5

Comparative analysis of continuous variables

'Independent evolution may be the ideal criterion for the comparative method' (Ridley 1983*a*, p. 18)

5.1 Introduction

Many of the questions of interest to comparative biologists involve comparing the values of characters that vary continuously, rather than discretely, among a number of species. Measures of morphology, physiology, life histories, and behaviour, such as walking, flying, or swimming, produce quantitative values. Comparative studies of characters that vary continuously have progressed from relying nearly exclusively on cross-species correlational analyses that ignore the historical relationships among species, to sophisticated techniques that incorporate information about phylogenetic relationships into the comparative test. We review briefly each of the major approaches in this chapter. Those comparative methods that compute sets of statistically independent comparisons, either across contemporary taxa, or between ancestral and descendant nodes, emerge as the best techniques currently available. This conclusion is supported by the results of computer simulation studies.

As with tests for discrete characters, tests for continuously varying characters are confronted with the problem that species form a nested hierarchy of phylogenetic relationships. For the reasons described in Chapter 2, closely related species are typically more similar than are more distantly related species. This means that species cannot be treated as independent units of information in statistical tests. The various comparative techniques that we review in this chapter can be distinguished by the procedures that they use to manage the effects of similarity associated with phylogenetic relationships. Each of the methods is based on a set of, often implicit, assumptions that comprise a null hypothesis of evolutionary change, and on a set of statistical techniques that apply those assumptions to

real data. The statistical techniques are designed to produce data points that can be treated as independent for statistical analysis.

Should we even bother with a separate class of methods for continuous variables? In a mathematical sense, the distinction between models for discrete variables and models for continuous variables hides a deeper equivalence between the two cases. Continuous variables, the topic of this chapter, are merely discrete variables in which we allow the width of the discrete units to become vanishingly small. Assume that change in a continuous metrical character proceeds by a series of discrete steps. The 'distance' moved each time is a fixed amount either 'forward' one unit (+1) or 'backward' one unit (−1) along the scale. Forward and backward movements are equally likely. Let there be a single step in each of a number of successive epochs of time, where a single epoch is denoted by τ. The position along the scale after n such epochs is the sum of all the preceding steps. These circumstances give rise to what is known as a random walk. Some random walks will have a preponderance of +1s, others a preponderance of −1s, others a more even number of pluses and minuses. Given that one of two possible outcomes occurs at each interval of time, there will be 2^n possible outcomes of the random walk after n steps.

Statistical theory informs us about the expected distribution of outcomes of such a process when we allow the unit of time τ to become vanishingly small, and thereby allow n to become large. It can be deduced from the Central Limit Theorem that, in the limit as τ goes to zero, the outcomes of the random walk after some unit of time t will be normally distributed with an expected value of zero, and variance of $\sigma^2 t$, where σ^2 is a positive constant. In continuous time, the discrete process is transformed into one of Brownian motion where the variance of change accumulates in direct proportion to the amount of time the process has been allowed to go. We make use of this result in the next section to illustrate various properties associated with phylogenies.

We begin this chapter with a discussion of the statistical problems that arise when data obtained from phylogenies are used in conventional statistical procedures. These are the same problems that arose in connection with methods for discrete characters, but now we discuss them using statistical models that are more convenient for continuously varying characters. We then show how each of the more recently developed techniques confronts those statistical problems. Although our account is somewhat historical, it also develops as a logical progression of ever more acceptable procedures. Our final sections present the most recently developed techniques, and then summarize selected results from computer simulations which serve to underline one of the main messages of this book: we make assumptions about the way evolution proceeds whenever we choose a comparative test.

5.2 Testing hypotheses on continuous variables

The simplest way to test for a relationship between two continuous variables is to treat species as independent data points and apply standard statistical techniques to characterize their relationship. This was also true for testing the relationship between two discrete variables. However, as we pointed out in Chapters 2 and 4, such an approach is bound to be flawed statistically because species are part of a hierarchical phylogeny. This fact virtually guarantees that each species will not have independently evolved the suite of traits that defines its phenotype, thus posing a critical problem for statistical methods which assume that the data points are independent.

To see why the preceding statement is true, it is necessary to think about the various ways that, for any given set of contemporary species, evolution could have arrived at the present. Consider the phylogeny of the eight contemporary species in Fig. 5.1. This figure is similar to Fig. 4.1, except now the species values represent two continuous characters.

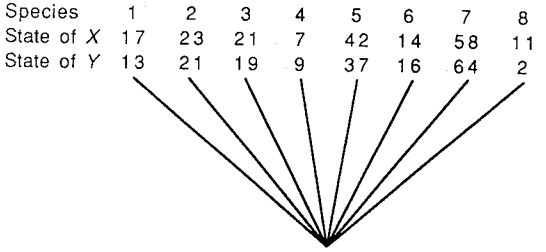


Fig. 5.1. Eight species simultaneously evolving from a common ancestor. Values of two continuously varying characters (X and Y) are given for each species. Figure 4.1 uses the same example with discrete characters.

For the purpose of discussion, we assume that characters evolve independently according to a Brownian motion process, following Edwards and Cavalli-Sforza (1964) and Felsenstein (e.g. 1981*b*, 1985*a*, 1988). This statistical model is appropriate for describing the random wanderings of a variable along a continuous dimension (see Section 5.1). Figure 5.2 plots four random walk sequences in which the value of the random walk at any point is the sum of all of the changes before it. The four sequences can be thought of as characterizing the evolutionary changes through time in any four species from Fig. 5.1. The displacements along the horizontal axis represent the value of a character, and the vertical axis is time.

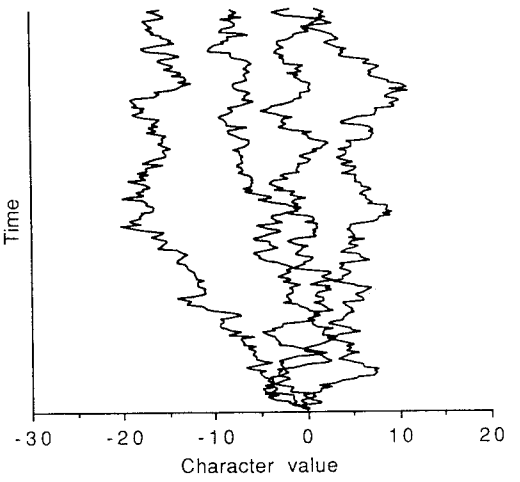


Fig. 5.2. Character change among four independently evolving lineages following random walks after splitting from a common ancestor at time zero. The value of each random walk at any point is the sum of all previous changes for that lineage. See also Felsenstein (1988).

Alternatively, the sequences in Fig. 5.2 can be thought of as iterations of the same random evolutionary process for a single species. This interpretation makes explicit what is meant by the expected variance of evolutionary change. If we were to re-run the random walk many times, the variance of the end-points would be an estimate of the expected variance of evolutionary change for that particular branch length. More generally, if the variance of a single step of the random walk is σ^2, then the expected variance of end-states after time period t will be $t\sigma^2$.

Thinking about the evolution of two or more species in terms of random walks can be used to illustrate why phylogenies might introduce correlations among species. If the phylogeny of species is like that in Fig. 5.1, then the random walks illustrated in Fig. 5.2 make the point that the evolutionary changes in the species since their common starting point have been independent. Thus, given the Brownian motion model, and the phylogeny of Fig. 5.1, we can treat species as independent points for statistical purposes. Furthermore, because the statistical properties of the random walk are known, we can say something about the expected change in each branch, as well as the expected variance of change in each branch.

But now consider that the eight species have the phylogeny shown in Fig. 5.3. This is, again, similar to a figure (Fig. 4.2) used in Chapter 4, except now the characters are continuous.

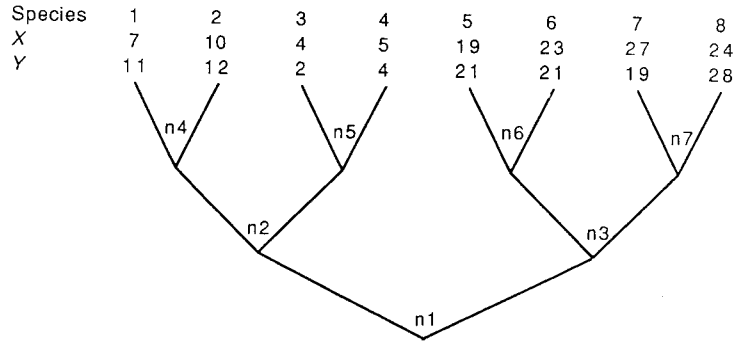


Fig. 5.3. A dichotomously branching phylogenetic tree showing the evolutionary history of eight species. Values of two continously varying characters (X and Y) are given for each species.

It will often be observed in such a phylogeny that the species in each pair of tips will be more similar to each other than to the other species. The Brownian motion model along with the phylogenetic structure can be used to illustrate this phylogenetic similarity. Figure 5.4 displays the results of a random walk that could correspond to species 1 through 4 from Fig. 5.3. At the beginning of the sequence there is but one path, indicating the common ancestry of all four species. At a later time the bifurcation corresponding to node n2 occurs. Later yet, nodes n4 and n5 occur. At the end of the sequence species 1 and 2 are closer to each other than to species 3 and 4. Their shared history introduces a correlation between them, even though the evolutionary changes along all branches of the tree have been independent. The sobering message from Fig. 5.4 is that phylogenetic similarity can arise from a completely random process. Imagine the similarity that arises when we allow for the processes described in Chapter 2!

Similarity associated with phylogeny causes statistical problems. Most statistical methods assume that the data points can be thought of as (1) having been sampled independently from (2) a normal distribution with some mean and variance. Phylogenetic similarity, however, by introducing a correlation among characters, invalidates the first assumption. The second assumption is slightly more complicated. Anticipating issues to be discussed later in this chapter, there are two main ways that points sampled from a phylogeny may have different expected variances.

First, we have assumed that the Brownian motion process proceeds at the same rate everywhere in the phylogeny. This guarantees that over any given amount of time, all lineages with a common starting point will have

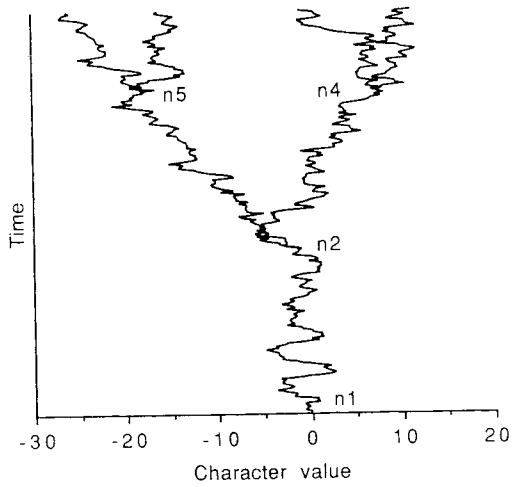


Fig. 5.4. Character change among four lineages following random walks after splitting from a common ancestor at time zero. The value of each random walk at any point is the sum of all previous changes for that lineage. Unlike the lineages in Fig. 5.2, the ones represented here have not evolved independently but have shared varying amounts of common ancestry since time 0, forming part of the phylogenetic tree shown in Fig. 5.3 There is a lineage split at n2, and two subsequent splits at n4 and n5. See also Felsenstein (1988).

the same expected variance of change. However, if we relax the biologically unrealistic assumption of equal rates of change (Simpson 1945), different expected variances of change occur in different lineages, thereby invalidating the assumption of equality of variances. Second, inequality of variances may also arise even if σ^2 is fixed if we compare two or more values derived from the phylogeny that do not have a common starting point, and which therefore may not have been evolving for the same length of time.

We have developed these points in detail because they illustrate all of the essential statistical and evolutionary considerations that must go into developing a test of a comparative relationship. We have statistical problems because phylogenetic relationships cause a lack of independence among the data points, and because unequal rates of change or differing time periods of change in different lineages may introduce inequality of variances. Fundamentally the same problems were encountered in Chapter 4. Either of these two problems—lack of independence or heterogeneity of variances—renders many standard statistical tests invalid. There are

statistical techniques for managing the effects of non-independence and unequal variances (Generalized Least Squares, see Draper and Smith 1981), but they depend upon being able to specify the nature of the correlation among species, and the expected variances of their characters. This is where the evolutionary considerations become central: our beliefs about the extent of the correlation among species and the extent to which variances are unequal depend upon our assumptions about how evolution proceeds. Does evolution move along at a constant rate in all branches or does it have different rates in different branches? Alternatively, is evolutionary change punctuational such that branch length is less important? Or, is it such a jumble of these two processes (and others) that it is futile to assume one or the other? The answers that we give to these questions are crucial and, as with similar questions in Chapter 4, they define the various different methods of comparative analysis.

There cannot be 'solutions' to the problems posed by comparative data, then, only approximations to solutions based upon our current understanding of evolution. Some approximations will be demonstrably better than others for particular situations but, in many instances, the validity of a technique will be unknown. This is an important point because it means that the choice of a particular approach for analysing comparative data will often depend less upon knowledge that one technique is superior to another, than on a set of beliefs about the workings of evolution for a particular set of species and variables.

In the sections that follow, we describe different methods that have been used to analyse comparative data on continuous variables. Each method can be characterized by the statistical techniques that it uses to produce (at least in theory) data points for comparative analysis that are independent, and that all have the same expected variance. Some of the methods attempt to manage the effects of phylogenetic similarity by estimating the extent of non-independence and heterogeneity of variance in the data on the basis of the phylogenetic tree. Other methods create independent data points by discarding the variation that is thought to reflect phylogenetic similarity. An additional class of methods based on equally plausible models can test the comparative relationship without discarding any of the information. This class of methods attempts to define a set of mutually independent comparisons calculated from the phylogeny. Comparisons can be either between species or between higher nodes, or between the beginning and end-states along a branch, in which case the direction of phylogenetic change is of interest (directional versus non-directional comparisons are discussed in Chapter 1). In either case, the comparisons are then scaled according to empirical or model-based rules that attempt to equalize their variances. We argue that these techniques are the best currently available for conducting comparative tests.

For the rest of this chapter, we shall make repeated reference to Figs 5.1–5.4 which are drawn, following convention, with extant species at the top. Unfortunately, there is another convention that we must follow, and that is to talk of 'higher' versus 'lower' taxa, and 'higher' versus 'lower' nodes. As far as phylogenetic trees are concerned, typified by Figs 5.1 and 5.3, higher taxa and nodes are usually drawn at the bottom of the tree and lower taxa and nodes are at the top! For example, a species is a lower taxon than a family. It seems unnatural to break either of these conventions, and appropriate care must thus be taken when translating between the text and the figures.

5.2.1 A cautionary note

Unless otherwise stated, we assume in our discussions of the various methods that the phylogenies are known without error. This, of course, will seldom be true, but it is an important assumption, given the uncertainties about whether a reconstructed phylogeny is the true phylogeny, even when the best methods are used (see Chapter 3). Some work has been done on the issue of placing confidence limits on estimates of phylogenies (see Felsenstein 1985*c*, and references therein). Little is known about how sensitive the conclusions of a particular study will be to the tree that is used. Until more work is done in this area, comparative biologists should be aware that their results may depend upon a particular reconstruction of a phylogeny. If several phylogenies are equally likely (or parsimonious, or compatible), then perhaps the analyses should be done using them all (see Björklund in Harvey 1991 for an example). If the conclusions vary widely, caution should be exercised in their interpretation.

5.3 Species analyses

Species values have been used as the units for statistical analysis in the vast majority of comparative studies that have analysed relationships between continuously varying characters[11]. We have stated that species cannot be assumed to be independent and that they may have different expected variances. So, what evolutionary assumptions are implicit in using species as the units of analysis? That is, what models of evolution would give rise to species being independent and having the same variances?

As was true for discrete characters (following Fig. 4.1), the model implicit in Fig. 5.1 provides one answer. Under this model of evolution, to use species as the units of analysis means that the investigator must be

[11] Examples from a variety of taxa and fields of biological enquiry include: Quiring (1941); Newell (1949); Hutchinson and MacArthur (1959); Southwood (1961); McNab (1963); Schoener (1968); Millar (1977); Clutton-Brock *et al.* (1977); Wooton (1987); Burt and Bell (1987); Clutton-Brock (1989).

willing to defend the belief that the phylogeny of the species can be represented by a simultaneous radiation of all of the species from a single common ancestor (see also Felsenstein 1985*a*). That is, the investigator must assume that there are no shared branches in the phylogeny. It must also be assumed that rates of evolutionary change are the same in each branch, an assumption guaranteeing that each datum has the same expected variance. These assumptions, if met, would mean that the different species could be treated as random samples from some common underlying distribution of possible outcomes. However, true simultaneous radiations may be rare in nature, and do not characterize the phylogenies of interest to most biologists.

Thus, to use species as independent data points in a comparative analysis requires that one ignores phylogenetic relationships. This should be anathema for anyone who believes in evolution. Nevertheless, species are used this way in comparative analyses, and so it is prudent to be aware of the consequences. The major problem is that the confidence limits on statistics are sensitive to the number of degrees of freedom declared in the analysis. Because species are typically not independent, confidence limits will be spuriously narrow. This can lead to rejection of a null hypothesis on false grounds. Allometric studies that employ species provide good examples.

Figure 5.5 shows a logarithmically scaled plot of home range size against body weight for 72 primate species. The slope of the model 1 regression line describing the best fit relationship between home range size and body weight is 1.26, with 95 per cent confidence limits from 0.95 to 1.58. A slope of 0.75, which might have been expected on energetic grounds (Kleiber 1961), lies outside these confidence limits. However, it is not valid to reject

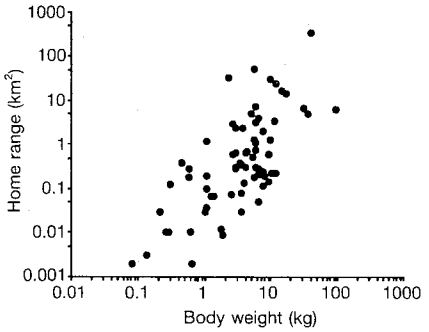


Fig. 5.5. Home range size increases with body weight across a sample of 72 primate species. When both axes are scaled logarithmically, the relationship is approximately linear with a slope of about 1.26. (Data from Harvey and Clutton-Brock 1981, and subsequent sources).

0.75 for this reason alone. The sensitivity of the analysis to the number of degrees of freedom is illustrated by the fact that if the sample size is reduced to 30, which is approximately the number of genera in the sample, the confidence limits widen to embrace 0.75.

5.4 Analysis of higher nodes

If species cannot be considered independent then perhaps some higher node can be? Crook (1965) suggested using genera and family means rather than species, but it was not until more than a decade later that an explicit statistical criterion was offered for this practice. Clutton-Brock and Harvey (1977), Harvey and Mace (1982), and Harvey and Clutton-Brock (1985) developed the use of the nested analysis of variance (Sokal and Rohlf 1969) to describe how the total variation among species in a continuous character is distributed among the taxonomic levels. The distribution of variance by taxonomic level is, in turn, used to identify which taxonomic level to use as the unit of analysis.

The nested ANOVA partitions the total variation among species into components representing each of the nested levels in a taxonomy:

$$\sigma^2_{tot} = \sigma^2_{s(g)} + \sigma^2_{g(f)} + \sigma^2_{f(o)} + \sigma^2_{o(c)}. \quad (5.1)$$

The term on the left is the total variance among species on the trait of interest. This total variation is then partitioned into, to adopt a simple taxonomy, a component representing the variation of species within their genera, the variation of genera within their families, families within orders, and finally orders within the class. At each level the mean of the values from the level immediately below is used.

If both sides of equation (5.1) are divided by σ_{tot}^2 and multiplied by 100, then the left hand side must be 100, and the terms on the right hand side become the percentages of variance found at each taxonomic level:

$$(\sigma^2_{tot}/\sigma^2_{tot}) \times 100 = [(\sigma^2_{s(g)} + \sigma^2_{g(f)} + \sigma^2_{f(o)} + \sigma^2_{o(c)})/\sigma^2_{tot}] \times 100. \quad (5.2)$$

These percentages of variance then can be compared among variables with different total variances. Another way to use the terms is to express them as cumulative proportions of variance moving from the highest level to the lowest level. Thus, if species is the lowest level of analysis then, by the time the $\sigma^2_{s(g)}$ term is added, 100 per cent of the variance will be accounted for. These cumulative percentages of variance have a precise statistical interpretation as *intra-class correlations*. The interpretation of an intra-class correlation is the correlation expected between any two data points selected at random from the same group. For example consider that species are the lowest taxonomic level represented and 75 per cent of the

variance is accounted for by the combination of orders within the class and families within the orders. Then, consider sampling repeated pairs of species where the first member of each pair is chosen at random from the data set and the second is another randomly chosen species from the same family as the first. The correlation among the pairs will be 0.75. The intra-class correlation coefficient might be viewed as a measure of the power of one species for predicting the value of another species in the same family.

For many continuous variables, nested ANOVA reveals that most of the variation in the trait occurs among orders nested within the class, and among families within orders (although, in principle, there is no constraint on how the variation is partitioned). Body weight in mammals provides a good example. Different orders tend to vary a great deal—compare Proboscidea (elephants) with Chiroptera (bats)—whereas species within genera tend to have relatively similar body weights. Table 5.1 displays the results of nested analyses of variance on several size and life history variables in mammals.

Table 5.1 Taxonomic distribution of life history variance among placental mammals. Tabulated values are percentages of total variance accounted for at successive taxonomic levels estimated from a nested ANOVA on logarithmically transformed species averages. (After Read and Harvey 1989).

Among: / Within:	species / genera	genera / families	families / orders	orders / class
Variance component:*	$\sigma^2_{s(g)}$	$\sigma^2_{g(f)}$	$\sigma^2_{f(o)}$	$\sigma^2_{o(c)}$
Adult weight	3	7	21	69
Neonatal weight	3	5	27	65
Gestation length	2	6	21	71
Age at weaning	8	11	19	62
Maximum reproductive life	10	10	12	68
Annual fecundity	5	7	14	74
Annual biomass production	6	8	18	68

*Following eqn (5.2), each variance component in the table has been multiplied by 100 and divided by σ^2_{tot}.

Harvey and Mace (1982) suggested that patterns similar to those in Table 5.1 might be interpreted to indicate that species within genera, and genera within families provide additional but probably not independent

data points in an analysis. That is, any one species' value will tend to be a good predictor of the other species in that genus, and the mean for a genus will tend to predict the other genera in a family relatively well. An alternative interpretation of cumulative percentages of variance (obtainable by adding the successive components of Table 5.1) as intra-class correlations supports this interpretation. Two randomly chosen individuals from the same family will typically correlate highly. Moving down a level to genera does not increase the correlation much: genera do not add substantially new information.

The nested ANOVA, then, provides a suggestion of the taxonomic level that should be used for analysis. Families or orders would be chosen as the units of analysis for the variables in Table 5.1, because around the family level there is a precipitous decline in the percentage of variance that a taxonomic level accounts for (statistical tests are available: Sokal and Rohlf 1981). However, any taxonomic level could, in principle, be chosen depending upon the distribution of variance in the characters being studied. Choosing a higher node greatly reduces the number of degrees of freedom in analyses. This accords with the belief that the additional 'degrees of freedom' obtained from lower taxonomic levels are not really very free at all. Thus, degrees of freedom and some variability are given up in return for what is hoped are increasingly independent data points.

The higher nodes method represents an early attempt to give a statistical justification for not treating species (or genera or even some higher taxa) as independent points. In addition, it has often been used to suggest a taxonomic level at which independence can be more or less assumed[12].

Harvey and Zammuto (1985) used the higher nodes method to investigate life history variation in mammals. These authors were interested in the idea that mortality patterns should be strongly correlated with the age at which individuals reach maturity, independently of adult body weight. Others had argued that this need not be so (e.g. Western and Ssemakula 1982). Across species, we may test the prediction that variation in mortality rates should correlate with the age at which different species reach maturity when body size is held constant: high mortality rates must be associated with early ages at maturity lest individuals die before successfully reproducing.

Harvey and Zammuto used Millar and Zammuto's (1983) data on age at maturity and life expectancy in 29 mammal species. Life expectation at

[12] The method has been used to investigate variation in life history, morphology, metabolism, sleep, and other behaviours including patterns of habitat utilization, particularly in birds and mammals, (e.g. Harvey and Clutton-Brock 1985; Gittleman 1986; Harvey *et al.* 1987; Elgar and Harvey 1987; Elgar *et al.* 1988; Bennett and Harvey 1985*a*, *b*, 1987; Read and Harvey 1989; Sherry *et al.* 1989; Promislow and Harvey 1990).

birth measured from natural populations of mammals living in approximately constant age-structured populations was taken to be an inverse measure of adult mortality rate. Harvey and Zammuto chose to analyse genera means instead of species because of concerns that the individual species did not represent independent points. When all variables were logarithmically transformed, life expectancy and age at maturity were both positively correlated with body weight (r = 0.87, 0.89, respectively, n = 25, P <0.001). However, the correlation between the two life history variables remained significant after controlling for the effects of body weight (r = 0.89, n = 25, P <0.001). This correlation may arise, however, because life expectancy is partly a function of age at maturity (demographic reality dictates that some individuals must survive to breed or the species would be extinct). So, Sutherland, *et al.* (1986) conducted the same test, this time using life expectancy *from age at maturity* instead of life expectancy at birth. The correlation controlling for body size remained significant and the two body-size-corrected measures are plotted against each other in Fig. 5.6.

Another example of a higher nodes approach comes from an analysis of the hippocamapal complex in birds reported by Sherry *et al.* (1989). Some bird species collect and store large numbers of food items, each in a separate place, before retrieving them at some later date. In contrast, many other bird species immediately consume the food that they gather.

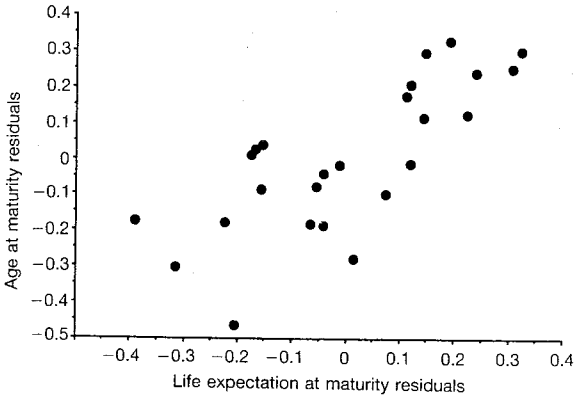


Fig. 5.6. The deviations from the logarithmic regression of age at maturity on body weight plotted against the deviations from the regression of life expectation at maturity on body weight. A positive deviation indicates that the life history variable was larger than would be expected on the basis of body weight, a negative deviation indicates the opposite. The two size corrected measures are highly correlated (r= 0.76, P<0.001). The data points are generic averages of constituent species values.

Sherry and his colleagues were interested in the idea that storing and retrieving food places demands on the memory of food-storing species not experienced by species that do not store their food. They suggested that food-storing birds might have enlarged hippocampuses as a result, because the hippocampal complex is thought to be involved in spatial memory.

Food-storing versus non-food-storing does not vary independently among species belonging to different families of birds. For example, many North American chickadees (Paridae), nuthatches (Sittidae), and jays and crows (Corvidae) store food while species in other passerine families and subfamilies do not. Sherry *et al.* (1989) measured the size of the hippocampal complex for 23 species from 13 passerine subfamilies, and performed their analysis across subfamily averages. They justified this by the fact that ten of the subfamilies in the data set were represented by only non-food-storing species while the other three subfamilies were represented by only food-storing species. The food-storing subfamilies had large hippocampal complex sizes relative to both their body weights and to the volumes of their telencephalons, which is the part of the brain within which the hippocampus complex is situated. Figure 5.7(a) plots body-size-corrected hippocampus complex volumes for each subfamily of birds in Sherry *et al.*'s data set.

Krebs *et al.* (1989) were able to use an interestingly expanded data set to tackle the same problem. Two of the food-storing families, the Paridae and the Corvidae, contain some species that do not store food. Krebs *et al.* measured the hippocampus size of species that store food and those that do not store food within each family. Do the non-food-storing members of these families have relatively smaller hippocampuses than their food storing relatives? They do, as can be seen in Fig. 5.7(b). Furthermore, the Troglodytidae which are non-food-storers have relatively smaller hippocampuses than the closely related food-storing Sittidae. Krebs *et al.*'s analysis goes beyond comparing higher taxon means by also examining variation within higher taxa.

What assumptions about evolution and the phylogeny of species are embedded in the higher nodes method? The method treats the higher nodes as independent points in analyses. Technically, this means that the higher nodes must have a phylogeny that forms a simultaneous radiation pattern like that in Fig. 5.1. Thus, for example, if families were the unit of analysis, the phylogeny must be such that all of the families simultaneously radiate from a single ancestor common to the entire class. All branches leading to families must be the same length to ensure equality of expected variances. Branch lengths can be ignored if the amount of evolution is believed to have been independent of time and yet equal in each branch (e.g. punctuational change). The branches leading to orders must have a length of zero: any order branch with a non-zero length would possibly

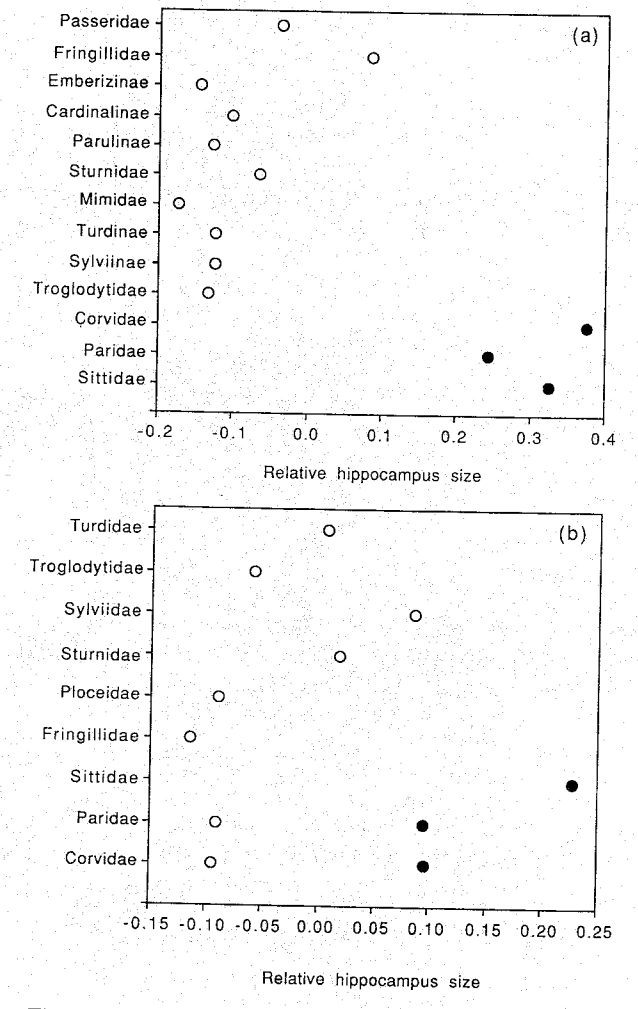


Fig. 5.7. The higher nodes approach illustrated by hippocampal complex size in passerine birds. (a) When body size effects are controlled for (relative hippocampus size is the deviation from the cross-subfamily regression of hippocampus volume on body weight), food-storing subfamilies of birds (●) have relatively larger hippocampal complexes than their non-food-storing relatives (○). (Data from Sherry *et al.* 1989). (b) When body size effects are controlled for, non-food-storing parids and corvids have relatively smaller hippocampuses than their food-storing relatives. Furthermore, the food-storing Sittidae have relatively larger hippocampuses than the closely related non-food-storing Troglodytidae. (Data from Krebs *et al.* 1989).

introduce phylogenetic similarity among its families (there are other models in which the families could be independent but this is the simplest and most general). Further assumptions are built into the family means themselves. By finding genera means first, and then family means as the average of their respective genera, all lower taxa are assumed to be simultaneous radiations with equal branch lengths.

It would be very unusual for any phylogeny to have the form required for the higher nodes method to be correct technically. In practice, however, this does not mean that conclusions drawn from all higher nodes analyses are incorrect. The most obvious worry in using a higher nodes approach is that the higher nodes may not be independent for the same sorts of reasons that species are not independent. One should, for example, be careful to examine the data to see if the result depends upon the contribution of any one cluster of points that share an immediate ancestor. The relationship can also be examined separately within taxonomic groups. For example, Krebs *et al.* (1989) were able to show that in all three cases where pairs of closely related taxonomic groups could be compared, the hippocampus volume was greater in the food-storing group. However, even if all of the assumptions of the higher-nodes method were met, it would still have the unavoidable limitation that information from lower taxonomic levels is lost, along with degrees of freedom for statistical tests.

5.5 Stearns' phylogenetic-subtraction method

A method for directly subtracting from the species' data the variation associated with phylogenetic similarity (or taxonomic similarity where taxonomy is used to stand in for phylogeny) was developed by Stearns (1983). Stearns' method, although motivated by the same concerns as the higher nodes method, proceeds in a manner opposite to it.

Stearns proceeds on the assumption that the portion of the total variation associated with differences among higher nodes represents lineage-specific variation that is not appropriate for testing questions about adaptation at the species level. Stearns statistically removes from the data the lineage-specific variation associated with higher nodes, and analyses the remaining variation. Stearns' (1983) analysis of mammalian life histories provides an example. Differences among orders and among families within orders were assumed to represent effects of phylogenetic similarity present in the species data points. Stearns simply subtracted from each species data point the mean value for its order. Species' values were then free of similarity associated with differences among orders. This procedure was repeated using family means, thus leaving a set of residual data points for species that were free of differences associated with families or orders. Stearns found that substantial covariation existed among

mammalian life history variables even after removing these taxonomic associations. Wooton (1987) employed a similar approach in his study of age at first reproduction in mammals.

Stearns' nested analysis of variance method for controlling phylogenetic effects is statistically equivalent to removing the variance using categorical codes in a multiple regression. Box 5.1 gives an example of the categorical coding required to remove order, family, and generic differences among sixteen species in a hypothetical symmetrically bifurcating phylogeny such as that in Fig. 5.3. If the *Y* variable (and any other variable to be analysed) is regressed onto the categorical codes, and residuals are found, these residuals will be statistically independent of their phylogeny. This has been done for a hypothetical data set in Box 5.1.

Box 5.1. Removing phylogenetic correlations from a continuously-varying character using multiple regression

S	O	F	G	O1	F1	F2	G1	G2	G3	G4	*Y*	Residual
1	1	1	1	1	1	0	1	0	0	0	29	2.0
2	1	1	1	1	1	0	1	0	0	0	25	-2.0
3	1	1	2	1	1	0	0	0	0	0	26	-1.5
4	1	1	2	1	1	0	0	0	0	0	29	1.5
5	1	2	3	1	0	0	0	1	0	0	24	3.5
6	1	2	3	1	0	0	0	1	0	0	17	-3.5
7	1	2	4	1	0	0	0	0	0	0	24	-2.5
8	1	2	4	1	0	0	0	0	0	0	28	2.5
9	2	3	5	0	0	1	0	0	1	0	15	1.0
10	2	3	5	0	0	1	0	0	1	0	13	-1.0
11	2	3	6	0	0	1	0	0	0	0	14	-2.5
12	2	3	6	0	0	1	0	0	0	0	19	2.5
13	2	4	7	0	0	0	0	0	0	1	12	4.5
14	2	4	7	0	0	0	0	0	0	1	3	-4.5
15	2	4	8	0	0	0	0	0	0	0	14	3.0
16	2	4	8	0	0	0	0	0	0	0	8	-3.0

Sixteen species are classified by phylogenetic relatedness into order (O), family (F) and, genus (G). The taxa are then given categorical codings to produce dummy variables, one for order membership (O1), two for family membership (F1, F2), and four for genus membership (G1, G2, G3, G4). Each categorical code is used as an independent variable in a hierarchical multiple regression (Draper and Smith 1981) to remove the taxonomic correlates of a continuously varying character (*Y*), resulting in a set of residuals that are not correlated with taxonomy. Thus, code O removes the differences between the means of the two orders, code F1 removes differences among the families 1 and 2 nested in order 1, and so on down to code G4 which removes differences between genera 7 and 8 nested in family 4.

The correlation between the residuals and any one or combination of the discrete phylogenetic dummy variables in Box 5.1 will always be exactly zero. Stearns (1983) removed taxonomic similarity due to orders and families. However, as Box 5.1 shows, it is possible to remove the effects of taxonomy down to the genus level, or more generally, down to one level above the lowest level represented in the data set. This procedure for removing variation associated with phylogeny can be applied to any number of variables, and then the relationships among them tested.

Returning to the conceptual model outlined at the beginning of this chapter, what phylogenetic structure is implicit in Stearns' approach? Stearns' approach assumes that, below some level (Stearns chose families), the phylogenetic or taxonomic groups are independent. The simultaneous radiation of Fig. 5.1 produces independence, but imagination is required to see how this model can be applied to groups that cannot be thought of as sharing an immediate common ancestor. The effect of removing variation associated with phylogeny is to make the mean value of the trait equal to zero within the lowest level groups controlled for in the analysis. Thus, in the example in Box 5.1, the mean value of the residuals within genera is exactly zero. In this statistical sense, then, the species can be thought of as having a common phenotypic starting point of zero. In Stearns' life history study, the family level was the lowest level controlled for, and so the implicit phylogeny must have all species in all genera radiating from a common starting point.

We have applied Stearns' approach to the Millar and Zammuto (1983) data set to illustrate the difference between this approach and the higher-nodes method. The data set contains six different orders and 18 families. Harvey and Zammuto (1985) and Sutherland *et al.* (1986) conducted their analyses of life expectancy, age at maturity, and body size across genera means ($n = 25$). We used multiple regression to remove the variation in these three variables that is associated with order and family differences among the 29 species. Then we repeated Harvey and Zammuto's analyses on the species data with order and family effects removed.

The dummy coding was done in a manner analogous to that in Box 5.1. Five dummy codes were required to control for the six orders. Only six additional codes were required to control for differences among families within orders because of the way that families were distributed among the orders. With additional codes for genera within families and species within genera, it is possible to reconstruct the nested analysis of variance table for this data set. The species are taxonomically diverse in this data set, so many families are represented by just one or two species. Because of this, the phylogenetic differences among orders and families account for a very large percentage of the total variation among the species.

Table 5.2 Taxonomic distribution of variance among the characters used for the analysis of life history variation and mortality among placental mammals. Tabulated values are percentage of total variance accounted for at successive taxonomic levels estimated from a nested ANOVA on logarithmically transformed species averages. (Data from Millar and Zammuto 1983).

Among: / Within:	species / genera	genera / families	families / orders	orders / class
Variance component:*	$\sigma^2_{s(g)}$	$\sigma^2_{g(f)}$	$\sigma^2_{f(o)}$	$\sigma^2_{o(c)}$
Body weight	1	2	12	85
Age at maturity	1	3	36	60
Life expectation at maturity	2	2	34	62

*Following eqn (5.2), each variance component in the table has been multiplied by 100 and divided by σ^2_{tot}

By using just the dummy codes for orders and families, the 29 species data points are statistically independent of phylogenetic variation associated with those levels. We then further controlled both life history variables for body size, and plotted the residuals against each other. Fig. 5.8 plots this relationship which although positive is not as strong as that in Fig. 5.6. Controlling for phylogenetic relationships and for body size, there is no longer a significant relationship between age at maturity and life expectation at maturity ($r = 0.35$, d.f. = 16, $P > 0.15$; note that the degrees of freedom are equal to the sample size minus the number of control variables minus 1: 29–12–1).

An investigator using Stearns' method on the Millar and Zammuto data set would have reached a different conclusion from that which Harvey and Zammuto (1985) and Sutherland *et al.* (1986) reached using a higher nodes approach. Which would be correct? Both would be correct for certain kinds of conclusions (Pagel and Harvey 1988*a*). The conclusion that there is not significant covariation between the life history variables after controlling for differences associated with phylogeny is correct for this data set. So is the conclusion that there is significant variation across genera. The interesting debate concerns what we should make of the difference.

Variation among higher nodes is removed by Stearns' method and variation at lower levels is retained. The higher-nodes method analyses the variation at higher levels and averages over the lower levels. Each method uses the information that the other method discards! Stearns assumed that

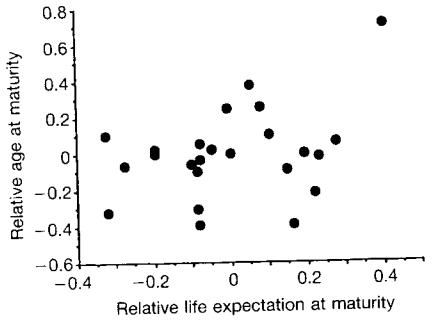


Fig. 5.8. The relationship between age at maturity and life expectation at maturity for mammals after controlling for body weight and phylogenetically correlated variation in each variable. The correlation is not significant [r=0.35, d.f.=16 (see text), P>0.15]. 29 species are represented on the figure, but 6 data points overlap at 0,0.

variation at higher levels should be attributed to lineage-specific trends and thus was inappropriate for testing questions about function. His point is that the adaptive variation is that which is independent of differences associated with phylogeny. The investigator using Stearns' method should explain why variation among higher taxonomic groups is thought to be inappropriate for testing this functional question. The higher-nodes investigator must provide an explanation of why it is felt that the differences among the higher taxonomic levels are not confounded by other taxonomic differences.

Consider, for example, that all of the rodents have early ages at maturity and short life expectancies from maturity not because of a causal connection between the two, but for some other reason associated with being a rodent. This would represent a taxonomic confound that the Stearns' method would remove. Thus, the higher nodes investigator using the order as the level of analysis should demonstrate that the relationship is found within orders, like the rodents, as well as across them.

Both sides of this debate have something to offer. However, we shall show in a later section how it is possible to get around the conflict between these two methods by using techniques that analyse the covariation between traits within taxonomic (or phylogenetic) groups. There is no need to throw out the variance at either the lower or the higher levels. All of the variation in the data can be used to test ideas about the correlation between traits.

5.6 Phylogenetic autocorrelation method

Cheverud *et al.* (1985) describe a method for partitioning phenotypic traits into phylogenetic and specific components that is conceptually similar to Stearns' method. Cheverud *et al.* predict a species' phenotype on the basis of the phenotypes of other species in the sample. More closely related species, such as those in the same genus, will typically be better predictors than more distantly related ones. Accordingly, the phenotype of a focal species that is closely related to many other species in the sample will be better predicted than one that does not have many close relatives. This fits with our intuitive feeling that closely related species do not each represent an independent instance of the evolution of their trait. However, species' phenotypes will not be perfectly predicted, even when they share many close relatives. Cheverud *et al.* (1985) use this *specific* portion of the phenotype to test for correlated relationships among variables. Their method, then, although conceptually similar to Stearns' (1983), differs by employing an explicit evolutionary model to estimate variation due to phylogenetic effects.

Cheverud *et al.*'s phylogenetic autocorrelation method uses a linear autocorrelation model to partition the total variance in a trait that is measured across species into the sum of phylogenetic and specific variances, plus the covariance between the phylogenetic and specific values of the trait. The model represents the trait y as a linear combination of phylogenetic and specific effects according to:

$$y = \rho W y + e. \tag{5.3}$$

where y is the vector of length n containing the n species' data points, ρ is a scalar 'phylogenetic autocorrelation coefficient', W is an n × n 'phylogenetic connectivity matrix', ρWy is a vector of predicted y values representing the phylogenetic portion of y, and e represents the vector of residual values of the trait that cannot be predicted by the vector ρWy. It is e that is used to analyse whether there is covariation between traits that is independent of phylogenetic effects. Cheverud *et al.*'s (1985) model assigns both parallel evolution and variance due to the interaction of the phylogenetic and specific components solely to the phylogenetic effect.

The matrix W is used to account for the phylogenetic relatedness of species. The phylogenetic portion of the attribute y, given by the vector ρWy, is just a weighted sum of the phenotypic trait values of each of the other species in the data set, scaled by the factor ρ. The scalar quantity ρ is roughly a measure of the correlation between the observed and predicted values of y, where the predicted values are equal to ρWy. Thus, if the actual values of y are largely predictable from phylogenetic relatedness, then ρ will be high.

The weights in W are assumed to be big for closely related species, and to decline for more distantly related species. In a worked example, Cheverud *et al.* (1985) arbitrarily assigned weights of 1.0 for congeners, 1/2 for species in the same family, 1/3 for the same superfamily, and 1/4, 1/5, and 1/6 for the same infra-order, suborder, and order, respectively. Thus, a trait value for a species that is closely related to a large number of other species in a data set is attributed primarily to phylogeny. The authors discuss several ways of estimating the relatedness weights; Gittleman and Kot (1991) report a method that allows an assessment of the weighting according to the variance in the data.

Cheverud *et al.*'s (1985), method can be illustrated for the phylogeny of Fig. 5.9.

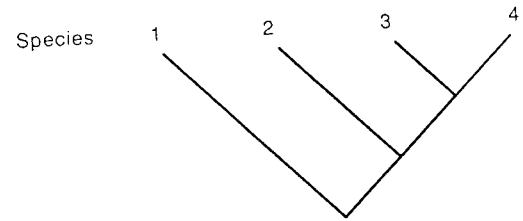


Fig. 5.9. A phylogeny used to illustrate Cheverud *et al.*'s (1985) phylogenetic autocorrelation method. Species 1, 2, 3, and 4 possess the character values 10, 8, 3, and 4 respectively.

We assume that the four species comprise a single order. Species 3 and 4 are congeners, species 1 is in a different family from 2, 3 and 4. A hypothetical reconstruction of the terms in eqn (5.3) is given below. This example is for illustrative purposes only and should not be taken to indicate the actual values that Cheverud *et al.*'s procedure would produce.

$$\begin{matrix} y \\ \begin{bmatrix} 10 \\ 8 \\ 3 \\ 4 \end{bmatrix} \end{matrix} = \begin{matrix} \rho \\ [0.49] \end{matrix} \begin{matrix} 1 \quad 2 \quad 3 \quad 4 \\ \begin{bmatrix} 1 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1 & 1/2 & 1/2 \\ 1/6 & 1/2 & 1 & 1 \\ 1/6 & 1/2 & 1 & 1 \end{bmatrix} \end{matrix} \begin{matrix} y \\ \begin{bmatrix} 10 \\ 8 \\ 3 \\ 4 \end{bmatrix} \end{matrix} + \begin{matrix} e \\ \begin{bmatrix} 3.88 \\ 1.55 \\ -3.21 \\ -2.21 \end{bmatrix} \end{matrix}$$

and thus

$$\begin{matrix} y & = & \rho\, Wy & + & e \end{matrix}$$

$$\begin{bmatrix} 10 \\ 8 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 6.12 \\ 6.45 \\ 6.21 \\ 6.21 \end{bmatrix} + \begin{bmatrix} 3.88 \\ 1.55 \\ -3.21 \\ -2.21 \end{bmatrix}$$

This example shows how the matrix product ρWy produces a predicted y score for each species, and the vector e is just the residual difference between y and the predicted y. Species that are closely related to each other have a greater amount of their trait attributed to phylogeny. Species 3 and 4 have the same predicted value of y because they are congeners. Our illustration uses an arbitrary value for ρ that makes the predicted values have the same mean as the actual values. In practice, Cheverud *et al.* use a maximum likelihood procedure for estimating ρ.

Cheverud *et al.* (1985) used their model to study body size dimorphism in primates. Four variables, body size, mating system, habitat, and diet, were studied for their association with the extent of sexual dimorphism in body size for 44 primate species. The phylogenetic autocorrelation procedure was used to define phylogenetic (ρWy) and specific (e) components for each trait. Then the eight components plus the phylogenetic portion of size dimorphism were used to predict size dimorphism. Table 5.3 shows the unique contribution of each variable in terms of the percentage of the variance in size dimorphism that it predicted.

Table 5.3 Proportions of total variance in sexual body size dimorphism explained by phylogenetic and specific sources of variance (after Cheverud *et al.* 1985). *Note*: inclusion of the specific proportion of sexual dimorphism would (arbitrarily) have led to 100 per cent of the variance being accounted for.

	Percentage of variance accounted for by sources of variance	
Trait	Phylogenetic	Specific
Mating system	<1	<1
Body weight	28	34
Sexual size dimorphism	1	–
Habitat	2	4
Diet	<1	2
Interaction among traits	19	2
Total	50	42

Fifty per cent of the variance in size dimorphism can be accounted for by the phylogenetic components, and an additional 42 per cent by the specific components.

As with Stearns' procedure, potentially the majority of the variation is assigned to phylogenetic components and is treated as non-adaptive. Investigators should be aware of this and provide an explanation of why they feel that variation among higher taxonomic groups should not be used to test questions about function and adaptation.

5.7 A maximum likelihood approach

Lynch (in prep.) reports a method that, like Cheverud *et al.*'s (1985) method relies on a statistical model to partition each species' phenotypic trait value into components associated with and components independent of phylogeny. Unlike Cheverud *et al.*, however, it is the phylogenetic component that Lynch uses to test the comparative relation.

Lynch borrows ideas from quantitative genetics to partition species' phenotypic values into three components, two representing phylogenetic variation, and one representing variation that is independent of phylogeny. Each species' phenotypic mean is seen as a combination of an overall phylogenetic effect, a component representing the 'heritable additive evolutionary value of the character', and a residual component. The overall phylogenetic effect is analogous to a grand mean on the trait. The interpretation of the additive heritable component is that it represents something akin to a breeding value: a species' additive effect is the expected phenotype of a descendant of that species. The sum of the first two components is the overall 'heritable component of a particular realization of the evolutionary process'. The residual effect represents non-additivity of genetic effects, environmental effects, and sampling error.

Knowledge of the across-species variance-covariance matrices of the additive effects and of the residual errors is required to estimate the additive effects and the overall effects. The matrix of additive effects, in turn, depends on a matrix that measures the true phylogenetic relationships among taxa. For example, all species are perfectly related to themselves, less so to sister taxa, and so on, in a fashion similar to Cheverud *et al.*'s (1985) matrix W. The phylogenetic relationships matrix is estimated from the phylogeny. The relatedness of two species is taken as the proportion of their total path lengths that they share. The additive effects and the overall effects are estimated by a recursive maximum-likelihood procedure. Using initial arbitrary values of the additive effects and residual errors, the variance-covariance matrices can be found. This, then, leads to new estimates of the additive and residual effects, and so on until a stable solution is reached. Lynch reports that the maximum-likelihood algorithm usually converges, or at least leads to a region of

results. Statistical convergence may also be stymied by multiple peaks in the likelihood surface.

Lynch's method, by using explicit statistical criteria to take into account the non-independence of taxa, has much to recommend it. Like Cheverud *et al.*'s (1985) technique it avoids altogether the problem of reconstructing ancestral character states, instead conditioning all tests on the variation among extant species. It is difficult to judge at this point how well Lynch's approach will work, and how it will manage with poorly known phylogenies. It is critical to get the estimates of the variance-covariance matrices correct in order to adjust the species additive effects for their strong phylogenetically based lack of independence. It may also be unnecessary, as we shall suggest below, to partition the species' phenotypes as Stearns (1983), Cheverud *et al.* (1985), and Lynch do. Nevertheless, maximum-likelihood methods such as Lynch's deserve more attention.

5.8 Independent comparisons methods

All of the methods reviewed so far, with the exception of a species regression, make a distinction between variation associated with phylogeny, and variation that is independent of phylogeny. The methods to be described in this and the following sections use all of the variation in a trait to test for a comparative relation, and they do so without partitioning the traits into phylogenetic and non-phylogenetic components. Independent-comparisons methods are able to make use of all of the data by recognizing that what is phylogenetic inheritance at one level of a hierarchy may constitute part of an adaptive difference at the next highest level.

This discussion is based on logic outlined by Felsenstein (1985*a*). Figure 5.3 shows a branching phylogeny for eight species. Focussing on a portion of this phylogeny, the range of values including the two species that split from node *n*4 and the two species that split from node *n*5 is largely the result of a phylogenetic difference that evolved once between *n*4 and *n*5. That is, most of the variation among these four species in a typical phylogeny would have already been present between the two higher nodes. However, there are three degrees of freedom among these four species: the difference between species 1 and 2, the difference between species 3 and 4, and the difference between nodes *n*4 and *n*5.

Assume that changes along the branches of the phylogeny can be modelled by a Brownian motion process such that, as above, successive changes are independent of one another, and that the expected total change summed over many independent changes is zero. Then, the three pairwise differences (species 1 versus 2, species 3 versus 4, and *n*4 versus *n*5) are independent of each other. This is because, for example, the difference between species 1 and 2 reflects only the evolutionary changes

that have taken place since they split from their common ancestor (*n*4). All similarity between species 1 and 2 that is due to their shared phylogenetic history will be, in effect, subtracted out. The same logic applies to species 3 and 4. Their difference in turn will be independent of the differences between species 1 and 2. Finally, the difference between nodes *n*4 and *n*5 reflects only the evolutionary events that have happened since they split from their common ancestor, and this difference will be independent of the other two.

The three comparisons together account for all of the variation among the four species by dividing the variation into three separate evolutionary events. Each event reflects the difference between the evolutionary changes in two branches of the tree. So, the advantage of independent-comparisons approaches is that, by partitioning the variation appropriately, it can all be used to assess the comparative relationship.

More generally, in a given branching phylogeny we might calculate the difference in Y and difference in X between the species within each of the lowest level clades, then again at the next highest level, and so on until we compare the two highest nodes of the tree. The important point is that each of these relationships represents under the null hypothesis an independent instance of the evolution of the relationship between Y and X. Thus, any covariation between Y and X that is present among the species sharing a common ancestor is phylogenetically independent of the covariation between Y and X among the species sharing a different common ancestor. The same argument applies to each similarly defined pair of higher nodes. The set of differences between Y and X provides a way to test whether changes in Y and X are correlated. Under the null hypothesis that evolutionary changes in Y and X are unrelated, a positive difference on X should be associated with a positive difference on Y no more often than with a negative difference. A preponderance of positive (or negative) relationships within taxa, then, is evidence against the null hypothesis.

Box 5.2 provides a simplified summary of the procedure used to produce and compare independent comparisons. Note, however that path lengths are ignored in the example and that higher nodes are calculated as the average value of lower nodes. We shall discuss both of these issues later.

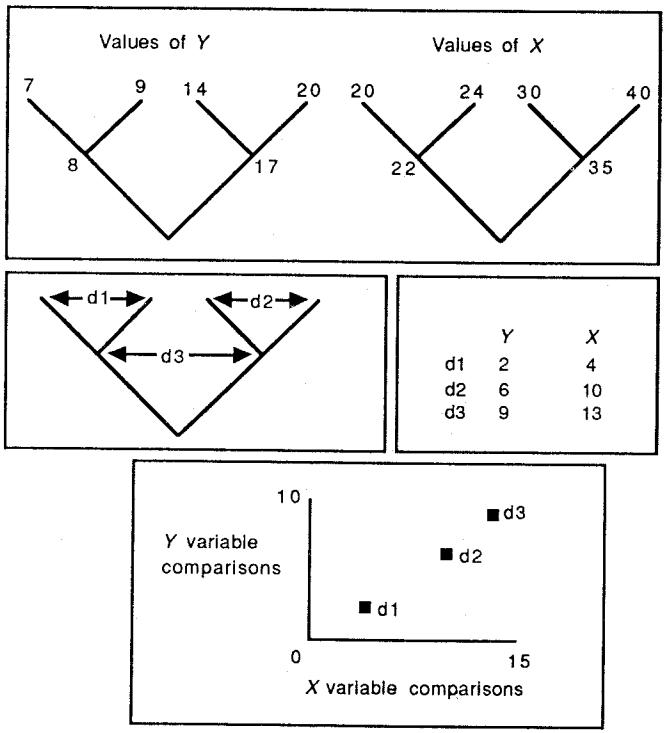
Later in this section we shall discuss three methods that use independent comparisons. The methods all rely on the same independent-comparisons logic but differ in assumptions and statistical manipulation of the data. Before discussing the three independent-comparisons procedures we make a brief digression to describe the nested analysis of covariance.

5.8.1 A brief digression: nested analysis of covariance

The nested analysis of covariance does not make use of the logic of independent comparisons, but it does exploit the fact that species naturally form a nested hierarchy. Following brief sorties by Dunham and Miles

Box 5.2. The independent comparisons method for two characters in a single phylogeny.

Under a Brownian motion model of evolution, d1, d2, and d3 provide independent comparisons. Path length differences are ignored in this illustration.



(1985) and by Martin and Harvey (1985), Bell (1989) suggests a novel use of nested analysis of covariance to analyse comparative data.

A nested analysis of covariance analyses the covariation, or equivalently, the correlation between two or more variables separately within each of the groups in a nested hierarchy. Applied to comparative data, the method finds a relationship between two or more variables separately within each taxonomically or phylogenetically defined group in a data set.

The separate within-group relationships can then be combined to form a pooled estimator of the within-group correlation between two variables. A pooled estimator of the correlation between two variables is analogous to a pooled estimator of the within-groups variance in an ordinary analysis of variance, except information about the covariation between two variables rather than information about the variation of one variable is used.

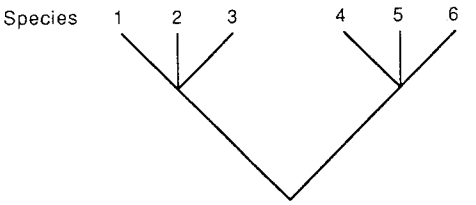


Fig. 5.10. A phylogenetic tree used to illustrate the nested analysis of covariance method.

Consider the hypothetical tree of Fig. 5.10 showing three species in each of two genera. Ignoring the genera classifications for a moment, let σ_{Pxy} represent the phenotypic covariance between Y and X across the six species. The covariance is defined as the correlation between X and Y multiplied by the standard deviations of X and Y:

$$\sigma_{Pxy} = \rho_{Pxy}\sigma_{Px}\sigma_{Py}. \tag{5.4}$$

Now, classifying the species into their genera, it is possible to partition the phenotypic covariance into two components, one representing the covariation between Y and X within genera, and one representing their covariation across genera:

$$\sigma_{Pxy} = \sigma_{Wxy} + \sigma_{Axy} \tag{5.5}$$

where the subscripts 'W' and 'A' refer respectively to 'within' and 'among'. The covariance within genera is the sum of the separate covariances in each of the two genera. The individual within-genus covariances are not influenced by phylogenetic differences among the species if the genus is a monophyletic group. The covariance among genera is calculated across the two genera means. This covariance expresses a phylogenetic difference between the two groups of species. Thus the overall covariance among the species is a combination of phylogenetic and

non-phylogenetic or evolutionary components. The within-groups covariance represents an evolutionary relationship between Y and X that is not influenced by differences in phylogeny among the species. This is true because, if Fig. 5.10 is an accurate representation of the true phylogeny, species 1 through 3 share their phylogenetic history as do species 4 through 6. This method can be applied to all levels of a phylogeny. For example, a pooled estimator of the covariance between Y and X calculated from genera means within families, has an analogous interpretation to the covariance among species within genera. Similar partitionings can be made at higher levels.

Bell (1989) applied the nested analysis of covariance method to study the relationship between litter mass and gestation length in mammals. Both variables are highly correlated with body weight in mammals and so, as a first step, Bell statistically removed the association of both variables with body weight. All analyses were then conducted on relative litter mass and relative gestation length, where the relative values are defined as the residuals from the respective regression lines for the two variables on adult body weight. Bell compiled 574 observations of litter mass and gestation length on 370 species, then conducted a nested analysis of covariance on the two body-weight-corrected measures. The results are given in Table 5.4 where, for ease of interpretation, Bell has converted each of the covariances to a correlation by dividing by the product of the two standard deviations. The correlations in Table 5.4 are found by pooling all of the individual within-group correlations at a given taxonomic level.

Table 5.4 Nested analysis of covariance results for relative litter mass and relative gestation length in 370 mammal species. The correlations are estimated for the taxonomic level indicated and are nested within the next highest level. For example, the within-species correlation is that for individuals within species. NE = not estimated.

Taxonomic level	Degrees of freedom	Estimated within-group correlation
Superorder	4	NE
Order	12	NE
Suborder	6	0.69
Family	52	0.52
Subfamily	47	−0.01
Genus	120	0.21
Species	128	−0.10
Within species	204	0.33

Bell was able to calculate a pooled correlation among individuals within species because more than one observation was available for many of the species. The results show substantial correlations within species, and then again among families within suborders, and suborders within orders.

The main difference between Bell's approach and the methods reviewed previously in this chapter is that Bell's analysis makes use of all of the variation in the data set to assess the comparative relation. Variation among species within genera, genera within families, and so on to the highest nested level is all used to investigate whether two variables are correlated. This method is very close to what we have called 'independent comparisons methods' in that it avoids phylogenetic influences by looking separately within taxonomically or phylogenetically defined groups. Provided that all of the members of a group share an immediate common ancestor, there is no phylogenetic variation within the group, and the correlation between two variables within that group represents evolutionary change since they diverged from their common ancestor. If the groups are not monophyletic then some phylogenetic differences will be included in the within-groups covariance.

Combining the information from different taxonomic groups and levels requires assumptions about phylogenetic branch lengths (assuming a fixed gradual model of change). The branch lengths leading from each common ancestor to their respective descendant taxa must be the same length in each group. If they are not, then the expected variation within a taxonomic group with longer branch lengths will be larger than that within a taxon having short branch lengths. Bell combines information only from the same taxonomic level, thus, having only to assume that all taxa at the same level have equivalent branch lengths.

The pooled estimators combine information from different groups. This gives more weight to groups that have more subtaxa. For example, a genus with seven species will contribute a larger share to the pooled estimator than a genus with two species. Such a weighting would be appropriate if the true phylogeny within the genus was a real simultaneous radiation of the n subtaxa. Then, there would be n–1 independent pieces of information among them. But consider that the true but unknown phylogeny is not a simultaneous radiation. Then the n subtaxa will not be independent and treating them as such will spuriously give more weight to larger groups (in the limiting case of no phylogenetic classification whatsoever, the nested analysis of covariance method would be identical to a species regression, with all species nested within a single higher node). Ideally, we should like to know how to weight such groups where we have reason to believe that the n subtaxa are not independent, but at the same time the true phylogeny is unknown.

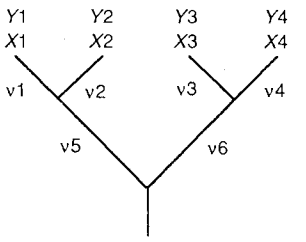
The methods that we discuss in the following sections address the

problem of weighting by allowing each taxonomic group to contribute only one piece of information that is not weighted by the number of subtaxa. We shall now discuss three methods that do this, each based on a different evolutionary model.

5.8.2 Felsenstein's method of pairwise independent comparisons

Felsenstein (1985*a*) was the first to develop a method for testing comparative relationships based on the logic of comparing pairs of species or higher nodes that share a common ancestor (Section 5.8). Felsenstein's method is based on a Brownian motion model (Section 5.2) of evolutionary change. Other models of change are possible, although most are intractable statistically (Felsenstein 1988). The method requires that the true branching phylogeny, including the lengths of the branches of the tree in units of expected variance of evolutionary change, is known. This information allows the calculation of a set of comparisons among pairs of data points, each of which has the same expected mean and variance under the null hypothesis.

Consider the phylogeny in Fig. 5.11. Following the logic developed above for independent comparisons methods, and drawing on Felsenstein's (1985*a*) article, the differences between the two pairs of species at the tips of the tree, represented by $(X1-X2)$, and $(X3-X4)$, will be independent of each other. By the same logic the difference between the higher nodes defined by $(X1+X2)/2$ and $(X3+X4)/2$ is another comparison, and it is independent of the first two. (Here we assume that $v1 = v2 = v3 = v4$, and that $v5 = v6$. If path lengths vary the comparisons change somewhat. Felsenstein's original paper gives formulae for the general case). Calculating the comparisons among the higher nodes this way ensures, given the Brownian motion model, that the comparisons are mutually independent



5.11. A phylogeny of four species. X_i and Y_i (i = 1–4) represent the states of phenotypic characters in species i. The v_j (j = 1–6) define the time in units of expected variance of evolutionary change spent evolving along each branch of the tree. (After Felsenstein 1985*a*).

statistically. Calculating the same three comparisons for the Y_i would yield three pairs of points that could be used to ask whether changes in the X variables go with changes in the Y variables. In general, with n species, we can find a set of $n-1$ mutually independent comparisons in a bifurcating tree.

Before discussing the use of the independent comparisons we should point out why such comparisons are used rather than those that measure the independent evolutionary change along each branch (which we referred to in Chapter 1 as 'directional' methods). We could use the latter if we had a method for reconstructing ancestral conditions that was independent of the descendant character states. But consider that, in Felsenstein's method for example, the higher nodes are roughly equal to the arithmetic means of the lower nodes. As the sum of the deviations of a set of scores around their mean must sum to zero, this means that we cannot use both scores for a pair of species. But we can use their difference.

The set of differences are independent under the null model, but they will not all have the same expected variance. Here is where the branch lengths and Felsenstein's evolutionary model are put to use. Felsenstein models the evolution of a character along its branch by a process of Brownian motion. If change is independent in each unit of time, then after one unit of time the character will have accumulated σ_x^2 units of variance, where σ_x^2 is the variance of the process (For ease of discussion, we shall assume that σ_x^2 is constant throughout the tree. But it need not be.). Then, over v units of time the variance will be $v\sigma_x^2$. This means that the various observations on X will have the same variance only if their branches are of the same length.

However, having knowledge of the variance makes it possible to scale each X score to have a mean of 0 and standard deviation of 1:

$$\frac{(X-0)}{\sqrt{v\sigma_x^2}} \qquad (5.6)$$

The same calculations can be performed for the Y variables also using the $v\sigma_y^2$ (however σ^2 for the X variable need not be equal to σ^2 for the Y variable). Then, each difference between a pair of species or higher nodes will also be a variate with a mean of 0 and a standard deviation of 1. If the evolution of the characters can be described by Brownian motion, then the set of comparisons on X and Y can be regarded as having been drawn from a bivariate normal distribution with means of 0, standard deviations of 1, and unknown correlation parameter, ρ. The null hypothesis is that ρ equals 0.

Sessions and Larson (1987) used Felsenstein's method to test whether genome size in plethodontid salamanders is related to developmental rate. The 'junk DNA' hypothesis predicts that junk DNA will accumulate in the genome until the costs of transcribing it impose too great a cost on the organism. This leads to the prediction that genome size (as measured by the *C*-value, the weight of the genome in picograms) should be inversely related to measures of the developmental rate of a species.

Sessions and Larson identified 18 independent pairwise differences or contrasts in the family Plethodontidae (Fig. 5.12). Higher nodes were reconstructed according to a parsimony procedure (Farris 1970; see Chapter 3). Difference scores were calculated for each of the 18 pairwise comparisons for three variables: *C*-value, and two measures of developmental rate (limb differentiation rate and limb growth rate). Estimates of the branch lengths in units of time were obtained from molecular data.

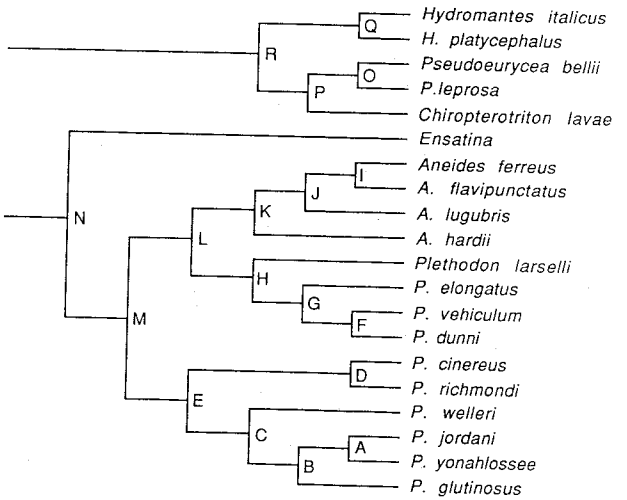


Fig. 5.12. The phylogenetic tree of the Plethodontid salamanders used by Sessions and Larson (1987) for their comparative analysis of limb differentiation rate and genome size. Letters assigned to nodes of the tree represent independent contrasts plotted in Fig. 5.13.

They tested their hypothesis for both variables by means of a rank correlation of the unstandardized pairwise comparisons. Limb differentiation rate but not growth rate was significantly inversely related to *C*-value. Large positive differences in the *C*-value within clades tended to go with large negative differences in the limb differentiation rate within clades ($r_s = -0.47$, $P < 0.025$; Fig. 5.13).

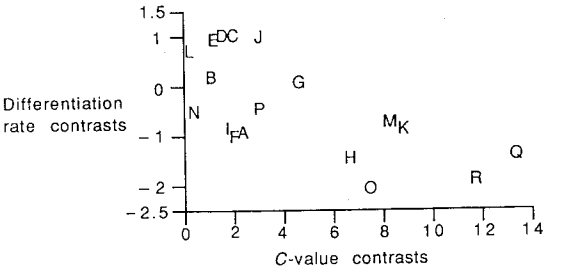


Fig. 5.13. Limb differentiation rate contrasts plotted against genome size (*C*-value, mass of DNA per haploid nucleus) contrasts for Plethodontid salamanders. The letters designate the nodes in Fig. 5.12. (After Sessions and Larson 1987).

Another application of Felsenstein's model is reported by Losos (1990), who studied locomotion and morphology in *Anolis* lizards. These lizards exhibit three locomotory patterns: running, jumping, and walking. Losos was interested in whether species that use one particular form of movement over another have morphological specializations for that behaviour. For example, species which typically run, such as sit and wait predators, may have evolved longer hind limbs.

Data on the percentage of movements attributable to running, jumping, and walking were collected from field observations of the 13 Jamaican and Puerto Rican *Anolis* species. Measures of fore limb, hind limb, and snout-vent length were obtained from 15 individuals from each species. Then, Losos used Felsenstein's method to analyse the relationship between morphology and the percentage of total movements that were walks. Twelve pairwise comparisons were calculated from the phylogeny in Fig. 5.14. Comparisons were standardized using branch length information obtained from literature sources. Hind limb length was negatively correlated with walking frequency controlling for body size (walking frequency as a percentage of all moves, $r = -0.75$, $P < 0.01$).

To summarize this section, Felsenstein's (1985*a*) method finds a set of independent pairwise differences or contrasts, each of which is scaled by its

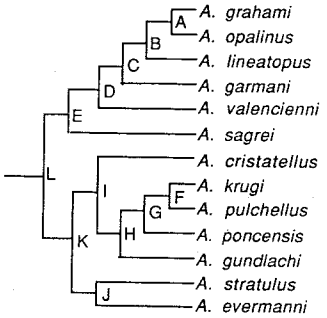


Fig. 5.14. Phylogenetic relationships among 13 species of Jamaican and Puerto Rican species of *Anolis* lizards. The 13 species allow 12 independent comparisons, labelled A to L, between the pairs of daughter taxa derived from each node. This phylogeny was used by Losos (1990) to examine whether evolution in limb morphology has been associated with evolution in locomotor propensities and movement (see also Section 5.10 and Fig. 5.22 in which this phylogeny is used to illustrate directional methods—each of the 12 nodes provides for two ancestor-descendant comparisons). Path lengths, not given here, were estimated by Losos using immunological and electrophoretic distance measures.

expected standard deviation. Expected standard deviations are derived by postulating an evolutionary model (Brownian motion) that translates branch lengths into units of expected evolutionary change. This illustrates how the choice of a comparative method is a choice of the way evolution proceeds as well as a choice of a set of statistical procedures. If, for example, a gradualist model of evolution such as this is assumed, and evolution has been punctuational (either everywhere or in some branches), or has proceeded at different rates in different branches, then the accumulation of variance will not be properly estimated, and the variates will not be scaled as intended (see Section 5.10, and Martins and Garland 1991).

These comments are not a criticism of Felsenstein's method. His procedures provide a way, in principle, of scaling the data according to whatever evolutionary model is proposed. Indeed, Felsenstein's procedures even allow for different rates of change in different branches.

5.8.3 Two Felsenstein-like approaches

Methods developed by Grafen (1989) and by Pagel and Harvey (1989*b*) follow Felsenstein's (1985*a*) idea of finding a set of independent comparisons. However, these methods can be applied to imperfectly resolved

phylogenies, such as might be the case if a taxonomy was used in place of phylogeny. Here we describe how the methods work on a simple bifurcating phylogeny. We take up the case of imperfect phylogenies in Section 5.8.6.

Grafen's method

Felsenstein's method assumes that branch lengths are known and uses a null-model of character change to derive expected variances of the observations. Grafen (1989) assigns branch lengths according to a counting rule, and assumes that the expected variance of change along a branch is proportional to its length (Brownian motion model). Branch lengths are found by first assigning to each higher node, one fewer than the the number of species below it in the tree. Thus, species receive a zero, the next higher node is assigned one fewer than the number of species in it, and so on. Then, branch lengths are calculated as the difference between successive nodes. For example, the branch lengths leading to two species sharing an immediate ancestor would be one. Alternatively, if branch lengths are known by some independent means they can be specified directly.

The initial branch lengths are then lengthened or compressed in response to a parameter ρ that is estimated from the data. A maximum likelihood procedure is used to estimate ρ which alters all path lengths in the tree by a positive power that can vary between 0 and 1. In its extremes (i.e. where ρ is close to 0 or to 1), this parameter alters the branch lengths so that the majority of the variation in the tree is placed either: (1) close to the species level making species relatively independent (that is, branches leading to species are lengthened) or (2) close to the root of the tree making higher nodes more independent. Alternatively, ρ may lie somewhere in between (a single value of ρ is estimated, thus this stretching or compression of the tree is the same for all variables regardless of their individual distributions of variance throughout the tree). Values of ρ near 1.0 assign greater variance to higher nodes by stretching the higher branches; values nearer to zero assign greater variance to lower nodes by stretching the lower branches.

Once the estimates of the variances of change for different branches are obtained, Grafen's procedure finds comparisons among pairs of species or nodes, for a bifurcating phylogeny, like Felsenstein's method. The comparisons are scaled according to their estimated branch lengths, where branch length is a measure of expected variance. The parameter ρ is used to remove any correlation between the magnitude of a standardized comparison and its estimated variance. Then, on the assumption that the Brownian motion model provides comparisons that have been properly scaled, relations among two or more variables can be studied by standard correlation and regression techniques. Grafen (1989) reports simulation

studies in which, for characters generated by a Brownian motion model of change, his method yields valid Type I error rates and has good statistical power. Stone and Willmer (1989) used Grafen's method to examine whether body size and thermal regime are related to warm-up rates in bees.

Pagel and Harvey's method

The method that we have developed finds the same set of comparisons as Felsenstein's or Grafen's methods but differs from those methods in the way that it scales comparisons for the expected variance.

The method assumes that rates of evolutionary change are likely to vary in different portions of the tree, and even within branches of the tree, and thus initially sets all branch lengths equal. For a bifurcating tree all comparisons are between two species or subtaxa. The method assigns all of these comparisons (differences) an expected variance proportional to two times the fixed branch length (variance of a difference between two independent data points is just the sum of their variances). Although this is implicitly a punctuational view of evolution, we do not mean necessarily to advocate that view. Rather, the method explores the possibility that arbitrarily fixed branch lengths can nevertheless yield acceptable statistical properties. At a later stage, patterns in the data are used to assess whether the scaling has in fact produced those properties (see p. 151) This method of scaling the comparisons has yielded approximately normal distributions of residual errors about regression lines in two recent studies (Harvey *et al.* 1990; Trevelyan *et al.* 1990).

As with the previous two techniques, the set of scaled comparisons can be analysed by standard regression and correlation techniques on the assumption that the scaling procedures result in a set of comparisons with equal expected means and variances under the null hypothesis[13].

5.8.4 Limitations of procedures for scaling comparisons

Many comparative studies of continuous variables combine in the same analysis, measures as disparate as mass, timing, and counts. It is reasonable to expect separate rates and distributions of change in these variables, or different rates and distributions of change in different regions of the phylogeny. If we had such information, it could then be applied

[13] The method has been used to examine relationships between clutch size and body size in birds (Blackburn 1990); metabolic rate and life history characteristics in mammals and birds (Harvey *et al.* 1990; Trevelyan *et al.* 1990); the size of the brain, its component parts, and ecological differences in mammals and birds (Harvey and Krebs 1990; Harvey in press); bird song, mating system, and life history (Read 1989); geographic range and habitat use in mamals (Pagel *et al.* 1991); parasite burden and geographical range in birds (see Harvey *et al.* 1991); and recombination rates and age at first reproduction in mammals (Eldred in press).

flexibly in Felsenstein's model. The computer application of Grafen's method, for example, allows the user to specify branch lengths (in units of variance of evolutionary change). However, lacking such information, we should be aware that the comparisons may not have the statistical properties we desire.

Formally, we should not use parametric statistics to analyse the comparisons derived from any of the three methods described in the previous two sections unless we know that the assumptions of those statistics have been met. Because the set of comparisons can reasonably be regarded as independent, the critical remaining assumption is that they all have the same variance. The Brownian motion model implicit in the methods does not guarantee this: equality of variances depends upon whether the Brownian motion model provides an accurate description of evolutionary change. In practice, correlation and regression techniques are quite robust to violations of their assumptions. However, this should not be taken as license to use them uncritically. We describe two techniques in this section that can be used to increase the chances that our statistical tests are valid: analysis of residuals, and randomization procedures.

Analysis of residuals

The relevant information for an analysis of residuals is the distribution of residual errors around the regression line formed by the relationship of the Y comparisons to the X comparisons. The residuals are found as the difference between the observed value of the Y variable and its predicted value as determined by the regression line. The frequency distribution of residual errors should be normal, with approximately 95 per cent of the points within two standard deviations of the mean. If the comparisons have not been scaled properly, then some will have larger true variances than others. This will manifest itself as heterogeneity of the variance of the residuals about the regression line.

We will concentrate on one technique for the detection of heterogeneity of variance among the residuals. Let s^2 be the variance and s the standard deviation of the residual errors about the regression line. It is a property of regression that the mean of the residual errors is zero. The vector

$$\mathbf{Z_e} = \frac{\mathbf{e}}{(s^2)^{1/2}} \tag{5.7}$$

is the vector of residual errors divided by the square root of the variance (the standard deviation) of residual errors. This transforms the residual errors to standard scores, that is, scores with a mean of zero and a standard deviation of one. Many regression procedures automatically calculate

'standardized residuals' according to eqn (5.7). If the residuals have all come from the same underlying normal distribution, then a histogram of the elements Z_e should be normally distributed with approximately 95 per cent of its observations falling in the interval −2 to +2. This is easily checked by means of a binomial test: the number of observations outside of the −2 to +2 range should not exceed that expected by chance under the binomial given that $P = 0.05$, where P is the probability of being < -2 or $> +2$. Other more sophisticated tests can be used to assess the shape of the distribution.

The residuals also should show no tendency to change systematically with the predicted value of Y obtained from the regression equation. The correlation of the residuals with the predicted Y is always exactly equal to zero. Nevertheless, the spread of residuals might increase (or decrease) with the predicted Y, or the residuals may show a curvelinear pattern against the predicted Y, and still have a zero correlation overall. Either of these patterns suggests heterogeneity of variance.

In the face of significant heterogeneity of variance what can be done? Standard methods are available on most statistical packages (SPSS, Minitab, SAS, GLIM, and BMDP all have methods for treating heterogeneity of variance). Heterogeneity of variance means that some form of weighted regression is required. Alternatively, one might attempt transformations of the raw data or of the comparisons as a way of removing the heterogeneity of variance. And here we come full circle to the methods for scaling comparisons that have been described previously in conjunction with the various methods. This is because weighted regressions, and (non-linear) transformations work, in effect, by stretching or compressing individual data points as a way of equalizing residual errors. A very large positive residual error may suggest that the data point needs to be scaled downward, a very large negative residual may suggest the opposite. The only real difference in doing it at this stage is that the weighting is conditioned on patterns in the data, rather than in response to an assumed model.

Randomization tests

In this section we are concerned not with heterogeneity but with the problem of not knowing what the null hypothesis distribution is. Randomization tests provide a way to estimate the null hypothesis sampling distribution from the data (Bradley 1968; Sokal and Rohlf 1981). Then, the result observed in the raw data can be compared against the simulated null-model distribution of outcomes to obtain valid statistical hypothesis tests (examples from comparative biology are given in Pickering 1980; Harvey 1986; Elgar and Harvey 1987; Pagel and Harvey 1988*a*; Blackburn *et al.* 1990).

The general procedure of a randomization test is repeatedly to shuffle a data set and calculate some summary statistic each time as a way of generating a frequency distribution of outcomes under a given null hypothesis. For example, the null hypothesis distribution for a simple correlation between pairs of independent comparisons on Y and X might be simulated by randomly reshuffling the Y comparisons or the X comparisons and calculating the correlation. The histogram of correlations obtained from doing this a large number of times becomes the null-hypothesis distribution. The actual obtained correlation is checked against the empirically derived distribution of correlations to see if it is sufficiently large to consider that it is not a chance result.

Elgar and Harvey (1987) used randomization tests to analyse data on the relationship between basal metabolic rate and diet. McNab (1986*a*, *b*) had argued that basal metabolic rate in mammals was associated with diet even after adjusting for body size. Elgar and Harvey's objection was that diet categories are not evenly distributed among mammalian taxa, and thus differences in basal metabolic rate might be associated with differences among taxonomic groups. The relationship between diet and metabolic rate must be shown to hold independently of taxonomic association. Elgar and Harvey (1987) employed a randomization procedure that shuffled metabolic rates (adjusted for body size) and diet categories among taxonomic groups. This unconfounded taxonomy from the other two variables. After each of 2000 shufflings, they calculated the relationship between metabolic rate and diet. They then compared the actual empirical result with the distribution of results from their randomizations. They were able to confirm McNab's claim for only two of the 10 diet categories.

Sessions and Larson's (1987) study of the relationship of genome size and limb differentiation rate (Fig. 5.13) can be used to provide a useful illustration of randomization procedures. We re-analysed their set of 18 comparisons by a randomization procedure designed to capture the null-hypothesis distribution of two types of correlation coefficient: one is the regular Pearson correlation (r), and the other is the coefficient of congruence (r^*). The coefficient of congruence is not sensitive to the direction of change within a comparison (Harman 1967). For example, it does not distinguish between a pair of contrasts that are both positive versus a pair that are both negative. This is potentially important because the sign of a contrast is arbitrary: we have no basis for deciding between whether to subtract A from B versus B from A, where A and B are two daughter taxa or species.

Our procedure first randomly re-ordered the set of limb differentiation comparisons against the set of C-value comparisons, then calculated the two correlations on the randomized data. This was repeated 2000 times and the frequency distribution of results obtained. Figure 5.15 displays the

frequency distributions for the two types of correlation. As would be expected, the frequency distribution for the Pearson correlation is centred roughly symmetrically around zero. The r^* distribution, however, has a smaller variance and is centred around –0.45. The slight skewing of the distributions is probably due to the distributional quirks of the 18 pairs of comparisons, and supports Sessions and Larson's expressed concern about testing the Pearson correlation against the tabled null-distribution values. However, the randomizations include these quirks in the data, and thus the appropriate probability values can be read right off the distributions. The obtained Pearson correlation of –0.65 is significant at the $P = 0.005$ (two-tailed) level. The r^* coefficient of –0.66 is also highly significant. These results agree with, but are slightly more extreme than, those reported by Sessions and Larson (1987).

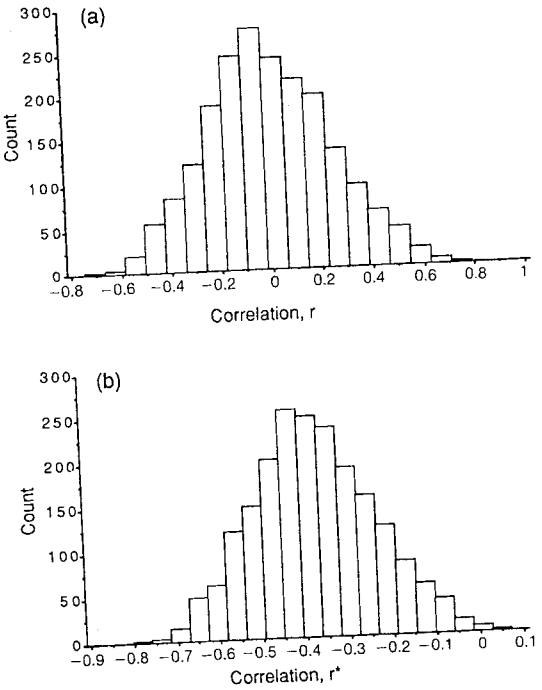


Fig. 5.15. The data from Fig. 5.13 have been randomized so that each differentiation rate contrast is paired with a random *C*-value contrast. The procedure was repeated 2000 times. On each occasion ('Count') a Pearson's product moment correlation (r) was calculated (a, above), as was a coefficient of congruence (r*) (b, above). The coefficient of congruence, as described in the text, is not sensitive to the direction of change within a comparison.

Other procedures for using the data to derive the null distribution exist and are slowly receiving more notice. The bootstrap (e.g. Efron and Gong 1983) is probably the best known of these re-sampling methods and is based on more rigourous statistical theory than simple randomization procedures. Lunneborg (1985) and Wu (1986) describe applications of the bootstrap to testing the correlation coefficient.

5.8.5 Summary of independent-comparisons methods

Of the procedures we have described for conducting analyses separately within groups, three (Felsenstein 1985*a*; Grafen 1989; Pagel and Harvey 1989*b*) extract from the data set a series of phylogenetically-defined independent comparisons which each bear on the comparative hypothesis. The comparative relationship is assessed via the number of times it has independently evolved rather than by the number of species or higher nodes that have come to inherit it. The three procedures will extract the same set of comparisons from a bifurcating phylogeny, but differ in how those comparisons are scaled. The illustrative example in Box 5.2 ignores the problem of scaling.

5.8.6 Comparisons on incompletely resolved phylogenies

The methods described in Sections 5.8.2 and 5.8.3 require that the true branching phylogeny is known. But, it is often the case that we are ignorant of the true (probably) bifurcating phylogeny. Good phylogenies based on molecular techniques are becoming available, but they often resolve the phylogeny only to the level of subfamilies or tribes (e.g. Sibley and Ahlquist 1985), and they may even lack resolution at these levels (see Sarich *et al.* 1989).

In many cases we will have either a poor phylogeny or even a taxonomy to represent the branching of species. These incompletely resolved phylogenies typically will contain many multiple-nodes, that is, nodes from which more than two daughter taxa are represented as direct descendants. With more than two tips in a node the logic of finding a simple difference breaks down. We need methods to cope with multiple nodes if we are to apply the logic of comparisons developed above. Our assumption is that multiple nodes are monophyletic but do not actually represent true simultaneous radiations. Rather we assume that multiple nodes conceal some unknown branching pattern (Grafen 1989; Maddison 1989).

If the multiple node is not a simultaneous radiation then the tips of the node will not be independent. For example, a multiple node with three species may actually conceal the phylogeny of Figure 5.16. Species 2 and 3 share a phylogenetic history that species 1 does not. Not knowing this we can nevertheless make the assumption that the multiple node conceals at least one evolved difference: for example, the difference between species 1 and the node from which species 2 and 3 descended. Ignoring branch

lengths, this difference can be represented as species 1−(species 2 + species 3)/2. This weighted difference score reduces the information in the three species points to a single point by subtracting the mean of species 2 and 3 from species 1. This represents our assumption that there is at least one evolved difference within the multiple node.

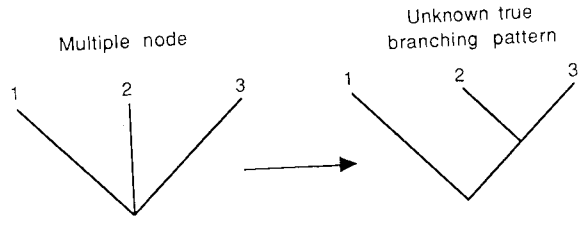


Fig. 5.16. A multiple node concealing an unknown true branching pattern. Taxa 2 and 3 share phylogenetic history that is not shared by taxon 1.

The example just given involves multiplying each species' value by a weight, and then finding the sum of the weighted values. If the weights are denoted by c_i, and the values of the character in each species denoted X_i, then the weighted sum can be written as $\Sigma c_i X_i$. Weights that have the property that they always sum to zero are known as *contrast coefficients*. Because the coefficients sum to zero, the weighted sum can be thought of as a weighted *difference* score. This logic can be applied to any number of points. The difference between two data points is the simplest case of a linear contrast, where the coefficients are equal to +1 and −1. In this case, then, the linear contrast is identical to what we would get by finding differences throughout a bifurcating phylogeny. Thus, we can think of simple difference scores as a special case of the more general problem of finding linear contrasts.

One way to view the contrasts coefficients is that they represent a hypothesis about the branching pattern of the unknown phylogeny. All of the tips that receive a positive weight are implicitly being treated as more closely related to each other than to the tips that receive a negative weight. The problem is that there are an infinite number of different linear contrasts for any given set of data points. Thus, we need to justify the methods for finding the contrast coefficients. We will describe the methods Grafen (1989) and Pagel and Harvey (1989*b*) chose.

Grafen (1989) chose contrast coefficients for multiple nodes by a procedure that gives greater weight to those species or nodes whose data points are not well explained by the phylogeny or by the other predictor

variables under consideration. Grafen's procedure in effect first regresses the Y variable on to a series of dummy codes representing phylogenetic membership, plus any other control variables chosen by the investigator. The residuals from the regression of the Y variable will have the property that they sum to zero within each taxonomically or phylogenetically defined group (see for example the residuals in Box 5.1 for Stearns' method). The residuals are then used as the weights in the linear contrasts on the original Y and X variables, and a contrast is found for each presumed monophyletic group. Values of Y that are not well explained by phylogeny and by the X variables used as controls will have larger residuals, and thus be weighted more heavily in the contrast. A phylogenetic interpretation of this is that observations that deviate from the regression line in the same direction are more closely related to each other than those that deviate in opposite directions.

The method that we (Pagel and Harvey 1989*b*; Pagel unpublished manuscript) use to derive the linear contrasts relies on the assumption that the X (or Y) variable can provide useful information about the hidden phylogenetic structure in the multiple node. Multiple nodes are divided into two sub-nodes according to the distribution of X; those above the mean on X within the multiple node are assigned to group 1, those below the mean on X within the node are assigned to group 2. This criterion assumes that phenotypically more similar species (or higher taxa) within a multiple node are phylogenetically more closely related. If the X variable does not provide useful phylogenetic information, the comparisons may lack efficiency.

All branches within a node are assumed to be the same length. The contrast coefficients are found as the reciprocal of the number of species or groups within each sub-node. Then, the coefficients for group 2 are given a negative sign. Thus, for example, if the phenotypic criterion assigned three species to group 1 and two species to group 2, the contrast coefficients would be 1/3, 1/3, 1/3, −1/2, −1/2. This procedure is a slight modification of that which was used in an earlier version of the method. Contrast coefficients assigned this way always sum to zero. The method gives greater weight to the subgroup with the fewer number of species. In the case of a bifurcation, the coefficients will always be 1, −1. The set of coefficients is applied to X, to Y, and to each of the control variables to find a weighted difference score on each. The value of each weighted sum is standardized by dividing by the sum of the absolute values of the contrast coefficients. This procedure is repeated for each multiple node. Pagel (unpublished manuscript) describes this procedure as well as a more general procedure for taking into account either known or estimated branch lengths.

The set of weighted 'difference' scores derived from the separate independent comparisons can be used to estimate not only whether Y and

X are related, but to estimate the nature of that relationship as well. As we shall see in Chapter 6, the slope of the relationship derived from the analysis of the linear contrasts across taxa will estimate the slope of the relationship between Y and X, provided that the contrast coefficients are found in such a way that they are uncorrelated with the residual variation in the Y variable. Estimating the slope from independent contrasts avoids the problems of using non-independent species values (see Sections 5.3 and 6.6.2).

5.8.7 An example of independent-comparisons with unresolved phylogenies

Earlier in this chapter, for illustrative purposes, we used Millar and Zammuto's (1983) data set to examine the relationship between age at maturity and life expectation among mammals. When we employed the higher nodes method (Section 5.4) there was a significant positive relationship between the two variables independently of body weight. But when we used Stearns' phylogenetic-subtraction method, the relationship was not significant (Section 5.5). We pointed out that each method discards the information used by the other, and that independent comparisons allowed the use of all available information. Is the relationship significant when independent-comparisons methods are used?

To develop the example, we have continued to use the simple taxonomy of species, genera, families and orders as in Section 5.5, and have calculated the linear contrasts and independent comparisons for body weight, age at maturity, and life expectation at maturity. The 29 species classified according to the standard mammalian taxonomy used in Section 5.5 (Corbet and Hill 1980) allowed the calculation of 17 independent comparisons for each variable. When size-independent life expectation comparisons are correlated with size-independent age at maturity comparisons, the relationship is highly significant ($r = 0.57$, n = 17, $P = 0.02$).

However, mammalian taxonomies provide only crude approximations of phylogenetic relationships. Using cladistic techniques on morphological data together with analyses of genetic variation recognised at the molecular level (see e.g. Benton 1988a, b), it is now possible to produce more accurate mammalian phylogenies which can be used with the independent comparisons methods. Using Millar and Zammuto's (1983) data set together with the most recent phylogenetic reconstructions drawn from many sources, we could distinguish 23 independent contrasts. With body weight effects controlled for, the correlation between life expectation at maturity and age at maturity is again highly significant ($r = 0.80$, $n = 23$, $P < 0.001$; Fig. 5.17).

The method can be used to reveal aberrant taxa or, as in this case, those most highly responsible for the relationship. The point at the bottom left of

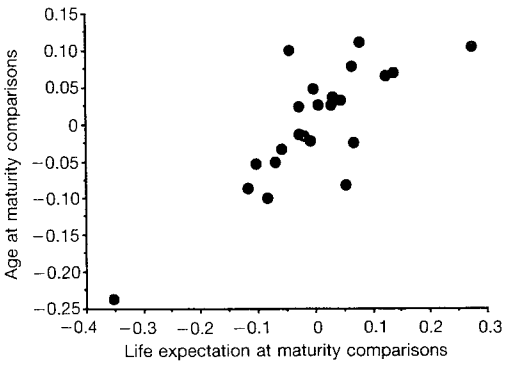


Fig. 5.17. Independent comparisons calculated from Millar and Zammuto's (1983) data. Age at maturity and life expectation comparisons are both corrected for body size in the figure, being given as a residuals from the regressions of each comparison on body weight comparisons. The correlation is highly significant (r = 0.80, n = 23, P <0.001). The bottom left point is a comparison between *Ochotona princeps* and *Sylvilagus floridanus*, and when this outlier is removed the correlation remains highly significant (r = 0.66, n = 22, P < 0.001).

Fig. 5.17 seems to be an outlier and represents a comparison between two lagomorphs, the northern pika *Ochotona princeps* and the Eastern cottontail *Sylvilagus floridanus*. The significance of the correlation does not depend on that comparison (with the comparison removed $r = 0.66$, $n = 22$, $P < 0.001$).

5.9 Testing hypotheses with independent comparisons

Under most circumstances, testing hypotheses with independent comparisons on Y and X proceeds as in Figs 5.17 and 5.18(a). When the relationship between the Y and X comparisons is positive and all or nearly all comparisons are positive, a simple linear regression or Pearson correlation (or non-parametric equivalent) can be used. If the X comparisons are positive and the Y comparisons are negative, Fig. 5.18(b), the same procedure can be used.

However, some patterns in the Y and X comparisons require more care in their interpretation. Specifically, it is necessary to test whether the slope *and the intercept* of the regression of the Y comparisons on the X comparisons differ from zero. Figure 5.18(c) depicts the case where all comparisons on Y and X are positive, but the magnitude of the Y comparisons does not change with changes in the magnitude of the X

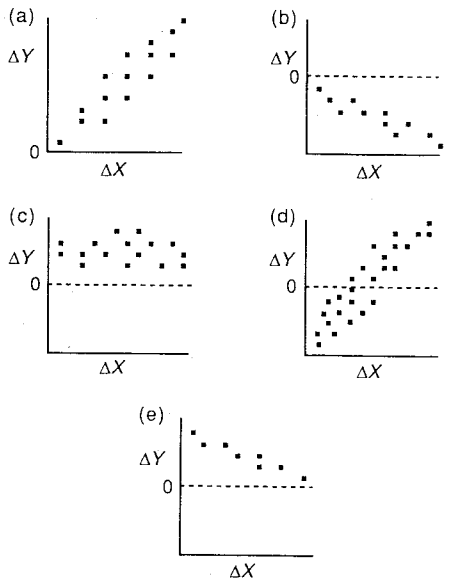


Fig. 5.18. Hypothetical patterns of relationships between sets of independent comparisons in which the comparisons on X ($\triangle X$) are constrained to be positive. Comparisons on Y ($\triangle Y$) can be either positive or negative, and can be positively or negatively correlated with comparisons on X.

comparisons (slope zero, intercept greater than zero). This pattern, nevertheless, is strong evidence against the null hypothesis of no relationship between changes in Y and X, which would predict an equal number of negative and positive comparisons on Y. A standard linear regression on these data would find no relationship. However, either a regression forced through the origin (Grafen 1989) or a simple binomial test will detect a significant relationship. If the former is used, however, the slope of the line should not be interpreted, only the sign of the slope is of interest. Figure 5.18(d) shows a relationship that appears significantly positive, but on reflection reveals that the relationship between Y and X goes in the positive direction in about half of the taxa and in the negative direction in the other half (slope and intercept differ from zero). A regression through the origin or a binomial test would correctly indicate that no significant relationship existed. Figure 5.18(e) shows an instance in which a standard linear regression would find a significant negative relationship (again, slope and intercept differ from zero). But, as all Y

comparisons are positive, it would be incorrect to interpret such a slope. Again, either a regression through the origin or a binomial test would treat this situation properly.

Many other patterns are possible and can be dealt with by applying the logic of one of the examples given here, but what realistic examples might give rise to patterns such as those depicted in Figs 5.18(c)-(e)? Figure 5.19 depicts a possible scenario.

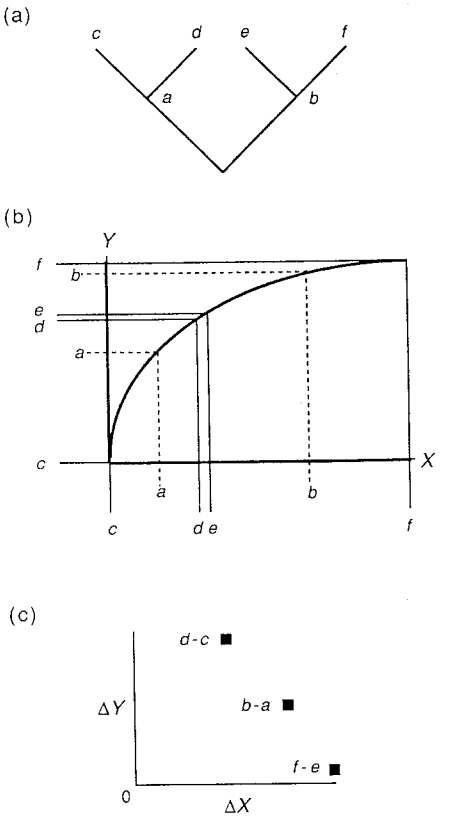


Fig. 5.19. The phylogeny in (a) shows four extant species and two ancestral nodes that can be used to make comparisons. Y and X are related by the curve shown in (b), with the values for nodal taxa a and b, and the four extant species c to f being depicted on the figure. Independent comparisons on X and Y ($\triangle X$ and $\triangle Y$) are shown on (c), and are negatively correlated even though a phylogenetic increase in X is associated with an increase in Y for each comparison.

Figure 5.19(a) shows a phylogenetic tree, Fig. 5.19(b) shows a non-linear relationship between Y and X in the raw data for the taxa in that phylogeny, and Fig. 5.19(c) shows a plot of comparisons similar to that in Fig. 5.18(e). The points a to f on the X and Y axes in Fig. 5.19(b) indicate the hypothetical values for the four species and the two higher nodes. Values of $b-a$, $d-c$, and $f-e$ form the three independent comparisons from the phylogeny. The relationship between Y and X in Fig. 5.19(b) is such that the small difference between d and c on the X axis (indicated by the vertical lines drawn up from those points) translates into a large difference on the Y axis. The large difference between e and f, on the other hand, translates to a small difference in Y, and the difference between b and a is intermediate on both axes. Note that the three changes in X are correlated with their average values of X. By appropriate non-linear transformation of the Y axis, this example can generate either of the relationships shown in Figs 5.18(a) and (c). Inverting the Y axis allows other relationships, including that in Fig. 5.18(b), to be generated. In practice, it is often possible to transform Y and X so that the relationship between them is linear and the difference between the values for taxa being compared is independent of their average value on the X axis (see Chapter 6). When such transformations are made, regression intercepts tend not to differ from zero and difficulties of interpretation are minimized.

5.10 Directional methods

The methods we have described so far do not test the direction of change in two or more variables along the branches of a phylogeny. Thus, we might have two species that differ in Y and X in the same direction, but this alone does not tell us the direction of change from ancestor to descendant. Both species could have evolved larger values of X and Y than their immediate ancestor, they could have both become smaller, or even evolved in opposite directions. We reported a similar distinction in Chapter 4 between Ridley's (1983*a*) and Maddison's (1990) methods. Huey (1987) points out that directional methods are useful for studying rates of change over evolutionary time, for assessing the ecological conditions that may have selected for derived traits, and for analysing whether changes in two traits are coincident. Chapter 1 discussed further distinctions between directional and non-directional comparative methods.

Huey and Bennett (1986, 1987) report a method for assessing directional changes in continuous characters along the branches of a phylogeny. They studied directional and non-directional trends (see Chapter 1) relating preferred body temperature and optimal temperatures for running speeds in Australian scincid lizards. Preferred body temperature is defined as that temperature selected by the animal when exposed to a thermal gradient.

Huey and Bennett were interested in whether the temperature at which a lizard species runs fastest has evolved to keep pace with preferred body temperatures.

To assess the direction and rate of evolutionary change in these two variables, Huey and Bennett first had to reconstruct the ancestral states of the lizard phylogeny. They used an iterative procedure (suggested by J. Felsenstein) that estimated each higher node as the average of its nearest three neighbours, subject to the provision that the final set of higher nodes minimized the squared change in the links of the tree, summed across all links (see Chapter 3). W. Maddison (personal communication) has shown that, if all branch lengths are assumed to be equal, this procedure also yields the maximum likelihood set of changes under a Brownian motion model. The phylogeny that Huey and Bennett used and the reconstructed ancestral character states for preferred body temperature are shown in Fig. 5.20.

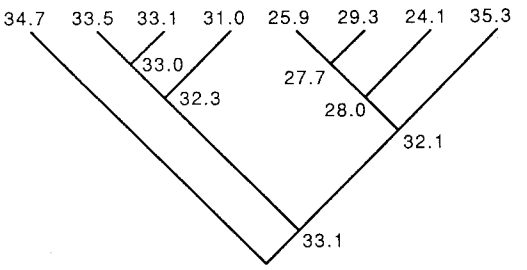


Fig. 5.20. A tentative phylogenetic tree of seven Australian skink genera. At the top left, the non-Australian *Mabuya* is used as an outgroup. Australian genera, listed from left to right are *Egernia, Tiliqua, Leiolopisma, Eremiascincus, Sphenomorphus, Hemiergis,* and *Ctenotus*. Numbers at tips are generic averages for thermal preferences (^{0}C). Numbers at nodes are presumed ancestral preferences generated by a minimum evolution method. (After Huey and Bennett 1987).

The directional relationship between optimal running temperature and preferred body temperature was assessed by calculating the regression of changes in optimal running temperature against changes in preferred body temperature. Changes were calculated as the difference between generic means and the nearest higher node. The authors chose to use genera because of concerns that species were not independent, although they acknowledge that the same problem, if not as extreme, may also apply to genera. Huey and Bennett (1986, 1987) did not adopt a model of evolutionary change from which they could derive the expected variances

of change along the branches. Instead, the changes along the branches were treated statistically as if they were drawn from the same null hypothesis distribution, an explicitly punctuational model of evolution. The regression of changes in optimal running temperature on to changes in preferred body temperature yielded a slope of 0.25, significantly less than 1.0, indicating that directional changes in optimal running temperature have lagged behind changes in preferred body temperature (Fig. 5.21).

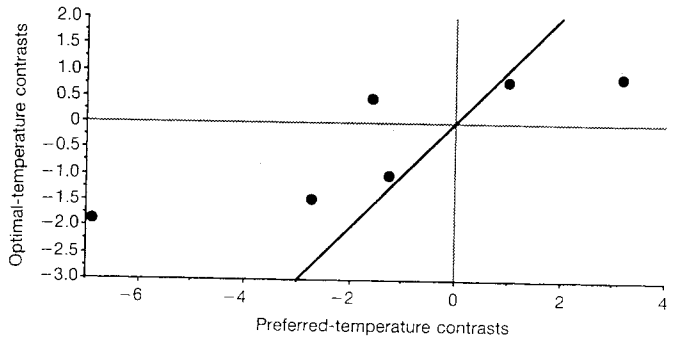


Fig. 5.21. Optimal running temperature versus preferred body temperature comparisons in Australian skinks. Change is measured as the difference between each of six generic values and its nearest node. Adaptation is seen to be partial because the change in optimal running temperature is less than that in preferred body temperature. The line represents perfect coadaptation where changes in optimal and preferred temperatures are the same. (After Huey and Bennett 1987)

Losos (1990) employed Huey and Bennett's procedure to study directional changes in locomotory behaviour and morphology in *Anolis* lizards (see Section 5.8.2 for a description of this study). Figure 5.22 plots the directional changes in locomotor behaviour versus directional changes in hind limb length, controlling for body size.

Huey and Bennett's procedure differs from the independent-comparisons methods by examining the changes between ancestors and descendants rather than between contemporary species or higher nodes. This allows tests of explicitly phylogenetic hypotheses about the direction and rates of evolutionary change that are not directly available to non-directional methods. In practice with this method, as the authors note, there will be some dependence among the lower and higher nodes because the higher nodes are estimated from the lower nodes. For example, in the extreme case, we might estimate the ancestral state of a genus as the arithmetic mean of the species. Because the sum of the deviations of the species about their mean must sum to zero, for a genus with k species only the first $k-1$ species in the genus can be independent. If changes along branches are computed as deviations from higher nodes, the property that the deviations sum to zero forces some species to appear to have increased and others to have decreased with respect to the ancestral condition. However, the effects of this non-independence do not appear to be critical (see Section 5.11).

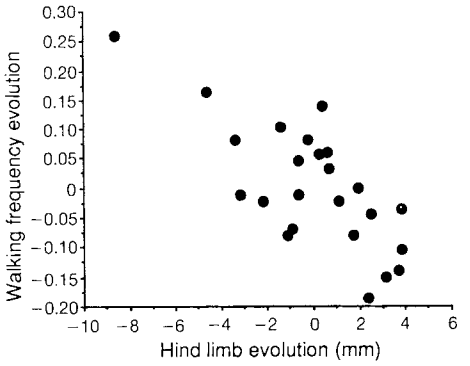


Fig. 5.22. The relationship between the evolution of hind limb length and walking frequency in *Anolis* lizards using Huey and Bennett's (1987) method. Walking frequency is walking as a proportion of all moves, and hind limb evolution is the residual of the change in hind limb length relative to the corresponding change in snout-vent length. Each point corresponds to the amount of evolutionary change in one ancestor-descendant pair. (After Losos 1990).

5.10.1 Accumulation of variance over time

Bell (1989) reports a novel use of the nested analyses of variance to study the accumulation across taxa of diversity in characters over time. This technique is useful for illustrating the rate of evolutionary diversification with time, and for identifying time periods of rapid evolutionary change.

As part of a larger study, Bell collected information on body weight and chromosome number in mammals. A nested analysis of variance was conducted on each character using seven taxonomic levels. Bell also collected information on the approximate times of divergence of the taxonomic levels, and then plotted the cumulative percentage of variance at each taxonomic level against time of divergence. The results are shown in Fig. 5.23.

If Bell's analysis is based on correct assumptions, approximately 19 per cent of the variance in body mass that was eventually to appear among contemporary species was already present among super-orders within the

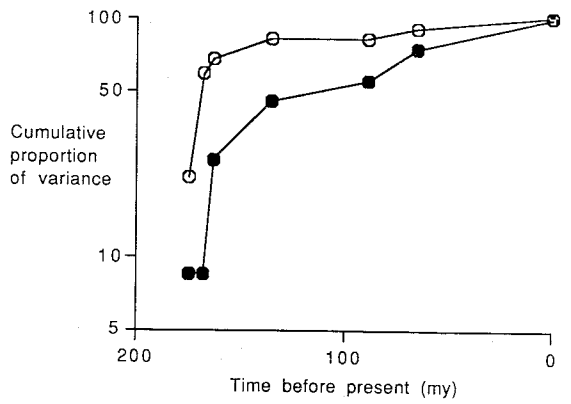


Fig. 5.23. The increase in variance of body mass (○) and chromosome number (●) through time in eutherian mammals. The eight taxonomic levels used for the analysis to produce seven nested variance components are infra-class, superorder, order, suborder, family, subfamily, genus, and species. (After Bell 1989).

infra-class Eutheria during early Tertiary times. Diversity in body mass among taxonomic groups increased sharply through the radiations of suborders by which time 78 per cent of the variance seen today was already present. The rate of diversification in body mass fell off sharply afterwards. Diversification in chromosome number lagged behind that for body mass at all stages. This means that the variation among taxonomic groups in chromosome number was not as great as that for body mass. One interpretation is that the variation in chromosome number within groups is always somewhat large compared to that across groups. Recalling the interpretation of the cumulative percentage of variance as an intra-class correlation coefficient (Section 5.4), the chromosome number of one member of a group will not be a very good a predictor of the chromosome number of another member of the same group. Bell (1989) suggests that chromosome number may have been more often associated with speciation events than was body size.

Figure 5.23 suggests that the amount of divergence among species is closely related to time since they shared a common ancestor. Bell tested this idea by plotting the variance of body mass among species within a genus against the age of the genus for 64 different genera. The plot, which showed a highly significant positive relationship ($P = 0.002$), confirms this idea.

5.11 Selected computer simulation results

We finish this chapter by presenting selected results from a series of computer simulations conducted by Martins and Garland (1991) to compare the performance of each of several different comparative methods under different evolutionary scenarios.

Our interest in Martins and Garland's work is not to use it to argue for the superiority of one technique over another. Such a conclusion would depend upon the models of evolution that were used to generate the simulated data sets being representative of the processes responsible for the diversity in real data sets. Rather, we use their results to emphasize that the success of a method depends strongly on whether the data being analysed were generated by the evolutionary model on which the method is based.

Martins and Garland analyzed the performance of five different comparative methods. The methods included a simple cross-species regression, Felsenstein's (1985*a*) method with standardized and unstandardized comparisons (i.e. the latter are not divided by an estimate of their standard deviation), and the method reported by Huey and Bennett (1987), again with standardized and with unstandardized changes. The standardized methods based on Felsenstein's approach were found according to two rules. The FL1G procedure standardized according to a Brownian motion gradual model of evolution in which variance accumulates additively over time. The FL1P, or punctuational, version allowed only one unit of variance per branch (i.e. all branches the same length). The non-standardized versions FL2G and FL2P are identical to the FL1 versions except they are not standardized.

Huey and Bennett's procedure was also represented by four versions. ME1G (for minimum evolution) calculates changes between all connected points on the phylogeny. Higher nodes were calculated as weighted averages of lower nodes, taking into account branch lengths. ME1P is the same as ME1G except all branch lengths are assumed to be equal. ME2G and ME2P replicate the ME1 versions except changes are calculated only between nodes and tips of the tree (ME2P corresponds most closely to what Huey and Bennett actually used). Felsenstein's, and Huey and Bennett's methods were then simulated for both punctuational and gradual models of change.

In the first of their simulations Martins and Garland studied Type I error rates, α, from simulations of a 'known' phylogeny of 15 species (see lower portion of Fig. 5.12, after Sessions and Larson 1987) with varying branch lengths. Data were generated by a Brownian motion process, with no correlation between Y and X. Branch lengths were measured in units of

expected variance. Thus, this data set conformed exactly to the model assumed by Felsenstein's method. Felsenstein's method yielded a Type I error rate very close to the nominal α of 0.05 (see Table 5.5) for this case. Interestingly, however, the other versions of Felsenstein's model had Type I error rates only about 3 to 4 per cent higher than the exact model, a difference which was significant statistically, but perhaps not very large in practice. The simple correlation across species had a very high Type I error rate. The minimum evolution models (after Huey and Bennett 1987), especially ME2G and ME2P, had somewhat higher Type I error rates, as might be expected given that changes along the branches are not independent (see Section 5.10).

Table 5.5 Type I error rates and power of several comparative methods to detect a false null hypothesis. For details of analysis models, see text. 'G' and 'P' refer to gradual and punctuational change respectively. 'Species' denotes a cross species regression. (After Martins and Garland 1991).

	Type I error rate		Statistical power			
Analysis model	Simulated G change	Simulated P change	Simulated gradual change for different values of ρ			
	α=0.05	0.05	ρ=0.25	0.50	0.75	0.90
FL1G	0.05	0.08	0.16	0.53	0.94	1.00
FL2G	0.07	0.06	0.13	0.40	0.87	1.00
FL1P	0.08	0.05	0.14	0.42	0.89	1.00
FL2P	0.09	0.06	0.13	0.39	0.86	1.00
ME1G	0.09	0.06	0.12	0.36	0.85	1.00
ME2G	0.14	0.09	0.10	0.28	0.75	0.98
ME1P	0.07	0.05	0.16	0.44	0.90	1.00
ME2P	0.10	0.08	0.14	0.38	0.85	0.99
Species	0.16	0.16	0.08	0.26	0.68	0.97

In contrast, when Martins and Garland simulated data according to a punctuational model of change, a different set of results emerged. In this set of simulations, a single punctuational change was allowed in each branch, regardless of its length. Here, Felsenstein's gradual model of change had slightly elevated Type I error rates, about 2.5 per cent higher than the two punctuational models, which had Type I error rates very close

to 0.05 (Table 5.5). The minimum evolution methods performed better with punctuational change than with gradual change but again, ME2G and ME2P did somewhat worse than the other minimum evolution models.

These results demonstrate a point that we have emphasized throughout this chapter: the success of a comparative method depends to a large extent on whether the assumptions that it makes about the way evolution proceeds are in fact true. However, it is important to point out that, aside from the species regressions, most of the Type I error rates, even for incorrect models, are not wildly elevated. Martins and Garland's simulations show that gradual and punctuational models provide reasonably approximate Type I error rates even when applied to data derived from the alternative model. However, in both cases the comparative method whose underlying model best matched that used to generate the data did significantly better, and not different from what would be expected theoretically.

Martins and Garland also conducted simulations to establish the relative power of the methods under punctuational and gradual models of change. For a range of simulated true correlations between Y and X, the punctuational and gradual versions of Felsenstein's method had similar power (Table 5.5). The minimum evolution methods generally had somewhat lower power.

5.12 Conclusions

We hope to have made clear throughout this chapter the inescapable connection between the statistical procedures that a comparative method employs to create a data set suitable for analysis by parametric techniques, and the evolutionary assumptions necessary to justify those procedures. As logic dictates, and simulation studies confirm, when a model's assumptions are true, the model performs well. When they are not met, the models do not perform as well. Nevertheless, models must be countenanced in spite of their weaknesses because, as we stated in the previous chapter, techniques which ignore them are liable to make implicit assumptions that are even less realistic than those made by the models. Comparative methods must either deal with assumptions directly in the form of explicit models which may be wrong, or acknowledge that they depend upon highly constrained evolutionary scenarios. Either way, we want to emphasise that the consequences must be dealt with statistically, and we have described methods of residual analysis to that end.

A niggling problem with independent-comparisons methods is that we do not as yet have a way of estimating ancestral characters that is independent of the distribution of extant characters. This can introduce some dependence among the members of a set of comparisons. Felsenstein

(1985*a*, p. 13), relaying an idea suggested by a student, suggested that 'we could use comparisons between pairs of species that we were fairly sure had a common ancestor not shared with any member of another pair, and that these contrasts could then be safely assumed to be independent'. Burt (1989) later reported the same technique, but without the evolutionary considerations with which Felsenstein imbued his model. This dependence does not seem to be a serious problem, however. If it were, we should not expect independent-comparisons methods to have performed so well in Martins and Garland's (1991) simulations.

On logical grounds, independent-comparisons methods and methods that use explicit ancestral character reconstruction to test the direction of evolutionary trends should be preferred over their cross-species rivals. The best developed cross-species methods unnecessarily discard large amounts of information that independent-comparisons and ancestral reconstruction (i.e. directional) methods approaches exploit. There is no good reason to discard as inappropriate for testing adaptive trends, the variance that is correlated with phylogenetic differences. We simply must know how to treat this variance, and independent-comparisons methods provide the way.

5.12 Summary

The branching structure of phylogenies ensures that species are not independent for statistical purposes. Various comparative methods differ in how they estimate and manage this non-independence. Some methods discard information in an attempt to create a set of independent points, while others which make use of all the variation in the phylogeny are to be preferred on logical grounds. These methods employ independent comparisons either to assess differences among species or higher nodes, or to assess the direction of evolutionary change. Evolutionary models implicitly or explicitly underpin all methods. The validity of a method depends upon whether the model on which it is based accurately describes the evolutionary processes that have generated diversity.

6

Determining the forms of comparative relationships

Allometry 'explains, among many other things, why large animals have relatively thick legs, why dachshunds can't be as large as elephants, [and] why flies can walk up walls' (Gould 1975, p. 245)

6.1 Introduction

Allometry has been variously defined as 'the study of size and its consequences' (Gould 1966, p. 587), 'the biology of scaling' (Calder 1984, p. 1), the study of 'the structural and functional consequences of changes in size or scale', (Schmidt-Nielsen 1984, p. 7), and perhaps more daringly as 'the study of the relationship between size and adaptation' (Fleagle 1985, p. 1). What is at issue in each of these quotations is the pervasive tendency for a variety of morphological, physiological, life history, and behavioural characteristics of organisms to show highly regular changes with changes in body size. Allometry is used to describe such relations quantitatively, and we use it here to illustrate a situation in which the comparative biologist's primary interest is not in whether a relationship between two variables exists but in the form that the relationship takes.

This chapter briefly reviews the field of allometry, and then moves on to describe how to estimate allometric relationships using the logic of independent comparisons that were described in Chapter 5. Later, we outline one approach to giving a theoretical basis to allometric slopes. Much of allometry is purely descriptive, and we suggest that an understanding of allometric relationships may be found in the details of the selective forces acting on the variables under study. We illustrate this idea with data on life history variation in birds and mammals. Finally, we sketch some of the difficulties that arise in estimating allometric relationships when the variable under study is caused by several factors.