J. theor. Biol. (1997) **184,** 7–14

A New Model for Discrete Character Evolution

ALAN GRAFEN[†] AND MARK RIDLEY[‡][§]

[†]*Department of Plant Sciences, South Parks Road, Oxford, OX*1 3*RA, U.K. and the* [‡]*Departments of Anthropology and Biology, Emory University, Atlanta, Georgia* 30322, *U.S.A.*

(*Received on* 14 *February* 1996, *Accepted on* 14 *February* 1996)

The paper provides an explicit justification for the principle that a uniform taxon should contribute only one datapoint in comparative analyses with discrete variables. The justification is that phylogenetic patterns in variables unincluded in the proposed test vitiate the assumption of independence, both at the level of species and at the level of branch segments. The consequence is that a uniform taxon cannot safely be counted as more than one datapoint. The arguments use a branching discrete Markov process in continuous time, with the new feature that the tested variables are only a subset of the evolving characters. This model is proposed as a useful criterion for measuring the merit of proposed tests, and illustrates the necessity for models in evaluating comparative methods.

1. Introduction

Most statistical tests rely on an assumption that all datapoints in an analysis are statistically independent. It has been recognized since Darwin (see Ridley, 1992) that comparative data do not meet this assumption. See Martins (1996) for a recent collection of papers on the topic. Species are generally more similar to congeners than to species that are merely in the same family, which are in turn more similar than species that belong only to the same order. It is the fact that this hierarchical pattern of similarity is to be expected that renders inappropriate the application of statistical tests that ignore phylogeny. The problem arises in a different guise for continuous and discrete characters. Here we are concerned only with the discrete case.

Ridley (1983) originally introduced the practice of counting each uniform taxon once in comparative tests. His reasoning was that the character states of species were not independent, but that independence could be found at a different level. He reconstructed the changes in character states, and inferred on which branch segments change took place. Ridley reasoned that the types of change that took place at those different places were independent of each other.

Later authors (Maddison, 1990 (on which see also Sanderson, 1991 and Sillén-Tullberg, 1993) and Harvey & Pagel 1991; Pagel, 1994) have drawn back from the principle that a uniform taxon should count only once, and have done so with the justification of models of evolutionary change. The purpose of the present paper is to look more closely at the kinds of independence that can occur in the evolution of character change, to construct a new model containing what we believe to be an essential component of non-independence for biological plausibility, and to argue that this more sophisticated model implies the "uniform taxon principle". Our work has been to use models to make more explicit the reasoning behind Ridley's (1983) arguments, and to reaffirm his conclusions.

An extra advantage of this model is that it can be used to create datasets against which different proposed statistical tests can be judged. Employing models does not, therefore, only allow reasoning to be made more explicit, it is also helpful in moving

[§]Present address: Department of Zoology, South Parks Road, Oxford OX1 3PS, U.K.

0022–5193/97/010007 + 08 $25.00/0/jt960269

towards the goal of having statistically justified tests for discrete comparative data.

2. The Statistics of Cross-species Data

2.1 A NEW FORM OF NON-INDEPENDENCE

Although the character states of phylogenetically related species are non-independent, the non-independence can be generated by evolutionarily independent events at a deeper level. A phylogeny shows the links between species. Each path segment that directly connects two nodes represents a species in a particular period of history. Independence of changes in different path segments leads to non-independence between species. The reason is that congeneric species share most of the path segments in their evolutionary history, and have only a few that are different. Distant species may share few or no path segments.

This underlying independence at the path-segment level is the basis of the one extant model of discrete character change, introduced into the comparative literature by Harvey & Pagel (1991); they cited the earlier results of Diamond & May, 1977; and see more recent developments by Pagel, 1994 or Maddison 1994). If independence at this level is accepted, then the uniform taxon principle is unnecessary, and, accordingly, that principle has not been applied in the methods suggested by Harvey & Pagel (1991), Pagel (1994), Maddison (1990) and Sillén-Tullberg (1993). In their methods, a more speciose uniform taxon provides stronger evidence than a less speciose one. Read & Nee (1995) have argued that these methods assume that all lineages with the same character state have the same chance of changing to another state—which is implausible and the methods therefore suffer from "pseudoreplication of lineage-specific factors".

Independence at the path-segment level is indeed biologically implausible. Rather, we expect when looking at a phylogenetic reconstruction of a discrete character to see changes clustered in particular parts of the tree. There is no statistical method extant by which we can put this expectation to the test on particular characters, but we adduce two lines of persuasive but not definitive argument in support. First, one of us (MR) has performed several comparative analyses, and the strong impression is that changes are clustered in the tree. Bell (1982, p. 339) reports clustering in the character of asexuality; for example asexuality is rare in beetles as a whole but has arisen many times in the Curculionidae. Another suggestive example is the multiple independent origins of eusociality in the Hymenoptera compared to other insects made famous by Hamilton (1964). It should be an uncontroversial claim among those who carry out comparative tests that changes do cluster. The second line of argument is a general expectation on biological grounds that changes will cluster. The ecological situations that influence selection pressures may be taxonomically clustered. For example, changes in adaptations to salinity will be common in estuarine and riverine fish, and rare in deep-ocean and freshwater fish. So long as some taxa are exclusively deep-ocean and others are exclusively freshwater, changes will not affect them and so will be clustered elsewhere. Another example is that polygyny is probably more fixed in mammals than in birds, because of underlying relatively fixed physiological characteristics of mammals such as female viviparity and lactation. Therefore, we should expect changes in polygyny to be less common in mammals and more common in birds.

We shall assume in what follows that changes in characters are clustered on the phylogeny, although as we have emphasized the case cannot at present be made watertight. It is in principle a matter of fact whether changes are clustered for single characters. A full defence of models based on independence at the path-segment level would require a demonstration that changes are not clustered.

2.2 IMPLICATIONS OF THE CLUSTERING OF CHANGES

If the changes in single characters are indeed clustered, it would have important implications for statistical tests of associations between characters. The logical mirror image of the clustering of changes is that higher taxa are more likely to be completely free of changes than if changes were unclustered. It follows that it is more likely that two unrelated characters will be uniform on a taxon than if changes were unclustered. Although on average the association will cancel itself out, because different taxa will be uniform but with different combinations of character states, the clustering of changes increases the "lumpiness" of the apparent association between unrelated characters. The increased lumpiness is parallel to discovering that animals in an experiment are divided into litters—a test that assumes independence at the individual level will discover statistical significance too frequently.

This argument shows that Harvey & Pagel's (1991) model assumes that changes are unclustered, and so the justification they offer for their test relies on that assumption. Maddison's (1990) method also bears consideration here. He reconstructs two discrete characters, say A and B, and in his test takes as given the reconstruction of A. Then he randomly reassigns on to the tree as many changes in B as are actually present in the real reconstruction of B. The numbers of changes of both types ($B = 1$ to $B = 2$, and $B = 2$ to $B = 1$) are recorded in each of the two subsets of the tree ($A = 1$ defines one subset, and $A = 2$ defines the other). By comparing these numbers (derived for each of many random reassignments) with the corresponding numbers in the real reconstruction of B, Maddison finds a significance level to test the null hypothesis that A and B are unrelated. Maddison's randomization procedure implicitly embodies the assumption of independence at the path-segment level. On our view, the changes even in the null case should be clustered within the phylogeny. We believe that the assumption of independence at the path-segment level will cause Maddison's procedure to yield statistical significance too often, unless changes really are not clustered in the phylogeny. The reason is that he will too seldom find no changes within a taxon uniform for A. The assumption of independence will lead him to expect results closer to the average than clustering would produce.

3. The New Model

3.1 THE NEW MODEL FOR ONE CHARACTER

Harvey & Pagel's (1991) model works as follows. Each character has a given initial state at the root of the tree. It then has a probability of changing per unit time along each path segment, as it evolves down the tree, dividing at each higher node. In the original form, there was one probability, α, of a change from $A = 1$ to $A = 2$, and another instantaneous rate of change, β, from $A = 2$ to $A = 1$. Pagel (1964) has generalized this to more than two character states. Independence, more precisely conditional independence (meaning conditional on current state), rules out any tendency for changes to cluster.

To represent clustered change, and include non-independence at the path-segment level, we have modified Harvey & Pagel's (1991) model as follows. We assume that a Harvey–Pagel process goes on for an underlying variable, say C, and that C determines the value of A, typically in a many-to-one fashion. For example, if C has 4 states we might have a rule of the form:

C	A
1	1
2	1
3	2
4	2

The biological reasons we gave above for expecting clustering referred to variables causally underlying the variable actually being observed. This table defines how an underlying variable (C) affects an observed variable (A). Interest lies in the implication of independence in the evolution of C for the observed distribution of A.

Suppose that the matrix of instantaneous rates of change for C is as follows:

	To 1	To 2	To 3	To 4
From 1	−0.01	0.01		
From 2	0.01	−0.21	0.2	
From 3		0.2	−0.21	0.01
From 4			0.01	−0.01

A species in state 1 will tend to stay in state 1, occasionally switching into state 2. A species in state 4 will tend to stay in state 4, occasionally switching into state 3. However, species in states 2 and 3 will tend to switch between these two states, occasionally relapsing into 1 or 4. We shall subsequently refer to states from which change is improbable as "freezer states", and states 1 and 4 are examples here.

Changes in C will have the property of conditional independence. Changes in A, on the other hand, will not have the property of conditional independence. The change of $A = 1$ switching to $A = 2$ will depend on whether the underlying variable C has state 1 (in which case the rate of switching is very low) or state 2 (in which case the rate of switching is high). The changes in A will be clustered.

3.2 THE NEW MODEL FOR TWO CHARACTERS

We can construct a model for the simultaneous evolution of two discrete characters by assuming that each follows an independent process of the kind just described, and this is a natural null hypothesis. There will be two underlying variables, say C underlying A and D underlying B. It is also possible to generalize this assumption slightly to produce a model that contains both the null hypothesis and a non-null hypothesis in which the two characters A and B are related. The generalization assumes that B is determined by D alone, but that A is determined by C and by B. A trivial case of joint determination, trivial because it represents the same situation as separate determination, corresponds to the null hypothesis. This can be shown as:

	$B = 1$	$B = 2$
$C = 1$	$A = 1$	$A = 1$
$C = 2$	$A = 1$	$A = 1$
$C = 3$	$A = 2$	$A = 2$
$C = 4$	$A = 2$	$A = 2$

A more interesting case is the non-null situation in which B does affect A, for example

	$B = 1$	$B = 2$
$C = 1$	$A = 1$	$A = 1$
$C = 2$	$A = 1$	$A = 2$
$C = 3$	$A = 2$	$A = 2$
$C = 4$	$A = 2$	$A = 2$

This provides a model in which (1) all four combinations of A and B states can occur, (2) there is an explicit stochastic element generating the data, and (3) there is a deterministic rule producing the A and B states. It is an alternative to Pagel's (1994) extension of the Harvey–Pagel model in which different rates are introduced for different transitions in multi-character states.

We have discussed here a relatively simple case with four states of C and of D. A third-variable character with four states can have two freezer states in which the observed character changes rarely, and another two states to allow more rapid evolution. For some other purposes a larger number of states is necessary. Suppose we expand the system from four to six states as follows:

$D = 1$	$B = 1$
$D = 2$	$B = 1$
$D = 3$	$B = 2$
$D = 4$	$B = 1$
$D = 5$	$B = 2$
$D = 6$	$B = 2$

There could be four subsets of characters, $\{1\}$, $\{2,3\}$, $\{4,5\}$, $\{6\}$, with transitions rare between, but common within, subsets. The two subsets $\{2,3\}$ and $\{4,5\}$ that give rise to variation could have different transition probabilities, which would allow for the possibility that one subset tended to have mainly $B = 1$ with some $B = 2$, and the other mainly $B = 2$ with some $B = 1$. This is only one possible development of the model.

In summary, the parameters of the model are as follows:

(1) The initial states of C and D, at the root.
(2) The instantaneous transition rates for C and D.
(3) The rule determining B from D.
(4) The rule determining A from B and C.
(5) The branch lengths.

When the model is used as a data generation process to scrutinize a statistical test, the data will be the observed states of A and B at the species tips. The only parameter of interest is (4), and within that rule only the effect B has on A. The other parameters are clearly inestimable from the data. A good statistical test will extract information about whether B determines A in a way that is reasonably robust against any values of the other parameters.

A computer package written in Mathematica (Wolfram, 1988) that implements this model is available from the first author. Requests should be accompanied by a Macintosh formatted floppy disk. One additional feature of the model is its treatment of polytomies, described in the Appendix below.

3.3 IMPLICATIONS OF THE MODEL

The major implication of the model is that it justifies the uniform taxon principle. If a taxon is uniform for A and B, it may be because each of the underlying variables is in a state in which change is extremely unlikely. The subtaxa of the uniform taxon then do not provide independent information about the association between A and B, and should not be treated as if they do. This overcounting is simply avoided by adopting the uniform taxon principle: in comparative analysis, a uniform taxon should count as one datapoint only.

There are many parameters in the model that are not of direct interest, such as the transition probabilities for C and D, and the branch lengths. A good test should be one that works well whatever the values of those parameters. Indeed, in real applications all that will be known is the values of A and B at the species tips. In particular, the nature of characters C and D will be unknown, as will the number of states they may belong to. A good test must therefore also work robustly with many different numbers of states for the underlying C and D characters. This kind of robustness is likely to be aided by principles of data reduction that exclude data of a kind that can be evaluated only in a highly model-dependent way. The uniform taxon principle is exactly the kind of data reduction that is needed for robust inference.

Pagel's (1994) approach could, in principle, be further generalized to incorporate clustering, by allowing the rates of evolution to vary across the phylogeny. (We are grateful to an anonymous referee for drawing this possibility to our attention.) A statistical approach that aimed to estimate all the parameters in a model would produce very different results when applied to this extension to Pagel's model, as compared with our model. In both Pagel's case and ours, the "estimate all parameters" approach would result in a complex test. The difference between the cases is unsatisfactory, because the two models are different idealizations of the same system, and we are unlikely ever to be able to tell which, if either, is right in any application. The complexity is unsatisfactory, because the test would be highly dependent on which particular form of a model was used. The uniform taxon principle, on the other hand, amounts to an attempt to discard information that might be contaminated by clustering; it avoids the need to rely on complex representations of exactly how clustering operates.

3.4 A BIOLOGICAL ILLUSTRATION

To clothe our conceptual skeleton with a little flesh, we now look at an example that we believe biologists in general will admit would lead to erroneous conclusions if independence at the path-segment level was assumed.

Consider the mating systems of birds and mammals; monogamy and polygyny might be two states of the observed character B ($B = 1$ for polygyny and $B = 2$ for monogamy, for example). Polygyny is probably more fixed in mammals, by its interaction with all the physiological characters of viviparity and lactation. $D = 1$ might stand for this combination of other reproductive character states that freeze the mating system; note the clear causal relation by which D determines B, with female viviparity and lactation tending to lead to the evolution of polygyny. In a polygynous bird, such as the red-winged blackbird, the mating system is probably evolutionarily more labile; the underlying factors (expressed in the model by $D = 2$) that led to its evolution can readily change and switch the mating system to monogamy. The character states represented by $D = 2$ perhaps include a habitat structure that provides territories of limited number and variable quality; but the detail is not the issue here. In a real biological example there will probably not be just one character D controlling the states of the observed character B; there will be several. Lactation and viviparity are a whole suite of character states.

Consider now another character that is likely to have a null relation with the mating system. Birds and mammals have "single" circulatory systems; aortal circulation happens to be left-handed in mammals and right-handed in birds. The circulatory pattern is probably evolutionarily frozen, partly because of its association with the other components of the anatomy and embryology of the circulatory system. In warm-blooded amniotes there is an association at the species level between a left-handed circulation and polygyny (in mammals), and a right-handed circulation and monogamy (in birds). The relation is probably not adaptive; monogamy and polygyny are equally plausible whichever direction blood is carried out of the heart. (At any rate, readers may allow such a possibility in a merely illustrative argument.) In mammals, one character state $B = 1$ (polygynous mating system) has been frozen by its association with one set of character states D (rest of reproductive system) and another character state $A = 1$ (left-handed circulation) has been frozen by its association with another set of character states C (rest of circulatory anatomy and embryology); birds are mainly $B = 2$ and $A = 2$, due to other states of C and D. The end product is a large clade with a certain character association, simply because the states of the two characters have been independently evolutionarily frozen.

This illustration shows that assuming independence at the path-segment level leads to a false evaluation of the strength of evidence that circulatory pattern and mating system are related in warm-blooded amniotes.

3.5 PUNCTUATIONISM AND GRADUALISM

The model as presented allows each path segment in the phylogeny to be given an arbitrary length, and this allows it to represent various assumptions about how evolution proceeds. If times of speciations are known, then each path segment's length could be assigned as the duration between the two speciations that delimit the segment. This may be thought of as a gradualist assumption.

On the other hand, various steps towards punctuationism can be made. Each path segment could be assigned equal length, and this is what punctuationism has meant to Harvey & Pagel (1991, p. 159) and others. If more information were available about extinct species, it would presumably be right on the punctuationist view to make the length of each path segment in the phylogeny of current species proportional to the number of speciation events that took place on it. Closer adherence to the theoretical suggestions of Eldredge & Gould (1972) would require us to know which offspring species was the parent species unaltered, and which had undergone rapid evolutionary change and deviated from the parental form. With this knowledge, one of the path segments below each speciation event would be set to zero, and the other could be set to one. However, things are more complicated still. There is no reason in the punctuationist approach to suppose that each burst of rapid evolutionary change should be of roughly equal size. It is possible, for example, that the longer a species has been constricted by genetic homeostasis, the larger the change that will result when the constraint is released. The extra information required to approach more closely to the punctuationist hypothesis will usually not be available. In view of extinct species, and the possibility that the magnitude of change at speciation is proportional to the time since the previous genetic revolution, it may well be that in our ignorance, a better approximation to true punctuationism is to be found in the superficially gradualist assumption that change is proportional to duration, than in the assumption that every path segment has equal length.

One important point emerging from this discussion is that the capacity to set the length of each path segment separately makes the model very general, and capable of representing a wide range of assumptions about the mode of evolutionary change.

4. Why Models Matter

This paper so far illustrates one important advantage of models. It has allowed us to dissect the reasoning behind different authors' arguments, to develop it, and to make what we believe to be an important point about a desirable property of statistical methods for discrete data.

Models have another rôle in statistics, which has so far failed to make an appearance in the literature on discrete comparative methods, though it has for continuous methods (Grafen, 1989, Miles & Dunham, 1993). They can be used to create artificial datasets, in which the truth or falsity of the null hypothesis is known, and therefore to subject proposed tests to a measurement of Type I and Type II error rates. To measure Type I error rates performance under the null hypothesis is examined, and a test is said to be "valid" if it produces 5% significance 5% of the time (and in general x% significance x% of the time). To measure Type II error rates the statistical power is examined, i.e. the ability of a test to reject the null hypothesis when it is in fact false. Because there are usually many different alternative hypotheses, and because a test must be valid to be acceptable, it is usual to concentrate first on Type I error rates, and we confine our discussion to them.

Conducting these simulations, employing a model, is, in principle, the only way to justify a statistical method (analytical proofs are mathematical constructions to show that the simulations would produce the correct Type I error rates, so avoiding the need to perform them). It follows that any claim that a statistical method is valid, or acceptable, or reliable, implicitly requires a model. How have authors of discrete comparative methods justified their proposed tests?

Maddison (1990) used no explicit model, but he did employ a set of "axioms" that embodied properties he believed a reasonable model would have. Specifically, he assumed independence at the level of path segments, and constructed his method so it would work with any model satisfying this axiom. As we have seen, we disagree with this axiom, but the form of his argument is correct and clear. Burt (1989) and Møller & Birkhead (1992) employ axiomatic arguments, though more implicitly than Maddison did.

Harvey & Pagel (1991) produced an explicit evolutionary model, but then used it only for discussion. They did not go on to apply it to see whether their proposed method had the properties they claimed for it.

The failure in all cases to pursue the justification fully, either to a formalized set of assumptions with a proof of correct Type I error rates, or to a simulation in which an explicit model creates artificial datasets, has contributed to the present uncertainty in the subject. There is a plethora of claims and counterclaims, with very few substantiated. The new model that we describe above was developed by us specifically to measure the Type I error rates of a number of existing methods (Ridley & Grafen, 1996; Grofen & Ridley, 1996).

We have argued that models are extremely desirable in the justification of methods proposed for general use. That does not mean, however, that all comparative research that lacks such a model is worthless. Quite the contrary. Biologists often invent individual, or *ad hoc*, tests to deal with particular datasets. Proctor (1991), for example, noticed that her data contained a feature that made existing techniques impossibly conservative, and accordingly invented a method that tests the trend in the data better than any of the available general methods could do. These kinds of *ad hoc* methods can be convincing without any reference to a formal statistical model, and they are a sensible research strategy while properly justified general methods have not been developed. However, if a method is to be recommended for general use, the justification cannot be tied to a particular dataset. An abstract model, and null data, become necessary.

Most earlier work that has aimed to justify general methods for discrete characters has lacked any model and been forced to offer verbal arguments about the merit of proposed tests: these arguments can be of varying quality and are not useless, but they are no substitute for manifest Type I and II error rates. Readers should be sceptical of claims made for the validity of proposed tests unless Type I error rates are provided, on the basis of a model that incorporates the biological assumptions they believe to be necessary.

Other models of discrete character change exist that produce no phylogenetic structure. The two main methods of simulating random character states considered by Maddison & Slatkin (1991) were to assign character state 1 or 2 with equal probability to each terminal taxon, and to shuffle at random the terminal taxa with character states 1 or 2.

This does, as Maddison and Slatkin say, make sense if the characters have a very rapid rate of evolution; but it would be an unrealistically rapid rate for most (or even all) comparative research, because evolution would be so rapid that the data had no phylogenetic structure at all.

5. Conclusions

Changes in a discrete character are likely to be clustered in certain parts of the phylogenetic tree, so that there is no independence at the level of path segments. We have developed a model of discrete character change that reflects this clustering, and conclude that statistical tests for discrete comparative data should follow the "uniform taxon principle" of Ridley (1983), that a taxon uniform for a character should count as only one datapoint in the analysis. Ridley's own method embodied this principle. Later authors have discussed the principle, but strayed from it in their own methods. The justification given here will, we hope, encourage an uncompromising adoption of the uniform taxon principle.

Models are useful for pursuing conceptual points such as this, and also for conducting explicit measurements of Type I error rates. The uncertain state of discrete comparative methods at present is partly due to a general reluctance to use models in this way. Our new model includes all the essential biological properties of which we are aware, and so we recommend its use for judging proposed statistical tests.

REFERENCES

Bell, G. (1982). *The Masterpiece of Nature*. London: Croom Helm.

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

Diamond, J. M. & May, R. M. (1977). Species turnover rates on islands: Dependence on census interval. *Science* **197,** 266–270.

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

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

Grafen, A. & Ridley, M. (1996) Statistical tests for discrete cross-species data. *J. theor. Biol.* **183,** 255–267.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *J. theor. Biol.* **7,** 1–52.

Harvey, P. H. & Pagel, M. M. (1991). *The comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Maddison, D. R. (1994). Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. *Annu. Rev. Entomol.* **39,** 267–292.

Maddison, W. P. (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics* **5,** 365–377.

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.

Maddison, W. P. & Slatkin, M. (1991) Null models for the number of evolutionary steps in a phylogenetic tree. *Evolution* **45,** 1184–1197.

Martins, E. P. (ed.) (1996). *Phylogenies and the Comparative Method in Animal Behavior*. New York: Oxford University Press.

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

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

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

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

Read, A. F. & Nee, S. (1995). Inference from binary comparative data. *J. theor. biol.* **173,** 99–108.

Ridley, M. (1983). *The Explanation of Organic Diversity*. Oxford: Clarendon Press.

Ridley, M. (1992). Darwin sound on comparative method. *Trends Ecol. Evol.* **7,** 37.

Ridley, M. & Grafen, A. (1996). How to study discrete comparative methods. In: *Phylogenies and the Comparative Method in Animal Behavior* (Martins, E. P., ed.) pp. 76–103. New York: 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. (1993). The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution* **47,** 1182–1191.

Wolfram, S. (1988). *Mathematica*: *A System for Doing Mathematics by Computer*. New York: Addison-Wesley.

APPENDIX

One feature of the model is unimportant to the main points made in the text, but is essential for other reasons in a model suitable for evaluating proposed tests. It concerns polytomies. A proposed test should be evaluated on the basis that a polytomy in the phylogeny represents not certainty that simultaneous branching took place, but rather uncertainty about the order of branching. In the sense of Maddison (1989), polytomies should be assumed soft not hard. The model answers this point by taking a random compatible binary refinement (in the sense of Grafen, 1989) of each polytomy. A new refinement for each polytomy is selected each time a dataset is created. Selection in the model is done as follows (this differs from the method described by Grafen, 1989). Consider a node with n daughters, where $n > 2$. The daughters are randomly re-ordered, and then divided into two subsets by placing the first i daughters into one subset and the remaining $n - i$ into the second. i is chosen to take values from 1 to $n - 1$ with equal probability. The process is applied recursively if either of the subsets has more than two members. The result is a dichotomous tree that is a random compatible refinement of the original polytomy.