and cannot infer about mechanism on the basis of comparisons of two artificially selected lines is found in a series of articles in the journal *Behavior Genetics* (vol. 22 [1992], pp. 1–42).

Acknowledgments

We thank R. B. Huey for motivation, many helpful discussions, and bringing references to our attention. C. D. Chevalier kindly provided a copy of his unpublished Ph.D. dissertation. R. J. Chappell, R. Diaz-Uriarte, C. S. Richardson, M. L. Zelditch, and various other colleagues provided helpful discussions. G. V. Lauder, C. Martinez del Rio, and D. L. Stern reviewed the manuscript, and the latter two suggested the multivariate approach for enhancing two-species comparisons. We recognize that not all of the foregoing may agree with all of the opinions expressed in the final version. T.G. was supported by National Science Foundation grants IBN-9111185, DEB-9157268, and DEB-9220872; S.C.A. was supported partly by the U.S. Department of Energy, Office of Health and Environmental Research, contract DE-FG02-88ER60633 to W. P. Porter.

Literature Cited

- ABU-GHALYUN, Y., L. GREENWALD, T. E. HETHERINGTON, and A. S. GAUNT. 1988. The physiological basis of slow locomotion in chamaeleons. *J. Exp. Zool.* 245:225–231.
- ADOLPH, S. C., and W. P. PORTER. 1993. Temperature, activity, and lizard life histories. *Am. Nat.* 142:273–295.
- BAILEY, J., D. SEPTON, and W. R. DRIEDZIC. 1991. Impact of an acute temperature change on performance and metabolism of pickerel (*Esox niger*) and eel (*Anguilla rostrata*) hearts. *Physiol. Zool.* 64:697–716.
- BARTHOLOMEW, G. A. 1987. Interspecific comparison as a tool for ecological physiologists. Pages 11–37 in M. E. FEDER, A. F. BENNETT, W. W. BURGGREN, and R. B. HUEY, eds. *New directions in ecological physiology*. Cambridge University Press, Cambridge.
- BAUM, D. A., and A. LARSON. 1991. Adaptation reviewed: a phylogenetic methodology for studying character macroevolution. *Syst. Zool.* 40:1–18.
- BEAUPRE, S. J., A. E. DUNHAM, and K. L. OVERALL. 1993. Metabolism of a desert lizard: the effects of mass, sex, population of origin, temperature, time of day, and feeding on oxygen consumption of *Sceloporus merriami*. *Physiol. Zool.* 66:128–147.

- BENNETT, A. F. 1994. Adaptation and the evolution of physiological characters. In W. H. DANTZLER, ed. *Handbook of comparative physiology*. Oxford University Press, Oxford (in press).
- BENNETT, A. F., and R. B. HUEY. 1990. Studying the evolution of physiological performance. Pages 251–284 in D. J. FUTUYMA and J. ANTONOVICS, eds. *Oxford surveys in evolutionary biology*. Vol. 7. Oxford University Press, Oxford.
- BENNETT, A. F., R. B. HUEY, and H. B. JOHN-ALDER. 1984. Physiological correlates of natural activity and locomotor capacity in two species of lacertid lizards. *J. Comp. Physiol.* 154B:113–118.
- BOGGS, D. F., and D. L. KILGORE, JR. 1983. Ventilatory responses of the burrowing owl and bobwhite to hypercarbia and hypoxia. *J. Comp. Physiol.* 149B:527–533.
- BRADSHAW, A. D. 1987a. Functional ecology = comparative ecology? *Funct. Ecol.* 1: 71.
- . 1987b. Comparison: its scope and limits. *New Phytol.* 106 (Supp.): 3–21.
- BROOKS, D. R., and D. A. MCLENNAN. 1991. *Phylogeny, ecology, and behavior. A research program in comparative biology*. University of Chicago Press, Chicago. 434 pp.
- BURGGREN, W. W. 1991. Does comparative respiratory physiology have a role in evolutionary biology (and vice versa)? Pages 1–14 in A. J. WORLIES, M. K. GRIESHABER, and C. L. BRIDGES, eds. *Physiological strategies for gas exchange and metabolism*. Cambridge University Press, Cambridge.
- BURGGREN, W. W., and W. E. BEMIS. 1990. Studying physiological evolution: paradigms and pitfalls. Pages 191–238 in M. H. NITECKI, ed. *Evolutionary innovations*. University of Chicago Press, Chicago.
- CALDOW, R. W. G., and R. W. FURNESS. 1993. A histochemical comparison of fibre types in the M. pectoralis and M. supracoracoideus of the great skua *Catharacta skua* and the herring gull *Larus argentatus* with reference to kleptoparasitic capabilities. *J. Zool. Lond.* 229:91–103.
- CANADY, R. A., D. E. KROODSMA, and F. NOTTEBOHM. 1984. Population differences in complexity of a learned skill are correlated with brain space involved. *Proc. Natl. Acad. Sci. USA* 81:6232–6234.
- CAREY, C. 1993. Vertebrate adaptations. *Science* 259:390–391.
- CHEVALIER, C. D. 1991. Aspects of thermoregulation and energetics in the Procyonidae (Mammalia: Carnivora). Ph.D. thesis. University of California, Irvine. 202 pp.
- COHAN, F. M., and A. A. HOFFMANN. 1989. Uniform selection as a diversifying force in evolution: evidence from *Drosophila*. *Am. Nat.* 134:613–637.
- DECKER, D. M., and W. C. WOZENCRAFT. 1991. Phylogenetic analysis of recent procyonid genera. *J. Mammal.* 72:42–55.
- DOHM, M. R., and T. GARLAND, JR. 1993. Quantitative genetics of scale counts in the garter snake *Thamnophis sirtalis*. *Copeia* 1993:987–1002.
- ENDLER, J. A. 1986. *Natural selection in the wild*. Princeton University Press, Princeton, N. J. 336 pp.
- FARACI, F. M., D. L. KILGORE, JR., and M. R. FEDDE. 1984. Attenuated pulmonary pressor response to hypoxia in bar-headed geese. *Am. J. Physiol.* 247 (Regul. Integrative Comp. Physiol. 16): R402–R403.
- FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 in N. I. PLATNICK and V. A. FUNK, eds. *Advances in cladistics*. Vol. 2. Proceedings of the

- Second Meeting of the Willi Hennig Society. Columbia University Press, New York.
- FEDER, M. E., A. F. BENNETT, W. W. BURGGREN, and R. B. HUEY, eds. 1987. New directions in ecological physiology. Cambridge University Press, New York. 364 pp.
- FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *Am. Nat.* 126:1–25.
- . 1988. The detection of phylogeny. Pages 112–127 in D. HAWKSWORTH, ed. Prospects in systematics. Systematics Association, Clarendon, Oxford. 457 pp.
- GANS, C. 1989. Morphology, today and tomorrow. Pages 631–637 in H. SPLECHTNA and H. HILGERS, eds. Trends in vertebrate morphology. Proceedings of the Second International Symposium on Vertebrate Morphology, Vienna, 1986. *Fortschritte der Zoologie*. Vol. 35. Gustav Fischer, Stuttgart.
- GARLAND, T., JR. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am. Nat.* 140:509–519.
- GARLAND, T., JR., and S. C. ADOLPH. 1991. Physiological differentiation of vertebrate populations. *Annu. Rev. Ecol. Syst.* 22:193–228.
- GARLAND, T., JR., and P. A. CARTER. 1994. Evolutionary physiology. *Annu. Rev. Physiol.* 56:579–621.
- GARLAND, T., JR., A. W. DICKERMAN, C. M. JANIS, and J. A. JONES. 1993. Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.* 42:265–292.
- GARLAND, T., JR., P. H. HARVEY, and A. R. IVES. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41:18–32.
- GARLAND, T., JR., R. B. HUEY, and A. F. BENNETT. 1991. Phylogeny and thermal physiology in lizards: a reanalysis. *Evolution* 45:1969–1975.
- GARLAND, T., JR., and C. M. JANIS. 1993. Does metatarsal/femur ratio predict maximal running speed in cursorial mammals? *J. Zool. Lond.* 229:133–151.
- GARLAND, T., JR., and J. B. LOSOS. 1994. Ecological morphology of locomotor performance in squamate reptiles. Pages 240–302 in P. C. WAINWRIGHT and S. M. REILLY, eds. Ecological morphology: integrative organismal biology. University of Chicago Press, Chicago.
- GRAFEN, A. 1989. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. Biol.* 326:119–157.
- GUREVITCH, J., L. L. MORROW, A. WALLACE, and J. S. WALSH. 1992. A meta-analysis of competition in field experiments. *Am. Nat.* 140:539–572.
- HARVEY, P. H., and M. D. PAGEL. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford. 239 pp.
- HILL, W. G., and A. CABALLERO. 1992. Artificial selection experiments. *Annu. Rev. Ecol. Syst.* 23:287–310.
- HINSLEY, S. A., P. N. FERNS, D. A. THOMAS, and B. PINSHOW. 1993. Black-tailed sandgrouse (*Pterocles orientalis*) and pin-tailed sandgrouse (*Pterocles alchata*): closely related species with differing bioenergetic adaptations to arid zones. *Physiol. Zool.* 66:20–42.
- HOCHACHKA, P. W., and G. N. SOMERO. 1984. Biochemical adaptation. Princeton University Press, Princeton, N. J. 537 pp.
- HUEY, R. B. 1987. Phylogeny, history, and the comparative method. Pages 76–98 in M. E. FEDER, A. F. BENNETT, W. W. BURGGREN, and R. B. HUEY, eds. New directions in ecological physiology. Cambridge University Press, New York.

- HUEY, R. B., and A. F. BENNETT. 1987. Phylogenetic studies of coadaptation: preferred temperatures versus optimal performance temperatures of lizards. *Evolution* 41: 1098–1115.
- . 1990. Physiological adjustments to fluctuating thermal environments: an ecological and evolutionary perspective. Pages 37–59 in R. MORIMOTO and A. TISSIERES, eds. *Stress proteins in biology and medicine*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- HULBERT, A. J., and T. J. DAWSON. 1974. Standard metabolism and body temperature of perameloid marsupials from different environments. *Comp. Biochem. Physiol.* 47A:583–590.
- HURLBERT, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54:187–211.
- JAMES, F. C. 1991. Complementary description and experimental studies of clinal variation in birds. *Am. Zool.* 31:694–705.
- JONES, J. H., and S. L. LINDSTEDT. 1993. Limits to maximal performance. *Annu. Rev. Physiol.* 55:547–569.
- KELLOGG, E. A., and H. B. SHAFFER. 1993. Model organisms in evolutionary studies. *Syst. Biol.* 42:409–414.
- LANDE, R., and S. J. ARNOLD. 1983. The measurement of selection on correlated characters. *Evolution* 37:1210–1226.
- 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. 1990. Ecomorphology, performance capability, and scaling of West Indian *Anolis* lizards: an evolutionary analysis. *Ecol. Monogr.* 60:369–388.
- . 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Syst. Biol.* 43:117–123.
- LOSOS, J. B., and D. B. MILES. 1994. Adaptation, constraint, and the comparative method: phylogenetic issues and methods. Pages 60–98 in P. C. WAINWRIGHT and S. M. REILLY, eds. *Ecological morphology: integrative organismal biology*. University of Chicago Press, Chicago.
- LYNCH, C. B. 1992. Clinal variation in cold adaptation in *Mus domesticus*: verification of predictions from laboratory populations. *Am. Nat.* 139:1219–1236.
- McKEAN, T., D. SCHMIDT, J. M. TINGEY, D. WINGERTSON, and L. W. SEEB. 1986. The use of glucose, lactate, pyruvate, and palmitic acid by muskrat and guinea pig hearts. *Physiol. Zool.* 59:283–292.
- MACMILLEN, R. E., and T. GARLAND, JR. 1989. Adaptive physiology. Pages 143–168 in J. N. LANE and G. L. KIRKLAND, JR., eds. *Advances in the study of Peromyscus* (Rodentia). Texas Tech University Press, Lubbock.
- MADDISON, W. P., and D. R. MADDISON. 1992. *MacClade: analysis of phylogeny and character evolution*. Version 3. Sinauer, Sunderland, Mass. 398 pp.
- MALHOTRA, A., and R. S. THORPE. 1993. An experimental field study of a eurytopic anole, *Anolis oculatus*. *J. Zool. Lond.* 229:163–170.
- MARKEN LICHTENBELT, W. D. VAN, and K. B. ALBERS. 1993. Reproductive adaptations of the green iguana on a semiarid island. *Copeia* 1993:790–798.
- MARTINS, E. P. 1993. A comparative study of the evolution of the *Sceloporus* push-up display. *Am. Nat.* 142:994–1018.
- MARTINS, E. P., and T. GARLAND, JR. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* 45:534–557.

- MILES, D. B., and A. E. DUNHAM. 1993. Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analyses. *Annu. Rev. Ecol. Syst.* 24:587-619.
- NAGY, K. A. 1987. Field metabolic rate and food requirement scaling in mammals and birds. *Ecol. Monogr.* 57:111-128.
- OTTE, D., and J. A. ENDLER, eds. 1989. Speciation and its consequences. Sinauer, Sunderland, Mass. 679 pp.
- PAGEL, M. D. 1992. A method for the analysis of comparative data. *J. Theor. Biol.* 156:431-442.
- . 1993. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. *J. Theor. Biol.* 164:191-205.
- . 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond.* 255B:37-45.
- PATTON, J. L., and P. V. BRYLSKI. 1987. Pocket gophers in alfalfa fields: causes and consequences of habitat-related body size variation. *Am. Nat.* 130:493-506.
- PETERMAN, R. M. 1990. Statistical power analysis can improve fisheries research and management. *Can. J. Fisheries Aquatic Sci.* 47:2-15.
- PLATNICK, N. I., and V. A. FUNK. 1983. Advances in cladistics. Vol. 2. Proceedings of the Second Meeting of the Willi Hennig Society. Columbia University Press, New York. 218 pp.
- PROMISLOW, D. E. L. 1991. The evolution of mammalian blood parameters: patterns and their interpretation. *Physiol. Zool.* 64:393-431.
- PURVIS, A., and T. GARLAND, JR. 1993. Polytomies in comparative analyses of continuous characters. *Syst. Biol.* 42:569-575.
- QUINLAN, M. C., and N. F. HADLEY. 1993. Gas exchange, ventilatory patterns, and water loss in two lubber grasshoppers: quantifying cuticular and respiratory transpiration. *Physiol. Zool.* 66:628-642.
- REEVE, H. K., and P. W. SHERMAN. 1993. Adaptation and the goals of evolutionary research. *Q. Rev. Biol.* 68:1-32.
- ROBERTSON, A., ed. 1980. Selection experiments in laboratory and domestic animals. Commonwealth Agricultural Bureau, Farnham Royal, Slough, U.K. 245 pp.
- ROHLF, F. J., and R. R. SOKAL. 1981. Statistical tables. 2d ed. W. H. Freeman, San Francisco. 219 pp.
- SINERVO, B., and S. C. ADOLPH. 1989. Thermal sensitivity of growth rate in hatchling *Sceloporus* lizards: environmental, behavioral and genetic aspects. *Oecologia* 78: 411-419.
- SOKAL, R. R., and F. J. ROHLF. 1981. Biometry. 2d ed. W. H. Freeman, San Francisco. 859 pp.
- SPARTI, A. 1992. Thermogenic capacity of shrews (Mammalia, Soricidae) and its relationship with basal rate of metabolism. *Physiol. Zool.* 65:77-96.
- THOMPSON, C. F., and A. J. NEILL. 1993. Statistical power and accepting the null hypothesis. *Anim. Behav.* 46:1012.
- ZAR, J. H. 1984. Biostatistical analysis. 2d ed. Prentice-Hall, Englewood Cliffs, N. J. 718 pp.