Phylogenetic Targeting of Research Effort in Evolutionary Biology

Christian Arnold[1,2,*] **and Charles L. Nunn**[1]

1. Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138; 2. Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany

Submitted April 18, 2010; Accepted July 13, 2010; Electronically published September 21, 2010

Online enhancements: appendixes, data file.

abstract: Many questions in comparative biology require that new data be collected, either to build a comparative database for the first time or to augment existing data. Given resource limitations in collecting data, the question arises as to which species should be studied to increase the size of comparative data sets. By taking hypotheses, existing data relevant to the hypotheses, and a phylogeny, we show that a method of "phylogenetic targeting" can systematically guide data collection while taking into account potentially confounding variables and competing hypotheses. Phylogenetic targeting selects potential candidates for future data collection, using a flexible scoring system based on differences in pairwise comparisons. We used simulations to assess the performance of phylogenetic targeting, as compared with the less systematic approach of randomly selecting species (as might occur when data have been collected without regard to phylogeny and variation in the traits of interest). The simulations revealed that phylogenetic targeting increased the statistical power to detect correlations and that power increased with the number of species in the tree, even when the number of species studied was held constant. We also developed a Web-based computer program called PhyloTargeting to implement the approach (http://phylotargeting.fas.harvard.edu).

Keywords: comparative method, phylogeny, correlated evolution, taxon sampling, pairwise comparison.

Introduction

The comparative method has played a major role in uncovering adaptive trait evolution in biological systems (Ridley 1983; Harvey and Pagel 1991; Pagel 1999; Martins 2000). The comparative method has, for example, revealed links between mating systems and sperm competition in primates (Harcourt et al. 1981) and other animals (Møller 1991; Hosken 1997). The comparative method also supports a model of sexual selection in which females choose males on the basis of their ability to resist parasites (Hamilton and Zuk 1982), and it has been used to probe the origins of both parasitic and symbiotic associations (e.g., Hugot 1999; Lutzoni et al. 2001). More recently, comparative methods have been applied to study phylogenetic community ecology (Webb et al. 2002), for example, in the context of the phylogenetic overdispersion of mammalian communities (Cooper et al. 2008). The comparative method can also be used to address conservation issues (Fisher and Owens 2004), such as questions involving the factors that influence rates of extinction (Purvis et al. 2000*b*) and how the phylogenetic clumping of conservation threat status can lead to greater loss of phylogenetic diversity when species become extinct (Purvis et al. 2000*a*).

A comparative analysis requires data on a set of species relevant to a hypothesis of interest. Usually, however, data are available for only a fraction of the species in a clade, and data collection in both the field and the laboratory is expensive and time consuming. A proper selection of species to study is a nontrivial and multifaceted problem (Garland 2001; Westoby 2002) that has rarely been addressed in a systematic way. Instead, species are often chosen either randomly or subjectively (Westoby 1999; Faustino et al. 2010) because they are of "particular (and perhaps irrational) interest" (Garland 2001, p. 119). Two problems are introduced when species are chosen in an unsystematic way. First, the full range of variation is not used to test the hypotheses. Second, taxonomic gap bias may occur, meaning that data collection has been focused on a few "popular" lineages. These different kinds of biases—incomplete variation and gap biases—can make a momentous difference to the conclusions one draws. In studies of primates, for example, results of comparative research are likely to change when the sample is tilted toward terrestrial species rather than those that live in trees, because terrestrial species possess larger body masses, exhibit different locomotor patterns, and live in larger social groups

* Corresponding author; e-mail: carnold@fas.harvard.edu.

Am. Nat. 2010. Vol. 176, pp. 601–612.
0003-0147/2010/17605-52091$15.00.
DOI: 10.1086/656490

(Clutton-Brock and Harvey 1977; Martin 1990; Nunn and van Schaik 2002).

To address these issues, methods are required that quantify potential biases in comparative data sets and identify the species that should be studied in the future. Indeed, it is common to read in articles of comparative research that further sampling is required to validate the findings, because either the sample sizes were small or the sample was biased toward particular species within a clade (e.g., in the study of sleep patterns; Roth et al. 2006; Capellini et al. 2009; Nunn et al. 2009). Unfortunately often, however, only general guidelines for this selection process have been given, and these guidelines are often specific to the question of interest (Westoby 2002). To our knowledge, no method yet exists that is flexible and specific enough to address the crucial task of prioritizing future research in light of specific hypotheses about the apportionment of variation in relation to one or more ecological factors.

Only a handful of studies have investigated ways of systematically identifying species to study. For example, Ackerly (2000) compared the performance of different taxon sampling strategies and found their statistical performances to differ substantially. One of the algorithms he examined is based on the pairwise comparison approach (Felsenstein 1985, p. 13; Møller and Birkhead 1992; Oakes 1992; Read and Nee 1995; Purvis and Bromham 1997; Maddison 2000) and identifies meaningful comparisons by selecting species pairs that differ by a certain amount in the independent variable, following the suggestion of Westoby (1999). Although it overestimates the magnitude of the correlation, Ackerly (2000) showed that this design increases the statistical power to detect correlated evolution (see also Garland 2001). Major weaknesses of the method are that the threshold for when differences are large is arbitrary, that it is dependent on the data set, and that it must be set manually, which limits its applicability considerably. Mitani et al. (1996) considered sampling strategies in relation to testing competing hypotheses, while Read and Nee (1995) discussed the need to identify pairs that contribute for or against hypotheses. Similarly, Maddison (2000) presented a methodology for choosing species pairs in which each pair is "a comparison relevant for the question of interest" (p. 198). However, his method is designed for binary rather than continuously varying data and can handle only fully bifurcating trees, and thus it does not provide enough flexibility for identifying meaningful comparisons with real data.

The method of pairwise comparisons has been used frequently to identify meaningful comparisons. Several reasons exist for using pairwise comparisons. For example, the method of pairwise comparison relies on fewer assumptions (Ackerly 2000; Maddison 2000; Hearn and Huber 2006) than other methods. Thus, unlike phylogenetically independent contrasts (PIC; Felsenstein 1985; Harvey and Pagel 1991; Garland et al. 1992), pairwise comparison does not require a specific model of evolution or the estimation of states at interior nodes. In addition, some sets of species within a larger clade might not be directly comparable in standard implementations of comparative methods, such as PIC. In regard to mammalian sleep, for example, some cetaceans sleep with only one-half of their brains (Lyamin et al. 2008), making it difficult to compare the measurements of sleep in cetaceans with those in other mammals. The method of selecting specific pairwise comparisons provides a way to limit comparisons so that cetaceans are compared only with other cetaceans and noncetaceans are compared only with noncetaceans. Similarly, some behavioral experiments might require similar sensory modalities or cognitive ability among species in the data set. Pairwise comparisons of some close relatives may be more appropriate for selecting species for focused comparative experiments that take these factors into account.

When using the method of pairwise comparisons, it is important that all pairs are phylogenetically independent, that is, that no branches are shared among the comparisons (Felsenstein 1985; Maddison 2000). In figure 2, for example, different sets of phylogenetically independent pairs (which we call a "pairing"; see Maddison 2000) are shown for each tree. Thus, when selecting phylogenetically independent pairs, the selection of a particular pair constrains which other pairs can be selected.

Here we present a new approach, which we call "phylogenetic targeting," to systematically identify the species to be studied. Phylogenetic targeting is a taxon sampling approach that aims to prioritize future research by identifying species that should be studied in a target-oriented way under consideration of the specific hypotheses and data. It is not a new way to analyze comparative data or a substitute for existing analysis methods, but rather it draws on existing methods in comparative biology. This method uses the pairwise comparison approach and is based on a scoring system that incorporates phylogeny and data on variables relevant to testing hypotheses, specifically involving the predictor and response variables in a comparative test. The predictor variables can include potentially confounding variables or variables relevant to testing alternative hypotheses for an association. If external information suggests that comparisons should be restricted taxonomically or in relation to existing data, one can use the method to limit the species to be compared.

After assigning a score for each pair of species, phylogenetic targeting uses a newly developed algorithm to select the set of phylogenetically independent pairs of species that offer greater statistical power to test the hypothesis when data have been collected on the dependent variable.

After data collection, pairwise contrasts for the targeted species pairs can be used to test hypotheses, or one can use standard comparative techniques for testing correlated character evolution (fig. 1). This decision is up to the investigator and depends on the actual hypotheses, data, and analysis preferences (see "Discussion"). We use computer simulations to assess the degree to which phylogenetic targeting, compared with random sampling of species, increases statistical power for detecting correlated trait evolution. We have also implemented the method online (http://phylotargeting.fas.harvard.edu). We anticipate that the general approach developed here for pairwise comparisons can be adjusted for use with additional comparative methods, such as PIC or generalized least squares approaches, and we discuss some of these potential extensions.

Methods

The method requires a phylogeny and one or more explicit hypotheses that offer predictions for how variation in one trait (X_1) correlates with variation in another trait that is common to all the hypotheses and, because it is not known in all the species, is the "target" of the analysis (Y_t; fig. 1). We call this association between Y_t and X_1 the primary hypothesis. Additional hypotheses, if desired, are implemented through traits $X_2 \ldots X_n$, which relate to competing hypotheses or potentially confounding variables. The goal of the method is to identify species that should be studied, with regard to Y_t, through the use of phylogenetic relationships and data already collected for the X traits. Thus, a species cannot be included in a phylogenetic targeting analysis if data on the X trait are lacking for that species. We assume that larger evolutionary changes in X_1 provide higher statistical power for comparative tests to test the hypotheses, because they increase the available range of variation (Westoby et al. 1998; Westoby 1999; Garland 2001; Garland et al. 2005). We also assume that the characters show a linear relationship. Different targeting analyses are likely to focus on a primary hypothesis and various combinations of alternative hypotheses, and both discrete and continuous traits can be used. Scores, such that higher values indicate more preferred species to study, are calculated on the basis of user-defined criteria involving control of confounding variables, testing of alternative hypotheses, and availability of data on Y_t for one or more species in a clade.

Calculating Pairwise Comparisons

The analysis starts by calculating all possible $n \times (n - 1)/2$ pairwise comparisons. In the tree shown in figure 2, for example, 15 comparisons can be constructed. The method as such does not rely on using only pairs of sister species, as pairs of more distantly related species could also offer compelling tests of the hypotheses (Read and Nee 1995; Westoby 1999; Maddison 2000). Pairwise comparisons with missing data in any of the traits except Y_t are excluded. In addition, certain species can be excluded manually from the analysis, for example, in cases where an experiment can be applied to only certain species on the tree.

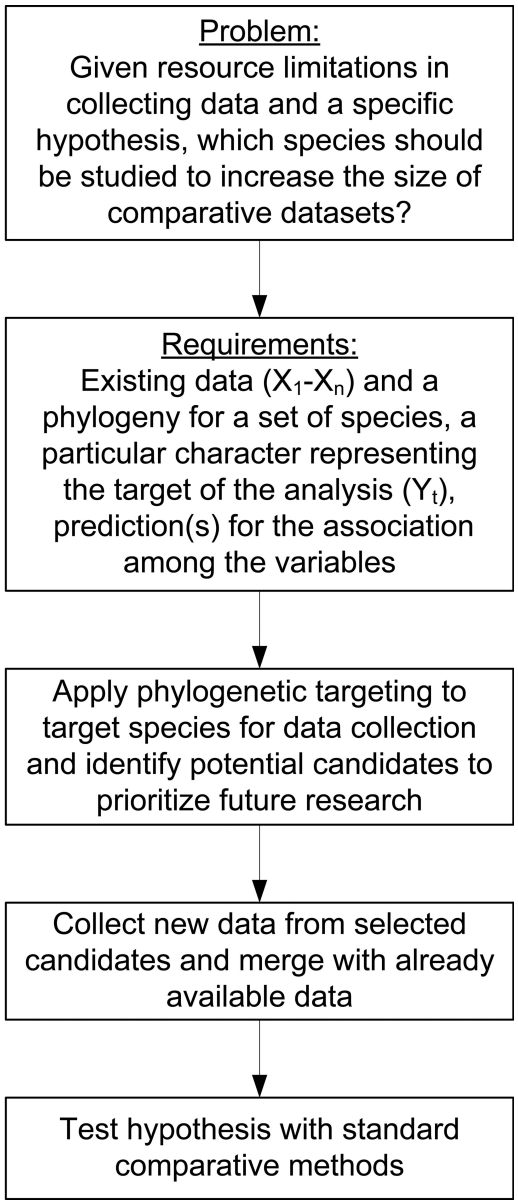


Figure 1: Flow chart for applying phylogenetic targeting. Phylogenetic targeting is essentially a taxon sampling technique to systematically guide future data collection.

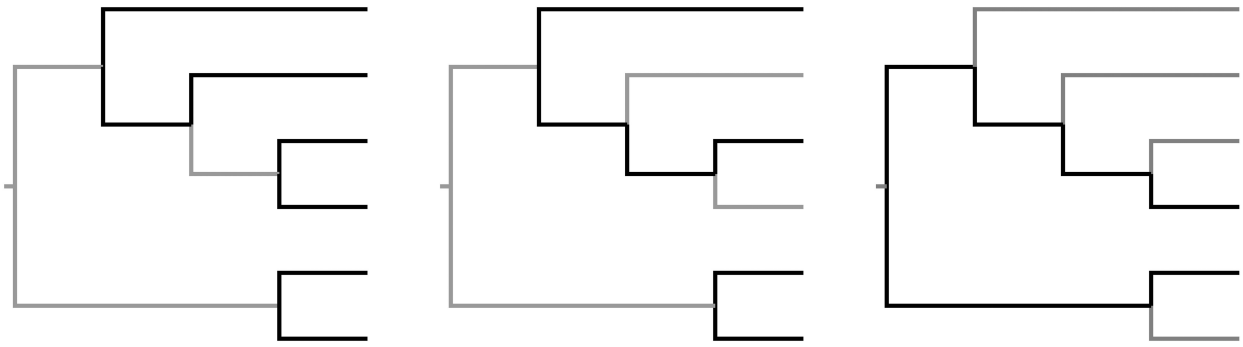


Figure 2: Three of the 15 possible pairings for an example tree. Paired species are highlighted in black. One pairing has three pairs, 10 pairings have two pairs, and four pairings have only one pair. In all pairings, pairs are phylogenetically independent, and no additional pair can be added without violating the requirement of phylogenetic independence.

If discrete characters with more than two possible states are used, they can be treated as ordered (costs between different pairs of states are different, as a particular sequence exists in which the states must occur through evolution) or unordered (every state change is equal, as each state can directly be transformed into any other state; Slowinski 1993).

Calculating Scores for Models with a Single Predictor (Y_t and X_1)

For predictions that involve only a primary hypothesis (i.e., only one independent variable), phylogenetic targeting uses a scoring system that maximizes the variability in X_1. In other words, species pairs that differ the most in X_1 are targeted. If we were interested in hypotheses that involve body mass as an independent variable, for example, phylogenetic targeting gives pairs with the largest differences in body mass higher scores. Thus, pairwise comparisons with large differences in X_1 are scored more positively, whereas smaller differences are scored less positively. These contrasts are then standardized to a scale of 0 to 1, with a difference of 0 assigned a score of 0 and the largest difference in all considered pairs assigned a score of 1. Note that even if no zero contrasts are found in the data, the method fixes the lowest contrast as 0. All other differences are assigned a score between 0 and 1 by applying a linear scaling transformation. We call this the score of X_1. If X_1 is an unordered discrete character, then the score will be either 0 or 1 regardless of the actual difference in character state assignments, whereas the difference is scored on an interval between 0 and 1 in the case of an ordered character, with the maximum number of character steps scored as 1.

Calculating Scores for Models with Covariates (Y_t, X_1, X_2 … X_n)

Models that incorporate additional traits enable the testing of different kinds of hypotheses (e.g., mutually exclusive and non–mutually exclusive), and they can be used to control for confounding variables. For each $X_2 \ldots X_n$, a separate scoring mechanism is defined in which larger contrasts have either a negative or a positive influence on the overall score. The decision as to whether larger differences in each of the variables X_2–X_n are scored higher or lower depends on whether the variables reflect confounding variables or a desire to distinguish among competing hypotheses. To simplify discussion, we consider a case in which only one additional variable is included; that is, $Y_t = f(X_1, X_2)$. Further details on the specifics of scoring are given below.

To control for confounding variables, the goal is to minimize variation in the predictor variable that corresponds to the confounding variable of interest, that is, X_2. Thus, pairwise comparisons in X_2 that make the absolute value of change in a particular confounding variable as small as possible are scored higher, whereas pairwise comparisons with larger differences are scored lower ($score_{NC}$, i.e., the score from standardizing the covariate for "no change"). The smallest pairwise contrast is assigned a score of 1, whereas the maximum pairwise contrast is assigned a score of 0. All other differences are assigned a score between 0 and 1.

To address mutually exclusive hypotheses, the goal is to maximize scores for X_2 that differ maximally from contrasts in X_1. Two different scoring options can be applied that both target large differences but differ in how they score these differences. The first option positively scores the differences in X_2 that are in the opposite direction from the difference in X_1, and it negatively scores differ-

ences in the same direction as X_1 ($score_{OD}$, i.e., the score from standardizing the covariate in the opposite direction). The largest difference in the opposite direction is assigned a score of 1, whereas the largest difference in the same direction is assigned a score of −1. A difference of 0 is assigned a score of 0. In analogy to models with a single predictor, the method fixes the lowest contrast as 0, even if no zero contrasts are found in the data. This ensures that all nonzero differences are assigned a score that is different from 0. All other differences are assigned a score between −1 and 1 by applying a linear scaling transformation, which is calculated separately for positive and negative contrasts. The second option is the opposite of the first option; that is, differences in the opposite direction from the difference in X_1 are scored negatively and differences in the same direction are scored positively ($score_{SD}$, i.e., the score from standardizing the covariate in the same direction). For example, this option might be useful if an increase in X_1 is predicted to reduce Y_t and an increase in X_2 is predicted to increase Y_t. Thus, it is necessary to give higher scores to contrasts in the same direction for X_1 and X_2 in order to distinguish among the hypotheses.

For models with covariates, the direction of change for $X_2 \ldots X_n$ always refers to the direction of change in X_1; for example, a positive value means that the direction of change is the same as in X_1. By doing so, we force the difference in X_1 (Δ_{raw}; see table 1) to be positive and achieve consistency with other widely used programs, such as CAIC (Purvis and Rambaut 1995) and PDAP-Mesquite (Midford et al. 2005). This "positivization assumption" also helps to make sense of the other trait differences and their directions when using the computer program, as it becomes possible to determine whether other pairwise comparisons are consistently positively or negatively associated with X_1 (e.g., if X_2 is positive, then it must be in the same direction as X_1). Although this is not strictly necessary for the algorithms implemented here, it helps to guide manual selection of contrasts in the Web-based implementation of phylogenetic targeting.

Summed Score and Standardizing Scores for Branch Lengths

For each pairwise comparison, the scores for all traits are summed up to define the summed score (see table 1 for a case involving X_2 as a confounding variable, i.e., $score_{NC}$). The summed score combines the information from all of the traits and thus represents the strength of a pair for testing the hypotheses. For models with Y_t and X_1 only, the summed score thus equals the score of X_1.

Regardless of the scoring model, summed scores can sometimes be uninformative when compared among different pairs because the more divergent two species are, the more likely it is that they have evolved larger differences. In other words, different pairs will have different expected amounts of change (i.e., variance). This problem can be overcome by normalizing the summed score by its expected variance (square root of the sum of the branch lengths that connect the two species; Felsenstein 1985; Gar-

Table 1: Illustration of the scoring system and the maximal pairing

Pairwise comparison	X_1		X_2				Summed score	Sum of branch lengths	Standardized summed score
	Δ_{raw}	Score	Δ_{raw}	$Score_{NC}$	$Score_{SD}$	$Score_{OD}$			
s1-s2[a]	.5	.385	−3	.831	−.171	.171	1.216	6	.496
s1-s3	.8	.615	−1.5	.916	−.086	.086	1.531	6	.625
s1-s4	1.3	1	−2.7	.848	−.154	.154	1.848	6	.755
s1-s5	1	.769	14.8	.169	.831	−.831	.938	8	.332
s1-s6	.6	.462	9.6	.461	.539	−.539	.922	8	.326
s2-s3	.3	.231	1.5	.916	.084	−.084	1.146	4	.573
s2-s4	.8	.615	.3	.983	.017	−.017	1.599	4	.799
s2-s5	.5	.385	17.8	0	1	−1	.385	8	.136
s2-s6	.1	.077	12.6	.292	.708	−.708	.369	8	.13
s3-s4[a]	.5	.385	−1.2	.933	−.069	.069	1.317	2	.931
s3-s5	.2	.154	16.3	.084	.916	−.916	.238	8	.084
s3-s6	.2	.154	−11.1	.376	−.634	.634	.53	8	.187
s4-s5	.3	.231	−17.5	.017	−1	1	.248	8	.088
s4-s6	.7	.538	−12.3	.309	−.703	.703	.847	8	.3
s5-s6[a]	.4	.308	5.2	.708	.292	−.292	1.016	2	.718

Note: As applied to figure 3. Δ_{raw} = raw difference of trait values (see fig. 3). See scoring section for details on $score_{NC}$, $score_{SD}$, and $score_{OD}$. Calculation of the summed score was based on the score of X_1 and the $score_{NC}$ scoring option for X_2; sum of branch lengths was calculated according to the tree in figure 2.

[a] Pairs that are selected in the maximal pairing.

land et al. 1992). We call this the standardized summed score. In this score, all pairwise comparisons have a common variance, which is required by most statistical tests (see also "Discussion").

Table 1 summarizes and applies the scoring system to the data set in figure 3, on the basis of controlling for X_2 as a confounding variable ($score_{NC}$). Different standardized summed scores would be obtained if we treated X_2 as representing a competing hypothesis and depending on the expected direction of X_2 in the context of competing hypotheses (see data for $score_{SD}$ and $score_{OD}$ in table 1).

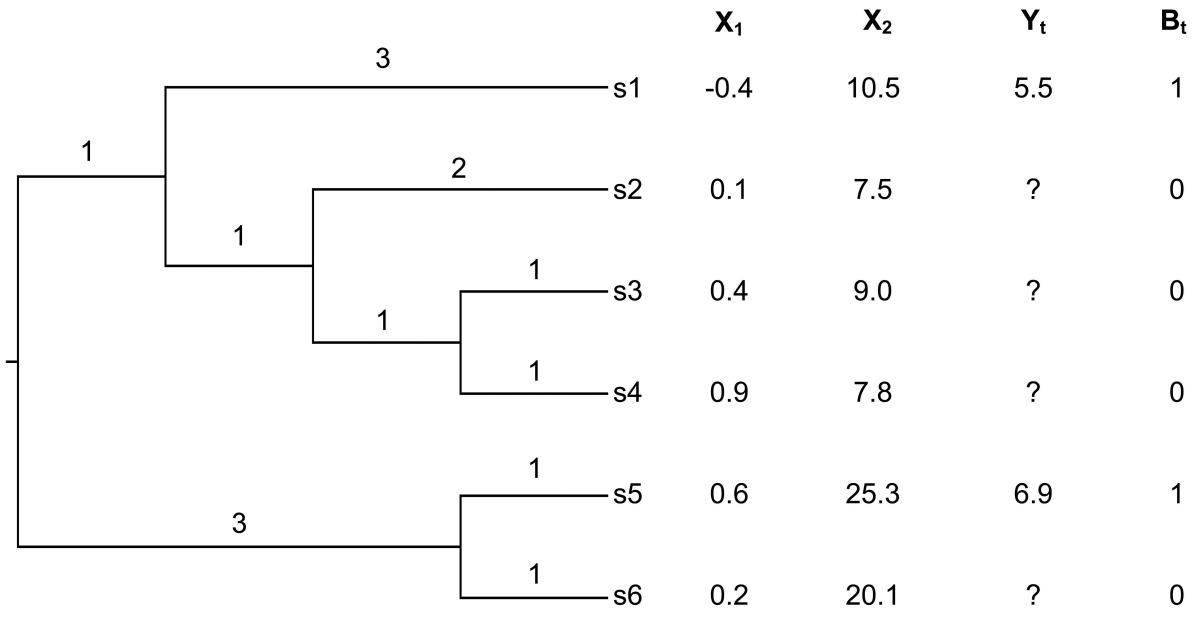
Availability Variable

In addition to manually excluding species from an analysis, it is possible to define an "availability variable" to automatically exclude species or pairs in relation to the availability of data for Y_t. One can thus use the availability variable to identify other species that should be studied in the context of existing data on Y_t. An availability variable also provides a way to quickly pinpoint where the missing data points are in a phylogenetic context, which can help to identify biases in the distribution of the studied species.

The availability variable must be a discrete binary variable that identifies whether data are available for Y_t for a particular species. For example, consider the scenario in figure 3, in which B_t is the availability variable. Possible options would be to consider only pairs where data are available for both species in the pair (exclusion of all pairs except s1-s5), for one species in the pair (exclusion of s1-s5 and all combinations of s2, s3, s4, and s6), for at least one species in the pair (as for when data are available for one species, but also including s1-s5), and for neither of the species in the pair (exclusion of the nine pairs with s1 and s5). Thus, this scoring procedure can be used in a variety of ways. For example, if the availability variable indicates that data are available for only a fraction of the species, then the majority of the pairs will be excluded if the option is chosen to consider only pairs where one species has already been studied and data are needed for the other species. In such a case, the pairs remaining are those containing one studied species and one that has yet to be studied. It can thus be seen as an additional selection factor that effectively constrains the species to be targeted.

Maximal Pairing Algorithm

The actual selection of species is performed by a dynamic programming algorithm that we call maximal pairing. The maximal pairing algorithm is a general optimization algorithm that selects pairs of species that are phylogenetically independent (see also Arnold and Stadler 2010). In contrast to PIC, where pairs can also involve internal nodes on the tree, the maximal pairing algorithm selects only



	X_1	X_2	Y_t	B_t
s1	-0.4	10.5	5.5	1
s2	0.1	7.5	?	0
s3	0.4	9.0	?	0
s4	0.9	7.8	?	0
s5	0.6	25.3	6.9	1
s6	0.2	20.1	?	0

Figure 3: Example data set and phylogeny for applying phylogenetic targeting. The tree shows continuously varying traits X_1, X_2, and Y_t, and a binary trait B_t, indicating whether the species has already been studied in relation to Y_t. Two species have already been studied regarding Y_t, and data on Y_t are missing for four species. The goal is to identify which of the four unstudied species should be targeted for studying Y_t.

pairs between the tips of the tree. The selection of pairs is based on the summed score for each pair, and the algorithm determines the set of phylogenetically independent pairs that maximizes the sum of the individual summed scores (table 1). This criterion is therefore assumed to maximize the power to test the hypotheses, given constraints on maintaining phylogenetic independence. With large data sets, it is difficult to find the maximal pairing manually, due to the large number of possible pairings and the complex phylogenetic dependence of pairs that must not share a branch (fig. 2). Despite some differences that involve execution time and representation of polytomies, the maximal pairing algorithm also works for polytomous trees (see app. A in the online edition of the *American Naturalist* for more details).

For models that involve only X_1, for example, maximal pairing generally selects pairs of closely related species that maximize differences in X_1; these pairs are often distantly related to the other pairs that are selected. In a comparative test, such a design is considered to be especially powerful (Garland et al. 2005). If, however, an additional trait X_2 is used to control for confounding variables (thereby scoring small differences in X_2 higher with score_{NC}), then the algorithm both maximizes differences in X_1 and minimizes differences in X_2. Conversely, if one aims to maximize differences in X_2 (thereby scoring larger differences in X_2 that are opposite in sign to X_1 higher with score_{OD}), then the algorithm maximizes differences in X_1 and maximizes differences in X_2 opposite in sign to corresponding differences in X_1. Similar logic applies to score_{SD}. It is worth noting, however, that because of the phylogenetic constraints and the standardizing of contrasts, maximal pairing does not simply select the pairs with the most extreme character differences; instead, pairs with small differences among closely related species are also frequently selected.

Simulations

We compared the performance of phylogenetic targeting with that of random selection of species, using simulations. The aim of the simulations was to generate data with known degrees of correlation between pairs of variables and then to select subsets of species either randomly or with phylogenetic targeting. To perform the simulations, we first generated phylogenetic trees and character data, using the GEIGER package (Harmon et al. 2008) in R (R Development Core Team 2009) according to a uniform birth-death process ($b = 0.15$, $d = 0$). We created 1,500 random phylogenies for a series of $N = 50$, 70, and 90 taxa. We then simulated character evolution for two continuously varying characters on each tree, using nine different models of evolution (table B1 in the online edition of the *American Naturalist*) with character states (0, 0) at the root of the tree. When simulating the non-Brownian-motion models of evolution, we transformed the tree in GEIGER (Harmon et al. 2008), simulated traits on the transformed tree, and then analyzed the data on the original tree, thereby simulating a case where the branch lengths failed to accurately reflect trait evolution (see app. B). Characters were simulated with a variance of 1 and correlations of both 0 and 0.5 for each tree. This procedure yielded 81,000 simulated data sets for analysis (1,500 trees across three sets of species, and simulating data under nine models of evolution and two levels of correlation, i.e., $1{,}500 \times 3 \times 2 \times 9$). For the Brownian-motion simulations reported in the text, analyses are based on 9,000 simulated data sets, with the remainder presented in the supplement.

Using these data and phylogenies, we selected subsets of species randomly and with phylogenetic targeting. In each simulation file, we selected the first simulated trait to be X_1; the second variable was assumed to be Y_t. We also standardized the scores. The maximal pairing was then calculated, and we selected the six highest-scoring pairs. We also randomly selected six phylogenetically independent pairs. To investigate whether the number of selected pairs impacts statistical performance, all analyses were repeated with 9 and 12 pairs.

To evaluate the statistical properties of both sampling approaches, we performed standard statistical tests on the basis of the selected pairwise comparisons. For this we used the character differences for X_1 and Y_t for the selected pairs and then standardized them by their expected variance (square root of the sum of the branch lengths that connect the two species). We tested for an association between the two characters based on correlation through the origin (Garland et al. 1992), using a t-test with $N - 2$ degrees of freedom and $\alpha = 0.05$. We determined Type I error rates (incorrectly rejecting a true null hypothesis of no association between traits) and statistical power (probability of rejecting a false null hypothesis) for both sampling approaches. Type I error rates were calculated as the proportion of significant results based on $P = .05$ for data sets in which $r = 0$, while statistical power was based on the proportion of significant results for data sets in which $r = 0.5$.

In addition to tests based on pairwise comparisons, we performed tests based on the full set of independent contrasts. We did this because many users may be interested in using a full set of contrasts, yet the method operates by examining pairwise comparisons. Thus, understanding the statistical performance of phylogenetic targeting when it is used with PIC is an important step and expands its application spectrum. After pruning the tree to the subset of the selected pairs, we calculated PIC (Felsenstein 1985), using the APE package (Paradis et al. 2004). We tested for

a significant correlation between the two characters, using the methods described above.

We also tested how the inclusion of randomly selected, nontargeted species affects the results. This simulates a common situation because data are often already available for some species but missing for others. Specifically, we examined how including k random species affects the results for tests based on pairwise comparisons and PIC (with values of k ranging from 2 to 10 in steps of 2). We included these additional species from the set of species that were not selected by phylogenetic targeting (and thus without using the availability variable).

Finally, we analyzed how much of the original range of variation in the simulated data was available after the data were pruned to the selected species. This gives insights to the range of variation that is available for hypothesis testing under the two sampling techniques.

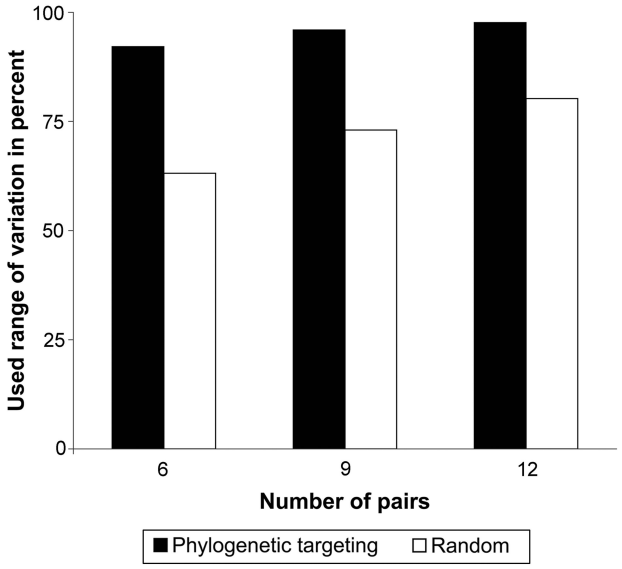


Figure 4: Simulation results for the percentage of the used range of variation for X_1 when species pairs are selected with phylogenetic targeting (*filled bars*) and randomly (*open bars*). The X-axis plots the effects of the number of pairs that have been selected (6, 9, and 12). Contrasts were standardized.

Results

PhyloTargeting Program

We created a freely available computer program, PhyloTargeting, that implements the phylogenetic targeting approach. It is Web based, taking the data as a Nexus file (Maddison et al. 1997) and providing a user-friendly, interactive, step-by-step interface, a variety of analysis options, and graphical visualizations of the results. The program is publicly available at http://phylotargeting.fas.harvard.edu.

Simulations

The simulations revealed that phylogenetic targeting substantially increases the range of biological variation that is sampled relative to random sampling (fig. 4). Phylogenetic targeting also provided substantially higher statistical power for detecting a true relationship (fig. 5). This held for both the pairwise tests and tests based on PIC. For the pairwise tests, Type I error rates for $\alpha = 0.05$ were elevated if the number of selected pairs was small, but they decreased to the expected level when more pairs were selected. For the tests based on PIC, Type I error rates were close to the expected level in all scenarios. Importantly, Type I error rates under random sampling and phylogenetic targeting were generally indistinguishable. All simulation results (including the results not highlighted in the article) are provided in an Excel file titled "Simulation Results," available in the online edition of the *American Naturalist*.

Increasing the number of pairs that are selected by the sampling algorithms increased statistical power, as was expected (fig. 5). For the pairwise tests, it also decreased Type I error rates. The number of taxa per tree, however, revealed a more surprising effect. Even when the number of pairs was held constant, the statistical power increased with the number of taxa in the clade under phylogenetic targeting, and Type I error rates did not increase (fig. 5). When species were selected randomly, however, power did not increase with increasing clade size.

When the true correlation was 0.5, mean values of r were elevated; moreover, they increased with the number of species per tree (see "Simulation Results" Excel file). Thus, a sampling regime based on phylogenetic targeting resulted in biased estimates of evolutionary trait correlations when $r \neq 0$, whereas a random selection of species resulted in no bias. Importantly, however, no bias was found when the true correlation was 0, as is shown in the results for Type I error rates. Furthermore, the bias decreased when additional, randomly selected species were included (see also "Discussion").

The results highlighted above are for a Brownian-motion process of character evolution. For the alternative models that we tested (see app. B), many of the results were comparable. However, for most of these analyses, Type I error rates were highly elevated and statistical power was reduced under the two sampling approaches and for PIC on the full tree (which we used as a control). Not surprisingly, the pairwise tests showed substantially less elevated Type I error rates if model assumptions were vi-

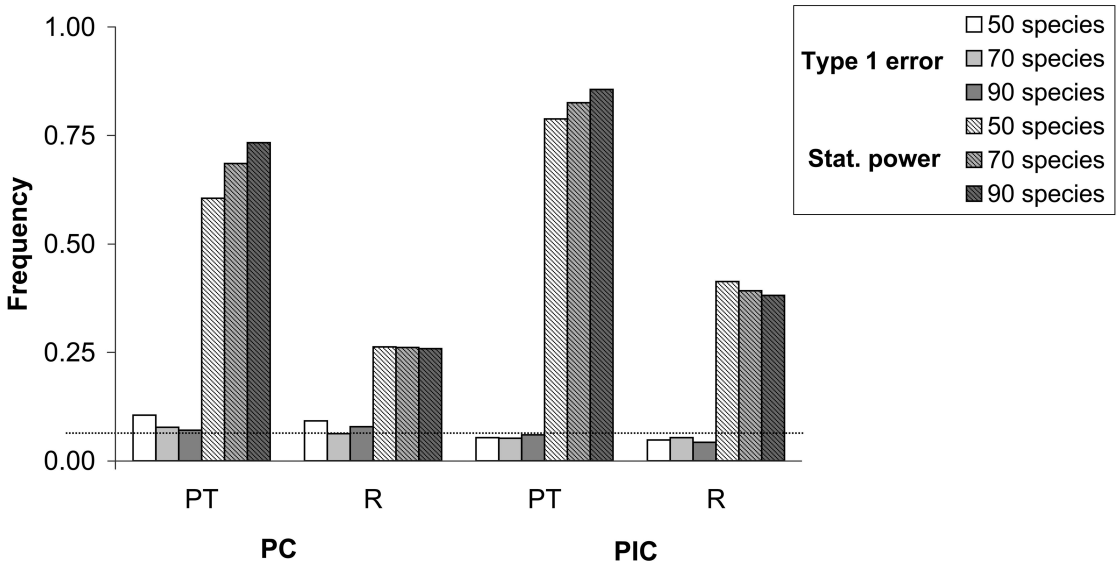


Figure 5: Selected results from the simulations under Brownian motion. Type I errors and statistical power for correlation tests based on pairwise comparisons (PC) and phylogenetically independent contrasts (PIC) are shown for phylogenetically targeted (PT) and random (R) taxon sampling. The first three bars in each category represent Type I error rates (based on 50-, 70-, and 90-species trees), and the last three bars represent statistical power (also based on 50-, 70-, and 90-species trees). Contrasts were standardized, and six pairs were selected.

olated, possibly because the method of pairwise comparisons relies on fewer assumptions.

Discussion

Comparative studies generally make use of available data. Here we show that the comparative approach can also be used to target species for future data collection. By applying the phylogenetic targeting concept, we can identify species that offer higher power to test predictions of a comparative hypothesis. Moreover, phylogenetic targeting provides a way to control for confounding variables when selecting species for further study or when testing competing hypotheses. The method will most likely be used to augment existing data, but it can also be applied to generate new data sets in the context of finite resources for data collection.

A major strength of the approach is that phylogenetic information is incorporated when selecting species to study (Garland 2001; Garland et al. 2005), thereby ensuring that the selected pairs are phylogenetically independent of one another. This makes it possible to analyze the data with standard statistical methods (i.e., pairwise tests). However, the simulations revealed that, when compared with PIC, statistical power here is reduced (see also Ackerly 2000). This may be due to the fact that for pairwise differences, the number of data points is reduced by a factor of approximately 2 because only the tips of the tree and not the interior nodes are contrasted. Furthermore, the bias in estimating the correlation coefficient is increased with pairwise comparisons. We therefore advise users to analyze the selected species with standard comparative methods on the basis of the full set of contrasts whenever possible, instead of using the differences for the selected pairs directly.

The simulation results revealed that phylogenetic targeting, compared with a random selection of species, provides many advantages for detecting correlated trait evolution. Statistical power was strongly increased in all of the cases we examined. Phylogenetic targeting used a higher percentage of the available range of variation for a character than did random sampling of species. We can therefore be more certain that the pattern holds generally across the clade of organisms rather than, for example, only among the species that are larger in body size or more amenable to study. Surprisingly, the simulations also revealed that statistical power increased with the number of species per tree, even when the number of taxa selected for study remained constant. Type I errors, however, were always close to the nominal level and were indistinguishable between phylogenetic targeting and random species sampling. Thus, applying the method to larger clades resulted in increased power without increasing the number of pairs examined, probably because having more taxa increased the magnitude of the differences that could be

selected overall (which in turn increased the ability to detect a correlation).

Phylogenetic targeting should be used with caution when one wants to determine the magnitude of a correlation. Like the pairwise approach of Westoby (1999), it overestimates the correlation coefficient (Ackerly 2000). This was true for both the pairwise tests and PIC, and the bias was stronger with the pairwise tests. The simulations also revealed that this overestimation increases with the number of species per tree, thus mirroring the increase in power. In the context of applying the method to real-world data in which data for Y_t are already available for some of the species, however, simulations confirmed that this bias decreases substantially with the number of randomly selected species for which data are already available. For most questions of interest that we envision, data are often available on Y_t for a number of species that often comprised a majority of the species in the data set. When such data are available, inclusion of already available data in subsequent analysis after applying phylogenetic targeting is highly recommended. Alternatively, users can implement the availability variable option described above to more fully integrate decisions about future data collection with already-studied species. Furthermore, as noted above, the bias is likely to decrease if additional traits representing confounding variables or alternative hypotheses are included in the analysis.

A few limitations and assumptions of phylogenetic targeting should be noted. Although the maximal pairing selects the set of species pairs that have the highest overall scores according to a user-defined scoring model, it may select species that are not directly comparable in relation to a particular test, such as an experiment that involves testing cognitive abilities. To overcome this possible weakness, our PhyloTargeting program provides a way for the user to manually select pairs in which particular comparisons are possible and to exclude other comparisons. Phylogenetic targeting must be used with caution if nonlinear relationships exist between the variables, and we advise users to critically examine the variables beforehand. Other critical issues are the phylogenetic tree, the representation of polytomies (see app. B), and the branch lengths on which the species selection is based. The selection of species can vary substantially between similar tree topologies due to the fact that the maximal pairing algorithm strictly maximizes the overall score, which can sometimes be heavily influenced by the topology. Branch lengths are assumed to be proportional to the expected variance in the amount of evolutionary changes along each branch (Brownian motion); this assumption becomes important in both phylogenetic targeting and subsequent analyses, particularly for PIC. If these assumptions are violated, Type I error rates are inflated and statistical power is reduced (Díaz-Uriarte and Garland 1996; Quader et al. 2004). Indeed, the simulations confirmed this effect; for almost all of the alternative models, Type I error rates were highly elevated. The only exception is the early-burst model, which yielded results that were very similar to those for Brownian motion ("Simulation Results" Excel file).

Because sister taxa will tend to be similar in many ways, confounding variables are expected to be less of a problem in comparisons of sister species (Harvey and Pagel 1991; Møller and Birkhead 1992). In our approach, however, more distantly related species pairs can also be selected. This can be critical, because other unmeasured, confounding variables may be introduced to the analysis. The comparison of distantly related species is comparable to an experiment with multiple uncontrolled variables (Garland and Adolph 1994; Garland 2001). The more distantly related two species are, the more likely it is that such an effect could bias the results. Including additional variables in the calculations makes it possible to control for some confounding variables when measurements are available.

We recommend that users standardize pairs to meet statistical requirements of subsequent statistical tests (i.e., equal variances among pairs). Standardization has not typically been implemented for pairwise comparisons, but it is necessary if one wishes to use parametric statistical tests that make assumptions about homoscedasticity. When contrasts are standardized, distantly related pairs are selected less often. This may be useful if large differences are informative only when the species are closely related (e.g., to control for possibly unknown confounding variables) or if comparisons should be made between closely related species (e.g., because of biological differences that limit comparability of experimental results). Standardization as such affects the selection of pairs.

Another argument for standardization is that fewer traits should change on shorter branches, and thus it helps to control for confounding variables. However, standardization may exaggerate evolutionary differences for close relatives when differences are due to sampling error or within-species variation (Purvis and Webster 1999). It can therefore overestimate the importance of certain species pairs if they are close relatives. We may sometimes expect a larger absolute change in some trait, regardless of its rate of change, to be more valuable in testing a hypothesis than a small change over a short branch. For example, brain size that increases by an order of magnitude might be a stronger test than a smaller amount of brain change, even if the small change occurs over a small branch. Using the program that we provide, the choice of standardization is left up to the user (with the default option to standardize scores) and is based on his or her preferences, the assumptions of subsequent methods, and the particulars of the biological system.

Phylogenetic targeting works best for continuous traits, but it can also be used with discrete traits. However, phylogenetic targeting that is based purely on discrete characters is more challenging because the number of distinct differences is typically smaller. In such cases, it is common to find that numerous pairs have the maximal possible score. This will ultimately result in multiple optimal solutions in the maximal pairing algorithm. However, because the current implementation returns only one optimal solution, it is difficult to evaluate its uniqueness. Possible work-arounds would be to add a continuous variable or to standardize contrasts; both of these would help to generate variation in the scores and thus aid in deciding among the possible pairs of taxa.

The maximal pairing algorithm falls in a class of general combinatorial optimization problems that are of considerable interest in comparative phylogenetics and, more generally, bioinformatics. Several modifications of this algorithm have practical importance as well. For example, the algorithm could be modified to select only a fixed number of pairs (given by the researcher), thereby incorporating the fact that limited resources are available to select species for future study. This important variant is described in more detail elsewhere (see Arnold and Stadler 2010). It might also be desirable to take into account the conservation statuses of different species to ensure that species are studied before they become extinct. More generally, the selection of species could be based not solely on pairwise comparisons but on the full set of contrasts, possibly in combination with examining the raw data space or regularly sampling character values along the entire range of a character of interest. Here we laid down the foundation for systematically identifying species for future study. Many possible extensions and modifications of the approach are possible, particularly as they relate to alternative ways of sampling species.

In summary, we provided a systematic method of selecting species for future study that offers greater statistical power to test adaptive hypotheses, as compared with a random selection of species. With this method of phylogenetic targeting, it is also possible to control for confounding variables, incorporate alternative hypotheses, and make use of existing data on the trait of interest. It therefore provides a way to guide the selection of species relative to a priori hypotheses. Through our Web-based computer program, other researchers are able to easily implement this approach in a flexible and user-friendly way.

Acknowledgments

We want to thank all of the people who contributed to this research, especially L. J. Matthews, L. Revell, and P. F. Stadler. This research was supported by grant BCS-0923791 from the National Science Foundation, the Max Planck Society, the University of Leipzig, and Harvard University.

Literature Cited

Ackerly, D. D. 2000. Taxon sampling, correlated evolution, and independent contrasts. Evolution 54:1480–1492.

Arnold, C. 2008. Phylogenetic targeting: a systematic approach and computer program for targeting research effort in comparative evolutionary biology. Diploma thesis. University of Leipzig, Leipzig.

Arnold, C., and P. F. Stadler. 2010. Polynomial algorithms for the maximal pairing problem: efficient phylogenetic targeting on arbitrary trees. Algorithms for Molecular Biology 5:25.

Capellini, I., P. McNamara, B. T. Preston, C. L. Nunn, and R. A. Barton. 2009. Does sleep play a role in memory consolidation? a comparative test. PLoS ONE 4:e4609.

Clutton-Brock, T. H., and P. H. Harvey. 1977. Primate ecology and social organization. Journal of Zoology 183:1–39.

Cooper, N., J. Rodriguez, and A. Purvis. 2008. A common tendency for phylogenetic overdispersion in mammalian assemblages. Proceedings of the Royal Society B: Biological Sciences 275:2031–2037.

Díaz-Uriarte, R., and T. Garland Jr. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Systematic Biology 45:27–47.

Faustino, C. E. S., M. A. Silva, T. A. Marques, and L. Thomas. 2010. Designing a shipboard line transect survey to estimate cetacean abundance off the Azores archipelago. Arquipélago. Life and Marine Sciences 27:49–58.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1–15.

Fisher, D. O., and I. P. F. Owens. 2004. The comparative method in conservation biology. Trends in Ecology & Evolution 19:391–398.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. American Naturalist 160:712–726.

Freckleton, R. P., M. Pagel, and P. H. Harvey. 2003. Comparative methods for adaptive radiations. Pages 391–407 *in* T. M. Blackburn and K. J. Gaston, eds. Macroecology: concepts and consequences. Oxford, Blackwell.

Garland, T., Jr. 2001. Phylogenetic comparison and artificial selection: two approaches in evolutionary physiology. Pages 107–132 *in* R. C. Roach, P. D. Wagner, and P. H. Hackett, eds. Hypoxia: from genes to the bedside. Advances in Experimental Biology and Medicine. Kluwer Academic/Plenum, New York.

Garland, T., Jr., and S. C. Adolph. 1994. Why not to do two-species comparative studies: limitations on inferring adaptation. Physiological Zoology 67:797–828.

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

Garland, T., Jr., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. Journal of Experimental Biology 208:3015–3035.

Hamilton, W. D., and M. Zuk. 1982. Heritable true fitness and bright birds: a role for parasites? Science 218:384–387.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Harcourt, A. H., P. H. Harvey, S. G. Larson, and R. V. Short. 1981. Testis weight, body weight and breeding system in primates. Nature 293:55–57.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford.

Hearn, D., and M. Huber. 2006. The ancestral distance test: what relatedness can reveal about correlated evolution in large lineages with missing character data and incomplete phylogenies. Systematic Biology 55:803–817.

Hosken, D. J. 1997. Sperm competition in bats. Proceedings of the Royal Society B: Biological Sciences 264:385–392.

Hugot, J. P. 1999. Primates and their pinworm parasites: the Cameron hypothesis revisited. Systematic Biology 48:523–546.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature 411:937–940.

Lyamin, O. I., P. R. Manger, S. H. Ridgway, L. M. Mukhametov, and J. M. Siegel. 2008. Cetacean sleep: an unusual form of mammalian sleep. Neuroscience and Biobehavioral Reviews 32:1451–1484.

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. Nexus: an extensible file format for systematic information. Systematic Biology 46:590–621.

Maddison, W. P. 2000. Testing character correlation using pairwise comparisons on a phylogeny. Journal of Theoretical Biology 202: 195–204.

Martin, R. D. 1990. Primate origins and evolution. Chapman & Hall, London.

Martins, E. P. 2000. Adaptation and the comparative method. Trends in Ecology & Evolution 15:296–299.

Midford, P. E., T. Garland, Jr., and W. P. Maddison. 2005. PDAP package of Mesquite, version 1.08(2). http://mesquiteproject.org/pdap_mesquite/.

Mitani, J. C., J. GrosLouis, and J. H. Manson. 1996. Number of males in primate groups: comparative tests of competing hypotheses. American Journal of Primatology 38:315–332.

Møller, A. P. 1991. Sperm competition, sperm depletion, paternal care, and relative testis size in birds. American Naturalist 137:882–906.

Møller, A. P., and T. R. Birkhead. 1992. A pairwise comparative method as illustrated by copulation frequency in birds. American Naturalist 139:644–656.

Nunn, C. L., and C. P. van Schaik. 2002. Reconstructing the behavioral ecology of extinct primates. Pages 159–216 *in* J. M. Plavcan, R. F. Kay, W. L. Jungers, and C. P. van Schaik, eds. Reconstructing behavior in the fossil record. Kluwer Academic/Plenum, New York.

Nunn, C. L., P. McNamara, I. Capellini, P. Preston, and R. A. Barton. 2009. Primate sleep in phylogenetic perspective. Pages 123–144 *in* P. McNamara, R. A. Barton, and C. L. Nunn, eds. Evolution of sleep: phylogenetic and functional perspectives. Cambridge University Press, Cambridge.

Oakes, E. J. 1992. Lekking and the evolution of sexual dimorphism in birds: comparative approaches. American Naturalist 140:665–684.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Price, T. 1997. Correlated evolution and independent contrasts. Philosophical Transactions of the Royal Society B: Biological Sciences 352:519–529.

Purvis, A., and L. Bromham. 1997. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. Journal of Molecular Evolution 44:112–119.

Purvis, A., and A. Rambaut. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Computer Applications in the Biosciences 11:247–251.

Purvis, A., and A. J. Weber. 1999. Phylogenetically independent contrasts and primate phylogeny. Pages 44–68 *in* P. Lee, ed. Comparative primate socioecology. Cambridge University Press, Cambridge.

Purvis, A., P. M. Agapow, J. L. Gittleman, and G. M. Mace. 2000*a*. Nonrandom extinction and the loss of evolutionary history. Science 288:328–330.

Purvis, A., J. L. Gittleman, G. Cowlishaw, and G. M. Mace. 2000*b*. Predicting extinction risk in declining species. Proceedings of the Royal Society B: Biological Sciences 267:1947–1952.

Quader, S., K. Isvaran, R. E. Hale, B. G. Miner, and N. E. Seavy. 2004. Nonlinear relationships and phylogenetically independent contrasts. Journal of Evolutionary Biology 17:709–715.

R Development Core Team. 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org.

Read, A. F., and S. Nee. 1995. Inference from binary comparative data. Journal of Theoretical Biology 173:99–108.

Ridley, M. 1983. The explanation of organic diversity: the comparative method and adaptations of mating. Clarendon, Oxford.

Roth, T. C., J. A. Lesku, C. J. Amlaner, and S. L. Lima. 2006. A phylogenetic analysis of the correlates of sleep in birds. Journal of Sleep Research 15:395–402.

Slowinski, J. B. 1993. "Unordered" versus "ordered" characters. Systematic Biology 42:155–165.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33:475–505.

Westoby, M. 1999. Generalization in functional plant ecology: the species sampling problem, plant ecology strategy schemes, and phylogeny. Pages 847–872 *in* F. I. Pugnaire and F. Valladares, eds. Handbook of functional plant ecology. M. Dekker, New York.

———. 2002. Choosing species to study. Trends in Ecology & Evolution 17:587.

Westoby, M., S. A. Cunningham, C. Fonseca, J. Overton, and I. J. Wright. 1998. Phylogeny and variation in light capture area deployed per unit investment in leaves: designs for selecting study species with a view to generalizing. Pages 539–566 *in* H. Lambers, H. Poorter, and M. M. I. V. Vuren, eds. Variation in growth rate and productivity of higher plants. Backhuys, Leiden.

Associate Editor: Dean C. Adams
Editor: Mark A. McPeek