Chapter 6

A parametric bootstrap approach to the detection of phylogenetic signals in landmark data

Theodore M. Cole III [1], Subhash Lele [2], Joan T. Richtsmeier [3,4]

[1]Department of Basic Medical Science, School of Medicine
University of Missouri, Kansas City
Kansas City, MO 64108, USA.

[2]Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta T6G 2G1, Canada.

[3]Department of Cell Biology and Anatomy
School of Medicine, The Johns Hopkins
University, Baltimore, MD 21205, USA.

[4]Department of Anthropology
The Pennsylvania State University
University Park, PA 16802, USA.

ABSTRACT

A phylogenetic signal is present in a morphometric data set if similarities in form reflect genealogical relationships. The degree to which such a reflection exists can be measured by comparing the topology of a morphometric-based hierarchical clustering with the topology of a cladogram that is specified a priori using other sources of data. A strong phylogenetic signal is indicated by a high degree of agreement between topologies. A lack of agreement is indicative either of data with a strong "alternative" signal (attributable to homoplasy) or of data with a lack of a signal of any kind. In considering the uncertainties inherent in morphometric data, we present a new method for detecting phylogenetic signals when form is described using landmark coordinate data. We provide a parametric bootstrapping algorithm that, while applied to landmarks, is general enough to be applied to any sort of morphometric data where a reasonable model of within-sample variation can be specified. We then demonstrate how the bootstrap data can be used to make topological comparisons between morphometric clusterings and the cladogram, using: 1) bootstrap proportions attached to cladogram nodes; 2) tree-comparison statistics; and 3) analysis of the frequencies of morphometric-based clusterings that occur when bootstrapping under the model. We then demonstrate our method by examining phylogenetic patterning in midfacial shape for ateline primates. We conclude by discussing topics where more research is needed, concentrating on efforts to partition morphometric data into homologous and homoplasious components.

INTRODUCTION

Within the past quarter century, the science of comparative biology has undergone a substantial transformation, centered on the advocacy of an explicitly phylogenetic (historical) perspective in the study of evolution. Such a perspective is now considered essential for testing hypotheses about adaptation, the evolution of biological roles, evolutionary covariances among characters, and general principles of organismal design (Gould and Lewontin 1979; Lauder 1981, 1990; Coddington 1988; Wake and Larson 1988; Larson and Losos 1996; Huelsenbeck et al. 2000). In addition, a phylogenetic framework is required for the study of the mechanisms underlying evolutionary transformations in form, such as heterochrony (e.g., Fink 1982; Wake and Larson 1988; McKinney and McNamara 1991). Within the same quarter century, there has been a proliferation of new methods in morphometrics, particularly where analyses of landmark data are concerned (Rohlf and Marcus 1993; Bookstein 1996). While the applications of morphometrics have varied widely, many of them address evolutionary questions, so that a combination of the "new comparative biology" and the "new morphometrics" seems natural. However, despite their simultaneous development, there has been little synthesis of the two fields, and we perceive substantial theoretical and methodological gaps between them.

The mandates of the new comparative biology should strongly influence the course of morphometric research, as the latter should be regarded as a tool for pursuing the goals of the former. Therefore, we are faced with the task of developing new methods (or retooling existing ones) so that morphometrics becomes more relevant to researchers who study the history of biological patterns and processes. The purpose of this paper is to take an initial step in meeting this challenge. We begin with a basic question that has broad relevance: what does it mean for a data set to have (or to lack) a phylogenetic signal? We then propose a method

for recognizing a phylogenetic signal in a set of morphometric data. We will be concerned specifically with the analysis of landmark data, although much of what we will present may be applied to other quantitative descriptors of biological form.

WHAT ARE PHYLOGENETIC SIGNALS?

In the most basic sense, we say that there is a phylogenetic signal in morphometric data when closely-related taxa are more similar to one another than they are to more distantly-related taxa. Because phylogenies are hierarchically organized, it is useful to refine this basic definition in terms of a comparison between two nested, hierarchical trees. The first tree is a phylogenetic one, represented by a *cladogram* that provides a history of the speciations that gave rise to the clade's member taxa. The second tree is a *phenogram*, which is constructed using the landmark data, where the hierarchical structure is described using a clustering algorithm. The strength of the landmark data's phylogenetic signal is reflected in the degree to which these two trees match. If the topologies of the cladogram and the phenogram are very similar, we can say that there is evidence of a strong phylogenetic signal contained within the landmark data. The morphometric affinities based on the landmark data would therefore reflect the genealogical relationships among taxa. If the two topologies are very different, we can conclude that form has evolved in a way that does not reflect phylogeny.

Now that we have established (in basic terms) what phylogenetic signals are, we can begin to think about how they come about and what they mean. We can also think about what it means if there is a strong, nonrandom signal that *does not* reflect phylogeny. Finally, we can consider why there may be no apparent signal at all, so that morphometric variation is randomly distributed with respect to phylogeny.

HOW DO PHYLOGENETIC SIGNALS ORIGINATE AND WHY ARE THEY INTERESTING?

Before discussing phylogenetic signals any further, we must state a fundamental assumption. Throughout this paper, we will assume that there is a preexisting estimate of the phylogenetic relationships among the study taxa (a cladogram) and that this estimate has been made without error (Pagel 1999; Huelsenbeck et al. 2000). To reduce the chance of making circular arguments, we also will assume that the cladogram has been estimated using data other than morphometric measurements (e.g., using molecular sequences, developmental patterns, aspects of behavior and life history, etc.). Naturally, as other researchers refine their estimates of the phylogenetic relationships among taxa, we will be faced with the task of revising our own morphometric analyses accordingly.

There are many possible scenarios where organisms evolve so that phylogenetically-patterned form variation is the result. As a first example, let us consider a scenario where natural selection plays no role, so that form evolves solely via stochastic (= random) processes (e.g., Felsenstein 1988). If we make some simplifying assumptions, we may find that the history of speciations largely determines the observed differences in form among terminal taxa. More specifically, the expected difference between any two taxa will be a function of the time that has passed since they last shared a common ancestor.[1] Consider the history of a four-taxon clade that is mapped onto a space where the horizontal axis is some measure of form (measured with morphometrics) and the vertical axis is time (Figure 1). The common ancestor for the radiation gives rise to two daughter lineages (labeled Lineage1 and Lineage 2), which, in turn, give rise to four terminal taxa. Taxa *A* and *B* arise from Lineage 1, while taxa *C* and *D* arise from Lineage 2. Whenever speciation occurs, we assume that the daughter taxa (whether lineages or terminal taxa) will evolve independently and at random

from that point onward (Cavalli-Sforza and Piazza 1975; Cheverud *et al*. 1985; Felsenstein 1985, 1988). We therefore expect taxa *A* and *B* to be the most similar because their evolutionary histories are identical, up to the point where they diverge from their common ancestor. From that point onward, they evolve independently; however, relative to the age of the entire clade, they have not had very much time for differences between them to evolve. In contrast, if we consider the difference between taxa *A* and *D*, we see that their last common ancestor is much closer to the base of the clade. Therefore, their common history is proportionately much shorter, and the time over which they have evolved independently is much longer. We thus expect them to have greater morphometric differences, simply because they have had more time to accumulate those differences through random processes.

We can also imagine scenarios where natural selection plays an active role in the production and maintenance of phylogenetic signals. As a simple example, suppose that the taxa in a clade have evolved different forms as specializations to different biological roles (Bock and van Wahlert 1965). Figure 2 shows a hypothetical situation where there is an association between organismal form and the occupation of two different adaptive zones. Now suppose that populations of the ancestral taxon for the clade encounter two novel sets of environmental conditions, so that a speciation occurs. Associated with the initial speciation, there may be a morphological divergence between the lineages, which may be especially pronounced if the lineages are entering novel adaptive zones where their respective forms can function as key adaptations to new biological roles (Harvey and Pagel 1991). Following the initial speciation, there may be a number of subsequent speciations in each clade (particularly if a key adaptation is involved), but the members of the two large clades may experience no further selective forces that would tend to force them outside of their respective adaptive zones. Their occupation of those zones is, therefore, very stable over time. As a result, the

phylogenetic signal is maintained at a relatively high taxonomic level, even though it may become obscured at lower levels. Harvey and Pagel (1991) refer to this pattern as phylogenetic niche conservatism. They also point out that reversals in the evolution of some complex forms and adaptations may be very unlikely, providing further reinforcement of phylogenetic signals once they have originated. Finally, Simpson (1961) points out that the separation between clades in such a case might be further reinforced by extinctions in the boundary between adaptive zones (i.e., in the valleys between adaptive peaks). Therefore, we might consider extinctions and the incompleteness of the fossil record as other potential factors that influence the expression of a phylogenetic signal in a data set.

There also may be interesting situations where there is a strong, nonrandom signal in the data that is *not* a phylogenetic signal. In this discussion, we will call such instances "alternative signals." Researchers who are interested in adaptation and the role of selection may find instances of alternative signals particularly valuable. For example, there may be instances where striking morphometric similarities between distantly-related animals are the result of convergence (Figure 3). If such similarities have arisen independently~~,~~ they can allow the construction of testable hypotheses about adaptation (Coddington 1988; Harvey and Pagel 1991; Wake 1991; Brooks 1996; Losos and Larson 1996). Homoplasies might also be frequent enough at lower taxonomic levels to be considered "rampant" in the measurements examined. While striking structural and functional similarities can evolve independently in distantly-related animals (e.g., the well-known similarities between some marsupial and placental carnivores), this pattern is also likely to occur in closely-related taxa with similar developmental programs (Sluys 1989; Brooks 1996; see also Alberch 1980, 1985). When taxon-specific developmental programs are essentially minor variations on the same theme,

different taxa to likely to find similar morphological solutions to similar biological problems, obscuring evidence of shared history.

Finally, there may be cases where there is no apparent signal of any kind. One familiar example is “star radiation” (Figure 4), where speciations occurred very rapidly, so that the cladogram’s nodes are all concentrated near the root. If there have been no further selective forces to cause taxa to evolve in parallel, the taxa will have evolved independently for nearly all of the clade’s history. As a result of this speciation pattern, nearly all of the morphometric differences that accumulate between taxa will be autapomorphic, so that “closely related species are no more likely to be similar than any two species picked at random” (Mooers et al. 1999, p. 250). If autapomorphies are ubiquitous, phylogenetic signals will be very hard to recognize. This will be especially true if each of the lineages experiences very different selective pressures, causing them to follow highly divergent evolutionary pathways.

Finally, there may be difficulties in detecting a phylogenetic signal if morphometric traits exhibit high evolvabilities (*sensu* Houle 1992), so that they are evolutionarily labile. In our initial depiction of the origins of a phylogenetic signal (Figure 1), the evolution of the clade was very neat and orderly, with the evolutionary paths of the taxa staying well-separated in space. However, suppose that each taxon covers a great deal of morphological ground as it evolves. If this is the case, the evolutionary pathways of the individual taxa may cross one another many times, so that form variation becomes randomly distributed with respect to the phylogenetic relationships between taxa (Figure 5). The reasons that morphometric data may be labile (so phylogenetic signals are therefore absent) are varied. On one hand, the lability may be an intrinsic characteristic of the organisms themselves. For example, high degrees of within-population genetic variation and low levels of stabilizing selection may both be

contributing factors (Houle 1992). On the other hand, the apparent lability of the traits may not be a quality of the organisms themselves, but of a rapidly fluctuating environment, where the organisms are chasing fast-moving adaptive peaks through the morphological space. Whatever their cause, a particularly interesting aspect of labile data is that the patterns they produce may be very difficult to distinguish from the homoplasy that results from parallel responses to selection (compare the positions of the taxa in Figures 3 and 5).

To conclude this section, we would like to consider the implications of strong signals (whether phylogenetic or alternative) for the way that we think about morphometric variation in a phylogenetic framework. To begin, we must recognize that morphometric data are phenetic data, and that phenetic similarities are, by definition, mixtures of homologous and homoplasious similarities (Cain and Harrison 1960; Simpson 1961; Sneath and Sokal, 1973; Felsenstein 1982; Cheverud et al. 1985). The difficulties of recognizing homologies in morphometric variables prior to construction of a phylogeny are widely recognized, and many investigators have been skeptical of the validity of continuous data as characters for estimating phylogenies (e.g., Pimentel and Riggins 1987; Cranston and Humphries 1988; Chappill 1989; Bookstein 1994). Much recent debate has focused on landmark data in particular (Zelditch et al. 1995, 2000; Fink and Zelditch 1995; Adams and Rosenberg 1995; Rohlf 1998). However, in looking at the distribution of morphometric variation relative to a phylogeny that has already been estimated using other data, we can approach questions about homology and homoplasy from a somewhat different perspective, because we already have a cladogram in place.

Discussions about homology frequently include detailed considerations of terminology, but for the purposes of our presentation we will opt for a fairly simple definition. We would consider a shared “state” of form or shape to be homologous if it is a shared-derived state that

characterizes a monophyletic group; this is Patterson's (1982) criterion of congruence (Zelditch et al. 1995; Chang and Kim 1996). If a shared morphometric state is, in fact, homologous, we expect to see congruence between phenetic and cladistic topologies, which returns us to our original definitions of a phylogenetic signal. Therefore, the presence of a strong phylogenetic signal suggests that morphometric similarities tend (in an overall sense) to be homologous, rather than homoplasious. If we judge some shared aspect of form to be homologous, we gain the advantage of discussing that morphometric similarity in terms that are familiar to phylogenetic systematists, including "symplesiomorphy" and "synapomorphy". However, what is most important about the provisional identification of a morphometric homology is that it provides us with a starting point for better understanding the processes that generate evolutionary diversity. As Sanderson and Hufford (1996, p. 329) succinctly state: "At issue in the study of homology is how character states become different despite their common origin."

In contrast, a strong "alternative" signal tells us something very different about how morphometric similarities tend to evolve from different beginnings (Sanderson and Hufford 1996). To recognize homplasy in discrete character states, Hennig (summarized by Brooks 1996) recommended that similar states should be provisionally considered as homologous. Following the construction of a cladogram, this assumption is reevaluated for each character, and similarities that are incongruent with the phylogeny are reclassified as homoplasies. For morphometric data, we could make the same assumption initially (that is, that all morphometric similarities are homologous). However, if we then found a strong alternative signal, we would have evidence that our initial assumption was incorrect. We would then have to consider the possibility that many of the morphometric similarities we observed were homoplasies. We might then be able to construct testable hypotheses about the biological

roles of these similarities and the factors that would tend to produce homoplasies of form (e.g., adaptations to similar environments, developmental constraints).

UNCERTAINTY AND THE BOOTSTRAP

Now that we have introduced some basic ideas about what phylogenetic signals are and how they might arise, we can turn our attention to how they can be studied using real data. As we stated earlier, we can evaluate the phylogenetic signal in morphometric data by comparing cladistic and phenetic hierarchical topologies. However, we first need to discuss the nature of the data that are used to construct them.

When we compare a phenogram to a cladogram, we assume that the cladogram is measured without error. However, this assumption cannot be made for morphometric phenograms, because we know that the measurements vary within populations. We also acknowledge that our sample sizes often may be small, particularly when we are studying rare organisms or fossils, so that sampling errors become an important consideration in estimating morphometric affinities. We therefore realize that *uncertainty* in statistical estimation threads its way throughout our study of phylogenetic signals from start to finish. The fact of this uncertainty has led us to use the bootstrap, which is a very versatile method for addressing issues of statistical uncertainty in interesting and informative ways.

The bootstrap technique was developed in the late 1970s and early 1980s by Efron and colleagues [see Efron and Tibshirani (1991, 1993) for reviews]. It is a frequently used method for working with statistics that have either very complex or unknown distributions, where intensive computational effort can be used to address problems that might otherwise be intractable (Efron & Tibshirani 1991). Phylogenetic applications of the bootstrap and related methods were introduced soon after the development of the bootstrap itself. As with other

bootstrap applications, the first phylogenetic uses were motivated by concerns about the uncertainties of statistical estimation. Felsenstein (1985) was the first to use the nonparametric bootstrap to address concerns about the uncertainty of sampling discrete characters for use in phylogeny estimation. His method of attaching bootstrap proportions to cladogram nodes is now widely used in the systematics literature (see below). At roughly the same time, Lanyon (1985) used the jackknife (a related method) as a means of dealing with the uncertainties of estimating genetic distances and the phylogenies inferred from them. Mueller and Ayala (1982) had previously suggested the use of the jackknife for this same purpose. More recently, Huelsenbeck et al. (1996:20) used the parametric bootstrap to model variation in DNA sequence data, demonstrating the versatility of the method. Their aims were to examine bias in phylogenetic estimation, to compare the support for competing phylogenetic hypotheses, to conduct power analyses, and to measure the repeatability of a tree's subclades. This last aim is the same as Felsenstein's (1985) and is probably the most frequent application of the bootstrap in phylogenetics.

Before we describe how we have applied the bootstrap in this study, we will provide a brief illustration of how the method generally works in applications other than phylogenetics. Suppose we have measured a sample of organisms using a continuously-distributed variable (for example, body mass or length), and we want use the data to estimate a parameter θ, which is a smooth function of the data. In addition to obtaining a point estimate of θ (a statistic called T), we want to say something about our uncertainty in making that estimate. The uncertainty involved in making point estimates is usually quantified using standard errors and confidence intervals. For many familiar statistics (such as means, regression coefficients, and correlation coefficients), there are analytical formulae for calculating these uncertainty measures. However, if there are no available analytical methods for the statistics that interest us, we can apply the bootstrap. If the sample has n observations, we can

construct a *pseudosample* by drawing *n* observations from the sample randomly and with replacement. By "randomly", we mean that all observations have the same probability of being selected. By "with replacement," we mean that any given observation can be sampled more than once and that some observations may not be sampled at all. From the pseudosample, we compute a bootstrap estimate of *T* and call it *T**. This process is then repeated for *M* independent pseudosamples, where *M* is a large number (usually between 200 and 1000), so that we get a *bootstrap distribution* of *T**. Once the bootstrap distribution is obtained, we calculate its mean $\hat{T}^* = \Sigma(T^*)/M$, which is the bootstrap estimate of *T*. The standard *error* of *T* is estimated by the standard *deviation* of the *T** estimates, and there are several ways that bootstrap confidence intervals for *T* can be computed (Efron and Tibshirani 1991, 1993; Davison and Hinkley 1997).

The type of resampling just described is called *nonparametric bootstrapping*, because no assumptions are made about the distribution of the data. However, suppose we can make reasonable assumptions about how the data are distributed (although the distribution of the statistic of interest may remain very complex or unknown). In that case, we can use *parametric bootstrapping*, where a fitted parametric model serves as the basis for generating random data sets that can serve as pseudosamples (Efron and Tibshirani 1991, 1993; Huelsenbeck et al. 1996; Davison and Hinkley 1997). Returning to our simple example, suppose we can assume that the measurement data are distributed as $N(\mu,\sigma^2)$; that is, the data are normally distributed with mean μ and variance σ^2. To perform a parametric bootstrap, we first obtain sample estimates of μ and σ^2, called $\bar{x}$ and s^2, respectively. We then generate *M* independent pseudosamples, each with *n* random observations that are distributed as $N(\bar{x},s^2)$. The parametric bootstrap estimates of the mean and standard error of *T* are then computed as with the nonparametric method. The primary advantage of parametric bootstrapping is that

the standard-error and confidence-interval estimates are generally more accurate than nonparametric estimators. This is an especially important concern when studying multivariate data, where the sample sizes required for precise nonparametric estimates increase exponentially with the number of variables (Silverman 1986).

A MODEL FOR DESCRIBING MORPHOMETRIC VARIATION USING LANDMARKS

An explicit model of within-sample variation is necessary for any application of the parametric bootstrap. Before we describe the statistical model and computations that we use in this study, we will present a more general picture of biological variation in landmark data, which is largely based on Lele (1999). Suppose we are interested in a sample of n organisms and we measure them using a series of K landmarks in D (= 2 or 3) dimensions. The mean for the population is described by a $K \times D$ matrix called **M**, where each row represents the D-dimensional coordinates of a landmark. While **M** is not directly observable, we can imagine the mean configuration of landmarks in a "Nature Space" (Figure 6), where within-sample variation arises. No single individual is likely to be identical in form to the mean, nor are any two individuals likely to be identical. This phenotypic variability is due to both genetic and environmental variation (Falconer and Mackay 1996). In the Nature Space, phenotypic variation is manifested as perturbations around **M** (Figure 6). Note that the dispersion patterns of these perturbations can vary in size and shape from landmark to landmark. Some landmarks are more variable than others, as indicated by the relative sizes of their dispersions. The perturbation scatters also can vary in shape from one landmark to another, with some being round and others being elliptical. Finally, there may be covariances between the landmarks, so that the relative positions of the observations at one landmark may be correlated with their positions at other landmarks.

We describe the phenotypic variation statistically with a *general perturbation model*. This model was used by Goodall (1991) in the development of superimposition (Procrustes) methods and by Lele (1993; Lele and Richtsmeier 1990; Lele and McCulloch 2001) in the development of Euclidean distance matrix analysis (EDMA). We can describe the landmark data for each observation in a sample with a $K\times D$ matrix called $\mathbf{X}_i$. Each $\mathbf{X}_i$ is related to $\mathbf{M}$ as follows:

$$\mathbf{X}_i = (\mathbf{M} + \mathbf{E}_i)\Gamma_i + \mathbf{t}_i$$

$\mathbf{E}_i$ is a $K\times D$ matrix of perturbations that describe how $\mathbf{X}_i$ differs from $\mathbf{M}$ *in the Nature Space*. For the population, these perturbations are assumed to have a multivariate normal distribution with a $K\times D$ mean matrix $\mathbf{0}$ and a covariance structure $\Sigma_K \otimes \Sigma_D$, where the $\otimes$ operator denotes a Kronecker product. Σ_K is a $K\times K$ matrix that describes the variances and covariances of the landmarks, while Σ_D is a $D\times D$ matrix that describes the variances and covariances of the perturbations with respect to the Nature Space's coordinate-system axes (*i.e.*, they describe the eccentricity and orientation of the perturbation scatters). The mean and the variance-covariance matrices are obviously of great biological interest. Unfortunately, they are not estimable because of the presence of the other terms in the equation, which are called *nuisance parameters* (Neyman and Scott 1948; Lele and Richtsmeier 1990; Lele 1993). The orthogonal $K\times K$ matrix Γ_i describes the rotation of $\mathbf{X}_i$ (as we measure it) relative to $\mathbf{M}$ (as it lies in the Nature Space), while the $K\times D$ matrix $\mathbf{t}_i$ describes the translation of $\mathbf{X}_i$ relative to $\mathbf{M}$. Unfortunately, these entirely arbitrary parameters are unobservable, which means that reconstruction of the Nature Space from empirical data is impossible (Lele 1993, 1999; Lele and McCulloch 2001).

Fortunately, there are some biologically interesting components of the model that are identifiable and can be estimated using method-of-moments techniques developed by Lele (1993; Lele and McCulloch 2001). While we cannot observe directly the coordinates of the population mean form (**M**), we can compute the coordinates of a consistent sample estimate of the mean, called $\hat{\mathbf{M}}$, up to translation, rotation, and reflection. While we cannot estimate the sample among-landmarks variance-covariance matrix (Σ_K), we can compute a consistent estimate of a singular version of it, called Σ_K^*. Finally, while neither the among-axes variance covariance matrix (Σ_D) nor its eigenvectors is estimable, its eigenvalues can be estimated (Lele and McCulloch 2001), describing the overall eccentricity of the perturbation scatters. Alternatively, we can make the simplifying assumption that $\Sigma_D = \mathbf{I}$ (Lele and Cole 1996). We should emphasize that all of these estimators are *coordinate-system invariant*, so that they are not affected by the positions and orientations of the observations in any arbitrary coordinate system (Lele 1993). If we assume that the landmark perturbations about **M** approximate a multivariate normal distribution, we can use $\hat{\mathbf{M}}$ and Σ_K^* to randomly generate pseudosamples under the model (Lele and Cole 1996). This data-generating procedure is at the heart of our parametric bootstrapping method, as described in the following section.

PARAMETRIC BOOTSTRAPPING UNDER THE MODEL

We now provide a description of the parametric bootstrapping algorithm that we use to assess a phylogenetic signal. It is illustrated schematically in Figure 7. The particular details of scale adjustments, dissimilarity metrics, and clustering algorithms may vary from one application to another, so we will only speak of them in general terms for the time being.

1. Using the sample-specific estimates of the mean form ($\hat{\mathbf{M}}$ for each sample), compute a matrix of the pairwise dissimilarities in form between taxa. As explained below, this matrix is called $\mathbf{F}_\Omega$ (or $\mathbf{S}_\Omega$ if the data are scale-adjusted). $\mathbf{F}_\Omega$ is used as the basis of a hierarchical cluster analysis, yielding a morphometric-based clustering that is referred to as the "empirical phenogram". If the goal is to study *shape*, rather than *form* (where information about scale is retained), the mean forms should be scaled first (see Lele and Cole 1996 for a discussion).
2. Again using the sample-specific estimates of $\hat{\mathbf{M}}$ and Σ_K^*, generate a set of pseudosamples of the appropriate sample sizes. An algorithm for generating random data under the model is provided by Lele and Cole (1996). After calculating the pseudosample means, construct a matrix of pairwise dissimilarities called $\mathbf{F}_\Omega$**** (or $\mathbf{S}_\Omega$*), where the asterisk indicates that it is computed from a set of pseudosamples.
3. Generate *M* sets of pseudosamples by repeating Step 2 a large number of times (e.g., 200 to 1000). Use the resulting bootstrap distribution of $\mathbf{F}_\Omega$**** (or $\mathbf{S}_\Omega$*) matrices and a hierarchical clustering algorithm to obtain a distribution of bootstrap phenograms.
4. Compare the bootstrap phenograms to the cladogram, examining the bootstrap proportions attached to cladogram nodes or using tree-comparison statistics. We might also examine the frequencies with which different bootstrap phenograms occur when resampling under the model. Each of these methods for comparison is described below.

CONSTRUCTING PHENOGRAMS BASED ON MORPHOMETRIC DISSIMILARITY

Given the collection of sample mean forms, we must determine which of the means are most similar and which are most different (Steps 1 and 2 of the bootstrapping algorithm). These similarities and differences serve as the basis for constructing our morphometric clusterings (the empirical and bootstrap phenograms). For landmark data, we can use *Euclidean distance matrix analysis* (EDMA; Lele and Richtsmeier 1991, 2000; Lele 1993; Lele and McCulloch 2001), which is a coordinate-system-invariant method of describing and comparing forms. The basis of all EDMA applications is the *form matrix* (**FM**). Suppose we have the mean coordinates of a taxon *A*, measured with *K* landmarks. The form matrix **FM**(*A*) is defined as:

$$\mathbf{FM}(A) = \begin{bmatrix} 0 & d(1,2) & \cdots & d(1,K) \\ d(2,1) & \ddots & & \\ \vdots & & \ddots & d(K-1,K) \\ d(K,1) & & d(K,1-K) & 0 \end{bmatrix}$$

where $d(i,j)$ is the Euclidean distance between landmarks *i* and *j*. This representation of form is coordinate-system invariant because the elements of **FM**(*A*) are always the same, no matter how *A* is positioned (=translation) or oriented (=rotation and reflection). Now suppose we have a second taxon *B*, with its own form matrix **FM**(*B*), and we want to compare the forms. We can define a *form-difference matrix*, called **FDM**(*A*,*B*), as follows:

$$\mathbf{FDM}(A,B)_{ij} = \frac{\mathbf{FM}(A)_{ij}}{\mathbf{FM}(B)_{ij}}$$

where $i,j = 1 \ldots K$ and with the convention that 0/0=0. Each element of the **FDM** is the ratio of like distances in *A* and *B*. If *A* and *B* are identical in form, all of the off-diagonal elements of the **FDM** will be one. If a given distance is greater in *A* than in *B*, the corresponding element of the **FDM** will be greater than one. Similarly, an instance where the distance in *B* is greater will be indicated by an element that is less than one. If all of the off-diagonal elements are equal and are different from one, *A* and *B* will have the same *shape*, but will differ in *scale*. Finally, if the off-diagonal elements are heterogeneous, *A* and *B* will differ in *shape*.

Importantly, the **FDM** shares the property of coordinate-system invariance with the form matrices, meaning that its elements are always the same, no matter how either *A* or *B* are translated, rotated, or reflected.

In the algorithm described above, we discussed dissimilarity measures in very general terms, but we can now define them more precisely. The form-difference matrix can be used as the basis for a dissimilarity measure, called F_Ω, between two taxa (Richtsmeier et al. 1998). Given **FDM**(*A*,*B*), we can calculate:

$$F_\Omega(A,B) = \sqrt{\sum [\ln(\mathbf{FDM}(A,B))]^2}$$

where the summation is over all of the below-diagonal elements. $F_\Omega(A,B)$ is a metric and is equivalent to the Q-mode Euclidean distance between *A* and *B* (Sneath and Sokal 1973, p. 124), when that distance is calculated using all ln-transformed interlandmark distances. If *A* and *B* are identical, then F_Ω will equal zero; otherwise, F_Ω will become increasingly positive as *A* and *B* become more different in form. If the taxa we are comparing vary substantially in scale, we may want to concentrate on variation in *shape*, rather than in form. Scaling of each sample mean is accomplished simply by computing the form matrix (**FM**) and then dividing each element by an appropriate scaling factor (derived from the elements themselves), yielding a shape matrix called **SM** (Lele and Cole 1996). Some examples of possible scaling factors include any single interlandmark distance (e.g., one that measures maximum length or breadth), the maximum distance, the median distance, or the geometric mean of all of the distances. Whichever scaling factor is chosen, we strongly recommend that the choice should be based solely on biological grounds (Lele and Cole 1996).

Given the means of multiple taxa, we collect all of the pairwise F_Ω statistics in a symmetric dissimilarity matrix called $\mathbf{F}_\Omega$. For example, if there are three taxa called *A*, *B*, and *C*:

$$\mathbf{F}_{\Omega} = \begin{bmatrix} 0 & F_{\Omega}(B,A) & F_{\Omega}(C,A) \\ F_{\Omega}(A,B) & 0 & F_{\Omega}(C,B) \\ F_{\Omega}(A,C) & F_{\Omega}(B,C) & 0 \end{bmatrix}$$

Alternatively, a dissimilarity matrix based on shape matrices would be called $\mathbf{S}_{\Omega}$. With such a dissimilarity matrix in hand, we can use it as the basis for hierarchical cluster analysis to obtain both the empirical phenogram and the distribution of bootstrap phenograms.

In outlining our algorithm, we have spoken of hierarchical clustering in very general terms. However, the choice of a clustering algorithm should be carefully considered. There are many different algorithms available for constructing hierarchical clusters, and these may yield different results, even when based on the same dissimilarity matrix (Sokal and Rohlf 1962; Sneath and Sokal 1973; Johnson and Wichern 1982). In deciding which method to use, we note that hierarchical clustering methods cannot summarize the relationships of taxa in multivariate space without introducing some kind of distortion, and some of the information about pairwise dissimilarities is invariably lost (Sokal and Rohlf 1962; Sneath and Sokal 1973). Therefore, the best choice of a clustering algorithm might be the method that introduces the least distortion and provides the most faithful summary of the information in the dissimilarity matrix. One way of measuring this accuracy is through use of the cophenetic correlation (Sokal and Rohlf 1962), which is the correlation between the elements of the original dissimilarity matrix and those implied by the hierarchical clustering. The "best" algorithm using this criterion will be the one with the cophenetic correlation that is closest to 1.0. Finally, we would strongly discourage selecting the method that produces the clustering that is most similar in topology to the cladogram; this practice would defeat the

purpose of our method by biasing the results toward the detection of a strong signal, where a weaker one may exist in reality.

MEASURING THE SIGNAL

After generating the empirical phenogram and a distribution of bootstrap phenograms, we can compare the topologies of all of them to that of the cladogram. Because we are interested in comparing the topologies of two different hierarchical trees, we have a choice of several different approaches. First, we can calculate bootstrap proportions (colloquially referred to as "bootstrap support") for each of the nodes of the cladogram. Second, we can employ any number of tree-comparison statistics for evaluating the agreement between cladistic and phenetic topologies. Finally, we can examine the frequencies of the different topologies that occur under when resampling under the model. Although this last approach does not involve actual tree comparisons, it can be very interesting and informative.

Bootstrap proportions: The distribution of bootstrap phenograms can be used to assign a bootstrap proportion (Felsenstein 1985; Efron et al. 1996) to each internal node (=subclade) of the cladogram. Suppose we are interested in the node corresponding to a subclade that includes three taxa *A*, *B*, and *C*. The associated bootstrap proportion is the percentage of bootstrap phenograms where *A*, *B*, and *C* cluster together to the exclusion of all other taxa. Note that there is no consideration of the internal structure of the clade, only of the identity of its member taxa. Note also that bootstrap proportions are *marginal* proportions (Felsenstein 1985), meaning that the proportions for different nodes are calculated independently. To relate bootstrap proportions to a phylogenetic signal's strength, we would say that a bootstrap proportion of 100% would indicate a perfect phylogenetic signal for the subclade, while 0% would indicate that the subclade's members never cluster together. We can attach bootstrap proportions not only to the cladogram, but to the empirical phenogram as well. In doing so,

we get an idea of the *repeatability* in the data (*sensu* Hillis and Bull 1993), so that we have an explicit picture of the uncertainty in estimating the phenetic clustering. This is similar to the concept of the "robust validity" of distance and correlation matrices that was used by Cheverud et al. (1989).

Finally, and as an aside, we would like to make a comment regarding terminology. While bootstrap proportions are popularly referred to as measures of "bootstrap support," we have avoided using that term. "Support" has a very specific statistical definition that is related to principles of likelihood (Edwards 1992), and this definition is different from what we measure using bootstrap proportions. Hillis and Bull (1993) similarly favor the use of "bootstrap proportions", while discouraging the use of "bootstrap *P* values."

Tree-comparison statistics: Tree-comparison statistics are commonly used to measure the degree of difference between two hierarchical structures, expressing this degree as a single number. There are many different statistics available [see reviews by Rohlf (1974, 1982), Hubert (1978), Hendy and Penny (1985), and Lapointe and Legendre (1990)], and bootstrapping can be used to generate estimates of their standard errors and confidence intervals. For this study, we have developed a simple tree-comparison statistic that we find useful because its interpretation is very straightforward. To compare a cladogram with a phenogram, we first represent their topologies in matrix form using *cardinality matrices* (Lapointe and Legendre 1992), where all of the information about branching sequences is retained, but where branch lengths are ignored. The dissimilarity between two taxa *A* and *B* is defined as the total number of taxa in the smallest clade/cluster containing both *A* and *B*. Therefore, if *A* is closest to *B*, the dissimilarity between them will be the minimum possible value of 2 (the size of the smallest possible clade/cluster). If *A* is most distant from *B*, the dissimilarity will be the maximal value, which is equal to the total number of taxa in the

study (which is the largest possible clade/cluster). Figure 8 provides three examples of how cardinality matrices are constructed. In the third example, note how cardinality matrices can be used with cladograms that have unresolved multifurcations.

To compare the cladogram with the empirical phenogram, suppose we represent the topology of the former with a cardinality matrix $\mathbf{C}_C$, and we represent the topology of the latter with a second cardinality matrix $\mathbf{C}_P$. We can compare the two cardinality matrices by subtracting one from the other, forming a *cardinality difference matrix* (**CDM**). Suppose we subtract the cladogram topology from the phenogram topology:

$$\mathbf{CDM} = \mathbf{C}_P - \mathbf{C}_C$$

If the two topologies are identical, then all of the elements of **CDM** will equal zero. If the topologies differ, then some or all of the off-diagonals of **CDM** will be non-zero. The elements of **CDM** are useful for defining an overall measure of topological dissimilarity. As one of a number of possibilities, we can use the absolute value of the off-diagonal element of **CDM** that is furthest from zero, calling that number *C*: $C = \max(\text{abs}(\mathbf{CDM}))$. The minimal value that *C* can take is 0, indicating that the topologies are identical (a perfect phylogenetic signal). The maximal value that *C* can take is the total number of taxa minus 2. An advantage of this particular statistic is that it gives us an idea of the *depth* of the topological dissimilarity (moving from the branch tips toward the root). If $C = 1$, then we know that the greatest difference between the topologies is concentrated near the branch tips. More specifically, if $C = 1$, we know that the greatest differences occur within one or more three-taxon subclades. [Note that *C* does not tell us how many three-taxon subclades differ, only that at there is at

least one difference of that magnitude.] If *C* is maximal, we know that the disagreement between topologies extends all the way to their roots.

Suppose we have used the parametric bootstrap algorithm to estimate *M* bootstrap estimates of *C* (each called *C**). When we look at the distribution of *C**, we see that it contains two types of information (Figure 9). First, the *mode* of the distribution is a measure of the *agreement* between the cladogram and the bootstrap phenograms. The closer the mode is to zero, the greater the agreement tends to be, as described above. Second, the *dispersion* of the distribution is a measure of the *precision* (= repeatability) of *C**. Here, we define precision in a way that is similar to Hillis and Bull's (1993:183-184) definition: "... the correspondence between multiple sets of bootstrap pseudosamples taken from the same initial sample". In our case, we are looking at the correspondence between bootstrap phenograms generated from the same set of sample estimates. Note that while the mode of the distribution can be seen as an indication of a phylogenetic signal, the dispersion does not necessarily lead us to that conclusion by itself. *C** can be highly precise (=repeatable) when bootstrapping under the model without the presence of a phylogenetic signal; it may instead be indicative of a strong alternative signal. By itself, a small dispersion of *C** simply tells us that the hierarchical clusterings that are based on the bootstrap data are highly repeatable. In contrast, if the bootstrap clusterings are highly variable, the precision may be too low to distinguish any repeatable structure in the morphometric data, so that no signal of any kind is detected.

Topology frequencies: We have found that it is instructive to look at the frequencies with which different topologies occur under bootstrap resampling. Here, we are looking at the same data that are used to compute bootstrap proportions, but without condensing them in that way. We simply count the number of times that each topology is observed and express

that number as a percentage of the total number of bootstrap resamples. We then know whether some topologies are more likely to be observed than others, and we can make observations about how these topologies tend to be similar or different.

Given a set of taxa, how many different topologies do we expect to observe when bootstrapping? Initially, we might guess that each possible topology will be observed with the same frequency, given enough resamples. However in practice, we tend to observe far fewer topologies than all those possible. There are two reasons for this. The first is that we would only expect to see each of the possible topologies if the sample means were all the same. If the means differ (as the usually do—otherwise, we would probably not be interested in carrying out the study in the first place), then the number of topologies observed will necessarily be limited. The second reason is that the among-taxon variation observed in interspecific studies tends to be substantially larger than within-taxon variation. Therefore, the relationships between well-separated means in multidimensional space will tend to remain stable under bootstrap sampling. As a result, most of the bootstrap phenograms will probably lie close together in the space of all possible phenogram topologies (see Efron et al. 1996), unless within-taxon variances are relatively very large.

In summary, we recommend the use of a combination of all three of the method just described here. Taken together, they can give a measurement of the repeatability of the data when resampling under the model, they can provide a picture of the taxonomic levels where homoplasies occur, and (primarily in the case of bootstrap proportions) they allow us to "localize" parts of the cladogram where the phylogenetic signal may be particularly strong or may be nonexistent.

A SIMPLE EXAMPLE

To demonstrate our methods, we will examine morphometric variation in the facial skeletons of ateline primates. The subfamily Atelinae is a small clade of Neotropical monkeys that are characterized by large body size and possession of prehensile tails. Despite their close phylogenetic relationships, there is a large degree of anatomical diversity within the clade, especially in skull form. This diversity makes them a particularly interesting group for studies of comparative functional anatomy. There are four living ateline genera: *Ateles* (spider monkeys), *Alouatta* (howler monkeys), *Lagothrix* (woolly monkeys), and *Brachyteles* (muriquis or woolly spider monkeys). Figure 10 shows the hypothesized genealogical relationships among the genera, following Rosenberger and Strier (1989). *Ateles* and *Brachyteles* are the two most closely-related genera. The *Ateles-Brachyteles* clade is then joined by *Lagothrix*, followed by *Alouatta*, which joins as the sister-taxon to the other three genera. This phylogenetic estimate is based on a variety of data (e.g., anatomy, genetics, life history, and behavior), but is not based on skull form. Our sample consists of adult specimens of *Ateles geoffroyi* (N=10); *Alouatta seniculus* (N=6); *Lagothrix lagothricha* (N=14), and *Brachyteles arachnoides* (N=7). To avoid the potentially confounding effects of sexual dimorphism, only females were examined. Variation in midfacial form was initially quantified by recording the three-dimensional positions of six homologous landmarks on the facial skeleton: 1) rhinion; 2) prosthion; 3) premaxilla-maxilla suture at alveolus; 4) inferior end of the zygomaxillary suture; 5) maxillary tuberosity; and 6) posterior nasal spine. Because ateline taxa vary in their adult sizes, we scaled the mean form matrix of each genus by the geometric mean of all interlandmark distances. This scaling measure seems to be a reasonable representation of the overall "volume" of the midfacial skeleton. Figure 11 shows the empirical phenogram that results when $\mathbf{S}_{\Omega}$ (the matrix of shape dissimilarities) is subjected to UPGMA clustering, with a cophenetic correlation > 0.99. Other clustering

methods (including single linkage, complete linkage, and neighbor joining) yielded the same phenogram, so that this particular data set seems robust to differences in the choice of a clustering algorithm. When the empirical phenogram is compared to the cladogram, we see a similarity in the position of *Alouatta*, relative to the *Lagothrix-Ateles-Brachyteles* cluster and clade. So, not only is it more distantly related to *Lagothrix*, *Ateles*, and *Brachyteles* than any of those taxa are to each other, it is the most distinct in shape. The difference between the topologies lies within the *Lagothrix-Ateles-Brachyteles* cluster and clade. In terms of shape, *Lagothrix* and *Ateles* are the most similar taxa, despite of the fact that *Ateles* and *Brachyteles* are the most closely related.

To quantify the effects of uncertainty in measuring shape differences, we generated 500 sets of parametric pseudosamples that were used to obtain a distribution of bootstrap phenograms. Within the bootstrap phenograms, only two topologies were observed (out of the 15 that are possible with four taxa). The most frequently observed topology (481/500 or 96.2%) matched the empirical phenogram, while the remaining topology (19/500 or 3.8%) matched the cladogram. Because we observed one topology an overwhelming majority of the time, and because that topology does not match the cladogram, we have convincing evidence for a strong alternative signal in the morphometric data. At the same time, we have evidence for a high degree of repeatability in the morphometric data, because the overwhelming majority of bootstrap phenograms are identical, matching the empirical phenogram.

Figure 11 shows the bootstrap proportions superimposed on the nodes of the cladogram. *Alouatta* was always the most distinct in shape, so that the bootstrap proportion associated with the *Lagothrix-Ateles-Brachyteles* clade is 100 per cent. Within that clade, we see the very low proportion associated with the *Ateles-Brachyteles* clade (3.8%), indicating that this

close phylogenetic relationship was not reflected in the strongly-patterned morphometric data. Finally, we come to the same conclusion when we look at the bootstrap distribution of the C^* statistic. C^* equals zero 3.8 per cent of the time, indicating that perfect matches between the bootstrap phenograms and the cladogram were rare. C^* equals one 96.2 per cent of the time and never equaled two (the maximum value). This indicates that the mismatches were always restricted to the structure of the *Lagothrix-Ateles-Brachyteles* cluster and never extended any deeper toward the cladogram root (so that *Alouatta* was always the most distinct).

Since our ultimate goal is to better understand how organisms evolve, we would like to suggest an evolutionary scenario that accounts for the strong alternative signal. As we mentioned previously, there is considerable intergeneric variation in ateline facial morphology, and much of this variation has been interpreted in terms of biomechanical adaptation to different diets (e.g., Rosenberger and Strier 1989; Anapol and Lee 1994). *Ateles* and *Lagothrix* are primarily frugivorous, with relatively small and low-crowned postcanine teeth and facial skeletons that are somewhat more gracile, especially for *Ateles*. In contrast, *Alouatta* and *Brachyteles* incorporate a far greater proportion of mature leaves in their diets. As a correlate, they possess very large molars with high shearing crests. Their faces are also substantially larger (relative to the rest of the skull) than in their frugivorous relatives. It is generally thought that the ancestral ateline was probably a generalized frugivore and that the most parsimonious hypothesis for the evolution of dietary specializations in the clade is a parallel acquisition of folivory in *Alouatta* and *Brachyteles* (Rosenberger and Strier 1989; Rosenberger 1992). Therefore, when we return to the empirical phenogram, we might hypothesize that the strong alternative signal is due in part to the symplesiomorphic (shared primitive) retention of a frugivorous diet by both *Ateles* and *Lagothrix*, with a correlated

retention of a primitive facial structure. An obvious question is why *Alouatta* and *Brachyteles* are not more similar, given our hypothesis that they have adopted similar dietary strategies in parallel. We can speculate that the lack of similarity may be is due to *Alouatta*'s unique and bizarre modifications in its hyolaryngeal apparatus (Rosenberger and Strier 1989), so that many of the diet-related morphological similarities that might have arisen between *Alouatta* and *Brachyteles* have simply been "swamped" by *Alouatta*'s many morphological autapomorphies that are not related to its diet. It is interesting to note that, as the analysis now stands, we cannot tell how much of the clustering of *Lagothrix-Ateles* and *Brachyteles* might be attributed to synapomorphies for those data and how much is due to the fact that *Alouatta* is simply so different. To make such a distinction, we would require an outgroup and a way to decompose the morphometric similarities into homologous and homoplasious components (see below under *Further research*).

Finally, we should emphasize again that our interpretations of evolutionary patterns in the atelines is strongly dependent on our assumptions of how the taxa are related. It is clear that our ideas could require substantial revisions if we were confronted with new evidence that would lead us to assume a different cladogram.

DISCUSSION

Uncertainty in constructing the cladogram: As we have mentioned several times, the results that we gain in applying our methods are always contingent on our *a priori* specification of a cladogram. As a result, our interpretations could be impacted substantially if the cladogram were different. In general, we will choose the cladogram that we believe to be correct, given our current knowledge, and we assume that there has been no uncertainty in its construction (Pagel 1999; Huelsenbeck et al. 2000). We realize, of course, that uncertainties also play a

role in estimating phylogenetic relationships and that our assumption is ultimately (though necessarily) a simplification. While considerations cladistic uncertainties are beyond the scope of this study, we are intrigued by the recent work of Huelsenbeck et al. (2000), who have taken a Bayesian perspective toward that problem. Using Markov chain Monte Carlo methods with DNA sequence data, they have presented a method of assigning weights to different cladograms, in an effort to determine which are more likely to be correct, given the data. For researchers who use morphometrics, such methods may ultimately prove to be extremely useful for evaluating the distribution of form variation relative to a set of "credible" cladograms, so that the implications of cladogram differences can be evaluated.

Parametric versus nonparametric bootstrapping: We have stressed parametric methods for pseudosample generation in this paper, but our methods are easily used with nonparametric bootstrapping. There are relative advantages and disadvantages to both resampling methods. As we mentioned previously, one of the primary advantages to using a fitted parametric model is that estimators of uncertainly will generally be more accurate, especially for modest sample sizes. For this reason, we recommend the parametric bootstrap for most applications. However, if we have doubts about the suitability of the parametric model, the nonparametric bootstrap may be preferable. In our ateline example, we believe that the assumption of normally-distributed perturbations is reasonable for modeling within-sample variation. If this assumption proved unreasonable, a nonparametric bootstrap would be a better choice, as the advantages of parametric bootstrapping only hold when the model is valid. Unfortunately, because the Nature Space is unobservable, we cannot test the adequacy of the parametric model directly. Therefore, it might be worthwhile to try nonparametric bootstrapping as a test of the parametric model's suitability. This is because we expect the results to converge (as the number of pseudosamples increases) if the parametric model is a good descriptor of the

observed variation. This type of comparison measures the *robustness of specification* for the parametric model (Davison and Hinkley 1997). If the results differ considerably, it might be advisable either to use the nonparametric results or to use another parametric model.

Generality of the method: While we have emphasized the use of parametric bootstrapping with landmark data and EDMA, the basic bootstrapping strategy that we have outlined is very general. It can be applied with any type of data where the data vary within samples and, preferably, where a reasonable model of within-sample variation can be specified. For example, a researcher might be interested in determining whether a phylogenetic signal is present in the postcranial skeleton for a group of organisms, and the data may consist of the maximum lengths of the limb bones. In that case, within-sample variations can be described by assuming multivariate normal distributions and using the sample mean vectors and variance-covariances matrices to generate pseudosamples. This approach can be taken with either unscaled (form) or scaled (shape) data. The measurements do not even have to be continuously- distributed. Meristic (= count) data can also be used, providing they exhibit within-sample variation. For example, Mosimann et al. (1978) show how a multivariate lognormal model can be used to describe the relative proportions of counts. This model could conceivably serve as the basis for parametric bootstrapping to examine phylogenetic signals in scale counts for reptiles or in fin-ray counts for fishes. In contrast to landmark-based applications, we can apply tests for multivariate normality in these cases to determine whether our model choices are reasonable; otherwise, nonparametric bootstrapping is always an option.

Further research: While signal detection is an important first step in the phylogenetic analysis of morphometric data, it is clearly not an end in itself. Rather, it is an indication that the data contain interesting information that should be investigated further. One particularly

interesting area for future research is the development of methods for studying "mosaic" evolution in a phylogenetic context. Our method for detecting phylogenetic signals is based on measures of morphometric dissimilarity (e.g., F_{Ω}), where differences in many interlandmark distances are summarized in terms of a single number that describes the "overall" degree of difference in form. However, a signal in overall form does not necessarily imply that there is the same pattern or strength of signal for each of the contributing measurements. In reality, complex organisms tend to evolve as "mosaics", where an organism's components or parts can potentially evolve at different rates and in response to different selective pressures (Lande and Arnold 1983). So, the distribution of the "whole" with respect to the phylogeny can be thought of as a consensus of the distribution of the "parts." We would naturally like to decompose this consensus into its component parts. For example, we would like to recognize those parts that are homologous and those that are homoplasious. We would also like the ability to distinguish different patterns and events of homoplasy (e.g., different episodes of convergence, parallelism, and reversal). While there are existing methods of "optimizing" changes in discrete character states over a cladogram (e.g. Swofford and Maddison 1987), there are no analogous methods for studying the evolution of multivariate, continuously-distributed "characters" in a phylogenetic context. We believe that the development of such methods should a high priority in further efforts to bridge the gaps between the "new comparative biology" and the "new morphometrics."

ACKNOWLEDGEMENTS

Thanks to Norman MacLeod and Peter Forey for inviting us to contribute to their Systematics Association symposium and to this volume. Much of this research was supported by NSF Grant DBS 9209083 to JTR and SL, as well as by Project II of NIH Grant 1 P50 DE11131-03 to JTR. Support for collection of the ateline data was provided to TMC by NSF Grant BNS

9020562, by Wenner-Gren Foundation Grant 5303, and by the Field Museum of Natural History. All of the analyses presented here were performed using the *WinEDMA* software package (Cole 2001), available via the Internet at no cost (faith.med.jhmi.edu or c.faculty.umkc.edu/colet).

Footnote

1. The assumptions we must make in this simple case are that all of the lineages evolve at the same rate and that the forms of sister taxa will tend to diverge following a speciation. From another point of view, these assumptions are the conditions that must hold if phenetic similarities are to be an accurate reflection of genealogical relationships (Colless 1970; Sneath and Sokal 1973; Cavlli-Sforza and Piazza 1975; Felsenstein 1982).

REFERENCES

Adams, D.C. and Rosenberg, M.S. (1998) 'Partial warps, phylogeny, and ontogeny: a comment on Fink and Zelditch (1995)', *Systematic Biology*, **47,** 167-172.

Alberch, P. (1980) 'Ontogenesis and morphological diversification', *American Zoologist*, **20,** 653-667.

Alberch, P. (1985) 'Developmental constraints: why St. Bernards often have an extra digit and poodles never do', *American Naturalist*, **126,** 430-433.

Anapol, F. and Lee, S. (1994) 'Morphological adaptation to diet in platyrrhine primates', *American Journal of Physical Anthropology*, **94,** 239-261.

Bock, W.J. and von Wahlert, G. (1965) 'Adaptation and the form-function complex', *Evolution*, **19,** 269-299.

Bookstein, F.L. (1994) 'Can biometrical shape be a homologous character?', in Hall, B.K. (ed) *Homology: the Hierarchical Basis of Comparative Biology*. San Diego: Academic Press.

Bookstein, F.L. (1996) 'Biometrics, biomathematics and the morphometric synthesis', *Bulletin of Mathmatical Biology*, **58,** 313-365.

Brooks, D.R. (1996) 'Explanations of homoplasy at different levels of biological organization', in Sanderson, M.J. and Hufford, L. (eds) *Homplasy: the Recurrence of Similarity in Evolution*. San Diego: Academic Press, pp. 3-36.

Cain, A.J. and Harrison, G.A. (1960) 'Phyletic weighting', *Proceedings of the Zoological Society of London*, **135,** 1-31.

Cavalli-Sforza, L.L. and Piazza, A. (1975) 'Analysis of evolution: evolutionary rates, independence and treeness', *Theoretical Population Biology*, **8,** 127-165.

Chang, J.T. and Kim, J. (1996) 'The measurement of homoplasy: a stochastic view', in Sanderson, M.J. and Hufford, L. (eds) *Homplasy: the Recurrence of Similarity in Evolution*. San Diego: Academic Press, pp. 189-203.

Chappill, J.A. (1989) 'Quantitative characters in phylogenetic analysis', *Cladistics*, **5,** 217-234.

Cheverud, J.M., Dow, M.M. and Leutenegger, W. (1985) 'The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates', *Evolution*, **39,** 1335-1351.

Cheverud, J. M, Wagner, G. and Dow, M.M. (1989) 'Methods for the comparative analysis of variation patterns', *Systematic* Zoology, **38,** 201-213.

Coddington, J.A. (1988) 'Cladistic tests of adaptational hypotheses', *Cladistics*, **4,** 3-22.

Cole, T.M. III (2001) *WinEDMA: Windows-based Software for Euclidean Distance Matrix Analysis*. Kansas City: School of Medicine, University of Missouri-Kansas City.

Colless, D.H. (1970) 'The phenogram as an estimate of phylogeny', *Systematic Zoology*, **19,** 352-362.

Cranston, P.S. and Humphries, C.J. (1988) 'Cladistics and computers: a chironomid conundrum?', *Cladistics*, **4,** 72-92.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and their Application.* Cambridge: Cambridge University Press.

Edwards, A.W.F. (1992) *Likelihood*. Expanded edition. Baltimore: The Johns Hopkins University Press.

Efron, B., Halloran, E. and Holmes, S. (1996) 'Bootstrap confidence levels for phylogenetic trees', *Proceedings of the National Academy of Sciences USA*, **93,** 13429-13434.

Efron, B. and Tibhshirani, R.S. (1991) 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy', *Statistical Science*, **1,** 54-77.

Efron, B. and Tibshirani, R.S. (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. Fourth edition. Harlow, England: Longman.

Felsenstein, J. (1982) 'Numerical methods for inferring evolutionary trees', *Quarterly Review of Biology*, **57,** 379-404.

Felsenstein, J. (1985) 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution*, **39,** 783-791.

Felsenstein, J. (1988) 'Phylogenies and quantitative characters', *Annual Review of Ecology and Systematics*, **19,** 445-471.

Fink, W.L. (1982) 'The conceptual relationship between ontogeny and phylogeny', *Paleobiology*, **8,** 254-264.

Fink, W.L. and Zelditch, M.L. (1995) 'Phylogenetic analysis of ontogenetic shape transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei)', *Systematic Biology*, **44,** 343-360.

Goodall, C.R. (1991) 'Procrustes methods in the statistical analysis of shape', *Journal of the Royal Statistical Society, Series B*, **53,** 285-339.

Gould, S. J. and Lewontin, R. (1979) 'The spandrels of San Marcos and the Panglossian paradigm: a critique of the adaptationist programme', *Proceedings of the Royal Society of London Series B*, **205,** 581-598.

Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Hillis, D.M. and Bull J.J. (1993) 'An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis', *Systematic Biology*, **42,** 182-192.

Hillis, D.M. and Huelsenbeck, J.P. (1992) 'Signal, noise, and reliability in molecular phylogenetic analysis', *Journal of Heredity*, **83,** 189-195.

Houle, D. (1992) 'Comparing evolvability and variability in quantitative traits', *Genetics*, **130,** 195-204.

Hubert, L.J. (1978) 'Generalized proximity function comparison', *British Journal of Mathematical and Statistical Psychology*, **31,** 179-192.

Huelsenbeck, J.P., Hillis, D.M. and Jones, R. (1996) 'Parametric bootstrapping in molecular phylogenetics: applications and performance', in Ferraris, J.D. and Palumbi, S.R. (eds) *Molecular Zoology: Advances, Strategies, and Protocols.* New York: Wiley-Liss, pp. 19-45.

Huelsenbeck, J.P., Rannala, B. and Masly, J.P. (2000) 'Accommodating phylogenetic uncertainty in evolutionary studies', *Science*, **288,** 2349-2350.

Johnson, R.A. and Wichern, D.W. (1982) *Applied Multivariate Statistical Analysis*. Englewood Cliffs, New Jersey: Prentice-Hall.

Lande, R. and Arnold, S.J. (1983) 'The measurement of selection on correlated characters', *Evolution*, **37,** 1210-1226.

Lanyon, S. (1985) 'Detecting internal inconsistencies in distance data', *Systematic Zoology*, **34,** 397-403.

Lapointe, F.-J. and Legendre, P. (1990) 'A statistical framework to test the consensus of two nested classifications', *Systematic Zoology*, **39,** 1-13.

Lapointe, F.-J. and Legendre, P. (1992) 'Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees', *Systematic Biology*, **41,** 378-384.

Larson A. and Losos J.B. (1996) 'Phylogenetic systematics of adaptation', in Rose, M. R. and Lauder, G.V. (eds.) *Adaptation*. San Diego: Academic Press, pp. 187-220.

Lauder, G.V. (1981) ‘Form and function: structural analysis in evolutionary morphology’, *Paleobiology*, **7,** 430-442.

Lauder, G.V. (1990) ‘Functional morphology and systematics: studying functional patterns in an historical context’, *Annual Review of Ecology and Systematics*, **21,** 317-340.

Lele, S. (1991) ‘Some comments on coordinate free and scale invariant methods in morphometrics’, *American Journal of Physical Anthropology*, **85,** 407-418.

Lele, S. (1993) ‘Euclidean distance matrix analysis (EDMA): estimation of mean form and mean form difference’, *Mathematical Geology*, **25,** 573-602.

Lele, S. (1999) ‘Invariance and morphometrics: a critical appraisal of statistical techniques for landmark data’, In Chaplain, M.A.J, Singh, G.D. and MacLachlan, J.C. (eds) *On Growth and Form: Spatio-temporal Pattern Formation in Biology.* Chichester: Wiley, pp. 325-336.

Lele, S. and Cole, T.M. III (1996) ‘A new test for shape differences when variance-covariance matrices are unequal’, *Journal of Human Evolution*, **31,** 193-212.

Lele, S. and McCulloch, C.E. (2001) ‘Invariance and morphometrics’, *Journal of the American Statistical Association*, submitted manuscript.

Lele, S. and Richtsmeier, J.T. (1990) ‘Statistical models in morphometrics: are they realistic?’, *Systematic Zoology*, **39,** 60-69.

Lele, S. and Richtsmeier, J.T. (1991) ‘Euclidean distance matrix analysis: a coordinate free approach to comparing biological shapes using landmark data’, *American Journal of Physical Anthropology*, **86,** 415-428.

Lele, S. and Richtsmeier, J.T. (2000) *An invariant approach to the statistical analysis of form*. London: Chapman & Hall, in press.

Losos, A. and Larson, J.B. (1996) ‘Phylogenetic systematics of adaptation’, in Rose, M.R. and Lauder, G.V. (eds) *Adaptation*. San Diego: Academic Press, pp. 187-220.

Mooers, A. Ø., Vamosi, S. M. and Schluter, D. (1999) 'Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae)', *American Naturalist*, **154,** 249-259.

Mosimann, J.E., Malley, J.D., Cheever, A.W. and Clark, C.B. (1978) 'Size and shape analysis of schistosome egg-counts in Egyptian autopsy data', *Biometrics*, **34,** 341-356.

Mueller, L.D. and Ayala, F.J. (1982) 'Estimation and interpretation of genetic distance in empirical studies', *Genetical Research (Cambridge)*, **40,** 127-137.

Neyman, J. and Scott, E. (1948) 'Consistent estimates based on partially consistent observations', *Econometrika*, **16,** 1-32.

Pagel, M. (1999) 'Inferring the historical patterns of biological evolution', *Nature*, **401,** 877-884.

Penny, D. and Hendy, M.D. (1985) 'The use of tree comparison metrics,' *Systematic Zoology*, **34,** 75-82.

Pimentel, R.A. and Riggins, R. (1987) 'The nature of cladistic data'. *Cladistics*, **3,** 201-209.

Rohlf, F.J. (1974) 'Methods of comparing classifications', *Annual Review of Ecology and Systematics*, **5,** 101-113.

Rohlf, F.J. (1982) 'Consensus indices for comparing classifications', *Mathematical Biosciences*, **59,** 131-144.

Rohlf, F.J. and Bookstein, F.L. (1987) 'A comment of shearing as a method for "size correction"', *Systematic Zoology*, **36,** 356-367.

Rohlf, F.J. and Marcus, L.F. (1993) 'A revolution in morphometrics', *Trends in Ecology and Evolution*, **8,** 129-132.

Rohlf, F.J. (1998) 'On applications of geometric morphometrics to studies of ontogeny and phylogeny', *Systematic Biology*, **47,** 147-158.

Rosenberger, A.L. (1992) ‘Evolution of feeding niches in New World monkeys’, *American Journal of Physical Anthropology*, **88,** 525-562.

Rosenberger, A.L. and Strier, K.B. (1989) ‘Adaptive radiation of the ateline primates’, *Journal of Human Evolution*, **18,** 717-750.

Sanderson, M.L. and Hufford, L. (1996) ‘Homoplasy and the evolutionary process: an afterword’, in Sanderson, M.J. and Hufford, L. (eds) *Homplasy: the Recurrence of Similarity in Evolution*. San Diego: Academic Press, pp. 327-330.

Simpson, G.G. (1961) *Principles of Animal Taxonomy*. New York: Columbia University Press.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*. San Francisco: Freeman.

Sokal, R.R. and Rohlf, F.J. (1962) ‘The comparison of dendrograms by objective methods’, *Taxon* **11,** 33-40.

Wake, D.B. and Larson, A. (1987) ‘Multidimensional analysis of an evolving lineage’, *Science*, **238,** 42-48.

Zelditch, M.L., Fink, W.L. and Swiderski, D.L. (1995) ‘Morphometrics, homology, and phylogenetics: quantified characters as synapomorphies’, *Systematic Biology*, **44,** 179-189.

Zelditch, M.L., Swiderski, D.L. and Fink, W.L. (2000) ‘Discovery of phylogenetic characters in morphometric data’, in Wiens, J.J. (ed) *Phylogenetic Analysis of Morphological Data*. Washington, DC: Smithsonian Institution Press, pp. 37-83.

List of Figures

specimen). These locations are phenotypic perturbations of the mean, which relfect underlying genetic and environmental variations. Note that the dispersion patterns differ from one landmark to the next. Some landmarks have roughly circular distributions (1, 2, and 3), while others are elliptical. Some landmarks (1 and 3) have relatively small dispersions, while others (2) are large. In addition, some of the perturbations may be correlated – note the similarity in the rank-order of perturbations (from upper left to lower right) for landmarks 4 and 5. As described in the text, the positions of the perturbations in Nature Space cannot be reconstructed; however, some descriptors of the dispersion patterns can be estimated.

Figure 7. Schematic of the parametric bootstrapping algorithm used in this study. Data from four taxa (*A*, *B*, *C*, and *D*) are used to estimate mean forms ($\hat{\mathbf{M}}$) and among-landmark variance-covariance matrices (Σ_K^*). The means are first used to estimate the empirical tree, shown at the bottom of the figure. The means and covariance matrices are then used to generate a large number (e.g. 500) of pseudosamples, assuming multivariate Gaussian perturbations. Each set of pseudosamples is, in turn, used to estimate a bootstrap tree. The intermediate step of computing a dissimilarity matrix is not shown. Finally, each of the bootstrap trees is compared with the cladogram, using bootstrap proportions, tree-comparison statistics, or some other measure of topological similarity.

Figure 8. Examples of how the topologies of phenograms and cladograms are expressed using cardinality matrices. See text for an explanation of how the matrix

elements are defined. For the cladogram at the bottom of the figure, note how unresolved multifurcations can be accommodated.

Figure 9. Hypothetical bootstrap distributions of the *C** statistic. A perfect agreement between topologies is indicated by a value of zero. At the upper left, there is a case where there is a strong phylogenetic signal (mode of zero) and high repeatability, indicated by the small dispersion. At the upper right, there is still evidence of a phylogenetic signal (mode of zero), but it is less repeatable, indicated by the greater dispersion. At the lower left is a strong alternative signal, indicated by the combination of a lack of agreement (mode very different from zero) and high repeatability. At the lower right, there is no apparent signal, with a large dispersion and no distinct mode.

Figure 10. Phylogenetic relationships of living ateline genera, following Rosenberger and Strier (1989).

Figure 11. Results of the bootstrap analysis. The cladogram (left) is shown with bootstrap proportions that measure the agreement between it and the structures of the bootstrap phenograms. The UPGMA phenogram (right) is shown with bootstrap proportions that measure its repeatability (see text). The bootstrap distribution for the *C** statistic is shown in the center.

Figure 1.

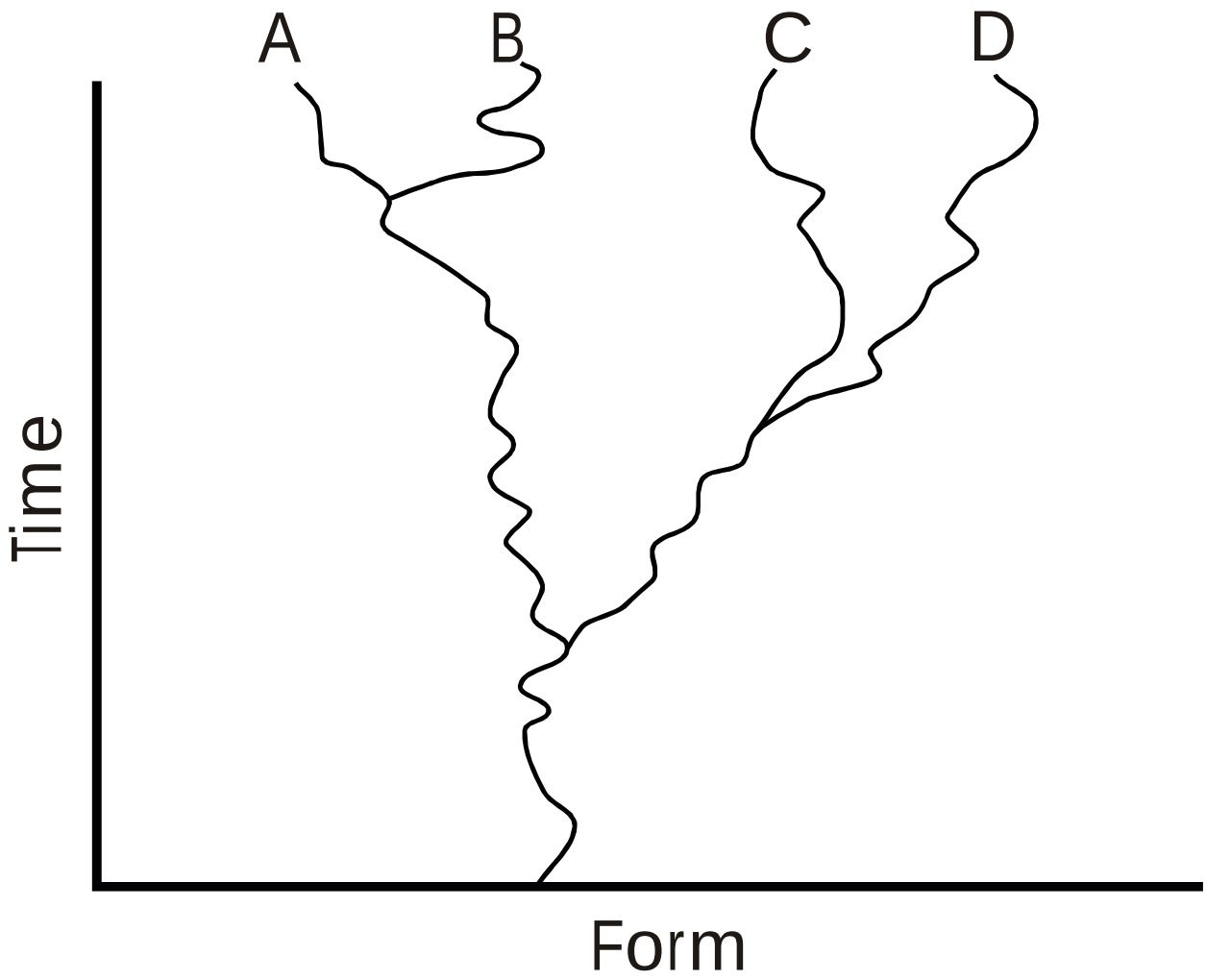


Figure 2.

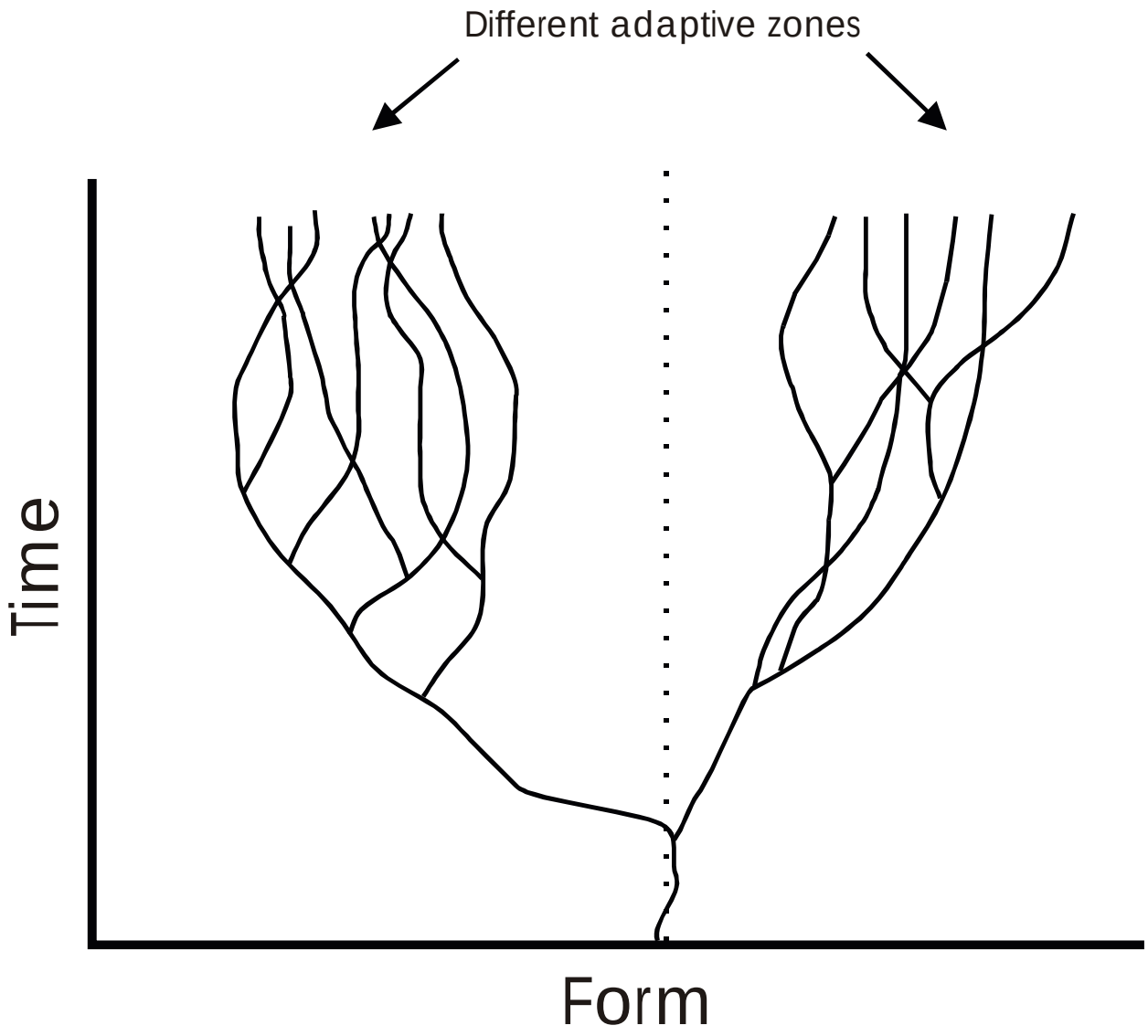


Figure 3.

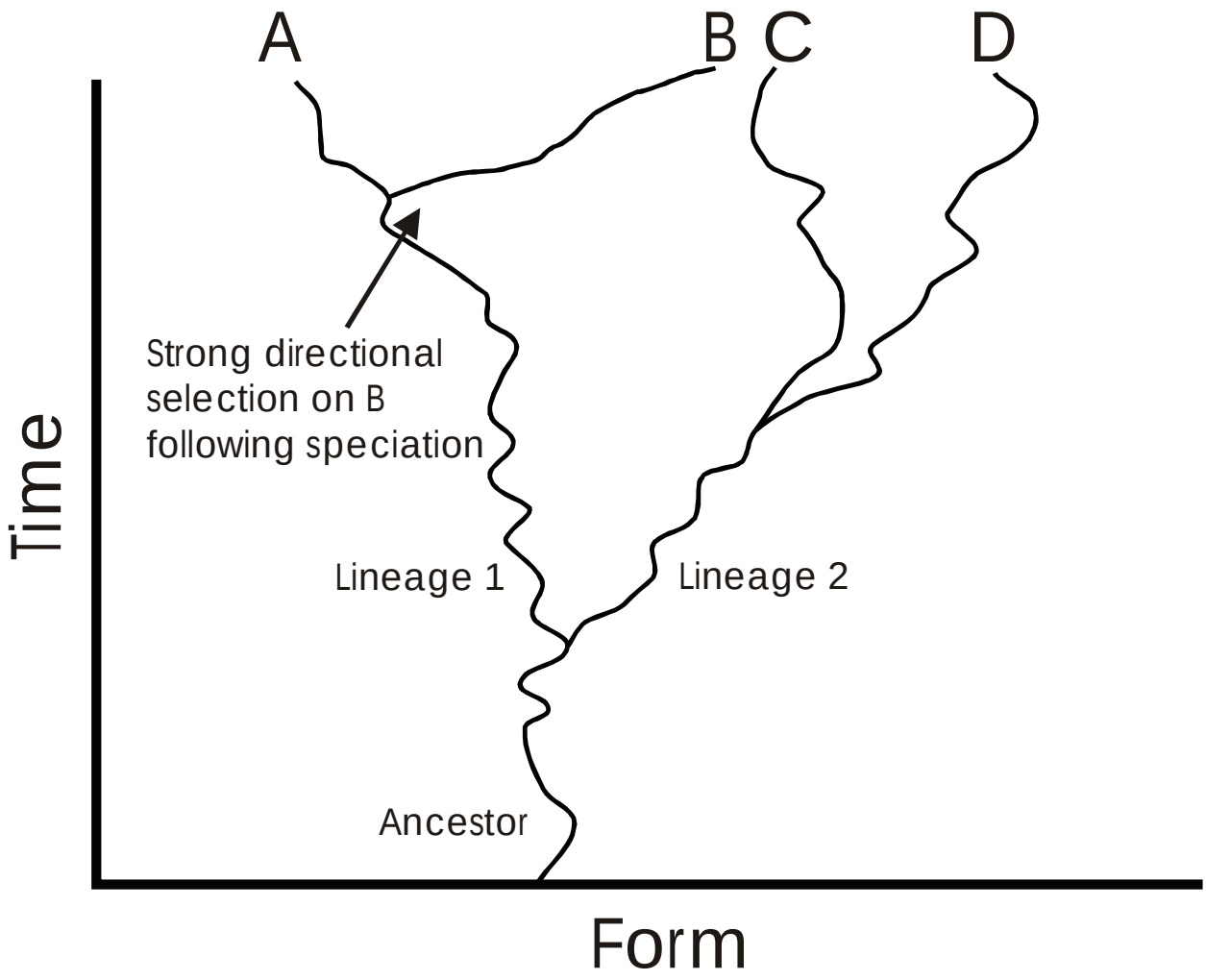

A
B
C
D
Strong directional
selection on B
following speciation
Lineage 1
Lineage 2
Ancestor
Time
Form

Figure 4.

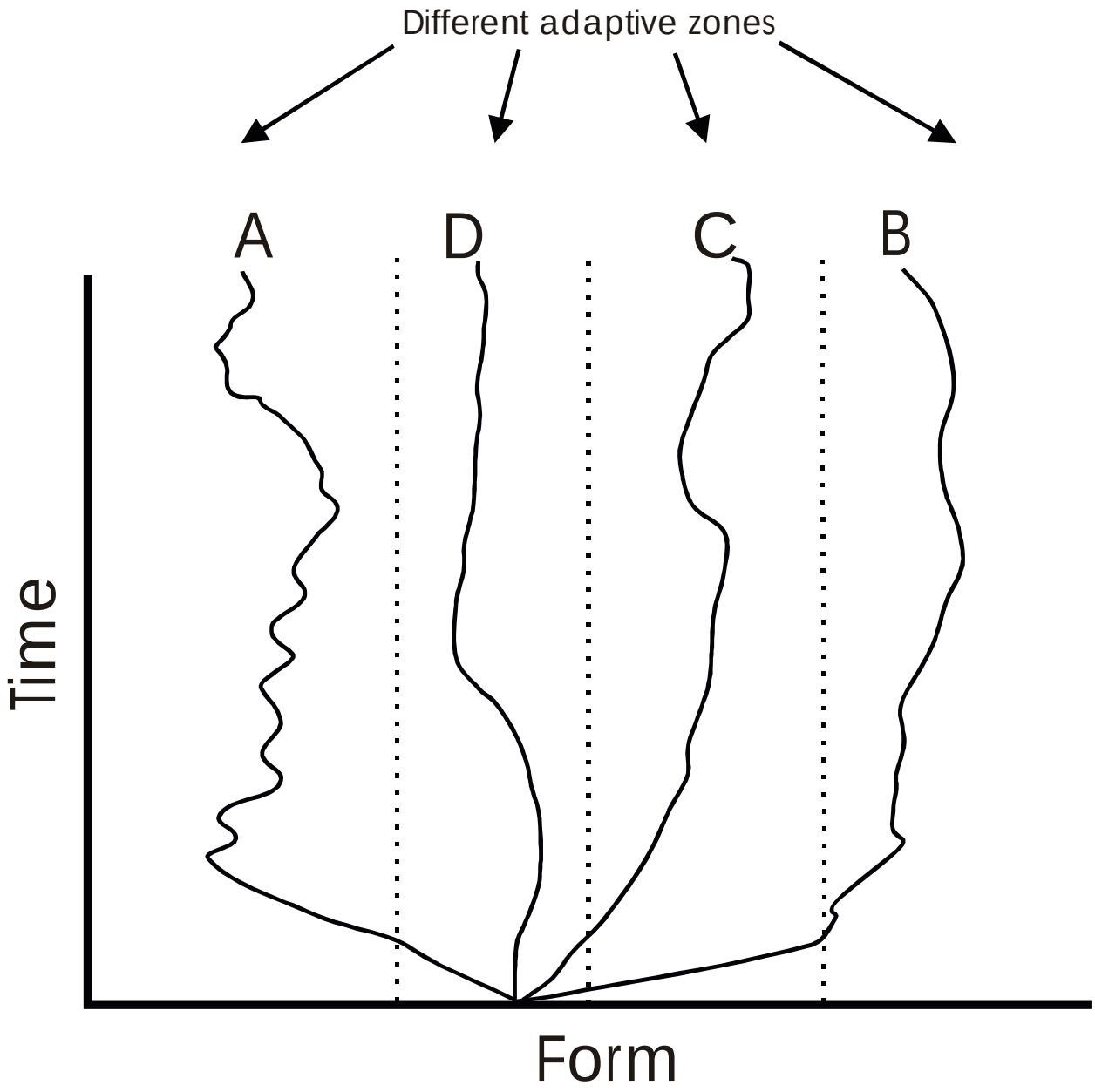


Figure 5.

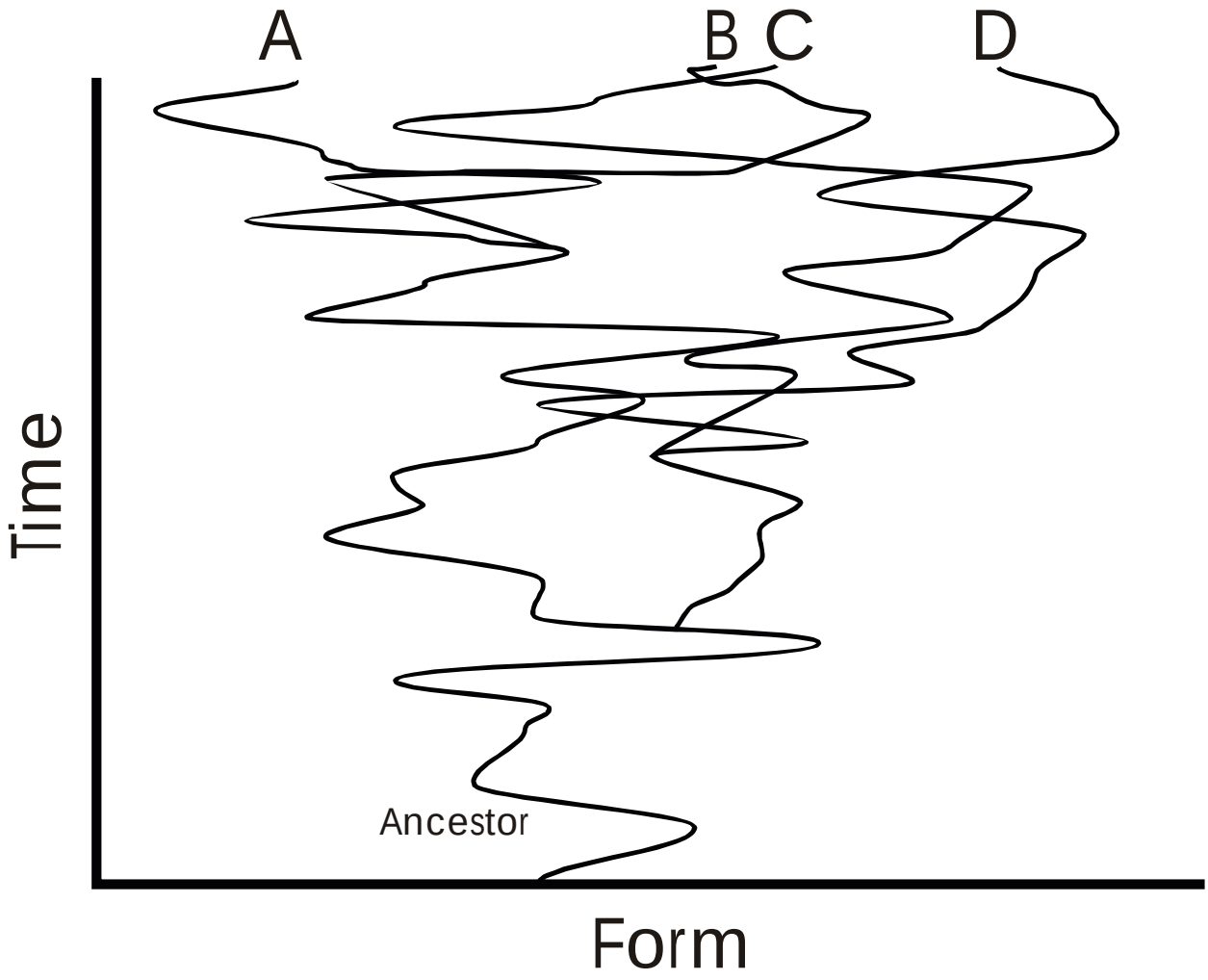


Figure 6.

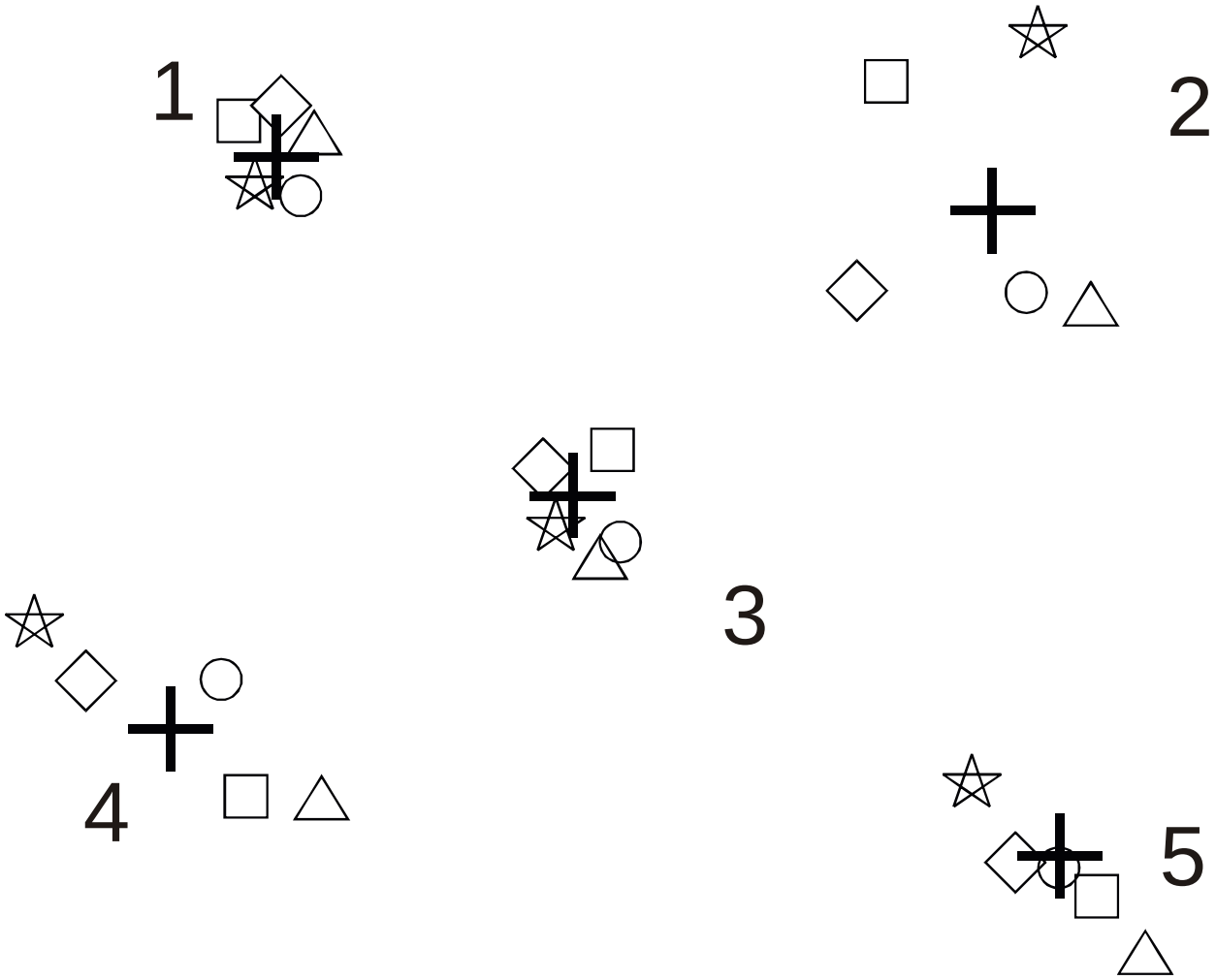


Figure 7.

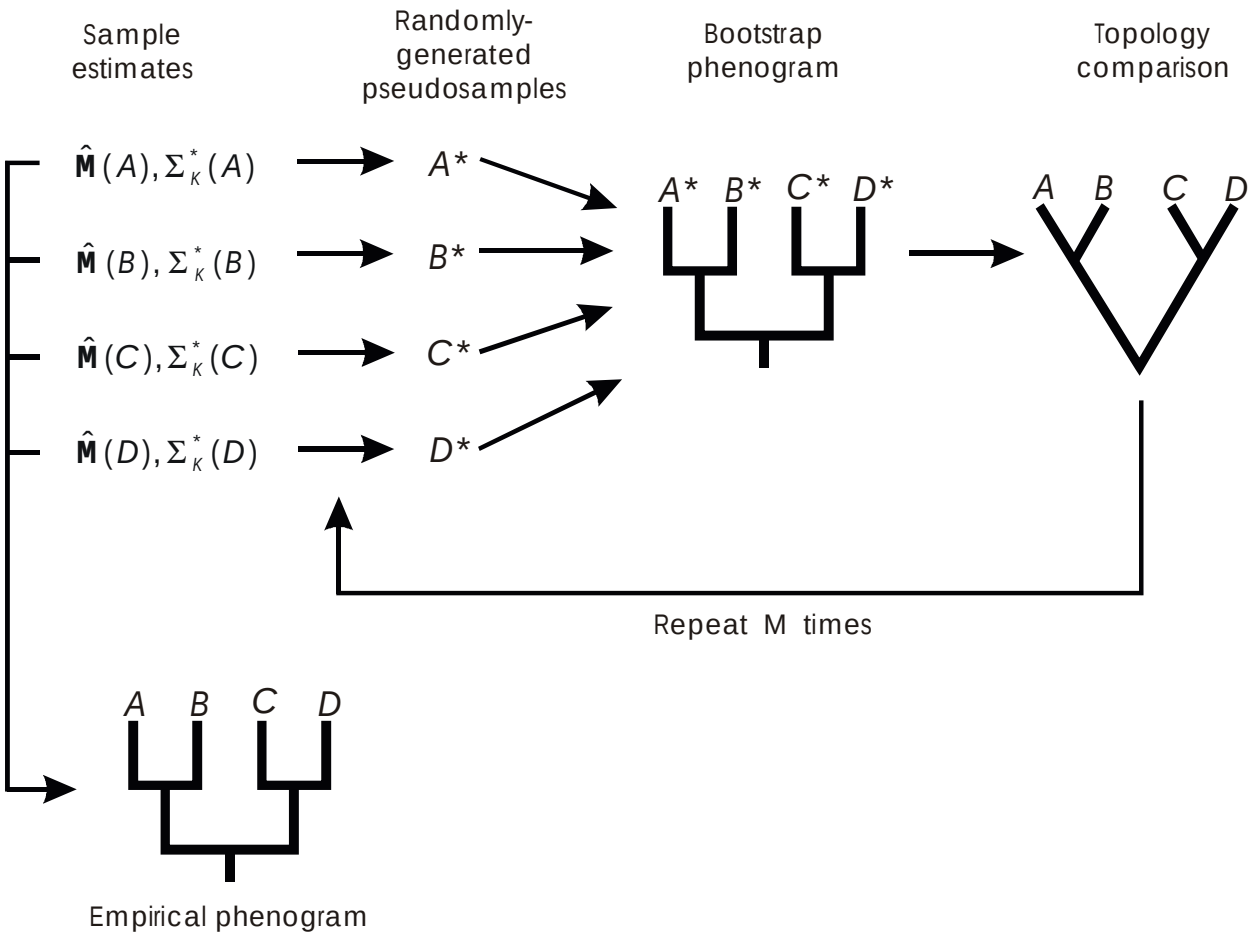
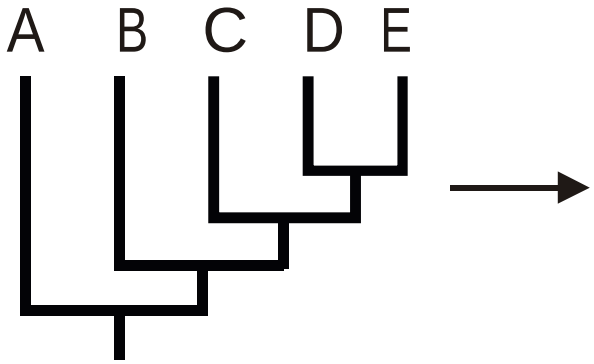
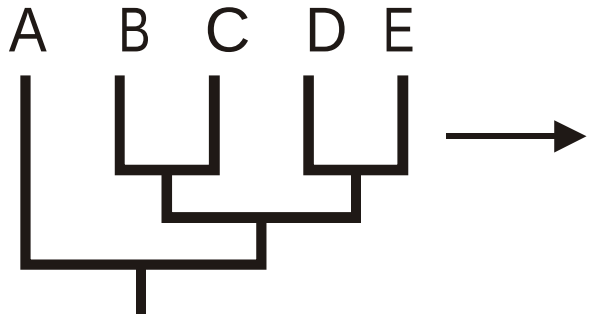

Sample estimates
Randomly-generated pseudosamples
Bootstrap phenogram
Topology comparison
$\hat{\mathbf{M}}(A), \Sigma_K^*(A)$
$\hat{\mathbf{M}}(B), \Sigma_K^*(B)$
$\hat{\mathbf{M}}(C), \Sigma_K^*(C)$
$\hat{\mathbf{M}}(D), \Sigma_K^*(D)$
A* B* C* D*
A B C D
Repeat M times
A B C D
Empirical phenogram

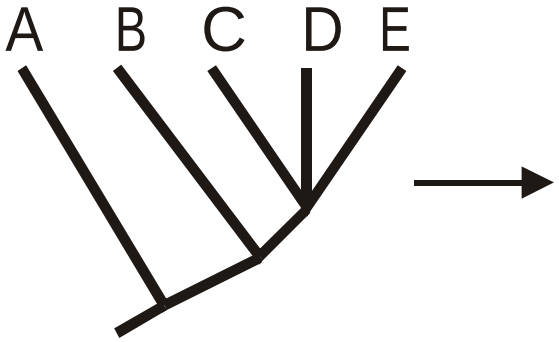
Figure 8.



	A	B	C	D	E
A	0				
B	5	0			
C	5	4	0		
D	5	4	3	0	
E	5	4	3	2	0



	A	B	C	D	E
A	0				
B	5	0			
C	5	2	0		
D	5	4	4	0	
E	5	4	4	2	0



	A	B	C	D	E
A	0				
B	5	0			
C	5	4	0		
D	5	4	3	0	
E	5	4	3	3	0

Figure 9.

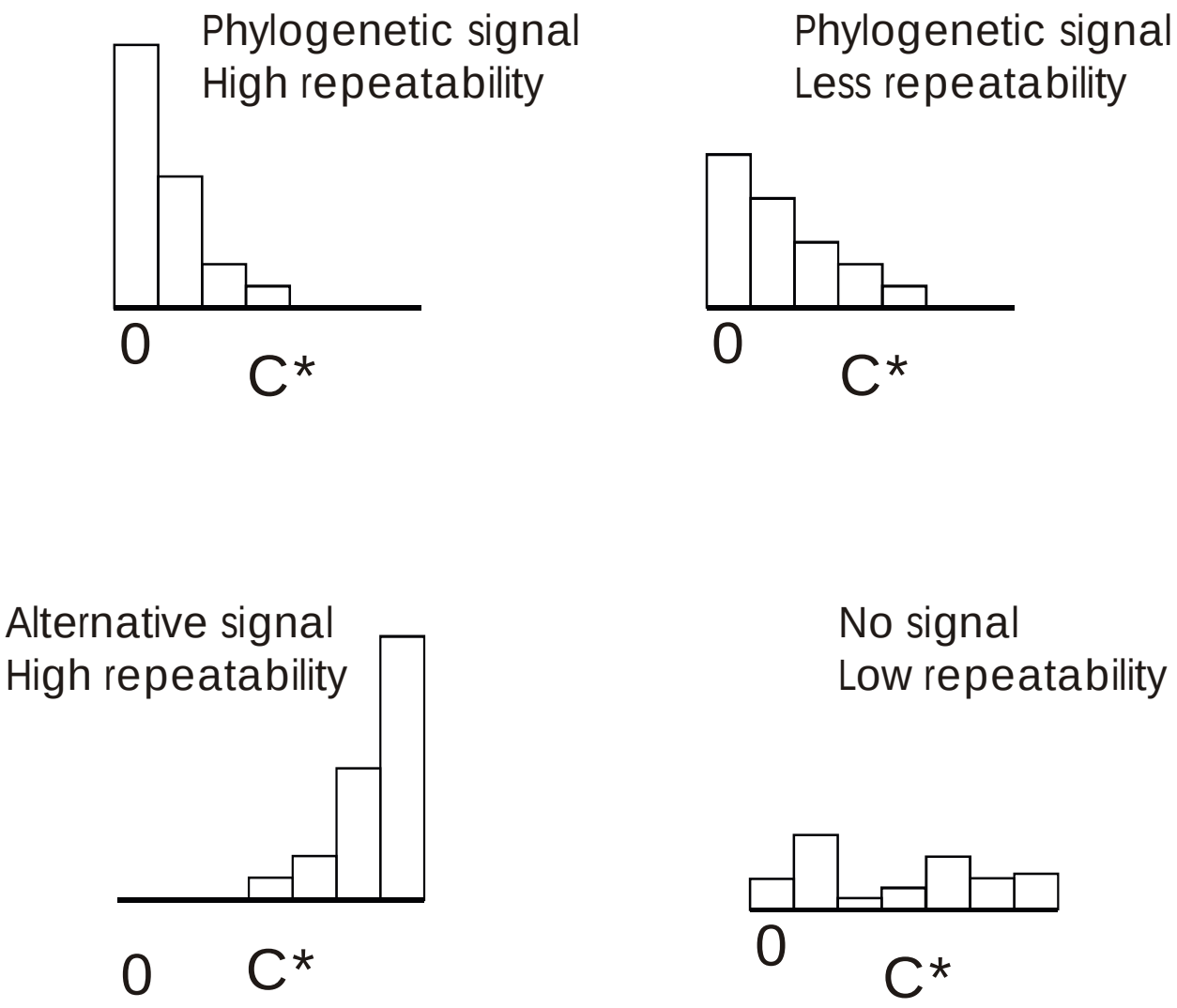

Phylogenetic signal
High repeatability
0
C*
Phylogenetic signal
Less repeatability
0
C*
Alternative signal
High repeatability
0
C*
No signal
Low repeatability
0
C*

Figure 10.

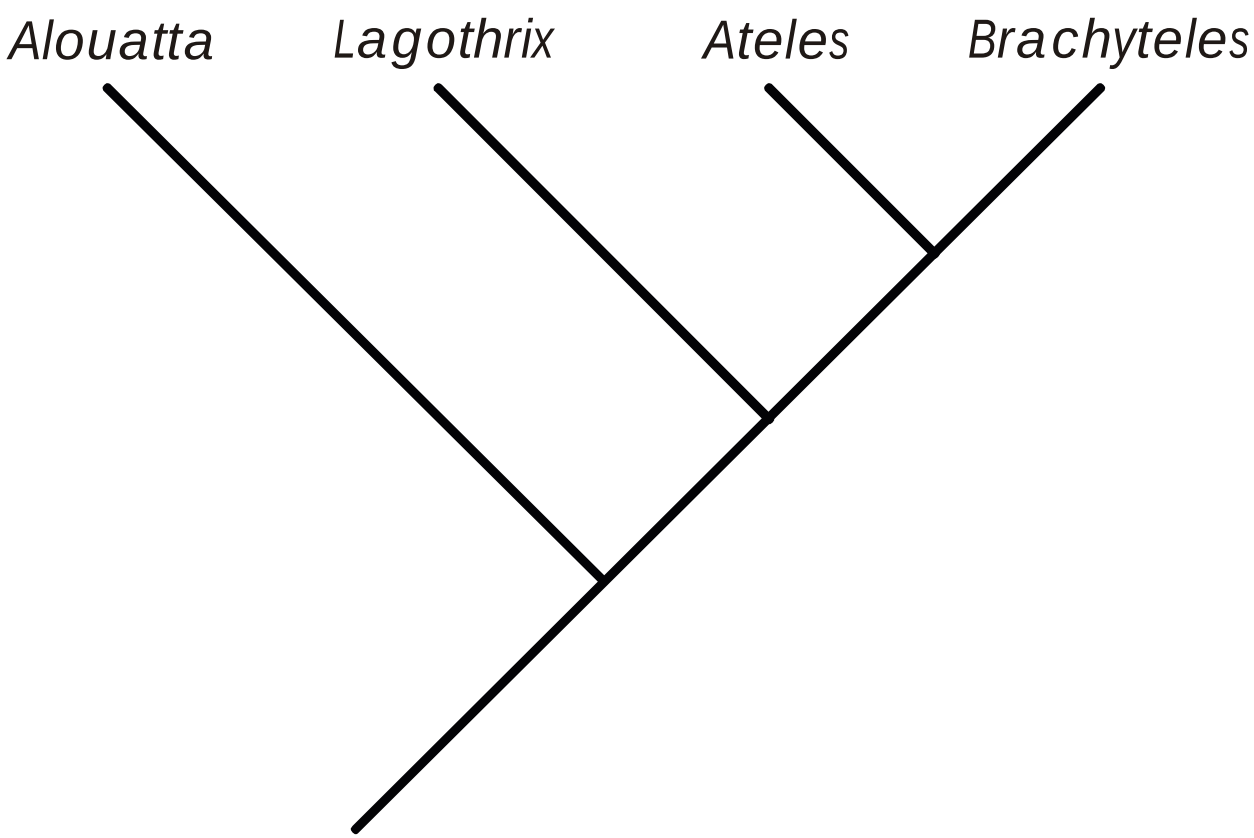
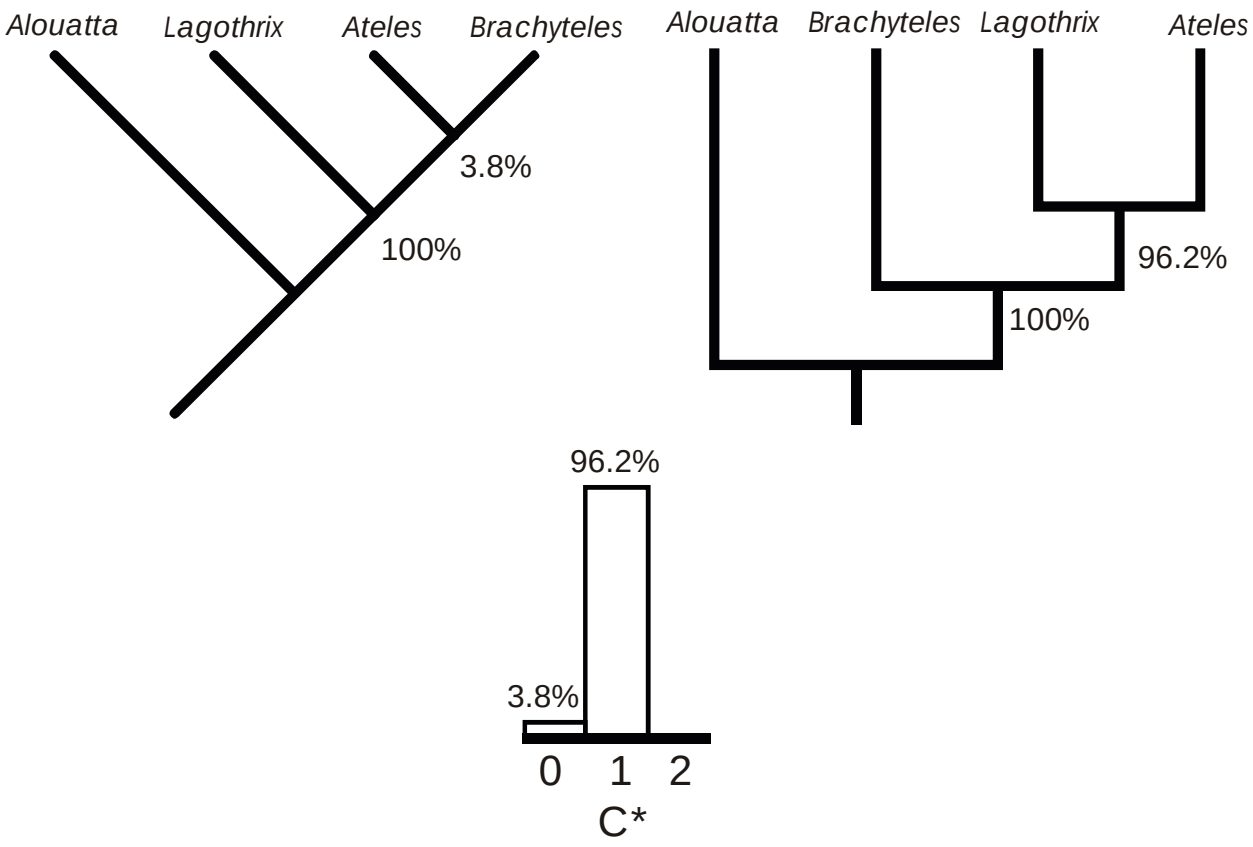


Figure 11.


Alouatta
Lagothrix
Ateles
Brachyteles
3.8%
100%
Alouatta
Brachyteles
Lagothrix
Ateles
96.2%
100%
96.2%
3.8%
0
1
2
C*