doi:10.1111/j.1558-5646.2010.00956.x

DELAYED-RESPONSE PHYLOGENETIC CORRELATION: AN OPTIMIZATION-BASED METHOD TO TEST COVARIATION OF CONTINUOUS CHARACTERS

Norberto P. Giannini[1,2,3,4] and Pablo A. Goloboff[1,5]

[1] *Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto Miguel Lillo, Tucumán, Miguel Lillo 205, CP 4000, San Miguel de Tucumán, Argentina*

[2] *Programa para el estudio de Biodiversidad Argentina*

[3] *Department of Mammalogy, American Museum of Natural History, New York, New York 10024*

[4] *E-mail: norberto@amnh.org*

[5] *E-mail: pablogolo@csnat.unt.edu.ar*

Received April 15, 2009
Accepted November 3, 2009

A new phylogenetic comparative method is proposed, based on mapping two continuous characters on a tree to generate data pairs for regression or correlation analysis, which resolves problems of multiple character reconstructions, phylogenetic dependence, and asynchronous responses (evolutionary lags). Data pairs are formed in two ways (tree-down and tree-up) by matching corresponding changes, Δx and Δy. Delayed responses (Δy occurring later in the tree than Δx) are penalized by weighting pairs using nodal or branch-length distance between Δx and Δy; immediate (same-node) responses are given maximum weight. All combinations of character reconstructions (or a random sample thereof) are used to find the observed range of the weighted coefficient of correlation *r* (or weighted slope *b*). This range is used as test statistic, and the null distribution is generated by randomly reallocating changes (Δx and Δy) in the topology. Unlike randomization of terminal values, this procedure complies with Generalized Monte Carlo requirements while saving considerable computation time. Phylogenetic dependence is avoided by randomization without data transformations, yielding acceptable type-I error rates and statistical power. We show that ignoring delayed responses can lead to falsely nonsignificant results. Issues that arise from considering delayed responses based on optimization are discussed.

KEY WORDS: Continuous characters, phylogenetic comparative methods, phylogenetic correlation, phylogenetic dependence, randomization testing.

Phylogenetic comparative methods (PCMs) are intended to analyze data measured on taxa, which become biological sampling units exhibiting various degrees of statistical dependence due to common ancestry. The treatment of the phylogenetic information and which transformation, if any, is applied to the original data, varies widely across methods (see reviews in Harvey and Pagel 1991; Martins and Hansen 1996; Diniz-Filho 2000; Martins 2000; Diniz-Filho and Bini 2008). In general, such methods can be seen as specifications of a linear model (Martins and Hansen 1997). Many of the common PCM applications involve the relationship of two quantitative variables (continuous or frequency data) measured on a number of taxa. Instances of such relationships include the functional response of home range (e.g., Garland et al. 1992; Haskell et al. 2002), basal metabolic rate (White and Seymour 2003), ontogenetic change (e.g., Reilly et al 1997; Klingenberg 1998), and behavioral traits (Dial et al. 2008) to body size in a given lineage, mass of fruit to number of seeds (Niklas 1994), and in general any specific relationships for which data are considered above the level of biological individuals (i.e., taxa).

Optimization is a tool used to determine the originations of a character state during evolution, one of the essentials in comparative biology (e.g., Brooks and McLennan 1991). Conceptual or practical difficulties inherent to this method (e.g., numerous possible reconstructions in case of ambiguities, dependence of assigned nodal states) and the lack of implementation in most phylogenetic programs may help explain why few currently available PCMs for quantitative data are based on optimization. We propose here a new PCM for the correlation/regression of two quantitative characters on a given tree that explicitly incorporates the particulars of optimization. This is possible given the recent development in full of continuous character optimization (Goloboff et al. 2006, implemented in the computer program TNT; Goloboff et al. 2003, 2008) as a natural extension of Farris' (1970) multistate character optimization. We discuss here how the many aspects involved in the reconstruction of a continuous character on a tree affect a correlation test for two continuous characters evolving together, as well as the type of test required given multiple reconstructions, and the general dependence that permeates all phylogenetically explicit problems. Optimization allows for tracking changes locally at nodes; when that is the case, it is possible to take into account potential evolutionary lags in a direct way. Thus, central to this new PCM is the question of how to deal with lagged responses, so delayed pairs of *x* and *y* data can be formed if required by the data. We provide estimations of performance, error rates and power, and discuss interpretations that impact our current understanding of PCMs.

The Problem

An interest common to evolutionary as well as ecological, behavioral, or physiological studies is whether one character responds to change in another. For instance, there are functional reasons to propose that home range depends on body size in organisms such as mammals and birds (e.g., Calder 1996; Haskell et al. 2002). This problem involves a relationship of a response variable (home range) and an explanatory variable (body size), both measured on continuous scales (in km^2 and kg, respectively) in related taxa (species). A conventional regression of the raw *x*–*y* pairs of values measured on the terminals may account for the functional relationship of the two variables, with a predicted slope based on functional analysis (e.g., Haskell et al. 2002). However, such an exercise would fail to recognize the relationships among taxa and how this affects parameter estimation, so the degrees of freedom of such a regression would be inflated (e.g., Harvey and Pagel 1991). Therefore, a form of regression to account for the functional relationship and the phylogenetic relatedness among taxa is needed such that the significance of the former can be correctly assessed given the dependence in the latter.

Mapping each character separately and then finding some sort of association between the evolutionary changes inferred for each character would be one reasonable approach within the framework of "realized evolutionary correlation" of Martins and Garland (1991). However, this encounters a number of problems in the context of ancestral character reconstruction, both at the level of the individual characters and the relationship in itself. First, the optimization of continuous characters may produce, as with any other type of character, ambiguous reconstructions at many nodes of the tree. This is not a specific weakness of optimization; it just follows from the recognition that observational data may imply more than one character reconstruction, irrespective of whether the original data are ambiguous or not. This has been a serious obstacle for using parsimony more widely in the comparative context. For instance, Martins and Garland (1991:538) choose to include in their performance tests only minimum evolution methods that minimize the sum of squared changes, thereby excluding parsimony, "to avoid complications due to multiple solutions." How should such ambiguity be dealt with? In addition, the number of possible values to assign to a given ambiguous node in the case of continuous characters can be exceedingly large—it is theoretically infinite within the inferred range of nodal values. Within one character, this applies not only to single nodes, but to all ambiguous nodes taken together, which is potentially a very large combinatorial problem. But there are two continuous characters to look at, so the real dimension of the problem is at the level of combinations of reconstructions of both characters. This can be intractable for real cases.

Second, there is no reason to assume that changes in an independent character (*x*) must produce, in every case, an immediate response in another character (*y*). Many cases of possible evolutionary lags have been suggested (see Deaner and Nunn 1999, and references therein). Prominent examples include evolutionary changes in brain size lagging behind changes in body size in primates (Deaner and Nunn 1999), evolutionary change in the number of males in a group lagging behind the change in the number of females in primates (Lindenfors et al. 2004), and the delayed morphological response in the horse lineage to the spread of North American grasslands during the Neogene (Strömberg 2006). This may be a common phenomenon because a trait may be in the process of responding to a recent selective force, different lineages may respond at different speeds, and not all new demands on the organisms are necessarily strong enough to require an immediate population response (see Discussion, and Deaner and Nunn 1999). As stated by Maddison (1990) for binary characters, a change in *y* may be a response to a change in x that occurred before (several nodes down the tree). If these delays actually occur, the obligatory same-node comparison of x and y prevalent in many PCMs breaks down and may frequently produce false negatives. A way to simultaneously track both immediate and out-of-phase paired changes across the nodes of the tree is needed.

Third, the optimization process generates final nodal assignments that are functions of character states of both descendant and parent nodes. In fact, expressing this inherent dependence of values measured in terminals is desirable as it explains similarity by common ancestry, and this is exactly what optimization does using a set of logical rules (Farris 1970; Fitch 1971; Sankoff and Rousseau 1975). However, that dependence invalidates conventional statistical inferences so randomization testing has been used in many comparative situations, from analysis of real datasets to simulations of error rates and violations of assumptions (e.g., Martins and Garland 1991; Diniz-Filho et al. 1998; Blomberg et al. 2003; Giannini 2003).

These and other problems led us to develop a correlation/regression approach for continuous characters as optimized on a tree, implemented in the script DELCOR. This script is written in the macro language of the TNT program for phylogenetic analysis (Goloboff et al. 2003, 2008) and it is made available for application and modification (see Appendix S1). In the following sections, we lay out the new method and provide examples illustrating its application and properties.

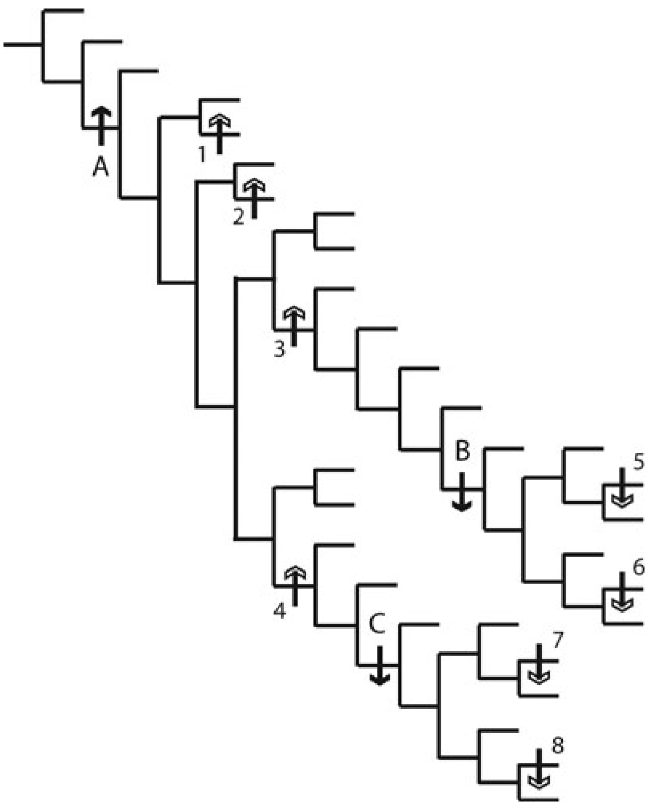


Figure 1. Artificial example of correlated evolution of two continuous characters, *x* and *y*, on a topology. Solid arrows represent changes in *x* (Δ*x*); open arrows represent changes in *y* (Δ*y*). Upward arrows represent inferred increases in either *x* or *y* characters; downward arrows represent decreases. "A" indicates an increase in *x* followed by delayed increases in *y* (tagged 1, 2, 3, and 4); "B" and "C" indicate decreases in *x* followed by delayed decreases in *y* (tagged 5 and 6 and 7 and 8, respectively).

FORMING THE *x–y* PAIRS: TREE-DOWN TESTING

Consider first the evolutionary response of each change of a continuous character y to changes in the continuous character x on a rooted tree. This is the approach taken for binary characters by Maddison (1990) in his method of concentrated changes: each change in y is potentially explained by a change in x, allowing the latter to happen simultaneously with, or earlier than, the change in y (i.e., at the same node or at a node further down the tree, respectively).

Let us consider a single reconstruction for each of two continuous characters, x and y, optimized on a given tree as in Figure 1. Both characters show several changes (signed increments) on the tree. Provided that x is an independent or predictor character (e.g., body size) and y is a dependent or response character (e.g., home range), it can be seen in Figure 1 that each increase/decrease in y is preceded in time by a corresponding increase/decrease in x some nodes down the tree. Although not a single one of those changes occurs at the same node for both characters, it is evident from this contrived example that in every case in which x changes, there is a change in y one or a few descendant nodes away of the same sign. A method based on same-node comparisons could easily fail to recognize such association. Therefore, it seems sensible to match a signed change in y, or Δy, with a signed change in x, or Δx, the latter occurring at the same node or at a parent node (i.e., delayed in time) in the tree with respect to Δy. We term this matching of a specific Δx and Δy a "delta pair", with the y response either delayed or not.

Systematically finding all the appropriate delta pairs requires traversing the tree in a down-pass (i.e., postorder traversal): starting at the terminals, the first Δy is tracked down the branches of the tree, until finding a change in x. The change in y is then accounted for and a delta pair is formed. The process continues from the next node down the tree until the root is eventually reached. If tracking a Δy does not result in finding a Δx in the path to the root, then a delta pair is formed with $\Delta x = 0$ (no change). As a result, a number of delta pairs are formed, some at the same node, some with Δx occurring at a parent node with respect to Δy. With this set of pairs, an observed point estimate of correlation r (or slope, b_1) can be calculated for the particular reconstruction. It must be emphasized that this r does not measure the correlation between x and y as observed but instead the correlation between changes

Table 1. Functions available in the DELCOR procedure.

Command	Default argument	Function
Tree	0	Select tree *N*
Chars	0 1	Specify independent (x) and dependent (y) characters from input data matrix
Minincx	0.000	Minimum (unsigned) increment in x to allow formation of a Δx–Δy pair (minincx$\geq$0)
Minincy	0.000	Minimum (unsigned) increment in y to allow formation of a Δx–Δy pair (minincy$\geq$0)
Radius	4.000	Maximum number of branches (or total sum of branch lengths) to be traversed searching for a y-change from the location of a x-change in either up or down tree direction (radius$\geq$0)
Delfac	1.000	Define delay factor ($DF\geq 0$). With default 1 down-weighting of delayed pairs is proportional to nodal or branch length distance between specific Δx and Δy
Sample	100	Maximum number of reconstructions to evaluate during calculation of observed correlation
Randwts	No	If yes, apply a new radius value to each node chosen at random between 0 and Radius. Else, use a fixed radius (default)
Cycles	5	If using "randwts," repeat random radius assignments *N* times ($N\geq 1$) for each reconstruction
Repls	100	Number of random replicates to generate the null distribution of r
Twotailed	No	If yes, perform two-tailed test on the observed r. Else, perform one-tailed test for either negative or positive r (default)
Testdown	No	By default, from a given Δx, traverse the tree in the root-to-tip direction in search of a Δy to form a pair (tree-up test). With "testdown," traverse in the tip-to-root direction in search of a Δx to form a pair (tree-down test)
Stasis	No	If yes, allow the formation of 0–Δy pairs (tree-up only). Else, skip nodes until $\Delta x\neq 0$ (default)
Userlengths	No	If yes, use supplied branch lengths to calculate weight of Δx–Δy pairs within defined radius. Else, use number of branches (default)
Notest	Test	With "notest," skip randomization, reporting only observed r and total number of combinations of reconstructions.

in x and y. An adequate test for correlation of these changes is proposed below.

A DELAY FACTOR

It can be argued that a y-response indicates a stronger degree of correlation if immediate (i.e., when an x-change and its y-response occur at the same node). Therefore, delays in the y-response can be penalized by weighting delta pairs. If the y-response occurs B branches higher up in the tree with respect to Δx, this delta pair is downweighted such that

$$w = \frac{1}{1 + (B \times DF)} \quad (1)$$

where w is the weight assigned to the delta pair, and DF is a delay factor varying between 0 and 1. Downweighting strength is thus determined by DF, with no attenuation of the contribution of the specific delta pair to the regression/correlation if $DF = 0$, and downweighting distant changes more strongly as DF increases. Default DF in our implementation is 1 (Table 1), which is roughly equivalent to downweighting on the basis of the distance (B) between Δx and Δy. DF only affects delayed delta pairs, given that for any value of DF, if the y and x changes both occur at the same node, the corresponding delta pair they form is given the maximum weight of 1 because $B = 0$ in equation (1).

The effect of weighting on the overall fit will depend on whether the downweighted pairs correspond to good or poor matches. Two delta pairs that are identical in sign and magnitude will be penalized differently according to how distant the y-response is. A poor delta pair formed at the same node (no delay) would decrease the overall fit (r^2) because it is not penalized ($w = 1$), but the regression/correlation would improve if the same poor match is established between distant nodes (because $w \ll 1$). Likewise, a "good" delta pair between distant nodes is penalized and therefore its contribution to the overall fit may be less than an average match at the same node.

ADDITIONAL CONSIDERATIONS

First, it may be necessary, in many cases, to define a minimum (unsigned) increase to be taken into account as a change in both x and y. With this in effect, only a Δy above the prefixed threshold is tracked and matched with a sizeable Δx to form a delta pair (this corresponds to the *minincx* and *minincy* variables in our DELCOR script, set by the user; Table 1). This is intended to prevent confounding effects from negligible changes (which can be common in optimization of continuous characters), and to allow for testing specific scenarios of evolutionary thresholds regarding the relationship between changes in x and y (see Discussion).

Second, the maximum distance (number of nodes) a Δy can be tracked down the tree seeking a Δx is limited to a given radius (optionally defined in the DELCOR procedure; (Table 1)). Note that this prefixed radius is an arbitrary value that applies equally to all nodes and therefore imposes a strong homogeneous component to the whole system under analysis. This assumption almost certainly does not hold, and of course the empirical value to be applied is unknown in the first place. This problem is ameliorated by randomizing the radius value at each node examined, making the radius to vary from 0 to a maximum value when calculating the observed *r*. To avoid making a given reconstruction too dependent on the particular random sample of weights affected by the radius so chosen, we suggest trying several "cycles" of random radius assignments per reconstruction (default is 5; Table 1). In this way, each reconstruction will have *n* random sets of radius values and associated *r*-values, thus ensuring a formation of delta pairs within a tree and across reconstructions that does not depend on one arbitrary choice of radius.

Third, as a consequence of the allowance for a delayed *y*-response, all Δx encountered within the prefixed value of radius are summed in the downtrack from Δy. That is, Δy is matched with the cumulative changes of *x* within the radius. In our implementation, the delta pair is downweighted with *B* equal to the number of nodes between the location of Δy and the location of the furthest Δx. So this composite, delayed response is the most penalized (with minimum weight), whereas the contribution of immediate, same-node response pairs is not penalized (maximum weight of 1).

TREE-UP TESTING

The approach presented above involves traversing the tree down from terminals to the root matching each Δy with a Δx in the same or at a parental node within the radius. This answers the question: are all changes in *y* explained by changes in *x*? There is an alternative question, however: do changes in *x* necessarily imply a change in *y*? These superficially similar questions are in fact subtly different, corresponding to asking whether a change in *x* provides either a necessary or a sufficient condition for change in *y*. The latter requires traversing the tree up from root to terminals (i.e., a preorder traversal), matching each Δx with all the descendants of its node with or without a *y* response. Given the bifurcating structure of the tree in this direction, a single Δx may cause more than one *y*-response, but also fail to elicit a response in some of the descendants. To consider these possibilities, it is necessary to locate the first Δx after the root, and visit all descendants above this node within the prefixed or randomized radius. If a Δy is found, a delta pair is formed and Δx is accounted for in that branch; else, a delta pair is formed with $\Delta y = 0$, either at a terminal or at the node at radius distance from Δx. The algorithm restarts above the node in which a change in *y* "paid" the change in *x*, until all the descendants within the radius are exhausted. This tree-up test, answering the question "Is change in *x* sufficient?" will generally produce more delta pairs than the alternative tree-down type of test answering "is change in *x* necessary?" A comparison of the tree-down and tree-up methods is given in Figure 2.

Two further issues remain. First, successive *x* changes may occur within the radius before encountering a Δy (or a nonresponse). These *x*-changes are summed so a cumulative Δx is matched along its path with either a Δy or a nonresponse. This final Δx value is downweighted with *B* equal to the number of nodes between the first Δx and the corresponding match. The underlying justification here is that a change in *y* may be triggered only after certain amount of change has been accumulated in *x*,

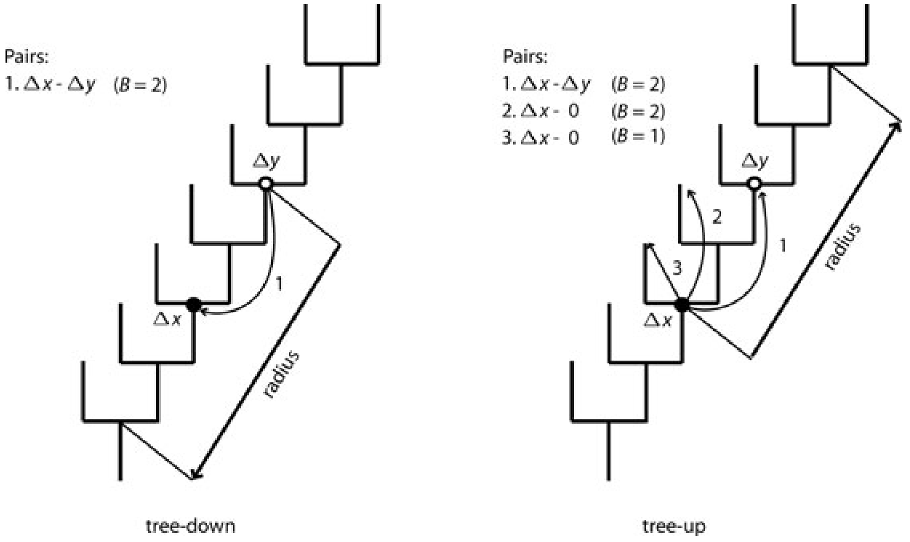


Figure 2. Comparison of the tree-down and tree-up matching schemes, showing the distinct formation of delayed delta pair under each method based on the same changes, with radius = 4. *B* is the number of branches between corresponding Δx and Δy. See text for details.

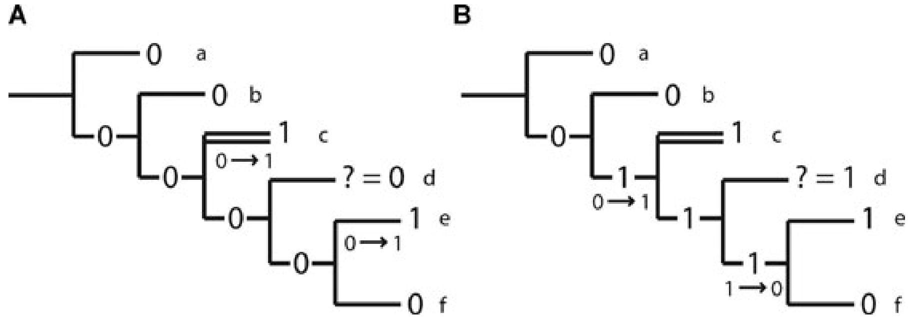


Figure 3. Alternative, equally parsimonious reconstructions in one binary character *x* originated in the ambiguity of the missing value in taxon d. Reconstruction A shows partial association between one of the two 0 → 1 changes in *x* and a single change in character *y* (double line in the tree). Reconstruction B implies no *x*–*y* association because none of the changes in *x* (0 → 1 and 1 → 0) is congruent with the single change in *y*.

or certain threshold is exceeded in that particular lineage before a response in *y* is produced.

The second issue is whether the fact that *y* does not change when there is no change in *x* should be taken as evidence of correlation. The answer may depend on particular cases, but a "stasis" function is provided optionally (Table 1) so that 0–0 pairs can be formed, as well as delta pairs with $\Delta x = 0$ (which are evidence against an effect of *x* on *y*).

MULTIPLE RECONSTRUCTIONS AND THE TEST STATISTIC

So far we have considered the association of two characters based on a single most parsimonious reconstruction for each character. The point estimate *r* calculated in the preceding sections (whether tree-down or tree-up) is in fact one of many possible values, each of which is associated with a particular combination of one reconstruction for character *x* and one reconstruction for *y*. The example in Figure 3 shows, for a binary character, why more than a single reconstruction has to be examined. A missing value generates ambiguity and two reconstructions are possible for character *x*; the unambiguous change in *y* partially matches the first reconstruction of *x*, but it does not match the second reconstruction. A sensible (and conservative) test must consider both possibilities.

Ideally, all possible combinations of most parsimonious reconstructions should be evaluated and an *r*-value calculated for each combination, which would produce an observed set or distribution of *r*-values. But, as anticipated, the total number of reconstructions for a continuous character can be very large–even with a restriction to use a few decimal places for each nodal value. We propose two heuristic solutions to approximate that distribution. First, we enumerate reconstructions for each of the characters (with the TNT command *iterrecs*, which generates all possible reconstructions of a character on a given tree) using as possible state assignments for each node only the limits of the nodal ranges (as opposed to using all values in between). This will necessarily contain the reconstructions with minimum and maximum possible values of Δx or Δy at each node, but it reduces dramatically the number of reconstructions to be examined for each ambiguous node for the characters while keeping the actual range of possible outcomes unchanged. Second, we select a random subset from all the possible combinations resulting from using the limits of the nodal ranges. As the delta pairs are formed between the reconstruction *n* of *x* and reconstruction *m* of *y*, the corresponding *r* value is stored. So in principle, in the correlation proposed here the test statistic is not a single *r*-value but a range of observed *r*-values obtained from the set of combinations of reconstructions examined.

A note on the test statistic

Many researchers are in fact more interested in the slope parameter (b_1) than in the correlation coefficient *r*. Our script reports both b_1 and *r* but our test is based on the randomized *r*. However, *r* and b_1 are equivalent test statistics from a randomization perspective (Manly 1997). That is, they yield the same *P* estimate because the only term of their respective formulae that changes as a consequence of randomization (the sum of the crossproducts) is the same in both (see p. 149 in Manly 1997).

THE PERMUTATION TEST

A permutation test for *r* in a conventional situation (see Manly 1997) involves randomizing the order of elements in the *x*- or *y*-vector, thus destroying the original pairing of *x* and *y* together with any within-character dependence (e.g., temporal, spatial). New *r*-values are calculated in every randomization cycle so that a null distribution of *r*-values is generated. Given that *r* varies between -1 and $+1$, counting the number of randomized *r*-values that happened to equal or exceed the observed *r* if $r > 0$, or the converse if $r < 0$, effectively approximates a test of significance provided that the number of permutations is large enough so the test is not biased by limitations of a small sample.

In our situation, the test statistic is not a single number but a range of r-values. A conservative test is thus derived by counting the number of times the randomized r-value equals or exceeds the smallest r-value of the observed range, if $r > 0$, or the largest r-value of the range, if $r < 0$. But what must be permuted?

An option in line with a conventional use of permutation is shuffling the original values, that is, exchanging at random x-values and/or y-values across taxa, reoptimizing both characters and recalculating (for each sampled reconstruction) each Δx and Δy, then weighting the delta pairs using the corresponding delay factor. However, the procedure seems not entirely appropriate given that the original x and y values are permuted but the test itself is in fact performed over the changes, Δx and Δy. It has two additional drawbacks: first, the number of nodes at which there is a change in x, or a change in y, as well as their magnitudes and signs, can be different as a result of this procedure; second, it requires reoptimization for every permutation, which can be computationally intensive. An alternative test involves permuting the signed increments themselves. The observed Δx are redistributed on the branches of the tree at random; the same is done with the observed Δy; the delta pairs are tracked (up or down); and a null r-distribution is generated. This procedure corresponds more closely to the intended test. In the case of a conventional permutation, the hierarchical correlation between the permuted character and the tree is destroyed. The null hypothesis being tested, however, is not that there is no correlation between the character(s) and the tree, but instead that the two characters (whether hierarchically correlated with the tree or not) have uncorrelated changes. The test based on permuted signed increments does not change whether the character is correlated with the tree–the amount of evolutionary change remains the same, but reassigned to new branches, independently of changes in the other character.

If the random radius option is chosen (Table 1), a single set of random assignments of radius values (different for each branch) is used in each replication, instead of the five (or more) cycles used for the observed r (see above). Although each replication is less intensely explored than the observed relationship, comparatively more replications can be done in a given time. The test is still conservative, because the worst correlation produced by the five (or more) cycles for the observed data is compared to the ones produced by randomization.

ARTIFICIAL EXAMPLE

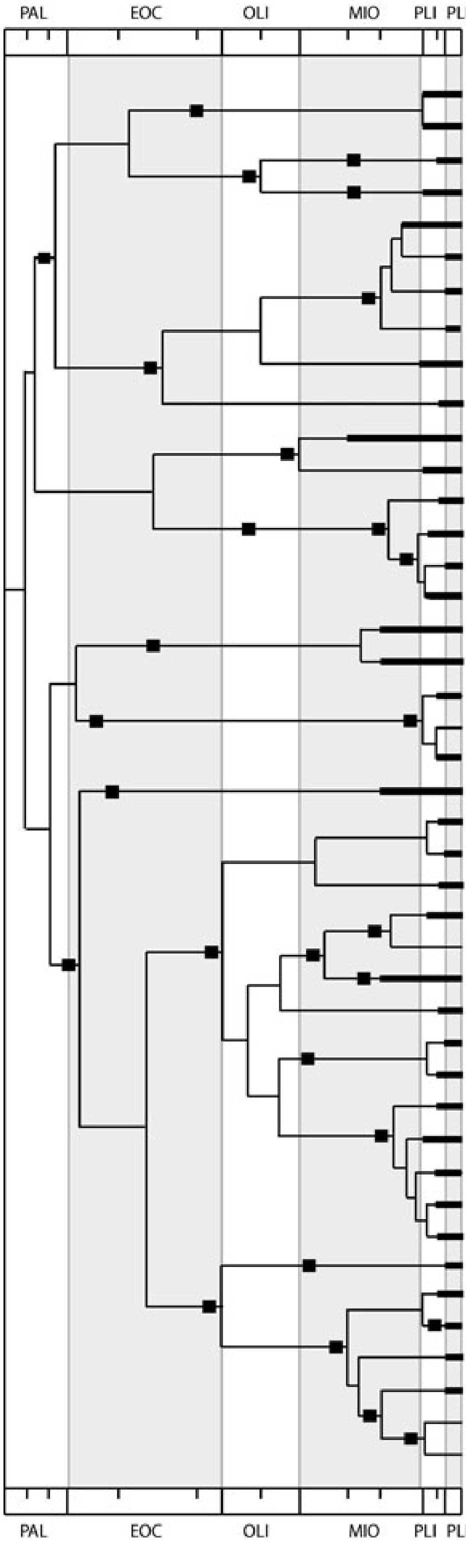
Simple examples show that, when the y-response is delayed, cases of obviously correlated changes may easily produce false negatives if analyzed requiring that changes in both x and y occur at the same node. The artificial dataset of Figure 1 was constructed such that a given change in y was always preceded by a change in x of the same magnitude at some parent node (delayed response) with the same sign (increase or decrease). In this example, the expectation is a strong positive correlation. However, correlation between terminal values is slightly negative and clearly nonsignificant ($r = -0.093$; $P = 0.500$). It may be thought that this is due to ignoring the phylogenetic relatedness of terminals, but in fact the same-node correlation between x and y states assigned to the internal nodes is even worse: $r = -0.286$; $P = 0.200$. This apparently negative correlation is because the delay in the response produces many cases in which x is small and y is large (or vice versa), which suggests negatively correlated values. Similarly misleading is a correlation between delta pairs on the same node ($r = 0.002$). Clearly, either of these ways to look at the data produces misleading results in the presence of a shifted y-response.

We tested this dataset with delayed-response correlation under a tree-up approach (radius = 4, $DF = 0.75$, 1000 replications) and obtained the expected significant result ($P = 0.039$) for an observed r-range of 0.779–0.841 (two possible combinations of reconstructions). As a test, we reversed the characters, that is, converting the original y into the explanatory character and the original x into the dependent character. This produced, as expected, a nonsignificant (and inversely proportional) result: observed $-0.474 \leq r \leq -0.451$; $P = 0.125$. As shown, our approach has the ability to cope with lagged responses, but immediate responses are also dealt with satisfactorily given that the weighting scheme in general favors them (weight, $w = 1$) over delayed responses (which are penalized, $w < 1$).

EMPIRICAL EXAMPLE

Garland et al. (1992) provided an example of character correlation between body size and home range in 43 ferungulate mammals (carnivorans + ungulates) on a composite tree built from different sources, which we reproduce with modifications in Figure 4. Body size varied from 2.5 to 2000 kg, whereas home range varied between 0.04 and 394 km^2. In this example, there were 28,080 combinations of most parsimonious reconstructions. We first tested this example under a tree-down approach, using a randomized maximum radius of four nodes with five cycles, with all branches set to unity, and sampling 100 combinations of reconstructions to calculate the observed r. The range of observed r-values included zero, so before doing any replication we concluded that there was no evident response of home range to changes in body mass in ferungulate mammals. We obtained the same result with a tree-up approach, and when stasis was in effect.

We then modified the ferungulate dataset by multiplying the body size of the carnivoran terminals by average pack size (see Fig. 4). This correction was applied only to carnivorans because, if social, they effectively collaborate to prey on the same item that they share, and defend the resource territory collectively so actual home range depends more on the group than on individuals. By contrast, ungulates in this example tend to socialize for



	Home range (km^2)	Body size (kg)	Corrected body size* (kg)
Ursus horribilis	82.20	251.30	251.30
Ursus americanus	56.80	93.40	93.40
Mephitis mephitis	0.09	2.50	2.50
Meles meles	0.87	11.60	11.60
Canis lupus	394.00	35.30	423.60*
Canis latrans	45.00	13.30	13.30
Lycaon pictus	160.00	20.00	180.00*
Canis aureus	9.10	8.80	8.80
Urocyon cinereoargenteus	1.10	3.70	3.70
Vulpes fulva	4.30	4.80	4.80
Hyaena hyaena	152.80	26.80	26.80
Crocutta crocutta	25.00	52.00	416.00*
Acinonyx jubatus	62.10	58.80	58.80
Panthera pardus	23.02	52.40	52.40
Panthera tigris	69.60	161.00	161.00
Panthera leo	236.00	155.80	1869.6*
Diceros bicornis	15.60	1200.00	1200.00
Ceratotherium simum	6.65	2000.00	2000.00
Equus hemionus	35.00	200.00	200.00
Equus caballus	22.50	350.00	350.00
Equus burchelli	165.00	235.00	235.00
Giraffa camelopardalis	84.60	1075.00	1075.00
Synceros cafer	138.00	620.00	620.00
Bison bison	133.00	865.00	865.00
Taurotragus oryx	87.50	511.00	511.00
Gazella granti	20.00	62.50	62.50
Gazella thomsonii	5.30	20.50	20.50
Antilope cervicapra	6.50	37.50	37.50
Madoqua kirki	0.04	5.00	5.00
Oreamnos americanus	22.75	113.50	113.50
Ovis canadensis nelsoni	14.33	85.00	85.00
Hippotragus equinus	80.00	226.50	226.50
Aepyceros melampus	3.80	53.25	53.25
Connochaetes taurinus	75.00	216.00	216.00
Damaliscus lunatus	2.20	130.00	130.00
Alcelaphus buselaphus	5.00	136.00	136.00
Antilocapra americana	9.65	50.00	50.00
Cervus canadensis	12.93	300.00	300.00
Alces alces	0.43	55.00	55.00
Dama dama	16.09	384.00	384.00
Rangifer tarandus	30.00	100.00	100.00
Odocoileus virginianus	1.96	57.00	57.00
Odocoileus hemionus	2.85	74.00	74.00

Figure 4. Empirical example from Garland et al. (1992). The tree represents relationships among some ferungulate mammals (carnivorans + ungulates) from diverse sources. We converted the tree to dated ultrametric by scaling branches to Myr based on fossil records of terminal taxa (thick branches) or clades (solid squares) using McKenna and Bell (1997) as primary source. The scale represents 65.5 Myr and each geological period is divided in Early, Late, and Middle (when applicable). Home range (km^2), body size (kg), and corrected body size (kg) are given for each terminal. Corrected body size is the product of individual body size by median pack size (from various sources) and is only calculated for those highly social carnivorans (marked *; see text). EOC, Eocene; MIO, Miocene; OLI, Oligocene; PAL, Paleocene; PLE, Pleistocene; PLI ,Pliocene.

reasons other than resource sharing (e.g., defense); although there exists extensive home range overlap among individuals in herding mammals (so that density affects resources and therefore effective home range size; Damuth 1981), they do not behave so strongly as cooperative consumers or communal territory defenders as social carnivorans do. The number of possible combinations of reconstructions in this modified dataset was 17,080. A random sample of 100 such reconstructions, each with five cycles of random radius assignments (maximum radius = 4), resulted in an observed r-range of 0.403–0.534 under the tree-down method. A thousand randomizations yielded $P = 0.029$ (28 instances of a randomized r equal to or greater than 0.403, plus the observed value). The tree-up method using the same settings also yielded a significant result, with r-range 0.404–0.542 and $P = 0.027$. Under stasis, a similar result was obtained ($0.399 \leq r \leq 0.530$; $P = 0.020$). It seems clear that (corrected) body size and home range are correlated in the lineage of ferungulate mammals. Differences among the three methods used (tree-down, tree-up, and tree-up with stasis) exist, but are minimal in this dataset, because there is a high number of changes in both x and y (and very few nodes without changes in the tree). Under these conditions, the three methods will converge in essentially the same results when a strong association is present.

We also performed tests on the ferungulate dataset and tree using branch lengths as estimated from the fossil record in each lineage (Fig. 4). The branches in the ultrametric tree in Figure 4 are thus scaled to time. Using the same settings as in previous analyses, the maximum radius of 4 becomes 4 million years (Myr), irrespective of how many nodes are traversed in the tree. This may seem too long a time to expect a character response; however, by way of example, Strömberg (2006) reported a delay of no less than 4 Myr in horses to respond morphologically to the spread of grasslands in North America. Therefore keeping such a radius for our exercise seemingly was within possible evolutionary delays. The tree-up procedure yielded an r-range of 0.388–0.565, which was still clearly significant ($P = 0.043$). The tree-up procedure yielded, as expected, a wider r-range (0.388–0.565) which was still clearly significant ($P = 0.043$). It must be noted that branches are not required to represent time, so the tree does not need to be ultrametric for the test to be performed. Branches may be scaled to substitution rate or any measure of character change.

TYPE-I ERROR RATE

To estimate the probability of rejecting the null hypothesis of no correlation when it is true (Type I (α) error rate), we simulated the evolution of continuous-character data on the tree topology from the preceding example of ferungulate mammals, and then tested the resulting data (final states on the terminals) using the DELCOR procedure with default settings. We generated independent data for x and y starting at the root of the topology and traversing the tree upward to the terminals. For each character, the root was assigned a value equal to the average between minimum and maximum possible most parsimonious assignments for the root on the observed data (1.469 for x, 3.550 for y). At each node, both characters x and y had a probability of change (equiprobably increasing or decreasing) equal to the proportion of branches of the tree on which there is some change for the observed data (0.476 for x, and 0.488 for y). This simulation could of course be done in other ways, but we have chosen to make the null hypothesis true (i.e., change in x and y uncorrelated) while at the same time using parameters as close as possible to those inferred from the observed data. Once the evolving characters reached the terminals, a final state was assigned and a new data matrix was generated. This matrix was submitted to the delayed-response correlations script, and the whole procedure was repeated 1000 times. The number of (incorrect) rejections of $r = 0$ is then an estimate of the α error rate, with the expectation that the null hypothesis of no correlation will be rejected no more than 5% of the times (testing at $\alpha = 0.05$) or no more than 1% (testing at $\alpha = 0.01$).

The smaller r-value at which the delayed correlations yielded a false significant result was 0.447, well above the median r-value (0.228) obtained in the simulations. Only 10 of 1000 replications incorrectly rejected the true null hypothesis at $\alpha = 0.05$, and only 1 at $\alpha = 0.01$. We conclude that the type-I error rate is highly satisfactory for a rather typical case (in terms of tree size and shape) as the ferungulate example.

POWER

The probability of rejecting the null hypothesis when it is false (power) was estimated as in the preceding section, with corresponding modifications. To make the null hypothesis false, the simulated characters x and y should be correlated, with evolution of y dependent on changes of x. Each character started with the same baseline probability P_b of change (with $P_{bx} = 0.476$ for x, and $P_{by} = 0.488$ for y, as in the previous example), but the correlation was simulated by increasing the probability of change in y following changes in x (inversely proportional to the distance to the node, if any, in which x had changed, and with probability of increase augmented if x had increased, and vice versa). That is, the probability of a change in y is augmented (baseline P_{by} + response P_r) after a change in x, and reset to its baseline value after the change in x is "paid" with a change in y. The actual formula used for P_r is

$$P_r = P_{by} + (P_{by}(2/(1 + (D_x/C)))), \qquad (2)$$

where D_x is the distance to the closest node in which x changes, and C is a correlation factor, making changes in y more strongly dependent of changes in x as it increases; we used C = 2. The

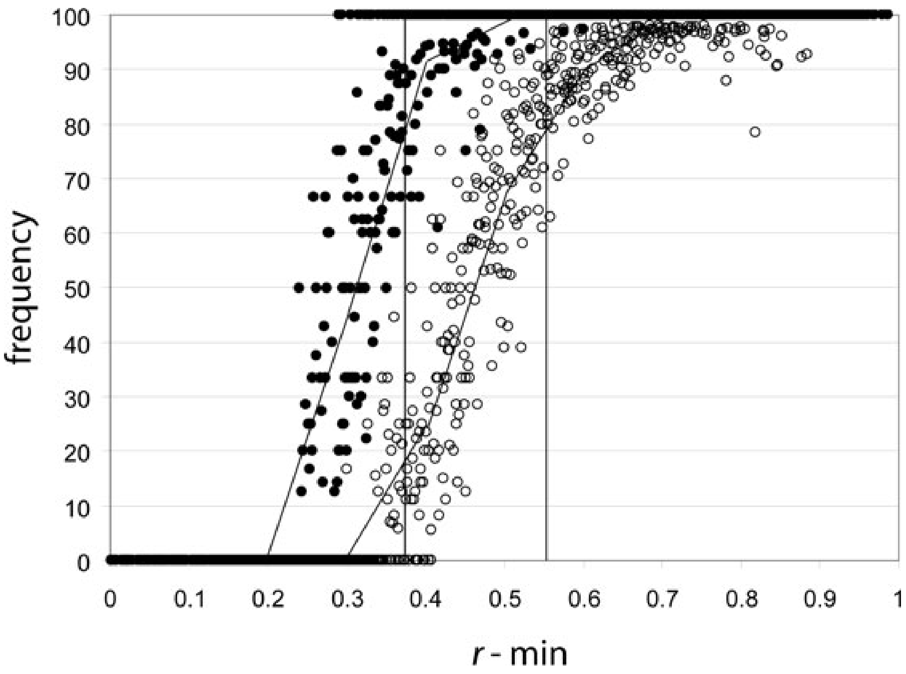


Figure 5. Power of a delayed correlation test represented as the percentage of correct rejections of simulated false null hypotheses of character coevolution at 5% alpha (solid circles) and 1% alpha (open circles) as a function of the *r*-values (i.e., effect size, on the *x*-axis) obtained in 15,000 replicates. Vertical lines indicate the *r*-level at which a conventional 80% power (*y*-axis) is achieved for each alpha level.

formula used for determining the probability that the change in *y* is an increase is

$$P_{\Delta y>0} = 0.5 \pm (0.5(2/(1 + (D_x/C)))) \tag{3}$$

using "+" if the change in the other character *x* represented an increase, or "−" if the change in *x* represented a decrease. As a result, a dataset of two correlated characters is generated and submitted to the DELCOR procedure with the expectation that delta pair correlations should be significant at a given α. However, power will be dependent on the effect size, that is, larger *r*-values will be more easily found significant. Therefore, we estimated the proportion of rejections of the false null hypothesis in each category of *r*-values thus generated (with categories 0.1–0.199, 0.2–0.299, . . . , 0.9–1.0). The procedure must be more computer intensive as compared to the type-I error rate simulation to ensure that enough cases are generated to cover each category of *r*-value.

Character coevolution was simulated 15,000 times and the most frequent *r*-value was 0.700. As expected, power was low for small *r*-values (Fig. 5), but power estimates showed a steep increase over the range of *r*-values so that a conventional 80% power (e.g., Zar 1995) was already achieved at $r \approx 0.37$ (range 0.314–0.469 testing at $\alpha = 0.05$) or $r \approx 0.55$ (range 0.461–0.607, testing at $\alpha = 0.01$). Larger *r*-values (i.e., $r > 0.6$) always yielded power estimates that were consistently >80% over the whole range of values up to $r = 1.0$ (Fig. 5). These power estimates are conservative as we used the lowest value of the *r*-range in each replicate.

Discussion

The goal of PCMs for continuous variables is to detect an association of two variables (in general, characters) whenever it exists, taking into account the evolutionary relationships among the sampled members of a given lineage. Here, we propose a PCM based on mapping the response of one continuous character *y* to changes in another, independent character *x*, using the correlation/regression of changes in both characters as optimized on a given tree, allowing delayed-response changes in *y* with respect to independent changes in *x*. The method escapes from phylogenetic dependence by applying a permutation test on the distribution of changes, showing a good performance in terms of Type-I error rates and power. Delayed correlations incorporate the complexities of character reconstructions, solving two types of related but subtly different problems that imply traversing the tree in different directions, the tree-up versus tree-down methods. Examples show the potential of delayed correlations to find patterns that, predictably, are not easily detectable with methods that either depend on same-node comparisons or rely exclusively on the values of the terminals. Here, we examine some of the properties and consequences of adopting such a strategy for a phylogenetic character correlation.

WHY DELAYED RESPONSES

Evolutionary responses need not be perfectly synchronous. Evolutionary lags have been considered in previous works in more or less explicit ways. One attempt at dealing with lags is that of Deaner and Nunn (1999), who proposed a variant of independent contrasts but applied only to selected same-node pairs (i.e., most of the tree and its potential phylogenetic information is discarded). Maddison (1990), in his method of concentrated changes for binary characters, allowed delayed associations in the sense that the response state change needed only to occur later on the tree, not necessarily in the same node. The method described here can be viewed as a variation of concentrated changes for continuous characters, but also with many additions including differential weighting of immediate versus delayed responses, inclusion of all possible combinations of reconstructions (or an unbiased random sample thereof), randomization testing based on the amount of evolution, and different testing procedures (tree-up versus tree-down).

Allowing delayed responses can be seen as a relaxation, and hence a generalization, of the same-node matching scheme imposed in most other methods, most prominently independent contrasts, that rely in some form of ancestral character reconstruction. As shown in our artificial example (Fig. 1), same-node comparisons may carry a serious decrease in power to detect correlation due to the confounding effect of shifting increase–decrease trends. We hypothesize that this may cause false negatives in tests of phylogenetic correlation or examining evolutionary lags. This aspect deserves further research, with careful scrutiny of the literature and documentation in reality. The same is demonstrably true for methods that rely only on values of the terminals (see also Martins and Garland 1991), which in fact represent an even more restricted subset of same-node correspondences. Whenever delayed responses occur, they will predictably decrease the power of same-node methods to the point of uselessness in the extreme case of a majority of delayed responses and changes that switch sign during evolution, as shown in our artificial example. Interestingly, delayed correlations are in theory not affected if lagged responses do not occur frequently in nature (i.e., if the *x* and *y* changes are normally synchronous) because the highest weight is given to immediate responses; so if all *y*-responses are indeed same-node responses the effect is one of maximum weighting of (highest reliance on) nodal values as reconstructed.

How frequent delayed responses occur in real datasets is unknown, but it seems clear that delayed responses can occur and may be common (e.g., Strömberg 2006). One reason is that each cladogenetic event does not necessarily cause a response in every evolving character. In other words, many intermediate nodes can be "silent" with respect to the evolution of a given character. Another reason is that obligatory same-node pairing requires the same frequency of evolutionary change in both characters; however, characters do not need to evolve in this way, if only because some characters are more complex than others. Deaner and Nunn (1999) provided an interesting example in this line: a discrepancy in primate brain mass and body mass has long been noted, which has been attributed to the fact that the brain is a complex organ whose changes may lag behind the relatively easier and faster changes in body mass. Using a same-node pairing, Deaner and Nunn (1999) failed to support this hypothesis; our new method to deal with evolutionary lags, which allows delayed responses in different nodes along the entire tree, reopens this question. Interestingly, Lindenfors et al. (2004) showed that the evolutionary change in the number of males in a group lags behind the change in the number of females, so females drive social evolution in primates through male competitive adjustment to changing female group size.

Another type of case concerns the response of morphology to a changing environment, like the evolution of the horse lineage that coped with the spread of grasslands in North America during the Neogene cooling. Strömberg (2006) has shown that the dental characters associated with grazing (hypsodonty) evolved at least 4 Myr later than the earliest record of continuous C3 grasslands in North America. Hypsodonty may have been a response to feeding in open habitats, as evidenced by functional analysis, but clearly horses responded with some delay (of millions of years) to the environmental change. According to Strömberg (2006:236), "explanations for the slow evolution of full hypsodonty may include weak and changing selection pressures and/or phylogenetic inertia." Diverse brachyodont (browsing) horses may have survived in increasingly marginal habitats until they finally developed full hypsodonty (or went extinct).

ON TESTING EVOLUTIONARY ASSOCIATIONS

One central principle underpinning our method to test delayed correlations is that evolutionary changes are the units of comparison among characters (as in character congruence in phylogenetic reconstruction). This means that terminal values are not the actual data and, consequently, the tree is the sole guide to find those data. That is, a different tree would produce different comparative data for the same terminal values. From this perspective, the comparative dataset is therefore a construct of interweaved dependencies not amenable to conventional statistical treatment, so that randomization is a sine qua non for testing patterns of character correlation. Keeping with this principle also means that changes themselves should be permuted rather than terminal values. This has two interesting consequences. One is that the amount of evolutionary change (steps on the tree) remains identical throughout the testing process. The second consequence is that terminal values, if estimated from the redistributed changes in each permutation, will frequently be different from the observed ones.

Combining these two aspects, observed terminal values become one set of possible values that may have occurred with a given amount of evolution on a particular branching pattern. This fulfills a crucial requirement of the generalized Monte Carlo testing framework; specifically, the potential for the null hypothesis to be true given that all possible values, including the observed ones, can be generated by rearranging the data in a systematic way (Manly 1997).

A not immediately obvious consequence of the principle outlined above is that detecting correlation in a phylogenetic context does not depend directly on amounts of homoplasy. Consider, for simplicity, two discrete characters evolving together, the aminoacid positions *h* and *m*, with the former responding to the latter at the same node or in a delayed fashion. It is possible that both aminoacidic positions evolve without homoplasy and still show a clear relationship if changes in *m* include many unique substitutions (e.g., A → C, D → K, etc.) followed by corresponding substitutions in *h* that are also unique. Therefore, provided that changes are tracked in a generalized way (i.e., allowing both synchronous and delayed matching), neither detectability nor conclusions on adaptation need to depend exclusively on homoplasy, as long as enough variation (i.e., alternative states or conditions) is present. This is in contrast with the view that homoplasy represents the best evidence for adaptations to common environments (Brooks and McLennan 1991; Brooks 1996; Blomberg and Garland 2002).

Correlations of any kind, including delayed correlations on a tree, can only tell whether an association of two characters is unlikely to occur at random. Therefore, delayed correlations cannot provide direct and conclusive evidence of adaptation or coevolution given that unknown/uncontrolled third factors may be at play. In the example above, two proteins seem to coevolve one in response to the other, when in fact both proteins may depend on changes on a third, unsampled protein that may limit the evolution of the two proteins of interest. If the third factor is known (e.g., body size simultaneously affecting both independent and response variables) the residuals of a regression can be used (Martins and Garland 1991). More often, comparative data of the two variables of interest are all we have, but this is not necessarily a hopeless situation: as stated by Maddison (2000), what matters is how effectively we can rule out alternative explanations. In this line, delayed correlations seem well prepared at least to stand the test of error rates, which makes a safe rejection or acceptance of the null hypothesis possible. As a consequence of the statistical decision, a hypothesis of adaptation may emerge from the strong test of delayed correlations, especially because the methods proposed here are conservative at several levels. But the ultimate test of association (or adaptation) will require additional background knowledge to complement the information packed in the scorings of the characters under test and this is beyond the reach of pure pattern recognition (Blomberg and Garland 2002).

ASSUMPTIONS AND LIMITATIONS

The new method depends on the properties of parsimony reconstructions. Although many researchers in the PCM field avoid parsimonious optimization methods (e.g., Farris 1970, 1983; Fitch 1971; Sankoff and Rousseau 1975) on the grounds that it may produce inconsistent results under models of evolution that assume homogeneity of rates for all characters (Felsenstein 1978, 1985), those concerns seem to have little justification in the present case, for three main reasons. First, even if a homogeneous model is assumed for continuous characters, because those characters can have large numbers of alternative states, parsimony is expected to converge to a maximum likelihood estimator (Felsenstein 1978; Steel and Penny 2000). Second, the concern that parsimony produces statistically invalid results is irrelevant in the context of mapping characters (and therefore in PCMs), because as shown by Tuffley and Steel (1997) the mappings of parsimony and likelihood are equivalent for individual characters. Third, it is clear that assuming an homogeneous model in the case of continuous variables is even less realistic than assuming it for DNA sequences: there is no compelling reason to accept homogeneous rates of change in characters like body size or characters in general (for a recent evaluation of the risks of such assumption, see Matsen and Steel 2007). Thus, a proper use of most parsimonious optimization is not expected to cause a systematic bias in ancestral-state reconstructions of continuous characters.

As with correlation in general, sample size may be limiting for the detection of delayed phylogenetic correlations. Power will obviously decrease when dealing with small trees (or trees with few changes). We established a cautionary lower limit to prevent execution of the DELCOR algorithm if less than five delta pairs are formed, but obviously a much larger sample of pairs will be needed to reliably detect a correlation (or to confidently accept the null hypothesis). This will depend on specifics of the dataset and we recommend a power analysis such as the one performed here for problems with relatively few actual changes as optimized in the tree.

The regression model used in DELCOR is least squares (LS, model I), which is fit for the problem of one variable responding to changes in another, explanatory variable. However obvious for body mass data and function-based analyses, this may not always be the case so a model II regression (e.g., a Reduced Major Axis approach, RMA) may be required. LS and RMA models are arithmetically connected via the product-moment correlation coefficient *r*, so the RMA slope readily obtains simply by dividing the LS slope by *r* (see Niklas 1994), both of which are reported by DELCOR (standard error is the same for both models). This effectively converts DELCOR in a model II method.

CONCLUSION AND FUTURE DIRECTIONS

Delayed correlations are intended to cope with the complex problems of multiple reconstructions of character evolution, asynchronous evolutionary responses, and inherent phylogenetic dependence. Type-I error rates, power estimates, empirical results, and properties and derivations described here suggest that this method has the potential to satisfactorily resolve many interesting problems in comparative biology. The methodology developed here is applicable to continuous characters, whereas Maddison's (1990) test considers discrete characters exclusively. Extending the current framework to allow the inclusion of combinations of discrete and continuous characters, as well as multiple explanatory characters, is work in progress. As we advance in this direction, we emphasize that these next steps will be greatly facilitated by the interconnections of general linear models, the power of randomization testing that fully comply with Monte Carlo requirements, as in the case of the DELCOR procedure, and the access to all reconstructions of any kind of character, most of which are made readily available by the present development. Therefore, we anticipate a wide range of situations to be covered under the guiding concepts of allowing weighted delayed pairing when required by the data and randomization testing of evolutionary changes as units of comparison.

ACKNOWLEDGMENTS

We thank J.A.F. Diniz-Filho and four anonymous reviewers for constructive criticism. We thank the Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET, Argentina). The program TNT is freely available at http://www.zmuc.dk/public/phylogeny/TNT, thanks to a subsidy of the Willi Hennig Society. NPG thanks partial support from National Science Foundation Grant EF-0629811. PAG acknowledges funding from CIUNT (26/G331), CONICET (PIP 6502), and ANPCyT (FONCyT 01314).

LITERATURE CITED

Blomberg, S. P., and T. Garland, Jr. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J. Evol. Biol. 15:899–910.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Brooks, D. R. 1996. Explanations of homoplasy at different levels of biological organization. Pp. 3–36 *in* M. J. Sanderson and L. Hufford, eds. Homoplasy. The recurrence of similarity in evolution. Academic Press, San Diego, CA.

Brooks, D. R., and D. A. McLennan. 1991. Phylogeny, ecology, and behavior. The Univ. of Chicago Press, Chicago, IL.

Calder, W. A., III. 1996. Size, function, and life history. Dover, Mineola, NY.

Damuth, J. 1981. Home range, home range overlap, and species energy use among herbivorous mammals. Biol. J. Linn. Soc. 15:185–193.

Deaner, R. O., and C. L. Nunn. 1999. How quickly do brains catch up with bodies? A comparative method for detecting evolutionary lag. Proc. R. Soc. Lond. B 266:687–694.

Dial, K. P., E. Greene, and D. J. Irschick. 2008. Allometry of behavior. Trends Ecol. Evol. 23:394–401.

Diniz-Filho, J. A. F. 2000. Métodos filogenéticos comparativos. Holos Editora, Riberão Preto, Brasil.

Diniz-Filho, J. A. F., and L. M. Bini. 2008. Macroecology, global change, and the shadow of forgotten ancestors. Global Ecol. Biogeogr. 17: 11–17.

Diniz-Filho, J. A. F., C. E. Ramos de Sant'Ana, and L. M. Bini. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19: 83–92.

———. 1983. The logical basis of phylogenetic analysis. Pp. 7–36 *in* N. Platnick and V. Funk, eds. Advances in cladistics, Vol. 2. Columbia Univ. Press, New York.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

———. 1985. Phylogenies and the comparative method. Am. Nat. 125: 1–15.

Fitch, W. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20:406–416.

Garland, T., Jr., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Giannini, N. P. 2003. Canonical phylogenetic ordination. Syst. Biol. 52:684–695.

Goloboff, P. A., J. S. Farris, and K. C. Nixon. 2003. TNT: tree analysis using new technology. Version 1.1. Available at www.zmuc.dk/public/phylogeny/TNT.

Goloboff, P. A., C. I. Mattoni, and A. S. Quinteros. 2006. Continuous characters analyzed as such. Cladistics 22:589–601.

Goloboff, P., J. S. Farris, and K. C. Nixon. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24:774–786.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology: Oxford series in ecology and evolution. Oxford Univ. Press, Oxford.

Haskell, J. P., M. E. Ritchie, and H. Olff. 2002. Fractal geometry predicts varying body size scaling relationships for mammal and bird home ranges. Nature 418:527–530.

Klingenberg, C. P. 1998. Heterochrony and allometry: the analysis of evolutionary change in ontogeny. Biol. Rev. 73:79–123.

Lindenfors, P., L. Fröberg, and C. L. Nunn. 2004. Females drive primate social evolution. Proc. R. Soc. Lond. B 271: S101–S103.

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.

———. 2000. Testing character correlation using pairwise comparisons on a phylogeny. J. Theor. Biol. 202:195–204.

Matsen, F. A., and M. Steel. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56:767–775.

Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in Biology, 2nd ed. Chapman and Hall, London.

Martins, E. P. 2000. Adaptation and the comparative method. Trends Ecol. Evol. 15:296–299.

Martins, E. P., and T. Garland, Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534–557.

Martins, E. P., and T. F. Hansen. 1996. The statistical analysis of interspecific data: a review and evaluation of phylogenetic comparative methods. Pp. 22–75 *in* E. P. Martins, ed. Phylogenies and the comparative method in animal behaviour. Oxford Univ. Press, Oxford.

———. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667.

McKenna, M. C., and S. K. Bell. 1997. Classification of mammals above the species level. Columbia Univ. Press, New York.

Niklas, K. J. 1994. Plant allometry. The scaling of form and process. The Univ. of Chicago Press, Chicago, IL.

Reilly, S. M., E. O. Wiley, and D. J. Meinhardt. 1997. An integrative approach to heterochrony: the distinction between interspecific and intraspecific phenomena. Biol. J. Linnean Soc. 60:119–143.

Sankoff, D. M., and P. Rousseau. 1975. Locating the vertices of a Steiner tree in arbitrary space. Math. Program. 9:240–246.

Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Strömberg, C. A. E. 2006. Evolution of hypsodonty in equids: testing a hypothesis of adaptation. Paleobiology 32:236–258.

Tuffley, C., and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.

White, C. R., and R. S. Seymour. 2003. Mammalian basal metabolic rate is proportional to body mass$^{2/3}$. Proc. Natl. Acad. Sci. USA 100:4046–4049.

Zar, J. H. 1995. Biostatistical analysis. 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

Associate Editor: S. Magallon

Supporting Information

The following supporting information is available for this article:

Appendix S1. Input format.

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.