NUMERICAL METHODS FOR INFERRING EVOLUTIONARY TREES

JOSEPH FELSENSTEIN

Department of Genetics,
University of Washington,
Seattle, Washington 98195 USA

ABSTRACT

Despite a century of evolutionary theory, only in the last few decades have clearly defined procedures for inferring phylogenies been stated. For discrete characters whose ancestral states are known, the prescriptions of Hennig are well defined, but they are applicable only when there is no incompatibility between different characters. This limitation has led to the elaboration of a number of methods for dealing with such incompatibilities. One category consists of the parsimony methods, which choose that phylogeny on which the fewest changes of character state need be assumed. Another category consists of the compatibility methods, which choose that phylogeny which is perfectly compatible with the largest number of characters, irrespective of how many changes need be assumed in other characters. Other approaches include the use of phenetic clustering algorithms and methods fitting trees to similarity or distance matrices.

Each method has a different set of implicit assumptions concerning the biology of the characters and the information available from the data. If the methods are considered in a statistical framework as different estimators of an unknown quantity (the phylogeny), these assumptions are more clearly seen. Standard statistical approaches, such as maximum likelihood, can be used to obtain methods whose properties are known and for which one can determine the amount of uncertainty in the resulting estimates of the phylogeny. Although existing statistical models are highly oversimplified and do not reflect the complexity of evolutionary processes, it is by viewing the problem as a statistical one that we can place all these methods in a common framework, within which their behavior and assumptions can be compared. It is essential that we not adopt a single method as a universal panacea, but that an attempt be made to understand the biological assumptions and statistical behavior of each method.

INTRODUCTION

FOR MOST OF THE past one hundred years, systematists have been content to leave the reconstruction of evolutionary trees (phylogenies) to biological intuition. Prescriptions for reconstructing phylogenies were not published, it being assumed that each practitioner would learn the process by apprenticeship. Simpson (1961) agreed with Cain’s statement (1959, p. 243) that “young taxonomists are trained like per-
forming monkeys, almost wholly by imita-
tion.” The scientific revolution in systematics
brought about by the acceptance of the neo-
Darwinian synthesis in the 1940s had its ef-
effect almost wholly at or below the species
level: the major proponents of the neo-Dar-
winian synthesis continued to leave inex-
plicit the methodology for reconstructing
phylogenies.

The textbook by Mayr, Linsley, and
Usinger (1953) is particularly instructive in
this respect. Of a total of 284 pages, less than
ten are devoted to discussing the methodol-
ogy for inferring phylogenies, and that dis-
cussion centers on the many pitfalls and dif-
ficulties in the process. No clear prescription
for inferring phylogenies is given, but in-
stead it is stressed that “because of the subjec-
tive nature of the problem, it is difficult to
lay down any hard and fast procedures for
attaining satisfactory results” (p. 168). Simp-
son’s book (1961) takes much the same ap-
proach, his concentration being on con-
structing a classification which only partly
reflects the phylogeny.

The availability of computers in the 1950s
led to the development of phenetic clustering
procedures (cf. Sokal and Sneath, 1963) for
classification, and this has in turn led to a
wave of interest in the development of expli-
cit numerical techniques for inferring phylo-
genies. That systematists were becoming
more receptive to numerical and algorithmic
approaches during this period may be seen
from the use in botanical work of Wagner’s
(1952, 1961) Groundplan Diagram, from
the impetus given to morphometrics by the
work of Olson and Miller (1958), and from
the increasing popularity of Hennig’s (1950)
approach to inferring phylogenies. The
availability of molecular data, starting in the
1960s, led to an independent line of develop-
ment of algorithms for inferring phylogenies.

**Hennig’s Method**

Hennig (1950, 1965, 1966) broke with tra-
dition by stating a reasonably clearly defined
procedure for inferring phylogenies. Hennig’s
method can be illustrated by reference to
Table 1. The table shows an imaginary set of
data involving five taxa, labelled A to E,
each of which has been scored for six char-
acters. Each character is found to occur in
two states, here called 0 and 1. Hennig’s pro-
cedure requires that we know for each char-
acter which of these states is ancestral and
which is derived. In each case we shall
assume that the ancestral state has been
called 0 and the derived state 1. For the mo-
moment we assume that only characters 1 to 4
are available to us.

Hennig takes each character as providing
evidence for the existence of one mono-
phyletic group in the phylogeny. This is based
on the further assumptions that each derived
state has arisen only once, and that it is im-
possible for a character having a derived
state to return to the ancestral state, evolu-
tion being irreversible in these characters.
These assumptions being accepted, char-
acter 1 provides evidence that AC is a mono-
phyletic group, in whose ancestral lineage
state 1 arose from state 0. Any other ar-
rangement would require two origins of state
1, or a reversion from state 1 to state 0.

Each of the other characters also delin-
eates a monophyletic group: character 2 sug-
gests group ACE, character 3 suggests BD,
and character 4 suggests D. A phylogeny can
be obtained by superimposing these mono-
phyletic groups: it is shown in Fig. 1. Note
that character 4 makes no real contribution
to the phylogeny, since we began by assum-
ing that each of the forms A through E was
itself monophyletic.

It should be noted that Hennig’s ter-
mindology differs from that used here. Hen-
nig (1966) referred to character states as
“character conditions,” to the derived states
as “apomorphous,” and to the ancestral states
as “plesiomorphous.”
This phylogeny is based on the first four characters of the data in Table 1. The bars mark the locations in which character state changes would occur if as few changes as possible are required in each character. The number of the character that is changing is shown beside each change.

Although Hennig has been the most influential exponent of this method, its essential logic was appreciated much earlier. For example, in their paper on the aardvark Orycteropus, Le Gros Clark and Sonntag (1926) stated that “Orycteropus shares primitive features with most Edentata, but these do not imply relationship” (p. 478), and “the features in which Orycteropus resembles Insectivores . . . are all primitive, and hence no test of affinity” (p. 479). The method is even older than this. Nelson and Platnick (1981) quoted extensively from a detailed description of it by Mitchell (1901), who described it as “merely a codification of criteria in common employment among naturalists.”

Hennig’s Dilemma and Its Resolutions

The chief difficulty with Hennig’s scheme is illustrated by the data in Table 1. If we include characters 5 and 6 in the data set, these both point to the existence of still another monophyletic group, BCE. But group BCE cannot be monophyletic as long as AC, ACE, or BD are accepted as monophyletic. There is no possible phylogeny according to which all of these groups can be monophyletic. Hennig’s method works only as long as there is no internal conflict in the data.

Hennig gave only one prescription for resolving such incompatibilities: to recheck one’s interpretation of the characters, presumably until all incompatibility in the data has been resolved by a more careful interpretation of the characters. Of course, it will not do to restrict one’s reinterpretation to one set of characters only (say numbers 5 and 6)—for they might be the ones giving correct information, whereas the others have been misinterpreted.

In certain cases no amount of reinterpretation may be able to resolve the incompatibility. In a protein sequence, the same amino acid may appear at the same position in widely separated lineages, and no amount of resequencing may be able to change an offending alanine into a glycine. Even at the DNA level, the codons for the two alanines may be the same. It is quite possible for the same codon to arise in two different lineages, and we are not justified in assuming that this is biologically impossible. It is necessary to have some prescription other than reinterpretation for resolving the incompatibility.

Two approaches to resolving incompatibilities seem possible. One approach allows the characters to have reversions from state 1 to state 0, or allows state 1 to arise more than once in each character, and finds that phylogeny which minimizes the number of these extra events. The other approach finds the largest set of characters that are mutually compatible, and uses only those characters to build the phylogeny. The first approach makes a compromise among the characters, a compromise with which no individual character may be entirely compatible. The second finds the phylogeny that is completely compatible with a plurality of characters, even though the remaining characters may be extremely incompatible with it.

There are existing numerical methods (parsimony and compatibility) which correspond to these two approaches. In the remainder of this article, we will explore these methods and other more complex procedures for inferring phylogenies, with particular attention to the implicit biological assumptions of each method. It is of particular importance to find some common logical
framework within which the various methods can be examined and compared. I will argue that statistical inference constitutes such a framework.

It is important to realize that we are not here discussing philosophies of classification. Hennig is known as much for his strict cladistic approach to classification as for his methods of inferring phylogenies. These are logically separable matters, and we address here only the latter. To avoid possible confusion, I will not use the ambiguous term "cladistics" in this article. Other review articles which the reader may find useful are those of Estabrook (1978), Moeller Anderson (1978), and Fitch (1973).

**PARSIMONY METHODS**

Parsimony or minimum-evolution methods were first proposed by Edwards and Cavalli-Sforza (1963, 1964), in an application using gene frequencies of human polymorphisms to make a phylogeny of human populations. This application is discussed below (pp. 384, 396). For the moment we concentrate on discrete character data.

(1) **Camin-Sokal Parsimony**

Suppose that we take the data in Table 1 and assume that reversions from state 1 to state 0 are impossible, but that there may be for each character more than one origin of state 1 in the phylogeny. Allowing multiple origins of state 1, we can attempt to compromise information from different characters by seeking that phylogeny which requires as few of these extra changes of character state as possible. This was the approach taken by Camin and Sokal (1965), who introduced the first discrete-characters parsimony method. They allowed their characters to have multiple states, and gave details of the algorithm which they used to search among possible topologies of the tree to find the one that required the fewest changes of character state. A set of FORTRAN programs to carry out all of Camin and Sokal's methods has been published (Bartcher, 1966).

In reading this literature, it is important to separate each procedure into three parts: (1) the quantity that is used to evaluate the merit of a particular phylogeny; (2) the algorithm for computing this quantity, given a particular phylogeny; and (3) the algorithm for searching among possible phylogenies to find the one of greatest merit. We will concentrate on the first of these, since it is the part of the procedure that embodies the biological assumptions. One of the commonest errors of newcomers to this literature is to pay too little attention to part (1), as a result of becoming bogged down in trying to understand the details of parts (2) and (3). In the method of Camin and Sokal (1965) the quantity being minimized is the number of extra changes (or the total number of changes—it does not matter which) required to explain the observed data on a particular phylogeny.

Fig. 2 shows two phylogenies, each with the locations of changes in each character shown. Fig. 2A shows the same phylogeny as Fig. 1 (the Hennigian phylogeny for characters 1 to 4), and Fig. 2B shows the parsimony solution using the Camin-Sokal criterion. The first of these requires 10 changes, the second 9.

(2a) **Wagner Parsimony**

Two of the restrictive assumptions of Hennig's method are (1) that we are able to specify for each character that state which is ancestral, and (2) that reversion from a derived state to the ancestral state is impossible. In real data we often will not know which state is ancestral, nor will we necessarily be confident of the irreversibility of state changes. It is possible to construct a parsimony method that does not require these assumptions. The original continuous-character minimum-evolution method of Edwards and Cavalli-Sforza (1964) does not specify ancestral states and does not assume irreversibility. The first such method for discrete-state data was devised by Eck and Dayhoff (1966, pp. 161–168) for protein sequence data, where the character states are the different amino acids possible at a site. They gave a partial description of their algorithms.

Kluge and Farris (1969) also stated this parsimony criterion, which they dubbed the "Wagner method," with morphological data in mind. They showed how to recode data so that one multi-state character, in which the order or connectedness of states is known,
FIG. 2. TWO PHYLOGENIES WITH THE NUMBERS OF CHANGES OF STATE SHOWN

Both of these phylogenies are based on the data shown in Table 1. For each of them, the changes of state have been assigned in such a way as to minimize the number of changes in each character. The changes of state are marked on the phylogenies by bars, the number of the character changing being shown next to each change of state. A, the same phylogeny as in Fig. 1; B, the result of applying the Camin-Sokal parsimony method.

For the data given in Table 1, the Wagner parsimony criterion finds a tree (Fig. 3) that is one step shorter than the Camin-Sokal criterion. In general, we may be able to find solutions under the Wagner criterion that require fewer changes of state, since we have fewer constraints on the result. An important difference between the criteria is that the Wagner method does not allow us to specify the location of the root: we obtain an unrooted tree (sometimes called a network). In Fig. 3 we could locate the root of the tree anywhere in the tree, and still obtain a tree requiring eight changes. To each possible placement of the root there corresponds an assignment of ancestral states to the characters. For example, if the root were placed at the fork which connects B to the tree, this would imply that the ancestral states in characters 1, 2, and 4 were 0, and the ancestral states in characters 3, 5, and 6 were 1. For this reason the position of the root is not indicated in the tree.

If the character states at the root are known in advance, a fictional taxon named (say) ANCESTOR can be introduced. When a Wagner tree is inferred, the position of this taxon indicates the position of the root. Kluge and Farris (1969) used this method, and Farris (1970) distinguished between Wagner Trees and Wagner Networks, on the basis of the availability of information about the ancestor.

(2b) Parsimony in Molecular Evolution

In its largely independent development in the study of molecular evolution, the Wagner parsimony criterion has had to cope with the degeneracy of the genetic code. Eck and Dayhoff (1966) counted changes in amino acids rather than changes in nucleotides. Subsequent workers have preferred to count the minimum number of base substitutions compatible with the observed amino acid sequences, although this is much more difficult and has generated a certain amount of work (Fitch, 1971; Moore, Barnabas, and Goodman, 1973; Moore, 1974; Fitch, 1974; Fitch and Farris, 1974; Moore, 1977).

In many cases it is necessary to allow for
This phylogeny is based on the data in Table 1. The locations of postulated changes of state are shown by bars, with the number of the character marked beside each. The phylogeny is intrinsically unrooted, as its root could be placed anywhere on this tree without affecting the numbers of changes of state reconstructed.

the possibility of insertions and deletions of nucleotides, in addition to nucleotide substitutions and amino-acid replacements (Sankoff, Morel, and Cedergren, 1973; Sankoff, 1975; Sankoff and Rosseau, 1975; Fitch and Yasonobu, 1975). A generalization of this problem is the task of simultaneously estimating the phylogeny and the sequence of gene-duplication events when dealing with a family of related proteins, such as hemoglobins, which have diverged from a single common ancestral gene. The problem that arises is how to weight gene-duplication events as compared to base substitutions (Goodman and Moore, 1973; Goodman, Moore, Barnabas, and Matsuda, 1974; Goodman, Czelusniak, Moore, Romero-Herrera, and Matsuda, 1979; Fitch, 1979; Goodman, Czelusniak, and Moore, 1979).

After they had chosen the best topology, Dayhoff and Eck (1968) computed estimates of the amounts of evolution in different branches of the phylogeny by correcting for presumed multiple amino-acid replacements in long branches. Various corrections have been used by Holmquist (1972a, b), Goodman, Moore, Barnabas, and Matsuda (1974), Moore (1977), and Langley and Fitch (1974). There has been considerable controversy over the adequacy of these correction methods (Moore, Goodman, Callahan, Holmquist, and Moise, 1976; Holmquist, Jukes, Moise, Goodman, and Moore, 1976; Tateno and Nei, 1978; Czelusniak, Goodman, and Moore, 1978; Holmquist, 1978a, b; Nei and Tateno, 1978; Karon, 1979; Nei and Tateno, 1979; Holmquist, 1979; Fitch, 1980; Holmquist and Conroy, 1981; Nei and Tateno, 1981; Holmquist and Jukes, 1981; Fitch, 1981b).

(2c) Parsimony with Continuous Variables

There is an analogy between the Wagner parsimony criterion and the minimum evolution criterion when applied to such continuous variables as gene frequencies or quantitative characters. Edwards and Cavalli-Sforza (1964) originally applied the minimum-evolution criterion to gene-frequency data, searching for the phylogeny whose segments had the smallest total length when drawn in a space whose coordinates were the gene frequencies. Thompson (1973) and Chang (1972) have discussed strategies of searching for the minimum-length phylogeny.

When the measure of distance between ends of a segment of the phylogeny is not the usual Euclidean distance but is, instead, the sum of absolute values of changes in characters, this “Manhattan metric” has the two-state Wagner Network method as a special case. If each coordinate of any point is either 0 or 1, the Manhattan distance measures the number of states that have changed in the segment, and we are attempting to minimize the total number of changes of character state that must be assumed. When there are many discrete states that can be arranged on a linear scale, the Manhattan metric continues to measure the number of state changes.

The difficulty that arises in attempting to apply the minimum-evolution criterion to quantitative characters lies in knowing how to combine differences in characters that may vary on extremely different scales. The simplest suggestion seems to be to scale the characters so that their within-population variances are equal, although the information necessary to do this is often lacking. Alternatives that have been suggested include (1) Kluge and Farris’s method of scaling each character so that the value of its total range between taxa is 1 (Kluge and Farris, 1969; Farris, 1969), and (2) the method
of "gap coding" employed by Mickevich and Johnson (1976), a method that takes a gap in the distribution of a character across taxa to have value 1. Both of these methods are sensitive to the distribution of characters across taxa, and thus might be affected by extinction of some lineages. None of the above methods takes correlations between characters into account.

Mickevich and Mitter (1981) have advocated turning gene-frequency data into discrete-state character data by coding each allele as 1 or 0 according to whether it is or is not present in the species. This method of coding has obvious limitations, in that it is unable to distinguish between extremely rare and extremely common alleles, and will be more likely to show a 1 the larger is the sample size from the species. No statistical examinations of this method have yet been done, so that it is not known how serious this loss of information is or whether it results in any bias in the resulting phylogeny.

(3) Dollo Parsimony

At least two other major parsimony methods are available. One, the Dollo parsimony method, is based loosely on Dollo's principle of the irreversibility of loss of complex phenotypes. If we assume that state 1 is a complex derived state and state 0 a simple ancestral state, it makes sense to try to find a phylogeny on which state 1 has arisen as few times as possible, and to explain the incompatibility of a character with a given phylogeny by allowing reversions from state 1 to state 0. The Dollo parsimony method allows no more than one forward change (0 1) but as many reversions (1 0) as are necessary. That phylogeny is preferred which requires the fewest reversions.

The Dollo method was suggested by Le Quesne (1974, 1977) and was formalized and clarified by Farris (1977a, b). I have pointed out (1979) a certain difficulty that arises in applying it to multi-state characters. The Dollo method can never find a solution with fewer changes of character state than the Wagner method, inasmuch as it imposes an extra constraint, namely, that no more than one change 0 1 is allowed per character.

Another parsimony method, the polymorphism parsimony method, allows incompatibilities to be explained by the existence of polymorphisms. In effect, there is assumed to be a third state, "01," which is a state of polymorphism. It is assumed that state 1 arose only once on the tree, and its appearance in different lineages is explained by postulating that after it arose, the population remained polymorphic for both states 0 and 1 for some time, and then that in some lineages state 0 was lost from the polymorphism, and in other lineages state 1 was lost. The criterion that is to be minimized is the total length of time during which polymorphism must have persisted, summed over all characters. Though this definition has been given in terms of two states, it is not hard to construct a multi-state version.

The polymorphism parsimony method was introduced by Farris (1978) and Felsenstein (1979). It is inspired by the problem that arises in inferring phylogenies from chromosome-inversion data in dipterans. Throckmorton (1965) argued that morphological characters in the Drosophila reproductive tract have followed a pattern of evolution similar to this model. Inger (1967) showed a phylogeny produced by a polymorphism-parsimony program using morphological data. Polymorphism parsimony methods may either assume or not assume a knowledge of which character state is ancestral. In the latter case, the tree produced is unrooted.

(4) Polymorphism Parsimony

Weighting Characters

In parsimony methods it is usual for all steps that are counted to be weighted equally. It is a simple matter, given the appropriate prior information, to construct weighted versions of these methods which sum the weights of all steps on the phylogeny. Kluge and Farris (1969) have shown how to incorporate weights into their Wagner Tree and Wagner Network criteria. The difficulty lies in acquiring the prior information as to what weights to assign to steps in different characters.

I have discussed the logic of character weighting, as illuminated by a statistical in-
ference framework (Felsenstein, 1981b). It turns out that unequal weights of changes in different characters are based implicitly on unequal probabilities of change in those characters. This accords well with the traditional view (which goes back to Darwin) that characters that are “conservative” should have high weight. A character whose probability of change in a segment of the phylogeny is $x$ has half the weight of one whose probability of change is $x^2$.

An implication of this correspondence between weights and probabilities of change is that as probabilities of change are taken to be small, weights in different characters become more and more equal. This conclusion seems to justify the common practice of assuming equal weights. It is interesting to note that it involves the implicit assumption of extremely low rates of change.

Farris (1969) has invented an interesting iterative weighting method, in which the weights of characters are changed according to how many changes they require to fit the tree. As the weights are changed, the tree can thereby be altered. Ultimately both the weights and the tree settle down into a final configuration. Using a statistical inference framework, I have arrived at a similar method, the threshold method (Felsenstein, 1981b). Both methods assume (one implicitly, the other explicitly) that weights differ among characters, but that we do not know which characters are the ones that have high weights.

**CORRELATED CHARACTERS**

A major assumption of all parsimony methods is that the characters evolve independently. There has been no attempt to provide methods of detecting character correlation and removing its effects from the analysis. In extreme cases it will have the same effect as improper weighting of characters. If two characters in a data table are actually repeated measurements of the same biological entity, so that they always evolve in concert, then each change of character state of the underlying character will appear in the analysis as two changes of character state. We will thus give too large a weight to each change, and too great an influence to the character.

The problem is not confined to parsimony methods—the same problem faces users of other methods of inferring phylogenies from character data, in particular the compatibility and direct statistical methods discussed below.

How to develop methods to remove the effects of character correlation is perhaps the most important unsolved problem facing phylogenetic inference. The absence of a solution is the greatest weakness of existing methods.

**Computational Difficulties**

The biological assumptions of all of the above methods affect the definition of the criterion used to evaluate phylogenies, but they do not necessarily affect the strategy used to search among all possible phylogenies for the one that maximizes (or minimizes) this quantity. The details of the search strategy are thus purely computational, but it should not be thought that the computation is an easy one. Since the beginnings of numerical work on estimating phylogenies there has been a persistent search for methods that would guarantee finding the phylogeny that maximizes (or minimizes) the criterion.

There is a simple method that would guarantee this: examine all possible trees one after another, while keeping a record of the best tree yet found. When all trees have been examined, we then know which one is best. If the phylogenies do not include estimated lengths of branches, but only topologies of branching events, then there are only a finite number of them, and the problem is solved. The difficulty is that the number of possible trees rapidly becomes enormous: for 20 taxa, allowing multifurcations, it exceeds Avogadro’s Number (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1978a)!

Hendrickson (1968) gave exact solutions to the problem of finding the best minimum-length tree under the Camin-Sokal criterion for two and three taxa, and pointed out the difficulty of finding a general solution. Estabrook (1968) made a great advance by stating a restriction on the solution and giving a recursive search procedure that is guaranteed to find the minimum-length tree, and that does not need to examine many of the
possible trees. Unfortunately, for most real data-sets Estabrook's method still requires searching too many trees to be practical, except when the data sets are small.

Nastansky, Selkow, and Stewart (1973) reinvented Estabrook's algorithm and added some further restrictions. Their extension of this algorithm (1974) yielded a more efficient method, but it is questionable whether it is practical for use with all but rather small data-sets. An interesting approach has been followed by Hendy, Penny, and Foulds (1978; Penny, 1978; Foulds, Hendy, and Penny, 1979; Foulds, Penny, and Hendy, 1979; Hendy, Foulds, and Penny, 1980; Hendy, 1980; Penny, Hendy, and Foulds, 1980). They divide the set of characters into subsets, for each of which they can determine the minimum number of steps possible.

The sum of these lower bounds must be less than or equal to the total number of steps on the tree of minimal length, and if they can find a tree of exactly that length, then it is guaranteed to be of minimal length. This method will not always yield a result, but it may be useful, particularly with small data-sets. Fitch (1975, 1977) had earlier considered other lower-bound methods of economizing on the search for the tree of minimal length. Although some of the above methods were designed to apply to the Camin-Sokal method and some to nucleotide-sequence data, which involves a multistate version of the Wagner Network criterion, most of them can be modified so as to apply to either.

The problem of finding phylogenies of minimal length is a special case of the general mathematical problem of finding Steiner Trees, minimal-length trees connecting points in a graph or space. Levin (1971) and Dreyfus and Wagner (1972) have given general algorithms for finding Steiner Trees in a graph of arbitrary structure, and Gilbert and Pollak (1968) have given conditions that aid the finding of Steiner Trees in Euclidean space, as have Sankoff and Rosseau (1975). The algorithms available in this mathematical literature are sufficiently slow as to not be helpful in the case of phylogenies.

In recent years is has become possible to prove that a given computational problem is in a class of problems (the “NP-complete” problems) that are all equally difficult. None of the NP-complete problems is known to have an efficient algorithm that solves it, and it is known that if one of them has such an algorithm, all do! Since these include most of the best-known intractable computational problems, in effect to show a problem to be NP-complete is a way of showing that there is no efficient algorithm that is guaranteed to solve it. This theory is summarized by Garey and Johnson (1979). A brief elementary description of the theory is contained in the article by Bland (1981).

One parsimony problem, the Wagner Network problem with two characters each of which has states on an integer scale, is known to be NP-complete (Garey and Johnson, 1977). The Wagner parsimony method with two states, 0 and 1, has also been proven to be NP-complete (Foulds and Graham, in press).

Fortunately, things are really not as bad as they seem. The rough-and-ready search methods used in existing programs are not guaranteed to find the minimal-length tree, but many of them allow us to try different starting points in the search. If multiple runs from different starting points are done on an actual data-set, the probability of finding among their results the minimal-length tree appears to be high. One should also keep in mind Sneath's question (in Felsenstein, 1975, p. 126): why are we so obsessed with finding an absolutely minimal tree when there may be others of nearly minimal length? Should having one more evolutionary step rule out a phylogeny?

As part of the process of evaluating the length of a tree in parsimony methods, we must often fill in hypothetical states in the interior forks of the phylogeny. For the Camin-Sokal method this is straightforward. For the Wagner criteria it was solved by Kluge and Farris (1969). For a multistate character such as a base in a DNA sequence an algorithm solving it in bifurcating trees was given by Fitch (1971), and for multifurcating trees by Hartigan (1973, 1975) and by Sankoff and Rosseau (1975). Moore, Barnabas, and Goodman (1973) solved the more complex problem of assigning nucleotide sequences to the interior nodes of the tree when the data at the tips are amino-acid sequences. Sankoff, Morel, and Cedergren
The major question that arises in considering parsimony methods is to determine why the criterion is appropriate. Is a total count of numbers of steps the appropriate measure of a phylogeny, and if so, why should we minimize it instead of maximizing or "mediumizing" it? There are two approaches to answering these questions. One considers the process as that of making a statistical estimate of an unknown entity (the phylogeny), and then justifies use of a parsimony method by reference to its properties as a statistical estimation procedure. This was the original approach of Edwards and Cavalli-Sforza (1964), both of whom are former students of R. A. Fisher. They made use of their method of minimum evolution as a rough approximation technique when, for purely technical reasons, they could not get the method of maximum likelihood to work.

Both Farris (1973; Farris in Hull, 1973, pp. 396–398; Farris, 1977a) and I (Felsenstein, 1973b) have appealed to statistical criteria as the fundamental justification for the use of parsimony techniques. The difficulty with this justification is that statistical criteria support the use of parsimony only under rather extreme conditions. Farris (1973) argued that there was a general correspondence between maximum likelihood methods and parsimony methods. I have argued (1973b) that this is only the case when we use maximum likelihood not only to estimate the form of the phylogeny but also the character states of hypothetical ancestors as well. This means that as we add new characters, we bring with them new parameters (their states in the ancestors). The amount of data per parameter then does not rise without limit as we add new characters.

Under those conditions it is known that maximum likelihood methods can fail to have desirable statistical properties such as consistency (convergence with certainty to the true answer if there is a large enough amount of data). I have argued that this might be the case (Felsenstein, 1973b), and I was later able to show (Felsenstein, 1978b) that parsimony and compatibility methods would in fact make an inconsistent estimate of the true phylogeny when evolutionary rates in different lineages were unequal and not small.

If we estimate only the phylogeny (its topology and times of branching, but not the states of characters in hypothetical ancestors), then the parsimony method is a maximum likelihood estimate, but only if we assume that rates of evolution are small (Felsenstein, 1973b). This makes a certain amount of intuitive sense: we are trying to find a phylogeny that requires us to assume as few changes as possible because each one strains our credulity farther. If we had instead expected a moderate amount of change, we would have had no reason to try to find a phylogeny that involved far less change than we had expected.

The difficulty with invoking maximum likelihood as a justification for parsimony methods is that the data often require far too many extra steps to support the implicit assumption that the rate of evolution is small. If little change were expected, then most characters should be invariant across taxa (though we might not have included those), and most of those that showed any change at all would require only a single change of state to explain the evolution of the taxa. This is manifestly not the case in most data-sets I have seen.

Cavender (1978, 1981) has made a serious investigation of the statistical properties of a parsimony method, the two-state Wagner Network method. He showed for a tree with four taxa how one could compute the number of steps by which a tree had to exceed the minimum number to be outside the confidence region containing the best trees. The number is rather large: it is always more than one-quarter the number of characters. For example, for 20 characters, the confidence region includes all trees which are within 9 steps of the best tree. Cavender's results have not yet been extended to five taxa, but it seems likely that this critical number of steps will grow as the number of taxa grows.
The alternative to using a statistical justification for parsimony is to take the word "parsimony" as implying that the number of extra steps required is a direct quantitative measure of economy of hypothesis, and to appeal to the principle of Ockham’s Razor. This argument usually also regards any character that requires an extra step on a given tree as rejecting the hypothesis that the tree is correct. This point has been particularly clearly made by Wiley, who prefers a phylogenetic method in which “the phylogenetic hypothesis which has been rejected the least number of times is preferred over its alternatives” (Wiley, 1975). Eldredge and Cracraft (1980, p. 70) stated that “the criterion of parsimony specifies our acceptance of the least rejected hypothesis,” and that “we prefer the cladistic hypothesis which minimizes convergence.”

Both in the book by Eldredge and Cracraft (1980) and in the discussion by Beatty and Fink (1979), it is made clear that each extra change of character state is regarded as requiring one extra hypothesis to account for the observed data. As we shall discuss below (when compatibility methods are considered), it is less than obvious that one can equate one extra hypothesis with each extra change of state. Even having accepted the equation, it is necessary to understand clearly that the advocates of this position do not take literally the notion of rejecting an hypothesis. Otherwise it would be difficult to understand why they allow themselves to accept an hypothesis which has been rejected even once.

The difference between the two positions can be illustrated by the following analogy. Suppose that we toss a coin five times and obtain three heads. We wish to know whether the coin is biased towards heads or towards tails. We may safely rule out the possibility that its probability of heads is precisely one-half. A statistician using maximum likelihood (or almost any other statistical technique) obtains an estimate of $p = 3/5$ for the probability of heads. With respect to the compound hypotheses that the coin is biased towards heads ($p > 0.5$), or towards tails ($p < 0.5$), the former has the higher likelihood. The statistical properties of this estimate (e.g., consistency, efficiency) are readily discovered.

The parsimony approach to the same case seems to involve arguing that if the coin is actually biased towards tails, then we must assume that the three heads are misleading (though not impossible) outcomes, and if it is biased toward heads, the two tails must be the misleading outcomes. It is therefore argued that the latter is preferable, having been “rejected” the fewer times. Note that the count of the number of rejections functions here in effect as a measure of goodness-of-fit. Alternatively, if the number of hypotheses is being counted rather than the number of rejections, a bias towards heads requires two extra hypotheses (that the two tails are misleading), whereas a bias towards tails requires three.

An alternative way of justifying parsimony is on empirical grounds. This was the approach of Camin and Sokal (1965), and a similar argument has been made by Farris (as reported by Mitter, 1980). Although biologists are most comfortable with empirical justifications, the difficulty with them in the present case is that there is usually no independent means of knowing whether the method being tested has obtained the right result. Even when this is possible, as with artificial data produced by computer simulation, one requires assurance that the processes that produced the data are a reasonable model of the processes acting in nature, since a given method may do well on data produced by one process and poorly on data produced by another.

Panchen (1982) and Friday (1982) have been critical of the concept of parsimony. The paper by Hull (1979) and the stimulating debate between Harper and Platnick (1978) will be of particular interest to a reader who is interested in the controversies over which hypotheses are testable and the extent to which they are testable. The equation of extra steps with extra hypotheses has yet to receive the careful consideration it deserves.

COMPATIBILITY

The second major approach to resolving Hennig’s dilemma is to prefer that phylogeny which has the largest number of characters perfectly compatible with it. The first appearance of a form of compatibility in a
numerical context seems to be in Camin and Sokal's 1965 paper. As a step in their algorithm for finding the most parsimonious tree, they prepared a table showing for each pair of characters how many steps each would need to evolve on a tree suggested by the other character. This is similar, but not identical, to the more usual notion of compatibility, one which calls two characters compatible if there exists some phylogeny on which both of those characters could evolve without any state having to arise more than once. Wilson (1965) proposed using compatibility to test whether a set of characters could have arisen on the same phylogeny in unique and unreversed fashion, and he specified how the phylogeny could be found, once it has been determined that all characters were compatible. Since to find the phylogeny he required complete compatibility and since he assumed that ancestral states were known, his proposal may be regarded as an analogue of Hennig's method.

The critical step in developing compatibility methods was taken by Le Quesne (1969), who proposed using the largest set of mutually compatible characters to infer the phylogeny. Le Quesne did not give an explicit algorithm for finding the largest clique of mutually compatible characters. He envisioned finding it by visual inspection of a table that showed which pairs of characters were compatible. He made no assumption that it was known which state in each character was ancestral. This meant that if each character had two states, 0 and 1, two characters were compatible with each other if three or fewer of the four possible combinations of states 00, 01, 10, and 11 occurred in the data. When both states have known ancestral states (say state 0), then two characters are compatible if two or fewer of the three combinations 01, 10, and 11 appear in the data.

The compatibility method was put on a firm logical foundation by George Estabrook and F. R. McMorris in a series of papers that provided formal mathematical definitions and proofs. The first two of these (Estabrook, 1972; Estabrook, Johnson, and McMorris, 1975) provided the formal definitions. The major theorem that is needed we may call the Pairwise Compatibility Theorem: that a collection of characters are all pairwise compatible if and only if they are mutually compatible, which means that there is a phylogeny on which all of them could evolve with each state arising no more than once.

For characters with two states, one of which is known to be ancestral, the theorem was proved (Estabrook, Johnson, and McMorris, 1976a; McMorris, 1975). When a character has several states, one of which is known to be the ancestral state, a method of recoding the states was developed by Kluge and Farris (1969) from the additive coding method of Sokal and Sneath (1963). This recoding method allows the data to be represented as a series of two-state characters. Estabrook, Johnson, and McMorris (1976b) proved that if this system of recoding is used, and if all the resulting pairs of two-state characters are compatible, then the two characters are compatible, and this finding allows us to use the Pairwise Compatibility Theorem for multistate characters. Estabrook and McMorris have more recently (1980) given a considerably simpler method of deriving these results.

When we have two-state characters but do not know the ancestral states, things are still fairly simple. McMorris (1977) proved the Pairwise Compatibility Theorem for this case, and thus justified Le Quesne's intuition that a table showing pairwise compatibilities is sufficient. McMorris also showed that two characters found to be compatible would remain compatible if in each character the commoner of the two states was taken to be ancestral.

In the case of multistate characters without the ancestral states known, a solution is more difficult. When any state may arise from any other, as with bases in DNA sequences, Estabrook and Landrum (1975) and Fitch (1975) gave a method of testing the compatibility of two characters, and Estabrook and McMorris (1977) proved it correct. Fitch (1975) however, had shown by counterexample that the Pairwise Compatibility Theorem was not true for such characters: it is possible for a collection of pairwise compatible characters not to have any phylogeny with which they are all compatible. Estabrook and Meacham (1979) gave a
method for deciding whether two characters that have their states connected in a treelike network, but whose ancestral states are unknown, are compatible. It follows from their method that the Pairwise Compatibility Theorem is true for such "undirected character state trees."

The application of Le Quesne's compatibility method to the data of Table 1 results in the tree shown in Fig. 2A. The first four characters form the largest clique. Its chief rival is a clique of three characters, numbers 4, 5, and 6.

One of the persistent difficulties that arises in the practical use of the compatibility method is that the largest clique may contain too few characters, so that the partial information contained in most characters is not used in defining the tree topology (Farris, 1969; Hill, 1975). Estabrook, Strauch, and Fiala (1977) have extended compatibility analysis to alleviate this difficulty. They use the largest clique to get an initial topology of the tree, and then divide the taxa into two groups in a way consistent with that topology. Within each of these groups the analysis is repeated. A character that is not part of the largest clique may then be readmitted to active consideration once such a division is made, because it may be incompatible with other characters within one subgroup but not within the other. By repeating the process and subdividing the group ever more finely, information from most characters is ultimately used. This process is reminiscent of the procedure advocated by Throckmorton (1968), who used clustering algorithms rather than compatibility considerations. It seems likely that the intuitive methods used by traditional taxonomists are close to this approach.

Another extension of compatibility methods involves an attempt to compute probabilities of compatibility and use these as guides to the evaluation of characters. Le Quesne (1972) gave an expression for the probability that two two-state characters will be incompatible if each is distributed at random among the taxa. Sneath, Sackin, and Ambler (1975) extended this calculation by computing the probability that k or fewer taxa would have to be reinterpreted to achieve compatibility between two characters. Meacham (1981) has extended the computation much farther, by computing probabilities of compatibility of multistate characters and producing a computer program to compute the probability of compatibility of a set of characters. Estabrook (1980) has pointed out the possibility of basing one's inference not on the largest clique, but on the most surprising clique, using Meacham's probability as the measure of how surprising it would be that the clique would form at random.

It should be kept in mind that the null hypothesis in these probabilistic calculations is that one of the two characters is distributed at random across taxa. This is the same as assuming that the characters have evolved at random on a tree, with so many reversals or multiple origins of character states that the character is in effect informationless. It is not the same as assuming that the character evolved on a tree of different form, with some reversals or multiple origins of character states. Thus, the probabilities do not allow testing of one tree against another, but they do appear to give a useful guide as to when characters could plausibly be totally informationless.

Sneath, Sackin, and Ambler (1975) have made another extension of compatibility methods by using them to divide the characters into two sets, within each of which there is much compatibility but between which there is incompatibility. This procedure is done with the objective of detecting hybridization events that would create a loop in the phylogeny and would result in different characters showing different and incompatible evolutionary patterns.

While our discussion has been entirely in terms of replacing the biological intuition of the taxonomist by a mechanistic method of choosing a tree (say by basing it on the largest clique), many advocates of the use of compatibility see the notion as advisory. In practice they will not always use the largest clique, but will employ external criteria (e.g., Estabrook and Anderson, 1978), or will limit themselves to using those characters which occur in several of the largest cliques (Strauch, 1978). A similar comment applies to many users of parsimony methods. At the risk of being unfair, we will con-
fined our attention to the more mechanical approach.

**The Justification for Compatibility**

When is it justifiable to use compatibility methods? As Sankoff (in Le Quesne, 1975, p. 426) has pointed out, compatibility and parsimony methods are very close in spirit. While the one criterion counts the total number of changes, the other counts the number of characters that do not perfectly fit the tree. The result in both cases is a quantity to be minimized. This close relationship is underscored when we look carefully at the justification of parsimony by arguments involving Ockham's Razor. When Wiley (1975) speaks of preferring “the phylogenetic hypothesis which has been rejected the least number of times,” the question immediately arises how we are to count the number of rejections.

If each extra change of character state is a rejection, so that a character with two extra changes of character state is considered to reject the hypothesis twice, then Ockham's Razor will be taken to support use of a parsimony method. If, on the other hand, a character with two extra changes of state is taken to reject the hypothesis only once, so that we count incompatible characters rather than states, then Wiley’s argument supports use of a compatibility method. Similar considerations apply if we count extra hypotheses rather than rejections: are we to consider each extra change to require one extra hypothesis, or to consider each character that conflicts with the others to require one hypothesis? This very ambiguity calls into question this approach to justifying either method: what principle are we to use to decide which is the correct way to count hypotheses? Are we not simply displacing all the difficult choices, rather than finding a way to make them?

The alternative approach is to consider the methods as statistical inference methods. Compatibility methods will be justified when they can be shown to have good statistical properties, and in particular when they are determined to be identical to a well-supported statistical method such as maximum likelihood.

In two recent papers (Felsenstein, 1979, 1981b), I have asked when compatibility methods are also maximum likelihood methods. It turns out that this will be the case if most characters change very slowly, so that they can be explained by assuming only a single origin of each derived state, but a few characters show multiple origins of each derived state. These few characters may be explained by presuming that they have been grossly misinterpreted by the systematist, so that they contain no usable taxonomic information and will show many origins of each state if an attempt is made to place the changes of state on a phylogeny (Felsenstein, 1979). Alternatively, these few characters may have very high rates of evolution, a condition that leads to the same result—namely, that they contain little usable information (Felsenstein, 1981b).

Thus both parsimony and compatibility can be justified as maximum likelihood methods, but under somewhat different circumstances. Both methods require that most characters have a low rate of change. Both in effect assume that homoplasy will be rare. If we expect homoplasies to be scattered at random over all characters, then a parsimony method is supported. If they are expected to be concentrated in a few characters, whose identities are not known in advance, then compatibility is supported.

The weakness of this justification for each of these methods is that assumptions are required that conflict with what is seen in many data-sets. In the first case, we expect to see data that require only a few extra changes of state. In the second, we expect to see data on which most characters are in the largest clique, and only a few need to be discarded. These expectations will remain true even if we assume that invariant characters are not included in the data-set.

Most data require too many changes of state and have too many incompatible characters to allow us to believe that we are in a situation in which parsimony or compatibility are maximum likelihood methods. As we shall see below, they may still have reasonable statistical properties, but it is no longer inevitable that they are the methods of choice.

We have already seen that parsimony methods can fail to have the property of con-
sistency if rates of change are not small. In fact, the same difficulty arises with compatibility methods (Felsenstein, 1978b), so that this defect does not give us a basis for preferring one method over the other, but is rather a reason for skepticism of both.

**CLUSTERING METHODS**

The first numerical methods used to find phylogenies were clustering methods based on pairwise similarity between taxa. Michener and Sokal (1957) intended to infer phylogeny by using an average-linkage clustering method. Any hierarchical clustering method can be used to obtain a phylogeny, provided that one interprets all the groups obtained as being monophyletic. It was immediately realized that this procedure would give a misleading result if rates of evolution varied among lineages. Michener and Sokal (1957) discussed the problem of deciding whether dissimilarities were the result of “prergroup” or “exgroup” relationships, the former correctly reflecting phylogeny and the latter being a consequence of a high rate of evolution in one lineage. They gave some corrections to their techniques, corrections intended to detect exgroup relationships and to correct the phylogeny.

The problem has also been discussed by Kirsch (1969) and particularly clearly by Farris (1971). The essence of the problem is as follows. Suppose that we were considering four species, a bird (B), a mammal (M), a crocodile (C), and a snake (S). Let us accept for the sake of the argument that the snake and the crocodile have changed less from the ancestor of all reptiles than have the bird and the mammal. The snake and crocodile will then be the two most similar of these species. A phenetic clustering algorithm applied to a measure of overall similarity, one that counts both shared ancestral and shared derived states, will place S and C together and will suggest that they constitute a monophyletic group. Even if we ignore the position of the root in the resulting topology, the wrong unrooted topology is obtained.

The conditions under which this error is likely to happen have been derived by Colless (1970) and in a different form by Moore (1971). Colless stated conditions on the amount of evolution in different lineages in a four-taxon example such that a phenetic clustering algorithm will arrive at the wrong topology. He assumed that the amounts of evolution add up to the dissimilarity between taxa. Colless’s conditions can be relaxed somewhat to get the conditions for obtaining the correct unrooted topology, when we consider \(((S, M), (B, C))\) to be the same topology as \((M, (S, (B, C)))\).

Colless found that there is no single clustering method that is best for recovering the phylogeny: the best clustering method depends on the exact pattern of convergences (homoplasies) expected. Moore’s conditions were stated in terms of the pattern of similarities to be expected, and led him to a preference for the average-linkage (UPGMA) clustering method. The hints of discrepancy between these two studies make a reconsideration and extension of them desirable. In any event, the two studies are in agreement, in that they both find that clustering methods are vulnerable to differences in rates of evolution in different lineages.

Because these difficulties have been known from the outset, there has not been much development of phenetic clustering methods for specifically phylogenetic purposes. Goodman and Moore (1971) used a divisive clustering method on immunological distances. Gibbs, Dale, Kinns, and MacKenzie (1971) used clustering on tables of pairs of adjacent amino acids derived from protein-sequence data. Their method would seen to involve some loss of information, though it may also be unusually robust in the presence of insertions and deletions.

Phenetetic clustering methods have been used extensively on genetic distances to generate trees from frequencies of electrophoretic alleles. The most widely used genetic distance measure, Nei’s distance (Nei, 1971), does not necessarily satisfy the triangle inequality. It will do so in expectation, if we assume that evolution proceeds according to an infinite-isoalleles neutral-mutation model and that there are many loci in the study. But for finite numbers of loci it will vary around this expectation and will occasionally fail to satisfy the triangle inequality. Much has been made of the occasional failure of Nei’s distance to be metric (cf. Farris, 1981), but it is not obvious why this
should be considered a fatal objection to it on any grounds other than esthetic ones.

It might seem that the use of phenetic clustering methods on electrophoretic data requires faith in the exact constancy of the molecular clock. It would seem from Colless's and Moore's studies that exact constancy is not required, only sufficient regularity to enable phenetic methods to work. Theoretical studies are clearly called for to determine how much constancy is necessary.

PAIRWISE METHODS

Closely related to clustering methods are the pairwise methods. These attempt to fit a tree to a matrix of pairwise differences or pairwise similarities between taxa. All pairwise methods attempt to solve essentially the same problem, in slightly different ways. It is assumed that the data have been reduced to a matrix of pairwise differences between taxa, and that we want to find a tree (usually an unrooted tree) that comes as close as possible to predicting these observed distances. The methods differ mostly in the way they measure the overall difference between observed and expected distances. Some general comments on pairwise methods will be found in the paper by Fitch (1982).

Some types of data come in the form of distances, such as immunological data and DNA hybridization data. Most others originate as coded sequences of characters, the distances being calculated from the original data. A common assumption of all methods is that this reduction to distances (or similarities) does not lose any significant amount of information. Some information is inevitably lost: the original character states cannot be reconstructed from the distances. For example, the following two data-sets have the same sets of distances between pairs of taxa:

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 1 1 1</td>
<td>1 1 1 0</td>
</tr>
<tr>
<td>B</td>
<td>0 0 1 1</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>C</td>
<td>0 1 0 1</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>D</td>
<td>0 1 1 0</td>
<td>0 0 1 0</td>
</tr>
</tbody>
</table>

While pairwise methods thus clearly lose some information, they preserve many of the important aspects of the data. The two data-sets above are alike in showing all pairs of taxa to be equidistant, and the distances reflect this.

A critical common assumption of all pairwise methods is that expected distances between taxa are sums of lengths of segments along the tree. In effect, what is being assumed is that expected phenotypic distance between taxa is proportional to time. This will nearly always be invalid unless great care is taken in selecting the measure of distance. For example, if there are two states, 0 and 1, in each character, and changes in either direction are equally frequent, then when two taxa are infinitely far apart on the tree they will be expected to differ in half their character states. If the hypothesis of additivity of lengths were correct, they would have to differ in more characters than they possess! Read (1975) has expressed skepticism that additivity is the appropriate assumption for immunological distances, and has suggested an alternative nonadditive relationship. Wilson, Carlson, and White (1977) have argued that additivity is an appropriate assumption for immunological distances.

If the additivity hypothesis can be accepted, however, then most pairwise methods have the desirable statistical property of consistency. As we accumulate more and more data, the observed distances between taxa will approach the expected distances. These latter will be perfectly fitted by the correct tree with the correct lengths, and it is a simple matter to show from this correspondence that we will be more and more certain of getting the correct result as we accumulate data.

The first pairwise method was that of Fitch and Margoliash (1967). They proposed finding the tree that minimized the sum over all observed distances of \((d - d')^2/d^2\), where \(d\) is the observed distance and \(d'\) the expected distance under the additive hypothesis. Goodman, Barnabas, Matsuda, and Moore (1971) gave some results obtained using the Fitch-Margoliash method. Beyer, Stein, Smith, and Ulam (1974) criticized those results for the failure of the expected distances to satisfy the triangle inequality.
This occurred because Fitch and Margoliash allowed negative segment lengths in their solutions. It is easy to restrict their method to avoid negative segments.

Cavalli-Sforza and Edwards (1967) suggested minimizing the sum of \((d - d')^2\), a procedure which had the merit that matrix algebra then enabled them to assign segment lengths to a given tree by solving a set of simultaneous linear equations. Kidd and Sgaramella-Zonta (1971) discussed this method further. Hartigan (1967) also suggested a simple least-squares approach, but confined the search to rooted trees whose tips were all an equal distance from the root.

Farris, Kluge, and Eckardt (1970) found a way of altering conventional clustering methods to cluster taxa by the number of shared derived states. If the phenotype of the ancestor is presumed to be known, a formula presented by them corrects the observed similarities between taxa so that they reflect only similarities in derived states. These authors then advocated that these similarities be used in a conventional clustering algorithm, so as to obtain a result approximating the result of the Camin-Sokal parsimony method. Klotz and co-workers (Klotz, Komar, Blanken, and Mitchell, 1979; Klotz and Blanken, 1981), and Li (1981) have discovered the same formula for correcting distances in order to eliminate the effect of shared derived states.

Farris (1972) presented a method, the Distance Wagner method, which has been widely used. Despite its name, it is not related to the Wagner parsimony method. It fits expected to observed distances, though without an explicit measure of goodness of fit. Farris restricts attention to trees in which \(d' \geq d\). In the absence of a well-defined criterion which is being maximized or minimized, the name Distance Wagner can be applied only to the particular algorithm given in Farris’s paper. Swofford (1981) has presented some variations on the Distance Wagner Method.

Beyer, Stein, Smith, and Ulam (1974) and Waterman, Smith, Singh, and Beyer (1977) preferred to minimize either the total length of the tree, the sum over all distances of \((d' - d)/d\), or the sum of \((d' - d)/d^2\). For these three criteria, they noted that the problem of assigning segment lengths for a given topology was a problem in linear programming that could be solved by existing numerical methods (although this still would not solve the problem of searching among topologies). This point has also been discussed by Wagner (1981). Fitch and Smith (1982) have given necessary conditions for the achievement of minimal length that facilitate finding the minimum length tree.

Another class of pairwise methods is based on the Four-Point Metric (FPM). This is a condition that will hold for every set of four taxa if the data fit the tree perfectly. It was first presented by Buneman (1971), who proposed (1) requiring that \(d' < d\) and (2) minimizing a quantity that measures departure from the FPM. Dobson (1974) gave conditions sufficient for the FPM to hold, and showed that if it is satisfied the taxa lie at the vertices of an unrooted tree. Waterman, Smith, Singh, and Beyer (1977) gave an algorithm for finding the tree from a set of data that happen to satisfy exactly the FPM condition. Sattath and Tversky (1977) proposed clustering taxa on the basis of the number of times the FPM is satisfied, but using the least squares criterion to set segment lengths. Fitch (1981a) has advocated a similar approach. Moore, Goodman, and Barnabas (1973) proposed using a quantity, the Moore Residual Coefficient, which is quite complicated but seems to be related to the FPM and will be zero if the data fit the tree perfectly.

A particularly interesting pairwise method is that of Chakraborty (1977), who used a simple model of amino acid replacements in a protein to compute the correlations between the pairwise distances between the resulting protein sequences. He then used a least squares criterion that takes these correlations into account. Although his method was presented as a means of estimating segment lengths within a given topology, one could as easily use his criterion to choose among topologies (R. Chakraborty, pers. commun.).

In many of the papers mentioned above, there is discussion of strategies for rearranging the tree in order to find the topology that is best by the criterion proposed. In no case
is there any more guarantee that one would be able to find the best topology using pairwise methods than when using other methods.

**Statistical Properties**

Faced with a profusion of suggested criteria of fit, most of which should result in a consistent estimate, we can appeal to statistical methods to indicate which of these methods will use the data most efficiently. The difficulty in using statistical methods is that it is rare for there to be a well-stated probabilistic model available that describes the expected distribution of the set of pairwise distances, given the tree. An exception is Hartigan's (1967) model, which assumes that each similarity value is independently drawn from a normal distribution with equal variance. This assumption cannot be strictly true unless negative similarity values or values greater than one are to be allowed, but it may be a good approximation.

We can make a model like Hartigan's, under which the pairwise distances are all independently drawn from normal distributions. Each distance $d$ is drawn from a normal distribution, with expectation $d'$ and some variance. If the variances are all the same, then the maximum likelihood estimate is Cavalli-Sforza and Edwards' (1967) least squares method. If the variances are functions of $d'$, so that small distances vary around their expectation less than long ones, we should expect weighted least-squares methods like the Fitch-Margoliash method to do well, though none of the existing methods is precisely the maximum likelihood method.

The difficulty with this variant of Hartigan's model is that it does not take into account that the distances may be correlated. There are two possible sources of correlation. With immunological distances or with DNA hybridization data, variations in the concentration of standards can affect in the same direction all the distances in a given row or column of the table. For sequence data or for immunological distances that reflect the amino acid sequence of a single purified protein, correlations will also arise as a result of random evolutionary events in interior branches of the tree, which thus affect several different distances at the same time. Chakraborty's method is a pioneering attempt to correct for this latter source of correlation (although with a somewhat oversimplified model of evolution). No one has yet attempted to correct for the former source of correlation.

Farris (1981) has recently criticized most distance methods, on the grounds that they infer impossible values for the branch lengths on the phylogeny. I have argued elsewhere (Felsenstein, in press) that Farris's critique is dependent on the branch lengths being achieved rather than expected amounts of change. If the expected distance interpretation is adopted, Farris's criticisms are not applicable.

Pairwise methods have an attractive simplicity, and may well be robust to variations in the model of evolution, but it is clear that they need much statistical development before we can feel assurance that they are making efficient use of the data. The dubious status of the additivity assumption is probably a more serious problem than the loss of information by reduction of the data to pairwise distances.

**EXPLICITLY STATISTICAL METHODS**

While we have seen that we can ask under what biological conditions existing methods can be justified as statistical methods, an attractive alternative is simply to try to implement statistical methods, such as the maximum-likelihood method, directly. This is what Edwards and Cavalli-Sforza (1964) were trying to do. Using gene frequencies of blood groups in various human populations, they wanted to infer a branching phylogeny. To implement a maximum likelihood method, it is necessary to have a probabilistic model that allows computation of the probability of the data given the hypothesis. The evolutionary force invoked by them was random genetic drift, which is quite plausible as the main force differentiating human blood-group gene frequencies.

They were faced with the problem that transition probabilities from one gene frequency to another are almost impossible to calculate. To simplify matters, they approximated the process by using the arc-sine square-root transformation and by assuming that the resulting variables were performing
a Brownian motion on an infinite scale. This has been the fundamental model underlying all work done since then on continuous-character evolutionary trees. When they tried to make maximum-likelihood estimation work, they ran into a technical problem. They were attempting to estimate not merely the phylogeny, but the phenotypes at the forks in the tree as well. This procedure led to singularities in the likelihood surface, and forced them to fall back on a parsimony method as a less desirable alternative.


By following some of Gomberg's suggestions, I was able to eliminate the estimation of ancestral phenotypes and so make maximum-likelihood estimation work, and that produced a slow but usable computer program (Felsenstein, 1968, 1973a). The likelihood method that worked best turned out to be a restricted maximum-likelihood (REML) method in which the absolute coordinates of the populations are ignored and only the differences between populations are used.

Thompson (1975), in a comprehensive monograph on the problem, developed a strict maximum-likelihood (ML) method. Thompson was able to find an iterative algorithm for maximizing this likelihood, and produced a workable computer program. The major limitation of this method is that it may not make a consistent estimate of the phylogeny. With each character that is added to the data, a new parameter must be added to the estimation problem. It is in such cases of “infinitely many parameters” that maximum likelihood often fails to have the property of consistency. More recently, I have developed an iteration method similar to Thompson's, one which greatly speeds up computations and results in a practical computer program to implement REML estimation of the phylogeny (Felsenstein, 1981c).

Cavalli-Sforza and Piazza (1975) have made a pioneering attempt to test the goodness of fit of a tree to a continuous-character or gene-frequency data-set (see also Astolfi, Piazza, and Kidd, 1978). For purely technical reasons (Cavalli-Sforza and Piazza, 1975, p. 145–146), their algorithms are not entirely correct, but the question they have raised is critically important in applying these methods to within-species gene-frequency data, where gene flow is an important complication. More complete methods extending their approach will be of great importance in the future, particularly at the intraspecies level. I have recently (1982) discussed unsolved problems that arise in microsystematic analysis within species.

**Nucleotide and Protein Sequence Data**

Application of likelihood methods to DNA or protein sequence data has been prevented largely by the difficulty of the computations. Neyman (1971) called the attention of statisticians to the problem of inferring phylogenies from DNA sequences. He presented a simple probabilistic model of evolution and showed how to compute likelihoods for three species. Kashyap and Subas (1974) proposed using the maximum-likelihood solutions for three species at a time in order to analyze data involving more than three species.

I have shown how to economize on the calculation of likelihoods with DNA sequence data, and have developed an iteration method of finding the maximum-likelihood phylogeny (Felsenstein, 1981a). The computations are still expensive, and one must hope that further progress can be made to reduce this problem, particularly in view of the large amount of DNA-sequence data that will soon be available.

Holmquist (1972a, b, c) has developed a probabilistic model of change in DNA sequences and protein sequences, although he did not present any statistical methods based on this model. The difficulty with a likelihood analysis of protein-sequence data is that the number of possibilities is much greater (64 codons and 20 amino acids as opposed to 4 bases), so that those likelihood
methods which are marginally applicable to DNA sequences become impractical for proteins. Even though one can easily see how the computations ought to be done, a large breakthrough is needed before they can be applied, in practice.

Using protein-sequence data, Langley and Fitch (1974) tested the hypothesis of constant rates of replacement of amino acids, and concluded that the rates are not constant. They used a parsimony method to reconstruct ancestral sequences; then, using these ancestral sequences as if they were observed data and making corrections for multiple changes in a single codon, they used a chi-square test of the numbers of changes in each segment of the tree. This is at best an approximate approach: a maximum-likelihood procedure could be used to construct a likelihood ratio test of the same hypotheses, if it were known how to do the computations in a reasonably finite amount of time.

Kaplan and Langley (1979) have developed a likelihood-estimation procedure for estimating times of divergence from restriction-enzyme fragment data. Their method is applied to only two species at a time. There is much data of this sort, and methods of analyzing it need to be developed to deal with multiple-species phylogenies. Restriction-site data have many of the same problems that protein-sequence data present: although there are only two observable states (site present or site absent), these reflect a much larger number of states at the DNA level (2,048 for a six-base recognition site).

Kaplan and Risko (1981) have extended the approach of Kaplan and Langley, and have compared it with the less statistical approach of Nei and Li (1979). They found a substantial agreement between the estimates of divergence time from the two methods, and they discussed the application of their method to multi-species data. Kaplan and Risko (1982) have stated a pairwise method for inferring phylogenies from nucleotide sequences, based on an approximate maximum likelihood divergence time for each pair of species. In this respect their method resembles Kashyap and Subas's (1974) approach.

**Gene Duplication Data**

One of the most successful applications of likelihood methods was that of Ferris, Portny, and Whitt (1979), who used a simple probabilistic model of the loss of expression of duplicated genes to analyze data on the catostomid fishes. They were able to evaluate the likelihood of the data under various hypotheses and to carry out likelihood ratio tests of hypotheses of constancy of rate of loss.

**CONCLUSION**

The development of numerical methods for inferring evolutionary trees has been plagued by dogmatic adherence to arbitrary criteria, which are often defended on primarily esthetic grounds. Some workers (e. g., Fitch, 1982) have adopted a more eclectic approach, although they have usually relied on empirical tests of methods. Such a strategy is difficult to apply in the absence of external information as to the true phylogeny.

We have seen that each category of methods lends itself to evaluation and analysis by applying statistical criteria, which serve as a common standard and allow the comparison of different methods. By focusing attention on the probabilistic model of evolution underlying each method, a statistical approach also serves to focus attention on the biological assumptions of each method, rather than on the esthetic or geometric properties that have played such a prominent role in justifications of numerical phylogenetic methods.

There are obviously major difficulties with a statistical approach: we have sufficient uncertainty about our models of evolution that the precision of a confidence interval may be misleading. There is much room for development of robust methods that may sacrifice some precision but apply to a larger class of models. Unfortunately, there has been virtually no work yet along these lines. It may even be the case that some existing methods are robust, but the proper proofs of this have not yet been provided.

In focusing our attention on the biological assumptions underlying each method, a sta-
statistical approach will promote a realization that there can be no such thing as an all-purpose method for inferring phylogenies. Since they are couched in a vocabulary that explicitly recognizes the existence of uncertainty about the result, statistical criteria should discourage dogmatism.

At the present time, non-statistical approaches are couched in a vocabulary that employs terms such as “rejection,” “refutation,” and “deduction.” This seems to reflect a perceived need in systematics for absolute certainty, a need that has been felt the more strongly, the less firm were the logical foundations of the procedures in common use.

The adoption of a methodology that explicitly acknowledges uncertainty is a paradoxical necessity if phylogenetic inference is to be placed on a firm scientific footing.

ACKNOWLEDGMENTS

I am grateful to Dr. Hezy Shoshani for pointing out to me the paper by Le Gros Clark and Sonntag (1926). I wish to thank Walter Fitch, George Estabrook, Christopher Meacham, and David Swofford for extremely helpful comments on an earlier draft of this paper. This work was supported in part by task agreement DE-AT06-76EV71005 of contract DE-AM06-76RL02225 between the U. S. Department of Energy and the University of Washington.

LIST OF LITERATURE


INFERRING EVOLUTIONARY TREES


GOODMAN, M., and G. W. MOORE. 1971. Immu-


KLOTZ, L. C., N. KOMAR, R. L. BLANKEN, and


