Annu. Rev. Genet. 1988. 22:521–65

PHYLOGENIES FROM MOLECULAR SEQUENCES: INFERENCE AND RELIABILITY

Joseph Felsenstein

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

CONTENTS

INTRODUCTION

The field of molecular evolution owes most of its existence to the possibility of sequencing proteins and nucleic acids. Molecular sequences provide us with precisely comparable characters, observed at or near the level of the

0066-4197/88/1215-0521$02.00

gene, which can be examined in diverse organisms. The amount of data is very large and rising rapidly. It enables us to work in two modes: we can either use our knowledge of the evolutionary history of the species to examine the mechanisms of evolution of the molecules, or we can use knowledge of the evolution of the molecules to infer the evolutionary history of the species. It is this latter, the inference of phylogenies, that is the concern of this review. However, the techniques used to do this are also relevant to the other task.

In either mode, we make use of a model of the evolutionary process. The central model of molecular evolution is one of random evolutionary changes, occurring at a stochastically constant rate. It was first introduced by Zuckerkandl & Pauling (144) in the form of the "molecular clock," which is the somewhat stronger assertion that the expected rate of change was the same in all lineages. Analysis of molecular data can often proceed without that strong an assumption.

Kimura (81) provided a population-genetic rationale for a molecular clock by propounding the neutral mutation theory of molecular evolution (see also 82). This provided a unified theory accounting for both genetic polymorphism at the molecular level and change of the molecules through time. The theory does not rule out natural selection against deleterious mutants, and it argues that most differences in the rate of evolution between different molecules and different parts of the genome are accounted for by conservation of biologically significant sequences.

Theories explaining evolutionary change and polymorphism by natural selection have been less well developed, partly because there are so many different possible kinds of selection that it is difficult to choose between them. Gillespie (59, 60, 61, 63) has argued that randomly varying selection coefficients, rather than neutral mutations, account for most polymorphism and molecular evolution.

The controversies between neutralists and selectionists have continued for 20 years with no clear resolution, primarily due to the low resolving power of the data—natural selection many orders of magnitude weaker than we can detect in the laboratory can be effective in nature. From the point of view of this review, it does not matter whether nucleotide substitutions are neutral or selective. Our very inability to resolve the controversy over neutrality is an advantage when it comes to estimating phylogenies, since we can use the neutral mutation theory as if it were true, confident that for the data we can collect, other theories would make indistinguishably different predictions.

ESTIMATING PHYLOGENIES

Numerical methods for inferring phylogenies from molecular data have existed for over 20 years, but there is still much confusion in the literature about

their assumptions and properties. For example, there is little coverage of them in textbooks of evolution or of molecular biology; that which exists is usually a brief and mechanical exposition of a particular method familiar to the author. As a result, the inference of phylogenies often seems divorced from any connection to other methods of analysis of scientific data.

Nor are most journal articles much help: molecular evolutionists who use methods for inferring phylogenies do not engage in much discussion of the properties of the methods they use since they focus on the difficult task of collecting the data. It is not unusual to see papers presenting phylogenies with little more than the most perfunctory description of how they were obtained. This lack of detail would not be tolerated in presentation of the biochemical methods in the same papers: editors take no comparable care to see that the phylogenetic methods are carefully described.

The most effective way of thinking about the inference of phylogenies is to adopt a statistical point of view, as with other kinds of data analysis. It is then seen simply as making an estimate of an unknown quantity, in the presence of uncertainty, and using a probabilistic model of the evolutionary process. Viewing the process in this way immediately emphasizes a limitation of most current discussion of methods for inferring phylogenies. They make a single estimate—a point estimate—but are not designed to tell us what other phylogenies might also be acceptable. This is partly because of the difficulty of doing so and partly because some exponents do not believe that a statistical framework is appropriate.

The importance of making some assessment of the statistical variability of the estimates of phylogenies is underscored by two recent studies. Miyamoto et al (95) studied 7.1 kB of DNA sequence from the $\psi\eta$-globin region in apes and human and found that the most parsimonious tree had chimpanzees and humans as most closely related. However, this conclusion could be based on only 13 positions at which there were "phylogenetically informative" patterns of nucleotide substitution or deletion/insertion events. Of those, eight backed a human-chimpanzee relationship, three a chimpanzee-gorilla relationship, and two a human-gorilla relationship. They concluded that their data "provide strong evidence . . . that human and chimpanzee are more closely related to each other than either is to gorilla." This conclusion is mandated if one adheres to the school of "phylogenetic systematics," or "cladism," which focuses on the most parsimonious tree to the exclusion of any statistical interpretation. An accompanying news article (90, p. 273) quotes Goodman as saying "if we had only our dataset, the question of a human-chimpanzee association wouldn't be decisive, and maybe putting all the datasets together still would leave some room for doubt." There is a discrepancy in the firmness of their conclusion in these two statements. Perhaps this is inevitable if one excludes statistical analysis as irrelevant but still has the good biological sense to regard the conclusions as uncertain.

Field et al (51) analyzed 22 animals for sequences of 18S rRNA, using a distance matrix method with distances derived from the sequences. They estimated the phylogeny of the metazoa, coming to suggestive and controversial conclusions (for example, that coelenterates are derived from protists independently of other metazoans). However, many of these conclusions are based on short internal branches of the tree, whose reality can only be judged if we have some measure of the variability of length of these branches. Field et al (51) are concerned about this, saying that "there are no simple measures of reliability for the position of given branch points" but arguing that their conclusions are reliably indicated by reproducibility of the branching order as different sets of species are used to make the tree. Assessment of the reliability of the results is thus central to any appreciation of the meaning of this study.

This review therefore focuses on the methods for assessing the reliability of phylogenies from molecular sequences, after describing briefly the three major families of methods for inferring phylogenies.

METHODS FOR INFERRING PHYLOGENIES

The three major families of methods for inferring phylogenies are the parsimony and compatibility methods, the distance methods, and maximum likelihood methods. Most other methods fit under one of these headings.

Parsimony and Compatibility Methods

PARSIMONY If each site in a set of sequences has changed only once in the evolution of a group, then the newly-arisen base will be shared by all species descended from the lineage in which the change occurred. If this were the case at all sites, then the sets of species having the new bases would be either perfectly nested or disjoint, never overlapping unless one set of species was included in the other. It would be possible to erect a tree on which we could explain the evolution of the group with only a single change at each site. This can be done by inspection of the sets of species defined at each varying site. If some of these sets of species overlap without being nested, then there is conflict between the information provided by different sites. Most of the interesting issues in phylogeny reconstruction are in how to resolve these conflicts.

A natural way is to count the minimum number of base substitutions that are required for each proposed tree, (leaving aside for the moment the issue of insertions and deletions). That tree requiring the fewest changes is preferred. This is the parsimony criterion. It was first introduced, in the context of estimating phylogenies from gene frequencies, by Edwards & Cavalli-Sforza

(17, 18), who called it the "method of minimum net evolution." The word "parsimony" was first associated with it when Camin & Sokal (6) published an influential description of this method for discretely coded morphological characters. Eck & Dayhoff (16) described the first application to molecular sequences. The algorithms for counting changes among states were given by Kluge & Farris (85; see also 28) on a linear or branched scale, and by Fitch (53) for nucleotides among which changes can occur from any one to any other.

The parsimony method is usually justified by the school of "phylogenetic systematics" by asserting that the count of extra state changes on a tree counts the number of ancillary hypotheses that must be erected to explain evolution in the group, and by identifying the criterion with William of Ockham's principle of parsimony, "Occam's Razor" (142). Along with this view goes the assertion that the use of parsimony requires no substantive assumptions about evolutionary processes, a position that when viewed from the standpoint of statistics, is questionable at best.

Normally, parsimony methods applied to nucleotide substitutions count only base substitutions. Sankoff et al (118) applied a method, later described by Sankoff & Rousseau (120) and Sankoff (119), that performs alignment of sequences at the same time as it estimates the phylogeny by minimizing a weighted count of substitutions and deletion/ insertion events. A more recent description of the class of methods is given by Sankoff & Cedergren (121). This process is computationally intensive but will receive more attention when sequence aligners realize, as they must, that multiple-sequence alignment is best carried out with explicit reference to the phylogeny and that one cannot simply treat all sequences symmetrically, when some may be near-duplicates of others. The realization of this will have a large impact on multiple-sequence alignment and may cause some embarassment when it is noted that David Sankoff and his colleagues understood the matter clearly in 1973.

The particular case of protein sequences has caused some difficulties. In Eck & Dayhoff's original parsimony method for protein sequences (16), they allowed any amino acid to be replaced by any other. Subsequently Dayhoff & Eck (14) used a set of weights that reflected the empirical probabilities of replacement for each possible change. Fitch (53) suggested counting not the number of amino acid replacements but the underlying number of base substitutions implied by the amino acid sequences. Because of the complexity of the mapping from codons to amino acids, this is not simple to compute. Algorithms for counting the number of base substitutions have been given by Moore et al (97), Moore (98), Fitch (54), Fitch & Farris (55), and Moore (99). In my own program for protein parsimony in the PHYLIP package, I have preferred to count only those base substitutions that also change the

amino acid, under the assumption that the synonymous changes are substantially more probable and should thus be deemphasized. This is more easily accomplished than counting all base substitutions.

COMPATIBILITY. A method closely related to parsimony is compatibility analysis (usually called a "clique method" by those who dislike it). It uses a different criterion for resolving conflict among characters. A character is compatible with a phylogeny if its evolution can be explained without assuming that any state arises more than once. Thus a site that shows three bases, A, C, and T, is compatible with a phylogeny if the observed data could arise with only two nucleotide substitutions. The compatibility method finds that tree on which the maximum number of sites are compatible with the tree.

The compatibility criterion was first proposed for discrete two-state morphological characters by Le Quesne (89). Estabrook & Landrum (23) and Fitch (56) showed how to determine whether two nucleotide sites are compatible with each other, in the sense that there must exist a tree on which they can both evolve with no extra changes. However, Fitch (56) also showed that a set of sites that are all pairwise compatible may not be jointly compatible, in that there may not exist one tree on which all can evolve without extra changes. This is in contrast to some classes of multistate morphological characters for which Estabrook et al (24, 25; see also 26) proved that when characters are all pairwise compatible, they must be jointly compatible, and the tree fitting all of them can be found very easily.

Although the absence of this pairwise compatibility theorem for nucleotide sequences makes it somewhat harder to find the tree with the most sites compatible with it, compatibility methods are no harder to use than parsimony methods. It should be apparent that the two classes of methods are closely related, although some authors, e.g. Wiley (142), have felt otherwise.

Distance Matrix Methods

Distance methods, the second major category, fit a tree to a matrix of pairwise distances between the species. For nucleotide sequence data the distances might, for example, be calculated from the fraction of sites different between the two sequences. The phylogeny makes a prediction of the distance for each pair as the sum of branch lengths in the path from one species to another through the tree. A measure of goodness of fit of the observed distances to the expected distances is used, and that phylogeny is preferred which minimizes the discrepancy between them as evaluated by this measure. There is a widespread misconception that distance methods assume a molecular clock, mostly because molecular evolutionists using these methods have also tended to make such an assumption and invoke it as the reason why their methods

work. It is possible to either assume or not assume a molecular clock when using distance methods.

Fitch & Margoliash (52) introduced the first distance matrix method, and Cavalli-Sforza & Edwards, (8) had independently produced another. Both were least squares methods. If the D_{ij} were the observed distances and the d_{ij} the expected distances computed from the tree, then the measure of lack of fit was

$$\sum_{i,j} w_{ij}(D_{ij} - d_{ij})^2,$$

which is a weighted least squares measure. The weights w_{ij} were $1/D_{ij}^2$ for Fitch & Margoliash's method, and 1 for Cavalli-Sforza & Edwards's method. These represent a different weighting of discrepancies for large and small distances.

Many other distance matrix methods have been introduced. Some such as Farris's (30) "distance Wagner method," Li's (91) method, Tateno et al's (135) "modified Farris method," and Saitou & Nei's (116) "neighbor joining method" are not defined in terms of a measure of lack of fit, but only as the result of following a certain algorithm which joins species and calculates branch lengths. The algorithms involved are designed to yield an exact result when there is a tree that perfectly fits the data, but it is less easy under this approach to see how different kinds of discrepancies from a perfect fit are weighted. This makes statistical analysis of the properties of these methods particularly difficult.

Chakraborty (12) has taken the opposite tack and derived a least squares method from a statistical model, one which tries to take into account the variances of the distances and the correlations between them, when protein sequences are used. Hasegawa et al (64, 67) have derived a distance method from statistical properties of nucleic acid sequences. Hogeweg & Hesper (70) have derived a distance from pairwise alignments of molecular sequences and have inferred phylogenies by using this in a distance matrix method. This differs from the approach of Sankoff et al (118) in that there need not be any consistency between the alignments for different pairs of species—Hogeweg & Hesper's method is thus necessarily more approximate.

The widely used UPGMA method, or "average linkage method" (Sokal & Sneath, 129) of constructing a tree from a distance matrix is also defined as the result of applying a certain algorithm. That algorithm would work perfectly only if the data were generated by a clocklike evolution—if the data were an exact fit to a nonclocklike tree the UPGMA method could give erroneous results (13, 29, 96). The UPGMA method is, however, not as arbitrary as might first seem. Farris (27) and Chakraborty (12) have pointed

out that it assigns the branch lengths (or node levels) so that the sum of squares of differences between observed and expected distances is minimized. The topology is found somewhat arbitrarily as a result of the clustering algorithm rather than by an explicit search among alternatives, but otherwise the relationship between versions of least squares that assume a clock and the UPGMA method is a close one.

Likelihood Methods

Maximum likelihood is the most general method of deriving statistical estimates. In essence it is quite simple—one has a model (M) and data (D). The likelihood of a tree (T) is the probability of the data given the tree and the model, P(D; T, M), considered as a function of the tree. The probability of all possible sets of data must add up to one, but when the data is held constant and the tree is varied, the different values of P(D; T, M) need not add up to one and are called likelihoods rather than probabilities. The maximum likelihood method simply chooses that tree T which maximizes the likelihood, thus maximizing the probability that the observed data would have occurred. Likelihood methods are not as widely known as they ought to be, because the computation of the likelihood frequently involves taking products of a large number of quantities or sums of logarithms. Before the existence of computers likelihoods were hard to compute, and methods based on them were regarded as arcane and impractical. They have only recently begun to make their way into the elementary statistics texts studied by biologists.

It was inevitable that maximum likelihood would be applied to estimating phylogenies. Edwards & Cavalli-Sforza (18) made the first attempt, with gene frequencies as the data. The first application to molecular sequences was by the famous statistician Jerzy Neyman (105), who used a simple model of symmetric change among amino acids or nucleotides, with changes occurring randomly and independently at different sites. This was closely similar to the model implicit in Jukes & Cantor's (75) formula relating the time of divergence of two species to the probability of net change in a base. It ignores differences in the rate of transitions and transversions, and it does not allow for different frequencies of the four bases or different rates of change at different sites. Neyman investigated only the case of data from three species.

Kashyap & Subas (77) wrestled with the problem of combining Neyman's three-species trees for all triples of species in a data set into one larger tree. Their methods were somewhat ad hoc. I gave (38) computationally effective methods of computing the likelihood for a tree with an arbitrary number of species, and of finding branch lengths that maximize the likelihood. The model used allows unequal base composition and does not assume a molecular clock. More recently it has been extended to allow differences between the rates of transition and transversion and to allow different prespecified rates of

change at different sites (J. Felsenstein, in preparation). Hasegawa and his colleagues have applied maximum likelihood to a number of nucleic acid sequence data sets (65, 66, 68).

Bishop and Friday (3a) have used several models of base substitution to construct a maximum likelihood method for inferring rooted phylogenies under the assumption of a molecular clock. They have applied these to some published nucleotide sequencies on mammals, and discuss extensively some of the changes that would have to be made in models to make them more realistic.

Barry & Hartigan (3) have developed a maximum likelihood method which, instead of assuming that the parametric form of the matrix of base changes is known in advance, estimates it from the data. This turns out to simplify computations considerably. The disadvantage is that it allows too great a flexibility in the probabilities of change between specific bases, so that what it gains in flexibility it may lose in power from having to estimate more parameters. Processes of base change probably do not differ much in related species, a factor Barry & Hartigan's method does not take into account. On the other hand, methods such as my own assume that the processes do not change at all in different parts of the tree. The truth must lie somewhere in between.

Saitou (117) has derived conditions under which maximum likelihood on a clocklike tree will give the correct results, and compared those to conditions for parsimony and UPGMA methods. For three and four species the likelihood method is found to behave similarly to UPGMA. It is not clear whether this will generalize to more species.

It is worth noting here that maximum likelihood methods have also recently been applied to restriction sites data (76, 15, 102, 92, 124) where they are needed to correctly account for the relative rates of parallel loss and gain of sites.

STATISTICS AND THE JUSTIFICATION OF METHODS

It is unsatisfactory to have several competing approaches if it is not understood how they differ in their assumptions, and thus when one ought to prefer one to another. The two main approaches to justifying phylogenetic methods are the hypothetico-deductive and the statistical. The former has been applied mostly to parsimony methods, under the belief that William of Ockham's principle that entities ought not to be multiplied unnecessarily (called "Occam's razor") is directly related to parsimony, which is said to measure the number of hypotheses that must be erected to explain a data set. That in turn is related, by authors such as Wiley (141, 142), to Popper's hypothetico-deductive model of falsification of scientific hypotheses. The

central flaw in this argument is that falsification is not absolute—when Wiley (141) says, "the phylogenetic hypothesis which has been rejected the least number of times is to be preferred over its alternates," he is trying to stretch the original Popperian argument to cover parsimony, which may have every possible phylogeny rejected by requiring extra changes of state in one or another character. Rejection then inevitably is not absolute, and statistical concepts must be admitted through the back door.

The other, preferable way to justify methods is to consider them as methods of statistical inference and investigate their statistical properties. The biological assumptions of a method may be found by asking which ones endow it with reasonable statistical properties. The issue is subtle because statisticians do not agree on the most important properties of a statistical method.

Consistency

A statistical estimation method is consistent if it approaches the true value of the quantity as larger and larger amounts of data are accumulated. For example, the mean of a sample from a normal distribution gets closer and closer to the quantity it estimates, the true population mean, as the number of data points increases. Statisticians differ on how fundamental a property consistency is: Bayesians and advocates of likelihood relegate it to a lesser role while most others consider it a fundamental desirable property of an estimation method.

Maximum likelihood methods are usually consistent, with the exception of certain cases where the number of quantities being estimated rises at least at the same rate as the number of data points. In the case of phylogenies, the parameters being estimated are the branch lengths of the tree but may also include the states of hypothetical ancestors that occur at interior nodes of the tree. If only branch lengths are estimated, the number does not change as more nucleotide sites are considered. However, if we are also estimating the nucleotide states in the interior nodes of the tree, the number rises proportionately to the length of sequences considered, and the estimate may be inconsistent. This will become relevant when we discuss the inconsistency of parsimony and compatibility methods.

CONSISTENCY AND DISTANCE MATRIX METHODS Distance matrix methods are consistent when the distances are derived from sequences and certain conditions are met. We expect that as the number of sites sequenced rises, the distances measured approach more and more closely to their expected values. If the expected values are the sums of the branch lengths through the true tree from one species to another, then in the limit there will be a perfect fit between the tree and the distance matrix, and the method will be consistent.

We must transform the distances so that their expected values are equal to the total branch lengths intervening between two species. This is an important criterion often overlooked when distance matrix methods are applied. Chakraborty (12) made an effort to correct the distances derived from protein sequences to achieve linearity. Olsen (106, 107) also has carried out such a correction when using distances derived from nucleotide sequences. Farris (31, 33, 34) and I (43, 46) have discussed different options for making this correction, either transforming the distances or using a nonlinear least squares method.

Simple use of a Fitch-Margoliash or other least squares method with a distance that measures the fraction of nucleotides different between sequences is inconsistent. The expected distance between two species rises at first nearly proportionally to the intervening branch length, but as we consider longer paths through the tree we expect more and more cases in which one substitution overlays or reverses another. For example, when we expect 10% nucleotide sequence difference between nodes A and B on a tree, and a further 10% between B and C, then under a simple symmetric model of change among four nucleotides (such as that of Jukes & Cantor, 75) we expect that 1% of the sites have been changed twice between A and C. One third of these double changes will cause reversion to the original nucleotide, so that the net difference between the sequences of A and B is expected to be not 20% (as would be predicted by adding up the branch lengths) but 19.67%. Thus the branch lengths will not be additive: the expected distances will be less than the sum of the branch lengths, particularly when that sum is large. To the extent that a distance method is trying to fit the tree to both long and short distances, it will make the branches too short as a result of this problem of overlaid substitutions.

This may not seem like a very serious problem with the example given, but it becomes severe with larger differences between sequences. As two DNA sequences become very far apart in the tree, the branch length between them should rise towards infinity, but their sequence difference cannot rise above 100%, and in fact will approach 75% under the Jukes-Cantor assumptions. With more realistic models of nucleotide substitution, involving unequal frequencies of the four bases, the problem becomes even worse. Branches in the tree may be substantially shortened in order to have the branch length between less closely related species fit a distance of 75% which actually reflects much larger amounts of nucleotide substitution.

The objections raised by Farris (31, 33, 34) to the use of distance matrix methods consist in part simply of pointing out this problem. In my responses (43, 46) I have agreed that this is potentially a problem, while emphasizing that there are ways to correct it. The remainder of Farris's critique is that the branch lengths estimated may not be achievable. Thus, we may estimate a

branch length of 0.17 in a case in which the data consists of sequences of length 50 nucleotides, while it is impossible that the actual sequences at the two ends of that branch differed by exactly 17%. I have pointed out (43, 46) that this causes no problem if we think of making a statistical estimate of the tree. The branch lengths are expected differences between two sequences; the expected difference is a weighted average of the distance over all possibilities, and as such need not be a quantity that is equal to any of the actual differences. Farris's objections do not apply if one adopts, as I am urging that we do, a statistical inference approach to inferring phylogenies.

CONSISTENCY, PARSIMONY, AND COMPATIBILITY If the issue of consistency of distance methods is complicated, the issue with regard to parsimony and compatibility methods is positively baroque. It interacts with the logical justification of parsimony and with the question of when parsimony and compatibility methods are equivalent to maximum likelihood methods.

Cavender (9) and I (36) discovered a simple case in which parsimony and compatibility methods would be inconsistent. The example involves a four-species case with unequal rates of evolution among two lineages. The sites are assumed to change independently. The original case involved two-state characters, but an equivalent example can be constructed for four-state characters such as nucleic acid sequences (38). The topology of the unknown true tree is of the form ((A, B), (C, D)). The branches leading to species A and D are long, and all the others are short, where by length we mean not time but expected amount of change, as no molecular clock is assumed. Random change along this tree, in accordance with the branch lengths, generates many sites that have parallel changes in the lines leading to A and D, as one quarter of cases in which both of those lines change result in the same nucleotide arising in both of the lineages.

If the internal branches of the tree are short enough, it generates fewer sites which are "phylogenetically informative" in the sense of having one base in common between species A and B, and another in common between C and D. The upshot is that we expect to have more sites providing false evidence that the tree topology is (A, D), (B, C) than provide evidence of the true topology (A, B), (C, D). As we collect more and more sites, the chance that a parsimony method will chose this particular wrong topology becomes higher and higher, ultimately approaching 100%. With four species there is no difference between parsimony and compatibility methods, which in these cases always give the same results; thus, this is a counterexample to the use of either parsimony, or compatibility.

I was able (36, 41) to derive conditions for some particular patterns of branch length, showing for what combinations of their lengths parsimony

methods would be inconsistent. There was a trade-off between inequality of the expected rates of evolution in different branches of the tree and the overall rates of change: with less change one needs more inequality of rates to have inconsistency, whereas with clocklike evolution no combinations of branch lengths led to inconsistency. With grossly unequal lengths of branches, inconsistency could occur even with little expected change. Hasegawa & Yano (66) carried out computer simulations of the evolution of DNA sequences and verified these patterns.

Hendy & Penny (69) have developed a clever method using matrix algebra to generate the expected frequencies of different patterns of characters, including some models of nucleotide sequence change for cases with more species. They could show that the same phenomena occurred in some five-species cases, but with a surprising difference. They found the same pattern that "the long branches of the tree attract each other," causing inconsistency when the long and short branches are sufficiently different in length. But they were able to find cases in which parsimony (and incompatibility as well) were inconsistent, even with a perfect molecular clock, which disproved my conjecture of a trade-off between clockness and inconsistency. Apparently parsimony and compatibility are even less well-behaved than I had inferred. The two patterns that continued to hold up were that the inconsistency arose when branch lengths were unequal, and the smaller the overall rate of change the more unequal the branch lengths need to be to cause inconsistency.

An intriguing modification of parsimony methods is proposed by Hendy & Penny (69). They suggest that instead of counting changes of state, we should use the number of observed changes in each branch of the tree to reconstruct the estimated actual number of changes. Thus if a branch shows 10 changes out of 20 characters, we can compute (for a simple four-state nucleic acid model) how many substitutions have not been seen because they have been reversed or overlaid by other substitutions. They suggest that trees be scored according to this augmented number of substitutions. In their consistency calculations they found that this augmented parsimony method was always consistent, even when ordinary parsimony was not. This is an interesting approach to avoiding the inconsistency problem entirely. There is as yet no proof that it always does avoid the problem, and there may be ambiguities as to where changes occur in the tree which affect the augmentation calculation. The method may not yet be fully developed, but it is certainly promising.

ARGUMENTS AGAINST THE COUNTEREXAMPLES The examples of the inconsistency of parsimony and compatibility have generated considerable controversy, because if they are accepted they create a problem for the

hypothetico-deductive approach to inferring phylogenies. Farris (32) has pointed to the unrealistic nature of the model under which the inconsistency is derived—independently evolving characters, all evolving at the same average rate symmetrically among four (or two) states. His line of argument is unusual: "This is not to say that parsimony requires no assumptions at all; it presumes, one might say, that Felsenstein's models are unrealistic. But as that assumption seems generally agreed upon, that is not much of a criticism of parsimony" (32). The difficulty with his argument is that it implicitly presupposes that the special assumptions of the model are responsible for the inconsistency, and that a more realistic model would not be inconsistent. There is in fact no evidence for that whatsoever; there is no reason for believing that the inconsistency would not also occur in more realistic models. In fact, it can easily be shown that variation of rates of evolution among characters and correlation of the characters in some patterns will leave the inconsistency unchanged. Farris's argument therefore is insufficient reason for ignoring the possibility of inconsistency. His assurance that no controversial assumptions are involved in using parsimony is wrong—there is in fact no guarantee that parsimony will work well in any realistic case.

Sober (127, 47) has taken me to task, with considerable justification, for overstating the implicit assumptions of the parsimony methods by saying that they require evolutionary rates (as reflected in the expected amounts of evolution in branches) to be small or nearly equal in different lineages for parsimony to be consistent. Hendy & Penny's (69) work shows that for five species the conditions for parsimony to be consistent seem even more stringent than my projection based on four species. Nothing general is known about the conditions for consistency for more general models.

Carpenter (7a) summarized the state of affairs after the debates between Sober and myself by saying that Sober has "at least wrung from Felsenstein the retraction of his claim that parsimony necessarily assumes low rates of evolution." I see the matter differently. We know what the conditions are for inconsistency of parsimony for some particular four- and five-species models, and these suggest that the problem may extend well beyond those cases. Is this reason for complacency on the part of users of parsimony methods? None of the advocates of the position that parsimony has no controversial assumptions has presented any general proof that this is so.

Likelihood as Justification

LIKELIHOOD JUSTIFICATIONS FOR PARSIMONY Sober (126, 127, 128, 47) has taken a different tack, rejecting the notion of consistency itself as a fundamental property a statistical estimator ought to have. There are statistical positions (notably Bayesian and likelihoodist positions) in agreement with him in this, so that the matter unfortunately involves the philosophical

foundations of statistics, which biologists are unlikely to resolve on their own. Many statisticians, probably a majority, accept consistency as a fundamental desirable property of an estimation method, and I think many biologists agree.

Sober argues against the relevance of the consistency property because he is defending the use of the parsimony criterion. It is fair to ask what positive properties of parsimony a supporter would invoke. Sober's advocacy is based on his assertion that parsimony is the same, under noncontroversial assumptions, as maximum likelihood. His basis for advocating parsimony is a likelihoodist position that takes maximum likelihood as fundamental, regardless of whether the resulting estimator is consistent. This is a well-known statistical position, so again biologists are unlikely to resolve the matter by themselves.

Sober's position depends on some proof that parsimony methods are generally identical to maximum likelihood methods. He has presented such a proof (126, 127) in three-species cases with two-state characters, but it contains a step in which a particular internal branch length in the 3-species trees being compared is assumed to be identical. Recently, he retracted this proof as flawed (47, 128). At present we have no general proof of a correspondence between likelihood and parsimony, so that even if one takes a likelihoodist position and rejects the relevance of the consistency property, there is no clear guide as to what method of phylogenetic inference is to be used, other than direct use of maximum likelihood.

I have presented one proof (35) that when rates of evolution per unit branch length are taken towards zero with the lengths of branches held constant, then for any two trees and with a fairly general model of change among character states, the tree of higher likelihood will be the one with the fewer changes of character states. This proof establishes an equivalence between likelihood and parsimony, but only for cases with low expected amounts of character state change. This at least makes intuitive sense: if we expect very little change, then that tree which requires the fewest of these improbable events will provide the most credible explanation of the data. The problem with using this argument as a justification of the use of parsimony methods is that in many data sets we see rates of evolution that are not small.

COMPATIBILITY AND LIKELIHOOD In the studies showing that parsimony methods can be inconsistent, the cases investigated do not discriminate between parsimony and compatibility—since the two methods always yield the same result in those four- and five-species cases, the proof of inconsistency applies equally to compatibility methods. The debate has centered around parsimony since it is in more widespread use, and the school of systematists most committed to a hypothetico-deductive approach to phylogenetic inference identifies that approach with parsimony.

When we consider instead the circumstances under which compatibility and likelihood methods are identical, we get a slightly different answer than we do for parsimony, and the differences illuminate the assumptions of the parsimony and compatibility methods. In my examination of sufficient conditions for likelihood to be identical to parsimony when there are two character states (37) I investigated a variety of parsimony methods, including Dollo, Camin-Sokal, and polymorphism parsimony, and also compatibility methods. For compatibility to be identical to likelihood, it turns out that the homoplasy (parallelism or convergence) should not arise from random evolutionary changes occurring at a slow rate in all characters, but rather that most characters should have a very low rate of change and a few should have a high rate of change or of misinterpretation.

Compatibility methods tend to ignore the information from those characters that do not fit a phylogeny, although they do consider various possibilities and try to ignore as few characters as possible. If we assume that all characters will change at a low rate and hence tend to fit the true phylogeny, except for a few that will be almost useless because of misinterpretation or high rates of evolution, this behavior becomes explicable. Once a character has exhibited more than one change on a tree, it becomes probable that it is one of these misinterpreted or rapidly evolving characters, whose distribution should have little or nothing to do with the phylogeny. These characters are expected to be rare, so that we should assume as few of them as possible.

Thus the different treatment by parsimony and compatibility methods of characters that do not fit the tree is different in a way that corresponds to a diffferent assumption about the source of the homoplasy. It is natural to suggest that compatibility methods implicitly assume this sort of pattern of evolutionary rates, but as with the case of parsimony, we can only say that they assume this in the few cases that have been investigated, without having a prooof of what they assume in general.

An interesting issue that arises with use of compatibility methods on nucleotide sequences is to determine when we are to consider a site to be incompatible with a tree. The usual definition is that if each nucleotide state arises no more than once, it is compatible. However, all the likelihood arguments suggest otherwise—that two changes in the same site, even if they lead to different nucleotides, should be counted as evidence that this site has a high rate of change and, hence, should be ignored in making the tree. When this is used as the criterion for compatibility, the pairwise compatibility theorem can be used and construction of trees becomes much more straightforward. As far as I know this approach has never been used.

CHARACTER WEIGHTING AND LIKELIHOOD I have discussed (39) the assumptions of compatibility and parsimony in the context of character

weighting. When different characters had different but small rates of evolution, it could be shown that a weighted parsimony method was identical in result to a maximum likelihood method, the weights being related to the negative logarithms of the rates of character change. Thus the faster a character is known to change, the less weight it should be given. The lower the rates of change the more equally the characters should be weighted, so that unweighted parsimony methods may implicitly assume a low rate of change.

If one also assumed that a small fraction of the characters evolved at such a high rate as to be completely devoid of information, then a parsimony method with a threshold emerged as equivalent to maximum likelihood. If the probability that a character has a high rate of change is the same as the probability that it has four changes of character state on the tree, then the threshold should be set so that in each character we count the number of steps up to four, counting four for that character no matter how many more steps there are. For these two-state characters, a threshold value of two turns out to be equivalent to using a compatibility method. A character is then simply evaluated as to whether it has more than one change, and the minimization of the count of changes modified by the threshold is identical to maximizing the number of characters that can be interpreted as uniquely derived. Thus we have a family of methods that smoothly connect parsimony and compatibility, showing that they are indeed closely related. A similar family was presented by Farris (27a), although without a likelihood justification. He (32) has commented on these issues at length. Another a posteriori weighting method was developed by Penny & Hendy (110).

STATISTICAL TESTS OF PHYLOGENIES

So far, all of the discussion has been in terms of consistency of the point estimate of a phylogeny, when the estimate is, or is not, identical to a maximum likelihood method, and what this may mean about the implicit assumptions of the methods. The question of how to obtain confidence intervals and carry out statistical tests is in a relatively primitive state by comparison, but it is of greater practical importance to the molecular evolutionist. We cover here the suggestions that have been made for tests and confidence intervals based on parsimony methods, distance methods, and likelihood methods, and then data resampling approaches such as bootstrap methods.

Tests Based on Parsimony Methods

CAVENDER'S CONFIDENCE INTERVAL The pioneering investigations of how confidence intervals could be constructed based on parsimony methods

have been described in papers by Cavender (9, 10). He examined the four-species case with characters having two states. I have reworked his calculations (41) for the case of four states, such as nucleotide sequences. Cavender used as his statistic the number of differences in substitutions between the most parsimonious phylogeny and its next best competitor. He asked for what true phylogeny there would be the most evidence favoring the wrong topology, as judged by parsimony. He discovered the inconsistency problem, that in the worst case all "phylogenetically informative" characters might be expected to favor the wrong topology.

In the nucleotide sequence case, with a simple symmetric model of base change, it turns out that 3/16 of all characters would be expected to be "phylogenetically informative" and favor the wrong tree (41). In the original two-state case Cavender found the corresponding number to be 1/3. Each of these "phylogenetically informative" characters creates a one-substitution difference between the wrong tree and the correct one. One can only conclude in favor of the most parsimonious tree if the evidence is stronger than that. Cavender therefore asked whether the number of steps favoring the most parsimonious tree over its next best competitor was significantly greater than one third of the number of characters. For the nucleic acid case one asks whether it is significantly greater than 3/16 the number of sites. Note that it is the total number of sites that is used, not the "phylogenetically informative" ones, 100% of which can favor the wrong tree in the worst case.

Table 1 shows the results recalculated for the nucleic acid sequences case. The third column gives the significant number of steps expressed not in terms of all sites but in terms of all varying sites, so that we have omitted those that have the same base in all four species. The calculation in terms of varying sites uses the fact that in the worst case 1/16 of the sites will be invariant, so that the expected fraction of sites which favor the wrong tree by one substitution is 3/15 per varying site rather than 3/16 per site.

THE CONFIDENCE INTERVAL ASSUMING A CLOCK Cavender's calculations assume no evolutionary clock. When a clock can be assumed, the bounds can be made much tighter. I have (45) used the fact that when there is a clock the worst case is no longer one that has all of the "phylogenetically informative" sites backing the wrong tree. With three species (or four, if one has an outgroup) the worst case is the trifurcation—this is the tree of one topology most likely to give evidence favoring another topology. Each "phylogenetically informative" site has a 1/3 chance of favoring each of the three possible tree topologies. For this worst case, one can, by considering all possible data outcomes in turn and working out the probability of each, tabulate the distribution of the number of steps by which an incorrect tree will be favored.

Table 1 95% point of distribution of difference in number of substitutions

Sites	All sites	Varying sites only	Informative (clock)
2	2	2	—
3	3	3	—
4	3	3	4
5	3	4	5
10	5	5	5
13	6	6	5
15	6	7	6
20	8	8	6
25	9	9	7
30	10	11	8
40	13	13	9
50	15	16	9
100	26	28	13
200	48	50	17
500	109	116	27
1000	209	222	
2000	405	431	
5000	984	1048	
10000	1940	2067	

The final column of Table 1 gives the 95% points of this quantity for various numbers of "phylogenetically informative" sites.

The application of these numbers can be illustrated using the data of Miyamoto et al (95). They examined 7100 sites of sequence, found 391 sites that varied, of which 13 were phylogenetically informative, having 8, 3 and 2 sites, respectively, that favored human-chimp, chimp-gorilla, and human-gorilla clades. Using Table 1, we find that with 7100 sites one would need to have the best tree favored by about 1385 steps to be significantly better than the next best. If we confine our attention to the 391 varying sites and use the second column, the required differential in the number of steps drops to about 95. This still leaves the result wildly insignificant. However, if we are allowed to assume a molecular clock, then we find that with 13 "phylogenetically informative" sites we need a differential of only 5 steps, exactly the number found. This indicates that these data favor human-chimp by an amount barely significant at the 95% level.

Another calculation could ask whether the number of sites favoring the best tree is significantly greater than 1/3 (42). The result of 8 out of 13 does not quite reach the 95% point, which is 9 sites. This is a different way of using the same data and points out that it is not obvious which statistic to use.

TEMPLETON'S PAIRWISE TEST Templeton (138) has used the same sort of data differently. He asks, for two given trees, whether the data supports one significantly more strongly than the other. Looking at the differences of numbers of substitutions at each site, he does a Wilcoxon signed-ranks test of the hypothesis that the sum of the number of substitutions is equal in the two trees (which is the same as saying that the mean number of substitutions is equal). For fewer than six species the number of substitutions per site cannot differ by more than one per site. One can therefore simplify Templeton's test by comparing the number of sites favoring the one tree with the number favoring the other, and test these for departure from one half by a sign test (A. Wilson, personal communication).

I have used the technique of enumerating all possible data outcomes in a three-species case (45) to check whether Wilson's simplified version of Templeton's test is conservative. It turned out that it was, provided that the sign test is done as a two-tailed test, rather than one-tailed as Templeton had recommended. This seems necessary because we do not know in advance which tree is going to be best; even if we examine them and order them by number of substitutions immediately before doing the test, that does not change the necessity for doing a two-tailed test. Applied to the Miyamoto et al (95) data, we test the best two trees against each other, comparing the 8 characters supporting one to the 3 supporting the other. We find that 8 out of 11, on an expectation of 1/2 has a two-tailed value of $P = 0.22$, so that the result is not significant. It should not be surprising that we get slightly different results using different statistics. All seem to be telling us that these data are near the level of significance but at most barely beyond it.

The advantage of Templeton's test, and Wilson's simplification of it, is that it is not restricted to the three-species case. We can test any two trees against each other to see which is significantly more strongly supported by the data. The test does not construct a confidence interval—it simply tests two predesignated trees. If both are ill-supported by the data we may find ourselves in the absurd position of proving that one bad tree is significantly worse than another. Later we see some more recent developments of this test in the direction of constructing confidence intervals. This family of tests has scarcely ever been applied, but note that Holmquist et al (71) report that Prager & Wilson have made use of the sign test in analyzing primate mitochondrial sequence data.

SNEATH'S DISTANCE TRIADS Sneath (125) has developed formulas for estimating variances and covariances of lengths of adjacent branches in trees computed from sequence data. His methods use triples of reconstructed branch lengths in the interior of the tree, computing their variances by several

approximate methods. Although he has performed some simulation checks, little is known about how accurate his methods will be or how they relate to the tests mentioned above.

Distance Methods

F TESTS When the trees are generated by distance methods, we can sometimes use classical least squares methods to test hypotheses about them. I have outlined (43) how to use least squares methods to test whether a tree assuming a molecular clock fits the data better than one that does not assume it. The test uses the F distribution and assumes that we have obtained the same tree topology under both assumptions—in effect the test is that the branch lengths satisfy the constraints imposed by a clock. Barry & Hartigan (3) have used a similar approach to hominoid DNA hybridization data. Rohlf & Sokal (111) have presented a closely similar test in a clustering context. The same test can be used, in another variant, to find confidence limits on the length of any one branch, or joint limits on the lengths of any two branches. I have also argued (46) that we can use the F test to conservatively test whether a tree topology adjacent to the best one can be rejected. As the discussion of tests based on likelihoods shows, it can be argued that this test is incorrect.

There is in any case a serious flaw in using the F test on distances derived from sequence data. For such tests to be valid we must be able to assume that the distances are statistically independent, which will essentially never be true if they are derived from sequence data. A random change in a sequence will affect the distance between that species and all others in the tree. For example, a random change in the sequence of the ancestor of all primates will affect the sequences of all primates and thus all the distances between primates and nonprimates. Statistical fluctuations of distances from sequences will not be independent. For this reason the F test is not useable for sequence data (or for distances derived from restriction sites, restriction fragments, or gene frequencies).

THE RELATIVE RATE TEST Sarich & Wilson (122, 123) introduced the "relative rate test" which they used to investigate whether there has been a change in the rate of evolution on one branch of a tree. An outside reference species is used, and descendants of two sister lineages compared. For example, we might use a baboon as outgroup and compare the gibbons with the other apes. The objective is to see whether the baboon-gibbon distances are different from the other baboon-ape distances. If the source of statistical error in the distances is purely measurement error, arising independently in each pairwise distance, then the test can be conducted. But when we have distances derived from sequence data, in which individual substitution events can affect

many of the distances simultaneously, the values become correlated and cannot be treated as statistically independent.

A substitution in the ancestor of the African apes (human, chimpanzee, and gorilla), for example, will inflate the distances of all of these to the baboon. They therefore cannot be treated as independent observations, as is implicit in the relative rate test. It therefore seems that the relative rate test is sensitive to the error structure in the data, and inappropriate for distances derived from sequence data, unless greatly modified.

VARIANCES OF BRANCH LENGTHS Nei et al (103) have presented formulas for computing variances and covariances of branch lengths in trees derived from distance matrices, taking into account the variances and covariances of the distances when those are generated from sequence data. Their methods apply to purely clocklike trees, in which the times of forks are estimated by averages of pairs of species whose last common ancestor was that fork. Their formulas are closely related to those used by Chakraborty (12), although different in methods of approximation. Nei et al state that their formulas become tedious to compute when large numbers of species are involved. They share this with Chakraborty's formulas, which compute all the variances and covariances of branch lengths, but only at the cost of constructing matrices of size $n(n-1)/2$ by $n(n-1)/2$ and inverting some of them. For example, a study with 15 species would require manipulation of matrices 105×105 in size. Nevertheless, it is probably worthwhile to compute the variances and covariances of branch lengths to get a clearer picture of the effect of statistical error on the estimate of the tree.

The above approaches use distances that have been logarithmically transformed so as to be approximately linear with time. A more sophisticated approach would be to use the untransformed distances but allow them to depend nonlinearly on time. This has been done by Hasegawa et al (64, 67) who developed an interesting nonlinear distance matrix method specifically adapted for distances from nucleotide sequences. They compute two distances, one from transition differences and one from transversion differences. These depend nonlinearly on time, and they use nonlinear equation-solving methods to find numerically the optimum branch lengths. They also present formulas for the variances and covariances of these estimates. Like those of Chakraborty (12) and Nei et al (103), these involve computations with large matrices.

As yet no one has adapted any of these methods to the case where no molecular clock can be assumed. This could be done, although it might be so much algebraic work that a bootstrap resampling approach would be easier (see below).

TEMPLETON'S DELTA-Q TEST Templeton (139) proposed a nonparametric method for distance matrix data to test whether one tree was significantly more supported than another. He first replaced the table of pairwise distances by their ranks, then derived a test statistic, delta-Q, from these. His method and its application to published hominoid DNA hybridization data has been criticized by Ruvolo & Smith (114), Saitou (115), and Fitch (57), and defended by Templeton himself (140). The fundamental criticism is that, by reducing the data to ranks, much of the statistical power in the original data can be lost. For example, for four species even the cleanest data, analyzed by the delta-Q method, is completely unable to discriminate between the true phylogeny and any other. This lack of power is the price one often pays for robustness when using nonparametric statistics—and it is too easily overlooked.

The robustness gained is not total. For example, the test assumes that the statistical variation of the distance values is independent. This may be true with DNA hybridization values but is certainly not true for distances derived from sequences, as already mentioned. For that reason the delta-Q test would need substantial revision to apply it to sequences.

Tests Based on Likelihood Methods

When we consider the likelihood "surface" that results from the likelihoods of all possible trees (including all possible combinations of branch lengths), there are two general approaches to assessing the statistical variability of the results. For a given tree topology, we can use the curvatures of the likelihood surface plotted as a function of branch lengths to compute approximate variances and covariances of the branch lengths. One need only compute a matrix of second derivatives for all pairs of branch lengths. The covariance matrix of branch lengths is the negative of the inverse of this matrix. That is a classical result in likelihood theory, but it is not quite as useful as it might seem. The result is asymptotic, valid only for large amounts of data, which in this case means very long sequences. In such a case there will be no ambiguity as to the tree topology—the covariances can be used to set up a simultaneous confidence interval on the branch lengths, with all the trees in the confidence interval having the same topology.

The issue of testing alternative tree topologies against each other, or of constructing a confidence interval that includes trees of more than one topology, is complex. It is best to discuss first the use of the likelihood ratio in other, simpler cases.

The likelihood ratio test (LRT) can be used to test whether some set of

parameters in a model are restricted to given values. If we have a tree and wish to test whether it is consistent with a molecular clock, for instance, that amounts to the assertion that certain branch lengths and sums of branch lengths must be equal. In Figure 1, we can obtain a clocklike tree like the one on the right by requiring that $v_1 = v_2$, $v_4 = v_5$, and $v_3 = v_4 + v_6$ (the requirement that $v_3 + v_8 = v_1 + v_7$ is not a restriction of parameters, since the tree on the left does not constrain v_7 and v_8 individually, as it is an unrooted tree). This means that a tree of 7 parameters, the branch lengths of the tree on the left, is constrained by the clockness assumption to have only 4 parameters, v_1, v_4, v_6, and v_7. In this case the LRT can be carried out: we find the best trees with and without the constraint. Twice the logarithm of the ratio of their likelihoods is (asymptotically) distributed under the null hypothesis of clockness as a chi-square variable with $7 - 4 = 3$ degrees of freedom. This test of clockness is perhaps the most complete test possible, but it has not yet been carried out on actual data. I have done such a test in the DNA hybridization case (48), but not in the sequence case.

There are other cases in which the LRT is useable. When we wish to test assertions about rates of evolution in different parts of a molecule, such as the assertion that evolutionary rates are different at the third codon position, we could find the maximum likelihood trees under both hypotheses and take the likelihood ratio. If, for example, we had one model in which all three codon positions changed at the same rate, and another in which there was a different rate parameter for the third position, the restrictive model constrains the two parameters to be equal. The LRT would then be applicable with one degree of freedom.

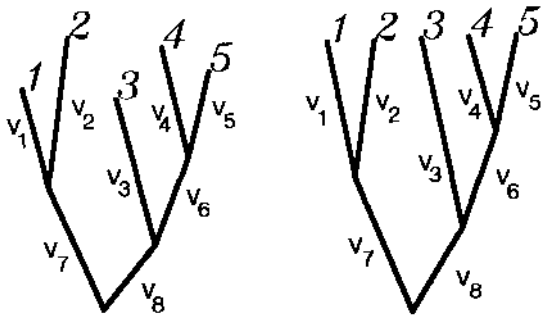

Figure 1 Phylogenies without (left) and with (right) the assumption of a molecular clock. Next to each branch of the trees is the branch length, which measures the expected amount of change, the product of the expected rate of change and time. In the tree on the right, the branch lengths are constrained so that all tips are level. The molecular clock is tested by testing for these constraints.

The LRT can also be used to improve on the confidence interval obtained from the curvatures of the likelihood. The likelihood ratio between the estimated tree and the true tree could be tested with a number of degrees of freedom equal to the number of branches ($2n-3$ if there is no clock assumed and there are n species). Thus by finding the likelihood that would be barely significant with this number of degrees of freedom, we can tell how far down the likelihood surface we should allow ourselves to go to define a confidence interval that would have a given chance of containing the true tree. This interval would probably be a better approximation than the one obtained from the curvatures, but it too is well-justified only when all trees in the interval have the same topology.

The LRT is applicable when the more restrictive hypothesis is a subcase of the less restrictive one, and when it is in the interior of the space defined by that hypothesis. In the third codon position example, the rate for the third codon could be either higher or lower than that for the others, so that the restrictive hypothesis is in the interior of the interval of possible rates. A more serious problem is that the LRT's justification is asymptotic. Technically it can be guaranteed to be correct only for very long sequences. In most cases statisticians ignore this requirement and hope that the LRT will behave well with smaller data sets. The phylogeny problems are not known to be any worse-behaved than others in this respect, but it would be desirable to have some verification, perhaps by simulation, of the adequacy of the LRT.

Testing different tree topologies against each other is much more difficult. When we test whether a particular branch could be of length zero, this is restricting a single parameter, and the LRT would have one degree of freedom. But the branch length cannot be negative, so the null hypothesis is on the boundary of the space. Owing to the continuity of the likelihood function in the vicinity of the trifurcation, this probably does not create a problem and we could still use the LRT (E. Thompson, personal communication). But when we wish to test one bifurcating tree topology against another, these hypotheses are not nested one within the other and have the same number of parameters. I have suggested in the case where the two topologies are adjacent (that is, they each have a branch which, if its length is shrunk to zero, results in the same trifurcation) that we could test one topology against another conservatively by pretending that there was one degree of freedom. I have used this test with DNA hybridization data (48).

Alan Templeton (personal communication) has pointed out that the logic I used in that argument is flawed, as the distribution assumes implicitly that the true tree has the trifurcation, whereas the intention is to use the test when it does not. It is possible that the conservativeness of this practice could be

proven, but at the moment the matter is not proven. Thus we are left with a situation in which many interesting hypotheses can be tested by likelihood methods, but alternative tree topologies cannot, at least without further work. Since the confidence interval consists of those trees that cannot be rejected, this also leaves open the question of how to construct a confidence interval.

APPLICATIONS AND EXTENSIONS Masami Hasegawa and his colleagues (65, 66, 68) have applied maximum likelihood to mitochondrial and RNA sequence data. They have used the approximate standard errors of branch lengths and have reported differences in log-likelihood among topologies. Ritland & Clegg (113) have applied likelihood analysis to a variety of problems in plant phylogenetics. They have extensively tested different models of base change against each other and have tested equality of rates of evolution in different regions of the genome, using the likelihood ratio test. They were cautious about interpretation of likelihood ratio differences among different tree topologies, turning to other methods (see below) when interpreting these. Barry & Hartigan (3) have also compared likelihoods, using their own models and allowing a different probabilistic process in each branch of the tree, between different tree topologies for hominoid DNA sequence data.

QUALIFICATIONS The main limitation of likelihood methods is that they require a precise parametric model of nucleotide (or amino acid) change. To the extent that this model is inaccurate, the inference drawn by using it may be wrong. Some authors, on hearing this point, have concluded that likelihood methods are particularly delicate. In fact, no such conclusion is justified. In the case of likelihood methods the model is explicit—for most other methods the model is implicit. Both kinds of method may be sensitive to violation of the model—it is just that in likelihood methods the model is more visible. There is no reason to believe that likelihood methods behave worse than parsimony or distance methods on real data, even when the model is not plausible.

That these models are not plausible should be apparent. Gillespie (62) has been witheringly skeptical, on empirical grounds, of all existing stochastic models of nucleotide sequence change. Heterogeneities of rate between different parts of the DNA are so extensive that it is impossible to believe in any of the models employed in likelihood analyses. Given that, and the hopelessness of finding a general and tractable model, should we not abandon attempts to use these highly parametric methods? We see below that we need not abandon them and that they can be greatly strengthened by being combined with empirical nonparametric methods.

Invariants

Closely related to likelihood methods are methods using invariants. These are functions calculated from the data that take one value for all trees of a given topology, irrespective of their branch length. The impetus to investigate invariants has come from the realization that parsimony methods can be inconsistent (as explained above) and that a method insensitive to branch lengths would have considerable advantages. The three papers on this subject are those of Cavender & Felsenstein (11) and Lake (87, 86). All investigate four-species trees, the smallest ones that have nontrivial differences. Cavender (11) discovered functions of the expected frequencies of different types of characters in a two-state case which were invariants, in the sense that they would be zero on the true topology. The functions were quadratic, and it was not easy to make a simple statistical method out of them. They were no longer zero if evolutionary rates varied from character to character, for example. What they did do was to express more explicitly what were the constraints on the expected frequencies of character outcomes that corresponded to having a tree of a given topology.

Lake (87, 86) found a different set of invariants with the property that they were nonzero only for the true topology. Lake's invariants are for a four-state case (modelling nucleotide sequences) and have the nice property of being linear rather than quadratic. This endows them with properties that avoid the problem of rate inequalities at different sites. If we consider four species and take all sites that are comparable in all four, some of these will show (for the four species, respectively) a pattern *xxzz*, where x and z are bases that differ by a transversion (such as $x = A$ and $z = T$). Some will be *xyzz*, where x and y differ by a transition and z from both by a transversion, and some will be *xyzw*, where x and y differ by transversions from z and w. Many sites will, of course, have other patterns such as *xxxx*, *xxxy*, *xzxz*, etc. Letting P(*xxzz*) be the fraction of sites which are expected to show pattern *xxzz*, and similarly for the others, the invariant is

$$\mathrm{P}(xxzz) + \mathrm{P}(xyzw) - \mathrm{P}(xyzz) - \mathrm{P}(xxzw),$$

which can be shown, under a suitably symmetric model of base change, to be nonzero for the tree topology ((A, B), (C, D)) and zero for the other two possible topologies. This means that the fraction of sites showing one of the patterns *xxzz* and *xyzw* should equal that showing one of the patterns *xyzz* and *xxzw*. Lake's statistical technique is to test this by counting these patterns in the data and doing a chi-square test of equality between these two classes of sites (an exact binomial test would work as well). Lake has also presented

(86) a method for estimating the lengths of the branches in the trees from similar calculations.

The advantage of linear invariants such as Lake's is that they make the inferences about tree topology in a way that is not sensitive to different branch lengths or different rates of evolution at different sites. This is a great advantage, but it is somewhat compromised by some disadvantages. The model of base change under which Lake derives his results has some built-in symmetries (when an adenine undergoes a transversion it must be equally likely to end up as cytosine or a thymine, and similarly for the other three bases) which may not reflect biological reality. It remains to be seen whether the method can be corrected for departure from that assumption. Secondly, by ignoring information from sites that do not have patterns such as *xxyz, xyzw,* and such, we inevitably lose some power. This is expected to be particularly pronounced in groups of closely related species, where transversion differences may be infrequent. With enough data, the method could be used even on fairly closely related species. Holmquist et al (72) have found Lake's invariants useful in discriminating among phylogenies of the higher primates using about 10 kb of sequence.

It is also not obvious how to extend the method to greater numbers of species. Lake (87) used an approximation to incorporate information from multiple sequences to ask whether a given interior node of a phylogeny exists. The approximation seems very rough; there should be a better way of using multispecies information.

Maximum likelihood methods do use information from multiple species correctly. They also make full use of all positions—even the invariant positions contribute to the estimation of overall evolutionary rate. However, the models employed may be unrealistic. The question of whether invariants are to be preferred to likelihoods thus depends on whether the models that underlie likelihood methods are likely to have broken down, without the symmetry assumptions of invariants having broken down. The matter is a subtle one and needs much further investigation.

THE BOOTSTRAP, THE JACKKNIFE, AND OTHER RESAMPLING METHODS

The Bootstrap and The Jackknife

Resampling methods have become popular in statistics in recent years. These involve using random sampling from one's own data to find out empirically the variability in the estimator. These methods, notable jackknifes and bootstraps, have been applied to phylogenies only recently. They provide us with a powerful way of escaping from some, if not all, of the restrictive assumptions of other methods. That is their great attraction—the conflicts between

information from different sites are assessed empirically. If different sites conflict, this will be reflected in a wider confidence interval. Conflict that goes well beyond what is expected under simple models of equal rates of change at all sites can easily be accommodated. Nonindependence of change at different sites cannot be so easily accommodated—in this respect these methods are not particularly robust.

MUELLER AND AYALA'S JACKKNIFE METHOD Mueller & Ayala (100) used the jackknife method to test the reality of a branch in a phylogeny, using gene frequency data. Their methods could be generalized easily to trees computed from sequences by distance methods. The jackknife is a resampling method in which one data point at a time is dropped. The estimate (in this case a branch length) is recomputed from the data left after the point is dropped. In Mueller & Ayala's case the points were loci; in sequence data they would be sites. Usually one drops all the sites in turn, but if the number of sites is large one could alternatively drop a random sample of sites, one at a time.

The collection of resulting estimates of the particular branch length are to be examined to see whether there is evidence that the branch length is greater than zero. It is important to realize that dropping one site will have a very small effect on the estimate, far smaller than the typical effect of sampling variability. In fact, we know how much smaller. If there are n sites, then the perturbation of the estimate by adding or dropping one site will typically be 1/n as large as the perturbation obtained by taking a completely new sample. In using the jackknife we compute "pseudovalues" of the estimate by taking the change in the estimate and extrapolating it by multiplying it by n. This is often left unclear because the formulas for the variance of the estimate incorporate the extrapolation factor without comment.

Mueller & Ayala drop one locus at a time and compute the variance of the pseudovalues (which have been extrapolated). They then want to use these to compute the variance of the branch length, where the branch length is obtained from a UPGMA clustering from the distance matrix. They give methods of taking into account the covariances of the distances with each other, using the linearity of the relationship between branch lengths and distances.

For phylogenies inferred by distance methods from sequence data, one need not use all of Mueller & Ayala's formulas. One could proceed more simply by dropping one site at a time, recomputing the distance matrix in each case, estimating the phylogeny from the resulting matrix, and recording the length of the branch of interest. If B is the branch length with all sites in the data and B' the estimate after dropping one site, the pseudovalue for the branch length is

$$S = n\,B - (n\text{-}1)\,B' \qquad (1)$$

which is the result of an n-fold extrapolation of the effect of dropping the one site since it can be rewritten as:

$$S = n\,(B - B') + B' \tag{2}$$

which amounts to an n–fold extrapolation of the effect of adding the site. Note that this is the effect of adding, not dropping the site; the extrapolation is in the opposite direction from the effect of dropping the site. To test the reality of the branch we need to test whether the mean of the pseudovalues is significantly different from zero. This could be done by a t-test or simply by seeing whether 95% of the pseudovalues are positive. Mueller & Ayala use some approximations involving gamma distributions that are special to the case of genetic distances. Their technique has not been applied to sequence data yet, but it is closely related to the bootstrap, which is used in a similar way.

LANYON'S JACKKNIFE Lanyon (88) has presented a completely different jackknife method for use with distance matrix data. Instead of dropping one site at a time, he drops one species at a time. A tree is constructed from the resulting reduced distance matrix. A group found in the original tree is regarded as confirmed if it shows up (with the exception, perhaps, of the species that has been dropped) in all of the resulting trees.

The difficulty with using Lanyon's method is that its statistical properties are completely unknown. The method is an exploratory tool for "distinguishing stable from unstable portions of phylogenetic trees" (88), but it is not a truly statistical method. The reason is that the entities sampled, species, are not independent. Their nonindependence results from evolution, from the existence of a phylogeny, and is the very fact we wish to study. Jackknifes and other resampling techniques usually assume that the data points are independently drawn from some distribution, an assumption that is not valid if species are the units of resampling.

Lanyon does not claim that his method can be used to create a confidence limit or test trees. There is no connection made in his paper between the assessment of whether a group is "stable" or "unstable" and any judgment of its statistical significance. This limits the technique to the status of a nonstatistical exploratory tool.

THE BOOTSTRAP I have (44) applied the bootstrap method of resampling (19, 20, 21) to phylogenies in a way parallel to Mueller and Ayala's use of the jackknife. The bootstrap dictates that we resample the data set by drawing points from it with replacement, until we get a data set of the same size as the original. Usually some points are sampled several times, others left out. The

estimates made from the resampled data set need not be extrapolated in any way. A confidence limit on any quantity can be constructed by the "percentile method" of simply discarding (for example) the upper and lower 2.5% of the distribution of that quantity to obtain a 95% bootstrap confidence limit.

The bootstrap assumes, as does the jackknife, that the points are independent and identically distributed. As with the Mueller and Ayala jackknife, the sites are used as the entities to be resampled, under the assumption that they can be regarded as independently evolved on the same phylogeny. When a site is chosen to be included, it is copied into the resampled data set, keeping the nucleotides associated with the same species. The species are not resampled in any way—all of them are included in the resampled data set.

The other choice, in addition to the method of resampling, is the quantity to be examined. A phylogeny is a complex multivariate entity with many discrete and continuous features, not just a simple number on a scale. It is not at all obvious how to take the cloud of estimates of the phylogeny, one or more for each bootstrap sample data set, and produce from them a confidence interval.

The method I have used is to assume that there is some particular group (set of species) in which we have declared a prior interest. For example, we may wish to know whether the monophyletic group (human, chimpanzee) is on the true phylogeny. We look among all the bootstrap estimates of the tree, and count what fraction this group is monophyletic. In effect, we are interested in a 0-1 variable which indicates the presence or absence of the group. If 95% or more of the trees have the group present, then the 95% bootstrap confidence interval on the 0-1 variable contains only 1's, so that we can declare the group significantly supported.

The easy way to find such groups is to take all the bootstrap estimates and construct a majority-rule consensus tree (94). This is a tree with all those groups that show up in more than half of the bootstrap estimates of the tree. It will therefore contain all groups that occur 95% of the time. The difficulty is that we may then declare all of them significant, tantamount to deciding after the fact which hypotheses we were interested in testing. It leads us to a multiple-tests problem. Among every 20 groups we might examine, one should be declared significant at the 95% level by this procedure, even if none are actually on the true phylogeny. When we examine only groups that have shown up at least once in a bootstrap estimate of the phylogeny, the chance of a spurious significance is even greater.

Thus we must either declare in advance which group we are looking for, or we must apply some correction for multiple tests. The proper multiple-tests correction has not yet been discovered. With n species there are 2^n-1 possible groups, and we may be interested in deciding whether each of them is significantly supported. For that matter, we may also be interested in whether

each of them is significantly opposed. If a group of prior interest is absent from 95% of the trees, it can be declared significantly opposed. Alternatively, we can use the upper and lower 2.5% points of the 0–1 variable, indicating presence or absence of the group to test in such a way that we can detect either the significant presence or the significant absence of a group. Again, this is subject to the multiple-tests problem if the group is not decided on in advance.

THE DELETE-HALF JACKKNIFE A resampling method similar to the bootstrap is to take a random half of the sites. This is a kind of jackknife in which, instead of dropping one character, we drop half of them. It is one of a family of jackknife methods advocated by Wu (143) for use in regression problems and has the advantage over other jackknife methods of not using any extrapolation factor. If one parameter is being estimated, there is no extrapolation at all—the variation between estimates from random halves of the data should be typical of the sampling variation of the estimate. The matter of how many parameters are actually being estimated is a complex one. If k parameters are being estimated, we are supposed to choose samples of size $(n + k - 1)/2$ to avoid extrapolation (143). However, if n is large, samples of size $n/2$ will be close to the correct size for any modest value of k.

I had also pointed out (44), much more crudely, that a jackknife with random halves would have this property. Like the bootstrap, it can be used to construct a confidence interval by the percentile method. The only investigation yet is Penny & Hendy's (110) empirical study using an actual data set, in which the delete-half jackknife (which they call "halflings" or the "method of Hobbits") shows about the same performance as the jackknife. It would be interesting to know under what conditions the one method is to be preferred.

PENNY & HENDY'S RESAMPLING METHOD Penny & Hendy (110) have used resampling methods, jackknifes and bootstraps, to show for a given data set how many characters would be needed to have the estimate accurately reflect the true tree. In the six protein molecules they used in different mammalian orders, there were 166 reconstructed "phylogenetically informative" nucleotide sites. They have resampled subsets of sites of various sizes, including bootstrap samples and jackknifes that delete various numbers of sites. They estimated trees (by parsimony or various kinds of weighted parsimony) for each resampled data set, without engaging in any extrapolation. They used a distance measure between trees, the partition metric, which measures the number of subsets of species that are different between the two trees. For the jackknifes they took nonoverlapping subsets of various sizes (up to half the sites) and measured the difference between the resulting trees. Extrapolating the results, they could show that about twice as many sites would have to be in the analysis for it to be reasonably likely that the most

parsimonious trees from different subsets be identical. One of the main objectives of their paper was to test various methods for weighting sites. They found that when positions were given less weight when they conflicted with others, the results were more reliable.

In effect Penny & Hendy were using the resampled data sets as an experiment on the statistical variability of resampling methods and to see how accuracy of estimation of trees was related to the number of informative sites. Their results give us a feel for the sizes of confidence intervals to be expected from different amounts of sequence, although they did not indicate how to extrapolate them to other groups of species, nor did they actually present a confidence interval on their estimate of the tree.

INTERVALS BASED ON PAIRED COMPARISONS OF SITES Templeton's (138) paired sites test has already been described. One might wonder whether it could be used to construct a confidence interval. Could one take all trees that fail to be significantly worse than the most parsimonious one, and call those a confidence interval? Since these tests are of different hypotheses, it is not obvious how to correct for the multiplicity of tests or the fact that some of them are of closely related hypotheses (if we reject tree T from the confidence interval, this is nearly the same test as the one that examines a closely similar tree). It seems that the naive procedure of taking all trees that do not fail the pairwise test could not possibly be valid. And yet there is some indication that it may be.

H. Kishino & M. Hasegawa (in preparation) have presented a variant on Templeton's test that uses likelihoods. They examine differences, site by site, between log likelihoods. They are then able to construct, from a Bayesian approach, an argument that an interval containing 95% of the posterior probability is found by taking, in effect, all trees that are not rejected compared to the maximum likelihood tree. This is not quite the same thing as a confidence interval, but it is related to it. They note that the same argument would apply to parsimony, using the Templeton test.

The difficulty with this method of constructing a confidence interval, apart from the question of whether it really is a confidence interval, is that one must examine trees one at a time to see whether they are rejected from the interval. This is a large computational task, although competing methods such as bootstraps are also computationally intensive.

SIMULATION STUDIES

Closely related to resampling is simulation. In fact, one bootstrap method, the "parametric bootstrap" (22) consists simply of taking the best estimate of the tree, simulating new data sets of the same size by evolution occurring along

that tree under the postulated model, and then using the variability among estimated trees from those simulated data sets to assess how much variability there was in the original estimate. This is one of the best uses of simulation, and should be done more frequently. It is not often done, partly because we may doubt that we have an accurate enough picture of the stochastic processes which bring about the data, but mostly because the potential users are intimidated by the complexities of computer simulation.

Most computer simulation studies have a different aim. They use simulation to test competing methods to see which does a better job of estimating the true tree. The problem common to such studies is uncertainty whether the results will continue to apply when the true tree is of a different shape or depth in time, the data set of a different size, or the model of evolutionary change different. Nevertheless, there is enough consistency among results to make comparisons interesting.

Many of the simulations carried out so far have been for gene frequency data rather than sequence data. These include those of Kidd & Cavalli-Sforza (78), Kidd et al (79) and Astolfi et al (1), Nei et al (101), Rohlf & Wooten (112) and Kim & Burgman (80). Others such as Fiala & Sokal (50) and Sokal (130, 131, 132) have modelled discrete morphological characters. Many of the patterns found in simulations of molecular sequences also are found in these simulations.

Simulations modelling molecular sequence data include the papers of Peacock & Boulter (109), Tateno et al (135), Blanken et al (5), Hasegawa & Yano (66), Tateno & Tajima (136a) Li et al (93), Sourdis & Krimbas (133), Sourdis & Nei (134), and Saitou (117). It is hard to come away from a reading of these papers with a clear overall consensus as to whether distance matrix or parsimony methods are better (none tested likelihood methods against these other two kinds). Peacock & Boulter (109) suggested that parsimony was better when sequences were little diverged, distance matrix methods better when divergence was more ancient. Sourdis & Nei (134) found a similar pattern. Blanken et al (5) found little difference between these methods. The gene frequency simulations cannot be compared readily to the nucleotide sequence simulations for this purpose. The simulations of discrete morphological characters can. Only that of Sokal (130), in which the organisms were not actually simulated but were "evolved" on paper by a biologist, can be directly compared: it had a moderate degree of divergence (as judged by changes per character) and was better estimated by parsimony than by the UPGMA distance matrix method.

There is one pattern, predicted by theory, that is found to be fairly clear in the simulations, though rarely commented on by the authors. This is the inconsistency of UPGMA clustering methods when rates of evolution in different lineages depart sufficiently from a clock. As noted above in the

discussion of distance matrix methods, UPGMA (average linkage) or other clustering methods can be inconsistent when we are not close to a clock. Examination of the conditions for this to happen in simple cases (42) show that UPGMA will fail considerably more readily than will parsimony, requiring only differences in rate between lineages of about a factor of two before it is expected to misbehave. Parsimony will be inconsistent when rates vary more extremely (although it will also fail, as Hendy & Penny (69) have discovered, in some clocklike cases where clustering methods will not). Distance matrix methods in which distances are properly transformed to reflect estimated divergence times and maximum likelihood methods are among those not expected to show the inconsistency.

Some of the simulation studies were conducted with a molecular clock assumed, others without. For this purpose we can compare the gene frequency, discrete morphological character, and nucleic acid simulations—any one in which a clustering method was compared to either a parsimony or a distance method. UPGMA was found to perform well in the clocklike simulations of Tateno et al (135), Nei et al (101), and Tateno & Tajima (136a). Fiala & Sokal (50) had an intermediate degree of rate variation, and found clustering to have mediocre performance. The studies in which rates varied considerably from a clock found clustering to perform badly, as expected. These include the studies of Blanken et al (5), Sokal (130), Sourdis & Krimbas (133), and Kim & Burgman (80).

The failure of parsimony to be consistent when rates vary is also seen when looked for. Hasegawa & Yano (66) used simulation to check the analytic results on inconsistency of parsimony for four species, and they found that parsimony did in fact fail in cases when likelihood did not. Kim & Burgman (80) carried out a similar test for gene frequency data and again clearly confirmed the expected misbehavior of parsimony methods. What is less clear is whether the weakness of parsimony methods when compared to distance matrix methods, for example in the studies of Sourdis & Nei (134), is a consequence of the inconsistency of parsimony. When cases with fewer or more sites are compared, one gets the impression that adding sites is helping parsimony methods less than it helps distance methods in identifying the correct tree.

Maximum likelihood and Lake's method of invariants have as yet received fewer simulation tests than the parsimony or distance methods. As mentioned above, Kim & Burgman (80) and Hasegawa & Yano (66) found it to converge on the correct tree when parsimony methods did not. Astolfi et al (1) found maximum likelihood to perform only moderately well for their gene frequency simulations. Rohlf & Wooten (112) found likelihood to become better than other methods when the number of loci simulated was large. Saitou (117) found that maximum likelihood with sequence data (for a small number of

species, with or without a molecular clock) had a lower probability of finding the correct tree than did a properly transformed distance matrix method. Since for sufficiently large amounts of data maximum likelihood should be more efficient than other methods, it will be interesting to see whether this low efficiency for intermediate amounts of data can be confirmed as a general property.

Li et al (93) found that Lake's method of invariants continued to select the correct tree when inequalities of branch lengths caused both parsimony methods and Saitou & Nei's (116) "neighbor joining" distance matrix method to be misled by inequalities of rates. (The distance matrix method was misled because the distances used were not transformed to correct for multiple changes). One would expect maximum likelihood methods to share this good behavior, and also distance matrix methods in which the distances were properly transformed.

AN OVERVIEW

This survey of methods for inferring phylogenies and assessing their reliability shows that the field is in an incomplete but interesting state. We have a number of different approaches: parsimony, distance matrix methods, and likelihood methods. The assumptions inherent in these methods are only sketchily known—we have hints but little in the way of comprehensive proofs that particular assumptions are required. It is clear from the failings of different methods in particular cases that they all have assumptions; no method allows one to make inferences about evolutionary patterns in a well-justified way without making any assumptions about evolutionary processes.

When it comes to assessing the reliability of the estimated phylogenies, we have only fragments of methods, each with many properties unknown. Parsimony methods can be inconsistent under a relatively unknown set of circumstances, of which we have only some hints. Distance matrix methods assume that we know how to transform the distances so that branch lengths are additive in expected distance. Maximum likelihood methods require specification of the probabilistic model of evolution, and it is not known how sensitive they might be to violations of the model, and how likelihood ratio tests can be performed to distinguish among tree topologies.

In the last few years a variety of quasi-empirical methods have been proposed for assessing the reliability of phylogenies, such as the jackknife, the bootstrap, and Templeton's pairwise test. Simulation methods are also available for the energetic. However, we have only the faintest notion of how well-behaved and how powerful these tests are.

FUTURE DIRECTIONS

While much remains to be done to complete the above picture, there are also related problems that have received only limited exploration. Two of these deserve particular attention. One is the integration of sequence alignment with phylogenetic inference, the other the integration of population-level processes with phylogenetic inference.

We have already mentioned the first of these, for which approaches have been pioneered by Sankoff et al (118), Sankoff (119), and Sankoff & Rousseau (120). Feng & Doolittle (49) have also recognized the need for a close relationship between sequence alignment and phylogenies. Of particular interest is the attempt by Bishop & Thompson (4) to place sequence alignment and phylogeny estimation both in a likelihood framework. In their case there were only two sequences, so that the phylogeny estimation reduced simply to estimation of the time of divergence between the sequences. In principle the approach of Sankoff & Cedergren (121) could be carried out using likelihood instead of parsimony, although the computational problems would be extreme. These computational problems have tended to divert attention from this approach. Even if it can never be made practical, it is important to consider, if only to gain perspective on what a complete integration of phylogeny estimation with alignment would look like. Overconcentration on practicality of methods has probably resulted in underestimation of the importance of these papers.

The second problem requires some further explanation. When we infer a tree by consideration of the sequence of one molecule, we are estimating the genealogy of the particular copies of the molecule that were sequenced (see the discussion by Nei (104) who calls this a gene tree). When the time scale is fairly long there will usually be no discrepancy between the genealogies of molecules and phylogenies of populations. It does not matter which individual crow, alligator, or mouse we choose—if we sequence cytochromes from any individual in each of three species, the genealogy of the genes should have the crow and alligator copies more recently descended from a common ancestor than either is from the mouse. One would gain little in the inference by sequencing other copies of this gene from any of these three populations.

When we work closer to the population level, matters become different. As Gillespie & Langley (58), Tajima (136), and Hudson (73) have emphasized, genes from different species, if traced backwards in time, both have ancestor copies in the population at the moment of speciation. But those copies are most likely not the same. We must trace back a further period of time, of average length $2N_e$ generations, before we find that these copies have

a common ancestor. N_e is the effective population size (for mitochondrial genomes the figure is instead N_{ef}, the effective number of females).

The result is that the genealogy of gene copies may actually differ from the phylogeny of the species. If we take my mitochondrion, yours, and a chimpanzee's, it may well be that when human and chimpanzee speciated, all three mitochondrial lineages were distinct. As we go back before that any two of the three might equally likely be the first to join—say mine and the chimpanzee's. Avise et al (2) have reviewed the problems that this "lineage sorting" creates when inferring geographic structure of species from mitochondrial genealogies.

There are two ways to correct for this random perturbation of the trees. One is to use more molecules, particularly those not closely linked to each other. One of my hemoglobin-β genes, one of yours, and a chimpanzee's may show a different genealogy than do the mitochondria. By collecting sequences from many loci, we should find that my gene copy and yours are more frequently sister lineages on the genealogy of gene copies than either of us is to the chimpanzee, thus indicating that our ancestors were in the same population for a time after the chimpanzee lineage split off. At the moment we totally lack a quantitative methodology for analyzing data like this—for reconstructing the phylogeny of populations, given a collection of sequences in which different loci are represented. Our interest is not in the genealogy of gene copies itself, but in the pattern of relatedness of populations. The relevant methods have not been developed mostly because this kind of data has only recently begun to be collected.

The other method of inferring the population phylogeny is to take multiple samples of the same locus from each population. This, in effect, is what Cann et al (7) did with mitochondrial DNA. They were able to see patterns in the genealogy of gene copies that suggested past population-level events such as a postulated bottleneck in the human species as it spread from Africa to the rest of the world. However, the phylogenetic methods they used were only able to estimate the genealogy of gene copies; the inferences about populations were made by informal and intuitive methods, there being no methods available for making them numerically. This is unfortunate—one would like to be able to make statistical statements about the reality and timing of the inferred bottleneck. Here again, there is a serious need for the development of methods, a need that has not been addressed mostly because the relevant data is only now being collected.

One can imagine how the inference could be done if practical computational considerations were not a barrier. Suppose that we want to evaluate a phylogeny of populations, where that phylogeny specifies not only times of splitting of populations, but effective population sizes as well. We have a series of loci, assumed unlinked, and for each a population sample of se-

quences. One can, in principle, use the phylogeny, T, to place a prior distribution on each possible genealogy of gene copies, G (where G specifies the exact times of splitting of lineages). Given the genealogy of gene copies, one can compute the likelihood of the data D, which is Prob(D;G), the probability of D given G. The overall likelihood is Prob(D;T), the probability of the data given the phylogeny. This is the weighted sum over all possible genealogies of gene copies that could be generated by that phylogeny:

$$\text{Prob}(D;\ T) = \sum_{G} \text{Prob}(G;\ T) \quad \text{Prob}(D;\ G), \tag{3}$$

the summation running over all possible genealogies of gene copies. The probabilities of the genealogies under the given phylogeny is obtained by consideration of the mathematics of genetic drift—the process in each population is the "coalescent" process of Kingman (83, 84), for a review of which see Tavare (137).

This idealized approach is not practical. The number of possible genealogies over which the likelihood must be summed is so great that there will have to be some breakthrough for it to be used. We would need either a major algebraic simplification, a major advance in computational methods, or an approximation that enabled much of the computation to be avoided. Nevertheless, the above formulation is important in giving us a clear picture of the inference problem and the most general form of solving it. Likelihood formulations frequently have this benefit even when they cannot be used in practice. Padmadisastra (108) has made calculations relevant to the parallel problem of inferring phylogenies from models of neutral alleles. The density of the mathematics in that paper, which treats the case of three populations, will be some indication of the long road ahead in the sequences case.

Of course, I have been assuming that there is a phylogeny of populations. When the populations are members of different species, this is uncontroversial. But when they are drawn from the same species, it is far from obvious that the genealogy of the populations is treelike. Migration creates loops in the tree, and in the extreme the genealogy no longer looks at all treelike, but instead takes on the form of the migration pattern among populations. I have outlined elsewhere (40) the difficult and unsolved inference problems that arise when we try to use gene frequency data to distinguish between treelike historical patterns of branching and nearly treelike patterns of migration among populations. In the years since that paper was published scarcely any advances have occurred in our understanding of this problem.

When we use sequence data the problem is at least as complex, but the data may have more power to discriminate among patterns of migration and

inheritance than do gene frequencies. I hope that there will be some attempts to address this problem before we are overwhelmed by data.

Another complication that arises with intraspecific data is recombination. When two loci are unlinked, one may approximate their behavior by saying that the genealogies of gene copies at the two loci are completely independently drawn from the set of possible genealogies within that population. When two sites are linked tightly, they follow the same genealogy. When the linkage is incomplete, they may follow the same genealogy in part, and in part different ones. The problems that this causes for inferring the genealogy have only begun to be addressed. It is naive to think that by constructing genealogies for different parts of a molecule we will simply be able to see all recombination events. Hudson & Kaplan (74) have discussed problems of inferring the number of recombination events in the ancestry of a sample, and they find that many recombination events will leave no trace.

In spite of all these difficulties, sequence samples of multiple loci from populations provide us with the most powerful data sets for looking at events in the past, giving us ways of detecting hybridization between species, and possibly even allowing us to see events in the speciation process itself. Whether this prospect can be realized will depend on whether the appropriate methods of analysis can be developed. The sequence data is beginning to pour in. It is just a matter of taking seriously the task of analyzing it.

ACKNOWLEDGMENT

This work was supported by National Science Foundation grant BSR-8614807. I wish to thank Elizabeth Thompson, Masami Hasegawa, David Penny, and Alan Templeton for discussions or correspondence on the subject matter of this review.

Literature Cited

1. Astolfi, P., Kidd, K. K., Cavalli-Sforza, L. L. 1981. A comparison of methods for reconstructing evolutionary trees. *Syst. Zool.* 30:156–69
2. Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., et al. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Ann. Rev. Ecol. Syst.* 18:489–522
3. Barry, D., Hartigan, J. A. 1987. Statistical analysis of hominoid molecular evolution. *Stat. Sci.* 2:191–210

3a. Bishop, M. J., Friday, A. E. 1985. Evolutionary trees from nucleic acid and protein sequences. *Proc. Roy. Soc. London* B226:271–302

4. Bishop, M. J., Thompson, E. A. 1986. Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* 190:159–65
5. Blanken, R. L., Klotz, L. C., Hinnebusch, A. G. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. *J. Mol. Evol.* 19:9–19
6. Camin, J. H., Sokal, R. R. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19: 311–26
7. Cann, R. L., Stoneking, M., Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–6

7a. Carpenter, J. M. 1987. Cladistics of cladists. *Cladistics* 3:363–75

8. Cavalli-Sforza, L. L., Edwards, A. W. F. 1967. Phylogenetic analysis: models

and estimation procedures. *Evolution* 32:550–70 (also published in *Am. J. Hum. Genet.* 19:233–57)
9. Cavender, J. A. 1978. Taxonomy with confidence. *Math. Biosci.* 40:271–80 (crratum 44:308)
10. Cavender, J. A. 1981. Tests of phylogenetic alternatives under generalized models. *Math. Biosci.* 54:217–29
11. Cavender, J. A., Felsenstein, J. 1987. Invariants of phylogenies in a simple case with discrete states. *J. Class.* 4:57–71
12. Chakraborty, R. 1977. Estimation of time of divergence from phylogenetic studies. *Can. J. Genet. Cytol.* 19:217–23
13. Colless, D. H. 1970. The phenogram as an estimate of phylogeny. *Syst. Zool.* 19:352–62
14. Dayhoff, M. O., Eck, R. V. 1968. *Atlas of Protein Sequence and Structure 1967–1968,* pp. 307. Silver Spring, Maryland: Natl. Biomed. Res. Found.
15. DeBry, R. W., Slade, N. A. 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. *Syst. Zool.* 34:21–34
16. Eck, R. V., Dayhoff, M. O. 1966. *Atlas of Protein Sequence and Structure 1966.* Silver Spring, Maryland: Natl. Biomed. Res. Found.
17. Edwards, A. W. F., Cavalli-Sforza, L. L. 1963. The reconstruction of evolution. *Ann. Hum. Genet.* 27:105 (also published in *Heredity* 18:553)
18. Edwards, A. W. F., Cavalli-Sforza, L. L. 1964. Reconstruction of evolutionary trees. In *Phenetic and Phylogenetic Classification,* ed. V. H. Heywood, J. McNeill, pp. 67–76. London: Systematics Assoc. Publ. No. 6
19. Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7:1–26
20. Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMS-NSF Reg. Conf. Ser. Appl. Math. No. 38. Philadelphia: Soc. Indust. & Appl. Math.
21. Efron, B., Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.* 37:36–48
22. Efron, B. 1985. Bootstrap confidence intervals for a class of parametric problems. *Biometrika* 72:45–58
23. Estabrook, G. F., Landrum, L. 1976. A simple test for the possible simultaneous evolutionary divergence of two aminoacid positions. *Taxon* 24:609–13
24. Estabrook, G. F., Johnson, C. S. Jr., McMorris, F. R. 1976. An algebraic analysis of cladistic characters. *Discrete Math.* 16:141–47
25. Estabrook, G. F., Johnson, C. S. Jr., McMorris, F. R. 1976. A mathematical foundation for the analysis of cladistic character compatibility. *Math. Biosci.* 29:181–87
26. Estabrook, G. F., McMorris, F. R. 1980. When is one estimate of evolutionary relationships a refinement of another? *J. Math. Biol.* 10:367–73
27. Farris, J. S. 1969. On the cophenetic correlation coefficient. *Syst. Zool.* 18: 279–85
27a. Farris, J. S. 1969. A successive approximations approach to character weighting. *Syst. Zool.* 18:374–85
28. Farris, J. S. 1970. Methods for computing Wagner trees. *Syst. Zool.* 19:83–92
29. Farris, J. S. 1971. The hypothesis of nonspecificity and taxonomic congruence. *Ann. Rev. Ecol. Syst.* 2:277–302
30. Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. *Am. Nat.* 106:645–68
31. Farris J. S. 1981. Distance data in phylogenetic analysis. In *Advances in Cladistics. Proc. 1st Meet. Willi Hennig Soc.* ed. V. A. Funk, D. R. Brooks, pp. 3–23. New York: NY Bot. Gard. 250 pp.
32. Farris, J. S. 1983. The logical basis of phylogenetic analysis. In *Advances in Cladistics, Vol. 2: Proc. 2nd Meet. Willi Hennig Soc.* ed. N. I. Platnick, V. A. Funk, pp. 7–36. New York: Columbia Univ. Press. 218 pp.
33. Farris, J. S. 1985. Distance data revisited. *Cladistics* 1:67–85
34. Farris, J. S. 1986. Distances and statistics. *Cladistics* 2:144–57
35. Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* 22:240–49
36. Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401–10 (Also reprinted in 1984 *Conceptual Issues in Evolutionary Biology, An Anthology,* ed. E. Sober, pp. 663–74. Cambridge, Mass.: MIT Press. 725 pp.)
37. Felsenstein, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. *Syst. Zool.* 28:49–62
38. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–76

39. Felsenstein, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol. J. Linnean Soc.* 16: 183–96
40. Felsenstein, J. 1982. How can we infer geography and history from gene frequencies? *J. Theor. Biol.* 96:9–20
41. Felsenstein, J. 1983. Inferring evolutionary trees from DNA sequences. *Statistical Analysis of DNA Sequence Data,* ed. B. S. Weir, pp. 133–150. New York: Dekker
42. Felsenstein, J. 1983. Parsimony in systematics: biological and statistical issues. *Ann. Rev. Ecol. Syst.* 14:313–33
43. Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* 38:16–24
44. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–91
45. Felsenstein, J. 1985. Confidence limits on phylogenies with a molecular clock. *Syst. Zool.* 34:152–61
46. Felsenstein, J. 1986. Distance methods: reply to Farris. *Cladistics* 2:130–43
47. Felsenstein, J., Sober, E. 1986. Parsimony and likelihood: an exchange. *Syst. Zool.* 35:617–26
48. Felsenstein, J. 1987. Estimation of hominoid phylogeny from a DNA hybridization data set. *J. Mol. Evol.* 26:123–31
49. Feng, D-F., Doolittle, R. F. 1987. Progressive sequence alignment as prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351–60
50. Fiala, K. L., Sokal, R. R. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. *Evolution* 39:609–22
51. Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., et al 1988. Molecular phylogeny of the animal kingdom. *Science* 239:748–52
52. Fitch, W. M., Margoliash, E. 1967. Construction of phylogenetic trees. *Science* 155:279–84
53. Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.* 20:406–16
54. Fitch, W. M. 1974. Response to the paper of Dr. Moore. In *Genetic Distance,* ed. J. F. Crow, C. Denniston, p. 117–19. New York: Plenum 195 pp.
55. Fitch, W. M., Farris, J. S. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. *J. Mol. Evol.* 3:263–78
56. Fitch, W. M. 1975. Toward finding the tree of maximum parsimony. In *Proc. Eighth Int. Conf. on Numerical Taxonomy,* ed. G. F. Estabrook, pp. 189–230. San Francisco: W. H. Freeman. 429 pp.
57. Fitch, W. M. 1986. Commentary (on papers by Ruvolo & Smith, Saitou, and Templeton). *Mol. Biol. Evol.* 3:296–98
58. Gillespie. J. H., Langley, C. H. 1979. Are evolutionary rates really variable? *J. Mol. Evol.* 13:27–34
59. Gillespie, J. H. 1982. A randomized SAS-CFF model of natural selection in a random environment. *Theoret. Popul. Biol.* 21:219–37
60. Gillespie, J. H. 1984. Molecular evolution over the adaptive landscape. *Evolution* 38:1116–29
61. Gillespie, J. H. 1984. The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA* 81:8009–13
62. Gillespie, J. H. 1986. Variability of evolutionary rates of DNA. *Genetics* 113: 1077–91
63. Gillespie, J. H. 1986. Natural selection and the molecular clock. *Mol. Biol. Evol.* 3:138–55
64. Hasegawa, M., Yano, T., Kishino, H. 1984. A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proc. Jpn Acad.* 60B:95–8
65. Hasegawa, M., Yano, T. 1984. Phylogeny and classification of Hominoidea as inferred from DNA sequence data. *Proc. Jpn Acad.* 60B:389–92
66. Hasegawa, M., Yano, T. 1984. Maximum likelihood method of phylogenetic inference from DNA sequence data. *Bull. Biometric Soc. Japan* 5:1–7
67. Hasegawa, M., Kishino, H., Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitchondrial DNA. *J. Mol. Evol.* 22:160–74
68. Hasegawa, M., Iida, Y., Yano, T., Takaiwa, F. Iwabuchi, M. 1985. Phylogenetic relationships among eukaryotic kingdoms inferred by ribosomal RNA sequences. *J. Mol. Evol.* 22:32–8
69. Hendy, M. D., Penny, D. 1988. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* In press
70. Hogeweg, P., Hesper, P. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J. Mol. Evol.* 20:175–86
71. Holmquist, R., Miyamoto, M. M., Goodman, M. 1988. Higher-primate phylogeny—why can't we decide? *Mol. Biol. Evol.* 5:201–16
72. Holmquist, R., Miyamoto, M. M., Goodman, M. 1988. Analysis of higher-primate phylogeny from transversion differences in nuclear and mitochondrial DNA by Lake's methods of evolutionary

parsimony and operator metrics. *Mol. Biol. Evol.* 5:217–236
73. Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–17
74. Hudson, R. R., Kaplan, N. L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111:147–64
75. Jukes, T. H., Cantor, C. R. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism,* ed. H. N. Munro, pp. 21–132. New York: Academic
76. Kaplan, N., Langley, C. H. 1979. A new estimate of sequence divergence of DNA using restriction endonuclease mappings. *J. Mol. Evol.* 13:295–304
77. Kashyap, R. L., Subas, S. 1974. Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *J. Theor. Biol.* 47:75–101
78. Kidd, K. K., Cavalli-Sforza, L. L. 1971. Number of characters examined and error in reconstruction of evolutionary trees. In *Mathematics in the Archaeological and Historical Sciences,* ed. F. R. Hodson, P. Tautu, pp. 335–46. Edinburgh: Edinburgh Univ. Press.
79. Kidd, K. K., Astolfi, P., Cavalli-Sforza, L. L. 1974. Error in the reconstruction of evolutionary trees. In *Genetic Distance,* ed. J. F. Crow, C. Denniston, pp. 121–36. New York: Plenum. 195 pp.
80. Kim, J., Burgman, M. A. 1988. Accuracy of phylogenetic-estimation methods using simulated allele-frequency data. *Evolution* 42:596–602
81. Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–26
82. Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge Univ. Press.
83. Kingman, J. F. C. 1982. On the genealogy of large populations. *Essays in Statistical Science,* ed. J. Gani, E. J. Hannan, pp. 27–43. London: Appl. Trust.
84. Kingman, J. F. C. 1982. The coalescent. *Stoch. Process. Appl.* 13:235–48
85. Kluge, A. R., Farris, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18:1–32
86. Lake, J. A. 1987. Determining evolutionary distances from highly diverged nucleic acid sequences: operator metrics. *J. Mol. Evol.* 26:59–73
87. Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: operator metrics. *Mol. Biol. Evol.* 4:167–91
88. Lanyon, S. 1985. Detecting internal inconsistencies in distance data. *Syst. Zool.* 34:397–403
89. Le Quesne, W. J. 1974. The uniquely evolved character concept and its cladistic application. *Syst. Zool.* 23:513–17
90. Lewin, R. 1987. My cousin the chimpanzee. *Science* 238:273–74
91. Li, W -H. 1981. Simple method for constructing phylogenetic trees from distance matrices. *Proc. Natl. Acad. Sci. USA* 78:1085–89
92. Li, W -H. 1986. Evolutionary change of restriction cleavage sites and phylogenetic inference. *Genetics* 113:187–213
93. Li, W -H., Wolfe, K. H., Sourdis, J., Sharp, P. M. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. *Cold Spring Harbor Symp. Quant. Biol.* 52:847–56
94. Margush, T., McMorris, F. R. 1981. Consensus n-trees. *Bull. Math. Biol.* 43: 239–44.
95. Miyamoto, M. M., Slightom, J. L., Goodman, M. 1987. Phylogenetic relations of humans and african apes from DNA sequences of the $\psi\eta$-globin region. *Science* 238:369–73
96. Moore, G. W. 1971. A mathematical model for the construction of cladograms. PhD thesis. North Carolina State Univ. Raleigh.
97. Moore, G. W., Barnabas J., Goodman, M. 1973. A method for constructing maximum parsimony ancestral amino acid sequences on a given network. *J. Theor. Biol.* 38:459–85
98. Moore, G. W. 1974. A counterexample to Fitch's method for maximum parsimony trees. *Genetic Distance*, ed. J. F. Crow, C. Denniston, pp. 105–116. New York: Plenum. 195 pp.
99. Moore, G. W. 1977. Proof of the populous path algorithm for missing mutations in parsimony trees. *J. Theor. Biol.* 66:95–106
100. Mueller, L. D., Ayala, F. J. 1982. Estimation and interpretation of genetic distance in empirical studies. *Genet. Res.* 40:127–37
101. Nei, M., Tajima, F., Tateno, Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.* 19:153–70
102. Nei, M., Tajima, F. 1985. Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. *Mol. Biol. Evol.* 2:189–205
103. Nei, M., Stephens, J. C., Saitou, N. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to

molecular data from humans and apes. *Mol. Biol. Evol.* 2:66–85
104. Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia Univ. Press. 512 pp.
105. Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, ed. S. S. Gupta, J. Yackel, pp. 1–27. New York: Academic.
106. Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harb. Symp. Quant. Biol.* 52:825–37
107. Olsen, G. J. 1988. Phylogenetic analysis using ribosomal RNA. *Methods in Enzymol.* In press
108. Padmadisastra, S. 1987. The genetic divergence of three populations. *Theor. Pop. Biol.* 32:347–65
109. Peacock, D., Boulter, D. 1975. Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. *J. Mol. Biol.* 95:513–27
110. Penny, D., Hendy, M. 1986. Estimating the reliability of evolutionary trees. *Mol. Biol. Evol.* 3:403–17
111. Rohlf, F. J., Sokal, R. R. 1981. Comparing numerical taxonomic studies. *Syst. Zool.* 30:459–90
112. Rohlf, F. J., Wooten, M. C. 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. *Evolution* 42:581–95
113. Ritland, K., Clegg, M. T. 1987. Evolutionary analysis of plant DNA sequences. *Am. Nat.* 130:S74–S100
114. Ruvolo, M., Smith, T. F. 1986. Phylogeny and DNA-DNA hybridization. *Mol. Biol. Evol.* 3:285–89
115. Saitou, N. 1986. On the delta Q-test of Templeton. *Mol. Biol. Evol.* 3:282–84
116. Saitou, N., Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–25
117. Saitou, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. *J. Mol. Evol.* 27:261–73
118. Sankoff, D. D., Morel, C., Cedergren, R. J. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. *Nature New Biol.* 245:232–34
119. Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28:35–42
120. Sankoff, D. D. Rousseau, P. 1975. Locating the vertices of a Steiner tree in arbitrary space. *Math. Progr.* 9:240–46
121. Sankoff, D. D., Cedergren, R. J. 1983. Simultaneous comparison of three or more sequences related by a tree. In *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, ed. D. Sankoff, J. B. Kruskal, pp. 253–63. Reading, Mass: Addison-Wesley. 382 pp.
122. Sarich, V. M., Wilson, A. C. 1967a. Rates of albumin evolution in primates. *Proc. Natl. Acad. Sci. USA* 58:142–48
123. Sarich, V. M., Wilson, A. C. 1967b. Immunological time scale for hominoid evolution. *Science* 158:1200–03
124. Smouse, P. E., Li, W.-H. 1987. Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution* 41:1162–76
125. Sneath, P. H. A. 1986. Estimating uncertainty in evolutionary triads from Manhattan-distance triads. *Syst. Zool.* 35:470–88
126. Sober, E. 1983. Parsimony in systematics: philosophical issues. *Ann. Rev. Ecol. Syst.* 14:335–57
127. Sober, E. 1985. A likelihood justification of parsimony. *Cladistics* 1:209–33
128. Sober, E. 1987. Parsimony, likelihood, and the principle of the common cause. *Philos. Sci.* 54:465–69
129. Sokal, R. R., Sneath, P. H. A. 1963. *Numerical Taxonomy*. San Francisco: Freeman. 359 pp.
130. Sokal, R. R. 1983. A phylogenetic analysis of the Caminalcules. II. Estimating the true cladogram. *Syst. Zool.* 32:185–201
131. Sokal, R. R. 1983. A phylogenetic analysis of the Caminalcules. III. Fossils and classification. *Syst. Zool.* 32:248–58
132. Sokal, R. R. 1983. A phylogenetic analysis of the Caminalcules. IV. Congruence and character stability. *Syst. Zool.* 32:248–58
133. Sourdis, J., Krimbas, C. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.* 4:159–68
134. Sourdis, J., Nei, M. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.* 5:298–311
135. Tateno, Y., Nei, M., Tajima, F. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related trees. *J. Mol. Evol.* 18:387–404
136. Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105:437–60

136a. Tateno, Y., Tajima, F. 1986. Statistical properties of molecular tree construction methods under the neutral mutation model. *J. Mol. Evol.* 23:354–61
137. Tavare, S. 1984. Line-of-descent and genealogical processes, and their application in population genetics models. *Theor. Pop. Biol.* 26:119–64
138. Templeton, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37:221–44
139. Templeton, A. R. 1985. The phylogeny of the hominoid primates: a statistical analysis of the DNA-DNA hybridization data. *Mol. Biol. Evol.* 2:420–33
140. Templeton, A. R. 1986. Further comments on the statistical analysis of DNA-DNA hybridization data. *Mol. Biol. Evol.* 3:290–95
141. Wiley, E. O. 1975. Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Syst. Zool.* 24:233–43
142. Wiley, E. O. 1981. *Phylogenetic Systematics. The Theory and Practice of Phylogenetic Systematics*. New York: John Wiley. 439 pp.
143. Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling plans in regression analysis. *Ann. Statist.* 14:1261–95
144. Zuckerkandl, E., Pauling, L. 1962. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry*, ed. M. Marsha, B. Pullman, pp. 189–225. New York: Academic.