# Non-independence in Statistical Tests for Discrete Cross-species Data

ALAN GRAFEN* AND MARK RIDLEY†

* *St. John's College, Oxford OX1 3JP, and the † Department of Zoology, South Parks Road,*
*Oxford OX1 3PS, UK*

(*Received on* 15 *August* 1996, *Accepted in revised form on* 23 *May* 1997)

The paper describes three previously undetected effects, due to biases and non-independence, that can arise in statistical tests for associations between character states in cross-species data. One kind, which we call the family problem, is general to all known methods. In phylogenetic data, the ancestral character state from which changes occur, or below which variation is found, is likely to be the same for many regions of the tree. The family problem interacts with two kinds of non-independence that arise because of the methods of reconstruction of character states that existing tests use. Different kinds of non-independence arise in methods that reconstruct joint, or single, character states, respectively. Methods, like Ridley's (1983), that work with joint character states suffer from the problem that a character state cannot change to itself with parsimony. Other methods that work with single character states suffer from the problem that within a locally variable region of the tree it is more likely with null data that there will be two single changes in the two characters in separate branches than one double change in both; associations opposite to the locally ancestral state are therefore likely to be found in more than 50% of the variable regions. In real data sets, the family problem acts to spotlight the other kinds of bias: if the family problem is large the bias in tests due to the way they reconstruct characters will be large, whereas if it is small, the local biases tend to cancel and disappear in the aggregate.

## 1. Introduction

Adaptation is often studied by looking for associations between characters in cross-species data: that is, by the comparative method. The chapters in Martins (1996) contain recent discussions of the topic. Cross-species data is well known to suffer from non-independence because of its phylogenetic structure: related species share characters states for phylogenetic reasons. Read & Nee (1995) and Grafen & Ridley (1997a) are recent discussions of the problem; Grafen & Ridley (1997a) also describe a model of character evolution that can be used to assess whether a proposed statistical method for discrete characters has successfully dealt with the non-independence due to phylogenetic structure in the data. We are not concerned with that problem here, and have nothing to say about "phylogenetic inertia" either in its original (Wilson, 1975) or subsequent (Ridley, 1983, p. 17–18) meanings. We are concerned with some other kinds of non-independence, which we believe have not been noticed before. These kinds of non-independence arise because of the way the methods of reconstructing character states treat the data; they are more imposed by the methods themselves than generated by the evolutionary process. The problems arise in existing methods as a combination of two factors, one that is common to all methods, and a second the nature of which depends on the kind of reconstruction used. We shall look at the common factor first and then move on to those that are peculiar to particular methods. We concentrate on tests for characters that have discrete character states (such as states $A$ and $a$ for character $A/a$).

## 2. The "Family Problem"

Tests for associations between character states look for regions of the tree where changes occur, or the character states are variable; they then see whether all

0022–5193/97/200507 + 08 $25.00/0/jt970484

the regions show a similar pattern of change, or association. The "family problem" arises when many of the regions share the same initial ancestral character states. At this stage in our argument the shared ancestral states are a phenomenon rather than a problem; the next section will reveal how it can become problematic. The reason why, as a matter of fact, many regions of the tree are likely to share the same ancestral states will usually be the rarity of evolutionary change. Several separate regions of the tree, within each of which there has been change, will all share the same initial ancestral states if there has been no character change between the deep ancestor at the root of the tree and those variable regions. The family problem could also arise if there is some other character state, which differs from the state at the deep root, that is shared in the ancestry of several variable regions of the tree.

The family problem arose in the simulations of Grafen & Ridley (1996) when the similarity of higher reconstructed states is due to common ancestry and fairly low rates of evolution. A referee has suggested that the family problem may also occur with extremely high rates of evolution, when each of two independent characters has a common and a rare state. Then, the methods of reconstruction of states might tend to create similarity of states at higher nodes. We do not know if this is true or not, but if it is similar arguments to those given below here will probably apply; it would also raise serious doubts about all methods like that of Burt (1989).

## 3. Single or Joint Character Reconstruction

All the methods that have so far been proposed to test for discrete character associations in cross-species data reconstruct by parsimony at least some character states in the phylogenetic tree. The methods start with a phylogenetic tree for the species, and the character states of the species at the tips of the tree. A parsimonious reconstruction can then be performed in either of two ways, using either joint character states or single character states. As we shall see, the kind of non-independence that arises differs in the two cases.

The method of Ridley (1983), formalized by Grafen & Ridley (1997b) as the "independent character evolution" (or ICE) test, works with joint character states. In the simplest case there will be two characters with two states each ($A/a$ and $B/b$); a species can then have one of four character states ($AB$, $Ab$, $aB$, or $ab$), and they are treated as four unordered states of a single character. Working back from the terminal species, each higher node is assigned a joint character state by parsimonious reconstruction.

Other methods, in particular those of Burt (1989) and a new method described by Grafen & Ridley (1997c), work with the single character states. The methods take each character in turn and work back from the terminal species until they find a higher node below which the character varies; the methods concentrate on nodes below which both characters vary. Burt (1989) and Grafen & Ridley (1997c) make no further reconstructions in the higher regions of the tree beyond these nodes. Pagel (1994) proposes still another kind of method, which does not easily fit our distinction between single- and joint-character reconstruction; we comment on it in Ridley & Grafen (1996).

We should clarify the meaning of the word "reconstruction". It can have what might be called a broad and a narrow meaning. A test relies on reconstruction in the narrow sense if it assigns character states to ancestral species and then makes inferences on the assumption that the assignments are correct. A test relies on reconstruction in the broad sense if it assigns character states to ancestral species but its inferences do not assume that the assignments are correct. Statements about reconstruction in this paper require no more than the broad sense, and all methods (except species counting) use some kind of reconstruction in this broad sense. Methods such as Burt's (1989) and Grafen & Ridley's (1997c), like the phylogenetic regression (Grafen, 1989, p. 148), work by conditioning on pattern rather than inferring ancestral states; they do not reconstruct ancestral states in the narrow sense. It can be seen that they do reconstruct character states in some sense; there is implicit reconstruction even in the recognition of uniform nodes. Ridley's (1983) test uses reconstruction in the narrow sense. The distinction is important because statistical tests need to take account of the uncertainty in ancestral assignments and therefore should not assume those assignments to be true. Methods can also be distinguished according to how far back in the tree they carry their reconstructions. Some methods (such as the phylogenetic regression and the ICE test) reconstruct all the way back to the root; others [such as the method's of Burt and Grafen & Ridley (1997c)] stop as soon as possible. "Reconstruction" therefore need not mean a complete reconstruction all the way back to the root. However, these distinctions are not the topic of this paper and this paragraph is included only to prevent misunderstanding.

## 4. A Character State Cannot Change to Itself With Parsimonious Reconstruction

We shall discuss the problem in terms of the ICE test; the problem is more general, however. Comparative methods for discrete characters that are proposed in the future may need to consider it.

The ICE test, in Grafen & Ridley's (1997b) version, contracts each reconstruction of the joint character states into a "character change tree" (illustrated in Ridley & Grafen, 1996, fig. 1). A region of the tree in which the joint character state does not change is collapsed into a single node. When all the uniform taxa have been collapsed in this way, the result is a tree (the "character change tree") in which the only branches are ones in which a character changed. The numbers of nodes with each character state are then entered in a contingency table and a test performed.

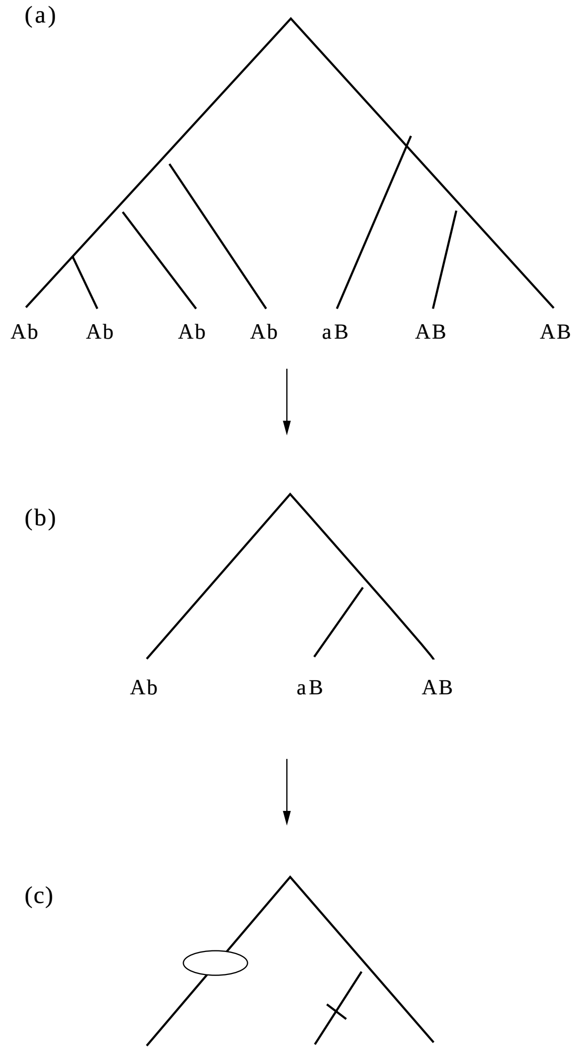

FIG. 1. Elemental depiction of character change. (a) The raw data. (b) Uniform branches collapsed. (c) A line indicates a change in character *A*/*a* and a flat circle a change in character *B*/*b*. In this case the state *AB* is ancestral at the top of the tree.

The events in the branches of the character change tree were, as Ridley (1983) argued, evolutionarily independent—in the sense that they were separate events in time or space—but they may not be statistically independent. Imagine any one node in the character change tree. It has some character state, such as *AB*. The method of parsimonious reconstruction then forces all the neighbouring nodes to be either *Ab*, *aB*, or *ab*; it is a parsimonious impossibility, even absurdity, for a neighbouring node to be *AB*. If a neighbouring node in the original phylogeny were *AB* it would have been collapsed, in the reconstruction, into the node in question. The only way an *AB* node in the character change tree can be joined to another *AB* node is via one of the other three kinds of node. The changes away from a node are "forced" into the three states other than the one at the node. The character changes through the tree are therefore statistically non-independent. Any method that assumes the changes are independent will be liable to find a spurious association in null data. Ridley (1983) and the ICE test (Grafen & Ridley, 1997b), for instance, enter the frequencies of the four character states in a contingency table and then apply a test, such as Fisher's exact test or the chi-squared test, that assumes independence.

The non-independence we have just described is for any one node. The magnitude of the problem in a real case depends on how many of the character changes in the whole data set occur from the same node in the character change tree. If the data set has a large family problem, most of the changes will be away from the same node—the node containing the ancestral state for the tree. (The character change tree is then star shaped: see Ridley & Grafen, 1996; Fig. 1). Then most of the changes in the tree have, because of parsimony not biology, to be to the other three non-ancestral states. Alternatively, it may be that there are a number of changes, one above the other, through the tree from the deep ancestor to the terminal species; the family problem is reduced. Now the changes will occur from a number of different nodes. (The character change tree looks like more of a string than a star.) Still, for any one node, the changes are forced into some set of three states, but the set differs between nodes and the biases cancel one another out. We no longer expect to find such severe invalidity in null data. Grafen & Ridley (1996) investigated the influence of tree shape on the Type I error rate of the ICE test and found that the test was reasonably valid for a realistically shaped phylogeny,

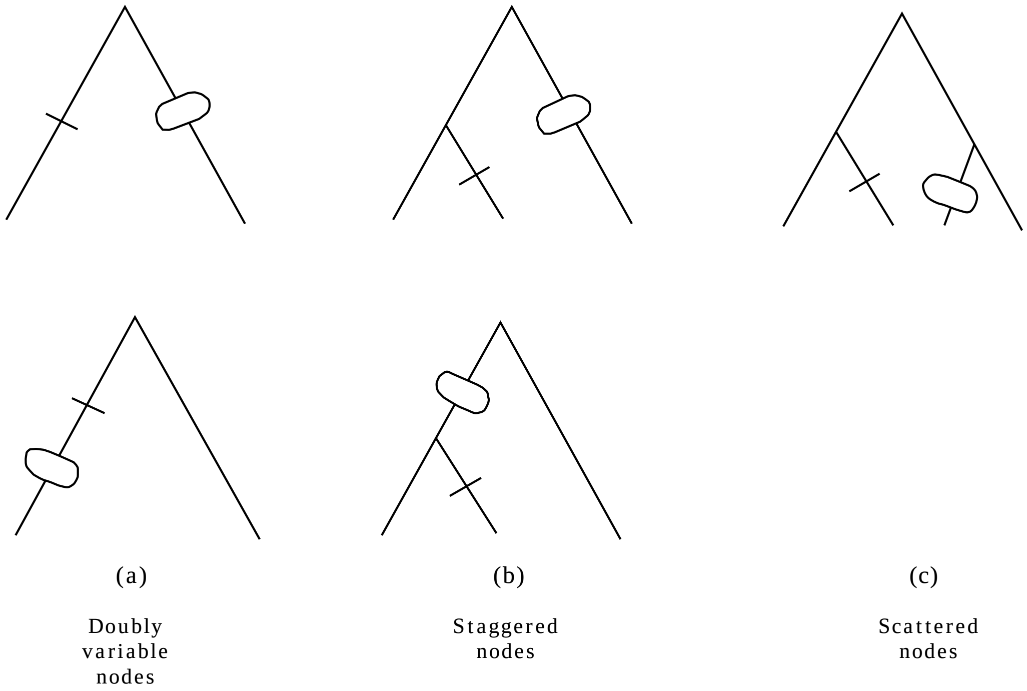

FIG. 2. Three kinds of variable node. Let an "A-node" be a node each of whose daughters is uniform for either "A" or "a", but not all of the daughters are uniform for the same state. Similarly for B-node and B/b. A variable node is the most recent common ancestor of an A-node and a B-node, with three logical possibilities. (i) A doubly variable node is itself an A-node and a B-node. (ii) A staggered node is itself an A-node, with a B-node somewhere in the subtree below it (or vice versa). (iii) A scattered node, which is itself neither an A-node nor a B-node. For conventions see Fig. 1.

but biased in the manner explained here (too many changes found away from the ancestral state in null data), when the phylogeny was symmetrical.

## 5. Two Separate Single Changes are More Likely than One Double Change

We now turn to tests that trace each single character back from the terminal species until they reach a node below which both characters vary. The general principle of the tests is that a number of these nodes may be found in a data set, and a comparative test can look to see whether the character association is consistent between them. The associations among the characters must have evolved independently in the different nodes, and the tests assume that they are statistically independent too.

At this point we should introduce a kind of figure. Figure 1 shows how the pattern of character changes can be expressed in elemental form: the uniform branches are all collapsed and the changes in the two characters indicated by symbols. The elemental diagrams can therefore represent any number of real species and branches. In these terms, when we trace back from the terminal species to find nodes below which both characters vary, it is possible for the variable nodes to be of three kinds (Fig. 2): both characters may vary below the same node (a doubly variable node); or one character may vary at a node below the node the other character is variable at (staggered variable node); or the two characters may vary below separate nodes that are not hierarchically arranged in the tree (scattered variable nodes). The comparative tests that have been proposed differ in which of these kinds of nodes they decide to admit as evidence: the decision matters, as we shall see.

We are going to discuss non-independence in terms of contingency tables, and it will help to have a conventional form. The species (or uniform blocks of species in Figs 1 and 2) can have any of four character states, giving a 2 × 2 contingency table. For a Fig. 2-type variable node, let the frequencies of species with states $AB$, $aB$, $Ab$, $ab$ be $n_{AB}$, $n_{aB}$, $n_{Ab}$, $n_{ab}$. For consistency we shall suppose that the ancestral state for the node is $AB$. The contingency table gives the numbers of species (or uniform blocks of species—it does not matter which here) at the bottom of the subtree; we assume for simplicity of exposition that there has been a single change in each character. If both changes are in the same branch, the node will have species with *ab* and species with *AB* (those that retain the ancestral state). The values of $n_{AB}$ and $n_{ab}$ will be positive; $n_{aB}$ and $n_{Ab}$ equal zero. The sign of $n_{AB}n_{ab} - n_{aB}n_{Ab}$ is positive. We shall call this a "positive" or "ancestral diagonal" contingency table. Alternatively, the changes may have been in separate branches: there will then be positive values for $n_{aB}$ and $n_{Ab}$, and usually for $n_{AB}$, but $n_{ab} = 0$. The sign of $n_{AB}n_{ab} - n_{aB}n_{Ab}$ is negative, and we call it a negative, or non-ancestral diagonal, contingency table.

The comparative tests working with single character states assume that with null data, in which the chances of change in $A/a$ and $B/b$ are independent, negative and positive contingency tables are equiprobable. But they are not: there is a bias in favour of the negative, non-ancestral diagonal, association.

The bias can be seen by working through all the shapes of tree below a variable node, with null data. We shall use the simplest trees to illustrate the problem, one for three species in any asymmetric tree (Fig. 3) and another for four species in a symmetric tree (Fig. 4); each tree has a single change in each character. The reader may like to translate the node shapes in these figures into the three kinds of variable node in Fig. 2. There is more than one way of obtaining some of the tree shapes, given the single change in each character, and their probabilities are given in the figures. What are the frequencies of negative and positive contingency tables? It is convenient to treat the asymmetric and symmetric trees separately. We assume that exactly one event takes place for each character, and that the probabilities of an event taking place in a given branch segment is proportional to its length. Suppose first that the height of the tall branch in the asymmetric tree has a height of 2 units relative to a unit length of the short branches. Now look at Fig. 3. If we count the frequency of contingency tables with positive and with negative signs, we find

chance positive contingency table = 11/25
chance negative contingency table = 14/25.

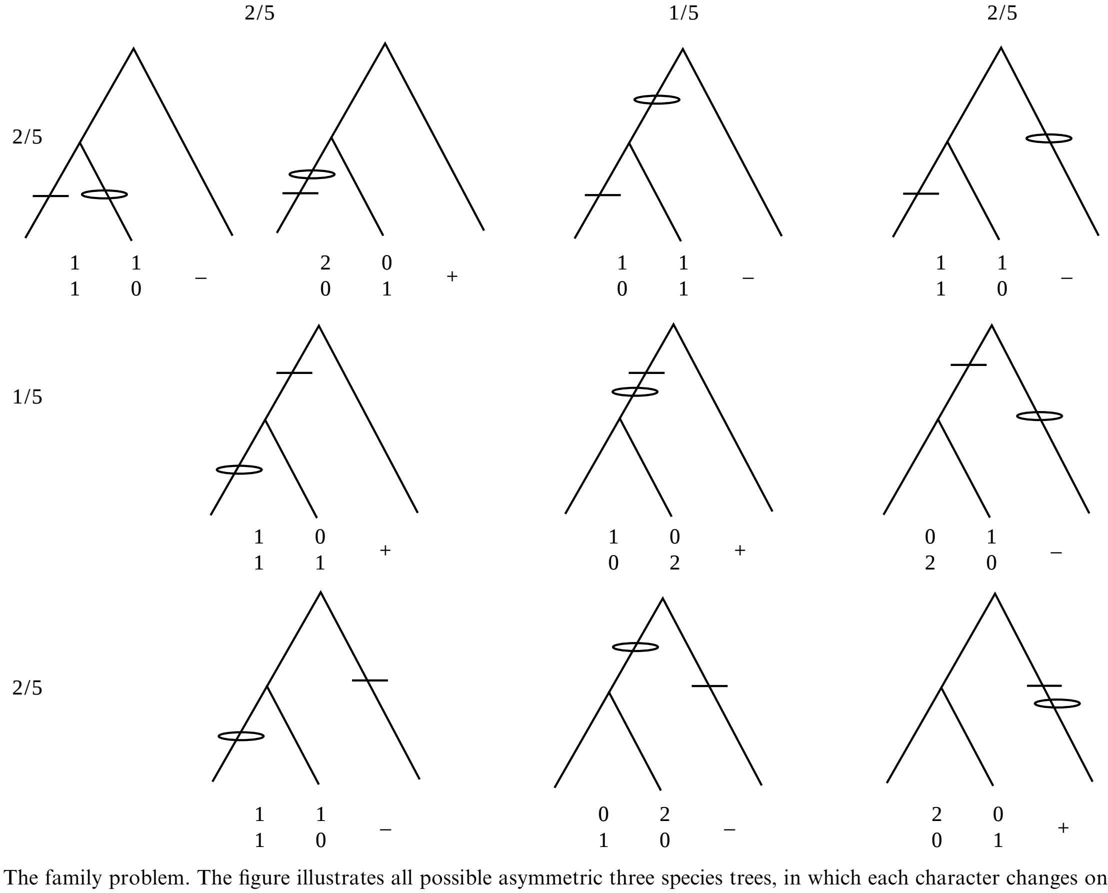

FIG. 3. The family problem. The figure illustrates all possible asymmetric three species trees, in which each character changes once. For conventions see Fig. 1. The chances of change for the rows and columns suppose that change is random and the height of the higher node is twice the height of the lower node. The contingency table and its sign are given below each tree. The fractions at the tops of columns and left of rows are the appropriate multiples, due to the number of ways each tree shape can arise by chance.

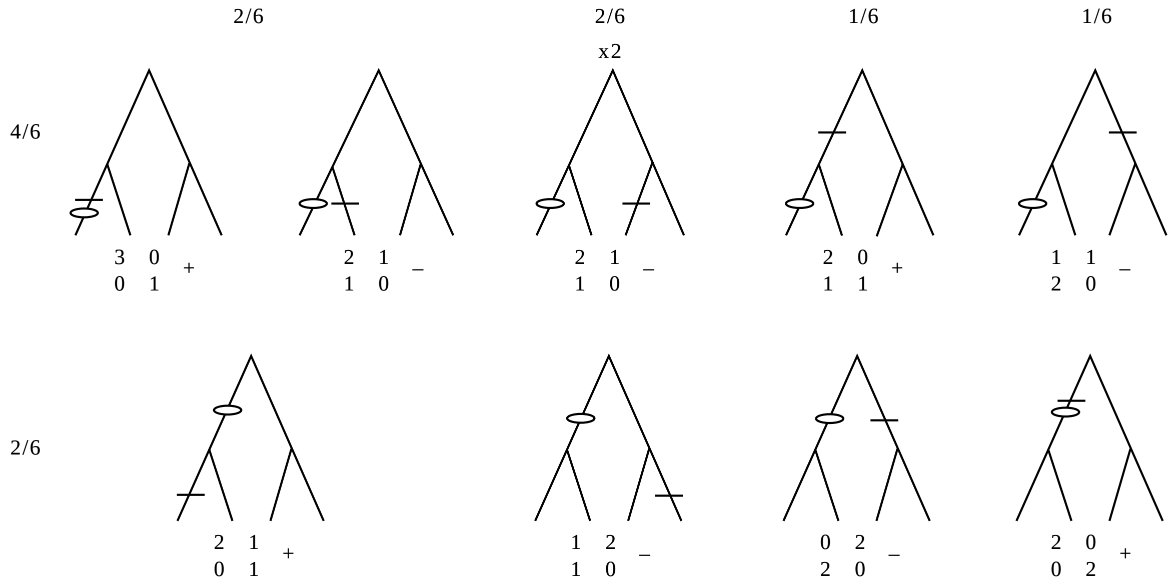

FIG. 4. The family problem. The figure illustrates all possible symmetric four species trees, in which each character changes once. For conventions see Fig. 1. The chances of change for the rows and columns suppose that change is random and the height of the higher node is twice the height of the lower node. The contingency table and its sign are given below each tree. The fractions at the tops of columns and left of rows are the appropriate multiples, due to the number of ways each tree shape can arise by chance.

Figure 4 shows all the possibilities for symmetric trees. Now

chance positive contingency table = 14/36
chance negative contingency table = 22/36.

Again the bias is in favour of negative results. The important fact is that the probabilities are not 50:50 in the null case. The reason is that, in these nodes, it is more likely that the changes in the two characters will be in separate branches than one above the other. Two single changes are more likely than one double change.

These fractions are for the special case in which the relative height of the three or four species tree is 2 units; but the result is general. If the height of the long branch is $h$ units relative to 1 unit in the short branches, then in the asymmetric three species case (Fig. 3),

$$\text{chance of positive contingency table} = \frac{2h^2 + 2h - 1}{(2h + 1)^2}$$

$$\text{chance of negative contingency table} = \frac{2h^2 + 2h + 2}{(2h + 1)^2}$$

The limit of these fractions, as $h$ goes to 1, are 1/3 and 2/3, and as $h$ becomes very large, 1/2 and 1/2. The bias is always towards too many negative contingency tables. Likewise, in the asymmetric four species case (Fig. 4), if the height of the whole four species clade is $h$ units relative to 1 unit in the lower branches,

$$\text{chance of positive contingency table} = \frac{h^2 + 2h - 1}{2(h + 1)^2}$$

$$\text{chance of negative contingency table} = \frac{h^2 + 2h + 3}{2(h + 1)^2}.$$

The limit of these fractions, as $h$ goes to 1, are 1/4 and 3/4, and as $h$ becomes very large, 1/2 and 1/2 again. The bias is slightly stronger than in the asymmetric trees, and still always towards too many negative contingency tables. In any particular case, there might be any mix of symmetric and asymmetric trees and the exact bias will depend on the proportion of each.

The fraction of positive and negative associations found by a method in null data will often differ from the exact fractions given here. One reason is that the frequencies of different node shapes are influenced by the rate of evolution: the frequency of doubly variable types [Fig. 2(a)], for instance, increases relative to the frequency of scattered types [Fig. 2(c)] as the rate of evolution increases. The exact bias, therefore, depends on the evolutionary rate. However, the bias always remains toward negative associations, as our analysis for general heights reveals. Another reason is that methods differ in how they select the kinds of variable nodes for analysis. Burt's (1989) method uses all three types in Fig. 2. The ICDE method of Grafen & Ridley (1997c) uses the doubly [Fig. 2(a)], and staggered [Fig. 2(b)] nodes, but ignores the nodes with scattered variation [Fig. 2(c)]. This matters because the scattered variable nodes always generate negative contingency tables and are a powerful source of bias. In the symmetrical trees of Fig. 4, all the bias is attributable to the node (in the middle of the top row) with the scattered variation pattern: if it is excluded the frequencies of positive and negative contingency tables is equal. That does not save the ICDE test from bias, however, because the symmetric trees of Fig. 3 are biased and contain no scattered variable nodes. In the simulations of Grafen & Ridley (1996) Burt's test mainly worked on scattered nodes and was vulnerable to severe bias in consequence, Grafen & Ridley (1997c) discuss further the relation between the ICDE and Burt tests.

So the methods differ in how they select nodes, and the nodes they exclude may tend to have positive or negative results. This is the key to understanding the statistical biases of the methods, and will probably be the key to improving them. One way to improve them would be to seek a method that selected nodes such that they had an equal chance of finding positive and negative contingency tables with null data, under all biologically reasonable shapes of phylogeny; perhaps no such method exists. Another way would be to seek to formulate better null expectations for methods of the existing form, which appear to find too many nodes with negative, non-ancestral associations. Consideration of the non-independence described here may assist either way.

The magnitude of the total bias in a real data set caused by the bias towards non-ancestral associations within each variable node depends on the extent of the family problem. If most of the nodes start with the same ancestral state, they are all biased in the same way, and the total bias is large. If, however, there are a number of changes between the deep root of the tree and the locally variable nodes that are used in the test, then although the changes within each node are still biased, they will be biased towards different character states and the aggregate bias of them all may cancel to zero. The simulations of Grafen & Ridley (1996) looked at two null data sets, in one of which there was a large, and the other a small, family problem (as measured by the proportion of variable nodes that began with the same, deep ancestral root, character states). Burt's test showed the strong bias predicted here in the data set with the strong family problem, but was better behaved in the other data set. Grafen & Ridley (1996) provide results and Grafen & Ridley (1997c) analysis of the behaviour of the ICDE test with the two data sets. Readers of Ridley & Grafen (1996) may note that we have slightly altered the meaning of the expression "family problem" here, to provide a more systematic analysis.

## 6. Discussion

What is the relation between the kinds of non-independence described in this paper? The bias in a test is a compound of two effects; the first (the family problem) acts to draw attention to the second. For example, Grafen & Ridley (1996) showed that tests like those of Burt (1989) and Ridley (1983) have reasonably valid Type I error rates for null data sets that have little family problem. That does not mean non-independence, or bias, is absent, however. In both tests a number of character changes, or nodes, contribute to the total result and within each of them changes will be biased in the manner described in sections 3–5. The bias disappeared because the biases of the local nodes pointed in varying directions. When there was a strong family effect the biases of the local nodes added up instead of cancelling and the tests proved strongly biased. Thus, the family problem acted to spotlight, or alternatively to blur, the biases of sections 3–5. The sections 3–5 biases are present whether or not there is a family problem. The family problem itself, however, is inevitable in phylogenetic data, and it is a challenge of the comparative method for discrete characters to devise a way of looking at the data such that the test is unbiased whether the family problem is strong or weak.

What is the relation between the kinds of non-independence described here and the well-known kind of non-independence due to the sharing of character states between phylogenetically related species? The answer is, it is an additional problem, at least for discrete characters. The tendency of related species to share character states is an inherent property of phylogenetic data. The problems identified here are not so much properties of the data as of the way the methods analyse it. The problem of phylogenetic structure itself was solved by what Grafen [1989, p. 125, following Ridley (1983)] called the radiation principle: that each non-uniform higher node contributes one datapoint. The problems that arise because of the way character states are reconstructed may also be soluble, by appropriate adjustments to the methods. But they have not been solved yet.

## REFERENCES

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

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

Grafen, A. & Ridley, M. (1996). Statistical analysis of discrete cross-species data. *J. theor. biol.*, **183**, 255–267.

Grafen, A. & Ridley, M. (1997a). A new model for discrete character evolution. *J. theor. biol.*, **184**, 7–14.

GRAFEN, A. & RIDLEY, M. (1997b). A formalization of the comparative method of Ridley (1983). In prep.

GRAFEN, A. & RIDLEY, M. (1997c). The ICDE test: a new method for analysing discrete cross-species data. In prep.

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

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

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: Oxford University Press.

RIDLEY, M. & GRAFEN, A. (1996). How to study discrete comparative methods. In: (E. P. Martins, ed.), *Phylogenies and the Comparative Method in Animal Behavior*, p. 76–103. New York: Oxford University Press.

WILSON, E. O. (1975). *Sociobiology*. Cambridge, MA: Harvard University Press.