CHAPTER 3

How to Study Discrete Comparative Methods

Mark Ridley and Alan Grafen

This chapter has three main purposes. The first is to show how to justify a proposed statistical method for discrete comparative data. We shall be particularly concerned with associative hypotheses. However, several authors have recently recommended methods for testing directional, rather than associative, hypotheses with discrete data. Our second purpose is to argue that the conceptual advantage of directional methods has been exaggerated and that the statistics of directional methods present difficult problems that have not been solved. Our third purpose is to compare the statistical approach we have taken with some alternatives — mainly from likelihood theory, though we also comment on bootstrapping — and we shall argue that, at least in their standard forms, they are inapplicable to phylogenetically structured data. We shall therefore set out our view of how discrete data should be studied — or, at least, of how potential methods to study discrete data should themselves be studied — and criticize the main alternative approaches being advocated in the current literature.

This chapter is concerned with discrete as opposed to continuous data. Data can be discrete for either of two reasons. One is that it is really continuous in its underlying form but has been discretized for a more or less observational reason. In an extreme case, a manifestly continuous variable like time might be divided into long/short categories for a test; in a subtler case, something we perceive

categorically might have an underlying quantitative genetic or embryologically continuous control. Alternatively, the data might really be discrete, as in the example suggested by Burt (1989) of modes of X/Y sex determination.

How to justify a comparative method

A. Problems that a discrete comparative method must solve

The minimal study will have two observed characters (call them A and B), each of which can have a number of states (from 1 up to n_A and n_B); in the simplest case they have only two states (A_1 or A_2 and B_1 or B_2, which we shall often write as A/a and B/b). The states of both these observable characters are known for a number of species, all at the tips of a phylogeny.

The main statistical problem is that the different species are nonindependent. Grafen and Ridley (1995) identify three sources of nonindependence. One, which we call phylogenetic overcounting, is well known (Clutton-Brock & Harvey 1977; Ridley 1983). It occurs when a group of related species, all sharing the same character state, are entered in the test as more than one trial (usually as many trials as there are species in the group). The other two had not (we believe) been identified before and are at least more of a problem with discrete than continuous data and may be exclusive to discrete data. One of them arises in Ridley's (1983) method. The method reconstructs combined character states throughout the entire tree back to the root. It then (in the formalization of Grafen & Ridley 1995) collapses each set of contiguous uniform nodes into single nodes to produce a "character change tree." Figure 1 is an example; it expresses all the changes in the joint character states in the tree; more than one such tree may be compatible with one data set. Changes in the tree suffer from statistical nonindependence because the changes away from any one node in the character change tree have to be away from the state of that node. If the node is AB, a change has to be to Ab, aB, or ab; it cannot be to AB because the tree has been reconstructed by parsimony, and if any neighboring nodes had the same character states, they would have been collapsed into one. It is a parsimonious impossibility for neighboring

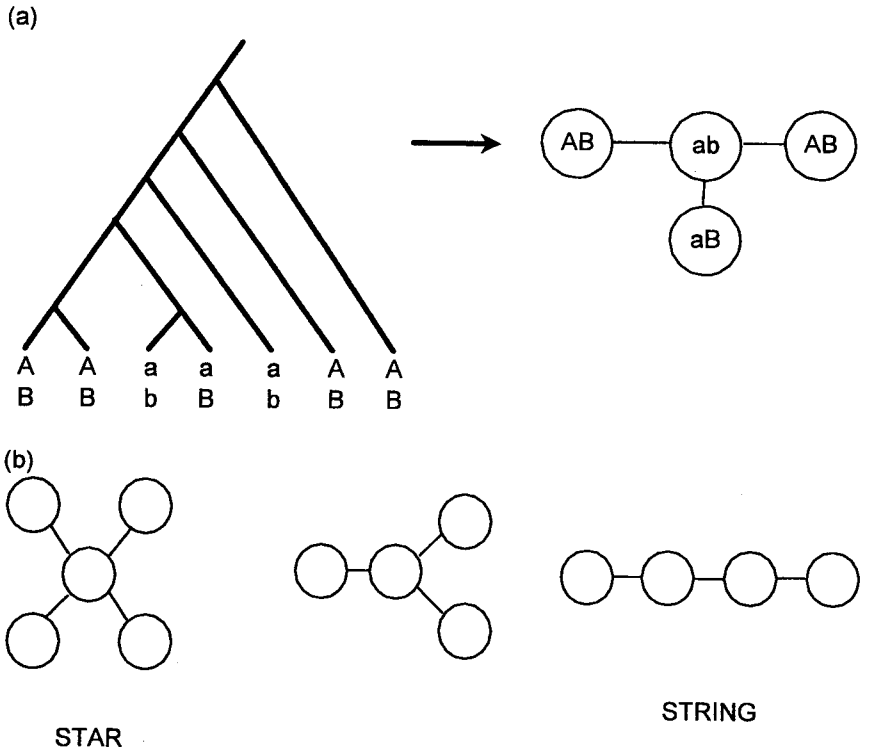

Figure 1. (a) A character change tree illustrates the character changes in a phylogeny, with all the contiguous uniform nodes collapsed into single nodes. (b) It can take on various shapes, depending on the pattern of evolutionary change.

nodes to have the same character state. This problem is likely to arise in any method that reconstructs ancestral states throughout the tree by parsimony and treats inferred changes as independent events. All the changes were separate evolutionary events, but the use of parsimony makes the resulting data points statistically nonindependent.

Other methods make more limited ancestral reconstruction and aim only to find regions of the tree — or partitions of the data — in which change is independent of other such regions. Then the pattern in all the regions is combined in a test. The independent contrasts method of Burt (1989) is one example; the randomization test described by Grafen and

Ridley (1995) is another. The pairwise comparison test of Møller and Birkhead (1992) focuses on variable regions in a tree in a similar, more restrictive, way. These tests all reduce or eliminate the amount of nonindependence due to parsimonious reconstruction because they make little or no use of it; but that is bought at the price of another kind of nonindependence, due to what Grafen and Ridley (1995) call the family problem. These methods identify separate regions in the tree in which evolutionary change must have occurred separately. In Figure 2, for instance, the *AB/ab* change in region 1 must have been a separate historical evolutionary event from the change in region 2. But although they were separate events, they can still be statistically non-independent. If a number of regions share the same local ancestral state, the transition in all those regions will be away from that ancestral state. A full analysis of the consequences of this fact is laborious (Grafen & Ridley 1995); but the net effect, provided that evolutionary change is rare, is an excess of associations within variable regions between the locally nonancestral states. There will usually only be one change in each character per variable region of the tree. If the locally ancestral state is *AB* for several regions, there will be an excess of them showing *Ab* and *aB* with null data. It is possible for there to be a double change (as in Fig. 2) to *ab*, giving *AB* and *ab* within the variable region, but this is less likely by chance than the non-ancestral association. The exact relative chance of finding *AB+ab* or *Ab+aB* depends on the rate of evolutionary change and on the way a particular method selects variable regions within the tree. However, methods that pick variable regions and do not reconstruct the higher ancestral states that are the cause of nonindependence among them are vulnerable to the bias in favor of the nonancestral associations.

These arguments illustrate the awkward fact that, with discrete data, it is not enough to identify separate events that occurred, in an evolutionary sense, independently: those events may still not be statistically independent. We suspect that almost any method for discrete data will have to face up to these problems of nonindependence; and no method that has been proposed so far has been shown, in terms of the logic of its operations, to avoid them successfully. Such are the problems: but how can we find out how much, in any given method, they matter?

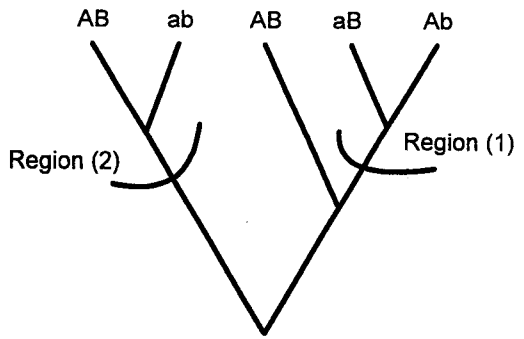

Figure 2. Identification of regions of character variation. Separate changes must have occurred in regions 1 and 2.

B. How to find out whether a proposed method solves the problems

There are two stages in the study of any statistical test. The first is to show that it has valid Type I error rates; the second is to compare the power of those tests that pass the first stage by studying Type II error rates.

Both stages require a model of evolutionary change through a phylogeny, which is a data-generation process. In the specific problem with which we are concerned, the model will be one in which the ancestral state at the root is given and the model specifies how the characters change along the branches of the tree to generate a pattern of character states at the tips. The model can be used to generate large numbers of data sets, which can then by used to scrutinize proposed tests, simply by having the tests analyze the data sets. The model is null if the states of the two characters do not influence each other's evolution, and the data sets can then be used to scrutinize Type I error rates. If the test is valid, it should find associations between the characters that are significant at, for instance, the 5% level 5% of the time, at the 1% level 1% of the time and so on. Alternatively, the state of one character may influence the evolution of the other (and vice versa) and the data sets can then scrutinize Type II error rates. It is crucially important that any method recommended for general use should be shown to be valid in this way. It is impossible to know

whether a test is worth using unless its validity has been assessed against null data generated by a model; indeed, the concept that a test is valid (or, in some looser sense, "good," or "worthwhile," or "well-behaved," or "satisfactory") assumes some model of data generation. In simple cases, such as standard regression theory, the need actually to generate random data sets can be short-circuited by analytical methods; in other cases, such as nonparametric tests, it is necessary only to make some assumptions about the data-generation process and not to construct it explicitly. But the principle is the same in both cases.

In setting out this procedure for justification, we do not mean to imply that no one apart from ourselves understands how to conduct statistical research in biology. The whole or parts of the argument will be familiar to many readers, not least because research with continuous comparative data is now well penetrated by it. The argument does need to be made, however, because it shows practically no influence in any recent research with discrete data. We shall cite all the theoretical work we know of from, say, the past 5 years, and none of it has been conducted by the method, entirely standard in statistics, that we have just described and elaborate on more below. If (which we doubt) work from an earlier period did use this method, it is not having any influence on modern work, and a restatement would be timely.

An alternative to simulational scrutiny is to show, analytically or logically, that a method is valid: this has been the aim of some applications of likelihood theory, applications that we find unreliable. In the current state of statistical theory, simulations are essentially the only reliable route to justification with the complicated problems of phylogenetically structured data.

C. A model for discrete character change through a phylogeny

A statistical test that has been justified by simulation has been shown to be valid only relative to the model that was used to generate the data. It is therefore desirable that the model contain as many of the biologically interesting features as possible. Harvey and Pagel (1991) were the first to introduce an explicitly phylogenetic model of discrete character change in the comparative method; their model (and Pagel's 1994 developments) has attractive features, but we prefer a model in which

the processes generating the phylogenetic structure in the data are even more explicit.

Let us return to the case of two observed characters with two states each. It is well known that if the numbers of species in the states *AB*, *Ab*, *aB*, *ab* are counted, the Type I error rates will be wrong, and too liberal; significant associations will be found in excessive frequency in null data. But why exactly is this? One version of the explanation is as follows. The chance that a character changes from one state to another will usually be influenced by the states of many other characters. We can concentrate on the null case, in which the state of *A* does not influence the evolution of *B* or vice versa. Other, unobserved characters will nevertheless be influencing the evolution of both *A* and *B* (if the other influential characters are observed, they become, for purposes of argument, like *A* and *B*). This can have consequences for the phylogenetic pattern of *A* and *B* in two ways. The simplest would be if a single "third variable" character *E* jointly influenced both *A* and *B*. This could cause phylogenetic pattern for causal reasons. Thus, one state of *E* might causally generate an association between states of *A* and *B* (for example if $E_1 \rightarrow A_1$ and B_1, and $E_2 \rightarrow A_2$ and B_2); this is a familiar kind of problem in comparative, observational, and some kinds of experimental inference (does an association of *A* and *B* mean they are causally related or they are both controlled by an unobserved variable?). However, counting independent trials will not help in this case. The relevant case is the null one, when the unobserved variable produces associations at the species level among the observed variables, but it is equally likely to do so for any of the four possible associations of character states. (In the causal case above, *E* produced particular combinations of *A* and *B*.) Suppose that *E* influenced the rate of evolution of a number of characters, including *A* and *B*; some states of *E* would be general "freezer" states that slowed down the rate of evolution of the other characters, whatever state those other characters were in. Then a block of species would end up with the same states of *A* and *B* — whatever the states were when the lineage entered the freezer *E* state — not because *A* and *B* were influencing each other's evolution but because *E* had brought all change to a stop. A test that counted species, or otherwise phylogenetically overcounted the evidence, would be liable to find spurious evidence of an association.

E is a single hidden character. An alternative is for there to be two different hidden characters (*C* and *D*) that separately interact with the observed characters *A* and *B* respectively. Again, *C* and *D* have states that can freeze the evolution of the character they interact with. But now *C* can freeze only states of *A* and *C* can freeze only states of *B*. For concreteness, suppose *C* can take on six states and the relation with states of *A* is as follows:

$C = 1$	$A = 1$
$C = 2$	$A = 1$
$C = 3$	$A = 2$
$C = 4$	$A = 1$
$C = 5$	$A = 2$
$C = 6$	$A = 2$

There is a perfect deterministic relation between the states of *C* and *A*. Now suppose that transitions between some *C* states are more frequent than others: transitions between $C = 2$ and $C = 3$ and between $C = 4$ and $C = 5$ are common; but transitions out of these states, and out of $C = 1$ and $C = 6$, are rare. $C = 1$ and $C = 6$ then act as "freezer" states for *C*: if a taxon is $C = 1$, the observed state of *A* will be 1 in a large block of species, or $A = 2$ if $C = 6$. Suppose, finally, that some similar relation exists between the states of *D* and *B*.

A model of this kind will again generate null associations between states of *A* and *B* at the species level, when there is no causal relation between them (or between *C* and *D*). As *C* and *D* evolve, it will happen from time to time that a species will evolve a freezer state for both *C* and *D*. Then there will be a block of species descended from it that are likely to share the same state of *A* and *B* and a species count will find evidence of a nonexistent association. We picked the relation between *C* and *A* above such that any one of the four states (*AB*, *Ab*, *aB*, *ab*) could be frozen in this way. In practice, the evolution of the observed characters will be influenced by more than one other character, but only one character is needed to represent the essence of the problem; the effects of a realistically complicated multicharacter set of influences could be represented as some version of the "*C*" character above. At least, that is the design.

A model with explicit third variables controlling the observed variables, as well as being concrete about the evolutionary processes of interest, also makes clear why phylogenetic overcounting is indeed overcounting. A taxon with freezer states of *C* and *D* will have a number of species uniform for states of *A* and *B*. A test that counts species as independent trials will inevitably have erroneous, and liberal, Type I error rates. Maddison (1990), Harvey and Pagel (1991), and Pagel (1994) have suggested that large uniform taxa can be evidence of an association between the characters; but in our model, the safest interpretation is that the taxon has evolved a freezer state of an interacting variable.

We have used a null two-character model (with two observed characters controlled by two unobserved characters like *C* and *D*) to scrutinize the behavior of a number of proposed tests for discrete cross-species data. *C* and *D* evolve in a stochastic manner (like the observed characters in Harvey & Pagel's model). The same setup could be used to investigate Type II error rates if the model had a non-null relation between the observed characters. For example, *C* could set the state of *A* as above and then *A* and *D* could determine *B*. But we have not deployed the model in this form, nor have we investigated single general freezer characters like *E* above.

D. Type I error rates of some proposed tests for discrete characters

In Grafen and Ridley (1995) we obtained the Type I error rates for several tests. Here we only summarize some of our findings and emphasize interim practical conclusions for people working with real data. We stress, however, that our work allows no general recommendation of one particular test.

Discrete tests have been proposed by Ridley (1983), Burt (1989), Maddison (1990), Harvey and Pagel (1991), Møller and Birkhead (1992), Pagel (1994), and Grafen and Ridley (1995). We did not test Maddison's, Harvey and Pagel's, or Pagel's on the simulated data sets, mainly because those tests could be seen to suffer from phylogenetic overcounting. The effects of this kind of nonindependence are understood, and it would be easy to invent data sets in which they would behave arbitrarily badly. We did not test Møller and Birkhead's pairwise test, which is restricted to a certain kind of data set and is in important

respects a refined version of Burt's test. Here we shall also exclude the randomization test proposed by Grafen and Ridley (1995) because that test is not familiar in the existing literature, it takes lengthy explanation, and we are unsure how good a test it is. See Grafen and Ridley (1995) for explanation of that test, and Type I error rates for it. We also scrutinized the phylogenetic regression (Grafen 1989), to see how it fared with discrete data, and we did naïve species counts for comparison. Table 3.1 contains a simplified summary of the results for two tree shapes; the table footnote gives some details of the simulations.

We draw attention to the following. First, Ridley's and Burt's tests are highly biased against the ancestral state in the tetratomous phylogeny. The reason is the nonindependence described above: both tests (for different reasons) find nonancestral associations more easily than ancestral ones. In the extreme cases of a "star" character change tree, it is practically impossible for either test to find a significant association on the ancestral diagonal because there is a maximum of one data point (the center of the star) in the ancestral state. The character change trees of the tetratomous null data are probably not simple stars, but they do have a large shadow of influence of the ancestral state. Both tests, however, have reasonably good Type I error rates with the more realistically shaped "Hennig" tree, though Grafen and Ridley show that the reason for the improvement in Burt's test in the Hennig tree was not that given as a rationale for the test.

The results have similar implications for the practitioner of either test. The behavior of the test is influenced by two properties. One is the shadow of influence of the ancestral state. It is not difficult to look at the character states on the phylogeny and see at a glance whether the root state permeates it all, with most evolutionary events being single changes from it. Under parsimony, the more resolved the tree, the less the influence of the ancestral state; with complete phylogenetic ignorance (that is, we can say nothing better than that all the species are equally related, in a star phylogeny) the character change tree has to be a star. The simulations show that in one realistic tree with approximately realistic rates of change, the shadow of influence of the root state has been reduced sufficiently for the bias in the tests to disappear. The condition of extensive shadow of influence of the root state is the condition for these two tests to be biased.

Table 3.1. Type I error rates for four comparative methods with discrete data:

	Tree Shape							
	Tetratomy Association on:				Hennig Association on:			
	ancestral diagonal		nonancestral diagonal		ancestral diagonal		nonancestral diagonal	
Chance of finding *p* value more extreme than:								
	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
Test:								
Species counts	0.239	0.150	0.172	0.083	0.286	0.211	0.325	0.239
Ridley (1983)	0.017	0.006	0.706	0.589	0.055	0.014	0.061	0.011
Burt (1989)	0.000	0.000	0.219	0.072	0.031	0.000	0.014	0.003
Grafen (1989)	0.055	0.008	0.072	0.011	0.044	0.006	0.039	0.014

Results of 360 null data sets, each containing 256 terminal species. The datasets were generated by the model described in the text; the parameter values were picked to make the amounts and distribution of change approximately realistic; there were about 25–30 events in the 256 species tree. The tetratomous tree contained 256 species arranged symmetrically in four member groups through four levels (4^4=256). The "Hennig" tree was abstracted from Hennig's (1981) phylogeny for the insects and aimed to have an approximately realistic amount of symmetry and frequency of polytomous and dichotomous branching. Associations on ancestral and nonancestral diagonals have the following meaning. Imagine a 2x2 contingency table for the A/a and B/b characters. If the ancestral state in the simulation was AB, the ancestral diagonal is the one with AB and ab; the nonancestral diagonal is the one with Ab and aB. See Grafen and Ridley (1995) for details and the full *p*-distributions.

When they are biased, the thing to look out for is whether the root state counts for or against the hypothesis under test. If the ancestral state supports the hypothesis, the bias (when it exists) is towards conservativeness. The test can be so conservative that it is almost unusable, and this was Proctor's (1991) reason for developing an alternative to Ridley's (1983) method; but when a significant *p* value is obtained with the ancestral state supporting the hypothesis, the evidence is more significant than the *p* value alone suggests. In practice, the method has mainly been used in cases in which the ancestral state supports the hypothesis, and this is likely to continue to be the norm; our simulations confirm that in this case Ridley's (1983) test is conservative. However, when the ancestral state counts against the hypothesis, both tests are liberal and a significant *p* value would not be impressive evidence for a hypothesis if that ancestral state had extensive shadow of influence through the tree. Thus, if the ancestral state supports the hypothesis, the tests can be used with confidence, in the knowledge that any bias is only making it more difficult to find support; when the ancestral state counts against the hypothesis, the tests should be used with caution. We should also note that the test of Møller and Birkhead (1992) does not suffer from the same difficulties and should give valid results under mild assumptions. However, it does throw away data, and few data sets will be large enough and well enough resolved for it to be practical.

Second, the phylogenetic regression had reasonably valid Type I error rates in Grafen and Ridley's (1995) simulations. We regard these results as promising, and as expected from the theory of the phylogenetic regression (Grafen 1989); but they are not as definitive as they might appear. When running the method on the data sets, we fixed the branch lengths at their known and correct values. This is justifiable as a first stage in scrutinizing the method (if it does not work when fed the correct branch lengths, then that is the end of it); but it gives the method an advantage relative to the other tests, which were not supplied with so much information. However, an interim conclusion would be that the phylogenetic regression is probably a reasonable method to use with discrete data when we are willing to make assumptions about branch lengths.

Table 1 also reveal that the species counts are predictably awful. The 1% significance level, for example, is exceeded in about 10 to 20% of the null data sets. Naïve, prephylogenetic comparative tests should be kept at the other end of a barge pole.

Association or direction of change

The methods discussed above all test for associations between characters and leave open the direction of causation. However, a number of authors have suggested that methods that test associations are inferior to methods that test hypotheses about causal direction. Indeed, they form an orthodoxy in recent thinking about discrete comparative data.

Pagel and Harvey (1989) noticed that associative tests "conflate" changes from several evolutionary directions; Donoghue (1989, p. 1141) noticed as a "difficulty [that...] simply recording the number of times that a particular combination of states appears in a cladogram may obscure information on the sequence of character origination"; and Miles and Dunham (1993, p. 600), in the most recent review of the subject, noticed as a "weakness" of an associative test that it "ignores the sequence of transitions." Maddison (1990), whose paper inspired those remarks, more neutrally suggested that different methods are needed for different questions; but only the recent paper of Frumhoff and Reeve (1994) really stands out against this stream of argument. We agree with Frumhoff and Reeve's main point and shall incorporate it below. Ridley (1983, pp. 34–40) discussed earlier writing on the subject, including Gittleman's (1981) study of parental care in fish, using an ad hoc taxonomic method, as well as how to infer causation in some special cases.

Causal direction is less often discussed for continuous methods, perhaps because it is less clear how to infer, from inferred ancestral states, the causes of a correlation between continuous characters. Lande's (1979) paper on brain-body size allometry in mammals is one interesting example; he inferred that the correlation is more likely to have evolved from selection on body size, with brain size being dragged along in a correlated response, than vice versa. The inference follows from the heritabilities of brain and of body size, which have been measured in mice, and the observed variation in brain and body size; it is, however, more curious than convincing, because modern laboratory murine heritabilities are unlikely to be representative of the broad sweep of 60 million years or more of mammalian evolution.

Huey (1987) and Harvey and Pagel (1991, pp. 162–164), when discussing Huey, used the term "directionality" to describe another kind of inference with continuous comparative data. Huey (1987) inferred the direction of change of two continuous variables (running speed, preferred environmental temperature) from inferred ancestral to modern states. But his interest was in associations between those changes: he was not trying to infer which character changes had causal priority. The critical remarks that follow do not apply to Huey's kind of directional study, which should not be confused with the topic of our discussion. By a directional test, or hypothesis, we mean one that tests whether, or

claims that, changes in one character (e.g. $a \rightarrow A$) precede and drive changes in another ($b \rightarrow B$).

There are two reasons we discern for preferring a directional test. The hypothesis may be explicitly historical and concerned with which character came first. We, however, are interested in functional hypotheses and not in history; indeed, it is doubtful whether the uncertainty in historical reconstructions, by their nature unique events, is really statistical in nature. The second reason is that a hypothesis may be inherently directional, claiming that one character causes a change in another, but not vice versa. Five points can be made.

1. A directional hypothesis predicts both an association and a direction of change; but a directional test is likely to be more powerful. It is therefore desirable, insofar as truly directional hypotheses exist, for people to work on developing directional comparative methods.

2. However, no formally justified directional test is available. No tests have, for example, reached the stage of justification that we set out above, and the existing tests trip up at earlier hurdles (see below, point 3). Until such a test is available, the fact that directional hypotheses also make associative predictions means that they can be tested with tests of association.

3. We expect that a valid and powerful directional test would be more difficult to construct than an associative test. One reason is the problem of phylogenetic overcounting, the difficulty of which is illustrated by the way Maddison's (1990) concentrated changes test — one of the most carefully constructed and thoroughly thought out of comparative methods — nevertheless suffers from the phylogenetic overcounting kind of nonindependence, as Maddison (1990) himself almost admitted (Grafen & Ridley 1995 show that his arguments on this point are in error) and Sanderson (1991) and Sillén-Tullberg (1993) have also discussed. Harvey and Pagel's (1991) method also phylogenetically overcounts, or seeks to compensate for overcounting in an arbitrary manner, and no one has yet devised a directional method that avoids the problem. That is not to say it is insurmountable.

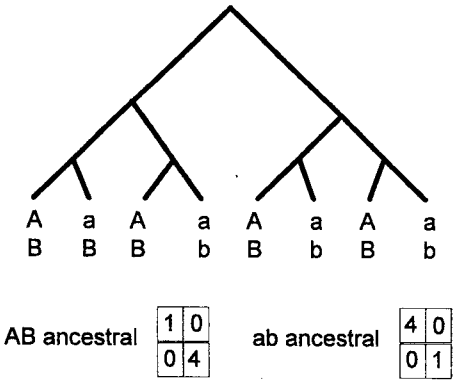

Figure 3. Influence of ancestral reconstructive error. Whether AB or ab is ancestral, the form and significance of the association is the same in Ridley's (1983) test. In this case there would be little or no effect on a directional test either, but in general directional tests are more sensitive to ancestral error than are associative tests.

The second reason is that directional methods will be more strongly affected by error in reconstructed ancestral states than are associative tests. This is essentially the criticism of Frumhoff and Reeve (1994), though they were not concerned to compare directional and associative tests. It also informs Höglund and Sillén-Tullberg's (in press) reply to Oakes (1992), in which they point out that Maddison's test, in Oakes' work, reconstructs many bird groups with lekking as the ancestral state even though it is a minority habit; such was Oakes' reason for concluding that sexual dimorphism often evolves after lekking. Höglund and Sillén-Tullberg suggest the reconstruction is erroneous and that sexual dimorphism is no more likely to evolve after lekking than after nonlekking.

Maddison's test reconstructs ancestral states in the phylogeny and counts the number of changes in the second character ($b \rightarrow B$ and $B \rightarrow b$) that occur in regions of the tree that are already a or A. Errors in the reconstruction of whether an ancestral state is A or a can have a large effect. The associative

tests, however, take a more relaxed view of reconstructed ancestral error. Ridley's (1983) test does reconstruct ancestral states throughout the tree, but it is often the case that many equally parsimonious reconstructions are consistent with the observations; then large changes in reconstruction, such as altering the state at the root, can have little or even no effect on the test. Figure 3 is an illustration. A reconstruction affects the test only to the extent that it alters the partitioning of species into subsets that share a state from a common ancestor, and even quite large changes in reconstruction may affect only one or two elements of such a partition. It would have a larger effect on the division of the interior of the tree into portions with one character state or another.

The other associative tests, strictly speaking, do not rely at all on ancestral reconstruction. Grafen (1989, p. 148) argued that the phylogenetic regression, although superficially appearing to condition on ancestry, actually conditions on modern pattern alone; Pagel's (1994, p. 37) unargued assertion to the contrary is in error. The apparent ancestral reconstruction is best understood as simply being part of the working: the method compares weighted averages with other weighted averages, which are chosen by knowing the phylogeny. However, it is also worth noting that the sampling error in these weighted averages is taken into account in the test. Even, therefore, if the weighted averages were interpreted as ancestral reconstructions, the method would not be assuming their correctness. Likewise, the apparent ancestral reconstruction in the randomization test of Grafen and Ridley (1995) and in Burt's (1989) test is only needed to find partitions in the data for purposes of randomization; the methods do not assume anything in particular about the history of the character states in those regions.

We are thus distinguishing three degrees of ancestral reconstructive dependency. Directional tests are likely to have the strongest dependency, and be most vulnerable to error in ancestral reconstruction. They use inference as data, in Felsenstein's (1984) phrase, in a strong sense. Of the associative tests, Ridley's (1983) test has the stronger dependency, but it is weaker than the directional tests; and the other class of

associative tests has only a weak and superficial dependency on ancestral reconstruction.

Simulations could reveal how much this argument matters. Maddison (1990, pp. 550–551) described some simulations to test the effect of parsimonious errors and found little evidence of damage, though there were imperfections. Maddison's simulations were a sensible method given the question he set out to answer; but two things are worth noting about them. One is that he held constant the areas of the phylogeny that had one or other state of the suppositious causal variable; he only randomized the suppositious caused variable. Reconstructive error is also possible for the causal variable, and it is error in that character that we are particularly concerned with in the argument given above. Secondly, Maddison's model of character change, like Harvey and Pagel's but unlike Grafen and Ridley's, had no "freezer" characters (like *C, D,* and *E* above). Every branch in Maddison's model had a separate independent chance of change, and therefore no taxa were frozen in a uniform state for an observed character. We should not expect problems with uniform taxa and reconstructive error in this model and doubt its usefulness for probing the problem. Thus Maddison's simulations are inadequate to contradict our logical point about the vulnerability of directional methods to reconstructive error. For the associative test of Ridley (1983), Grafen and Ridley (1995) obtained Type I error rates by simulation and, as we saw above, found that they were reasonably realistic in one phylogeny and conservative in another — which implies the method is robust to errors in ancestral inference. However, the relative vulnerability of directional and associative tests has not been measured in a direct comparison. Such an exercise would be premature at this stage because of the absence of a directional test that is free from non-independence due to phylogenetic overcounting.

4. Although many evolutionary hypotheses are expressed in a directional way and may seem at first sight to be directional, they may really be associative. That is, causation often works both ways and not just in one. Consider the link between gregariousness and aposematism among lepidopteran larvae,

analyzed by Sillén-Tullberg (1988) in a directional way, and discussed by Maddison (1990) and Harvey and Pagel (1991) as a directional hypothesis. The theory is that gregarious species may undergo selection for aposematism because predators that try one aposematic larva may then avoid its relatives.

However, there will be other selective forces at work. Aposematic species will sometimes undergo selection toward solitariness and gregariousness for other reasons. The aposematic species will have the balance tipped toward gregariousness because an aposematic species that is gregarious will experience less predation. The postulated selective force can therefore act in either direction and will be most appropriately investigated by testing for an association between gregariousness and aposematism. Many biological hypotheses are, we believe, of this form and — sometimes despite first appearances — make an associative and not a directional prediction.

This is a case in which the experimental analogy for comparative studies can mislead. In experiments there is usually unidirectional causality. If you add nitrogeneous fertilizer, the crops grow more — but that does not mean that if crops grow more for some other reason it will stimulate a nitrogeneous rain from the environment. Adaptive relations between two characters are rarely like that. If there is a selective equilibrium, in which two characters show an adaptive association, it will be quite peculiar for causality not to work in both directions. Sometimes a particular theory will suggest a hypothesis in a form that implies one causal direction rather than the other; but further thought (as in the gregariousness example) will often reveal that suggestion belongs to the context of discovery and not the logic of the hypothesis. Likewise, economy of exposition may constrain the proponents of a hypothesis to explain it as if it works in one causal direction; but an adaptive hypothesis in simplified expository form should not be confused with a truly directional hypothesis.

It may only be necessary to think about the two characters in the hypothesis by themselves in order to see that a truly directional hypotheses is implausible. That is, if *ab* and *AB* form adaptive associations, a little thought will usually show both that

a tends to change to *A* in species that are *B* and that *b* tends to change to *B* in species that are *A*. But the multifactorial nature of selective influences on characters makes directional hypotheses even less likely. The characters studied in comparative research are typically influenced by a number of other factors. The gregariousness example drew on this point: here is a further example.

Harvey and Pagel (1988) discussed the relation between mating system and sexual dimorphism as a directional hypothesis. If a species is polygynous it is more likely to evolve sexual dimorphism. The same idea was at work in the exchange between Höglund (1989), Oakes (1992), and Höglund & Sillén-Tullberg (in press). The causal tendency for polygynous species to evolve sexual dimorphism is well known and follows simply from the theory of sexual selection. However, sexual dimorphism is influenced by other factors too, including species recognition and ecology. If sexual dimorphism evolves for some other reason, it would probably influence the future evolution of the mating system. If the sexual dimorphism were such that males had a higher mortality rate, that would automatically direct the species down the evolutionary road to polygyny, simply through its effect on sex ratio.

Maddison (1990) introduced directional tests from a different perspective, that of release from constraint. Imagine the *A/a* character exerts a constraint on the evolution of another character (*B/b*), such that the evolution of *B* is possible in an *A* species, but *a* constrains things and makes the evolution of *B* difficult. Then in *A* zones there will be more evolution from *b* to *B*. The constraint may operate directionally in this manner, but that does not show the evolutionary relation between the two characters is directional; it could again be that thought about change in the other direction too would reveal that the system can move in either direction. If *AB* is an adaptive association it is likely that *A/a* will change in *B* species.

A similar point can be made about "permissive" comparative hypotheses. Pagel and Johnstone (1992) tested the relation between the *C* value of the DNA of a species and the cell cycle time. The two variables are known to be correlated, but Pagel and Johnstone argued this is more likely to be a permissive than a causally adaptive relation: if the cell cycle is longer for other reasons, more junk DNA will accumulate (in contrast to the idea that DNA increases in order to slow the cell cycle). Again, it is also theoretically possible that the process can work either way. If some species have lower, and others higher, *C* values, and some are selected for longer and others shorter cell cycle times, then the species with higher *C* values will be more likely to be the ones that end up evolving longer cell cycles.

It is not our purpose to deny the logical possibility of directional adaptive hypotheses. In other areas of evolutionary biology, directional ideas do exist, such as in the literature about Dollo's Law (Bull & Charnov 1985), and Godfray's (1987) work on parasitoid clutch sizes. However, the hypotheses that have been discussed in the literature about directional comparative biology seem to us to be associative. There are good theoretical reasons to think that most hypotheses about the adaptive relations between characters will really be associative, and the theoretical discussions of directional comparative methods have underestimated the generality of bidirectional causal influences in adaptive associations.

5. Frumhoff and Reeve (1994), crediting Armbruster (1992), make a further point, that a directional causal process, even a one-way example, will often produce an association in the observed species with real data. If *B* is selected for after *A*, and the selection pressures are not very weak, then all observed species will be either *AB* or *ab*. There will be no telltale *Ab* group whose location could reveal in what direction evolution occurred.

We conclude that directional comparative methods have yet to be developed and justified, face formidable technical and inferential difficulties, and have a narrower range of application, relative to associative methods. Associative methods exist and should at present be the methods of choice in discrete comparative inquiry.

Other statistical approaches

The procedure we discuss above for investigating a test is to generate random hypothetical data sets using a "data-generation process," and to apply the proposed test to many such data sets to discover initially Type I and later Type II error rates. This is the standard statistical approach, though often analytical work can remove the need for actual simulation. In this section we discuss two other justificatory methods that might be thought to apply to comparative statistics (Harvey & Pagel 1991; Pagel 1994; Felsenstein 1988). We shall mainly be concerned with likelihood theory and briefly mention bootstrapping. We discuss their relationship with the fundamental criterion just outlined and offer reasons why we believe they can be of limited if any use in the present state of statistical theory.

The older of the two alternatives is the likelihood approach, essentially introduced by Fisher (1922) and later much developed and enhanced. [Cox & Hinkley 1974, Chap. 2, particularly section 4(vii) and the bibliographic notes, is a discussion intended for statisticians.] The likelihood approach begins by writing down a data-generation process or statistical model. Instead of actually generating data and applying a proposed test to it, it is possible by applying standard techniques to derive a test from the data generation process. If certain assumptions are upheld, this "likelihood-ratio test" enjoys many enviable properties. It might therefore be thought that a test could be written down directly, and simulations avoided, and this was the approach of Pagel (1994). However, this is possible only if those assumptions are indeed upheld nearly enough, and we shall come to the question of whether they are.

Most standard tests can be derived as likelihood tests. For example, multiple regression and analysis of variance, and indeed the "general linear model," can be derived as likelihood tests. It is natural, therefore, when devising new tests, to attempt the likelihood approach. Cox and Hinkley (1974, pp. 47–48) express a consensus that the ultimate criterion, to be applied to all tests including those suggested by likelihood theory, is the repeated trials criterion, which we have applied using our simulated data sets.

As new tests have been devised for more complex problems, statisticians have increasingly realized that the straightforward likelihood approach is no panacea. Special sub-branches of likelihood

theory are devised to cope with specific problems: for example, partial likelihoods were invented by Cox (1972) for life table analysis (for a longer discussion, see Kalbfleisch & Prentice 1980). Let us turn to the kinds of difficulties that can lie in the way of standard likelihood theory. The following quotation from McCullagh (1991, p. 286) sets the scene:

> Classical likelihood-based analysis for complex problems may be intractable or impossible. In many instances this is due to the fact that a complete probabilistic specification of the model is not available or not suitable .... Related difficulties with standard likelihood-based analysis are the accommodation of possibly large numbers of nuisance parameters, and the choice of a reference set for computation of the relevant probabilities.

The first problem is the availability of a suitable complete specification of the model. In our case this means a data-generation process that we are prepared to accept as true. Thus acceptance of Pagel's (1994) likelihood method depends inter alia on acceptance of his model of character change. We gave reasons above for not agreeing with the model, so we would not accept his method. A broader point emerges from this need for an agreed model, and it is that we may never know enough to agree on a model. In simple statistical problems, the questions of interest can be answered by including only material of immediate interest. In more complex problems, answering a question (such as whether there is a functional relationship between two variables) may necessitate including in the model a lot of information of no direct interest (such as the phylogeny, its topology, and branch lengths). We may need to agree that the model for the whole lot is precisely right in order to agree that the test for the small part of interest is a good one. Any parameters of the model (unknown quantities that require estimation from the data, such as the intercept and slope in simple linear regression) that are not of direct interest are the "nuisance parameters" that McCullagh refers to in the quotation above.

In the case of discrete comparative methods that reconstruct character states at higher nodes, those reconstructions are nuisance parameters. Now the assumptions under which likelihood tests are justified do allow for nuisance parameters, but there is an important restriction. The general result about likelihood tests is asymptotic, and for our informal purposes we may think of it in the form "As the number of data points increases, the likelihood test becomes as close as

you like to validity." (For a more technical treatment, see Stuart & Ord 1991, particularly pp. 658–661 and p. 870.) Now the restriction is that the number of nuisance parameters should be fixed as the number of data points increases. In a regression problem where an x-variable must be controlled for, this is easily accommodated. The one extra nuisance parameter is the coefficient for that extra variable, and the likelihood test for the coefficient of interest is thus justified. In phylogenetic problems, the number of ancestral states to be estimated increases as the tree increases. Indeed, the number of ancestral states to be estimated remains in all interesting cases a sizable fraction of the number of data points. This matters greatly. The essence of the justification of likelihood tests is that there is enough information to estimate the nuisance parameter as precisely as we like. We cannot assume in phylogenetic problems that the nuisance parameters of ancestral states can be estimated as precisely as we like. They will in general remain poorly estimated in data sets of any size.

How is Pagel's (1994) method affected by these considerations? His likelihood function does not have nuisance parameters, but a structural aspect of the likelihood function creates a parallel difficulty. A likelihood function in a typical simple case is a product of terms, with each term corresponding to one data point. Further, the terms are all defined in the same way, and differ from each other only because the values of the observations differ. Once logged, such a likelihood function is a sum of terms, where each term has the same relationship to its own data point. Readers may recall that the central limit theorem deals with sums of identically distributed random variables, and indeed the asymptotic behavior of likelihoods is derived from the central limit theorem by, for example, Stuart and Ord (1991, pp. 658–661).

Pagel's (1994) likelihood function is not a simple product. Instead it is recursive sum of products of sums of products and so on, in a structure conforming to the phylogeny. When we increase the size of a data set in an ordinary case, we simply add extra terms to the sum representing the log-likelihood. The asymptote, as the data set increases, then corresponds to the asymptote in the central limit theorem that guarantees normality of the sum. But when we increase the size of a phylogenetic data set, the structure of Pagel's likelihood changes and the analogy with the central limit theorem breaks down. Pagel envisages an asymptotic justification for the test, but it would need to be established what kind of asymptote was being considered. Specifically, what shape the phylogeny would have for each different-sized data set. Not only would standard results probably not apply, but the conclusion that the likelihood ratio has the required chi-squared distribution might well not be true. The recursive nature of Pagel's likelihood function therefore means that standard likelihood theory does not apply, and his method is unjustified.

Pagel suggests a simulation test whose validity he claims would not depend on asymptotic arguments. It involves estimating parameters for the characters on the basis of maximum likelihood and on the assumption of the null hypothesis of independent evolution. Then simulated data sets are created using those parameter values, and the distribution of the likelihood ratio test built up empirically by repetition of this process. We remark that this test, although it involves maximum likelihood and simulations, has no logical justification in terms of statistical theory, and Pagel offers no simulational evidence that it is a valid test. The reasonable-sounding procedure is dubious if the distribution of the likelihood ratio depends strongly on the parameter values. If it does not, it hardly matters whether they are fitted by maximum likelihood or not. If it does, it is not clear why the distribution generated by the maximum likelihood estimates is of sole concern. Nearby parameter values may well have given rise to the data, and may have importantly different distributions of the likelihood ratio. The question which needs to be answered to evaluate this simulation version of Pagel's test is, therefore, whether the distribution of the likelihood ratio under the null hypothesis depends much on the parameters.

The difficulties of handling likelihood methods in the presence of nuisance and incidental parameters are discussed by Wetherill (1986, sec. 13.4). On page 279, Wetherill singles out the related problem of reconstruction of phylogenies as being "even more slippery" than the already problematic example he is discussing. We did begin work on a likelihood based test. It included nuisance parameters representing freezing of character states in parts of the phylogeny. Two remarks are worth making about it. The nuisance parameters reintroduced into the problem all the complexities that have to be grappled with in the standard approach, so the choice of approach does not change the core of difficulties that have to be overcome. Indeed its behavior most closely resembled that of Ridley's (1983) method among those

discussed in this paper. Second, the reason we abandoned this test, despite our attraction to its model of character change and freezing, was that there is no extant statistical justification for a likelihood test that necessarily involves so many nuisance parameters.

It follows that the likelihood approach to comparative statistics must await possible new theoretical developments before it can be relied upon. The possibility should not be discounted; as we mentioned above, the theory of partial likelihoods was developed by Cox (1972) to help construct methods for the now burgeoning field of tests for survival data, but at present it is only a possibility and we know of no relevant work on such an extension.

One possible attraction in using the likelihood approach is that the reconstruction of a phylogeny could be combined with testing a functional hypothesis in a grand likelihood scheme (e.g. Pagel 1994, pp. 38 and 42). We believe this would not be a helpful approach. Modern statistics does not tend to treat complicated problems with straightforward likelihood methods. Indeed an important strand in the recent history of statistics has been a sophistication of likelihood methods to handle special kinds of nuisance and incidental parameters that arise in particular kinds of applications. A grand likelihood test would be at least as technically dubious as its constituent parts are in the present state of likelihood theory.

Bootstrapping might be thought to be a second alternative approach. At present, the theory of bootstrapping is well worked out only for cases in which there is a sample whose members can be assumed to be independently drawn from some population (see for example Wu 1988). In inferring phylogenies, Felsenstein (1985) has taken the characters to be the random sample. The independence of molecular characters is certainly more plausible than the independence of ordinary phenotypic characters. In our case, however, the characters are fixed. The usual approach to bootstrapping is to resample the data points (i.e., species). However, it is not obvious what phylogeny to assume for a set of species, and furthermore the assumption of independence is not even close to being met. This probably explains why, to our knowledge, no one has yet suggested applying bootstrapping methods to the comparative method.

The approach we have taken to scrutinizing the behavior of a test is to form a data generation process or model, generate many data sets, and

try the test out on those data sets. The alternative approaches of likelihood theory and bootstrapping lack statistical foundation in the complex setting of phylogenetic problems. The attraction of the methods would be superficial, for they gloss over rather than solve real statistical difficulties in the problem. At present, any test suggested by those two approaches needs to be scrutinized by the logically prior and statistically routine method which we adhere to in this chapter.

References

Armbruster, W. S. 1992. Phylogeny and the evolution of plant-animal interactions. *BioScience*, 42, 12–20.

Bull, J. J. & E. L. Charnov. 1985. On irreversible evolution. *Evolution*, 39, 1149–1155.

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

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

Cox, D. R. 1972. Regression models and life-tables. (with discussion). *J. R. Stat. Soc. B*, 34, 187–220.

Cox, D. R., & Hinkley, D. V. 1974. *Theoretical Statistics*. London: Chapman and Hall.

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

Felsenstein, J. 1984. Review of "The explanation of organic diversity" by M. Ridley. *Nature*, 308, 565.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution,* 39, 783–791.

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

Fisher, R. A. 1922 On the mathematical foundation of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222, 309–368.

Frumhoff, P. C., & Reeve, H. K. 1994. Using phylogenies to test hypotheses of adaptation: A critique of some current proposals. *Evolution.*, 48, 172–180.

Gittleman, J. 1981. The phylogeny of parental care in fishes. *Anim. Behav.*, 29, 936–941.

Godfray, H. C. J. 1987. The evolution of clutch size in parasitic wasps. *Am Nat..*, 129, 221–233.

Grafen, A. 1989. The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B*, 326, 119–157.

Grafen, A., & M. Ridley. 1995. Statistical tests for discrete cross-species data. *J. Theor. Biol.* (submitted)

Harvey, P. H., & M. D. Pagel. 1991. *The Comparative Method in Evolutionary Biology*. Oxford, England: Oxford University Press.

Höglund, J. 1989. Size and plumage dimorphism in lek-breeding birds: a comparative analysis. *Am. Nat.*, 134, 72–87.

Höglund, J. & B. Sillén-Tullberg. (in press). Does lekking promote the evolution of male biased size dimorphism in birds? On the use of comparative approaches. *Am. Nat.*.

Huey, R. B. 1987. Phylogeny, history, and the comparative method. In: *New Directions in Ecological Physiology* (Ed. by M. E. Feder, A. F. Bennett, W. W. Burggren, & R. B. Huey), pp. 76–98. Cambridge, England: Cambridge University Press.

Kalbfleisch, J. D., & R. L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution*, 33, 402–416.

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.

McCullagh, P. 1991. Quasi-likelihood and estimating functions. In: *Statistical Theory and Modeling* (Ed. by D. V. Hinkley, N. Reid, & E. J. Snell), pp. 265–286. London: Chapman and Hall.

Miles, D. B., & A. E. Dunham. 1993. Historical perspectives in ecology and evolutionary biology: the use of phylogenetic comparative analysis. *Annu. Rev. Ecol. Syst.*, 24, 587–619.

Møller, A. P., & T. R. Birkhead. 1992. A pairwise comparative method as illustrated by copulation frequency in birds. *Am. Nat.*, 139, 644–656.

Oakes, E. J. 1992. Lekking and the evolution of sexual dimorphism in birds: comparative approaches. *Am. Nat.*, 140, 665–694.

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

Pagel, M. D., & R. A. Johnstone. 1992. Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the *C*-value paradox. *Proc. R. Soc. Lond. B*, 249, 119–124.

Pagel, M. D., & P. H. Harvey. 1989. Comparative methods for examining adaptation depend on evolutionary models. *Folia Primatol.*, 53, 203–220.

Proctor, H. 1991. The evolution of copulation in water mites: a comparative test for nonreversing characters. *Evolution*, 45, 558–567.

Ridley, M. 1983. *The Explanation of Organic Diversity*. Oxford, England: Oxford University Press.

Sanderson, M. J. 1991. In search of homoplastic tendencies: statistical inference of topological patterns in homoplasy. *Evolution*, 45, 351–358.

Sillén-Tullberg, B. 1988. Evolution of gregariousness in aposematic butterfly larvae: a phylogenetic approach. *Evolution*, 42, 293–305.

Sillén-Tullberg, B. 1993. The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution*, 47, 1182–1191.

Stuart, A., & J. K. Ord. 1991. *Kendall's Advanced Theory of Statistics*, 5th ed. London: Edward Arnold.

Wetherill, B. G. 1986. *Regression Analysis with Applications*. London: Chapman and Hall.

Wu, C. F. J. 1988. [Contributor to discussion of papers by Hinkley and Diciccio and Romano.] *J R. Stat. Soc. B*, 50, 338–354.