Cladistics 2(4):297-317

STATISTICAL COVARIANCE AS A MEASURE OF PHYLOGENETIC RELATIONSHIP

MALCOLM R. FORSTER[1]

[1]*Mathematics Department, Monash University, Melbourne, Australia*

Abstract—The method of parsimony in phylogenetic inference is often taken to mean two things: (1) that one should favor the genealogical hypothesis that minimizes the required number of homoplasies (matchings of independently evolved derived character states), and (2) that symplesiomorphies (matchings of primitive character states) have little or no evidential value for phylogenetic relationship. This paper shows both theses to be false by undermining recent likelihood arguments for them and by providing a more secure likelihood proof of a new method, which is incompatible with both (1) and (2).

Arguments Concerning Parsimony

This paper has two parts. The first and second sections summarize the main theses and arguments of the whole paper in a non-technical fashion, whereas the remaining five sections substantiate those claims in terms of the theoretical framework used by Sober (1983, 1984, 1985) in his (unsuccessful) attempt to provide a likelihood justification of parsimony.

In phylogenetic inference, cladistic methods of parsimony hold that synapomorphies (matchings of derived character states) provide the only taxonomic evidence of a shared phylogenetic history. One immediate question is how this method treats 'inconsistent' data sets for which it is not possible to explain every synapomorphy as homologous (as derived from a common ancestral species). This question may even be viewed rhetorically as suggesting that cladistic methods are in practice unworkable, for any sufficiently large data set is likely to be 'inconsistent' if enough homoplasies actually occur in nature. Any cladist might reply along the lines of Farris (1983:9):

> 'Observing' a falsifier of a theory does not prove that the theory is false; it simply implies that either the theory or the observation is erroneous. It is then seen that the only implication that can be derived from the falsification of every genealogy is that some of the falsifiers are errors—homoplasies.

The only other implication is that parsimony is false. Farris wants to avoid that conclusion by suggesting that some taxonomic evidence should be 'written off' as 'erroneous' so that we are left with a 'consistent' data set pointing uniquely to a single phylogenetic tree. Two questions then arise: what, if any, precedence or justification does this sort of strategy have in science generally, and how should the idea be applied unambiguously to the case at hand (given that different genealogies will be obtained by writing off different parts of the data)?

Farris has an interesting way of killing both these birds with one stone. He freely admits that treating evidence as 'erroneous' is ad hoc and should be avoided in science whenever possible. Prima facie this should refute parsimony, yet Farris cleverly turns this around in his defence by presenting the minimization of ad hoc hypotheses as precisely the principle that can make the method work. It is not a mark against cladistic methods that they require the use of ad hoc hypotheses, because genealogical hypotheses are not meant to explain everything in our data set. So, treating some evidence as 'erroneous' is justified. Given that homoplasies actually occur in nature (if they don't there's no problem), some apomorphic similarities are not explainable by facts of phylogenetic relationship, whatever views one has on taxonomic methodology. If two similar traits evolve independently then this similarity is due to chance. So, phylogenetic explanation must allow such facts to go unexplained (unexplained, that is, in terms of common ancestry). In this way homoplasies are 'errors of evolution' and should not be seen as errors of parsimony methodology. As Farris put it:

> A genealogy does not explain by itself why one group acquires a new feature while its sister group retains the ancestral trait, nor does it offer any explanation of why seemingly identical features arise independently in distantly related lineages. . . . A genealogy is able to explain observed points of similarity among organisms just when it can account for them as identical by virtue of inheritance from a common ancestor. Any feature shared by organisms is so either by reason of common descent or because it is a homoplasy. The explanatory power of a genealogy is consequently measured by the degree to which it can avoid postulating homoplasies. (1983:18).

Ad hoc hypotheses are unavoidable in the case of phylogenetic inference, but general scientific methodology does require that we minimize their number, and thereby maximize the amount of data explained by the theory. In biological circles, this is known as the principle of parsimony.

This is exactly the principle that also enables cladists to kill the other bird. For the principle unambiguously favors the hypothesis read off from the largest 'consistent' set of data obtained by dismissing the fewest synapomorphies as homoplasious 'errors.' This line of reasoning leads to the following parsimony principle (rule 1), as stated by Sober (1983:335):

> Parsimony stipulates that the investigator must be able to distinguish between the ancestral (plesiomorphic) and the derived (apomorphic) form of every characteristic used. Given this information, the *preferred genealogical hypothesis is the one that requires the fewest homoplasies.*

Along the same line of argument it also seems to follow that symplesiomorphies have no particular evidential value, because all distant ancestral species are assumed to have started out with the ancestral trait. This means that any matching of plesiomorphic characters can be seen as arising from common ancestral origins on any genealogical hypothesis, so one is as good as another in explaining those facts. This leads to a second rule of parsimony (Sober, 1983:335):

> Parsimony may also be described as holding that synapomorphies—matches with respect to derived characteristics—count as evidence of a phylogenetic relationship, but that symplesiomorphies—matches with respect to ancestral characteristics—do not.

The obvious way of making Farris' view of homoplasies as 'evolutionary errors' precise is by modelling evolutionary processes stochastically, or probabilistically. Instead of the strong (and unrealistic) assumption that past character states determine future character states, we believe the weaker assumption that past character states determine the chance or probability of future character states evolving, where it is allowed that these 'rates of evolution' may vary from species to species, from characteristic to characteristic, and from time to time. Homoplasies then arise in evolution as chance co-occurrences of seemingly identical traits in distantly related species. Therefore, in defence of cladism, we might conclude that the existence of 'inconsistent' data sets does not refute the method of parsimony, but instead refutes the idea that evolution is deterministic.

The purpose of this paper is to argue that this line of defence backfires. The fact of indeterminacy in evolution actually undermines the arguments cited by Farris in support of parsimony, and shows their conclusions to be doubtful. I agree with Farris that ad hoc hypotheses should be minimized, but I disagree that postulations of homoplasy are ad hoc. So, it does not follow that the number of homoplasies required by a genealogy should be minimized.

A New Method of Phylogenetic Inference

Farris' claim that homoplasies are ad hoc is mistaken. But to substantiate this claim we need to reconstruct the reasoning behind it more carefully. Let us do this in terms of one of Sober's own examples (Sober 1983:213), which I will call example 1:

		Characteristics									
		1	2	3	4	5	6	7	8	9	10
	A	1	1	1	1	1	1	1	1	1	0
Species	B	1	1	1	1	1	1	1	1	1	1
	C	0	0	0	0	0	0	0	0	0	1

Here we have observed 10 characteristics of three species with the observed character states recorded in the table above ('1' stands for the derived or apomorphic state, and '0' for the ancestral or plesiomorphic form). The competing genealogies are conveniently labelled as (ABC), (AB)C, and A(BC) as diagrammed in Figures 1, 2, and 3, respectively.

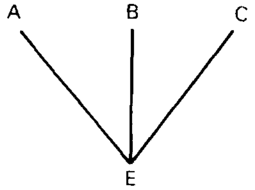

Figure 1: (ABC)

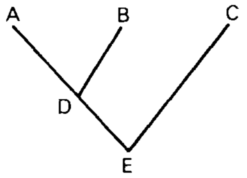

Figure 2: (AB)C

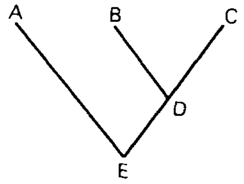

Figure 3: A(BC)

In this example the observed character distribution is 'inconsistent,' but if we discard some 1-1 matchings as homoplasious "errors" then we can be left with a consistent set. But there are two ways of doing this: either we can discard the 1-1 matching in characteristic 10, or we retain that and discard those in characteristics 1 to 9. The first alternative requires only 1 ad hoc dismissal of data, whereas the second requires 9, so the hypothsis (AB)C is more parsimonious than A(BC). The hypothesis (ABC) fares worst of all on this account, for it must dismiss all characteristics as providing "erroneous" information.

We will now examine the argument here more closely. What is the minimum number of homoplasies required by a genealogical hypothesis in its explanation of the data, and why? Let us look at one characteristic in two (arbitrary) sister species X and Y whose nearest common ancestor is the species Z according to the model. Suppose that we observe the apomorphic trait for both X and Y in this characteristic. Is this apomorphy homoplasious or homologous (synapomorphic) in the model? There are two cases to consider: (a) the ancestral species Z has the plesiomorphic form of that character; in this case, the apomorphy shared by X and Y is homoplasious. (b) Z is in the apomorphic state. In this case the apomorphy need not be homologous, because the population may have reverted back to the plesiomorphic state in evolving from Z to X or from Z to Y. But there is no homoplasy required in this case, because we can charitably assume that reversal might not have occurred.

Returning to example 1, (AB)C says that all apomorphies shared by A and C and by B and C must be homoplasies because E is assumed to be in the plesiomorphic state for all characteristics. But no apomorphies shared by A and B are required to be homoplasies, even if it is probable that some of them are. This is because we can assume that D is in the apomorphic state for those characteristics, and we can further assume that no reversal took place from D to A or from D to B. Thus, (AB)C requires only 1 homoplasy, although it is logically consistent with all 10 shared apomorphies being homoplasious. A similar argument tells us that A(BC) requires 9 homoplasies. Finally, (ABC) requires all 10 apomorphic matchings to be homoplasious, because of the assumption that E is in the plesiomorphic state for all characteristics.

The most telling objection to this approach brings into question the premise that the prediction of a high number of homoplasies by a model is an indication of its inadequacy. Homoplasies are nothing more than the matching of two features that have independent origins. Such things happen all the time, and it should count in favor of a theory, not against it, that it predicts such occurrences. Suppose we toss a pair of coins on 48 different occasions and 12 of these land a double heads. There are two theories that can account for this observation. The first theory says that the result for coin 1 is completely independent of that for coin 2, for any toss. The 12 double heads are explained as arising independently by chance—as being nothing more than expected accidents. The second theory asserts that some unseen mechanism determined that both coins would land the same on these 12 occasions. Assuming that the remaining tosses produced fairly random results, the first explanation is clearly the best despite the fact that it ''writes off'' all 12 ''synapomorphic'' matchings of heads as ''evolving'' independently from independent causal origins. Therefore, we have a case in which ''synapomorphic'' matchings provide no evidential support for common causal origins. Something is wrong with the idea that we should minimize the number of homoplasious matchings predicted by a theory.

Ironically, the example provided by Sober is exactly like this. In order to see that the independent evolution of species A, B, and C from E is a good explanation of the data, consider an urn analogy. Imagine that we have three large urns A, B, and C, full of marbles. 90% of the marbles (characteristics) in A are white (apomorphic), whereas all in B are white, and 10% in C are white. The rest are black (plesiomorphic). When we let '1' stand for 'white,' and '0' for 'black,' the data in example 1 is typical of what we would expect for 10 trials of an experiment consisting of one random draw from each urn. There is no reason to suppose that the draws were not independent of each other. All matchings are adequately explained as expected random chance co-occurrences. There is no need to postulate a common cause to explain the high number of matching pairs drawn from A and B because these are expected from the fact that those urns contain a high proportion of white marbles to start with. Analogously, the taxonomic data are adequately explained on the basis of a high expected number of matchings arising from high rates of evolutionary change experienced by both species, as evidenced by the fact that most traits are in the apomorphic state. In this example, hypothesizing the occurrence of homoplasies is fully justified, and is not ad hoc. There is no reason here to minimize the predicted occurrence of homoplasies.

It might be objected that this argument assumes that A and B actually have high rates of evolution. Indeed, if we were to assume otherwise, then the observed data would favor (AB)C over its rivals. But the assumption of high rates of change is, in fact, justified by the empirical evidence. For as Sober himself acknowledged (1984:223), the best estimate of the proportion of white marbles in each urn is given by the relative frequencies observed so far:

> I draw twenty balls from a large urn; you see that all twenty are red. You infer that this observation favors the hypothesis that all the balls in the urn are red over the hypothesis that 50% are red. The former hypothesis is better supported because it makes the observation more probable.

When we apply this to example 1, there can be no objection to the assumptions of my argument. Because the observed relative frequencies of apomorphic traits in A and B are high, the best guess is that the probabilities of a 0–1, transition from E to A and from E to B are both high. This entails that the probability of 1–1 matchings is high, without any need nor motive to posit a recent ancestor shared by A and B. This defence of the genealogy (ABC) clearly refutes the judgment of parsimony in this example.

What alternatives to parsimony are within the general cladistic philosophy? The lesson from this example is that there is no good justification for introducing common ancestral causes to explain apomorphic matchings when each trait can be seen as arising independently. But when do we have positive evidence for a non-trivial taxonomic grouping such as (AB)C or A(BC)? To answer this question, take the coin tossing example again, and suppose that out of 48 tosses roughly half are double heads while the other half are double tails. Here we have a strange phenomenon that is not so well explained on the assumption that each toss is independent of the other (although it does not disprove that hypothesis in the strict sense). So what is it that makes it difficult to explain these double heads on the assumption of independence, which does not apply to the previous example? The answer is that on independence we expect that roughly ½ x ½ (i.e., ¼) of all tosses will land double heads. In the last example this means that 12 out of 48 tosses are expected to be double heads. But we actually observe 24 double heads in 48 tosses, so 12 of these are not so well accounted for on the basis of independence (also note that the high frequency of double heads cannot be explained by supposing that the coins are biased towards heads, because this would leave the high frequency of double tails unexplained). The theory that some hidden mechanical device conspired to produce the results (i.e., that the tosses were not random) does now receive empirical support (we might have other evidence against it of course).

The utility, therefore, of postulating common causal origins of matching events lies in their ability to explain the number of matchings observed above that expected on the assumption of independence. Let the proposition $A=1$ say that species A is in the apomorphic state and $A=0$ say that it is not, and similarly for B and C (we are assuming that each characteristic is in one of two possible states). Then we can denote the observed proportion (relative frequency) of synapomorphies between A and B out of all characteristics observed as $r(A=1.B=1)$. Letting the total number of observed characteristics be N, it follows that $N.r(A=1.B=1)$ is the total number of synapomorphies. Let $r(A=1)$ and $r(B=1)$ be the observed proportions of apomorphic characters for species A and B respectively. The expected proportion of synapomorphies on the assumption of independence is the product $P(A=1).P(B=1)$, where P denotes the expected relative frequencies on the assumption that the independence hypothesis is true. But the best estimate of $P(A=1)$ and $P(B=1)$ on the evidence will be $r(A=1)$ and $r(B=1)$, as was mentioned above, so the best estimate of $P(A=1).P(B=1)$ will be $r(A=1).r(B=1)$. The expected number of synapomorphies, given independence, will be $N.r\ (A=1).r(B=1)$, where N is the number of characteristics observed. In the coin tossing example, $N.r(A=1).r(B=1)$ is 48x½x½, which is 12, whereas the actual number of double heads ($= N.r[A=1.B=1]$) was 24. The difference $[N.r(A=1.B=1)-N.r(A=1).r(B=1)]$ is the number of synapomorphies between A and B that can be better explained with the assumption that A and B have a recent common cause than without that assumption. So, it is only when this difference is large and positive that we have good evidence of common descent from a recent ancestral species D. In other words, the existence of the species D only really explains the number of synapomorphies that occur above the number that can be expected to arise on the assumption of independent causal origins. This idea is captured by the following rule:

> *Rule 3:* The method of comparing covariances: the best genealogy is the one that explains the most synapomorphies above those which are expected to occur on the assumption of independence.

It should be realized that this rule has nothing to do with minimizing homoplasies, even though it does require that we maximize the number of synapomorphies explained. The reason is that, contrary to Farris, we no longer equate "explaining a synapomorphy" with "showing that it is, or may be, homologous and therefore not, or probably not, homoplasious." Another caution concerns the meaning of "best." In a case in which the difference N.[r(A=1.B=1)−r(A=1).r(B=1)] is positive but not very large, rule 3 says that a common cause explanation is the "best," but this refers only to its level of evidential support. Other considerations, such a desire for simplicity, may lead us to retain the null hypothesis of independence in this case, or to seek further information.

Let us apply rule 3 to example 1. The data here may be characterized in terms of the relative frequency of each possible state of affairs. Let r(111) denote the relative frequency r(A=1.B=1.C=1), and so on. Here we observe that r(111)=0, r(110)=.9, r(101)=0, r(100)=0, r(011)=.1, r(010)=0, r(001)=0, and r(000)=0. These relative frequencies (proportions) all add up to 1, as they must. The number of each type of observation is obtained by multiplying each relative frequency by N (=10). The observed covariances, denoted by Cov(A, B), Cov(A, C), and Cov(B, C) are defined as the differences:

$$
\begin{aligned}
\mathrm{Cov}(A,B) &= {}^{df}r(A=1.B=1)-r(A=1).r(B=1)\\
\mathrm{Cov}(B,C) &= {}^{df}r(B=1.C=1)-r(A=1).r(B=1)\\
\mathrm{Cov}(A,C) &= {}^{df}r(A=1.C=1)-r(A=1).r(C=1)
\end{aligned}
$$

In this example each of these values is zero. Rule 3 therefore correctly captures the lessons of the urn analogy—there is no evidence in this example of common ancestry between A and B, despite the high number of apomorphies shared by them. Also note that rule 3 not only contradicts parsimony judgments but also the method of grouping in terms of overall similarity. In example 1, the degree of raw similarity (measured by the number of shared apomorphies plus plesiomorphies) is the same as that for parsimony (the number of shared apomorphies). So rule 3 contradicts both of those methods.

Example 2: Suppose we look at 100 characteristics of 3 species A, B, and C, finding that the first 10 are exactly as recorded in example 1, while the other 90 characteristics are all in their plesiomorphic states for all species. Letting n(000) be the number N.r(000), etc, we have n(000)=90, n(110)=9, n(011)=1, and the rest 0. Both the methods of parsimony and similarity cannot recommend a different genealogy on the basis of such additional data. For parsimony, the reason is that there are no synapomorphies in the last 90 characteristics, while the method of overall similarity adds the same scores to each genealogy. But the method of comparing covariances is sensitive to the addition of uniform data. In this example the observed covariances are now changed to (approximately): Cov (A,B)=.08, Cov(B,C)=0, Cov(A,C)=0. That is, there are now close to 8 synapomorphies between A and B that are better explained by the hypothesis (AB)C than by (ABC), or by A(BC). The judgment of rule 3 now concurs with the judgment of parsimony, but the rationale behind that judgment is completely different. This provides a simple explanation of why cladists tend to believe that (AB)C is supported by the data in example 1. They tend to assume, as will usually be the case, that there are many other characteristics in a plesiomorphic state for all species. But by their own rule 2, this additional data is irrelevant, so it is not recorded. If indeed, there is such unrecorded evidence then the covariance method will agree with parsimony in this case, but not for the same reasons. According to the covariance rule 3, both parsimony rules 1 and 2 are wrong. In this example the two errors happen to cancel out to produce the right answer. But this does not always happen, as the next two examples will show.

Example 3: Suppose we examine 10 characteristics of species A, B, and C; the results of which are as tabulated below:

		Characteristics									
		1	2	3	4	5	6	7	8	9	10
	A	0	0	0	0	0	0	0	0	1	1
Species	B	1	1	0	0	1	1	0	0	1	1
	C	0	1	0	1	0	1	0	1	0	1

Here r(111)=.1, r(110)=.1, r(101)=0, r(100)=0, r(011)=.2, r(010)=.2, r(001)=.2, and r(000)=.2. The observed covariances are: Cov(A,B)=.08, Cov(B,C)=0, and Cov(A,C)=0. What this means is that there are 0.8 synapomorphies observed between A and B above what we would expect on the assumption of independence, but no synapomorphies between A and C or between B and C above what we would expect on the basis of their independent evolution. The method of covariances in rule 3 judges the genealogy (AB)C the best explanation of the data, whereas A(BC) and (ABC) come out second equals. Admittedly, 0.8 of a synapomorphy is hardly a significant number, but if we were to observe exactly the same pattern for 100 characteristics, then the observed covariances would be exactly the same but the number of synapomorphies would then be 8. This is then significant evidence in favor of (AB)C. But the judgment of parsimony is clearly in favor of the genealogy A(BC). Moreover, rule 3 will not agree with parsimony even when there is unrecorded data for other characteristics in the plesiomorphic state for all species. The covariance between B and C will remain 0 in such a case, and so all shared apomorphies are still adequately explained as arising from independent causes.

Example 4: Suppose that the results are the same as in example 3, except for the 8th and 9th characteristics:

		Characteristics									
		1	2	3	4	5	6	7	8	9	10
	A	0	0	0	0	0	0	0	1	0	1
Species	B	1	1	0	0	1	1	0	0	1	1
	C	0	1	0	1	0	1	0	0	1	1

Although these results may appear to be similar to those in example 3, the covariances exhibited in the data are importantly different. In particular, Cov(A,B)=–.02, and Cov(B,C)=+.1; Cov(A,C) is again 0. Neither (AB)C, nor A(BC), can explain the negative covariance between A and B, but A(BC) can better explain the positive covariance between B and C whereas (AB)C and (ABC) cannot. Rule 3 comes down in favor of A(BC), and agrees with parsimony in this case, unlike example 3. If we compare example 3 with example 4, we see that covariances depend on rather subtle features of the data, for they differ in only 20% of the characteristics, yet the method of covariances gives different answers.

In summary, examples 1 and 3 both show that the method of comparing covariances is incompatible with parsimony's rule 1. My analysis of the situation is that the nature of probabilistic explanation needs to be fully appreciated here. If I draw marbles from two urns independently, I can still expect a certain proportion of the pairs to match without doubting that the draws were in fact independent. It is only when marbles match more frequently than would be expected on the basis of independence that one is justified in doubting that initial "null" assumption. Any probabilistic treatment of evolutionary change has the same consequence: it is only when characteristics of two sister species match more frequently than expected on the assumption of independent evolution that I am empirically justified in doubting the "null" hypothesis that all shared apomorphies are homoplasies.

Sober's (1983, 1984, 1985) arguments to the opposite conclusion will be critically examined below, where several objections are listed. After that, I provide a rigorous likelihood justification of covariance methods as formulated in rule 3. The technical argument against parsimony's rule 2 is postponed until the final section.

Likelihood Proofs of Cladistic Methods

As indicated above, the obvious way of making Farris' view of homoplasies as 'evolutionary errors' precise is by modelling evolutionary processes stochastically, or probabilistically. Instead of the strong (and unrealistic) assumption that past character states determine future character states, we adopt the weaker assumption that past character states determine the chance or probability of future character states evolving, where it is allowed that these 'rates of evolution' may vary from species to species, from characteristic to characteristic, and from time to time. Once a particular evolutionary model specifies these transition probabilities, it will also assign a probability (given other assumptions) to the occurrence of any possible character distribution among the sister species under consideration. Most importantly, the model will assign a probability to the observed character distribution, which allows us to compare different evolutionary models against the collected taxonomic evidence by means of the statistical notion of likelihood (defined, for our purposes, as the probability of the data given the theory—see Edwards, 1972). The likelihood of the model is an accepted measure of how well the genealogy conforms to the data.

In fact, Farris stated that his minimization of ad hoc hypotheses of homoplasy is "no more escapable than the general requirement that any theory should conform to observation; indeed, the *one derives from the other*" (Farris, 1983:17; emphasis added). Maximizing explanatory power and minimizing the need for homoplasies are both a consequence of maximizing evidential support, on his view. The suggestion is that parsimony can be justified from likelihood principles.

In some models quite the opposite is the case. Felsenstein (1978, 1979) presented stochastic models in which the genealogy assigning the highest probability to the data is not the most parsimonious. Only when the rates of evolutionary change are very small, under which conditions the models also predict very little homoplasy, is the maximum likelihood tree also the most parsimonious. From this we might conclude that parsimony requires rarity of homoplasy as a precondition of its valid application.

Farris (1984) counter-attacked by pointing out that the models used in the argument contain, by Felsenstein's own admission, unrealistic assumptions about evolution. The shoe is actually on the other foot. "If those assumptions do not apply to real cases, then, so far as Felsenstein can show, the criticism of parsimony need not apply to real cases either" (Farris, 1983:16). The original charge against parsimony—that it makes unrealistic assumptions—has now been turned around as a criticism of unparsimonious methods.

The defenders of parsimony have even taken that offensive at this point by producing their own likelihood proofs for parsimony rather than against it. In a series of fascinating articles, Sober (1983, 1984, 1985) defined and argued for a realistic theoretical framework in which he claims to provide a justification of parsimony principles, as encapsulated in rules 1 and 2 above. His framework is carefully designed so as to make minimal assumptions about the actual mechanisms of evolutionary change, and in this I believe he succeeds. By eliminating as many empirically unjustified assumptions about evolution as possible, his framework conforms to the theoretical frugality of cladism (and I agree that this is a virtue).

But there is a twist to the story. In his effort to make the assumptions of his model as weak as possible, Sober has run into a different type of problem. Because his models are very weak, they do not assign a definite numerical probability value to the observations, so they cannot be straightforwardly compared in terms of likelihood. His solution is to invoke a questionable (nonbiological) assumption about how the comparison should be carried out. The weakening of the assumptions about evolution requires a strengthening of his prescriptions on theory comparison, and this gets Sober into trouble. His method of comparing 'weak' biological models does not work, as I will show below.

I will provide a method for comparing the evidential support of weak theories that avoids these problems, and show that parsimony methods fail on the basis of Sober's own model. Then, this paper will go one step further and argue for the new method of comparing covariances on the basis of Sober's frugal cladistic framework.

The Theoretical Model

Suppose that we are faced with the task of classifying three sister species A, B, and C. The three competing genealogical hypotheses are (ABC), (AB)C, and A(BC), represented by the phylogenetic trees in Figures 1, 2, and 3 respectively. Consider now an arbitrary number of characteristics (N, say), each of which can have either the plesiomorphic form or one possible apomorphic form (so there are assumed to be only two possible character states, for the sake of simplicity). A useful symbolism is to introduce variables A_1, A_2,...,A_N, B_1,..., B_N, C_1,...,C_N, where A_j, for instance, ranges over the possible states of the jth characteristic of species A. The letter refers to the species, the subscript denotes the characteristic, while the value of the variable gives the character state. Let the value 0 denote the plesiomorphic state, and 1 denote the apomorphic state. Then $A_1=1$ tells us that species A has the apomorphic form of the first characteristic, while $A_1=0$ says that the same characteristic in species A is in the plesiomorphic state, and so on. The statement $(A_j=1.B_j=1)$ indicates that there is an observed synapomorphy between species A and B in the jth characteristic and $(A_j=0.B_j=0)$ similarly indicates a symplesiomorphy.

These variables become random variables (in the technical sense of mathematical statistics) when it is meaningful to speak of the probability of all statements that can be formed from these variables, such as $(A_j=1.B_j=1.C_j=0)$. The notion of probability intended here is very much a theoretical notion, and is not defined in terms of relative frequencies, although we do eventually want relative frequencies to provide evidence for statements about probabilities. But for now it suffices to think of probabilities as purely theoretical, in the sense of being assigned on the basis of a particular phylogenetic hypothesis.

Of course, the structure of a phylogenetic tree, such as those in Figures 1, 2, and 3 by itself does not determine the probability value of any possible state of affairs, but it will place constraints on what values can be assigned. Different trees will place different constraints on the probabilities. Sober's idea is that under certain conditions these different constraints may be sufficient in order to compare different probabilities qualitatively without knowing what the values are. What constraints are placed on the probabilities by the trees (ABC), (AB)C, and A(BC)?

We will be considering only one characteristic, so the subscript of the random variables will be dropped. Sober makes the following theoretical assumptions, which apply generally to all phylogenetic hypotheses (Sober, 1984:224):

(i) *Intermediate probabilities:* all probabilities are strictly greater than 0 and less than 1.

(ii) *Conditional independence:* a common cause screens off one joint effect from another, and a more proximal cause of an effect screens off a less proximal cause from the effect. That is, a tree is Markovian and singly connected (Sober, 1984:224). In our case this condition is that for all i, j, and m:

$$P(X=i.Y=j/Z=m)=P(X=i/Z=m).P(Y=j/Z=m), \qquad (1)$$

where Z is either the variable D or E, and X and Y refer to any two of the sister species A, B, or C (see Figure 5).

(iii) *The required inequality:* for all transition probabilities from past events to later events of the form $P(X=i/Z=m)$, we have that

$$P(X=1/Z=1)-P(X=1/Z=0)>0 \qquad (2)$$

For example, this requires that [for hypothesis ((AB)C)] $P(A=1/D=1)-P(A=1/D=0)>0$, and $P(D=1/E=1)-P(D=1/E=0)>0$ (from which it follows that $P(A=1/E=1)-P(A=1/E=0)>0$, so the requirement is consistent). In the language of statistics, there is a positive regression (and therefore a positive covariance) from any variable to any 'later' variable linearly connected to the first in the tree.

(iv) *Primitive ancestry:* the character state at the root of all trees is always in the primitive state. That is, $E=0$.

We are now in a position to analyze the essential differences between any hypothesis constrained by the trees (ABC), (AB)C, and A(BC). Take any phylogenetic tree in which X and Y occur, and consider the subtree connecting E (at the root of the tree) to X and Y. This subtree can only be of two possible forms; the V-tree (as in Figure 4), or the Y-tree (as in Figure 5).

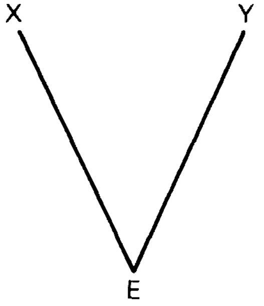

Figure 4: V-tree

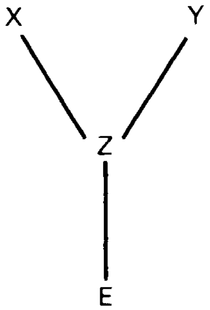

Figure 5: Y-tree

Take any probability function P that conforms to the constraints of the tree being considered. Letting $p=P(Z=1)$and $q=P(Z=0)$,

$$P(X=1.Y=1)=pP(X=1.Y=1/Z=1)+qP(X=1.Y=1/Z=0)$$
$$=pP(X=1/Z=1)P(Y=1/Z=1)+qP(X=1/Z=0)P(Y=1/Z=0),$$

and

$$P(X=1)P(Y=1)=p^2P(X=1/Z=1)P(Y=1/Z=1)+pqP(X=1/Z=1)P(Y=1/Z=0)+pqP(X=1/Z=0)P(Y=1/Z=1)+q^2P(X=1/Z=0)P(Y=1/Z=0).$$

Therefore,

$$P(X=1.Y=1)-P(X=1)P(Y=1)= pq[P(X=1/Z=1)-P(X=1/Z=0)].[P(Y=1/Z=1)-P(Y=1/Z=0). \qquad (3)$$

For the subtree in Figure 4, we can apply this result if we think of Z as being E, with the constraint that $P(Z=1)=0$ [assumption (iv)], to get

$$P(X=1.Y=1)-P(X=1).P(Y=1)=0. \qquad (4)$$

If the subtree is as in Figure 5, then $P(X=1.Y=1)\neq P(X=1)P(Y=1)$. But now assumption (iii) tells us that $P(X=1/Z=1)>P(X=1/Z=0)$ and $P(Y=1/Z=1)>P(Y=1/Z=0)$, and so;

$$P(X=1.Y=1)-P(X=1)P(Y=1)>0. \qquad (5)$$

Combining (4) and (5), we have proven the general result:

$$P(X=1.Y=1)-P(X=1)P(Y=1)\geq 0. \qquad (6)$$

This says that the expected covariance between the characteristics of any two species is always positive or zero, no matter which particular genealogical hypothesis happens to be true. This does not mean that an observed negative covariance is impossible, but that it is not so likely to occur, especially when the number of examined characteristics is high. But when a significant negative covariance is observed, Sober's assumption (iii) means that all genealogical hypotheses explain it badly. Genealogical hypotheses are not designed to explain negative covariances, and all hypotheses must write them off as arising from chance fluctuations. If such negative covariances were observed far more often than expected from chance, then this fact would reflect badly on assumption (iii)—so this assumption does have some empirical import, and is testable.

Let P_0, P_1, and P_2, be arbitrary probability functions applying to trees (ABC), (AB)C, and A(BC) respectively. Then we have the results: for all i, j, and k in $\{0, 1\}$;

$$P_0(A=i.B=j.C=k)=P_0(A=i)P_0(B=j)P_0(C=k), \qquad (7)$$

$$P_1(A=i.B=j.C=k)=P_1(A=i.B=j)P_1(C=k), \qquad (8)$$

$$P_2(A=i.B=j.C=k)=P_2(A=i)P_2(B=j.C=k), \qquad (9)$$

and, subsequently, the results;

$$Cov_0(A,B)=0, Cov_0(B,C)=0, Cov_0(A,C)=0 \qquad (10)$$

$$Cov_1(A,B)>0, Cov_1(B,C)=0, Cov_1(A,C)=0 \qquad (11)$$

$$Cov_2(A,B)=0, Cov_2(B,C)>0, Cov_2(A,C)=0 \qquad (12)$$

where $Cov_0(A,B)$ is the covariance on the basis of (ABC) and is defined as $P_0(A=1.B=1)-P_0(A=1)P_0(B=1)$, and similarly for the other terms.

Sober's Argument for Parsimony

The tree (ABC) is represented by the set of all probability functions $\{P_0\}$ that satisfy the constraints (7) and (10); (AB)C by the set $\{P_1\}$ satisfying (8) and (11), and A(BC) by $\{P_2\}$ satisfying (9) and (12). Any particular probability function is determined by the transition probabilities, together with Sober's theoretical assumptions (i) to (iv). A correspondence between probability functions in (ABC), or in (AB)C, or in A(BC), can be defined, therefore, by setting up a correspondence between transition probabilities as they apply to (ABC), (AB)C, and A(BC) respectively. This is what Sober does. The rule he uses is to identify phylogenetic "paths" and the transition probabilities associated with them as illustrated in Figures 6, 7, and 8.

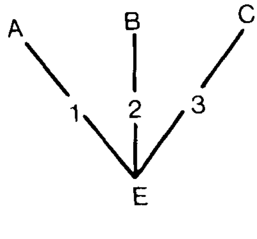

Figure 6: (ABC)

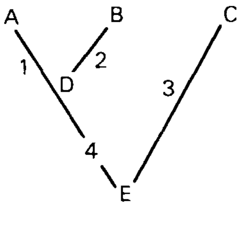

Figure 7: (AB)C

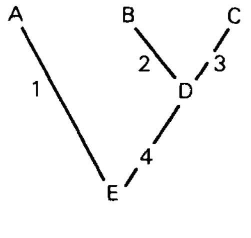

Figure 8: A(BC)

For path 1, for example, Sober assumes that $P_0(A=1/E=0)=P_1(A=1/D=0)=P_2(A=1/E=0)=e_1$, while for path 2 $P_0(B=1/E=0)=P_1(B=1/D=0)=P_2(B=1/D=0)=e_2$ and $P_1(B=1/D=1)=P_2(B=1/D=1)=q_2$, and so on. Once such cross-identifications of probabilities has been made, a simpler notation can be introduced. For any path, call it i, suppose that Z labels the species at the foot of the path, and Y the species at the end of the path (Y being later than Z in time). Then let

$$P(Y=1/Z=1)=q_i \text{ and } P(Y=1/Z=0)=e_i, \qquad (13)$$

and consequently,

$$P(Y=0/Z=0)=1-q_i \text{ and } P(Y=0/Z=0)=1-e_i. \qquad (14)$$

The subscript has been dropped from the probability function, not because the probability functions P_0, P_1, and P_2 are the same (they are not) but because they are the same for these conditional probabilities (by assumption). The general rule of path identification, of which this is a special case, is roughly that the path from any species A, say, back to its nearest ancestor shared with another species, is identified in both cases. In (AB)C the path from A goes back to D only, but in the (ABC) and A(BC) trees it must be traced back to E.

Both P_1 and P_2 can now be expressed as (different) functions of the common parameters q_i and e_i, for $i=1, \ldots, 4$. There are two cases of particular interest to us: the probabilities of the observations $(A=1.B=1.C=0)$ and $(A=0.B=0.C=1)$.

$$P_0(110)=e_1e_2(1-e_3), \qquad (15)$$

$$P_1(110)=[(1-e_4)e_1e_2+e_4q_1q_2](1-e_3), \qquad (16)$$

$$P_2(110)=e_1[(1-e_4)e_2(1-e_3)+e_4q_2(1-q_3)], \qquad (17)$$

$$P_0(001)=(1-e_1)(1-e_2)e_3, \qquad (18)$$

$$P_1(001)=[(1-e_4)(1-e_1)(1-e_2)+e_4(1-q_1)(1-q_2)]e_3, \qquad (19)$$

$$P_2(001)=(1-e_1)[(1-e_4)(1-e_2)e_3+e_4(1-q_2)q_3]. \qquad (20)$$

Sober is now able to prove that no matter what the values of these parameters are, provided only that assumptions (i) and (iii) hold (i.e., $1>q_i>e_i>0$, for all i, in this new notation), $P_1(110)$ is strictly greater than $P_2(110)$, and $P_1(001)$ is strictly less than $P_2(001)$.

These results are the key to Sober's likelihood justification of parsimony. The result that $P_1(110)>P_2(110)$ is intended to support the judgment of parsimony in the sense of rule 1, since on the evidence of one apomorphy, shared by A and B, and none between A and C or B and C, (AB)C is favored over A(BC) and (ABC). But note that Sober's argument is incomplete because it does not follow from $P_1(110)>P_2(110)$ that (AB)C is favored by likelihood when A and B share most, but not all, apomorphies. We will see later in this section that this last step is, in fact, fallacious (objection 3).

The second conclusion, that $P_1(001)$ is less than $P_2(001)$ and $P_0(001)$, together with the fact that the ordering between $P_2(001)$ and $P_0(001)$ is indeterminate, is seen by Sober as supporting the view that observed symplesiomorphies have no particular evidential value at all. This is Sober's second argument for parsimony, in the sense of rule 2, and will be rebutted in the last section of this paper.

In what follows, I will argue that both arguments are subject to a number of objections, and all of these faults should be blamed on the path identifications, as exemplified in Figures 6, 7, and 8. Once these cross-identifications of transition probabilities are given up, both of Sober's arguments dissolve, and their conclusions must be re-evaluated. This will be done in the last section.

The first two objections are designed to raise doubts about the "naturalness" of the path identifications codified in Figures 6, 7, and 8. The third objection shows that even if we suppress our doubts about that, it does not do the job of justifying the method of parsimony even in clear-cut cases.

Objection 1: The mapping is not one-to-one. Referring to Figure 7, it is obvious that the transition probability $P_1(C=1/E=1)$ plays no role in determining the value of $P_1(A=i.B=j.C=k)$ because it has been assumed that $P_1(E=1)=0$. But the cross-identification of transition probabilities requires that $P_2(C=1/D=1)=P_1(A=1/E=1)$, and $P_2(C=1/D=1)$ does help determine the value of $P_2(A=i.B=j.C=k)$. Therefore, if we vary the value of $P_1(A=1/E=1)$ we can change the value of P_2 without changing P_1. This in itself need not be damaging, but it is very odd that the values of $P_1(C=1/E=1)$ should play a role in determining the likelihood ratios, when it is also assumed that $P_1(E=1)=0$.

Objection 2: The mapping is not invariant under extensions of the model. Both (AB)C and A(BC) provide incomplete models of the phylogenetic history of the three species A, B, and C, in the sense that their relationship with other species is not considered. Suppose we include one other species, F, in the picture. There are at least three different ways of extending the hypothesis (AB)C to include F. These include ((AF)B)C, (A(FB))C, and (AB)(CF). Similarly, A(BC) may be extended to (AF)(BC), A((BF)C), or A(B(FC)). Conversely, if we choose to ignore F, the first three hypotheses reduce to (AB)C and the last three to A(BC).

Suppose we compare hypothesis (AB)(CF) with (AF)(BC) as diagrammed in Figures 9 and 10 respectively:

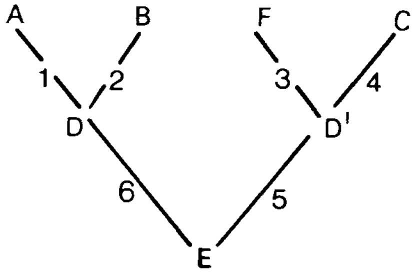

Figure 9: (AB)(FC)

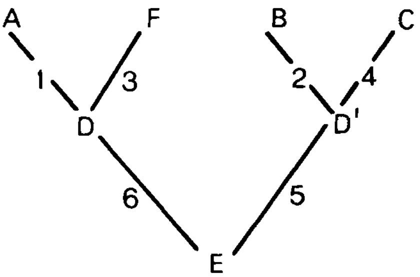

Figure 10: (AF)(BC)

From the assumptions (i), (ii), (iii), and (iv) stated earlier, the constraints on any probability function P_9 conforming to (AB)(FC), and P_{10} conforming to (AF)(BC) are, respectively:

$$P_9(A=i.B=j.C=k.F=1)=P_9(A=i.B=j)P_9(C=k.F=1), \tag{21}$$

$$P_{10}(A=i.B=j.C=k.F=1)=P_{10}(A=i.F=1)P_{10}(B=j.C=k). \tag{22}$$

If we now choose to ignore F, then (AB)(FC) reduces to (AB)C and (AF)(BC) reduces to A(BC), and indeed the constraint (21) reduces to (8) and constraint (22) reduces to (9). For example,

$$\begin{aligned}P_9(A=i.B=j.C=k)&=P_9(A=i.B=j.C=k.F=1)+P_9(A=i.B=j.C=k.F=0)\\&=P_9(A=i.B=j)P_9(C=k.F=1)+P_9(A=i.B=j)P_9(C=k.F=0)\\&=P_9(A=i.B=j)P_9(C=k).\end{aligned}$$

Given this natural reduction of the constraint for (AB)(FC) to the constraint for (AB)C and that for (AF)(BC) to A(BC), it is also natural to require that Sober's mapping between (AB)(FC) and (AF)(BC) should reduce to that between (AB)C and A(BC) when we eliminate F. But Sober's mapping does not fulfill this requirement, and this makes a difference. It suffices to prove that the ordering between $P_9(001)$ and $P_{10}(001)$ is different than between $P_1(001)$ and $P_2(001)$. In particular,

$$P_9(001)=[(1-e_6)(1-e_1)(1-e_2)+e_6(1-q_1)(1-q_2)][(1-e_5)e_4+e_5q_4], \tag{23}$$

$$P_{10}(001)=[(1-e_6)(1-e_1)+e_6(1-q_1)][(1-e_5)(1-e_2)e_4+e_5(1-q_2)q_4], \tag{24}$$

from which we can prove that

$$P_9(001)-P_{10}(001)=(e_2-q_2)[e_6(1-e_5)(1-q_1)e_4-(1-e_6)e_5(1-e_1)q_4].$$

For many values of the parameters e_6 and e_5 (e.g., whenever $e_6=e_5$), we can prove that

P_9 (001) is strictly greater than $P_{10}(001)$, whereas $P_1(001)$ is invariably less than the corresponding $P_2(001)$. This shows that the correspondence set up between P_9's and P_{10}'s has different properties entirely. Hence, the mapping set up between the probability functions in (AB)C and A(BC) is not invariant under extensions of the models, in the sense defined, and any conclusion based upon such a mapping is suspect. Looking at the situation in Figures 9 and 10, we can justify the conclusion that a symplesiomorphy between A and B is positive evidence of a close phylogenetic relationship between A and B, contrary to the conclusion arrived at by comparing Figures 7 and 8. This fact reflects badly upon Sober's method of likelihood comparison.

Objection 3: The likelihood comparison between (AB)C and A(BC) is not universal for all parameter values, q_i and e_i, for clear-cut cases such as example 3. It will be remembered that Sober's argument for choosing (AB)C over A(BC) in any particular example is that for every P_2 in A(BC), the corresponding P_1 in (AB)C will give a higher likelihood. In example 3, however, it is clear that Sober's notion of parsimony favors hypothesis A(BC), but contrary to Sober's intention, it is not true that for every P_2, the corresponding P_1 (determined by Sober's mapping) will give the data higher likelihood. Assuming that each characteristic conforms to the same probability distribution, and that the characteristics are mutually independent, then

$$\text{Likelihood}[(AB)C] = \prod_i \prod_j \prod_k P_1(ijk)^{n(ijk)}. \tag{25}$$

Similarly:

$$\text{Likelihood}[A(BC)] = \prod_i \prod_j \prod_k P_2(ijk)^{n(ijk)}. \tag{26}$$

No comparison can be made until the probabilities have been assigned numerical values. Parsimony clearly judges in favor of A(BC) in this example, but the method of comparing covariances legislates unequivocally in favor of (AB)C. Moveover, the best P_1 in (AB)C gives the data a higher probability then all P_2's in A(BC), as will be proved below. So, this P_1 must give a higher probability to the data than its corresponding P_2 under Sober's mapping (or any mapping for that matter). Therefore, Sober's method of likelihood comparison cannot give a universal judgment in favor of A(BC) over (AB)C, and thus cannot justify the judgment of parsimony in this example. Sober has not succeeded in justifying the method of parsimony in terms of likelihood. In fact, no method based on Sober's general strategy can succeed in justifying parsimony, for it must either agree with the covariance method in this example, or remain neutral.

The next section will abandon Sober's idea of mapping transition probabilities between different phylogenetic trees, and instead use the more direct method of comparing the best explanation of each hypothesis. This method provides a strong proof of the method of comparing covariances as described in rule 3, thereby strengthening the argument against parsimony.

A Likelihood Justification of the Covariance Method

Sober's strategy of justifying parsimony by comparing every probability function, P_1, in (AB)C with its "natural" counterpart in A(BC) does not work because the mapping fails to meet some basic desiderata. These faults might be corrected by providing a more satisfactory cross-identification of paths, and their associated transition probabilities, than depicted in Figures 6 and 7. Fortunately, this patching-up is unnecessary, for there is a second more direct strategy available. All we need to do is compare the best explanation of the data in terms (AB)C with the best explanation in terms of A(BC).

There are two distinct steps in this justification process. First, we need to identify the closest fitting probability function consistent with a particular genealogical hypothesis, such as (AB)C or A(BC), where this is defined as the one with maximum likelihood. Second, we see whether the best from (AB)C has greater likelihood than the best from A(BC), and make our choice on that basis. If this choice corresponds with the judgment of rule 3 under general conditions, then we have a powerful likelihood justification of that method.

The relationship between theoretical probabilities and observed relative frequencies can be properly established on the basis of additional assumptions. If we assume that the same probability function applies to all characteristics, so that for all j and k;

$$P(A_j =1.B_j =1.C_j =1)=P(A_k =1.B_k =1.C_k =1), \tag{27}$$

and so on, then the observed relative frequencies for a large number of characteristics will provide reliable information about the probability function P. If we further assume that each characteristic is independent of the others, then we can expect (in the probabilistic sense) relative frequencies of observed characteristics to conform to the true theoretical probabilities. The degree to which we can expect this will increase with increasing N (the number of observed characteristics), and the observed relative frequencies will almost always exactly fit the (true) theoretical values as N becomes infinite. This theorem of probability is known as the law of large numbers.

Assumption (27) says, in effect, that the rates of evolution are the same for each characteristic. Fortunately, this strong assumption can be weakened. As long as we insist that all character transitions are indeterministic—i.e., no transition probabilities are exactly 0 or 1 [Sober's assumption (i)]—then the desired result will still follow: define observed covariance as the statistic

$$\mathrm{Cov}(X,Y)=\frac{1}{N}\sum_{i=1}^{N} X_i.Y_i-[\frac{1}{N}\sum_{i=1}^{N} X_i][\frac{1}{N}\sum_{J=1}^{N} Y_j]. \tag{28}$$

If this statistic is less than or equal to 0, then this fact will still lend greater support to hypotheses with subtrees as in Figure 4, whereas a positive value will give greater support to trees corresponding with Figure 5. As long as the "theoretical" covariance for every character individually is greater than 0, then the expected value of Cov (X,Y) (on that hypothesis) will also be greater than 0. So, when Cov (X,Y) is observed to be positive, this fact will still lend greater support to the hypothesis that "expects" this to occur. The only real effect of weakening (27) might be to alter the expected sampling variation of Cov(X,Y) from its mean value.

The assumption that characteristics are mutually independent is equally innocuous. Its satisfaction is ensured so long as we individuate characteristics correctly. If two traits are not independent (e.g. pleiotropic traits) then they should be grouped together and counted as one trait, and so on, until we are left with a set of mutually independent characters (the practical problems involved in this are shared by any taxonomic method). Then the likelihood relative to the total evidence is obtained by multiplying the likelihoods obtained for each character separately.

Intuitively, the best explanation of the observed facts will be the hypothesis whose theoretical probabilities are exactly equal to the observed relative frequencies. This assertion can be mathematically justified by showing that such a hypothesis will have the maximum possible likelihood of all possible stochastic models. Because the proof of this is of basic importance to any likelihood proof, it is worthwhile stating the result explicitly. Suppose that we observe numbers n(ijk) of different character patterns (ijk), where i, j, and k are either 0 or 1. Then the total number of characteristics observed is the sum of these numbers:

$$\sum_i \sum_j \sum_k n(ijk)=N. \tag{29}$$

The likelihood of any (AB)C or A(BC) hypothesis, with probabilities $P_1(ijk)$ and $P_2(ijk)$ respectively, has been written in equations (25) and (26). Now, it is often convenient to compare two likelihood values by comparing the difference in their logarithms. Following Edwards (1972), we will write log (Likelihood[P]) more simply as Support [P]. The logarithm is a monotonically increasing function of its argument, so that Support $[P_1] \geq$ Support $[P_2]$ if, and only if, Likelihood $[P_1] \geq$ Likelihood$[P_2]$. Another property of the log function is that it turns a product into a sum in the sense that

$$\begin{aligned}\text{Support}[P] &= \log[\prod_i \prod_j \prod_k P(ijk)^{n(ijk)}] \\ &= \sum_i \sum_j \sum_k \log P(ijk)^{n(ijk)} = \sum_i \sum_j \sum_k n(ijk)\log P(ijk).\end{aligned} \tag{30}$$

But now assume that the other probabilities $R(ijk)$ conform exactly to the relative frequencies, i.e., for all i, j, and k, $R(ijk)=n(ijk)/N=r(ijk)$. Making use of the well known properties of the log function, we have

$$\text{Support}[R]-\text{Support}[P]=N \sum_i \sum_j \sum_k r(ijk)[\log R(ijk)/P(ijk). \tag{31}$$

It is one of the basic theorems of information theory that this last expression is always greater than or equal to zero, and equal only if $P(ijk)=r(ijk)$ for all i, j, and k.

THEOREM: Support$[R] \geq$ Support$[P]$, and Support$[R]$=Support$[P]$ if, and only if $P(ijk)=r(ijk)(=R(ijk))$. The only assumption used in the proof is that $R(ijk)$ and $P(ijk)$ are both probability functions (the proof is given in any comprehensive textbook on information theory; the proof I have seen is in Williams, 1980:133).

What this says is: if we want to maximize Support[P], we should choose $P(ijk)=r(ijk)$, if we are free to do so. This proves that the likelihood of the hypothesis whose theoretical probabilities exactly match the relative frequencies exhibited by the data cannot be exceeded by any other hypothesis whatsoever, and this will prove to be a useful result in what follows.

Any given set of data can be tabulated in terms of the relative frequencies $r(A=1.B=j.C=k)$, where i, j, $k=0$ or 1. We can write the Support function for a probability P_1 satisfying constraint (8) for tree (AB)C as:

$$\begin{aligned}\text{Support}[P_1] &= N.\sum_i\sum_j\sum_k r(A=i.B=j.C=k)\log P_1(A=i.B=j.C=k) \\ &= N.\sum_i\sum_j\sum_k r(A=i.B=j.C=k)\log P_1(A=i.B=j)+N.\sum_i\sum_j\sum_k r(A=i.B=j.C=k)\log P_1(C=k) \\ &= N.\sum_i\sum_j r(A=i.B=j)\log P_1(A=i.B=j)+N.\sum_k r(C=k)\log P_11(C=k).\end{aligned}$$

We now wish to choose P_1 such that Support $[P_1]$ is maximal (and this will automatically maximize likelihood $[P_1]$ as well). This is achieved by maximizing each of the two terms in the expression above individually because each is independent of the other.

The THEOREM provides the solution to this problem. In order to maximize the first term of Support$[P_1]$, we should choose $P_1(A=i.B=j)=r(A=i.B=j)$ and we choose $P_1(C=k)=r(C=k)$ to maximize the second term. Hence, the maximal function P_1 obeying constraint (8) is given by:

$$P_1(A=i.B=j.C=k)=r(A=i.B=j)r(C=k). \tag{32}$$

As it stands, the result is not completely general, for Sober's assumption (iii) further restricts the choice of $P_1(A=i.B=j)$ to satisfy the constraint

$$P_1(A=1.B=1)>P_1(A=1).P_1(B=1). \tag{33}$$

If the relative frequencies satisfy this constraint, i.e., if the observed covariance $\mathrm{Cov}(A,C)>0$, then we can consistently choose $P_1(A=i.B=j)$ to be equal to $r(A=i.B=j)$. But in the case where the observed covariance $\mathrm{Cov}(A,B)\leq 0$, (32) is not an option.

So let us consider the case in which Cov(A,B) is negative. We are still free to choose $P_1(C=k)=r(C=k)$ as before, but the optimum value of $P_1(A=i.B=j)$ must be chosen to satisfy (33). Towards this end, write:

$$\sum_i \sum_j r(A=i.B=j)\log P_1(A=i.B=j)$$

$$=\sum_i \sum_j r(A=i.B=j)\log[P_1(A=i.B=j)/r(A=i)r(B=j)]+\sum_i \sum_j r(A=i.B=j)\log[r(A=i)r(B=j)]. \quad (34)$$

The last term is constant, and does not enter into the maximization problem. We have assumed that $\mathrm{Cov}(A,B)=-K$, where $K>0$, so we can write:

$$\begin{gathered} r(A=1.B=1)=r(A=1)r(B=1)-K \\ r(A=1.B=0)=r(A=1)r(B=0)+K \\ r(A=0.B=1)=r(A=0)r(B=1)+K \\ r(A=0.B=0)=r(A=0)r(B=0)-K \end{gathered}$$

The first term in (34) now expands to:

$$\sum_i \sum_j r(A=i)r(B=j)\log\frac{P_1(A=i.B=j)}{r(A=i)r(B=j)}+(-K)\log\frac{P_1(A=1.B=1)P_1(A=0.B=0)}{P_1(A=1.B=0)P_1(A=0.B=1)}\,.$$

The first term of this expression is ≤ 0 by the THEOREM, and the second term is $\leq$ 0 because

$$P_1(A=1.B=1)P_1(A=0.B=0)>P_1(A=1.B=0)P_1(A=0.B=1),$$

as follows from constraint (33). Both of these terms take on their maximum values, 0, when $P_1(A=i.B=j)=r(A=i)r(B=j)$. Therefore, our sought solution for the case Cov(A,B) ≤ 0 is:

$$P_1(A=i.B=j.C=k)=r(A=i)r(B=j)r(C=k). \quad (35)$$

Strictly speaking (35) is inconsistent with (33), since (35) implies that $\mathrm{Cov}_1(A,B)=0$, whereas (33) says that $\mathrm{Cov}_1(A,B)>0$. But P_1, satisfying (33), can come arbitrarily close to the solution (35) without violating (33), so this blemish does not matter.

Likewise for A(BC), the best fit with the data when $\mathrm{Cov}(B,C)>0$ is given by:

$$P_2(A=i.B=j.C=k)=(r(A=i)r(B=j.C=k) \quad (36)$$

and when $\mathrm{Cov}(B,C)\leq 0$

$$P_2(A=i.B=j.C=k)=r(A=i)r(B=j)r(C=k). \quad (37)$$

Notice the interesting consequence of these solutions that

$$\begin{gathered} P_1(A=1)=r(A=1)=P_2(A=1) \\ P_1(B=1)=r(B=1)=P_2(B=1) \\ P_1(C=1)=r(C=1)=P_2(C=1). \end{gathered}$$

This means that the best of (AB)C and the best of A(BC) agree on what rates of evolution apply to each species, as we would hope in light of the fact the values of $r(A=1)$, $r(B=1)$, and $r(C=1)$ provide the only empirical information we have about what these rates are.

We also automatically satisfy the desideratum of objection 2 of the last section. For suppose we have an extended set of data with relative frequencies $r(A=i.B=j.C=k.F=l)$, for i, j, k, and $l=0$, 1. By the same argument as before, we can prove that, if Cov(A,B)

>0, the optimal fit for the genealogical hypothesis (AB)(CF) is given by either

$$P_8(A=i.B=j.C=k.F=l)=r(A=i.B=j)r(C=k.F=l)$$

or

$$P_8(A=i.B=j.C=k.F=l)=r(A=i.B=j)r(C=k)r(F=l),$$

and in both cases the optimal solution for (AB)(CF) reduces to the optimal solution for (AB)C, when we eliminate F by summing over l.

The second step in this likelihood justification of covariance methods is to compare the best of (AB)C, as given by (32) or (35), with the best of A(BC), as given by (36) or (37). This is done by considering the 4 possible cases in turn:

Case 1: Cov(A,B)≤ 0 and Cov(B,C)≤ 0. Here, P_1 is given by (35) and P_2 by (37), and they are the same; and so there is nothing to decide between (AB)C and A(BC) on the basis of likelihood. This agrees with the judgment of covariances since there are no synapomorphies occurring above those that are expected to occur on the assumption of independence.

Case 2: Cov(A,B)>0 and Cov(B,C)≤ 0. The optimal probability for (AB)C is given by (32), while (37) gives the best fit for A(BC). (AB)C is favored over A(BC) because the following expression is greater than 0:

$$\sum_i \sum_j \sum_k r(A=i.B=j.C=k)\log\frac{r(A=i.B=j)r(C=k)}{r(A=i)r(B=j)r(C=k)}=\sum_i \sum_j r(A=i.B=j)\log\frac{r(A=i.B=j)}{r(A=i)r(B=j)}>0.$$

This follows from THEOREM, and so (AB)C is always preferable to A(BC) in this case. Again this conclusion agrees with the covariance method, since the only shared apomorphies occurring above what we expect on the assumption of independence are between A and B.

Case 3: Cov(A,B) ≤ 0 and Cov(B,C)>0. Analogous to case 2, the best of A(BC) has the higher likelihood, and this accords with the method of comparing covariances.

Case 4: Cov(A,B)>0 and Cov(B,C)>0. This case is more difficult and the covariance method is only vindicated here within certain limits. Let Cov(A,B)$=K_1>0$, and Cov(B,C)$=K_2>0$. Rule 3 favors (AB)C if, and only if $K_1>K_2$. Likelihood will favor (AB)C exactly if the following expression is >0:

$$\sum_i \sum_j \sum_k r(A=i.B=j.C=k)\log\frac{r(A=i.B=j)r(C=k)}{r(A=i)r(B=j.C=k)}=$$

$$\sum_i \sum_j r(A=i.B=j)\log\frac{r(A=i.B=j)}{r(A=i)r(B=j)}-\sum_j \sum_k r(B=j.C=k)\log\frac{r(B=j.C=k)}{r(B=j)r(C=k)}.$$

By the THEOREM, the first term is positive and the second term is negative; so the likelihood decision depends upon the comparative magnitude of the two terms.

By writing $r(A=i.B=j)=r(A=i)r(B=j)\pm K_1$, etc., the above expression expands to

$$\sum_i \sum_j r(A=i)r(B=j)\log\frac{r(A=i.B=j)}{r(A=i)r(B=j)}+K_1\log\frac{r(A=1.B=1)r(A=0.B=0)}{r(A=1.B=0)r(A=0.B=1)}-$$

$$\sum_j \sum_k r(B=j)r(C=k)\log\frac{r(B=j.C=k)}{r(B=j)r(C=k)}-K_2\log\frac{r(B=1.C=1)r(B=0.C=0)}{r(B=1.C=0)r(B=0.C=1)}.$$

Now consider the special case in which $r(A=1)=r(C=1)$, so that $r(A=1)r(B=1)=r(B=1)r(C=1)$. In other words, assume that the yardstick against which the number of shared apomorphies are designated as "explainable" in terms of common ancestry is the same for A and B as for B and C. Now, if $K_1>K_2$, then the first term of the expression above is greater than the third term and the second term is greater than the fourth term, and the total expression is positive, thereby favoring (AB)C over A(BC). Similarly, if $K_2 > K_1$, the total expression is negative, favoring A(BC) over (AB)C. This agrees exactly with the judgment of rule 3.

When the proportion of apomorphic states for species A is approximately the same as that for C, then likelihood agrees with rule 3, in case 4. In all other cases, likelihood agrees with rule 3 unconditionally. In addition, it is worth noting that likelihood always accords with rule 3 at least qualitatively in the following sense: if we fix K_2, say, but imagine K_1 to vary, then as K_1 increases, the weight of evidence, as measured by likelihood, will shift towards (AB)C and away from A(BC), and vice versa. This shows that covariances are qualitatively the correct measures of evidential support in all cases, even though the 'neutral' point at which (AB)C and A(BC) have the same likelihood does not necessarily occur exactly when $K_1=K_2$ [as is possible if $r(A=1)$ differs greatly from $r(C=1)$]. The significance of the quantities Cov(A,B) and Cov(B,C) in theory choice as dictated by rule 3 has been completely vindicated, at least if we accept Sober's theoretical framework.

Conclusions

The starting point of this paper has been to treat taxonomic data as essentially statistical, and as such, that data should be given a probabilistic explanation. Sober has provided a theoretical framework in which the postulated phylogenetic tree imposes constraints on what probabilistic explanations are possible. The fact that different trees place different constraints on what probabilities apply leads to a way of choosing between phylogenetic hypotheses on the basis of maximizing likelihood. If this likelihood judgment coincides with any particular cladistic method in all cases, then we have an airtight likelihood justification of that method.

Farris (1983) made a vital contribution in pointing out that no genealogical hypothesis is designed to explain observed apomorphies that arise independently from parallel or convergent evolution. Within Sober's probabilistic framework, this point has a clear explication. The analogy of drawing marbles independently from two different urns shows that, on the basis of independence, we can expect a certain proportion of marbles to match. But such matchings are in no need of any explanation, unless they occur far more frequently, for large numbers of trials, than would be expected on the assumption of independence. It is only an observed covariance that requires a "common cause" explanation, and not matchings per se.

This led to a new cladistic method of classification requiring us to explain as many synapomorphies as possible above what would be expected on the basis of independent evolution. The surprising consequence of this is that the minimization of homoplasies has no special explanatory importance. This new taxonomic principle is incompatible with the method of parsimony and the minimization of homoplasies.

How do we decide which view is correct? Sober argues for a number of theoretical assumptions, (i), (ii), (iii), and (iv), from which he claims to justify the minimization of homoplasies as a correct rule. I agree with the assumptions, but not with the conclusion. The THEOREM states that the closer that the predicted probabilities fit the observed relative frequencies, the better the explanation of the data. It is not a trivial matter to prove that this reduces to the dictum that we should match covariances as closely as possible, but this is the result that has been proven in the last section.

The key to that problem was deciding how to make likelihood judgments about genealogical hypotheses in the first place, for these do not determine a unique theoretical probability function, but only a class of probability functions defined by some constraint. The problem is to compare one class of probabilities with another. Sober's solution was to pair the probability functions of one class with those of a second class and compare the members of each pair. No such approach to the problem can ever provide a likelihood justification of parsimony within Sober's theoretical framework. My alternative is to compare the best of one class of probability functions with the best of the other. This has led to some positive results in the preceeding section, where it was shown that the likelihood decision between (AB)C and A(BC) corresponds exactly with rule 3 in most cases.

In his justification of parsimony, Sober directed his attention to two cases of single observations of one characteristic: (a) $(A=1.B=1.C=0)$, and (b) $(A=0.B=0.C=1)$. His arguments for parsimony consisted in arguing that, for all possible transition probabilities (a) $P_1(A=1.B=1.C=0)>P_2(A=1.B=1.C=0)$ (b) $P_1(A=0.B=0.C=1)<P_2(A=0.B=0.C=1)$. From (a) Sober concluded that likelihood vindicates rule 1; and from (b) he hoped to justify rule 2. Sober's argument for (a) and (b) has been discredited, but are these conclusions correct anyway?

The "best of the best" likelihood strategy of this paper makes a judgment, not on the basis of one single observation by itself, but only in the context of a larger set of data. On my own assumptions we can prove that within any larger data set

$$P_1(A=1.B=1.C=k)\geq P_2(A=1.B=1.C=k), \tag{38}$$

for $k=0$ or $k=1$. Inequality (38) agrees with Sober's conclusion, (a), except that the inequality is no longer strict. So, within the context of a larger set of data, a synapomorphy is indeed positive (or at least non-negative) evidence of a phylogenetic relationship. But this result in no way vindicates parsimony's rule 1, because it is not true that total evidence from all (independent) characters is obtained by simply adding the total number of synapomorphic characters (cf. Felsenstein, 1981).

The next result totally overturns conclusion (b):

$$P_1(A=0.B=0.C=k)\geq P_2(A=0.B=0.C=k), \tag{39}$$

for $k=0$ or $k=1$. Sober's second conclusion (b) is always false, and so his defense of parsimony in the sense of rule 2 cannot be upheld. In fact, just the opposite conclusion must be drawn, namely, that symplesiomorphies do count as positive evidence for phylogenetic relationship, just as synapomorphies do. The underlying reason for this symmetry between symplesiomorphies and synapomorphies arises from the mathematical fact (true of any probability function) exhibited in equation (40):

$$P(X=1.Y=1)-P(X=1)P(Y=1)=P(X=0.Y=0)-P(X=0)P(Y=0) \tag{40}$$

In words, the number of synapomorphies above those expected on the basis of independence is equal to the number of symplesiomorphies above the number of those expected on the assumption of independence. Anyone who says that we should explain as many synapomorphies as possible (above the numbers expected on the assumption of independence), will automatically account for the excess number of symplesiomorphies (above the numbers expected on the assumption of independence) exhibited in the data as well. In matching a theoretical probability P to the observed relative frequencies, such that

$$P(X=1.Y=1)-P(X=1)P(Y=1)=r(X=1.Y=1)-r(X=1)r(Y=1),$$

We will automatically ensure that

$$P(X=0.Y=0)-P(X=0)P(Y=0)=r(X=0.Y=0)-r(X=0)r(Y=0).$$

When we maximize the number of explained synapomorphies, we automatically maximize the number of explained symplesiomorphies. But it is agreed, with rule 2, that symplesiomorphies do not provide evidence in addition to that provided by synapomorphies—to do so would be to count the same evidence twice. The present paper sides with parsimony in opposing classificatory methods that use 'overall similarity' as a measure of phylogenetic relationship.

Let me end with a conciliatory remark. In the majority of real examples, I believe that parsimony arrives at the right answer, but for the wrong reasons. Example 2 is the type of case I have in mind. In that example we observe 9 synapomorphies between A and B and 1 between B and C and many other characteristics in the ancestral states for all species under consideration. Advocates of parsimony reason as follows: first apply rule 2 to disregard the latter characteristics on the assumption that symplesiomorphies have no evidential meaning, and then count up synapomorphies to conclude that (AB)C is the best explanation. That conclusion is correct, but the rationale for it is not. Both rules 2 and 1 are wrong, but applying them together gives the right answer in such cases (the two errors cancel out). The correct rationale, according to this paper, is that the observed covariance between A and B calculated from the total data (symplesiomorphies included) is significantly higher than for B and C.

Acknowledgments

I have benefited greatly from extensive and helpful corespondence with Elliott Sober, and from the careful review of an anonymous referee.

LITERATURE CITED

EDWARDS, A. 1972. Likelihood. Cambridge Univ. Press, Cambridge.

FARRIS, J. S. 1983. The logical basis of phylogenetic inference *In* Platnick, N. I., and V. A. Funk (eds.), Advances in cladistics, vol. 2: Proceedings of the second meeting of the Willi Hennig Society. Columbia Univ. Press, New York, pp. 7-36.

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

FELSENSTEIN, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool. 28: 49-62.

FELSENSTEIN, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. Jour. Linn. Soc. 16: 183-196.

SOBER, E. 1983. Parsimony in systematics: Philosophical issues. Ann. Rev. Ecol. Syst. 14: 335-357.

SOBER, E. 1984. Common cause explanation. Philos. Sci. 51: 212-241.

SOBER, E. 1985. A likelihood justification of parsimony. Cladistics 1: 209-233.

WILLIAMS, P. M. 1980. Bayesian conditionalisation and the principle of minimum information. Brit. Jour. Phil. Sci. 31: 131-144.