Syst. Biol. 64(1):127–136, 2015

DOI:10.1093/sysbio/syu070
Advance Access publication September 9, 2014

The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters

WAYNE P. MADDISON[1,*] AND RICHARD G. FITZJOHN[2,3]

[1]*Departments of Zoology and Botany and Beaty Biodiversity Museum, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z4;* [2]*Departments of Zoology and Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z4 and* [3]*Department of Biological Sciences, Macquarie University, Sydney NSW 2109, Australia*
[*]*Correspondence to be sent to: Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z4; E-mail: wayne.maddison@ubc.ca.*

Received 5 December 2013; reviews returned 4 July 2014; accepted 12 July 2014
Associate Editor: Mark Holder

When comparative biologists observe that animal species living in caves also tend to have reduced eyes, they may see such correlation as evidence that the traits are adaptively or functionally linked: for instance, selection to maintain eye function is relaxed when light is unavailable. Such cross-species correlations cannot give definitive tests of evolutionary mechanisms, but nonetheless offer important insights into biological relationships among traits in realms as diverse as ecology (e.g., Paradis et al. 1998; Purvis et al. 2000) and genomics (e.g., von Mering et al. 2002; Barker and Pagel 2005).

However, the last few decades have taught us that among-species correlative tests should take into account evolutionary relationships (Felsenstein 1985; Ridley 1989; Harvey and Pagel 1991). If phylogeny is not taken into account, an interpreted correlation may have a trivial explanation different from the biological relationship we claim. There is a correlation among species in the distribution of fur and bones in the middle ear—species with fur also have three bones in the middle ear, and vice versa. These two traits are characteristics of mammals, and absent outside the mammals. Using their shared distribution as evidence of an interesting biological relationship between fur and middle ear bones would be considered a mistake, however, for reasons understood long ago by Darwin (1872):

> We may often falsely attribute to correlated variation structures which are common to whole groups of species, and which in truth are simply due to inheritance; for an ancient progenitor may have acquired through natural selection some one modification in structure, and, after thousands of generations, some other and independent modification; and these two modifications, having been transmitted to a whole group of descendants with diverse habits, would naturally be thought to be in some necessary manner correlated.

In modern vocabulary, we would say that the thousands of species of mammals are not statistically independent because of their shared history (Felsenstein 1985; Ridley 1989), instead being pseudoreplicates. Many methods have been developed to avoid pseudoreplication while assessing cross-species correlations by accounting for phylogenetic relationships (e.g., Felsenstein 1985; Maddison 1990; Harvey and Pagel 1991; Garland et al. 1992; Pagel 1994; Read and Nee 1995; Martins and Hansen 1997; Huelsenbeck et al. 2003; Hadfield and Nakagawa 2010). These phylogenetically aware correlation methods have been applied with enthusiasm, although there have been skeptics questioning their need (Westoby 1999) and cautions expressed over their use (Ricklefs and Starck 1996; Freckleton 2009).

Not all is healthy with the paradigm, however. For categorical characters a special concern has been raised: commonly used and well-respected methods (e.g., Pagel 1994), do not eliminate pseudoreplication, as they are susceptible to an effect from a single evolutionary event (Maddison 1990, 2000; Read and Nee 1995; Ridley and Grafen 1996; Grafen and Ridley 1997). As a result, a significant statistical association between traits inferred by these methods can mean very little in some circumstances, misleading biologists about the relationship between traits.

Although this concern has been raised, there has been little effect on the practice of comparative studies, we think because the issue has not been well understood. We do not, alas, have a solution. Our purpose here is to explore the issue in depth in part to bring caution to comparative studies, in part to characterize how our methods should behave in hopes of provoking appropriate solutions.

THE TARGET: ADAPTIVE/FUNCTIONAL RELATIONSHIP

It is useful to begin by outlining the limits of what we can hope to learn from comparative data. Comparative

data may indicate a correlation between variable X and variable Y, but this could be caused by the influence of a third variable—a streamlined body may not lead to tolerance to anoxia among mammals, but an aquatic habitat could lead to both. With comparative data alone, we cannot rule out the existence of a third variable that influences the two variables of interest. Credible mechanisms and controlled, manipulative experiments are needed to support precise causal hypotheses (Westoby 1999).

This means that *any* comparative method, no matter how robustly applied, must be satisfied with a relatively weak conclusion: that the two variables of interest appear to be part of the same adaptive/functional network, causally linked either directly, or indirectly through other variables. When teased out of patterns in cross-species data, even this weaker conclusion is likely interesting and satisfying to biologists.

This article is not simply a reminder that "correlation does not imply causation". Rather, it is to highlight the fact that in cross-species data, a correlation may not even give us the weaker conclusion—it may not even imply that variables are part of the same functional network. The perfect co-distribution of fur and three middle-ear bones among species would not be expected to occur by chance in a universe in which each species were independently evolved, but that is not the universe we live in. Instead, as Darwin explains, such co-distribution can be interpreted as a simple consequence of independent origin of two traits in a lineage, followed by co-inheritance by descendant species. It is not important whether this is described as no correlation (compatible with a null hypothesis involving simple phylogenetic descent) or as a correlation explained by coincidence. In either perspective, the pattern does not provide evidence for an adaptive or functional relationship among the traits.

Our target, and that of most comparative biologists, is not merely a general sense of correlation that reaches statistical significance, but rather a special sense of correlation that gives evidence for an adaptive or functional relationship between the traits. If we were satisfied with the former, we might as well do a Fisher's exact test on the species without regard to the phylogeny. If we want the latter, we should design our statistical tests to reveal adaptively or functionally interesting correlations. What we mean by "adaptive/functional relationship" (or "adaptive/functional network") is a set of influences among traits that arise from how the traits act or behave. Thus, we mean that one trait influences the other, or a third trait influences both, through some genomic, developmental, physiological, ecological, or other effect, whether immediate or through adaptation in evolutionary time. The influences need not involve adaptive function or selection pressure; merely that there is an effect. The effect need not be specified, and indeed a comparative test alone cannot reveal the precise nature of the adaptive/functional relationship. However, a comparative test on its own *can* provide evidence that there exists an adaptive/functional relationship of some sort, even without manipulative experiments or information about mechanism. Many biologists seek such evidence from comparative data. We want our comparative tests to respond to patterns that would be difficult to explain unless the traits were part of the same adaptive/functional network, and respond only to such patterns.

We would like to know therefore whether a positive result from a comparative test justifiably supports the conclusion that the variables are part of the same adaptive/functional network, or whether we are being misled by coincidence or biased sampling. The problem of concern in this paper is that tests such as Pagel's (1994) and Maddison's (1990) in some circumstances appear to give such support, when in fact the comparative data offers no support for a biologically interesting association (Maddison 1990; Read and Nee 1995; Ridley and Grafen 1996; Grafen and Ridley 1997).

What is Good Evidence for an Adaptive/Functional Relationship?

In order to consider the performance of particular comparative methods, we outline four scenarios to serve as litmus tests (Fig. 1). We argue that two of the scenarios (Fig. 1a,b) provide good evidence for an adaptive/functional relationship among the variables, while the other two (Fig. 1c,d) do not. Thus, we argue, comparative methods should indicate a correlation for Figure 1a,b (or at least, for scenarios of these types with sufficiently many origins), while they should not for Figure 1c,d.

Replicated Co-distribution

Figure 1a shows a scenario in which comparative methods should indicate correlation: a pattern of coincident origins of the black state in X and black in Y, replicated across multiple clades. Although the illustrated six origins may be too few to yield statistical significance (depending on the test used), larger patterns of this type would provide strong evidence for an association. It is reasonable to conclude that X and Y are adaptively or functionally linked in some way, even if indirectly through a third variable. Otherwise, why should the association between X and Y repeat in independent clades?

Replicated Bursts

Figure 1b shows another scenario in which an adaptively or functionally interesting correlation can be inferred. In each of multiple clades bearing an origin of the black state in X, there are multiple events of black evolving in Y. This suggests that changes in Y can be facilitated or inhibited depending on the state in X, but that the effect is neither necessary nor rapid. This pattern of nested changes reminds us that statistical methods need to be tuned to the

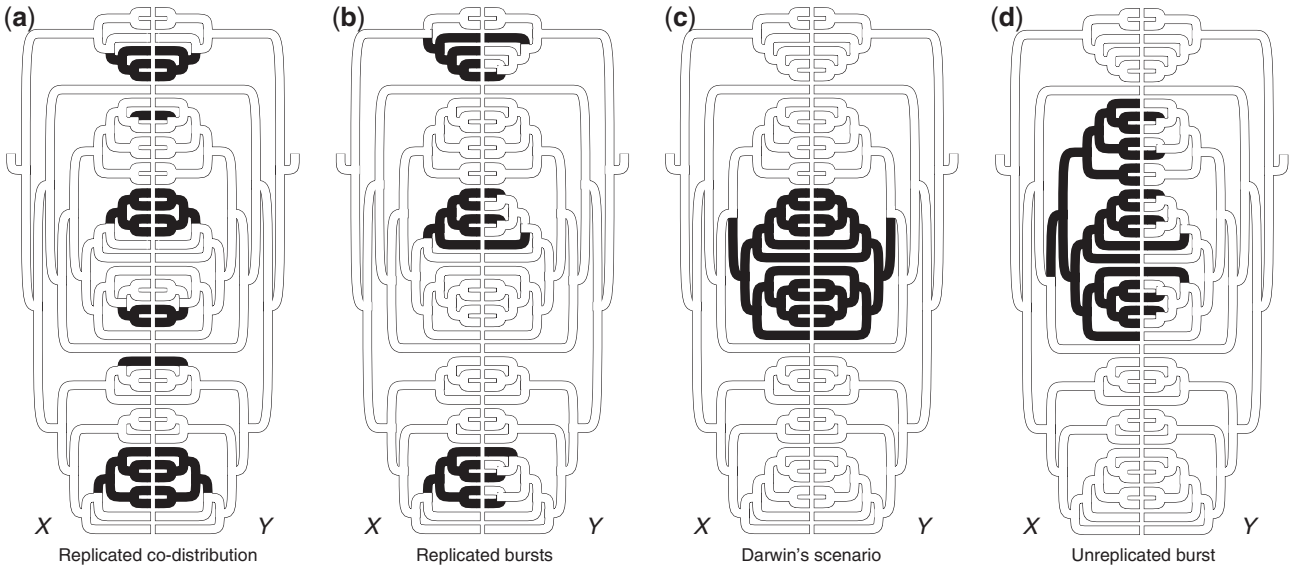

FIGURE 1. Four scenarios for the evolution of states of characters *X* and *Y*. In each, the same phylogeny is mirrored to show *X* at left and *Y* at right. State 0 = white; State 1 = black. (a) Replicated co-distribution. (b) Replicated bursts. (c) Darwin's scenario. (d) Unreplicated burst. Panels (a) and (b) provide good evidence for an interesting adaptive/functional relationship between *X* and *Y*; panels (c) and (d) do not.

particular biological hypotheses proposed, as different processes might predict different patterns (Maddison 1990). As with replicated co-distribution, it is reasonable to conclude that *X* and *Y* are adaptively/functionally linked in some way.

Darwin's Scenario

Figure 1c shows an example matching Darwin's (1872) scenario, with both *X* and *Y* having a state limited to a particular clade, like fur and middle ear bones in mammals. This pattern will be seen whenever a clade has more than one known synapomorphy. Despite the perfect co-distribution, it offers no evidence, on its own, for an adaptive/functional relationship. These traits could have changed millions of years apart along the long lineage ancestral to mammals, each by its own independent causes, and then been maintained, each by its own causes, in the descendants. There is no reason to think that any two of these traits are adaptively/functionally correlated to each other any more than to any one of the thousands of other distinct changes in the genome that would have occurred in this ancestral mammal lineage. Different parts of the genome can and do change independently within a lineage.

While a single coincident origin of both variables provides no evidence for an adaptively/functionally interesting relationship, it also provides no evidence against one (Westoby 1999). Feathers have a clear functional role in bird flight. The fact that there is only a single origin of feathers, co-distributed among extant archosaur species with active flight, is not evidence against a functional link between the characters. A good argument can be made for a functional relationship between flight and feathers, but it comes from experimental and other data, rather than from their co-distribution.

Unreplicated Burst

Figure 1d shows a more subtle scenario, where there is phylogenetically replicated change in one variable, but not the other. That is, there is only a single origin of the black state of character *X*, but multiple origins of black in *Y* within that same clade. An argument can be made that *something* special is happening to character *Y* in that clade, but this does not imply an interesting correlation with *X*. Many other features in the genome likely evolved in the ancestor of the clade and were inherited to become co-distributed with *X*—any one of them could just as well be the factor underlying *Y*'s enhanced rate of change. There is no good reason to conclude any direct or indirect adaptive/functional link between *X* and *Y*: their apparent correlation could be mere coincidence. For example, a gliding membrane stretching from the forelimbs to the thorax has originated at least six times in mammals (Jackson 2000), just once outside mammals, in the pterosaurs (Dudley et al. 2007). Do we attribute this concentration of origins in mammals to their fur? Milk? Enucleate erythrocytes? Were we to imagine replaying the evolutionary process, we should have no confidence that the evolution of enucleate erythrocytes would be associated with an increased rate of evolution of gliding membranes.

OUR METHODS DO NOT CORRECT PSEUDOREPLICATION

A primary motivation for developing tree-based quantitative methods for detecting correlated trait evolution is to avoid phylogenetic pseudoreplication, so

we would hope that they would detect no significant association between traits in Darwin's scenario (Fig. 1c) and the unreplicated burst scenario (Fig. 1d), given that the apparent association is not replicated. It might therefore come as a surprise that the most widely used phylogenetic method to test for correlation of categorical characters, Pagel's (1994) test, tends to find a significant association between traits in both cases.

To assess the behavior of Pagel (1994) under Darwin's scenario, we used Mesquite (Maddison and Maddison 2011) to simulate 100 birth–death trees, each with 100 species. Mesquite was then used to traverse through the tree, and at the first clade of 40–60 species found, set the state of the character outside the clade to 0, inside the clade to 1. This character was duplicated to make two identically distributed characters (X and Y), with distributions similar to those in Figure 1c. One such pair of characters was generated for each of the 100 trees. We then used the diversitree package (FitzJohn 2012) in R (R core team 2012) to fit models using maximum likelihood where each character evolved independently, or where rates of evolution of each character depended on the other character. In all of the 100 cases simulated for Darwin's scenario, Pagel's method returned a P-value <0.05 (Fig. 2).

Pagel's (1994) method also indicates correlation in the unreplicated burst scenario. To test this, we used the same 100 birth–death trees. For the X variable, we used the character generated as described above in mimicking Figure 1c. For the Y variable, we used Mesquite (Maddison and Maddison 2011) to simulate a binary character evolving on the tree with reasonably high rate. After this Y character's states were evolved, its states outside the clade marked by X were set to 0. Although this is not directly an evolutionary simulation, it is nearly the same as a model in which the Y variable starts at the root with state 0 and very low rate, and then suddenly increases its rate of change in the marked clade—the only difference is that in that model the marked clade would have started with Y in state 0, but in our construction it might not. Using diversitree (FitzJohn 2012) to perform Pagel (1994) test, we found that of the 100 cases simulated, 83 resulted in a significant result with $P<0.05$ (Fig. 2).

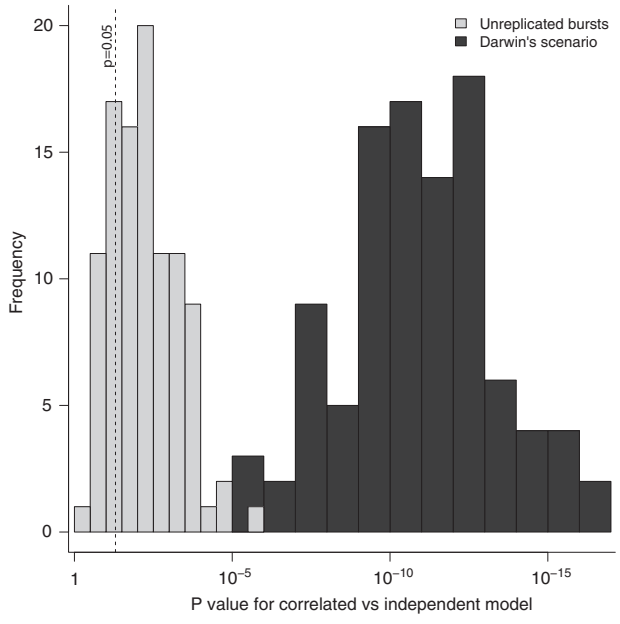

FIGURE 2. Pagel's (1994) test applied to 100 simulated cases like Fig. 1c (Darwin's scenario, dark grey) and like Fig. 1d (unreplicated bursts, light gray). Frequencies of log likelihood difference of correlated versus independent model $2(\ln(P(\text{data} \mid \text{correlated})) - \ln(P(\text{data} \mid \text{independent}))$. Vertical dashed line shows $P=0.05$.

Thus, Pagel's test is susceptible to yielding significant results from the effects of a single change in one of the characters, in both Darwin's scenario and the unreplicated burst. Other tests suffer the same problem, which we will call "within-clade pseudoreplication". Maddison (1990, p. 555) comments that his concentrated changes test is also susceptible to the effects of a single clade, the unreplicated burst scenario, and thus cannot rule out variables simply co-inherited with those of interest. It will also give a significant result with Darwin's scenario, if the test is adjusted to focus on a single branch. The correlation test of Huelsenbeck et al. (2003), a simple extension of stochastic character mapping (Nielsen 2002), measures in a given mapping the amount of time that lineages spend with states 1 in both X and Y, state 0 in both, and 0 in one and 1 in the other. In Darwin's scenario, the reconstructed mappings tend to show long stretches of lineages that hold state 1 in both X and Y, and elsewhere lineages that have 0 in both characters. This is interpreted by the method of Huelsenbeck et al. as a strong signal of an association.

The methods of Maddison (1990), Pagel (1994), and Huelsenbeck et al. (2003) suggest there are significant correlations in Figure 1c,d even though there is only a single evolutionary change in one or both characters. Pseudoreplication is still present, and coincidental correlations can be mistaken for interesting ones. It seems that we haven't progressed as far as we had thought in "correcting for phylogeny".

Model Failure and Ascertainment Bias

Our first response to these results is to defend the methods. They are, as mathematical constructs, merely operating under the assumptions they are built around, doing what they were designed to do. Thus, one might suggest that Pagel's test finds a significant correlation in Darwin's scenario because its assumptions are violated. If a character's distribution is inconsistent with the model of evolution assumed by a correlation test—for instance a continuous character with jumps violated the Brownian motion assumption of Felsenstein's (1985) test—then we could simply judge the test inapplicable to the character, and not use it. Perhaps character distributions like those of X and Y in Figure 1c,d cannot be reconciled with Pagel's model of character evolution (although this has not been demonstrated).

Pagel's test would not be at fault; it would simply be inapplicable. However, even if the test could be excused for the violation of its assumptions, that does not yield a solution. We need to be able to detect when the test's model is violated, and to devise a better test to handle more appropriate models. Although all models are inaccurate to varying degrees, we suspect that most biologists do not realize that this method and others fail to handle an iconic example of the need for phylogenetic corrections.

A second possible defence is that these tests are in fact *correct* to assign significance to Darwin's scenario. After all, if we imagine choosing two characters at random, it is very unlikely that their synapomorphies would fall on the same branch. This is a superficially compelling argument.

However, do we choose characters "at random"? One could scan genetic variation across many species' genomes and find genes with significantly concordant changes, as long as one corrected for multiple comparisons of the many genes. This, however, is not (usually) what is done in comparative analyses. Usually, only a few characters are considered in studies of correlation. While some may have been selected on purely functional grounds without prior knowledge of their distributions, others may have been selected precisely because we notice a trait characterizing a clade and wonder what effect it had on its bearers. That is, we may gravitate toward studying traits that we know *a priori* characterize clades of special interest. If this is the case, we cannot be surprised to see concordant distributions.

Thus, our mistakenly significant results may stem partly from ascertainment bias. When we react to Darwin's scenario as poor evidence for correlation, at least part of our negative judgment may come from an expectation that clade-biased character choice is possible or likely. Of course, if non-random choice of characters on which to focus (and, for that matter, clades to study) is rampant in evolutionary biology, then correlation tests will not be the only inferences that suffer.

However, even if random sampling of characters had been used, and Darwin's scenario or the unreplicated burst had been revealed in the data, we argue that one should still dismiss the pattern as evidence for an adaptive/functional association. Our very concept of phylogeny, of change and inheritance along lineages, predicts that there should be many traits throughout the genome that evolved along an ancestral lineage and were inherited by its descendants. Common descent generates many sets of co-distributed characters that would be unlikely if species evolved independently. Many of these would be sufficiently closely co-distributed to fur and middle ear bones—for example, milk or enucleate erythrocytes—so as to be alternative explanations for any purported correlation. This is not just an effect of genomes being so big that there will, by chance, be variables with concordant distributions. Because of inheritance along lineages, there are likely hundreds or thousands of other traits with distributions that could equally be claimed as correlated, even if we have not yet discovered them. How can we pin our functional or adaptive story on just one of these traits? Do we just hope that other biologists do not discover the other traits before we publish?

On the other hand, simple inheritance from common ancestry does *not* predict that there should be many other traits co-distributed with X in Figure 1a,b, unless these traits have some special relationship to X. It is indeed difficult to explain Figure 1a,b without invoking an adaptive/functional relationship.

The problem with accepting Darwin's scenario or the unreplicated burst as implying correlated evolution is that it would seem to open the door to innumerable papers claiming interesting correlations, significant by comparative methods, between pairs of synapomorphies for the same clade (Darwin's scenario) or between one synapomorphy and a character with a locally high rate of change (unreplicated burst). One might argue that most of these papers would be rejected for lack of a plausible mechanism between the arbitrarily chosen characters. Indeed, we have used examples such as milk and fur, or enucleate erythrocytes and middle ear bones, precisely because they seem ridiculous: a causal link seems implausible. However, the imaginations of scientists are good, and there are many characters to choose from, so many could pass the test of plausibility. Also, such an argument places almost all of the burden on the plausibility of the mechanism, with the comparative pattern contributing almost nothing to support an adaptive/functional correlation between variables. In this article we are concerned only with what support is given by the comparative patterns.

Pseudoreplicaton Within a Lineage

Another explanation for the failure of Maddison's (1990) and Pagel's (1994) tests is that they mistakenly treat adjacent branches or adjacent infinitesimally small sections of lineages as independent, when in fact they can share common factors. Indeed, this assumption of independence is at the heart of the Markov process that model-based approaches use. This criticism was first raised by Read and Nee (1995) and Grafen and Ridley (1997). In essence, our methods should be careful to count separate *origins* as independent, and recognize that the homologous instances along a branch are pseudoreplicates.

Consider Pagel's (1994) method applied to the unreplicated burst (Fig. 1d). The method estimates parameters for the instantaneous rates of joint change among the character states, and uses likelihood ratios to test the hypothesis that the rates of change of one variable (say, Y) do not depend on the state of the other variable. To assess likelihoods, the method sums probabilities for possible scenarios of ancestral states and parameters. The probability of an instance of character Y changing from state 0 to 1 is counted according to the context in which it occurs: whether it happens in the context of state

0 or in the context of state 1 in character *X*. However, if different events of a change to state 1 in *Y* occur in the context of state 1 in *X*, the method does not pay attention to whether these instances of state 1 in *X* are homologous or not—they could all be homologous, coming from the same clade. As pointed out by Grafen and Ridley (1997), if the different contexts of state 1 in *X* are homologous, then they represent the same instances of state 1, and there is not as much evidence as there appears to be for a correlation. Read and Nee (1995) refer to this as "pseudoreplication of lineage-specific factors".

It is conceivable that this "pseudoreplication of lineage-specific factors" is simply another way to describe the ascertainment bias discussed above, insofar as both make reference to third, unstudied variables characterizing a larger lineage. However, the pseudoreplication explanation makes no appeal to non-random character choice, and so we suspect that this effect would remain even if the problem of ascertainment bias were solved. Indeed, we expect within-clade pseudoreplication would remain even if the stochastic model underlying Pagel's (1994) test were adjusted to be consistent with characters evolving as *X* in Fig. 1c,d. To solve the problem of pseudoreplication of lineage-specific factors, it would seem that a rather different approach to modeling is needed.

Vigilance Is Not Enough

When we discuss these issues with colleagues, we tend to get three alternative responses. Some share our concern that biologists may be seriously overestimating evidence for correlation, others consider within-clade pseudoreplication to be a minor problem and easily fixed by vigilance, while others are unconcerned, holding that significant results in Figure 1c,d do indicate biologically meaningful correlations. We have already presented a case against the last response: we have argued that making adaptive/functional interpretations merely from the observation that fur and milk are correlated in tetrapods is little different from making grand interpretations comparing just two data points (Garland and Adolph 1994).

But, can we easily guard against over-interpretation simply by being vigilant, filtering and discarding the obvious problem cases? Surely, a good biologist should be aware enough of the dangers of Darwin's scenario and the unreplicated burst that they would not claim a significant result based on such patterns. A good biologist could, in addition to doing a statistical test, simply inspect the data to see how many origins are contributing to the pattern. If there is only a single origin, then the result should be discounted. If there are sufficiently many separate origins, the result should be accepted.

However, even if we knew exactly how many origins there were, what would be our rule for decision? Figure 3 shows four more scenarios that highlight why we cannot get an acceptable procedure simply by adding vigilance to current methods. In Figure 3a, there are two perfectly concordant examples of change in *X* and *Y*. Are two origins enough for significance, and if so, why? What if the change is not precisely concordant (Fig. 3b)? Does that tip the balance to insignificance? And for replicated bursts, what if the independent origins of *X* are clustered on the tree (Fig. 3c)? Even if we knew with certainty that there were three origins of black in *X*, the enhanced rate of origins in *Y* could easily be explained by an unrelated variable that changed in the larger clade marked by asterisk. Are the origins of *X* far enough apart in 3d to rule out this alternative explanation?

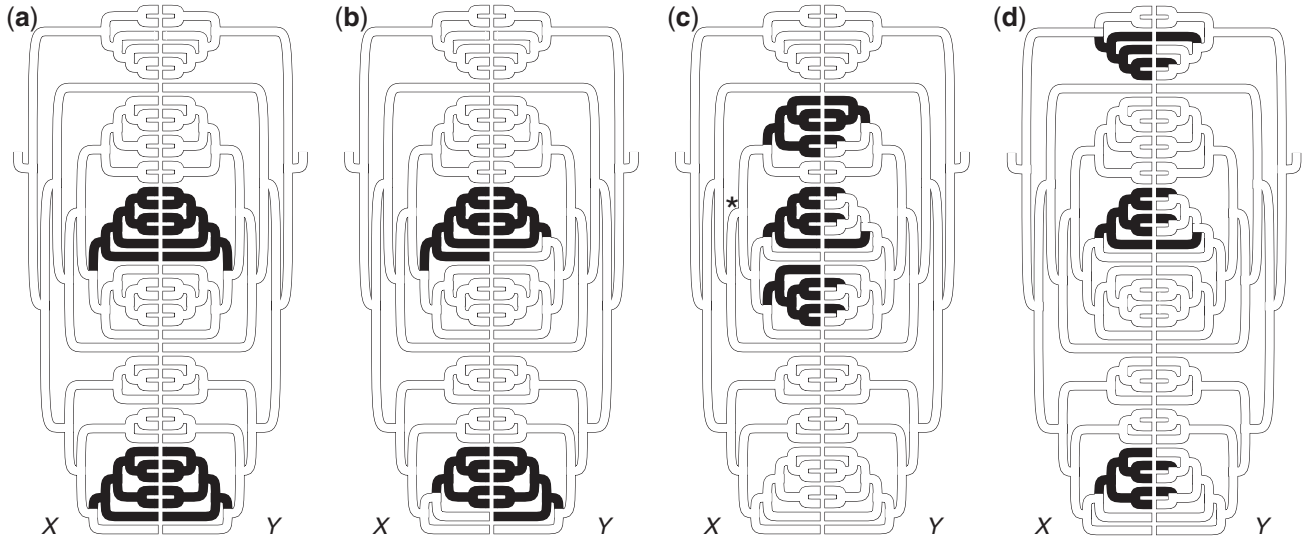

Figure 3. Four more scenarios for the evolution of states of characters *X* and *Y*. In each, the same phylogeny is mirrored to show *X* at left and *Y* at right. State 0 = white; State 1 = black. (a) Duplicated co-distribution. (b) Approximate co-distribution, duplicated. (c) Replicated bursts, but in close phylogenetic proximity. (d) Replicated bursts, more distant phylogenetically.

In each of the scenarios of Figure 3, we expect that tests like Pagel's (1994) or Maddison's (1990) would give strong support to a statistical association, but we also know that the result is contaminated, to an unknown degree, by the problems of unreplicated effects within each of the clades. Neither the tests nor our intuition tell us how serious are these problems, and thus whether to consider any correlation significant. As the number of independent origins rises, these problems would diminish, but how fast we do not know. We have no quantitative correction to apply to these methods, nor even a clear intuition as to whether there is sufficient replication, except in cases that are so clear that statistics are not needed to convince.

It appears unlikely that a satisfactory approach could be devised by adding to existing tests a step of counting origins. It would be far better to have all the evidence, including from the number of origins, summarized in a single well-defined quantitative method.

A Broadly Relevant Problem

The problems we outline go beyond tests of correlation between characters, and are suffered by other likelihood-based comparative approaches as well. Maddison et al. (2007) note that their BiSSE method, which investigates whether speciation and extinction rates are higher or lower in the context of a particular character state, likewise suffers from within-clade pseudoreplication. A significant result can arise from a single clade, and thus BiSSE can conclude only that the diversification rates depend on that character or any other character that might be co-distributed with it (Maddison et al. 2007, p. 708). FitzJohn (2010) found that a significant relationship between body size and speciation rate in primates inferred with the QuaSSE method could just as easily be explained by an unreplicated change in speciation rate attributable to the Old World monkeys. Thus, in our effort to develop methods to study diversification that use more of the information in the tree than was used by the older methods using sister-clade comparisons, we have lost the requirement for replication.

This susceptibility of QuaSSE to unreplicated patterns shows that our concerns are not restricted to categorical data. Could tests for correlations between continuous variables be susceptible as well? Felsenstein's (1985) method of phylogenetically independent contrasts can be misled by a single extraordinary event, but this is best considered a violation of the Brownian motion model, and can easily be detected by a contrast that stands as a distinctive outlier (Jones and Purvis 1997). Otherwise, the method is expected to be immune to the problems outlined here, because a significant result requires many separate sister clade comparisons that show an evolutionary event of change in both characters; independent contrasts intrinsically focuses on replication. Similarly, phylogenetic least squares (e.g., Martins and Hansen 1997) and related methods should be immune to our concerns, as they are equivalent to Felstenstein's under the Brownian motion model (Blomberg et al. 2012). However, we need to be aware that comparative methods detecting associations between continuous variables could in principle be susceptible. If a method were designed to detect an association between the values in one continuous variable and the rate of change in a second continuous variable, then a single clade of high value could lead it to report an association. Indeed, we suspect that any comparative method that responds to the effect of a state, rather than the effect of a change, will be susceptible to within-clade pseudoreplication.

Solutions?

Our characterization of the problem of within-clade pseudoreplication would have been more satisfyingly precise had we a solution in hand. Lacking a solution, we will discuss possible paths toward one. We are hopeful that a solution can be found, because even though our methods struggle, we can intuitively distinguish data that give strong evidence (Fig. 1a,b) from data that give no evidence (Fig. 1c,d).

The method of pairwise comparisons (Read and Nee 1995; Maddison 2000) is one possible solution. It chooses multiple phylogenetically independent pairs of species or clades to see if the difference in one character consistently predicts a difference in a second character. It can avoid the pitfalls of being influenced by a single origin of a character state by choosing pairs of taxa that contrast in the states of both variables (Read and Nee 1995). Because any patterns found are based on multiple cases of differences in both characters, they cannot be explained by a single factor in a broader lineage. In this way, pairwise comparisons avoids "pseudoreplication of lineage-specific factors". However, the method uses only a subset of taxa or branches, discarding much of the data (Felsenstein 1985), and so would likely have low power to detect correlations (Grafen and Ridley 1996). In addition, there are many arbitrary ways to choose pairs (Maddison 2000), some of which could bias the results. Pairwise comparisons may be an acceptable option for now, but it is not an ideal method.

Related to pairwise comparisons would be a method that decomposes the phylogeny into a series of clades in each of which there is a mix of both states of both variables, and performs a Pagel (1994) test or other test in each clade. Even if a single application of the test might be misled by a single origin, if the many applications indicated a consistent trend of correlation in the selected clades, we would have confidence that the correlation is not likely due to chance. This converts the analysis into a meta-analysis of small cases. It could be applied to a single group with multiple origins of the traits of interest, but it could also be applied by finding isolated clades scattered around the Tree of Life in which the two characters could be studied (e.g., Mayrose et al. 2011). Such approaches would suffer from inefficient use of information (as not all of the branches in the tree

can be used) and there is an arbitrary element to the decomposition. As noted above, more satisfying would be a single quantitative test or summary of evidence.

An ideal method would use the data as efficiently as likelihood, but would give credit to Figure 1a,b for having the black state of *X* distributed in three separate clades, and penalize Figure 1c,d for having the black state of *X* all arise from a single evolutionary event. The method needs to obtain power by extracting information from the whole tree, and yet avoid finding significant associations in cases like Figure 1c,d.

Given that the problem with Figure 1c,d is the possibility of a third character co-distributed by chance, one route to a solution may be to model hidden third characters explicitly. Grafen and Ridley (1997) develop a model with extra characters that underlie those observed, but this model has been used only to explain within-clade pseudoreplication, not to develop a statistical method that would provide a solution. Covarion methods (e.g., Penny et al. 2001; Beaulieu et al. 2013) could be adapted to model the effect of hidden variables. It is unclear to us whether modeling of hidden variables would provide a better null and surmount some or all of the problems. Such a method may need to account for the fact that our observed characters might have been chosen (intentionally or not) not by chance but in part based on their known distributions. Modeling the vagaries of character choice by biologists will not be easy.

The pairwise and decomposition approaches remind us that a good approach will likely incorporate a sense of contrast among sister clades. Thus, one possible route would be to seek a categorical character version of Felsenstein's (1985) independent contrasts method. Felsenstein (2012) has developed a method to measure covariances between discrete characters, where the categorical states arise by thresholds on underlying continuous characters. This method has yet to be developed and explored as a test, but is a promising avenue to pursue.

The Curvature of Biodiversity Time

According to the narrative of the field, the shift from non-phylogenetic to phylogenetic methods of studying character correlation was in essence a shift from counting *species* as if they were independent sample points, to counting *evolutionary events* as sample points. However, the problems of the current methods, and the field's failure to notice them, suggest that this shift to a phylogenetic paradigm remains incomplete.

It may be that we are so bound to the thin slice of time in which we live our lives that it is difficult for us to fully intuit a purely phylogenetic perspective. Maddison and Pérez (2001) explain this by analogy to the paradigm shift from a Newtonian cosmology to an Einsteinian cosmology. Since Einstein, space-time is said to be curved by mass, and so a meteor approaching Earth falls down this "gravity well". But, in a relativistic perspective, space-time only *appears* to be curved—we are misled by our Newtonian intuitions. On its own terms, space-time is flat; gravity defines a different sense of "straightness". The meteor that appears, in our view of space, to be curving into Earth is in fact simply continuing a natural undisturbed motion that is straight in space-time (Fig. 4a).

Biodiversity-time is curved by phylogeny. Before comparative methods became phylogenetic, they treated the straight-line comparison between extant species as the direct difference between them, from one living

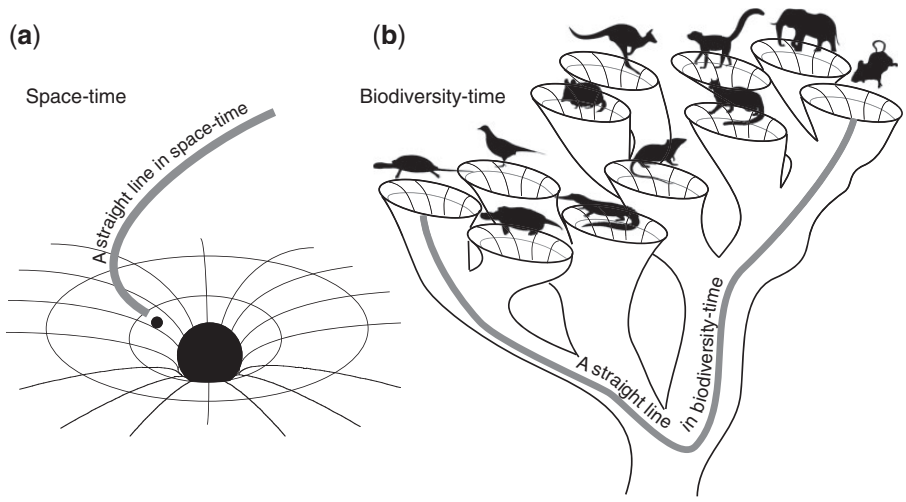

Figure 4. A relativistic perspective on motion in space-time (a), compared with a phylogenetic perspective on comparisons among lineages (b). Our intuition of a straight-line motion or comparison is incorrect in both cases. We are tempted to compare extant species directly, seeing differences in the thin horizontal slice of time we experience, but in fact a straight-line comparison in biodiversity-time is vertical, following changes along the lineages joining the species. Based on a similar figure in Maddison and Pérez (2001). Silhouettes of animals from http://phylopic.org; last accessed December 1, 2014. ©Michael Keesey under a Creative Commons 3.0 Unported or Share-alike Unported license, except the felid, which is in the public domain.

species to another. In using them, we restricted our perspective to the here-and-now, building our theories on patterns of differences, rather than patterns of change. The phylogenetic revolution in comparative methods moved them toward the new paradigm, following the lines of descent, modeling change along lineages. In this way they respected the natural straight-line comparisons between species, along the evolutionary lineages to the common ancestor and back up (Fig. 4b).

Yet, in some fundamental way, the problematical methods discussed here (including Pagel 1994) do not follow the lineages in biodiversity-time correctly. They treat phylogenetic lineages or lineage lengths as if they provide sample size regardless of whether change happens on them or not, counting small sequential pieces of lineage along lines of descent as independent. Pre-phylogenetic methods also counted small pieces of lineages as independent, namely the current populations of extant contemporaneous species. It appears that they share the same basic mistake, namely to count pieces of lineages, rather than to count evolutionary changes.

This we suspect is the fundamental mistake ("pseudoreplication of lineage-specific factors", Read and Nee 1995), compounded by non-random selection of characters and the dubious biological realism of the constant rates Markov model. They lead our methods to conclude significant association where mere coincidence is an available explanation. Given that many hundreds of other traits in the genome could provide equally good explanations for the evolution of a character, very little can be concluded by tests in such circumstances. We need new methods for categorical characters that avoid pseudoreplication and accurately assess strength of evidence from independent phylogenetic origins.

As long as we (or our methods) continue to see an association between X and Y when we look at Figure 1c,d, we will have failed to grasp the phylogenetic paradigm fully. Our current methods do not demand phylogenetic independence as much as we think they do. That we have missed their susceptibility to such a basic effect of shared inheritance suggests we are not thinking as phylogenetically as we should. When we learn how to build methods to compare properly along lineages, we will have fully come to the phylogenetic paradigm. We will see Figure 1c as revealing nothing other than a natural undisturbed motion in biodiversity-time, with no hint of interesting associations to be explained by evolutionary forces.

Summary: The Problem of Within-Clade Pseudoreplication

When the presence of a trait in a lineage is accompanied by a second trait's evolutionary change, we might be tempted to conclude that there is an interesting adaptive or functional relationship between them. However, if the first trait evolves only once, the apparent association can easily be attributed to coincidence followed by co-inheritance, because any other synapomorphy of the same clade is equally available as an explanation for influencing the second trait. We have known for decades that we need to use phylogeny to find independent evolutionary replicates to demonstrate correlations, and yet methods for assessing correlation of categorical characters (Maddison 1990; Pagel 1994; Huelsenbeck et al. 2003) will indicate association even if one of the traits arose only once. This flaw could be ascribed to a simple case of oversimplified models, with the troublesome character distributions being inconsistent with the stochastic model of evolution used. We suggest, however, that even were the models adjusted to predict such distributions, they would still be susceptible to the effects of single evolutionary changes. They would still lead to "pseudoreplication of lineage-specific factors" (Read and Nee 1995), by counting many events toward the likelihood even when they occur all with the same homologous instance of one of the traits. A second contributing problem could be non-random character choice, with our studied characters enriched in those that characterize well-known clades. While biologists may be wise enough not to publish when there is only a single origin, scenarios with a few or nearby origins show that the problem could often be more subtle and difficult to recognize. Within-clade pseudoreplication is a problem beyond categorical character correlation, as it also affects methods to study diversification and continuous characters (Maddison et al. 2007; FitzJohn 2010). We need to reform our methods, and it may require a rather different approach to modeling character evolution.

Acknowledgements

We thank Sally Otto and Arne Mooers for important discussion over the years on this conundrum. Mark Westoby helped us to understand alternative perspectives on comparative analyses. John Acorn suggested the example of gliding in mammals. We thank Sally Otto, Mark Holder, David Maddison, Heather Proctor, Arne Mooers, Matt Pennell, and two anonymous reviewers for many helpful comments on the manuscript. Of course, there is no implication that they would agree with everything (or anything) we say.

Funding

This work was supported by a Vanier Commonwealth Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) [to R.G.F.] and a Discovery Grant from NSERC Canada [to W.P.M.].

References

Barker D., Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1:24–31.

Beaulieu J.M., O'Meara B.C., Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. Syst. Biol. 62:725–37.
Blomberg S.P., Lefevre J.G., Wells J.A., Waterhouse M. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol. 61:382–91.
Darwin C.R. 1872. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 6th ed; with additions and corrections. London: John Murray.
Dudley R., Byrnes G., Yanoviak S.P., Borrell B., Brown R.M., McGuire J.A. 2007. Gliding and the functional origins of flight: biomechanical novelty or necessity? Annu. Rev. Ecol. Evol. Syst. 38:179–201.
Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.
Felsenstein J. 2012. A comparative method for both discrete and continuous characters using the threshold model. Am. Nat. 179:145–156.
FitzJohn R.G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633.
FitzJohn R.G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3:1084–1092.
Freckleton R.B. 2009. The seven deadly sins of comparative analysis. J. Evol. Biol. 22:1367–1375.
Garland T. Jr, Adolph S.C. 1994. Why not to do two-species comparative studies: limitations on inferring adaptation. Physiol. Zool. 67:797–828.
Garland T., Harvey P.H., Ives A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.
Grafen A., Ridley M. 1996. Statistical tests for discrete cross-species data. J. theor. Biol. 183:255–267.
Grafen A., Ridley M. 1997. A new model for discrete character evolution. J. theor. Biol. 184:7–14.
Hadfield J.D., Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J. Evol. Biol. 23:494–508.
Harvey P.H., Pagel M.D. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.
Huelsenbeck J.P., Nielsen R., Bollback J.P. 2003. Stochastic mapping of morphological characters. Syst. Biol. 52:131–158.
Jackson S.M. 2000. Glide angle in the genus Petaurus and a review of gliding in mammals. Mamm. Rev. 30:9–30.
Jones K.E., Purvis A. 1997. An optimum body size for mammals? Comparative evidence from bats. Functional Ecology. 11:751–756.
Maddison W.P., Maddison D.R. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. Available from: URL http://mesquiteproject.org (last accessed December 1, 2014).
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. 2000. Testing character correlation using pairwise comparisons on a phylogeny. J. Theor. Biol. 202:195–204.
Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.
Maddison W.P., Pérez T.M. 2001. Biodiversidad y lecciones de la historia. In: Hernández H.M., García Aldrete A., Álvarez F., Ulloa M., editors. Enfoques contemporáneos para el estudio de la biodiversidad. Instituto de Biología, UNAM, Mexico. p. 201–220.
Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.
Mayrose I., Zhan S.H., Rothfels C.J., Magnuson-Ford K., Barker M.S., Rieseberg L.H., Otto S.P. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257–1258.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51:729–739.
Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. Biol. Sci. Ser. B 255:37–45.
Paradis E., Baillie S.R., Sutherland W.J., Gregory R.D. 1998. Patterns of natal and breeding dispersal in birds. J. Anim. Ecol. 67:518–536.
Penny D., McCormish B.J., Charleston M.A., Hendy M.D. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53:711–723.
Purvis A., Gittleman J.L., Cowlishaw G., Mace G.M. 2000. Predicting extinction risk in declining species. Proc. R. Soc. Lond. Biol. Sci. B 267:1947–1952.
R core team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Read A.F., Nee S. 1995. Inference from binary comparative data. J. theor. Biol. 173:99–108.
Ricklefs R.E., Starck J.M. 1996. Applications of phylogenetically independent contrasts: a mixed progress report. Oikos 77: 167–172.
Ridley M. 1989. Why not to use species in comparative tests. J. theor. Biol. 136:361–364.
Ridley M., Grafen A. 1996. How to study discrete comparative methods. In: Martins E.P., editor. Phylogenies and the comparative method in animal behavior. Oxford University Press, New York and Oxford. p. 76–103.
von Mering C., Krause R., Snel B., Cornell M., Oliver S.G., Fields S., Bork P. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417:399–400.
Westoby M. 1999. Generalization in functional plant ecology: the species-sampling problem, plant ecology strategy schemes, and phylogeny. In: Pugnaire F.I., Valladares F., editors. Handbook of functional plant ecology. New York: M. Dekker. p. 847–872.