Evolution, 51(5), 1997, pp. 1623–1635

TESTING FOR UNEQUAL AMOUNTS OF EVOLUTION IN A CONTINUOUS CHARACTER ON DIFFERENT BRANCHES OF A PHYLOGENETIC TREE USING LINEAR AND SQUARED-CHANGE PARSIMONY: AN EXAMPLE USING LESSER ANTILLEAN *ANOLIS* LIZARDS

MARGUERITE A. BUTLER[1] AND JONATHAN B. LOSOS[2]
Department of Biology, Campus Box 1137, Washington University, Saint Louis, Missouri 63130
[1]*E-mail: butler@biodec.wustl.edu*
[2]*E-mail: losos@biodec.wustl.edu*

Abstract.—Although a large body of work investigating tests of correlated evolution of two continuous characters exists, hypotheses such as character displacement are really tests of whether substantial evolutionary change has occurred on a particular branch or branches of the phylogenetic tree. In this study, we present a methodology for testing such a hypothesis using ancestral character state reconstruction and simulation. Furthermore, we suggest how to investigate the robustness of the hypothesis test by varying the reconstruction methods or simulation parameters. As a case study, we tested a hypothesis of character displacement in body size of Caribbean *Anolis* lizards. We compared squared-change, weighted squared-change, and linear parsimony reconstruction methods, gradual Brownian motion and speciational models of evolution, and several resolution methods for linear parsimony. We used ancestor reconstruction methods to infer the amount of body size evolution, and tested whether evolutionary change in body size was greater on branches of the phylogenetic tree in which a transition from occupying a single-species island to a two-species island occurred. Simulations were used to generate null distributions of reconstructed body size change. The hypothesis of character displacement was tested using Wilcoxon Rank-Sums. When tested against simulated null distributions, all of the reconstruction methods resulted in more significant *P*-values than when standard statistical tables were used. These results confirm that *P*-values for tests using ancestor reconstruction methods should be assessed via simulation rather than from standard statistical tables. Linear parsimony can produce an infinite number of most parsimonious reconstructions in continuous characters. We present an example of assessing the robustness of our statistical test by exploring the sample space of possible resolutions. We compare ACCTRAN and DELTRAN resolutions of ambiguous character reconstructions in linear parsimony to the most and least conservative resolutions for our particular hypothesis.

Key words.—Ancestral reconstruction, Brownian motion, gradual evolution, phylogenetic comparative method, simulation, speciational evolution, statistical tests.

Received August 13, 1996. Accepted May 7, 1997.

In the past decade, a number of ancestral character state reconstruction methods have been developed for testing evolutionary hypotheses (Huey and Bennett 1987; Swofford and Maddison 1987; Maddison 1991; Martins and Garland 1991; Maddison and Maddison 1992). Previously, hypotheses that involved evolutionary change imbedded in the past history of a lineage were only accessible by examination of fossil evidence. Because fossil evidence is limited for most taxa and characters (e.g., coloration, most behavioral and physiological characters, and soft tissues), ancestral character state reconstruction methods have broadened the realm of possibilites for evolutionary investigations.

Application of these reconstruction methods has been hampered by two main difficulties. First, each of the available methods has its own assumptions, and we lack good criteria for choosing among the methods and their many variants. Second, statistical significance of tests using these methods is difficult to evaluate because reconstructed character data (inferred values at internal nodes of a phylogenetic tree) are not independent data points, and the appropriate number of degrees of freedom to use in such analyses is unclear.

These problems are particularly troubling for continuous characters for which a number of methods have been proposed to reconstruct ancestral states. The two most commonly used are linear parsimony (Swofford and Maddison 1987; Maddison and Maddison 1992), which minimizes the sum of the absolute value of the amount of change inferred on each branch of the phylogenetic tree; and squared-change parsimony and various modifications (Huey and Bennett 1987; Maddison 1991; Martins and Garland 1991), which minimize the sum of the square of the change on each branch of the phylogenetic tree.

Squared-change and linear parsimony can give substantially different reconstructions of character evolution because linear parsimony concentrates all the change on a few branches of the tree, whereas squared-change parsimony distributes evolutionary change among most or all branches. Most recent studies have employed squared-change parsimony for pragmatic reasons: it reconstructs a single value for each ancestral node (e.g., Losos 1990a; Garland et al. 1997). By contrast, linear parsimony often provides ambiguous reconstructions, in which the most parsimonious reconstruction of an ancestral node can lie anywhere within a range of values. Because the characters are continuous, and can take any value within the allowed range, an infinite number of most parsimonious character state sets is possible which may lead to very different evolutionary interpretations.

One solution to the statistical interpretation problem is to simulate character evolution on a phylogeny many times and use the results of these simulations to construct a null distribution of test statistics. Related approaches for testing the association between categorical variables are Maddison's

(1990) concentrated changes test, which permutes evolutionary transitions among the branches on the phylogenetic tree; and Sillén-Tullberg's (1988, 1993) contingent states test, which calculates the probability of association between two characters in a manner analogous to contingency tables. However, simulations have rarely been used in studies focusing on continuous variables (but see Martins and Garland 1991; Garland et al. 1993; Díaz-Uriarte and Garland 1996), and have never been applied in studies using linear parsimony to reconstruct ancestral character states, probably because of the many ambiguous reconstructions that are produced.

An additional level of uncertainty involves deciding how to conduct the simulations, because many parameters need to be specified in modeling character evolution. Garland et al. (1993) provided several simulation options, but to date, only one study (Díaz-Uriarte and Garland 1996) has investigated whether statistical tests are qualitatively affected by the use of different options.

We have three goals in this study. First, we investigate how the use of different ancestral reconstruction methods affect our hypothesis that character displacement has occurred. Second, we examine the extent to which different assumptions underlying the simulations alter the outcomes of the analyses. Last, we propose a means of using simulation methods to explore uncertainty resulting from ambiguous character reconstructions in studies using linear parsimony and to assess statistical confidence in the results.

Character Displacement in Caribbean Anolis Lizards

As an example, we reexamine a previous investigation of body size evolution in the *Anolis* lizards of the northern Lesser Antilles. Previous workers (Schoener 1970; Lazell 1972; Williams 1972) had suggested that differences in size among these species resulted from character displacement. That is, initially intermediate-sized species came into sympatry and evolved in opposite directions, producing large and small species. One of us (Losos 1990b) previously investigated this hypothesis by first using squared-change parsimony to reconstruct the evolution of body size, and then using the Mann-Whitney *U*-test to evaluate the support for character displacement. Three instances of transition from occupation of a one-species island to occupation of a two-species island were inferred along the phylogenetic tree using linear parsimony (Losos 1990b, Fig. 1). These three branches are hereafter referred to as "transition branches." The specific hypothesis tested was that evolutionary change in body size was greater on transition branches, compared with branches in which no such transition occurred. The result of this test was marginally significant ($P < 0.056$; Losos 1990b).

This analysis can be criticized on two counts. First, this analysis was conducted before the development of simulation methods for comparative analyses, and therefore used standard nonparametric tests counting each branch of the evolutionary tree as a data point. This is problematic because a phylogeny with N species was used to infer changes on $2N$-2 branches of the phylogenetic tree. Thus, 23 data points were used to estimate 44 changes, so that the degrees of freedom are inappropriately inflated (Huey and Bennett 1987; Martins and Hansen 1996). A related, but underappreciated, problem is that the values reconstructed for one node are not independent of values reconstructed for other nodes, which violates the assumptions of all parametric and nonparametric tests (Felsenstein 1985). The nonindependence will result in an an altered distribution of test statistics.

Second, linear parsimony methods were not used to reconstruct ancestral character states. Subsequently, Miles and Dunham (1996) reanalyzed these data using linear parsimony. Because many ancestral nodes are reconstructed with ambiguity, Miles and Dunham (1996) chose two of the possible equally parsimonious reconstructions using the ACCTRAN and DELTRAN algorithms (Swofford and Maddison 1987) to resolve nodes at which the inferred value was ambiguous. DELTRAN (delayed transformation) resolutions favor changes as late in the evolutionary tree as possible (resulting in more frequent instances of parallel evolution), whereas ACCTRAN (accelerated transformation) resolutions favor evolutionary change as early as possible (increasing the frequency of evolutionary reversal).

Miles and Dunham (1996) found that some reconstructions supported the hypothesis of character displacement, whereas others did not. However, just as in the previous analysis of Losos (1990b), the data were not independent and the degrees of freedom were overestimated. Furthermore, ACCTRAN and DELTRAN reconstructions were employed because it was presumed that they represent the extremes of the ambiguity in the character states relative to the hypothesis of character displacement. To assess whether different resolutions qualitatively affect the outcome of the analysis, however, one should examine the reconstructions that are most and least favorable to the hypothesis at hand, which may not be represented by ACCTRAN and DELTRAN (Swofford and Maddison 1987; Maddison and Maddison 1992). The most conservative resolution, for example, might include minimizing body-size evolution along transition branches (DELTRAN-like), and maximizing change elsewhere (ACCTRAN-like).

Materials and Methods

Overview

We reconstructed body size evolution in northern Lesser Antillean anoles using squared-change and linear parsimony. Figure 2 illustrates the general overview of the simulations and reconstructions used to test the character displacement hypothesis. The comparisons and various options are indicated by shaded boxes.

For each test comparison, we simulated character evolution on the phylogeny 500 times, producing values for each of the "tip species." Then for each simulation, we used weighted squared-change parsimony to reconstruct body size evolution simulated with a gradual model of evolution, and both squared-change and linear parsimony to reconstruct changes simulated with a speciational model of evolution (i.e., a Brownian-motion model with equal branch lengths, Fig. 2). The linear parsimony reconstructions resulted in ambiguities at many nodes. We resolved ambiguous nodes in four ways: by using the DELTRAN and ACCTRAN resolutions and by using Maximum-Transition and Minimum-Transition (customized resolutions that are potentially the most and least favorable to the character displacement hypothesis, as explained below, Fig.

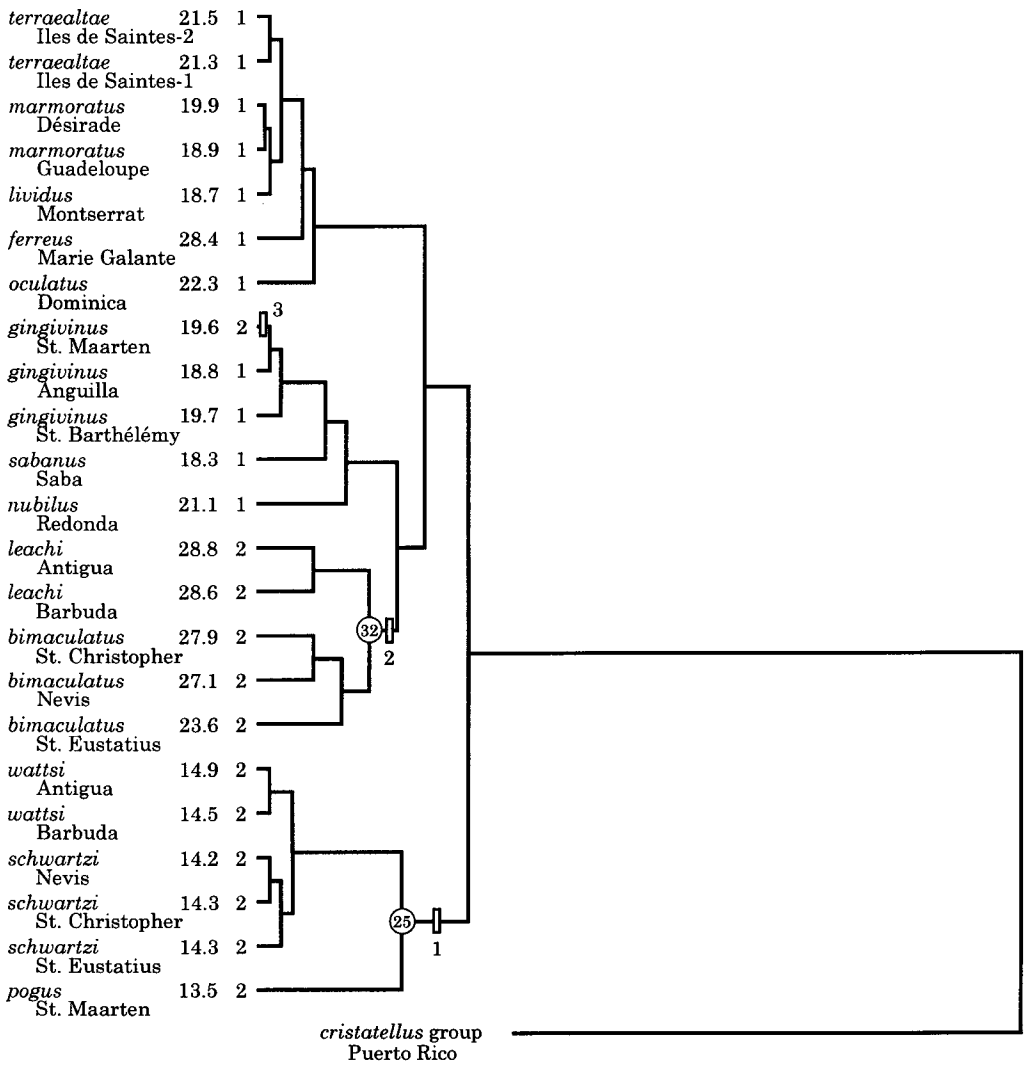

FIG. 1. The hypothesized phylogeny for lizards of the *Anolis bimaculatus* series of the northern Lesser Antilles (adapted from Losos 1990a, with additional branch length information from Hass et al. 1993). Ones and twos indicate which species occur on one- or two-species islands, respectively. The transition branches are indicated by open boxes and labeled 1–3, with descendant taxa for 1 and 2 labeled (node 25 and 32, respectively). The Body size values (mean jaw length in mm for the upper one-third of male specimens in mm) for the extant taxa are given. The *Anolis cristatellus* group of Puerto Rico was used as the outgroup. The actual extant values for the *cristatellus* group were used to reconstruct a single outgroup value to use in the simulations (17.7 mm for squared-change, 17.0 mm for weighted squared-change, and 19.2 mm for linear parsimony; see end of Methods section for rationale). The horizontal distances are drawn proportionally to the branch lengths used in the study (see text).

3). Tests using squared-change parsimony follow similar logic to that of Figure 3, except that resolution options are unnecessary because no ambiguous resolutions exist.

Once the reconstructions were completed, we conducted the hypothesis tests by focusing on the three transition branches. We used Wilcoxon Rank-Sums to test the hypothesis that the amount of evolutionary change was greater on the transition branches than the nontransition branches versus the null hypothesis that the amount of change did not differ between these two sets (rank test, Fig. 2). (Although tests based on the Mann-Whitney *U*-statistic and Wilcoxon Rank-Sums are logically equivalent, WRS are easier to compute; Hollander and Wolfe 1973.) Character displacement also restricts the direction of evolutionary change; if two species colonize an island, they cannot both increase or both decrease in size (direction test, Fig. 2). We incorporated this constraint into the overall hypothesis test by assigning a test statistic of zero to simulations which violated the direction test. To be conservative, reconstructions of zero change (i.e., both lineages remaining constant in body size, or only one lineage changing in body size) did not count against the direction test. Thus, for each reconstruction, we compared the test statistic from the analysis using real data with a distribution of test statistics produced from 500 simulations. The results of these analyses were considered significant if the observed test statistic was greater than the test statistics in 95% of the simulations.

Simulation Parameters

We generated simulations using the PDAP computer package (Garland et al. 1993). Both gradual and punctuational

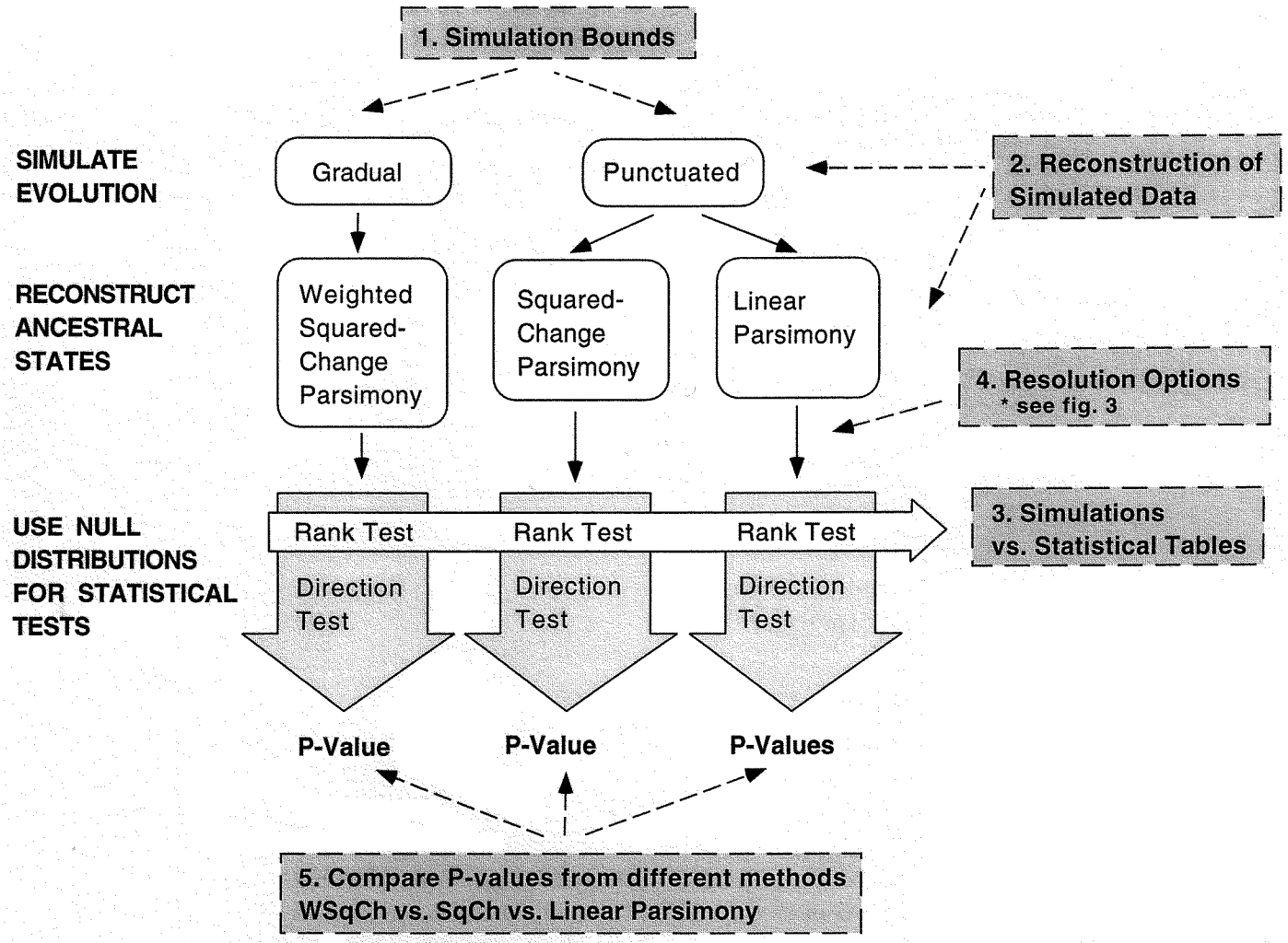

FIG. 2. A flowchart of the study. Observed values for amounts of body size change at internal branches were obtained by applying ancestral character state reconstruction methods (not included in diagram). A null distribution for the hypothesis test was constructed by simulating evolution to obtain body size data for the terminal taxa, reconstructing ancestral character states based on the simulated values, comparing observed scores for the ranks and signs of the reconstructed values, and comparing observed scores to the distribution of simulated scores. The shaded boxes represent variations on the basic protocol to investigate the effects of simulation parameters and reconstruction methods on statistical tests (steps 1, 4, 5), and to investigate differences between testing reconstructions against statistical tables versus simulated null distributions (step 2, 3). Only the rank tests were used to investigate the difference between using standard statistical tables versus simulated null distributions (step 3), whereas both rank and direction tests were used to test the character displacement hypothesis.

models of change were simulated using PDSIMUL. Garland et al. (1993) suggested that one might want to constrain the simulations to produce phenotypes within the range observed in the real taxa because some biological constraints might limit the phenotypes attainable. Five boundary condition options (i.e., what happens in a simulation when the boundary level is surpassed; see Díaz-Uriarte and Garland 1996 and Garland et al. 1993 for details) were tested to see how sensitive the results were to the choice of these parameters (Fig. 2, step 1): Unbounded, Flip, Hard Bounce, Soft Bounce, and Truncate. When bounds were chosen (i.e., all options except Unbounded), the upper bound was set to the largest value of an extant species plus 10%, and the lower bound was set to the lowest extant value minus 10%. (Note that selection of any of the boundary options except Unbounded violated the Brownian-motion model.) The Unbounded simulations maintained the mean and variance that was observed in the original dataset (using a modification of the algorithm that R. Nordheim developed for Martins and Garland 1991; T. Garland, pers. comm.). For gradual evolution (Brownian-motion model) simulations, we qualitatively estimated branch lengths of the phylogeny (Fig. 1) based on electrophoretic (Gorman et al. 1980; Gorman et al. 1983) and immunological (Shochat and Dessauer 1981) studies.

We used the "tip" values from the simulations to reconstruct the ancestral nodes using either squared-change or linear parsimony. We used CMMEANAL (part of the CMAP computer package of Martins and Garland 1991, with slight modification to the output) to obtain squared-change and weighted squared-change reconstructions (i.e., changes

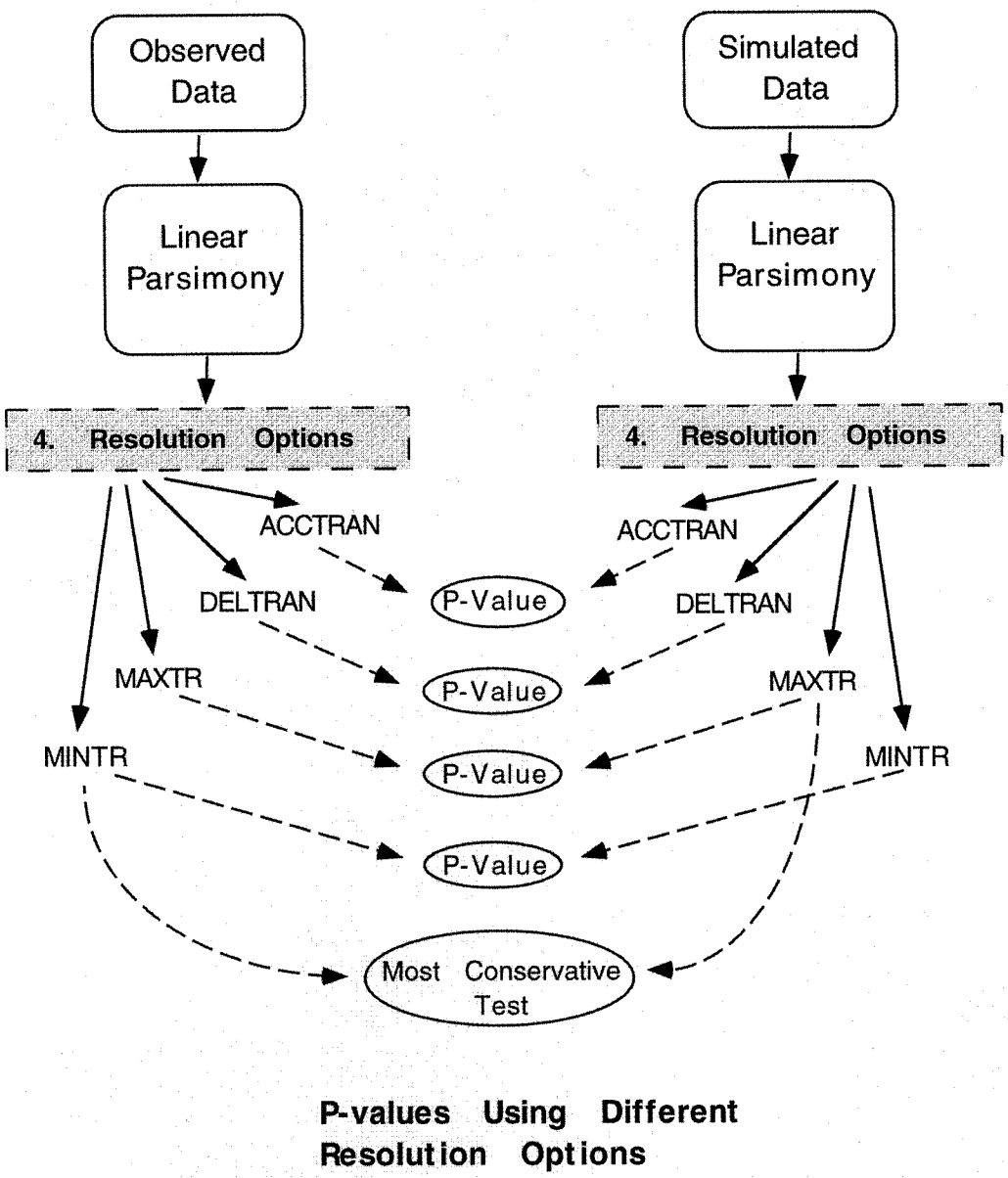

FIG. 3. Hypothesis tests using linear parsimony. Ambiguities occurred at many nodes, which were resolved using four methods. In addition, an extremely conservative test was applied: using MINTR to resolve the observed ambiguities, and MAXTR to resolve the simulated nodes. The tests for squared-change and weighted squared-change parsimony were similar, except that there was only one comparison for each method.

weighted by branch lengths as described by Huey and Bennett 1987). Forty iterations per simulation were sufficient to calculate nodal values identically to six decimal places. Weighted squared-change parsimony was used with the gradual simulations and unweighted squared-change parsimony was used with speciational simulations (Fig. 2).

The PDSIMUL and CMMEANAL programs were run on an IBM 386–compatible computer. Each program generated extremely large output files (on the order of 100 kilobytes of disk space for 500 simulations with 24 taxa). The output files were then transferred to a DECstation 5000/200 for further analysis. A program called CHDISP.C (written in C) was used to conduct the character displacement hypothesis test. All programs are available upon request.

Resolving Ambiguities in Linear Parsimony Simulations

To calculate minimum amounts of change for each simulation using linear parsimony, we modified the CONTRAST.C program from the PHYLIP package (Felsenstein 1993) to implement the algorithm of Swofford and Maddison (1987). This new linear parsimony program is called MAXBRANCH.C. Linear parsimony often results in an infinite number of equally parsimonious reconstructions. In the simulations, we used the same four algorithms for resolving ambiguities as we employed on the real dataset: ACCTRAN, DELTRAN, Maximum change on Transition branches, and Minimum change on Transition branches (hereafter, ACC, DEL, MAXTR, and MINTR, respectively; Fig. 3). We used

the *cristatellus* group as an outgroup (MAXTR and MINTR do not require an outgroup, although in this case, using the outgroup reduces the number of ambiguous nodes). Note that our use of terms such as "basal node" does not imply that an outgroup was actually necessary but is used only to indicate the starting node for the algorithms.

To obtain unambiguous resolutions using ACCTRAN and DELTRAN, a value must be assigned to the basal node of the entire clade. If the basal node to the ingroup is ambiguous, and it is not possible to resolve the ambiguity by including an outgroup, then the range of possible values must be sampled. In theory, MAXTR and MINTR resolution methods are not affected by ambiguity at the basal node, as they apportion the greatest or least amount of change allowable by linear parsimony. However, our program MAXBRANCH.C does not automatically accommodate ambiguous basal nodes. In such cases, we suggest "rerooting" the tree so that the program can begin with a nearby unambiguous node as the basal node.

Usually, an outgroup is used to fix the basal node. In our case, the outgroup is not a single species, but the *cristatellus* species group, which presented an additional problem for the simulations. Although the statistical analysis only examines the branches among the northern Lesser Antillean anoles (the *bimaculatus* series), the simulations maintain the mean and variance of the input data, and the *cristatellus* group lacks the upper range of variation in body size as compared with the *bimaculatus* series (13.2–19.8 mm in jaw length vs. 13.5–28.8 mm). Thus, the inclusion of the *cristatellus* group in the simulations will change the mean and variance, which may bias the results. To circumvent this problem, we reconstructed the value of the most basal node of the *cristatellus* group and used this value as a single outgroup. The method used to assign a value to this outgroup matched the method used in the simulations; for example, when linear parsimony was used to reconstruct ancestral values for the simulation runs, it was also used to calculate the outgroup value.

MAXTR resolutions are calculated by starting at the base of the tree and fixing ambiguous nodes successively up the tree so that size change is maximized on transition branches and minimized on the remaining branches. As each node is fixed, the possible character states are reevaluated for the portion of the tree above the fixed node (i.e., all the descendant taxa of the fixed node). Once the descendant nodes on transition branches one and two (nodes numbered 25 and 32, Fig. 1) are fixed, all remaining ambiguities on descendant branches are resolved using DEL. This has the desired effect of maximizing change at transition branch three (the most terminal transition branch), as well as fixing all the remaining nodes. In an analogous manner, the MINTR algorithm minimizes change at the transition branches and uses ACC to resolve remaining ambiguities.

Results

Bounds in the Simulations

The choice of boundary options influences the distribution of the simulated changes in both speciational and gradual evolution simulations (Fig. 2, step 1). The mean change was similar in all cases (approximately 0.0), except for the Flip option in the speciational simulations (which was shifted slightly to the left, Fig. 4a). However, the shapes of the distributions generated using boundary options are slightly altered as compared with the Unbounded simulations (Fig. 4). The bounded speciational simulations are platykurtotic, and the gradual simulations are more peaked at the mean with fewer observations in the tails. The standard deviations and ranges of the changes are smaller in the bounded simulations, especially in the gradual simulations.

The slight kurtosis introduced by using bounds in the simulations had different effects on the distributions of reconstructed change. When we used squared-change (SqCh) or weighted squared-change (WSqCh) parsimony reconstructions, the simulations using boundary options produced even more peaked distributions of reconstructed changes relative to Unbounded simulations (Fig. 5a,b). However, when we used linear parsimony (LinP) reconstructions, the leptokurtosis is very slight in bounded reconstructions in comparison to the Unbounded reconstructions (Fig. 5c).

Differences Between Simulated and Reconstructed Data

In the simulations, we examined how accurately the various methods reconstructed evolutionary changes on the phylogenetic tree by comparing the reconstructed changes with the actual values (as produced by the simulations, Fig. 2, step 2). Both methods of ancestral character state reconstruction produce distributions of changes that are significantly different from the "actual" changes (Fig. 6). Figure 6a compares the distributions of changes produced in the speciational Unbounded simulations (indicated by a line) with the distribution of changes reconstructed by linear or squared-change parsimony (indicated by bars). Changes for the Unbounded speciational simulations are normally distributed (Kolmogorov-Smirnov test $P > 0.15$), whereas every other distribution, either of simulated or reconstructed changes, differs significantly from normality. The distribution of changes for SqCh reconstructions is highly altered (more peaked with fewer observations in the tails) in comparison to the 'actual' (simulated) changes. The distribution of LinP reconstructions is also altered, but in this case, it is extremely leptokurtotic (strongly peaked with a large number of observations in the tails). When we conduct the analysis with branch length information included, the distribution of WSqCh reconstructions is again more peaked with fewer observations in the tails in comparison to the Gradual Unbounded simulations (Fig. 6b).

Additionally, the distribution of changes produced by SqCh and LinP are radically different (Fig. 6a). Both of the squared-change methods produce many more small, but nonzero, evolutionary changes than linear parsimony. By contrast, linear parsimony produces considerably more branches with large evolutionary changes, and an extraordinarily large number of branches with no change. Our observation concurs with earlier predictions (Losos 1990a; Maddison and Maddison 1992) that squared-change parsimony spreads change over the entire tree, whereas linear parsimony restricts change to fewer branches, but when change is inferred, the magnitude of the change is considerably larger. The reconstruction methods alter the shape of the tail area of the distribution of

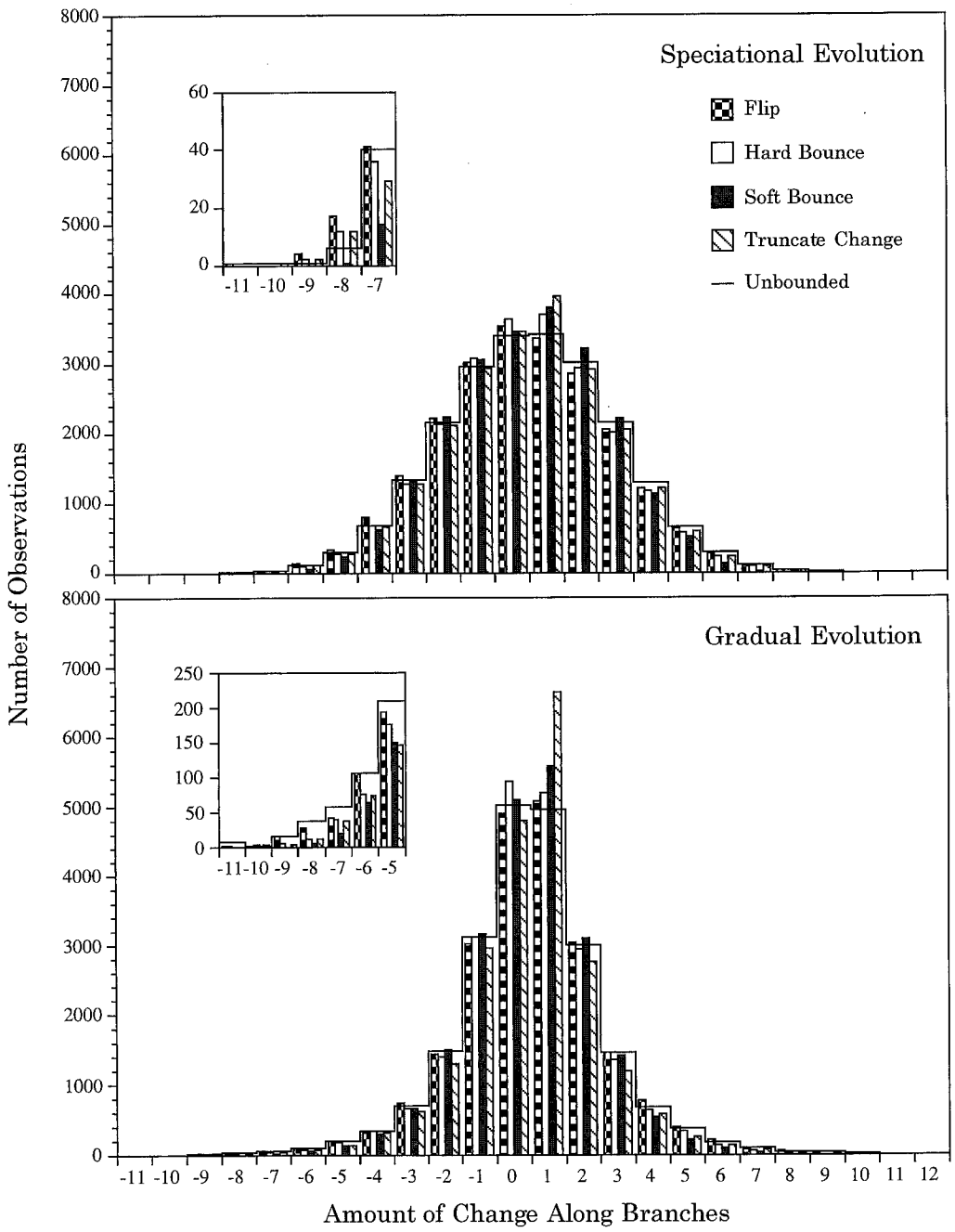

FIG. 4. The effect of boundary options in the speciational (upper) and gradual (lower) evolution simulations. The distribution of changes produced by the Unbounded option is used as the baseline (indicated by the line). These are the "real" changes produced by simulation.

changes (fewer observations in tail regions using squared-change reconstructions and greater number of observations in linear parsimony reconstructions, Fig. 6 insets).

Testing Hypotheses with Standard Statistical Tables versus Simulations

Squared-Change Parsimony

For the final test of the character displacement hypothesis, we used both the rank and direction tests (Fig. 2, step 5). However, to compare *P*-values resulting from standard statistical tables versus simulations, we used only the rank tests (Fig. 2, step 3) because there is no standard test that is equivalent to combining a rank and direction test.

Using any of the three methods for reconstructing change in the real dataset, the character displacement hypothesis is supported: change on the three branches on which the transition occurs from one-species to two-species islands is greater than changes that occur on the other branches, and changes along transition branch 1 are in the opposite direction from transition branches 2 and 3 (Table 1). In the rank-only analyses, we can compare observed changes with standard sta-

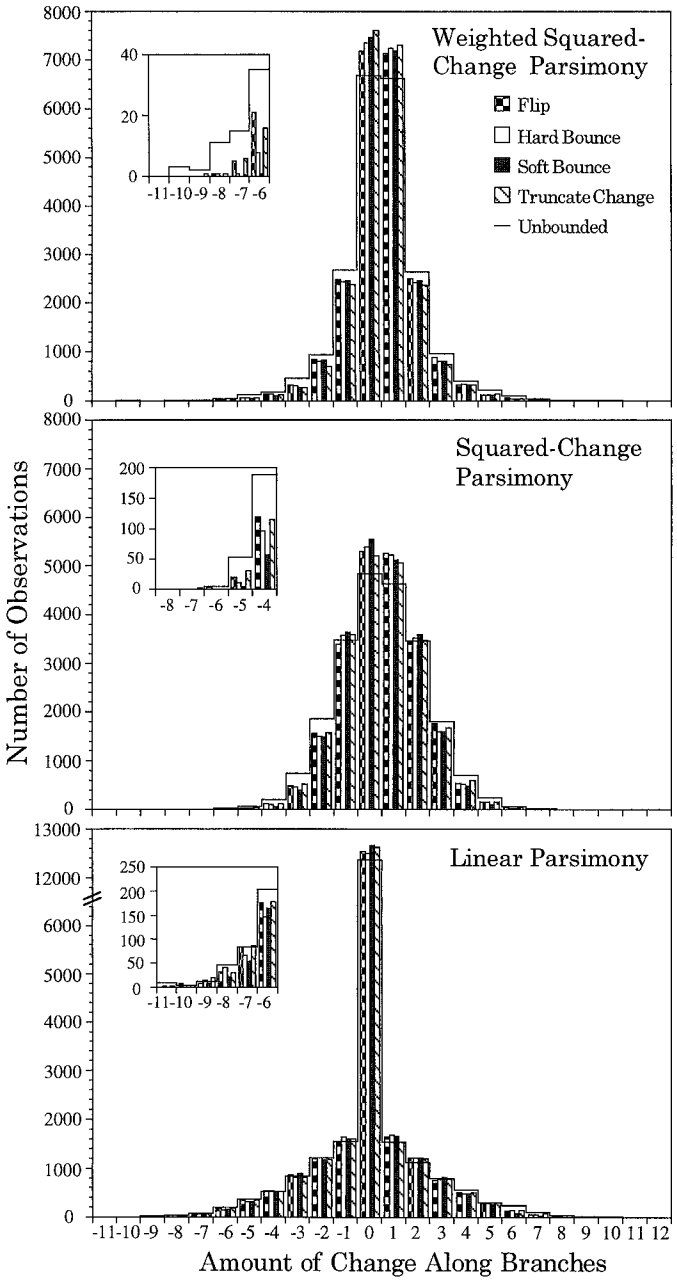

FIG. 5. Comparison of the effect of boundary options in the simulations on the distribution of reconstructed changes. In each case, the Unbounded option is used as the baseline. (a) gradual evolution simulations with weighted squared-change reconstructions; (b) speciational evolution simulations with squared-change reconstructions; (c) speciational evolution simulations with linear parsimony reconstructions (all using the MAXTR resolution method).

tistical tests. The rank-only analyses of the reconstructed character states for weighted and unweighted squared-change parsimony yield similar results. For unweighted squared-change parsimony, the sum of the ranks for the three transition branches is 99, whereas for weighted squared-change parsimony, the sum of the ranks is 96. If we ignored the statistical problems discussed above, both analyses would provide marginally nonsignificant one-tailed P-values using the Wilcoxon Rank-Sum test ($0.05 < P < 0.10$; in 1990b using the unweighted analysis, Losos reported a P-value < 0.056).

By contrast, the simulation studies suggest that the observed patterns are considerably less likely to have occurred by chance. In the Unbounded simulations for both the squared-change and weighted squared-change analyses, the probability of obtaining a test statistic as large as that observed in the real data was $P = 0.040$ and $P = 0.028$ (squared-change and weighted squared-change, respectively, Table 2). The addition of the direction test in conjunction with the simulations resulted in even lower probabilities ($P = 0.010$ for squared-change and $P = 0.014$ for weighted squared-change, Table 3). When the simulations are bounded by choosing one of the boundary options, there is very little difference in P-values (Table 3), despite differences in the distributions of nodal values reported above (Fig. 3).

Linear Parsimony

Reconstruction of the real dataset using linear parsimony results in ambiguous character state reconstructions at ancestral nodes. The four resolution rules yielded the following rank-sums: MAXTR = 96.5, MINTR = 90, ACCTRAN = 96.5, and DELTRAN = 93.5 (ignoring statistical problems, these rank-sums yield nonsignificant P-values from standard statistical tables [$0.07 < P < 0.14$]; Table 2). As with the squared-change reconstructions, using the simulations to assess probabilities for the rank-sums results in lower probabilities for all cases (Table 2). Including the direction test in conjunction with the rank-sums in the simulations to evaluate P-values, all methods of resolving ambiguities yield significant results regardless of simulation parameters (Table 3). When the simulations are conducted without bounds, the resolution methods ranked in P-values: ACC < MINTR < DEL < MAXTR, although the differences between the four methods are small. When the boundary options are considered, the only pattern that is evident is that the resolutions that allow the maximum amount of change along the transition branches are the most conservative resolutions (Table 3). The P-values of the remaining combinations of resolution methods and boundary options are not ordered in any regular manner. The distributions of the reconstructed changes were not affected by the choice of resolution options (Fig. 7).

Resolving Ambiguities in Linear Parsimony Reconstructions

All four methods of resolving the ambiguities in character states created by using linear parsimony resulted in significant P-values, as described above. All of the previous hypothesis tests compared values reconstructed in the same way for real and simulated data. We wanted to push the exploration one step further and design the most conservative test possible. This is to compare the reconstruction of the observed data that is least favorable with the hypothesis of character displacement (MINTR) to the reconstructions of the simulated data that is most favorable to the hypothesis (MAXTR). This would produce the absolute minimum difference between observed and expected values. Using UNBOUNDED simulations, this procedure results in a P-value of 0.082. The MINTR resolutions of the observed data compared with the other possible resolution options of the sim-

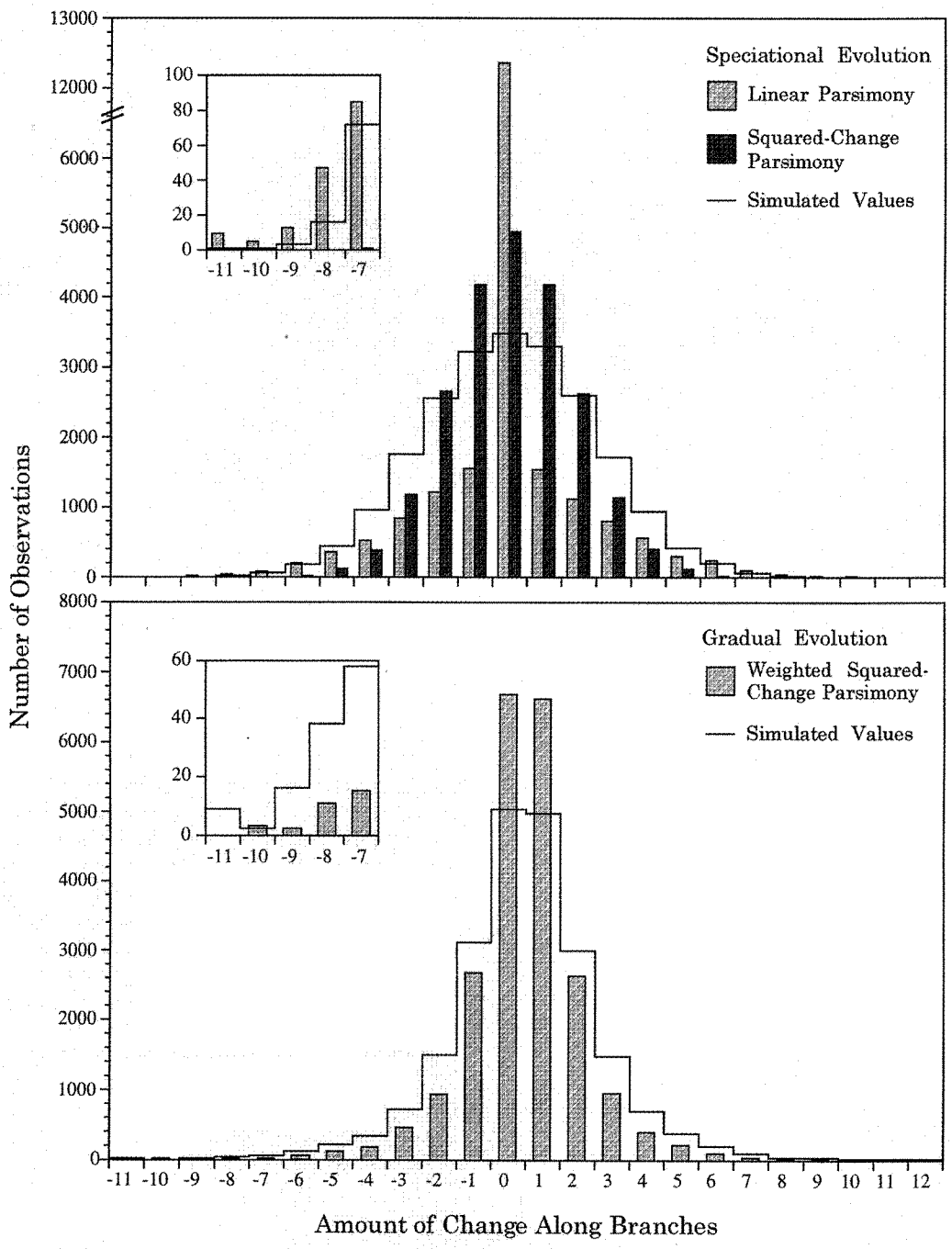

FIG. 6. Histograms of the simulated changes along branches compared with the reconstructed changes. (a) speciational evolution simulation model with linear and squared-change parsimony reconstructions; (b) gradual evolution simulation model with weighted squared-change parsimony reconstructions (both used the Unbounded option). The simulated values are indicated by the line, the respective reconstruction values are indicated by bars.

ulated data yielded *P*-values ranging from 0.018 to 0.052 (MINTR-obs vs. DEL-sim: 0.052; MINTR-obs vs. ACC-sim: 0.044; MINTR-obs vs. MINTR-sim: 0.018).

DISCUSSION

Certain evolutionary hypotheses, such as character displacement, require predictions about the amount of evolution on a particular branch of the phylogenetic tree ("directional" evolutionary hypotheses, sensu Harvey and Pagel 1991). Previous workers have addressed these hypotheses by reconstructing ancestral character states, inferring the minimum amount of evolutionary change that occured between ancestral and descendant taxa, and then testing the statistical significance of inferred changes via comparison to standard statistical tables (but see Garland and Adolph 1994 and McPeek 1995 for different approaches). However, using inferred

TABLE 1. The body size changes reconstructed along transition branches in the observed data using linear, squared-change, and weighted squared-change parsimony. The average amounts of change on the remaining (non-transition) branches are also included. The changes produced by the different resolution methods for linear parsimony are presented.

	Changes along transition branches			Avg. change on other branches
	1	2	3	
Linear parsimony				
MAXTR	−4.9	6.0	0	−0.044
MINTR	−4.7	1.3	0	−0.261
ACCTRAN	−4.9	4.8	0	−0.020
DELTRAN	−4.7	2.5	0	−0.285
Squared-change parsimony	−2.6	2.6	0.28	0.054
Weighted squared-change parsimony	−2.5	2.7	0.29	0.071

TABLE 2. Comparison of simulated versus normal-approximation *P*-values for the Wilcoxon Rank-Sums test in the Unbounded simulations. These values only test that the magnitude of the evolutionary changes along the transition branches are greater than changes on the non-transition branches. They do not include the direction test. In all cases, the nonparametric Wilcoxon Rank-Sums test statistic was used, but the statistical significance was assessed in two ways. Normal-approximation *P*-values are what would be obtained from using a large-sample approximation to the normal distribution, whereas simulated *P*-values are those obtained from comparing the observed rank-sums to a null distribution obtained by phylogenetic simulation.

	Observed rank-sums	*P*-values using the normal approximation	*P*-values from the simulations
Squared-change parsimony	99	0.071	0.040
Weighted squared-change parsimony	96	0.093	0.028
Linear parsimony			
MAXTR	96.5	0.088	0.0146
MINTR	90.5	0.142	0.044
ACCTRAN	96.5	0.088	0.062
DELTRAN	93.5	0.113	0.076

changes or character states in standard hypothesis tests is not statistically valid, and in our case, leads to under-estimation of statistical significance.

The statistical nonindependence problems created by using minimum evolution (or parsimony) methods to infer evolutionary changes or character states has long been recognized (Felsenstein 1985; Maddison 1991; Swofford and Maddison 1987). Much of the recent discussion has focused on the problem of inflated degrees of freedom, with the implication that these methods will result in Type I error (i.e., that studies will find greater significance than is warranted by the data). However, it is not widely appreciated that the scope of the nonindependence problem is much larger, and cannot be easily fixed by reducing the degrees of freedom or by using nonparametric tests (discussed by Felsenstein 1985). As we demonstrated in our character displacement study, it is also possible to have increased Type II error (i.e., that studies will fail to find significance when it is actually warranted by the data; see below) when using ancestral character state reconstruction methods. Although we have no way of knowing how general a problem this might be, it seems an equally likely source of error as the degrees of freedom problem.

Both of these problems are solved by using computer simulation to assess the significance of character state reconstruction data. In this way, we can address directional evolutionary hypotheses while circumventing the problem of phylogenetic nonindependence of the data. In addition, we can evaluate the robustness of our tests given the ambiguity in ancestor reconstruction methods and range of available simulations parameters (e.g., Díaz-Uriarte and Garland 1996; Garland et al. 1993; Martins and Garland 1991). We tested the robustness of our hypothesis test at various levels of analysis: simulation boundary options, choice of simulation model (gradual brownian motion vs. speciational [i.e., Brownian-motion with equal branch lengths]), reconstruction method (squared-change or weigthed squared-change vs. linear parsimony), and resolution methods used with linear parsimony. In our case study of character displacement in body size, the null hypothesis of random evolution was rejected in six of the seven combinations of simulation parameters and reconstruction methods. The one exception had a $P < 0.082$, and is discussed below.

TABLE 3. Comparison of squared-change versus linear parsimony *P*-values. Phylogenetic simulations were conducted with the five boundary options: Unbounded, Flip, Hard Bounce, Soft Bounce, and Truncate. Evolution in body size was simulated using a gradual model (G) for use with weighted squared-change parsimony, and simulated using a speciational model (S) for use with squared-change and linear parsimony. Ambiguous character states obtained using linear parsimony were resolved in one of four ways (MAXTR, MINTR, ACCTRAN, and DELTRAN, as described in the text). Overall *P*-values (as determined by simulation, including magnitude and direction tests) are tabulated.

	UNBOUNDED	Flip	Hard bounce	Soft bounce	Truncate
Squared-change parsimony (S)	0.010	0.004	0.012	0.012	0.014
Weighted squared-change parsimony (G)	0.014	0.010	0.014	0.012	0.024
Linear parsimony (S)					
MAXTR	0.032	0.046	0.042	0.042	0.050
MINTR	0.018	0.024	0.034	0.026	0.020
ACCTRAN	0.016	0.016	0.014	0.018	0.022
DELTRAN	0.026	0.014	0.020	0.024	0.020

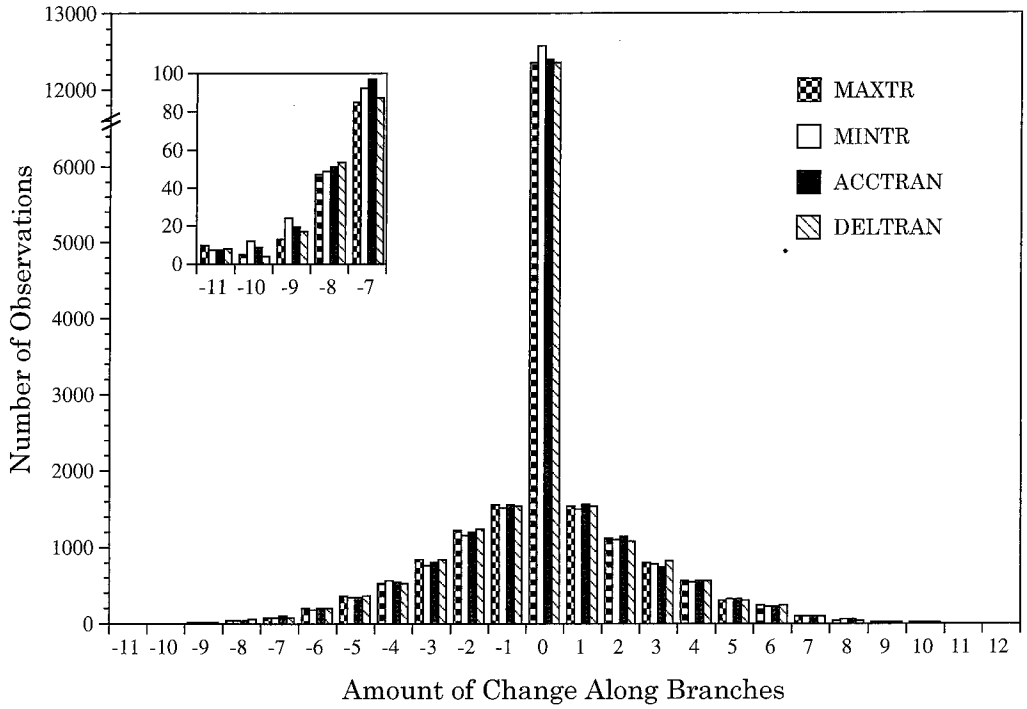

FIG. 7. A comparison of the resolution methods used in linear parsimony reconstructions on Unbounded speciational evolution simulations. All boundary options yielded very similar distributions.

Bounds in the Simulations

The choice of boundary options (Unbounded, Flip, Hard Bounce, Soft Bounce, and Truncate) made minor, though nonsignificant, differences in the statistical tests (Table 3). The distributions of the changes produced by these options are, however, nonnormal when simulations are conducted with equal branch lengths (Speciational Evolution). The Brownian motion model of evolution requires that the changes be normally distributed (with variance proportional to time since divergence or branch lengths), so the use of boundary options violates the assumptions of the Brownian motion model (see also Díaz-Uriarte and Garland 1996). Because the boundary options did not make any difference to the statistical tests, we will discuss only the Unbounded option to reduce the number of comparisons (however, if we had chosen a much narrower range for the upper and lower bounds, we would have found a stronger effect on the analysis).

The use of branch length information in the analysis did not make a large impact on the ultimate results. The *P*-values from the gradual evolution/weighted squared-change reconstructions were consistently slightly higher than those of the speciational evolution/squared-change reconstructions (Table 3).

The Importance of Differences between Simulated and Reconstructed Data

The distribution of reconstructed changes is very different from the distribution of "real" changes (from the simulations), upon which they were based. In the case of squared-change parsimony, this results in a distribution that is highly leptokurtotic relative to the changes in the "real" data (Fig. 6). A heuristic way to explain this is as follows (J. Cheverud, pers. comm.): When there is actually a large change on an internal branch of a tree, squared-change parsimony algorithms will assign most of the change on the correct branch, but will also assign some of the change to neighboring branches. Thus, the large change is concentrated on the correct branch, but 'bleeds out' onto neighboring branches. Large changes tend to be made smaller, resulting in a greater number of small changes and fewer large changes.

Linear parsimony produces a distribution of changes that is extremely different from the distribution of changes produced by a Brownian motion–based simulation. This is to be expected because we know that weighted squared-change parsimony (and not linear parsimony) is a good estimator for the nodal values under Brownian motion (Maddison 1991). An evolutionary model has not yet been linked with linear parsimony, so we cannot assess to what degree the linear parsimony algorithm accurately reconstructs changes when evolution was simulated in the situation in which the method performs optimally.

This discussion is not meant to imply that we expect evolution to occur parsimoniously. Rather, it serves to underscore the point that it is not statistically valid to use inferred values from any parsimony criterion in parametric or nonparametric hypothesis tests because parsimony algorithms inherently bias the data, and this bias is not incorporated into standard statistical test distributions. If one ignores this statistical problem for the *bimaculatus* dataset, a comparison of the *P*-values obtained from standard statistical tables versus those obtained from simulations shows that the simulated *P*-values are always lower (Table 2). The extent to which we may generalize from this result to other studies is difficult to pre-

dict. However, this clear demonstration of what may happen when standard statistical tests are inappropriately applied to parsimony methods reinforces the necessity of using simulation methods for hypothesis testing.

Testing Hypotheses with Standard Statistical Tables versus Simulations

A previous study (Losos 1990b) compared scores obtained from reconstructed values with standard statistical tables which assume that the data (reconstructed changes on each branch) are statistically independent. This method has two potential problems. First, because $2N$-2 branches are reconstructed on a phylogenetic tree of N species, using the number of branches as the sample size leads to overestimation of the degrees of freedom, which potentially inflates the rate of Type I error.

However, there may be a more influential source of error in comparing scores obtained from reconstructed values and comparing them to standard tables. The reconstruction methods alter the shape of the distribution of changes, especially in the tail region, which is critical for assessing the statistical significance of our results. Support for an hypothesis of character displacement requires the detection of large changes. However, squared-change methods generally underestimate the frequency of large changes. Thus if large changes occur, then squared-change methods may underestimate the evidence in favor of the character displacement hypothesis resulting in Type II error. The effect of linear parsimony is less predictable, because of the lack of an appropriate evolutionary model, as mentioned above. However, we can compare the distribution of reconstructed values with the normal distribution of simulated changes (which the reconstructions are based on, Fig. 4a). Linear parsimony reconstructs more values in the far extremes of the tail than what is present in the normal distribution, but far fewer reconstructed values in the intermediate range of values. Thus, generally speaking, a P-value read from a standard statistical table will be inaccurate.

For both squared-change and linear parsimony, our results using null distributions from simulations were much more significant than the corresponding analyses which relied on standard statistical tables and used numbers of branches to calculate degrees of freedom. This finding suggests that the second concern stated above is the most important in this example: by minimizing large and intermediate changes, ancestral reconstruction methods dilute the evidence for character displacement when compared with standard statistical tables. However, this is not a problem with the simulation approach because we employed the same reconstruction method for both the observed data and in creating the null distribution, cancelling out the effect.

Resolving Ambiguities in Linear Parsimony Reconstructions

Linear parsimony presents the added technical difficulty of often producing many (actually, an infinite number of) sets of most parsimonious solutions for continuous characters. Previous workers have used ACCTRAN/DELTRAN as a means for examining maximally disparate resolutions of ambiguous nodes, but ACCTRAN and DELTRAN do not necessarily represent the most and least conservative sets of resolutions for any given evolutionary hypothesis (Maddison and Maddison 1992; Swofford and Maddison 1987). In our case, we chose two methods of resolving ambiguous nodes, Maximum Transition and Minimum Transition, (note that these are not the Max State and Min State of Swofford and Maddison 1987) that could produce the most liberal and most conservative resolutions. In the real dataset, MAXTR resulted in the greatest amount of change apportioned to the transition branches, and MINTR the least. We used the same methods for resolving ambiguities in the observed and simulated datasets (i.e., MAXTR observed vs. MAXTR simulated, and MINTR observed vs. MINTR simulated). Using these methods, MINTR produced the most significant results and MAXTR the least significant results.

Although this result seems counter-intuitive, the explanation probably rests in the structure of our data. In the reconstructions of the real dataset (observed), only the second transition branch has an ambiguity that can vary substantially (range 1.3–6.0; Table 1). The first transition branch is restricted to varying between 5.1 and 5.3, and the third transition branch is fixed at 0.0; consequently, the difference between the MAXTR and MINTR reconstructions is not very great. In the simulations, there are enough instances in which the branches are allowed to vary more considerably, giving more liberal (relative to our hypothesis) simulated resolutions in the case of MAXTR and more conservative simulated resolutions in MINTR. Thus, the observed MAXTR resolutions tend to have a lower score than more of the MAXTR simulated resolutions, and the observed MINTR resolutions tend to have a higher score that more of the MINTR simulated resolutions. MAXTR is actually the most conservative and MINTR the most liberal method when one compares observed data with simulations in which the same resolution method is used in both.

The most conservative test imaginable is to compare the MINTR resolutions of the observed data with the MAXTR resolution of the simulated data. Given the extremely conservative nature of the test, the P-value (0.082) was rather low.

Except for this highly conservative test, our tests used the same method of resolving ambiguities in both the observed data and simulated data, and can be thought of as representing a consistent pattern of evolutionary change that has occurred in the history of the group. For example, if some aspect of the characters under consideration make evolutionary reversals rare, then DELTRAN would be the most appropriate resolution method for both the observed and simulated datasets (Maddison and Maddison 1992; Swofford and Maddison 1987). Because we do not know how evolution actually proceeded, it seems reasonable to explore a variety of different resolution rules which span the possible range. In our case, the maximum or minimum amount of evolutionary change on the critical transition branch Number 2 is never found using ACCTRAN or DELTRAN (Table 1), which motivated the development of MAXTR and MINTR. Customized resolution rules such as these can be developed for any hypothesis, making the exploration of the uncertainty more rigorous.

Implications for the Hypothesis of Character Displacement in Northern Lesser Antillean Anoles

We found that all of the resolution method tests used with Linear Parsimony were highly significant, irrespective of the resolution method used (ACC, DEL, MAXTR, and MINTR); even the most conservative test was near significance. Both squared-change parsimony and weighted squared-change parsimony also resulted in highly significant results. This is strong support for the character displacement hypothesis.

With all of the ambiguities encountered in using ancestral reconstruction methods, why would one wish to use them? One potential benefit is that these methods allow great flexibility in testing compound hypotheses. In testing historically based hypotheses, we will never be able to completely eliminate uncertainty. Thus, whether particular methods produce only one solution is an arbitrary criterion for choosing among phylogenetic comparative methods. Rather, we should systematically explore the level of confidence that we have in our results using a variety of disparate methods. The ability to test whether our data fit the assumptions of one method better than those of other methods would be very useful.

ACKNOWLEDGMENTS

Many thanks are due to J. Cheverud, J. Felsenstein, and T. Garland for helpful comments which greatly improved the manuscript; to E. Martins, J. Cheverud, and W. Maddison for discussions on reconstructions methods and associated evolutionary models; and S. Sawyer for statistical advice. J. Felsenstein, T. Garland, and E. Martins generously provided source code and gave suggestions on programming modifications. We thank the National Science Foundation (DEB 9318642 and DEB 9423473) and the David and Lucile Packard Foundation for financial support.

LITERATURE CITED

DÍAZ-URIARTE, R., AND T. GARLAND JR. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from brownian motion. Syst. Biol. 45:27–47.

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

———. 1993. PHYLIP (phylogeny inference package). Vers. 3.5c. Distributed by the author. Department of Genetics, Univ. of Washington, Seattle.

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

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

GARLAND, T., JR., K. L. M. MARTIN, AND R. DÍAZ-URIARTE. 1997. Reconstructing ancestral trait values using squared-change parsimony: plasma osmolarity at the amphibiam-amniote transition. Pp. 425–501 *in* S. Sumida and K. L. M. Martin, eds. Amniote origins: completing the transition to land. Academic Press, New York.

GORMAN, G. C., D. BUTH , M. SOULÉ, AND S. YANG. 1980. The relationship of the *Anolis cristatellus* species group: electrophoretic analysis. J. Herpetol. 14:269–278.

GORMAN, G. C., D. BUTH , M. SOULÉ, AND S. YANG. 1983. The relationships of the Puerto Rican *Anolis*: electrophoretic and karyotypic studies. Pp. 626–642 *in* A. Rhodin and K. Miyata, eds. Advances in herpetology and Evolutionary Biology. Museum of Comparative Zoology, Cambridge.

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

HASS, C. A., S. B. HEDGES, AND L. R. MAXSON. 1993. Molecular insights into the relationships and biogeography of West Indian anoline lizards. Biochem. Syst. Ecol. 21:97–114.

HOLLANDER, M., AND D. A. WOLFE. 1973. Nonparametric statistical methods. Wiley, New York.

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

LAZELL, J. D. 1972. The anoles (Sauria: Iguanidae) of the Lesser Antilles. Bull. Mus. Comp. Zool. 143:1–115.

LOSOS, J. B. 1990a. Ecomorphology, performance capability, and scaling of West Indian *Anolis* lizards: an evolutionary analysis. Ecol. Monogr. 60:369–388.

———. 1990b. A phylogenetic analysis of character displacement in caribbean *Anolis* lizards. Evolution 44:558–569.

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

———. 1991. Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Syst. Zool. 40:304–314.

MADDISON, W. P., AND D. R. MADDISON. 1992. MaClade: analysis of phylogeny and character evolution. Vers. 3.0 Sinauer, Sunderland, MA.

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

MARTINS, E. P., AND T. HANSEN. 1996. Phylogenetic comparative methods: the statistical analysis of interspecific data. Pp. 22–75 *in* E. Martins, eds. Phylogenies and the comparative method in animal behavior. Oxford Univ. Press, London.

MCPEEK, M. A. 1995. Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. Am. Nat. 145:686–703.

MILES, D. B., AND A. E. DUNHAM. 1996. The paradox of the phylogeny: character displacement of analyses of body size in island *Anolis*. Evolution 50:594–603.

SCHOCHAT, D., AND H. C. DESSAUER. 1981. Comparative immunological study of the albumins of *Anolis* lizards of the Caribbean islands. Comp. Biochem. Physiol. 68A:67–73.

SCHOENER, T. W. 1970. Size patterns in West Indian *Anolis* lizards. II. Correlations with the sizes of particular sympatric species—displacement and convergence. Am. Nat. 104:155–174.

SILLÉN-TULLBERG, B. 1988. Evolution of gregariousness in aposematic buttlerfly larvae: a phylogenetic analysis. Evolution 42: 293–305.

———. 1993. The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. Evolution 47:1182–1191.

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

WILLIAMS, E. E. 1972. The origin of faunas. Evolution of lizard congeners in a complex island fauna: a trial analysis. Evol. Biol. 6:47–89.

Corresponding Editor: L. Leamy