

**Character construction in Morphological Phylogenetics, and
the Affinities of Turtles.**

Simon R. Harris PhD Thesis, University of Bristol, UK, 2004.

Full thesis available at http://www.ncl.ac.uk/microbial_eukaryotes/simonr_harris.html

Chapter 4: Compatibility Methods and Their Uses

4.1 Introduction

Virtually all published morphological phylogenies in recent years have been produced using parsimony analysis. Parsimony is a model of evolution that assumes that the simplest possible explanation for character evolution in terms of tree steps represents the best estimate of phylogeny. Such an assumption, although seemingly rational, is not necessarily true, and therefore may not produce the true phylogeny. Alternative methods, such as maximum likelihood, based on different hypotheses of evolution are also available, and are often employed for analysis of molecular data. Attempts have been made to produce likelihood methods for use with morphological data (Lewis, 2001). Currently such methods are only rarely implemented, although Bayesian inference (which is based on a likelihood foundation) is becoming more popular. One alternative to parsimony and likelihood for investigating both morphological and molecular data is compatibility analysis.

The aims of this chapter are to briefly review current compatibility methods, their uses and discuss their limitations. Two new compatibility methods are introduced:

- 1) **Fuzzy compatibility.** This is a method for quantifying how incompatible two incompatible characters are based on the minimum number of taxa that must be rescored to make the characters compatible. An algorithm for calculating this measure is presented.
- 2) **Boildown bootstrap.** This is a method that provides a way of deciding when to halt a boildown procedure. Two types of boildown bootstrap are introduced.

New ideas concerning potential problems for compatibility analyses due to polymorphisms, linked characters and inapplicable data within data matrices, and biases towards unbalanced trees in both compatibility and parsimony are discussed.

4.2 Compatibility

The concept of character compatibility has been integral to the process of phylogenetic systematics since its inception. Darwin (1859) realised the importance of character compatibility with statements such as “The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance” (:401-402). The first discussion of a true compatibility method

for estimating phylogeny was by Wilson (1965), who described the ideal phylogenetic character as one that “both uniquely defines a set of species and has not been reversed in evolution, so that all existing species which possess this state can be said to have descended from one species in the past that evolved the state”. He described a test to identify “unique and unreversed” characters. In the same year the term “character compatibility” was first coined by Camin and Sokal (1965) who independently developed essentially the same concept as Wilson. Both of these methods simply compared the compatibility of individual characters with a tree topology, a method utilised in parsimony analyses. In 1969, Le Quesne defined the ‘uniquely derived character’ concept with the statement “if one is studying the taxonomy of a group, a character that has evolved only in one direction on a single occasion in its history is more likely to give an unambiguous indication of its phylogeny”. Le Quesne (1969) discussed pairwise compatibility of characters, rather than looking at whether single characters can simply be mapped onto a specified tree without homoplasy. Two characters are pairwise compatible if it is possible to draw a tree representing a phylogenetic hypothesis upon which both characters can be mapped without homoplastic changes.

Compatibility methods are useful in attempting to separate phylogenetic signal from noise in a dataset. It is always hoped that in a data matrix the true phylogenetic signal is present in at least some of the characters. Other characters will represent, to varying degrees, nothing more than phylogenetically uninformative noise. Phylogeny reconstruction methods aim to pluck out the true signal from the noise in order to return the true phylogeny. Unlike some other methods, compatibility analysis involves no assumptions about the model of evolution. The only assumption necessary is the most basic assumption that there is a true tree for the taxa under study and that the ancestor-descendant relationships among these taxa may be represented as a tree (Estabrook, 1983). Even this assumption is not always true. Estabrook (1983) identified hybridisation and gene flow as potential pitfalls. In cases where a true phylogeny can be assumed, however, compatibility methods use patterns of compatibility and incompatibility in data to attempt to identify those characters representing the true phylogeny. True characters, as defined by Estabrook (Estabrook *et al.*, 1975; Estabrook, 1984), cannot conflict, and thus all uniquely derived characters must be mutually compatible (see Meacham, 1984). Noisy characters are randomly distributed among taxa, and are likely to be incompatible not only with the true signal, but also with each other. It is hoped that the true phylogenetic signal will be represented by a large block of compatible characters, while noisy characters will have a

generally low compatibility with the rest of the matrix. Large groups of compatible characters do not necessarily represent the true phylogeny, however. Compatibility can also be the result of logical dependence between characters, or functional, ecological or developmental correlation (e.g. see Meacham, 1984). A useful property of compatibility methods is that they are tree-independent, meaning they are an a priori assessment of the data rather than a measure of fit to a phylogenetic hypothesis. This means that their results are partly independent of any parsimony analysis carried out on the same data, so that their results will not necessarily agree. Although it seems that parsimony is generally quite successful at identifying phylogenetic patterns in data (e.g. Wiens and Hillis, 1996; but see Harcourt-Brown, 2002), compatibility, used in conjunction with parsimony can provide useful supplementary information. If the two methods do agree, support for the hypothesis they propose may be considered strong. However, if the two methods disagree, one or both methods must be producing erroneous results, suggesting that the data should be investigated further and homology assessments reviewed.

Algorithms for identifying incompatibility in raw data (in the form of taxon x character matrices) have been devised, although some compatibility-based computer programs are useful only with binary data. This is because identifying incompatibility between binary characters is simple. Le Quesne (1969; 1972) observed that, given two binary characters, A and B, each with two character states, 0 and 1, only four combinations of character states are possible (A_0, B_0 ; A_0, B_1 ; A_1, B_0 and A_1, B_1). If all four of these state combinations are present in the taxa selected for analysis, then the characters cannot be mapped onto the same tree without at least some homoplasy. So, at least one (or possibly both) of the characters is not uniquely derived. It must be stressed, however, that if fewer than four of the state combinations are present, the two characters are compatible, but neither is necessarily uniquely derived. Identifying incompatibility between multistate characters is slightly more complex. Farris (1973) introduced a method by which multistate characters could be split into a number of binary characters, allowing the application of Le Quesne's methods (Estabrook *et al.*, 1976; Le Quesne, 1982). However, the simplest method for finding incompatibility in multistate characters, devised by both Estabrook and Landrum (1975) and Fitch (1975) and mathematically proven by McMorris (1975; Estabrook and McMorris, 1977), involves the creation of a character state by character state matrix for the two characters in question (see Estabrook, 1983 for a simplified description). In these matrices, henceforth called state combination matrices, the states of the two characters in question are plotted on the two axes. Cells in the matrix are marked

(here with an X) if the corresponding character state combination is present in at least one taxon. For example, if one or more taxa possess state 0 for one character (character A) and state 1 for another (character B) (both characters having three states), then the following matrix is produced and the cell corresponding to the state combination A_0, B_1 is marked with an X.

		Character A		
		0	1	2
Character B	0			
	1	X		
	2			

The state combinations of all taxa included in the analysis are inserted into the state combination matrix in this way. If at least four of the Xs in the matrix can be joined by horizontal and vertical lines to form the corners of at least one continuous loop (i.e. a path of horizontal and vertical lines that can be followed between Xs and returns to a cell that has previously been visited without reversing direction), then the characters are incompatible. It is not necessary for all Xs to be part of the loop, but all corners of a loop must coincide with an X. Loops may cross themselves in cells without Xs, but these are not regarded as corners. The following are two examples of pairs of incompatible characters with loops present indicated in red.

		Character A				
		0	1	2	3	4
Character B	0				X	
	1			X		X
	2	X				
	3		X			
	4			X		X

		Character A				
		0	1	2	3	4
Character B	0	X			X	
	1		X		X	
	2	X		X		
	3		X			X
	4			X		X

Most compatibility measures employ a strict cut-off, so that a pair or group of characters are simply classified as compatible or incompatible. This limits the power of current compatibility tests. Estabrook *et al.* (1975) defined a ‘true cladistic character’ as a divergent character that is a partial estimate of cladistic history and should meet the following three criteria:

- 1) A character state should contain its own most recent common ancestor.
- 2) If one taxon is the ancestor of a second, then the state of which the first is a member must be equal, or should be ancestral in the character state tree (a ‘tree’ showing the proposed ordering of character states), to the state of which the second is a member.
- 3) If one character state is ancestral to another in the character state tree, then the most recent common ancestor for the one state should be ancestral to the most recent common ancestor of the other.

As reiterated by Meacham (1984), ‘true cladistic characters’, as defined by Estabrook *et al.* (1975), cannot conflict with one another, and therefore all true cladistic characters must be mutually compatible. This logic cannot be faulted, but it might be taken to suggest that any characters that conflict with the true signal are not useful in phylogenetic reconstruction. On the contrary, many such are useful indicators of phylogeny. It appears that it is not uncommon for homoplasy (in the form of convergences and/or reversals) to occur in characters that provide phylogenetically useful (and important) information. Such characters, although providing some misleading evidence, are still desirable, because they may provide evidence of relationships that are not identified by other characters. Parsimony analyses aim to use evidence supplied by additional characters to reject homoplastic relationships, so that any useful information in a character is still utilised (e.g. see Farris *et al.*, 1996; Källerajő *et al.*, 1999). Most methods of compatibility analysis, however, will show such characters as incompatible with many ‘true cladistic characters’

because of their homoplastic change(s). The hope in such analyses is that if a character possesses few homoplasies, it will still be compatible with many other characters, including some ‘true cladistic characters’, whereas characters that are no more than noise are unlikely to be compatible with any other characters. However, in practice it is possible that many characters that contain useful phylogenetic information will be classified as ‘poor’ and will not be distinguished from noise.

In order to attempt to further separate noise from homoplastic characters that contain phylogenetically useful information, it may be useful to quantify how incompatible two incompatible characters are. One possible measure is fuzzy compatibility, described for the first time here.

4.2.1 Fuzzy compatibility

Fuzzy compatibility attempts to quantify the level of incompatibility between two or more characters. It is a measure of the minimum number of taxa that must be rescored in order to make the characters compatible (or the minimum number of taxa that must have been misscored in creating the original data matrix). Guise, Peacock and Gleaves (1982) described a method based on a similar idea that they called “labelling”. Their method identified incompatible pairwise binary character comparisons in which only one taxon possessed one of the four possible state combinations. They suspected that the primary homology assessment of one of the characters for that taxon was wrong. All such potentially homoplastic scorings were recorded and at the end of the analysis the particular score most likely to be homoplastic could be identified. Fuzzy compatibility differs from this method in that it calculates the minimum number of taxa that must be rescored to make all incompatible character pairs compatible, not just those in which a single taxon must be rescored. A similar method, called the minflip algorithm (Chen et al., 2003), was devised for supertree construction. Minflip uses heuristic algorithms to flip states in order to find the smallest number of states in the matrix that must be changed in order to make the entire matrix compatible.

In the case of two binary characters, calculating fuzzy compatibility is simple. As stated above, binary characters can only be incompatible if at least one taxon possesses each of the four possible character state combinations for the two characters. Therefore, if the two characters are incompatible, the minimum number of taxa that must be rescored to make the characters compatible is equal to the number of taxa scored as possessing the least common character state combination.

With multistate characters, finding the smallest number of taxa that must be rescored to make the characters compatible is more complicated. Again, the simplest method involves the use of a character state combination matrix as described above. However, in this case, instead of simply marking any state combinations possessed by at least one taxon with an X in the matrix, the number of taxa possessing each state combination is inserted into the cell. For example, consider the following two characters, each of which exhibits four states:

	Char A	Char B
Taxon 1	0	0
Taxon 2	0	0
Taxon 3	0	0
Taxon 4	0	0
Taxon 5	0	1
Taxon 6	1	1
Taxon 7	1	1
Taxon 8	1	2
Taxon 9	2	2
Taxon 10	2	2
Taxon 11	2	3
Taxon 12	3	3
Taxon 13	3	1
Taxon 14	3	3

This combination of character states gives rise to the following state combination matrix:

		Character A			
		0	1	2	3
Character B	0	4	0	0	0
	1	1	2	0	1
	2	0	1	2	0
	3	0	0	1	2

In this example, cells containing non-zero values form the corners of a single continuous loop, indicating that the characters are incompatible. Two cells, A₀, B₀ and A₀, B₁ are not integral parts of any loop, and therefore are not the cause of the incompatibility in the data. The taxa possessing these state combinations can be disregarded when identifying the minimum number of taxa that need to be rescored to make the characters

compatible. For the characters to be made compatible, the loop must be broken. It is a rule that **any single loop will be broken by removing any one of its corners. Which corner is removed is not important.** Therefore, **the minimum number of taxa that must be recoded to make two characters compatible is equal to the smallest number of taxa present in a cell at the corner of the loop.** In the above example, there are three equally minimal ways to make the characters compatible. State combinations A_1, B_2 ; A_2, B_3 and A_3, B_1 are all corners of the loop, and are all present in only one taxon. By referring back to the data matrix, it can be shown that taxon 8 (A_1, B_2), taxon 11 (A_2, B_3) or taxon 13 (A_3, B_1) can be rescored to make the two characters compatible. Therefore, the minimum incompatibility value for these two characters is one.

It is possible for multistate characters to produce state combination matrices that contain more than one loop. For example:

		Character 1			
		0	1	2	3
Character 2	0	3	1	0	0
	1	2	1	0	1
	2	0	1	2	0
	3	0	0	1	2

In this example, there are two loops that must be broken to make the characters compatible, shown in red and green. In order to break two loops, the smallest number of corners that must be removed is two. Methodologically, this can be achieved by first removing the corner of either loop that corresponds to the lowest number of taxa (i.e. in this example, one taxon must be rescored, either at position A_1, B_0 ; A_1, B_1 ; A_1, B_2 ; A_2, B_3 or A_3, B_1). When one corner is removed, the loop in which it was a corner is broken, and the remaining corners of that loop can be disregarded, provided they are not also corners of further loops. In the above example, if the taxon possessing state combination A_1, B_0 is rescored, the taxa at A_0, B_0 and A_0, B_1 can be disregarded. The taxon at A_1, B_1 , however, still forms the corner of the second loop, and therefore cannot be disregarded.

		Character 1			
		0	1	2	3
Character 2	0	3	-	0	0
	1	2	1	0	1
	2	0	1	2	0
	3	0	0	1	2

Removal of any corner when two loops are present will always leave a single loop, no matter which corner is chosen. **It is impossible to break two loops by removing only one corner.** In the above example, position A₁, B₁ is a corner of both loops. It therefore may seem that recoding the taxa possessing this state would break both loops simultaneously. However, as shown below, removing this ‘corner’ simply creates a single loop that crosses itself at the position vacated by the recoded taxon.

		Character 1			
		0	1	2	3
Character 2	0	3	1	0	0
	1	2	-	0	1
	2	0	1	2	0
	3	0	0	1	2

Once the first loop is broken, the process can be repeated to break the second loop. Again, the minimum number of taxa that must be rescored to break this loop is one (at position A₁, B₁; A₁, B₂; A₂, B₃ or A₃, B₁). Therefore, the minimum number of taxa that must be rescored to make these characters compatible is two, one to break each of the two loops. **one corner must be recoded for each loop present in a state combination matrix in order to make the two characters compatible.**

4.2.2 Uses of Compatibility

Simply measuring the number of pairwise incompatibilities for each character in a dataset is generally not considered sufficient for most purposes. Often, incompatibility values are used to identify the relative ‘strength’ of characters in a dataset. This necessitates comparisons of incompatibility values between characters. Simply counting

the number of pairwise incompatibilities for a character in a dataset has been recognized as being problematic for such comparisons (Le Quesne, 1969; 1972). Some characters are more likely than others to be compatible by chance alone, so that they may misleadingly appear ‘stronger’ than those other characters when in truth they are not. It is possible to calculate the probability of two binary characters with no missing data being incompatible by chance alone using the following formula derived from that published by LeQuesne (1972; see also Meacham, 1984).

$$P = 1 - \frac{n_0! \times (n_T - n_s)!}{n_T! \times (n_0 - n_s)!} - \frac{n_1! \times (n_T - n_s)!}{n_T! \times (n_1 - n_s)!},$$

where 0 and 1 are the two states of the two binary characters, A and B; n_s is the state assigned to the smallest number of taxa (i.e. the smallest number of 0s or 1s in character A or B); n_0 and n_1 are the number of states 0 and 1 in the character not including n_s ; and n_T is the total number of taxa.

For example, consider two binary characters scored for ten taxa with no missing data. If both characters have eight taxa scored as state 0 and two taxa scored as state 1, then the probability of their being incompatible by chance alone can be calculated using these values with the formula: $n_s = 2$; $n_0 = 8$; $n_1 = 2$ and $n_T = 10$.

$$P = 1 - \frac{2! \times (10 - 2)!}{10! \times (2 - 2)!} - \frac{8! \times (10 - 2)!}{10! \times (8 - 2)!}$$

$$\therefore P = 1 - \frac{80640}{3628800} - \frac{1625702400}{2612736000}$$

$$\therefore P = 1 - \frac{1}{45} - \frac{28}{45} = \frac{16}{45} = 0.3555.$$

Similarly, if one character has five 0s and five 1s and the other eight 0s and two 1s, then the probability of these two characters being incompatible by chance alone is 0.5555, and if both characters have five 0s and five 1s, the probability of their being incompatible by chance alone is 0.9921. This shows clearly that the characters most likely to be incompatible by chance alone are those with more equal numbers of the two states, and that those with only two taxa possessing one of the states are the least likely to be incompatible by chance alone. This may cause problems when comparing compatibility scores between characters, because characters with an unequal number of each state are more likely to be compatible with other characters by chance alone. Therefore, random, noisy, phylogenetically uninformative characters with unequal numbers of each state are more likely to be wrongly considered ‘strong’ characters. A number of methods have been

devised in an attempt to counteract this problem. Some of these methods are discussed below.

4.2.2.1 The Coefficient of Character State Randomness (CCSR)

The CCSR (Le Quesne, 1969; 1972; 1982) is a measure that attempts to correct the relative ‘strength’ of characters by taking into account the probability of such a character being compatible by chance alone. The CCSR is simply a ratio of the observed number of pairwise incompatibilities of a character with the rest of a matrix divided by the number of pairwise incompatibilities expected by chance alone for a randomly permuted version of that character. For binary characters it is relatively simple to calculate the expected value mathematically (see above). However, for more complex, multistate characters the exact expected number of incompatibilities is more difficult to calculate, and is therefore often approximated using the average number of incompatibilities of a large number of random permutations of the character. The character must be permuted as opposed to randomised so that the numbers of individual character states in the original character are preserved. In cases where some taxa cannot be scored for a character, it is recommended (Wilkinson, 1995) that missing data should be held constant during permutation to reduce the number of variables in the permutation test. The CCSR is calculated for each character in the data matrix, and a character-by-character matrix can be constructed containing the pairwise CCSR values (Le Quesne, 1969). The total CCSR value for each character can then be calculated by simply summing the values for all pairwise character comparisons containing that character. CCSR analyses can be carried out using either the normal or fuzzy compatibility methods of measuring pairwise incompatibilities between characters.

4.2.2.2 The Normal Deviate (NDev)

The NDev (Le Quesne, 1972; 1982) is similar to the CCSR, but has the advantage of showing whether the difference between the observed and expected number of incompatibilities for a character is statistically significant. It is a measure of where the observed number of incompatibilities falls in terms of the number of standard deviations away from the mean of the distribution of expected values. If the value is positive, then the observed number is lower than that expected by chance, and vice versa. The NDev for binary characters can be calculated using the following formula:

$$NDev = \frac{P_s - N_x - 0.5}{\sqrt{\left(\frac{P_s(n_v - P_s)}{n_v}\right)}}$$

where P_s is the sum of all values of P (the probability of two characters being incompatible by chance alone) for comparisons of the character with all other characters; n_v is the number of valid comparisons (i.e. comparisons with characters that are not phylogenetically uninformative); and N_x is the observed number of character pairings in which all four character state combinations are found. The sign of the last term in the numerator becomes '+' if $P_s < N_x$ (Le Quesne, 1989). This is a modification (Yates' correction) of the formula presented by Le Quesne (1972), which corrects for problems associated with analysing small datasets.

4.2.2.3 Le Quesne Probability (LQP)

The LQP measure was devised independently by both Wilkinson (1992; 1995; 1997a) and Meacham (1994), who called it the Frequency of Compatibility Attainment. Here the name LQP is employed because it was proposed first. LQP is a simple randomisation method for testing the null hypothesis that a character is no more compatible with other characters in a data matrix than is a random, phylogenetically uninformative character. The method involves first calculating the number of pairwise incompatibilities in the dataset for each character in its original form, and then doing the same for a number (usually 99 or 999) of random permutations of that character. The LQP is the probability of a randomly permuted character having an equal or lower number of total incompatibilities (the sum of all its pairwise incompatibilities) than the original character. It is calculated by simply counting the number of random permutations of a character that have the same or fewer incompatibilities with the dataset than the original character. This value is then divided by the total number of replicates (number of permutations + 1) to find a p-value between 0 and 1 with which to test the null hypothesis. Characters that are highly compatible with the matrix will have an LQP value close to 0, whereas characters representing random noise should have an LQP value closer to 0.5. The LQP is advantageous over other compatibility measures, such as the CCSR, because, like the normal deviate, it gives a measure of the level at which a character is significantly better (or worse) than random noise. It is especially useful because it sets up a significance test for which the significance level can be defined by the user. As with most randomisation tests, the cut-off between significance and lack of significance is usually

taken at the 5% level. The LQP test is a one-tailed test, because its aim is to find characters better than random, not just different from random. This means that any LQP values below 0.05 indicate that a character is significantly more compatible than random at the 5% level.

4.2.2.4 Clique Analysis

A collection of mutually compatible characters that is not a subset of a larger collection is termed a clique (Estabrook *et al.*, 1976; 1977) from the equivalent graph theory concept. Estabrook *et al.* (1976) provided proof that any collection of pairwise compatible, ordered binary characters are mutually compatible. However, this concept does not hold for unordered multistate characters (Fitch, 1975; 1977) or characters that include missing entries (Wilkinson, 1994b) or uncertainties. Wilkinson (1994a) specified the definition of clique analysis as “the discovery of cliques” in order to distinguish it from other compatibility methods, with which the term had often previously been synonymized. The original hypothesis behind clique analysis is that because all ‘true’ characters must belong to the same collection of mutually compatible characters, then the largest such collection contains them (Estabrook *et al.*, 1977). In other words, the largest clique is assumed to contain the ‘true’ characters. Estabrook *et al.* (1977) proposed simply identifying the largest clique and using the tree that it supports as the best hypothesis of phylogeny. Unfortunately, as noted by Wilkinson (1994a), largest cliques are often too small and the characters they contain too similar to resolve large sections of the tree (see also Felsenstein, 1982). They can also produce phylogenetic hypotheses that are inconsistent with the results of maximum parsimony methods (Wilkinson, 1994a). However, Wilkinson (1994b) showed that clique analysis and parsimony analysis are equivalent when a dataset does not include any characters with a maximum length greater than two, a situation which can be attained by use of the three-taxon statement method of character representation (Nelson and Platnick, 1991). A number of attempts have been made to devise methods using clique analysis as a starting point to find maximally parsimonious trees for a dataset without the need for conventional, time-consuming parsimony analysis (e.g. Penny, 1982; Lorenzen, 1993). Unfortunately, most of these methods appear to be impractical when large numbers of characters are present in the data, where conventional parsimony analyses are far more efficient (Wilkinson, 1994a).

4.2.2.5 Boildown

Le Quesne (1969; 1972) introduced a number of methods by which noisy characters can be eliminated from a data matrix in an attempt to increase signal. These methods included:

- Eliminating the character or characters with the highest incompatibility count.
- Assuming that the character with the lowest incompatibility count, lowest CCSR or highest positive NDev value is uniquely derived and eliminating characters incompatible with that character.
- Assuming the largest group of completely correlated characters (the largest clique) is uniquely derived and eliminating characters incompatible with it.
- Eliminating all characters with an NDev value below +2 or +3.

Whichever method is employed, once the ‘worst’ character(s) are identified and eliminated, the procedure can be repeated until all incompatibility has been removed from the data (Le Quesne, 1969; 1972). Gauld and Underwood (1986) first introduced the term ‘boil down’ for an analogous procedure that they used to identify groups of mutually compatible characters. They calculated CCSR values (which they alternatively called randomness ratios), deleted the character with the worst value (highest CCSR) and then recalculated. This procedure is repeated until no incompatibility remains. All characters left in the dataset are pairwise compatible, and, if all are binary and contain no missing data, mutually compatible, although not necessarily an entire clique. These methods are effectively tree-independent methods of extreme character weighting, in which characters deemed the ‘worst’ during each repetition are assigned a weight of zero. Many compatibility measures can be used to select the characters for removal at each stage of a boildown procedure, including the CCSR, NDev and LQP. Fuzzy compatibility can also be used in a boildown procedure.

As previously intimated, a boildown procedure is usually halted when no incompatibility remains in the matrix. However, the remaining clique may not be sufficient to resolve the phylogeny of the taxa under study (Felsenstein, 1982; Gauld and Underwood, 1986). Gauld and Underwood (1986) suggested ranking characters not in the maximum clique based on their compatibility with it, and using these to resolve the rest of

the tree. As an alternative to using methods for adding characters to a maximum clique to improve resolution (see also Penny, 1982; Lorenzen, 1993), it may be preferable to halt the boildown before all incompatibility is removed, in order to preserve signal that may be present in characters that show low incompatibility with the matrix. This is effectively an attempt to remove noise without removing too much useful phylogenetic signal. One possible way of doing this is by using the LQP as the measure of character strength for the boildown. The boildown procedure can then be stopped when all characters remaining in the matrix are significantly more compatible with each other than is a random permutation of that character (that is, when the null hypothesis for the LQP test can be rejected for all characters remaining in the matrix).

Alternatively, the effects of every step of the boildown process can be monitored to study how character elimination affects the tree produced by analysis of the data. Assuming that the true phylogeny of the taxa is represented by a large number of relatively compatible characters and that noisy characters are relatively incompatible, the signals provided by noisy characters should be in conflict and should not produce a combined signal strong enough to overpower the true tree. If this is true, by removing incompatible, noisy characters using the boildown, the resulting phylogeny should not change significantly other than possibly to lose resolution in areas of the tree in which initial resolution was provided solely by poor characters. If, however, there is a large amount of change in the topology through the boildown process, this suggests that the characters being removed by the boildown are playing a major role in the topology of the tree produced by parsimony analysis of the complete data. This could be due to the presence two or more competing signals in the data, possible caused by convergences due to functional similarities between unrelated taxa. In such cases the original data might be re-examined and hypotheses of homology checked.

Other methods similar to the boildown, by which characters are reweighted based on compatibility values such as the CCSR, have been presented by Penny and Hendy (1985; 1986) and Sharkey (1989). Sharkey (1994) proposed an improved method of character reweighting based on what he termed discriminate compatibility, and described a reduction routine similar to the boildown, which could be used to build trees. His discriminate compatibility method relied on a new procedure for determining the polarity of characters. However, as noted by Wilkinson (1997b), this method is feasible only with a dataset containing a small number of characters. When used to polarise binary characters there are 2^n possible polarisations to consider, where n is the number of characters. Therefore, with a

moderately sized matrix of just 100 characters, 2^{100} (= more than 10^{30}) alternative polarisations must be examined, rendering the process impractical even for powerful computers (Wilkinson, 1997b).

4.2.2.6 *Boildown Bootstrap*

Bootstrap analyses (Felsenstein, 1985) are widely used to investigate the strength of support for clades in a resulting phylogeny. Felsenstein's (1985) bootstrap involves producing a number of datasets of identical proportions (the same number of characters and taxa) to the original data using a resampling with replacement strategy, in which characters to be included in the bootstrap replicate dataset are selected randomly from the original data. When a character is added to the new dataset it is not excluded from being selected again, so that some characters in the original data can be in the bootstrap replicate dataset more than once, and others not at all. Each bootstrap replicate is then analysed using parsimony, and the MPTs of each replicate are saved. A majority-rule consensus tree is then produced from the MPTs of the replicates, in which the majority-rule values on each node of the tree represent the bootstrap support for that clade.

Here, the same methodology is, for the first time, applied to the boildown procedure in an attempt to answer the unanswered question of when to halt a boildown procedure. This method, which measures the support for clades present in the trees produced at each stage of the process, is named the boildown bootstrap. Two alternative types of boildown bootstrap can be performed.

The first type, called type 1 boildown bootstrap simply involves carrying out bootstrap analyses on the reduced datasets produced at each stage of a boildown procedure. The bootstrap trees produced should include the best-supported compatible groups (in terms of compatibility with the 50% majority rule bootstrap tree) present in less than 50% of the saved trees. This produces one bootstrap tree for each boildown step.

The type 2 boildown bootstrap, is more complex, more labour intensive and far more time consuming. First, a number of bootstrap replicate data matrices of the entire data matrix to be analysed are produced using resampling with replacement. A boildown is then carried out on each of these datasets. After each character removal during these boildowns a parsimony analysis is carried out on the reduced dataset and the resulting MPTs saved. Therefore, at the end of all boildown analyses a set of MPTs for each replicate has been saved for each character removal step. For example, in the first boildown, when one character is removed, the MPTs of the analysis of this reduced dataset are saved to a

treefile. In subsequent boildowns, the MPTs of analyses carried out after one character has been removed are appended to this treefile. Each set of MPTs in the tree file is then assigned a weight inverse to the number of trees that it contains. This ensures that there is not a bias towards sets of MPTs containing more trees when a consensus of all sets is taken. For example, each tree in a set of 20 MPTs is assigned a weight of 1/20, while trees in a set of five MPTs are assigned a weight of 1/5 each. Finally, a majority-rule consensus tree with compatible resolutions under 50% included is calculated for each of the treefiles, so that there is one majority-rule tree for every character removal stage of the boildown. These majority-rule trees are equivalent to bootstrap trees, with the majority-rule values on nodes equivalent to bootstrap values.

Whichever type of boildown bootstrap is used, the point during the boildown at which the tree produced by parsimony is best supported can be assessed. By summing the majority-rule values of all nodes on each bootstrap tree (= majority-rule consensus tree for type 2 boildown bootstraps), a measure of the strength of support for that tree is obtained. This is essentially a measure of the total bootstrap value for the tree. If the bootstrap tree with the highest total bootstrap value is not produced by analysis of the complete data, it can be suggested that up to this point in the process the boildown method has removed noise from the data without sacrificing signal.

4.3 Potential Problems with Compatibility Methods

4.3.1 Uncertainty and Polymorphism Within Leaves

All current compatibility methods suffer from an inability to cope with leaves containing uncertainty (leaves for which the state present for a character is uncertain between two or more possibilities) or polymorphism (or leaves in which taxa are present containing two or more states of a character). These two types of intra-leaf variation cause problems in compatibility analysis for two reasons.

The problem regarding the treatment of uncertainty lies solely in the efficiency of any compatibility analysis program. In theory it is possible to run a compatibility analysis in which taxa coded as uncertain for two or more states can be included in all possible configurations to see which is most compatible and which most incompatible. From this, a range of compatibilities for each character can be recorded. With only a few uncertain taxa this is not a problem, but when a larger number of taxa are coded as uncertain for a number of characters, the number of permutations of possible character states that must be analysed

increases exponentially. This causes problems for the efficiency of the analysis, leading to extremely long calculation times for even relatively small datasets. Currently no compatibility programs treat uncertainties in this way. Instead, as an approximation, they generally replace taxa coded with uncertainties with missing data (?), which is equivalent to uncertainty between all states. Although this situation is not perfect, treating uncertainty as missing data should, in most cases, lead to analytical results very similar to those that would be found if uncertainties could be treated individually.

A more difficult problem to solve is that of intra-leaf polymorphisms. These represent instances where the taxa included as a leaf in an analysis possess more than one of the states of a character. This is a major headache in compatibility analysis, because it is possible that a single character containing polymorphic leaves can essentially be incompatible with itself, and therefore in any pairwise compatibility test would seem incompatible with every other character. If a leaf is scored as possessing two states of one character and those states are also both present in one or more other taxa, it is not possible to draw a tree on which the character can be mapped without homoplasy. This phenomenon is here called intra-character incompatibility, and must be the result of errors in homology assessment, errors in the determination of the constituent taxa of leaves or of true homoplasy in the character. There are two ways in which polymorphism within a leaf can occur. The first is a situation where two or more states of a character are identified within a single individual. Such characters fail Patterson's (1982) conjunction criterion of homology assessment and should not be included in analyses (see chapter 1). The second is a situation where a leaf is made up of more than one taxon, and these groups vary in their scoring for an individual character. By creating a leaf that contains more than one taxon, authors are effectively proposing a hypothesis of relationships that they think is strong enough to assume without further testing. This is necessary when analysing relationships of basal groups, because otherwise an analysis would become too large for current methods to compute. For example, in order to evaluate the relationships within the amniotes it is not feasible to include every known amniote in a phylogenetic analysis. Decisions must be made a priori about which taxa to include as individual leaves and which to group together. Such decisions are based on knowledge gained from previous analyses, so that well-justified clades containing many taxa may be used in more inclusive analyses. In such circumstances, authors are faced with a choice of how to code these multi-taxa leaves. They can be coded either on the basis of a groundplan version of the taxa included in the leaf or using one or more exemplars from the group (see Yeates, 1995). It is only when

more than one exemplar is used to code these leaves that polymorphisms can result. If taxa are incorrectly grouped, so that unrelated taxa are forced together into one leaf, then it is not surprising that homoplasy results. A second possible explanation for intra-character homoplasy in characters containing polymorphisms is error in homology assessment. Patterson (1982), in his discussions of primary and secondary homology, described the conjunction criterion, which must be passed in order for the states of a character to be considered homologous. This criterion is that two states of one character cannot be present in a single taxon. By this, Patterson was suggesting that two states of a character cannot be homologous if a specimen is found possessing both states. Although in the case of polymorphism the two states are not always possessed by a single specimen, the problems caused are similar. The possession of more than one state by a leaf (whether that leaf is a single specimen or a group of taxa) usually leads to homoplasy in that character, so that homology of the states cannot be concluded. The only exception to this occurs when at least all but one of the character states in the polymorphism are unique in that taxon.

At present, in most compatibility analyses, polymorphic taxa are rescored as missing data. This is not a satisfactory method of coping with the problem, because it discounts homoplasy within characters containing polymorphisms, and artificially increases their 'strength'. In extreme cases, where many leaves are polymorphic, a character could contain large amounts of intra-character homoplasy, yet seem compatible with all other characters when these polymorphisms are replaced with missing data. Unfortunately, the alternative is to treat characters containing intra-character homoplasy as incompatible with all other characters. In the case of fuzzy compatibility analyses, a possible alternative is to replace polymorphisms with missing data, but increase the fuzzy incompatibility value to acknowledge the weakness of the character. For example, characters containing intra-character homoplasy could have 1, or the minimum number of recodings necessary to make the character compatible with itself added to each of their pairwise fuzzy incompatibility scores.

4.3.2 Logical linkage and Inapplicable data

Characters that are completely logically linked cannot be incompatible. This can cause problems in compatibility analyses, because, as in parsimony analyses, the linkage of these characters effectively increases the weights given to any hypotheses of relationship they support. In compatibility analyses, any group of completely linked characters will be

mutually compatible, which may artificially increase their own compatibility values and decrease those of ‘true’ characters.

In Chapter 2, I discussed the problem in character construction of characters relating to data that are inapplicable to some taxa under study. With the ‘missing’ method of coding inapplicable data, which is the method I suggested was preferable for parsimony analyses, there is usually one character describing the presence or absence of a structure and one or more characters describing attributes of that structure in taxa that possess it. Taxa that do not possess the structure are scored as unknown for these characters. These characters cannot be incompatible with the character describing presence or absence, leading to problems in compatibility analysis similar to those of logically linked characters. Therefore, in compatibility analysis it may be better to include inapplicable data characters using the ‘multistate’ coding method (see chapter 2).

Gauld and Underwood (1986) suggested another possible solution. They dealt with multistate characters in their data by splitting them into additive binary characters and labelled each of these additive binary characters so that they would not be compared with one another in analyses. Such a method could also be applied to inapplicable data. However, although their method does stop linked characters upweighting each other’s compatibility values, it does not necessarily correct for the effect a large number of additive binary characters would have on other characters in the data. If a large block of linked characters is present in the data, other characters in the data that are compatible with the original multistate character from which the additive binary characters were produced will show an artificially high compatibility relative to a character that is incompatible with the original multistate character. There are, however methods that could potentially negate this bias, for example, by weighting the results accordingly.

4.3.3 Tree Balance

A recurring problem in many phylogeny reconstruction methods is that of a bias towards unbalanced tree topologies. It has been shown in empirical (Mooers *et al.*, 1995) and simulation studies (Huelsenbeck and Kirkpatrick, 1996; Harcourt-Brown, 2002) that parsimony and other reconstruction methods favour unbalanced trees. Guyer and Slowinski (1991; 1993) provided an explanation for this bias in parsimony methods using the “equiprobable model” (that all tree topologies are equally probable). Under this model, they showed that there are more arrangements of taxa on unbalanced tree topologies, effectively meaning that there are more possible unbalanced topologies. Huelsenbeck and

Kirkpatrick (1996) provided a simple example. Consider a tree containing eight taxa. There are 40,320 ($8!$) ways that the taxa can be permuted onto (the leaves of) the tree. However, more of these ways of placing taxa onto the leaves are isomorphic (identical topologies, but with nodes rotated) on symmetrical trees. In fact, with 8 taxa, there are 20,160 ($\frac{n!}{2}$, where n is the number of taxa, as long as it is possible to create a completely balanced tree from that number of taxa) different rooted maximally unbalanced trees, whereas there are only 315 ($\frac{n!}{2^{(n-1)}}$) different maximally balanced trees. With so many more possible unbalanced topologies (64 times as many with 8 taxa), it is highly probable that a tree chosen at random will be unbalanced. This means that any noisy signal that has an effect on the resolution of the most parsimonious tree is likely to introduce more imbalance to that tree.

In recent years, a major emphasis has been placed on attempting to create a tree of life in which all known taxa can be placed and their affinities known. To do this using traditional phylogenetic analysis methods is not currently feasible due to the immense computing time required, so alternative methods have been sought. One solution is to use supertree methods to combine results of smaller analyses. These methods take the consensus trees output by analyses (the source trees) and attempt to join them together, as long as they have at least three (or two plus the root in rooted trees) taxa in common. The most commonly implemented supertree method is matrix representation using parsimony (MRP) (Baum, 1992; Ragan, 1992). This, like many other supertree methods, joins trees together by first representing them as a taxon by ‘pseudocharacter’ matrix representation (MR) (Farris, 1973). Each node of each source tree is coded as a single ‘pseudocharacter’ where taxa within the clade defined by the node are scored as 1 and those outside as 0. Any taxa in the matrix that are not present in the source tree are scored as unknown (?) for all ‘pseudocharacters’ based on that tree. MRP then employs parsimony to create one or more supertree(s) from the data matrix. However, the MRP method has been much criticised (Purvis, 1995; Wilkinson *et al.*, 2001; Goloboff and Pol, 2002). Among other problems, it suffers from two major biases that make its results unreliable at best. These biases are: that it produces output trees more similar to large source trees (those with more taxa) (Purvis, 1995), and to unbalanced source trees (Wilkinson *et al.*, 2001). The bias towards large source trees is simply due to the larger number of potential nodes that are present in trees with a larger number of taxa. This leads to more ‘pseudocharacters’ in the MR of large

trees, and therefore adds weight to the hypotheses supported by large source trees. The bias towards unbiased trees was illustrated nicely by Wilkinson *et al.* (in prep). They showed that, even using the same taxon list in the source trees, if a completely balanced and a highly incongruent completely unbalanced source tree are submitted to MRP, then the output supertree is more similar to the unbalanced source tree. The reason for this bias is not so simple to understand as the bias towards large source trees. Thorley and Wilkinson (2003) noted that one major problem with the MR method is that the tree-to-supertree distance that is the basis of the methods objective function does not obey the symmetry axiom. This means that the distance between each source tree and the supertree produced from them is not always the same. In Wilkinson *et al.*'s (in prep) example, if the pseudocharacters produced by MR of the balanced tree are mapped onto the unbalanced tree, the fit (in parsimony steps) is not the same as if the characters produced by the unbalanced tree are mapped onto the balanced tree. Wilkinson *et al.* (in prep) suggested that this asymmetric distance measure is responsible for the bias of MRP towards producing unbalanced supertrees

One artefact of the MR method is that more balanced source trees lead to MRs with a high ratio of primitive states to derived states, whereas unbalanced trees lead to an equal number. This is because balanced trees are formed by bifurcating nodes so that each node contains half the number of taxa contained by the previous node in the tree (the node further down the tree). Unbalanced trees are formed in a hierarchical way, where each node simply contains one taxon less than the previous node. For example, take the two source trees used in the example of Wilkinson *et al.* (in prep) (Fig 4.1 a and b). The matrix representations of these trees are shown in tables 4.1 and 4.2.

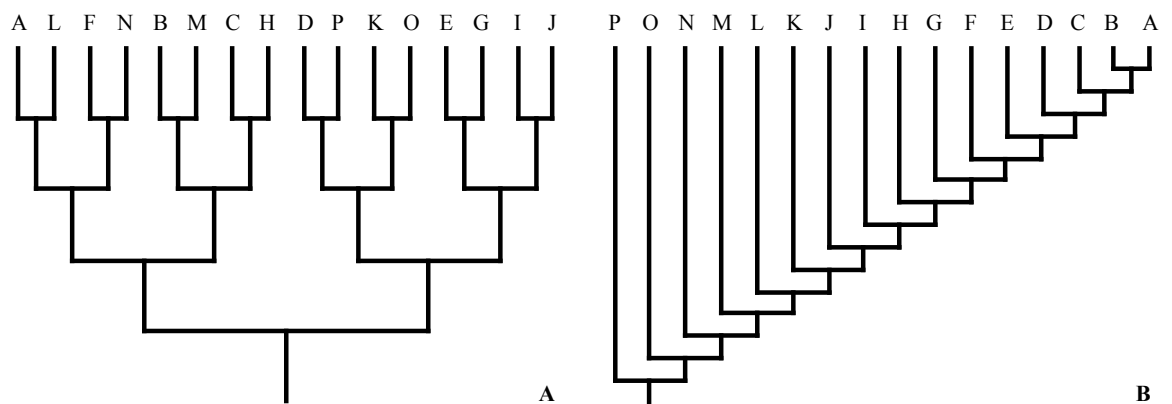


Figure 4.1. The two incongruent source trees used by Wilkinson *et al.* to illustrate the problem of a bias towards unbalanced tree topologies in MRP analysis. A) A completely balanced tree. B) A completely unbalanced tree.

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
A	1	1	1	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	1	1	0	0	0	0	0	0	0
C	1	0	0	0	1	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	1	1	0
F	1	1	0	1	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	1	1	0
H	1	0	0	0	1	0	1	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	1	0	1
J	0	0	0	0	0	0	0	0	0	0	1	0	1
K	0	0	0	0	0	0	0	1	0	1	0	0	0
L	1	1	1	0	0	0	0	0	0	0	0	0	0
M	1	0	0	0	1	1	0	0	0	0	0	0	0
N	1	1	0	1	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	1	0	1	0	0	0
P	0	0	0	0	0	0	0	1	1	0	0	0	0

Table 4.1. Matrix representation of a maximally balanced tree containing 16 taxa (tree A in Fig. 4.1). Matrix representation contains 160 (77%) primitive states (0s) and 48 (23%) derived states (1s).

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
A	1	1	1	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	1	1	1	1	1	1	1
C	0	1	1	1	1	1	1	1	1	1	1	1	1
D	0	0	1	1	1	1	1	1	1	1	1	1	1
E	0	0	0	1	1	1	1	1	1	1	1	1	1
F	0	0	0	0	1	1	1	1	1	1	1	1	1
G	0	0	0	0	0	1	1	1	1	1	1	1	1
H	0	0	0	0	0	0	1	1	1	1	1	1	1
I	0	0	0	0	0	0	0	1	1	1	1	1	1
J	0	0	0	0	0	0	0	0	1	1	1	1	1
K	0	0	0	0	0	0	0	0	0	1	1	1	1
L	0	0	0	0	0	0	0	0	0	0	1	1	1
M	0	0	0	0	0	0	0	0	0	0	0	1	1
N	0	0	0	0	0	0	0	0	0	0	0	0	1
O	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.2. Matrix representation of a maximally unbalanced tree containing 16 taxa (tree B in Fig. 4.1). Matrix representation contains 104 (50%) primitive states (0s) and 104 (50%) derived states (1s).

The unbalanced tree (Fig 4.1b) leads to a far more symmetrical MR (Table 4.2). Even if character polarity is unknown, and we code the most common state as state 0 and the least common as state 1 for the unbalanced tree, there are still more 1s in the data than in the MR of the symmetrical tree (Table 4.3). From here on I call this ratio of the less common state to the more common state in a character the state ratio (SR).

An SR of one indicates an equal number of each state in a character, and ratios closer to zero indicate unequal numbers of each state. Characters with a higher SR will also have a higher maximum number of steps that they can take in parsimony analyses. The average SR of the MR of the unbalanced tree is 0.43, compared with 0.3 for the balanced tree, showing that the maximum number of parsimony steps of the MR of the unbalanced tree is greater than that of the balanced tree. However, since all characters are binary, the minimum number of steps each MR can take in parsimony analysis is the same (26 each). Therefore, the difference between the maximum and minimum number of steps is greater for the MR of the unbalanced tree. Figure 4.2 shows the lengths of the two MRs (Tables 4.1 and 4.2) on 1,000,000 randomly generated trees. It is likely that by chance alone the balanced MR will have a lower tree length when plotted onto a random tree. This shows that it is more parsimonious to plot a matrix with a low SR onto a random tree than one with a high SR.

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
A	1	1	1	1	1	1	0	0	0	0	0	0	0
B	1	1	1	1	1	1	0	0	0	0	0	0	0
C	0	1	1	1	1	1	0	0	0	0	0	0	0
D	0	0	1	1	1	1	0	0	0	0	0	0	0
E	0	0	0	1	1	1	0	0	0	0	0	0	0
F	0	0	0	0	1	1	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	1	0	0	0	0	0	0
J	0	0	0	0	0	0	1	1	0	0	0	0	0
K	0	0	0	0	0	0	1	1	1	0	0	0	0
L	0	0	0	0	0	0	1	1	1	1	0	0	0
M	0	0	0	0	0	0	1	1	1	1	1	0	0
N	0	0	0	0	0	0	1	1	1	1	1	1	0
O	0	0	0	0	0	0	1	1	1	1	1	1	1
P	0	0	0	0	0	0	1	1	1	1	1	1	1

Table 4.3. Matrix representation of a maximally unbalanced tree containing 16 taxa (tree B in Fig. 4.1) with the most common state for each character coded as 0 and the least common coded as 1. Matrix representation contains 146 (70%) 0s and 62 (30%) 1s.

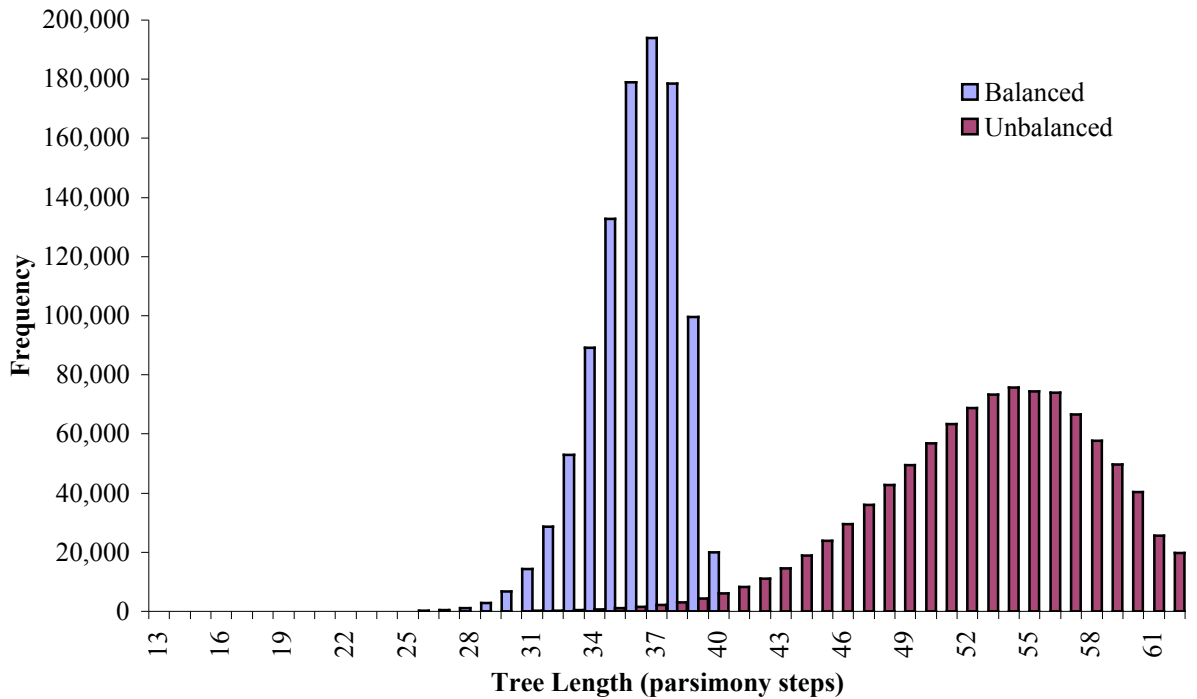


Figure 4.2. Histogram of parsimony lengths of the MRs of the balanced and unbalanced source trees mapped onto 1,000,000 randomly generated trees

Figure 4.3 shows the total tree lengths of the two MRs (Tables 4.1 and 4.2) on 10,000 randomly generated trees against the difference in length between the lengths of the unbalanced and balanced MR on that tree. It can be seen that at high tree lengths the unbalanced MR is much longer than the balanced MR. As the tree length decreases, the difference in length between the two MRs decreases until, on many of the shortest trees, the balanced MR is longer than the unbalanced MR despite its maximum possible length being much greater. This clearly demonstrates that it is more parsimonious to plot the balanced MR (high SR) onto trees more similar to the unbalanced tree than vice versa. This suggests that the bias in MRP towards unbalanced supertrees is caused by the higher SR value of MRs of unbalanced trees.

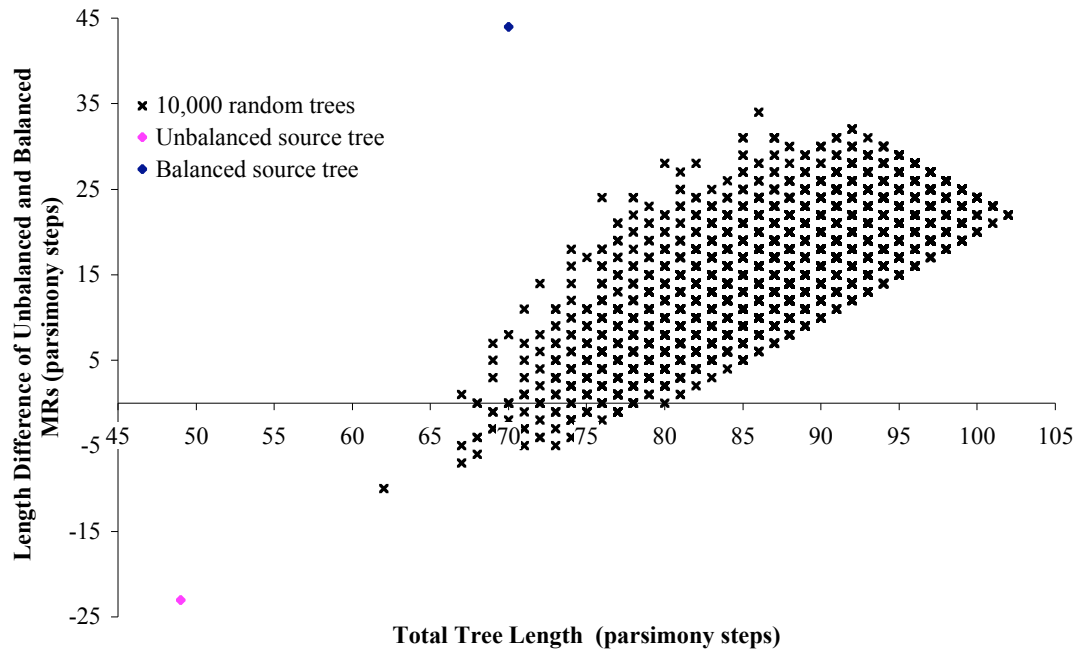


Figure 4.3. Scatter plot of total tree lengths of the MRs of the balanced and unbalanced source trees mapped onto 10,000 random trees. The y-axis represents the difference in length (parsimony steps) of the MRs of the two source trees when plotted onto the random trees. This value is simply the length of the unbalanced MR – the length of the balanced MR. Values on the y-axis greater than 0 indicate that the unbalanced MR is longer than the balanced MR, and vice versa. Points on the plot may represent many occurrences of the same value. The lengths of the two MRs mapped onto the two source trees are also marked.

Interestingly, as pointed out by Wilkinson *et al.* (in prep), a second type of MRP, called Purvis MRP (Purvis, 1995), shows the opposite bias, towards balanced trees. Purvis MRP uses one matrix element to represent each clade splitting its members from the members of its sister group and the root. Any other leaves in the tree are scored as missing (?). The Purvis MRs of the two source trees are shown in tables 4.4 and 4.5.

With Purvis MRP, the MR of the balanced tree has an SR of 1, whereas the MR of the unbalanced tree has an SR of 0.125. Therefore, with Purvis MRP the bias is still towards the tree with the higher SR, which in this case is the balanced tree.

If unbalanced trees are more likely to be expressed in the results of MRP than balanced trees when an equal number of characters supporting each hypothesis are present, it is unclear what happens in a normal parsimony analysis if competing phylogenetic signals are present. If one signal supports a balanced hypothesis and another supports an unbalanced hypothesis, even if there are equal numbers of characters supporting each hypothesis, it is likely that the output tree will be relatively unbalanced. This could explain

the often reported bias of incorrectly reconstructed phylogenies towards unbalanced topologies (e.g. Mooers *et al.*, 1995).

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
A	1	1	1	0	0	?	?	?	?	?	?	?	?
B	1	0	?	?	1	1	0	?	?	?	?	?	?
C	1	0	?	?	1	0	1	?	?	?	?	?	?
D	0	?	?	?	?	?	?	1	1	0	0	?	?
E	0	?	?	?	?	?	?	0	?	?	1	1	0
F	1	1	0	1	0	?	?	?	?	?	?	?	?
G	0	?	?	?	?	?	?	0	?	?	1	1	0
H	1	0	?	?	1	0	1	?	?	?	?	?	?
I	0	?	?	?	?	?	?	0	?	?	1	0	1
J	0	?	?	?	?	?	?	0	?	?	1	0	1
K	0	?	?	?	?	?	?	1	0	1	0	?	?
L	1	1	1	0	0	?	?	?	?	?	?	?	?
M	1	0	?	?	1	1	0	?	?	?	?	?	?
N	1	1	0	1	0	?	?	?	?	?	?	?	?
O	0	?	?	?	?	?	?	1	0	1	0	?	?
P	0	?	?	?	?	?	?	1	1	0	0	?	?

Table 4.4. Purvis matrix representation of a maximally balanced tree containing 16 taxa (tree A in Fig. 4.1). Matrix representation contains 40 (50%) primitive states (0s) and 40 (50%) derived states (1s).

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13
A	1	1	1	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	1	1	1	1	1	1	1
C	0	1	1	1	1	1	1	1	1	1	1	1	1
D	?	0	1	1	1	1	1	1	1	1	1	1	1
E	?	?	0	1	1	1	1	1	1	1	1	1	1
F	?	?	?	0	1	1	1	1	1	1	1	1	1
G	?	?	?	?	0	1	1	1	1	1	1	1	1
H	?	?	?	?	?	0	1	1	1	1	1	1	1
I	?	?	?	?	?	?	0	1	1	1	1	1	1
J	?	?	?	?	?	?	?	0	1	1	1	1	1
K	?	?	?	?	?	?	?	?	0	1	1	1	1
L	?	?	?	?	?	?	?	?	?	0	1	1	1
M	?	?	?	?	?	?	?	?	?	?	0	1	1
N	?	?	?	?	?	?	?	?	?	?	?	0	1
O	?	?	?	?	?	?	?	?	?	?	?	?	0
P	?	?	?	?	?	?	?	?	?	?	?	?	?

Table 4.5. Purvis Matrix representation of a maximally unbalanced tree containing 16 taxa (tree B in Fig. 4.1) with the most common state for each character coded as 0 and the least common coded as 1. Matrix representation contains 13 (11%) primitive states (0s) and 104 (89%) derived states (1s).

4.3.3.1 Bias Towards Unbalanced Trees in Compatibility

Like parsimony, many compatibility analysis methods may show a bias towards unbalanced topologies. The reason is again due to the relatively low SRs of characters supporting unbalanced trees. However, the reason for the bias is slightly different than in parsimony. As pointed out by many compatibility workers, binary characters with equal numbers of 0s and 1s (high SR) are more likely to be incompatible by chance alone (e.g. Le Quesne, 1969; 1972). Thus, if a character with a high SR is randomly permuted it is more likely to be incompatible with another character than is a character with a low SR. Therefore, if two characters are compatible, it is much more likely that the compatibility is attributable to chance alone if the characters have low SR values. This is analogous to a type I error in statistical tests. In this case, the null hypothesis would be that the two characters are random, and therefore unlikely to be compatible. If they are compatible, we can reject the null hypothesis and conclude that the characters are not random, and that they do contain phylogenetic signal. The type I error is that we have rejected the null hypothesis when it is true; i.e. when the character is random but by chance alone is compatible with other characters. To try to correct for this problem, many compatibility tests use random permutations to calculate the probability of the character being compatible by chance alone (calculating the type I error rate). They then use this value to downweight the significance of compatibility between characters that are likely to be compatible by chance alone. However, such methods have disadvantages. By correcting for the possibility that compatibility is due to chance alone, characters that are truly compatible due to phylogenetic signal can be downweighted simply because they have a high SR. As noted by Gauld and Underwood (1986), it is possible for this method to lead to characters with the same number of incompatibilities showing very different CCSR scores, and some characters can even show higher CCSR scores than other characters which have more incompatibilities with the rest of the matrix. For example, imagine a dataset that contained fifteen characters, ten of which were mutually compatible, but incompatible with the remaining five characters, which were themselves incompatible with all other characters in the data. Simply counting the number of pairwise incompatibilities of the characters would show that the first ten characters are all incompatible with five characters, while the last five characters are each incompatible with fourteen characters. Therefore, each of the ten compatible characters is scored equally by simply counting the number of character with which they are incompatible. If it is now assumed that one of the

group of ten characters has an SR of 1 (equal numbers of states 0 and 1) and a second character in the group of ten has a minimal SR (e.g. just two taxa coded state 1 and all others as state 0), then problems arise with some compatibility methods. Simply counting incompatibilities still gives the same result for the two characters: an incompatibility score of five for each. However, if CCSR, NDev or LQP methods are used, an expected incompatibility value is needed for each character. The expected compatibility value for the low SR character will be greater than that of the high SR character. This means that, when using CCSR, NDev or LQP, it is likely that a low SR character will be considered less compatible (weaker) than a high SR character simply because of the distribution of character states it has, when in fact both characters are compatible with an equal number of other characters in the data. In practice, applying methods such as the boildown to these compatibility methods can lead to characters with low SRs being downweighted or removed from the data preferentially. One side-effect of this is that, as previously mentioned, datasets describing asymmetrical trees generally have higher SRs. Therefore, given a dataset containing characters supporting two hypotheses, one being a balanced tree and the other an unbalanced tree, many compatibility methods will favour the unbalanced tree. Methods that involve simply counting the number of incompatibilities or finding maximal cliques, are not subject to this bias. Other methods, such as the CCSR, NDev and LQP, were designed to correct for the chance of compatible characters being compatible simply by chance. However, our goal in carrying out phylogenetic analyses is to find the tree representing true phylogenetic signal, and thus the evolutionary history of a group. To do this we might reasonably assume that the strongest signal in the data will be more likely to be the correct one. Any noisy characters that are compatible with the true signal will not change this result, although they may cause incorrect resolution of parts of the tree that is not resolved by phylogenetically useful data. It seems preferential to me to accept that some noisy characters are compatible with the true tree, and leave such characters in our data than to downweight some characters simply because they describe a synapomorphy of a small number of taxa.

BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**:3-10.

- CAMIN, J. H., and R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. *Evolution* **19**:311-326.
- CHEN, D., L. DIAO, O. EULENSTEIN, D. FERNÁNDEZ-BACA, and M. J. SANDERSON. 2003. Flipping: a supertree construction method. Pages 135-160 in *Bioconsensus*, volume 61 of DIMACS: Series in Discrete Mathematics and Theoretical Computer Sciences (M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). American Mathematical Society, Providence.
- DARWIN, C. 1859. *The origin of species*. 1968 Edition, Penguin Books Ltd., Harmondsworth.
- ESTABROOK, G. F. 1983. The causes of character incompatibility. Pages 279-295 in *Numerical taxonomy* (J. Felsenstein, ed.) Springer-Verlag, Berlin.
- ESTABROOK, G. F. 1984. Phylogenetic trees and character state trees. Pages 135-151 in *Cladistics: perspectives on the reconstruction of evolutionary history* (T. Duncan, and T. F. Stuessy, eds.). Columbia University Press, New York.
- ESTABROOK, G. F., C. S. JOHNSON JR., and F. R. MCMORRIS. 1975. An idealized concept of the true cladistic character. *Mathematical Biosciences* **23**:263-272.
- ESTABROOK, G. F., C. S. JOHNSON JR., and F. R. MCMORRIS. 1976. A mathematical foundation for the analysis of cladistic character compatibility. *Mathematical Biosciences* **29**:181-187.
- ESTABROOK, G. F., and L. R. LANDRUM. 1975. A simple test for the possible simultaneous evolutionary divergence of two amino acid positions. *Taxon* **24**:609-613.
- ESTABROOK, G. F., and F. R. MCMORRIS. 1977. When are two qualitative taxonomic characters compatible? *Journal of Mathematical Biology* **4**:195-200.
- ESTABROOK, G. F., J. G. STRAUCH JR., and K. L. FIALA. 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects. *Systematic Zoology* **26**:269-276.
- FARRIS, J. S. 1973. On comparing the shapes of taxonomic trees. *Systematic Zoology* **22**:50-54.
- FARRIS, J. S., V. A. ALBERT, M. KÄLLERAJÖ, D. LIPSCOMB, and A. G. KLUGE. 1996. Parsimony jackknifing outperforms neighbour-joining. *Cladistics* **12**:99-124.
- FELSENSTEIN, J. 1982. Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* **57**:379-404.

- FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.
- FITCH, W. M. 1975. Toward finding the tree of maximum parsimony. Pages 189-230 in Proceedings of the 8th international Freeman Conference in numerical taxonomy (G. F. Estabrook, ed.) W. H. Freeman, San Francisco.
- FITCH, W. M. 1977. On the problem of discovering the most parsimonious tree. *American Naturalist* **111**:223-257.
- GAULD, I., and G. UNDERWOOD. 1986. Some applications of the LeQuesne compatibility test. *Biological Journal of the Linnean Society* **29**:191-222.
- GOLOBOFF, P. A., and D. POL. 2002. Semi-strict supertrees. *Cladistics* **18**:514-525.
- GUISE, A., D. PEACOCK, and T. GLEAVES. 1982. A method for identification of parallelism in discrete character sets. *Zoological Journal of the Linnean Society* **74**:293-303.
- GUYER, C., and J. B. SLOWINSKI. 1991. Comparison of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution* **45**:340-350.
- GUYER, C., and J. B. SLOWINSKI. 1993. Adaptive radiation and the topology of large phylogenies. *Evolution* **47**:253-263.
- HARCOURT-BROWN, K. G. 2002. Phylogenetic tree shape with special reference to the phylogeny of the Cretaceous globotruncanid foraminifera. Ph.D. Thesis, University of Bristol.
- HUELSENBECK, J. P., and M. KIRKPATRICK. 1996. Do phylogenetic methods produce trees with biased shapes? *Evolution* **50**:1418-1424.
- KÄLLERAJÖ, M., V. A. ALBERT, and J. S. FARRIS. 1999. Homoplasy increases phylogenetic structure. *Cladistics* **15**:91-93.
- LE QUESNE, W. J. 1969. A method of selection of characters in numerical taxonomy. *Systematic Zoology* **18**:201-205.
- LE QUESNE, W. J. 1972. Further studies based on the uniquely derived character concept. *Systematic Zoology* **21**:281-288.
- LE QUESNE, W. J. 1982. Compatibility analysis and its applications. *Zoological Journal of the Linnean Society* **74**:267-275.
- LE QUESNE, W. J. 1989. The normal deviate test of phylogenetic value of a data matrix. *Systematic Zoology* **38**:51-54.
- LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological characters. *Systematic Biology* **50**:913-925.

- LORENZEN, S. 1993. The role of parsimony, outgroup analysis, and theory of evolution in phylogenetic systematics. *Zeitschrift für Zoologische Systematik und Evolutionsforschung* **31**:1-20.
- MCMORRIS, F. R. 1975. Compatibility criteria for cladistic and qualitative taxonomic characters. Pages 399-415 in *Proceedings of the 8th international Freeman Conference in numerical taxonomy* (G. F. Estabrook, ed.) W. H. Freeman, San Francisco.
- MEACHAM, C. A. 1984. Evaluating characters by character compatibility analysis. Pages 152-165 in *Cladistics: perspectives on the reconstruction of evolutionary history* (T. Duncan, and T. F. Stuessy, eds.). Columbia University Press, New York.
- MEACHAM, C. A. 1994. Phylogenetic relationships at the basal radiation of Angiosperms, further study by probability of character compatibility. *Systematic Botany* **19**:506-522.
- MOOERS, A. Ø., R. D. M. PAGE, A. PURVIS, and P. H. HARVEY. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. *Systematic Biology* **44**:332-342.
- NELSON, G., and N. I. PLATNICK. 1991. Three-taxon statements: a more precise use of parsimony? *Cladistics* **7**:351-366.
- PATTERSON, C. 1982. Morphological characters and homology. Pages 21-74 in *Problems of Phylogenetic Reconstruction* (K. A. Joysey, and A. E. Friday, eds.). Academic Press, London and New York.
- PENNY, D. 1982. Graph theory, evolutionary trees and classification. *Zoological Journal of the Linnean Society* **74**:305-328.
- PENNY, D., and M. D. HENDY. 1985. Testing methods of evolutionary tree construction. *Cladistics* **1**:266-272.
- PENNY, D., and M. D. HENDY. 1986. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution* **3**:403-417.
- PURVIS, A. 1995. A modification to Baum and Regan's method for combining phylogenetic trees. *Systematic Biology* **44**:251-255.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1**:53-58.
- SHARKEY, M. J. 1989. A hypothesis independent method of character weighting for cladistic analysis. *Cladistics* **5**:63-86.

- SHARKEY, M. J. 1994. Discriminate compatibility measures and the reduction routine. *Systematic Biology* **43**:526-542.
- THORLEY, J. L., and M. WILKINSON. 2003. A view of supertree methods. Pages 185-193 in *Bioconsensus* (M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.). DIMACS series in discrete mathematics and theoretical computer science, American Mathematical Society, Providence.
- WIENS, J. J., and D. M. HILLIS. 1996. Accuracy of parsimony analysis using morphological data: a reappraisal. *Systematic Botany* **21**:237-243.
- WILKINSON, M. 1992. Consensus, compatibility and missing data in phylogenetic inference. Ph.D. Thesis, University of Bristol.
- WILKINSON, M. 1994a. The permutation method and character compatibility. *Systematic Biology* **43**:274-277.
- WILKINSON, M. 1994b. Three-taxon statements: When is parsimony analysis also a clique analysis? *Cladistics* **10**:221-223.
- WILKINSON, M. 1995. PICA₉₅ software and documentation. School of Biological Sciences, University of Bristol. Bristol
- WILKINSON, M. 1997a. Characters, congruence and quality: A study of neuroanatomical and traditional data in caecilian phylogeny. *Biological Reviews of the Cambridge Philosophical Society* **72**:423-470.
- WILKINSON, M. 1997b. Limitations on the use of compatibility methods for polarizing ordered characters. *Systematic Biology* **46**:230-231.
- WILKINSON, M., J. L. THORLEY, D. T. J. LITTLEWOOD, and R. A. BRAY. 2001. Towards a phylogenetic supertree for the Platyhelminthes? Pages 292-301 in *Interrelationships of the Platyhelminthes* (D. T. J. Littlewood, and R. A. Bray, eds.). Chapman-Hall, London.
- WILSON, E. O. 1965. A consistency test for phylogenies based on contemporaneous species. *Systematic Zoology* **14**:214-220.
- YEATES, D. K. 1995. Groundplans and exemplars: Paths to the tree of life. *Cladistics* **11**:343-357.