*J. theor. Biol.* (1995) **173,** 99–108

# Inference from Binary Comparative Data

ANDREW F. READ† AND SEAN NEE‡

† *Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, U.K. and* ‡*Department of Zoology, University of Oxford, Oxford OX1 3PS, U.K.*

(*Received on* 5 *July* 1994, *Accepted on* 18 *July* 1994)

A variety of null hypotheses have been used to test for associations which might be construed as evidence of biologically interesting relationships between binary character states exhibited by taxa. These models assume that particular regions of a phylogenetic tree are independent with respect to their probabilities of character evolution, but they differ in the regions they specify. Early analyses specified terminal taxa (usually species); more recent developments have specified all or a subset of branches, or infinitesimally short sections of branches. Yet the central problem for comparative biologists is that branches throughout a phylogeny are not independent with respect to their evolutionary possibilities. Tests which assume that they are may indeed provide evidence for non-random association between characters according to the particular model of randomness used, but do not necessarily provide the basis for rationally inferring the existence of a biologically interesting link between the characters. In particular, they suffer from pseudoreplication of lineage-specific factors.

By way of contrast, we resurrect in this context a different model of randomness, the random assignment of treatments, which we argue provides a rationally acceptable basis for inference. States are assumed to be randomly assigned amongst sister taxa exhibiting different states of both variables. This model allows that the probabilities of character evolution vary throughout the tree, but does not require that these be specified, nor assumptions to be made about how evolution occurs. We illustrate this approach with reference to controversial associations between (i) warning coloration and larval gregariousness in butterflies, for which we find some support, and (ii) hybrid fitness and heterogamety (Haldane's Rule), for which we find no support, in contrast to Ridley's method which demonstrates the opposite Rule.

## 1. Introduction

There are two types of question one can ask about the association of character states of two traits ($x$ and $y$) over a range of species. First, "is $x$ non-randomly associated with $y$?" and, second, "is $x$ connected to $y$?" The second question concerns the issue of whether the comparative data provide evidence for a biologically meaningful or interesting connection (either mechanical or evolutionary, direct or indirect) between $x$ and $y$. An affirmative answer to the first question can only be used to address the second if we are persuaded that our ability to draw reasonable inferences is not seriously compromised by the way in which the assumptions of our model of randomness are likely to be violated.

Here we are concerned with models of randomness used to demonstrate biologically interesting associations between binary comparative variables. It is not our primary intention to present yet another comparative method, but rather to consider the foundations on which such tests are constructed. We begin by discussing the central problem to be overcome: non-independence of cross-taxa data (Clutton-Brock & Harvey, 1977; Ridley, 1983; Felsenstein, 1985; Grafen, 1989; Harvey & Pagel, 1991; Harvey & Purvis, 1991). This problem is well known in the context of cross-species analyses, and arises because numerous factors (such as morphology, developmental patterns, environments, history) ensure that different lineages are likely to exhibit rather different probabilities of possessing particular states of $y$ for

0022–5193/95/050099 + 10 $08.00/0

reasons that may have nothing to do with $x$. We argue that existing techniques developed to overcome phylogenetic non-independence in the analysis of binary data do not, in fact, do so any better in principle than do cross-species correlations: incorporating phylogenetic information does not in itself make a method a reasonable one for inference. We go on to suggest a different basis for the analysis of this type of data, which we believe is sensible for null hypotheses in comparative tests of this sort.

## 2. Non-independence of Cross-species Data

Consider a hypothetical clade in which there is a perfect association between the distribution of two binary variables $x$ and $y$, so that all taxa with $x1$ exhibit $y1$ and all with $x2$ exhibit $y2$ [Fig. 1(a)]. An across-species analysis using Fisher's exact test provides evidence of a significant association between the characters across species. Such an analysis has randomly shuffled the character states amongst the species in order to create the null distribution. This assumes that the species each had an equal and independent probability of exhibiting the states (subject, of course, to the conditionality constraints of the test). Under this assumption, the result is indeed evidence of a non-random association between $x$ and $y$. On its own terms, this analysis is perfectly acceptable. However as biologists, we would be reluctant to infer from this analysis that there is a biologically interesting relationship between the two variables, because the model of randomness is considered unacceptable. We know that a particular state of $y$ is more likely to be shared by closely related species than by species drawn at random from the whole sample. Why this should be so does not directly concern us here (for suggestions, see Gould & Lewontin, 1979; Dobson, 1985; Felsenstein, 1985; Grafen, 1989; Harvey & Pagel, 1991), but there is no doubt that it is so. The probability of a species having feathers, for example, clearly depends on whether that species is a bird, a reptile, a mammal or a plant; Donoghue (1989: 1148–51) describes several well-documented but less obvious cases for plant traits. Just as we would not add

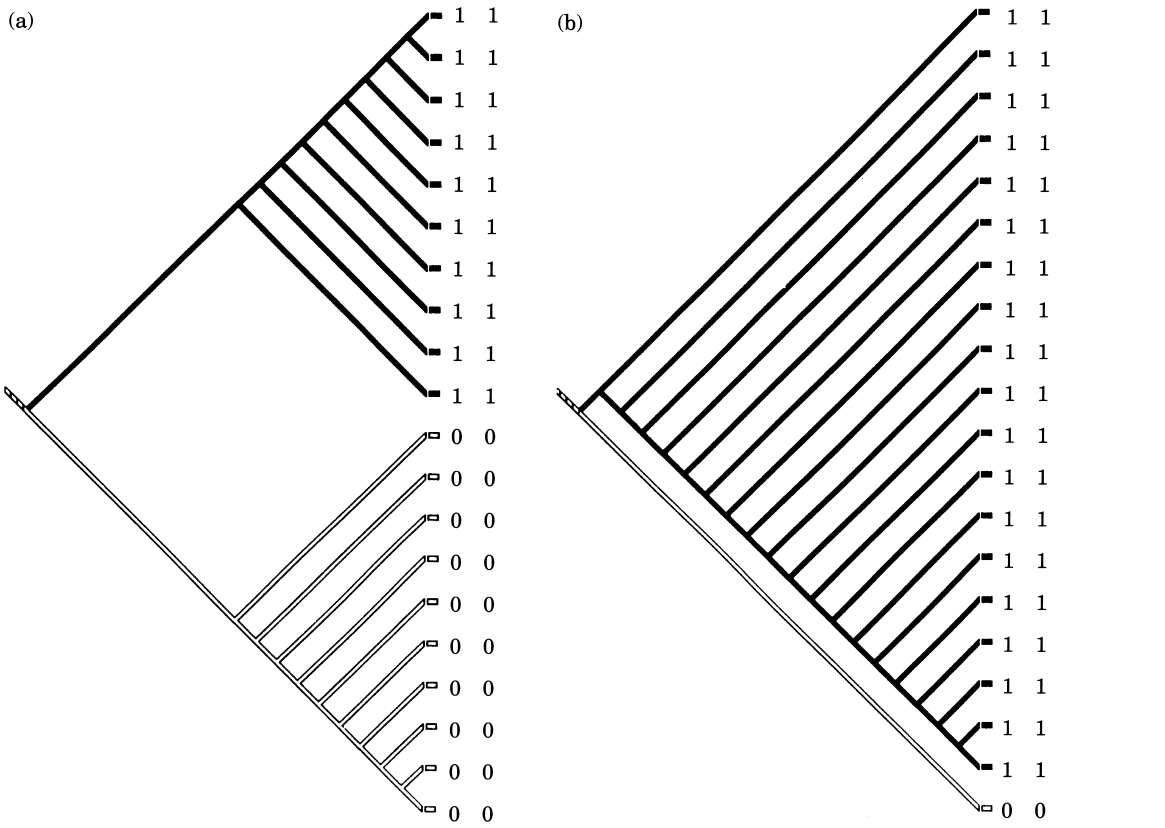


FIG. 1. Hypothetical phylogeny showing states of two binary characters. In both cases, two major clades are represented and the same number of terminal taxa have been sampled, but in (a), equal numbers of species have been sampled in two clades, whereas in (b), one of those clades is represented by only one species, with the remaining species coming from the other clade. Ancestral character states reconstructed by standard procedures (Maddison & Maddison, 1992). Shaded branches, $x = 1$, $y = 1$; open branches, $x = 0$, $y = 0$; hatched branches, equivocal. Branch lengths are arbitrary.

repeated samples of species-specific traits exhibited by *individuals* to a cross-species analysis, *ad hoc* sampling of species from the same lineage does not add independent trials relevant to a comparative test. To do so is to repeatedly sample the multitude of lineage-specific factors which affect the probability of the event in that lineage. This is true whether or not the trait in question is evolving repeatedly within that taxon (e.g. following each speciation event) or has evolved only once in that taxon and been maintained for whatever reason. Species are still not independent with respect to any lineage-specific factors affecting the state of $y$. For example, if the taxon consisted of two equally speciose sister taxa, which differed in the state of both $x$ and $y$, the across-species test would demonstrate a between-lineage effect. There is, however, no rational basis for inferring from that that it is due to any *particular* difference between lineages. Intuition or other information may suggest such an inference, but that sort of comparative evidence provides no rational basis for it. Inferences based on such analyses are in effect based on pseudoreplication (*sensu* Hurlbert, 1984). Just as we would not wish to test the effects of a fertilizer on plant biomass by comparing repeated samples from a single fertilized plot in the UK with those from a single unfertilized plot in the US, so too should we be wary of implicitly and unjustifiably equating taxon-specific effects with particular traits of that taxon.

Relaxing the assumption that the probabilities of exhibiting character states are equal (and developing some rational basis for assigning these probabilities within taxa) does not rescue across-species correlations, since the assumption of independence remains. As Felsenstein pointed out (1985: 13–14), all members of a clade may adaptively respond in the same way to some environmental change, for example, because of some shared feature which has nothing to do with the characters under investigation. Given this, the assumption of independence itself is not acceptable, and clade bushiness will arbitrarily determine probability values.

The central problem of comparative biology is thus to avoid the problems of phylogenetic non-independence and it is that problem which is the subject of this paper. Modern comparative biologists typically take great pains to identify phylogenetically independent units for analysis which would provide a rational basis for inferring evidence of biologically interesting non-random associations. Having identified such units, attempts are frequently made using those units to statistically control for the effects of confounding (third) variables (Harvey & Pagel, 1991). How that should best be done is an interesting but quite different issue to that on which we are focused: how to deal with non-independence introduced by phylogenetic similarity.

The basis of recent attempts to deal with non-independence of binary comparative data reviewed by Harvey & Pagel, 1991) is Ridley's (1983) critical insight that the fundamental unit of analysis for this type of comparative data should be evolutionary events or transitions (here defined as the gain or loss of states of a binary character). Methods based on this insight have the common feature that they employ more complex models of randomness utilising phylogenetic information. But "taking into account" phylogenetic relationships does not in itself provide a sound basis for rational interference. Here we argue that all contemporary methods actually extend the assumptions of the basic cross-species model up the phylogenetic tree from the species level. Their instigators and others have frequently worried that their underlying null models might not provide a suitable basis for rational inference, but have suggested that, nevertheless, their approaches are less likely to lead to erroneous conclusions than cross-species analyses. We will now argue that this is not obviously true.

## 3. Recent Developments

Ridley (1983) and Maddison (1990) attempt explicitly to incorporate phylogenetic structure into their null hypotheses. They do this by reconstructing ancestral character states on branches throughout the tree from the observed states of terminal taxa (using techniques reviewed by Harvey & Pagel, 1991; Maddison & Maddison, 1992). [Note the problems of independence we are discussing arise even if we could know the character state of ancestral nodes without estimating them from species values (cf. Harvey & Purvis, 1991; Oakes, 1992).]

Ridley's approach (Ridley, 1986: 1854; Pagel & Harvey, 1988: 419; Harvey & Pagel, 1991: 84) is to pay attention only to those branches on which there has been a change in the state of one or both characters. The end states of those branches are counted and the number of times each of the combinations of the two characters appears is entered into a $2 \times 2$ contingency table to test for a non-random association. In contrast, Maddison's method (termed the "concentrated-changes" test—Maddison & Maddison, 1992) is based on a model of randomness developed by designating one trait as the independent variable and then randomly distributing the observed changes in the state of the dependent variable across the tree. Evidence that a particular state of $y$ is found on branches with a particular state of $x$ more often

than expected by chance alone is taken as evidence that $x$ is connected to $y$ in some biologically interesting way (perhaps, for example, that a specified state of $x$ selects for or enables the evolution of a particular state of $y$).

A number of variants of these tests have been proposed (Proctor, 1991; Sillén-Tullberg, 1993), but all involve the same general approach and the points we make below apply equally to them. The tests of Maddison and Ridley or their derivatives have been used to investigate biological questions of the type "is $x$ connected to $y$" by Ridley (1983, 1986, 1988), Sillén-Tullberg (1988, 1993), Donoghue (1989), Höglund (1989), Maddison (1990), Proctor (1991), Oakes (1992), Sillén-Tullberg & Møller (1993), Martins (1993) and Brookfield (1993) amongst others.

The tests of both Maddison and Ridley randomize character states across branches, but they diverge in which branches are included in the randomization. Maddison's test incorporates all branches; Ridley's just those with change in one or both characters. Having decided which branches to include, both tests make the critical assumption that each of the branches they designated as relevant in the phylogeny has an equal and independent probability of exhibiting those changes or states. In their own terms, each test does identify non-random distributions of character states. But just as randomizing states across branches leading to terminal taxa (usually species) is inappropriate for comparative analysis because the probability of change on them is not equal and independent, branches higher in the tree also differ in their probabilities of change for the same reasons. Consequently, neither approach fully overcomes the central problem of comparative biology: both can incorporate the pseudoreplication of lineage-specific effects.

Consider, for example, a taxon in which there is no change in $x$ but $y$ changes from an ancestral to a derived state on $n$ independent occasions. That taxon would then contribute $n$ data points to the final contingency table in Ridley's test, each of which would be assumed to be independent. Yet the state of $x$ has not independently arisen in each of them: they are the same. The use of branches with changes in the state of one variable but not in the other can merely repeatedly sample the same taxon-specific state of the unchanged variable. Anything *at all* that members of that taxon share, as well as their state of $x$, could be responsible for changes in $y$. For example, eusociality has arisen many times in the Hymenoptera (Seger, 1991). But these do not represent many independent trials of the hypothesis that haplodiploidy promotes the evolution of eusociality: they represent one.

As an extreme case of this difficulty, consider Haldane's Rule. This states that where there are sex differences in hybrid fitness, the least fit sex will be the heterogametic sex (Haldane, 1922). A large number of crosses have been performed and the vast bulk of those support the rule. However, comparative analysis of these data using Ridley's approach supports the opposite rule! Almost all modern data show that, with few exceptions, birds and butterflies have female heterogamety and female hybrid inviability, whereas in all other sampled taxa males have those traits. (Read & Nee, 1991, 1994; Wu & Davis, 1993). Thus only two branches have change in either character and are therefore relevant to Ridley's method: those leading to birds and butterflies. These are consistent with the rule, but with just two data points, significance can not be reached. But what of the few "exceptions" to the general pattern? Heterogametic sex is invariant within the taxa for which crosses have been performed, but the sex with greatest hybrid unfitness varies somewhat, so that in a minority of crosses the least fit sex is the homogametic sex. According to Ridley's test, many of these "exceptions" should count as independent evolutionary events to be added to the contingency table on the diagonal contradicting Haldane's rule. In fact, despite the hundreds of crosses consistent with Haldane's rule, as few as five independent reversals in birds and butterflies, and five anywhere elsewhere would, according to Ridley's method, be sufficient to produce evidence directly contradictory to Haldane's Rule. At least that many "exceptions" do exist (Wu & Davis, 1993). One conclusion might be that the effects of heterogamety on hybrid inviability are indeed opposite to that proposed in Haldane's Rule and apparently contradicted by the vast bulk of existing comparative data. This conclusion might perhaps be best described not as counter-intuitive, but as counter-rational. In any case, reanalysis using the null model we advocate below provides no support of a significant association between heterogamety and hybrid viability (Read & Nee, 1991, 1993).

Problems of non-independence underlie similarly counter-rational conclusions which can emerge from Maddison's test. Consider a correlated-changes test of the taxa in Fig. 1(a) (as performed by *MacClade*—see Maddison & Maddison, 1992): as it stands, there is an equal probability that the single transition in one of the characters will appear on a branch exhibiting a particular state of the other. Quite rightly, the test emphasises that it could be a single accident of history that a transition to $y2$ occurs on an $x2$ branch, even though the resultant pattern would yield a significant result in an across species analysis. However, if the

same total number of species had been sampled but this time only a single species had been sampled in one of the taxa with the remainder coming from the other taxon [Fig. 1(b)], there would now be, according to the concentrated-changes test, a significant non-random association between $x$ and $y$ ($p = 0{\cdot}026$). This illustrates that the test is primarily influenced by the topology of the tree, rather than the independent information in it, and that, as with the cross-species analysis, a single accident of history can by this method become the basis for statistical inference of a biologically interesting link between two traits.

Several of the difficulties of the concentrated-changes test discussed by Maddison himself stem directly from the problematic nature of the assumption that all branches on a phylogeny are born equal and independent. First, the test does not distinguish independent replicates of taxon-specific traits from repeated independent origins of associations between traits (Maddison, 1990: 555). For example, the test would judge as equally improbable a few transitions in a dependent variable occurring only in a single small clade characterized by a particular character state and nowhere else in a very large tree, and the same number of transitions coincident with the same (and equally rare) character found in several widely separated regions in a tree with the same topology.

Second, the final probability value is sensitive to the inclusion and exclusion of taxa (Maddison, 1990: 554). According to Maddison (p. 555, see also Sanderson, 1991: 357; Harvey & Pagel, 1991: 111), "the decision as to how inclusive or detailed to make the tree for such an analysis depends on the extent of knowledge of the groups and whether all are enough alike to be considered as having 'all else [more or less] equal'". That there are taxa which are sufficiently unalike to be excluded emphasises our point. And exclusion is at best an attempt to categorize a continuum of variation in the probability of change; opinions about which taxa should be included can differ widely and affect the conclusions reached (e.g. Sillén-Tullberg, 1993; Brookfield, 1993). Indeed, if we had the sort of knowledge required to make such decisions on a rational basis, we would probably already know so much about the characters that a comparative test would be unnecessary.

Finally, branches in a tree are likely to represent variable periods of time. Several authors have suggested that evolution may be more likely on a longer branch (e.g. Pagel & Harvey, 1989; Maddison, 1990, Harvey & Pagel, 1991; Maddison & Maddison, 1992), and procedures based on Maddison's concentrated-changes test which attempt to take this into account exist (Pagel & Harvey, 1989; but see Maddison & Maddison, 1992: 315). Actually, there may be more evolutionary events on shorter branches (e.g. Peterson & Burt, 1992), but in any case, branch length is but one variable affecting probability of change, as Pagel and Harvey's development of Maddison's method makes explicit (Pagel & Harvey, 1989; Harvey & Pagel 1991). Its importance relative to other lineage-specific factors is hard to determine, but even were branch length of primary importance, there is every reason to suspect it would be so in a lineage-specific manner.

A set of comparative methods for discrete characters have been developed from attempts to take branch lengths into account [Pagel 1994, Milligan, in preparation (see acknowledgements of Pagel, 1994)]. These are based on Markov transition reasoning, lucidly described by Harvey & Pagel (1991: Section 4.6.1) for example, which has been used in models of island biogeography (Diamond & May, 1977) and neutral molecular evolution. In the present context, this reasoning makes the assumption (analogues of which are quite reasonable for island biogeographic processes and neutral molecular evolution) that evolutionary events at each infinitesimal time interval on each branch in a phylogeny are dependent only on the character state or states at the previous time interval and the probabilities of transitions between states, and are independent of everything else. Such methods assume that, following bifurcation, two sister taxa follow an independent evolution and that the transition probabilities modelling their evolution are the same as those of other taxa. The points made above during our discussion of the methods of Ridley and Madison apply equally to this approach. In addition we make the following observation. Methods for analysing binary data based on Markov processes bear a superficial similarity to modern comparative methods for the analysis of continuous traits based on, for example, Brownian motion models of character evolution (e.g. Felsenstein, 1985; Grafen, 1989; Pagel, 1992). There is a fundamental difference, however. In the case of continuous traits, it is the *differences* between sister taxa which evolve subsequent to their bifurcation which are taken as being independent items of information. In the case of these discrete methods, the characters of extant taxa are considered to be independent items of information, even if they are the same. It was the recognition that such an assumption leads to fallacious inference which originally prompted the development of comparative methods.

## 4. An Alternative Approach: Principles

A different approach, which we have used elsewhere (Read & Nee, 1991, 1993) and here more fully

develop and justify, allows that the probabilities of evolutionary change vary throughout the tree, but does not require that we specify either them or a model of evolution in order to develop an appropriate model of randomness. We suggest that this approach offers the possibility for a rational basis for inference from discrete comparative data. It closely parallels the approach used for all recent attempts to control for phylogenetic non-independence in the analysis of continuous variables (what Grafen, 1989, terms "the radiation principle"; Felsenstein, 1985; Grafen, 1989; Harvey & Pagel, 1991; Harvey & Purvis, 1991). In this section we discuss the principles involved; in the Appendix we give a worked example.

In medical experimentation, it is understood that when there is wide variation in personal characteristics that is relevant to the subject under inquiry, one desirable experimental design is a matched pair experiment, in which pairs of similar patients are constructed and each member assigned a different treatment. With comparative "experiments" we face the analogous problem of wide, and unknown, variation in taxon characteristics relevant to the subject under inquiry. Fortunately, we can fall back on the same solution, a matched pair design, since the lineages (or clades) on either side of a node constitute a natural matched pair: until they diverged they shared everything. Of course, we are not suggesting that evolution could have allocated treatments on a truly random basis; had it done so we would have evidence of causality not just correlation. Nevertheless, by using such a null model, one can identify phylogenetically independent points for analysis. The same approach is frequently used in medical research where experiments are impossible (e.g. pairing smokers and non-smokers matched as closely as possible for other factors, and then determining which member of the pair gets cancer). Comparative methods are discussed in the general context of inference from non-experimental data by Nee *et al*. (1995).

Given a set of matched pairs, one now forms the null hypothesis that, for each pair, the state of $x$ is irrelevant to the state of $y$. Using the reasoning of matched pair randomization tests (for a full account of this approach to hypothesis testing, see Cox & Hinkley, 1992), we suppose that, in a pair with different states of both $x$ and $y$, the observed configuration of the binary traits has a probability of 0.5. This is because, according to our null hypothesis, each taxon might have been "assigned" by evolution the other's state of $x$, with no effect on the evolution of $y$. Thus, for the data relevant to Haldane's Rule, birds and mammals are sister taxa, and under the null model, there was an equal chance of female heterogamety being assigned to Aves (the taxa exhibiting female hybrid unfitness) or to Mammalia (the taxon exhibiting male hybrid unfitness).

Only bifurcations which exhibit variation in both $x$ and $y$ affect the statistical evaluation of the hypothesis. This is a desirable feature. In a matched pair medical study, both patients in a pair might get cancer perhaps because they were both 102, or erroneous pairing might mean both members of a pair smoke. In neither case does that pair provide information about the possible differences between the two treatments which is the subject of inquiry. If both sister taxa share the same value of $x$ although they vary in $y$, this tells us simply that there are factors other than $x$ involved in the evolution of $y$, which we already knew. If both exhibit the same $y$ although they vary in $x$, it is impossible to determine whether this is because $y$ is unaffected by $x$, or because any of the large number of features that they share reduce the possibility of change in $y$ to zero.

Thus, in contrast to both Ridley and Maddison, we are focusing not on branches as the units of comparative analysis, but rather on sister taxa differing in the expression of both $x$ and $y$. Pseudoreplication of underlying taxon-specific factors is avoided because, under the null model, the alternate states of $x$ are randomly allocated in each sister-taxon comparison, and are thus independent of what is happening elsewhere in the tree. In practice, the approach is similar to matched-pair analyses (Pagel & Harvey, 1988) used to test for associations between continuous traits (e.g. Felsenstein, 1985; Burt, 1989) and between continuous and binary traits (e.g. Brown, 1961; Krebs *et al*., 1989; Read, 1991), although the null models underlying those tests frequently make additional assumptions. Indeed, generalizing from matched-pair tests of continuous variables, Burt (1989: 42) suggested a test for binary comparative data which, in practice, is similar to the approach derived from the null model we advocate here.

When implemented in practice, our approach, like all comparative methods, requires assumptions about how evolution works with which one may or may not agree (Harvey & Pagel, 1991; Harvey & Purvis, 1991). However, in contrast to other approaches, these assumptions are used only in generating the data (phylogenetic reconstruction and comparison of the relative change in $y$ in sister taxa). The model of randomness we use to overcome the fundamental problem of comparative biology makes assumptions which are unlikely to be precisely true (like all models), but these are not assumptions about how evolution occurs.

We note that as always with correlations, associations revealed by our test have many potential interpretations. In particular, third variables may be responsible, so that $x$ and $y$ covary because they are jointly but independently affected by $z$. But even were an association revealed by our test generated in that way, it would not be a consequence of taxonomic non-independence; it would not be coincidental. The effects of $z$ on $x$ and $y$ would have to happen on many different (independent) occasions in order to generate an association. In principle, the effects of third variables can be tested directly, but only once appropriate units for comparative analysis have been identified.

Finally, comparative biology is not just about single tests implemented without thought. Sister taxa which share the state of one or both traits are, for the reasons outlined above, irrelevant to the specific test. However, such sister taxa may be of interest for other reasons. For example, a hypothesis which purports to be relevant to a large clade (e.g. vertebrates) but which draws support from sister taxa in only a small part of the tree (e.g. birds) may have less generality than is claimed. Similarly, clades that have some sister taxa differing in both traits and others that differ in only one trait may be used to investigate other factors which affect the evolution of $x$ or $y$, and they may allow the investigation of possible third variables influencing any correlation between $x$ and $y$. Note, however, that these nodes remain irrelevant for inferring a non-coincidental correlation between $x$ and $y$ in the first place.

## 5. Too Conservative?

Critics may argue that our approach is more conservative than those of Ridley or Maddison because we may be using less of the tree in our test (e.g. daughter taxa that do not differ in both $x$ and $y$ do not affect the final $p$-value). The same sort of criticism is frequently used against attempts to root out pseudoreplication in ecological experiments (Hurlbert, 1984), and of course statistical methods should not be compared only in terms of the apparent degrees of freedom they generate. More importantly, our approach makes explicit that less of the tree is relevant to testing for the effect of $x$. Where there is a correlated lack of change in both traits, there is no way to distinguish between a lack of change in $x$ in any other lineage-specific factors as the reason for the lack of change in $y$. Ridley's approach avoids that sort of pseudoreplication (requiring as it does, the inclusion of branches with changes in at least one character), but by considering branches in which, for instance, $x$ changes but $y$ does not, it runs the risk of counting replicates of the same state of $y$ in the final test. In principle, it may be possible to incorporate information on lack of change into a model on which to base statistical inference if one knew enough about determinants of trait evolution. However, we suspect that if one knew enough to justify that, one would not bother with a comparative analysis in the first place.

It may often be the case that the comparative data cannot conclusively support or refute an evolutionary hypothesis, and we must accept that the natural experiments of evolution are often too poorly "designed" for us to draw rational inferences from their results. However, the use of appropriate null models has the potential to provide a lens on the world which can demonstrate relationships which might not be obvious in other tests. For example, using appropriate methods for the analysis of continuous data, Nee *et al*. (1992) found relationships between body size and abundance not apparent across species and, as we discuss in the Appendix, reanalysis of Sillén-Tullberg's (1988) data provides evidence for an association between aposematism and gregariousness not uncovered by the concentrated-changes test.

J. Brookfield's assertion that our analysis of Haldane's Rule (Read & Nee 1991) was inconsistent with contemporary comparative methods prompted this paper. We thank N. Barton, J. Brookfield, K. Fowler, P. Harvey, M. Pagel, L. Partridge, M. Ridley, A. Skorping and M. Whitlock for comments, and R. May and J. Maynard Smith for their support. Order of authorship arbitrary. We are funded by BBSRC Research Fellowships.

## REFERENCES

Brookfield, J. (1993). Haldane's rule is significant. *Evolution*, **47,** 1885–1888.

Brown, H. S. (1961). Differential chiasma frequencies in self pollinating and cross pollinating species of the genus *Gilia*. *El Aligo* **5,** 67–81.

Burt, A. (1989). Comparative methods using phylogenetically independent contrasts. In: *Oxford Surveys in Evolutionary Biology* Vol 6. (Harvey, P. H. & Partridge L. eds) pp. 33–53. Oxford: Oxford University Press.

Clutton-Brock, T. H. & Harvey, P. H. (1977). Primate ecology and social organisation. *J. Zool*. **183,** 1–33.

Cox, D. R. & Hinkley, D. V. (1992). *Theoretical Statistics*. London: Chapman and Hall.

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

Dobson, F. S. (1985). The use of phylogeny in behaviour and ecology. *Evolution* **39,** 1384–1388.

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

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat*. **125,** 1–15.

Felsenstein, J. (1988). Phylogenies and quantitative methods. *A. Rev. Ecol. Syst*. **19,** 445–471.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon.

GOULD, S. J. & LEWONTIN, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205,** 581–598.
GRAFEN, A. (1989). The phylogenetic regression. *Phil. Trans. Roy. Soc. Lond. B* **326,** 119–157.
HALDANE, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12,** 101–109.
HARVEY, P. H. & PAGEL, M. D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.
HARVEY, P. H. & PURVIS, A. (1991). Comparative methods for explaining adaptations. *Nature, Lond.* **351,** 619–624.
HÖGLUND, J. (1989). Size and plumage dimorphism in lek-breeding birds: a comparative analysis. *Am. Nat.* **134,** 72–87.
HURLBERT, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54,** 187–211.
KREBS, J. R., SHERRY, D. F., HEALY, S. D., PERRY, V. H. & VACCARINO, A. L. (1989). Hippocampal specialization of food-storing birds. *Proc. natn. Acad. Sci. U.S.A.* **86,** 1388–1392.
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. & MADDISON, D. R. (1992). *MacClade: Analysis of Phylogeny and Character Evolution. Version* 3.0. Sunderland, MA: Sinauer Associates.
MARTINS, E. P. (1993). A comparative study of the evolution of *Sceloporus* push-up displays. *Am. Nat.* **142,** 994–1018.
NEE, S., READ, A. F., GREENWOOD, J. J. D. & HARVEY, P. H. (1992). The relationship between body size and abundance in British birds. *Nature, Lond.* **351,** 312–313.
NEE, S., READ, A. F. & HARVEY, P. H. (1995). Why phylogenies are necessary for comparative analysis. In: *Phylogenies and the Comparative Method in Animal Behaviour* (Martins, E. P. ed.) Oxford: Oxford University Press.
OAKES, E. J. (1992). Lekking and the evolution of sexual dimorphism in birds: comparative approaches. *Am. Nat.* **140,** 665–684.
PAGEL, M. D. (1992). A method for the analysis of comparative data. *J. theor. Biol.* **156,** 431–442.
PAGEL, M. D. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255,** 37–45.
PAGEL, M. D. & HARVEY, P. H. (1988). Recent developments in the analysis of comparative data. *Q. Rev. Biol.* **63,** 413–440.
PAGEL, M. D. & HARVEY, P. H. (1989). Comparative methods for examining adaptation depend on evolutionary models. *Folia Primatol.* **53,** 203–220.
PETERSON, A. T. & BURT, D. B. (1992). Phylogenetic history of social evolution and habitat use in the *Aphelocoma* jays. *Anim. Behav.* **44,** 859–866.
PROCTOR, H. C. (1991). The evolution of copulation in water mites: a comparative test for nonreversing characters. *Evolution* **45,** 558–567.
READ, A. F. (1991). Passerine polygyny: a role for parasites? *Am. Nat.* **138,** 434–459.
READ, A. F. & NEE, S. (1991). Is Haldane's rule significant? *Evolution* **45,** 1707–1709.
READ, A. F. & NEE, S. (1993). Haldane's coincidence: a reply to Brookfield. *Evolution*, **47,** 1888–1889.
RIDLEY, M. (1983). *The Explanation of Organic Diversity: The Comparative Method and Adaptations for Mating*. Oxford: Oxford University Press.
RIDLEY, M. (1986). The number of males in a primate troop. *Anim. Behav.* **34,** 1848–1858.
RIDLEY, M. (1988). Mating frequency and fecundity in insects. *Biol. Rev.* **63,** 509–549.
SANDERSON, M. J. (1991). In search of homoplastic tendencies: statistical inference of topological patterns in homoplasy. *Evolution* **54,** 351–358.
SEGER, J. (1991). Cooperation and conflict in social insects. In: *Behavioural Ecology. An Evolutionary Approach* (Krebs, J. R. & Davies N. B. eds), 338–373. Oxford: Blackwell Scientific.
SILLÉN-TULLBERG, B. (1988). Evolution of gregariousness in aposematic butterfly larvae: a phylogenetic analysis. *Evolution* **42,** 293–305.
SILLÉN-TULLBERG, B. (1993). The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution* **42,** 293–305.
SILLÉN-TULLBERG, B. & MØLLER, A. P. (1993). The relationship between concealed ovulation and mating systems in anthropoid primates: a phylogenetic analysis. *Am. Nat.* **141,** 1–25.
WU, C.-I. & DAVIS, A. W. (1993). Evolution of postmating reproductive isolation: the composite nature of Haldane's Rule and its genetic bases. *Am. Nat.* **142,** 187–212.

## APPENDIX

Here we illustrate how our null model might be developed into a statistical test in situations more complex than Haldane's Rule (Read & Nee, 1991). We do so using some data from Sillén-Tullberg's (1988) study of aposematism and gregariousness in butterfly larvae. Sillén-Tullberg hypothesized that the evolution of warning coloration increased the probability of gregarious behaviour evolving. The two traits are thus coloration (warning/cryptic) and gregariousness (solitary/gregarious). Relevant data for one of the butterfly families she considered are summarized in Fig. A1. Our approach might be implemented as follows.

(1) Beginning at the tips of the tree, sister taxa which differ in the state of the putative independent variable are identified. Note that it is not critical to have a fully resolved phylogeny, although this obviously affects the amount of information contained in the tree. Four sister taxa differing in coloration can be recognized in Fig. A1; nodes defining them are numbered (i–iv).

(2) From these are derived the sister-taxon comparisons. These must be independent. Thus, each terminal taxon must appear in only one comparison, and paths linking sister taxa in the phylogeny must not cross. The four matched pairs in Fig. A1 generate four sister-taxon comparisons: (i) {*P. brassicae* + *P. cheiranthi*} versus {*Artogeia rapae* + *A. manni*}; (ii) {*A. napi*} versus {*Pontia* spp.}; (iii) {*Anthocharis* spp.} versus {*Euchloe* spp}; (iv) {*Aporia crataegi*} versus {Coliadinae}. Note that taxa involved in comparisons (i) and (iii) form part of the sister clades involved in the other two comparisons, but are excluded from those in order to ensure independence. Felsenstein (1988), Burt (1989) and Read (1991) discuss these procedures in the context of continuous variables.

(3) For each of the comparisons, we ask whether the sister taxa exhibit different levels of the putative dependent variable. From Fig. A1, it is obvious that the level of gregariousness exhibited by a taxon differs

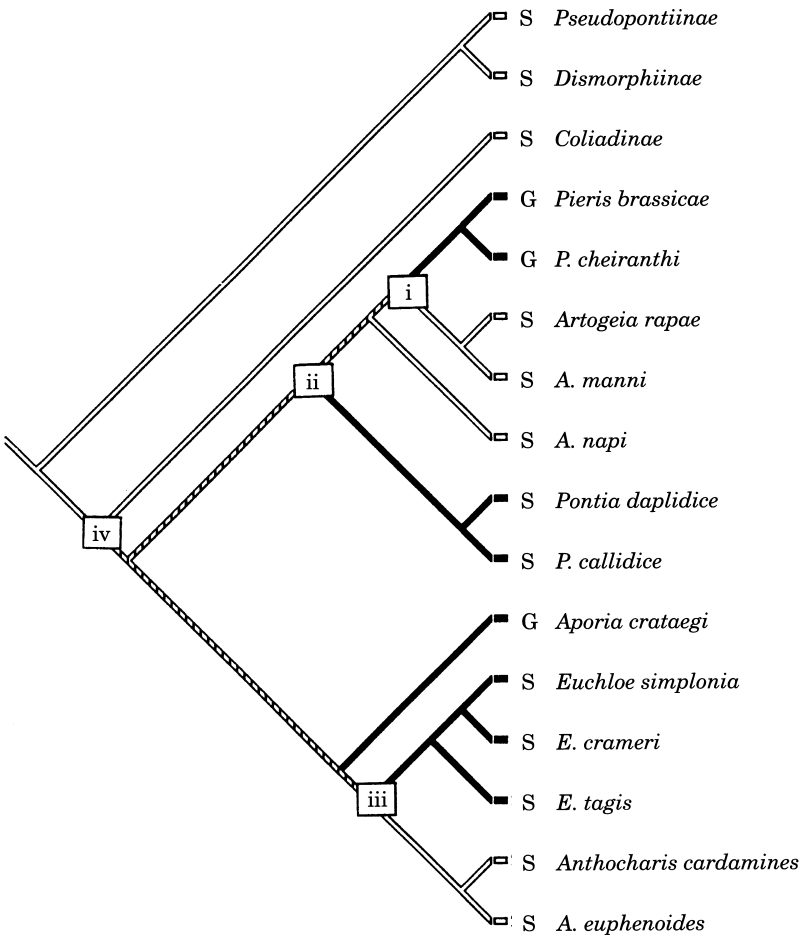


FIG. A1. The butterfly family Pieridae, showing, for clarity, the coloration of larvae as reconstructed from the states of terminal taxa (Maddison & Maddison, 1992). Black, warning coloured; open, cryptic coloration; hatched, equivocal. Solitary (S) or gregarious (G) terminal taxa marked. Numbered nodes are those defining sister groups differing in coloration, as discussed in text. Data and phylogeny from Sillén-Tullberg (1988: 298); branch lengths are arbitrary.

from that exhibited by its sister taxon in comparisons (i) and (iv). However, all species in comparisons (ii) and (iii) are solitary. These comparisons can therefore provide no evidence for or against the hypothesis and do not contribute to the calculation of the probability of the association arising by chance. However, some lineages in uniformative comparisons may be relevant in other comparisons if they form part of a clade used higher in the tree. For instance, in Fig. A1, the genera *Pontia* and *Euchloe*, which form a part of the uninformative comparisons (ii) and (iii) respectively can be incorporated in warning coloured taxon of comparison (iv), making that comparison {*Euchloe* spp. + *Pontia* spp. + *Aporia* sp.} versus {Coliadinae}. Such a procedure may make additional comparisons informative, or provide more data from which to compare the state of $y$ in matched pairs closer to the root of the tree. Of course, care must be taken to ensure that sister taxa used in the final test remain independent.

(4) Among the informative comparisons, the state of $x$ in the lineage exhibiting more of a particular state of $y$ is determined. Using a binomial test we can then ask whether a particular configuration is occurring more often than expected by chance under our null model of random allocation of the state of $x$. Both relevant comparisons in Fig. A1 support Sillén-Tullberg's hypothesis: in each case, the warning coloured taxon is the more gregarious. Under our null model of random allocation of coloration, that configuration had a 0.5 probability of occurring once and hence a probability of $0.5^2 = 0.25$ of it occurring in both cases.

If one was willing to define explicitly models of discrete character change it may be possible to develop more sophisticated statistical tests of the null model we have proposed, perhaps using the quantitative difference in the state of $y$ in the two taxa. However, given current understanding, the non-parametric binomial test is less difficult to justify.

Where there is non-simultaneous change in two traits, it may be possible to infer the sequence of evolutionary events and investigate the direction of any causal association between two discrete characters (Sillén-Tullberg, 1988; Donoghue, 1989; Maddison, 1990; Harvey & Pagel, 1991). It may be the case, for example, that in more comparisons than expected, the taxon with a particular state of $x$ exhibits, after several bifurcations, a predicted change in $y$. Indeed, which regions of the tree are informative can differ depending on which variable is treated as the causal variable. Differences in the number of possible comparisons may thus provide support for one causal direction but insufficient evidence to support the other. Below we give a real example where this is the case. Were one not interested in directionality, the test may have to be performed twice, assuming each variable was in turn the causal variable. Where character transitions in the two traits tend to occur on the same branches, the conclusions will be similar and it will be difficult to determine causal direction (as it should be under the null model, because temporal sequence cannot be discerned).

Perhaps the biggest difficulty in implementing the model of random allocation of treatments in this context is determining which taxon in each matched pair exhibits more of one state of $y$. For example, the number of times a trait has evolved, or the number or proportion of species, taxa or branches exhibiting a particular state could be compared. In many instances it will be obvious and in others it will not matter, since we are interested only in determining which taxon exhibits more of one state, not how much more. Where it does matter, we suspect that the best approach will depend on the characters under investigation and the nature of the hypothesis being tested; there is no general prescription. We note that any way of doing it must incorporate implicit assumptions about how evolution is thought to have occurred, which may be controversial and should therefore be made explicit. They do not affect the principle of the null model.

## AN EXAMPLE REVISITED

Sillén-Tullberg's (1988) hypothesis that warning coloration promotes the evolution of gregariousness (because group living is beneficial to prey if predators more quickly learn to avoid prey after some encounters) is contrary to what is generally thought to be the suggestion by Fisher (1930) that distastefulness and warning coloration evolves only where there are gregarious associations of kin groups. Using Ridley's (1983) approach, Sillén-Tullberg (1988) found evidence to support her hypothesis. Maddison (1990) suggested this was because much of the tree was warning-coloured and thus, by chance alone, transitions to gregariousness from the solitary ancestral state would tend to occur in warning coloured taxa. The concentrated-changes test apparently confirmed that view (Maddison, 1990), although Sillén-Tullberg (1993) has recently argued that Maddison's analysis was performed on a tree on which warning coloured branches were over represented because cryptically coloured clades were not resolved to the same extent as warning coloured clades.

From Sillén-Tullberg's (1988) cladograms and those derived from the descriptions in her text, there are 14 possible comparisons of sister taxa differing in their coloration (including the four in our Fig. A1). Both sister taxa in four of these exhibit the same state of gregariousness and are thus uninformative. Of the remainder, at least one of the sister taxa in eight of the pairs exhibits only a single state of gregariousness and so differences in $y$ can be unambiguously assigned. In the other two cases, sister taxa contain both solitary and gregarious subtaxa and the problem is to say which exhibits greater gregariousness. We did so by comparing the proportion of each state exhibited by terminal taxa and weighting the contributions of daughter lineages equally within each taxon (to reduce the noise introduced by random differences in clade bushiness). Taken together, eight comparisons support Sillén-Tullberg's contention that the evolution of warning coloration increases the likelihood of evolving gregariousness and two contradict it, a marginally significant result (one-tailed binomial $p = 0.0547$).

What about the reverse proposition, which is Fisher's (1930) idea that gregariousness affects the probability of warning coloration evolving? A total of 28 sister taxa differ in gregariousness, but coloration differs in only six of these. In four cases, the more warning coloured taxon is the more gregarious; in the other two it is not. This is no better than expected by chance (one-tailed binomial, $p = 0.35$). This is different from the conclusion reached above because the majority of transitions between character states do not occur on the same branch so that only three comparisons in the two tests are of the same sister taxa. A formal power test may show that there are too few comparisons in the second case to rigorously test the causal direction. Nevertheless, we note that our comparative analyses provide some support for Sillén-Tullberg's belief that the evolution of warning coloration promotes the evolution of gregariousness, but none for the idea that gregariousness provides the conditions necessary for the evolution of warning coloration.