Syst. Biol. 44(1):3–16, 1995

APPROACHES FOR ASSESSING PHYLOGENETIC ACCURACY

DAVID M. HILLIS

Department of Zoology, University of Texas, Austin, Texas 78712, USA[1]

Abstract.—Accuracy of phylogenetic methods may be assessed in terms of consistency, efficiency, and robustness. Four principal methods have been used for assessing phylogenetic accuracy: simulation, known phylogenies, statistical analyses, and congruence studies. Simulation studies are useful for studying accuracy of methods under idealized conditions and can be used to make general predictions about the behavior of methods if the limitations of the models are taken into account. Studies of known phylogenies can be used to test predictions from simulation studies, thus providing a check on the robustness of the models (and possibly suggesting refinements for future simulations). Statistical analyses allow general predictions to be applied to specific results, facilitate assessments as to whether or not sufficient data have been collected to formulate a robust conclusion, and indicate whether a given data set is any more structured than random noise. Finally, congruence studies of multiple data sets can be used to assess the degree to which independent results agree and thus the minimum proportion of the findings that can be attributed to an underlying phylogeny. These different methods of assessing phylogenetic accuracy are largely complementary, and the results are consistent in identifying a large class of problems that are amenable to phylogenetic reconstruction. [Phylogeny; accuracy; simulations; experimental evolution; statistics; congruence; consistency; efficiency; robustness.]

Phylogenetic analyses have become commonplace throughout the biological disciplines during the past few decades. This increased emphasis on evolutionary history is a direct result of the realization of the importance of understanding phylogenetic background as a prerequisite to interpreting virtually any biological system in a comparative context. However, the increased utilization of phylogenetic approaches has been driven at least as much by technological and methodological advances as by conceptual advances. In particular, advances in algorithm development, computer technology, and molecular biology have made phylogenetic analyses feasible for almost any problem involving biological lineages, from viral epidemics in extant human populations (e.g., Ou et al., 1992) to the origins of the earliest lineages of life (e.g., Olsen, 1987). Phylogenetic applications depend on accurate reconstructions of phylogenetic trees; therefore, it is natural that systematists should wonder about the accuracy of their reconstructed trees. This issue of *Systematic Biology* contains reviews of the four approaches that systematists have explored to examine phylogenetic accuracy and assess confidence in their results: evolutionary simulations (Huelsenbeck, 1995), exploration of known (observed) phylogenies (briefly reviewed here), statistical evaluations (Li and Zharkikh, 1995), and congruence studies (Miyamoto and Fitch, 1995).

In evaluating phylogenetic accuracy, there are two common goals: one may ask about general properties of phylogenetic methods or about a specific phylogenetic estimate. A systematist may address how well a given method works under different circumstances (e.g., different evolutionary conditions, different amounts of information, different types of trees, etc.). Such studies may ask about phylogenetic performance of a single method or may compare several methods. This approach primarily involves numerical simulations and investigation of known phylogenies. On the other hand, statistical and congruence studies tend to address specific questions of phylogenetic accuracy, i.e., how much confidence can be placed in a specific phylogenetic result? Of course, the distinction is not always clear because simulations can be used to address confidence in a particular empirical result (e.g., Hillis et al., 1994a) and general conclusions about the relative accuracy of phylogenetic methods can come from statistical or congruence studies (e.g., Penny et

[1] E-mail: hillis@bull.zo.utexas.edu.

al., 1982; Allard and Miyamoto, 1992; Miyamoto et al., 1994). Nonetheless, the distinction is valid for the majority of studies to date.

The purpose of this paper is to review progress in investigations of phylogenetic accuracy, to introduce the major approaches that have been developed, to explore the logical relationships among these methods, and to address their possible advantages, disadvantages, and future directions.

Criteria for Comparing Methods

Penny et al. (1992) identified five criteria for comparing phylogenetic methods: consistency, efficiency (called "power" by Penny et al.), robustness, computational speed, and discriminating ability. Hillis and Huelsenbeck (1994) added versatility to this list. Although all of these criteria may be important for selecting a method, the relative weight an individual investigator applies to each criterion is likely to vary depending on the desired application. In the current context, the focus is on the accurate reconstruction of branching relationships, so the first three criteria are of the greatest relevance.

Consistency

A phylogenetic method is consistent for a given evolutionary model if the method converges on the correct tree as the data available to the method become infinite. All methods are consistent when their assumptions (explicit and implicit) are met, and all methods are inconsistent when these assumptions are violated sufficiently. Perhaps because it is relatively easy to evaluate consistency of a method under a given model of evolution, consistency has been emphasized relative to other criteria in comparing phylogenetic methods (Felsenstein, 1978, 1983b; DeBry, 1992; Sidow, 1993). Certainly, it is of interest to identify particular conditions that may lead to inconsistency for a given method, particularly if the conditions that result in consistency are highly restrictive. However, knowing that a method will obtain a correct tree given an infinite amount of data when its assumptions are met perfectly is probably of less interest to most systematists than knowing how the method will perform given limited data under more realistic conditions (Hillis et al., 1994b). Obviously, it makes no sense to say that one method is consistent and another is not without reference to a particular tree and model of evolution. Stated another way, consistency studies provide a means for identifying the underlying implicit assumptions of phylogenetic methods.

Efficiency

Statistical efficiency is a measure of how quickly a method converges on the correct solution as more data are applied to the problem. In the case of phylogenetic methods, efficiency may be measured in terms of the number of characters required to find the correct solution at a given frequency or in terms of the frequency of correct solutions at a given sample size. It may seem intuitive that consistency and efficiency are closely related, but this intuition is wrong. Two methods may both be consistent for a given tree and model of evolution (i.e., they will both converge on the correct solution given infinite data), but nonetheless they may differ dramatically in the amount of information (e.g., length of nucleotide sequence) needed to find the correct solution at high probability. Hillis et al. (1994b) presented an example of a simple four-taxon tree with all branches of equal length evolving under a Kimura model of evolution (Kimura, 1980). Even with relatively high rates of evolution and a strong transition bias, all studied methods of phylogenetic inference are consistent for this tree and model. However, the number of nucleotides needed to find the correct tree (with a probability of >0.99) ranges from about 200 to more than 10^9, depending on the method used! In this case, knowledge of the relative efficiency of the methods is clearly much more important than knowledge of their consistency.

Robustness

Perhaps of greatest interest to practicing systematists is the relative robustness of phylogenetic methods. All methods are based on explicit or implicit assumptions about the evolutionary process, and yet we know these assumptions are violated to one degree or another in real data. For instance, virtually all methods assume that individual characters are evolving independently, and yet sources of nonindependence are known for both molecular and morphological data (Wheeler and Honeycutt, 1988; Dixon and Hillis, 1993). We also know that real patterns of substitution frequencies often differ significantly from the simple models assumed by many phylogenetic methods (Gojobori et al., 1982; Li et al., 1984; Moriyama et al., 1991; Hillis et al., 1994a). However, the degree to which these violations of assumptions will affect performance of phylogenetic methods is still largely an open question. Although specific departures from assumed models can be examined through simulations, biological data are required to compare the expectations of performance under ideal conditions to the limitations of the real world. Thus, experimental phylogenies and congruence studies are critical for evaluating robustness of methods in the real world.

Methods for Assessing Performance

Simulations

> The major problem in studying the relative efficiencies [of phylogenetic methods] is that the true tree is usually unknown for any set of real organisms or any set of real DNA sequences, so that it is difficult to judge which tree is the correct one. However, this problem can be avoided if we use computer simulation. (Nei, 1991:90)

> The evolutionary models used in many simulation studies are exceedingly simple, and even though they will surely become more sophisticated (e.g., more "realistic") in the future, such studies will still face a credibility gap. (Miyamoto and Cracraft, 1991:11)

One of the most common methods of comparing phylogenetic methods is through numerical simulations under explicitly stated evolutionary models (e.g., Peacock and Boulter, 1975; Blanken et al., 1982; Tateno et al., 1982, 1994; Sourdis and Nei, 1988; Jin and Nei, 1990; Rohlf et al., 1990; Nei, 1991; Huelsenbeck and Hillis, 1993; Kim, 1993; Kim et al., 1993; Schöniger and von Haeseler, 1993; Charleston et al., 1994; Hillis et al., 1994a, 1994b; Kuhner and Felsenstein, 1994). Simulations are useful because they can exhaustively explore the effects of models of evolution, tree topologies, relative or absolute rates of evolution, or any other parameters that are thought to affect the performance of phylogenetic methods. Although the models of evolution always will be gross oversimplifications of actual evolutionary processes, the goal of simulations is to detect generalizations about the performance of methods that will be widely applicable to real world situations. An example of such a generalization was the conclusion that long branches attract each other in many phylogenetic methods, leading to a bias in favor of trees with connected long branches (Hendy and Penny, 1989). This discovery has resulted in caution on the part of systematists when faced with estimated trees that unite long branches (e.g., Allard and Miyamoto, 1992). It is appropriate to ask in these circumstances if the apparent signal is greater than could be explained by the bias in the methods alone.

Many systematists dismiss the results of simulation studies because the conclusions of such studies all too often seem to match preexisting preferences of the authors. This problem arises because all methods have conditions for which they work well and other conditions for which they work poorly. It is relatively easy to identify the optimal conditions of a favorite method and then to present simulation results that compare competing methods only at this optimum. Such results are of very limited interest, but the conclusions drawn from such studies often are presented as if they were general. For instance, the UPGMA algorithm is now well known to be highly sensitive to unequal rates of evolution among the branches of a tree, and numerous simulation studies have shown that this method performs very poorly in comparison to most other competing methods if even relatively small differences in rates of evolution are introduced (e.g., Saitou and Nei, 1987; Rohlf et al., 1990; Nei, 1991; Hillis et al., 1994a, 1994b).

Nonetheless, it is possible to find simulation studies that conclude UPGMA is "generally superior to the other methods" (Heijerman, 1991:96). Similar general claims can be found for almost every major method, with simulation studies to back up the disparate conclusions. Obviously, then, such conclusions are not very general, and some source or sources of bias must exist in the individual studies.

Studies often may be biased in the selection of parameters that describe the simulated tree; branching order, branch lengths, and number of terminal taxa each have a strong influence on simulation results. For some simple situations, such as the four-taxon, two-rates problem first outlined by Felsenstein (1978), it is possible to examine the entire parameter space for combinations of branch lengths (see Huelsenbeck, 1995). For more complex trees, it is not as obvious how tree space can or should be delimited. Even within the parameter space of the simple situations, however, different regions can be identified where the rankings of methods switch based on their relative efficiency (Huelsenbeck and Hillis, 1993; Hillis et al., 1994b). Studies that only examine trees within one region of this parameter space are likely to draw conclusions that are not generally applicable (e.g., Tateno et al., 1994).

The evolutionary model also can be a source of bias. Most simulations assume a model of evolution that perfectly matches the assumptions of one or more methods. This is a useful starting point, because it allows the investigator to assess performance of the method under a best-case scenario. The robustness of the method can then be examined in a systematic fashion by violating the assumptions of the method one at a time. This approach leads to problems only when the conclusions drawn from the study are purported to be more general than they really are. For instance, it makes no sense to simulate a tree using a gamma distribution of evolutionary rates across sites and then to conclude that a method that assumes a gamma distribution is generally better than a method that does not because the gamma-assuming method performs better under this simulation. The simulation does not demonstrate that nature obeys a gamma distribution.

An alternative method of simulation that attempts to incorporate information from the real world is known as parametric bootstrapping (Efron, 1985; Felsenstein, 1988; Bull et al., 1993a). In this approach, a model of evolution and a model tree are constructed based on parameters estimated from data, and then replicates are generated through simulation (differences among replicates occur because of stochastic variation). The method should not be confused with nonparametric (traditional) bootstrapping, in which the pseudoreplicates are not independent and thus introduce bias into the distributions generated from the subsamples (Efron, 1979, 1987; Hillis and Bull, 1993). Parametric bootstrapping offers a method of producing independent replicates of observed data sets, which can be used to test the performance of competing methods (e.g., Hillis et al., 1994a) or to extend the conclusions of an experimental study (e.g., Bull et al., 1993a).

A less obvious source of bias in simulations results from poor implementation of a method. For instance, many simulation studies that compare clustering algorithms and optimality criteria involve highly inefficient searches for optimal trees. Many studies that supposedly compare the results of maximum parsimony, maximum likelihood, or minimum evolution (all optimality criteria) with those of a clustering method such as neighbor joining (a heuristic algorithm for approximating minimum evolution trees; Nei, 1991) actually are using approximate solutions to the optimality-based approaches. Although these approximate solutions may be necessary because of the complexity of the problem, it should be made clear in such cases that the optimal solutions have not necessarily been obtained, and the algorithms used to approximate the solutions should be clearly specified.

Some methods are more likely than others to find multiple, equally good solutions, making the treatment of ties a potential source of bias. Some authors count a method as incorrect if it finds more than one solution, even if one of the solutions is the correct tree. Other authors count a method as completely correct if the correct tree is among the solutions. (There is also variation among authors as to whether they score the percentage of time an entire tree is correct or the proportion of a tree's component clades that are correct.) Clearly, the two extreme methods of counting ties will bias the results either against or in favor of the methods that find more than one solution, respectively. An obvious way to avoid such bias is to use the average number of correctly resolved components across all optimal solutions to score each method. Thus, if a method finds two optimal two-component trees, one with both components correct and one with only one component correct, this scoring system would indicate a 75% success rate for the method (as opposed to 0% or 100% success rates indicated by the extreme scoring methods).

Despite the limitations of simulation studies, they have been very useful for formulating hypotheses about the behavior of phylogenetic methods under a wide range of model conditions. Obviously, there is a need to explore more complex models of evolution than have been examined to date, as well as to investigate the effects of more complex and larger trees. But a more basic question concerns the degree to which the results based on idealized evolutionary models of simulations match the results expected from real evolving organisms. To answer this question effectively, the predictions generated through simulations must be tested empirically.

Known Phylogenies

There are some fundamental philosophical and empirical differences between simulations of fictitious taxa and their DNA sequences, on the one hand; and real-world taxa and their sequence characteristics, on the other. (Miyamoto and Cracraft, 1991:11)

Although I am skeptical that the results of [experimental phylogenies] "directly support the legitimacy of methods for phylogenetic estimation," it remains to be seen what experimental phylogenetics can teach us about the problem of phylogenetic inference. (Sober, 1993:89)

Simulations are best suited for assessing consistency and efficiency of methods when their assumptions are met perfectly or for examining robustness of methods when the assumptions are violated in very specific (but perhaps not realistic) ways. However, no simulation will approach in complexity the evolutionary constraints and processes experienced by real organisms (Hillis et al., 1993b). Of course, the specific evolutionary processes will differ among each group of organisms and will also vary through time in response to variation in the external environment. Simulations allow us to describe the behavior of methods in an ideal world that we know differs from the real world in most of the details, but the hope is that we can make generalizations from simulations about the relative behavior of methods that will apply to the real world. Known phylogenies of actual organisms allow direct experimental tests of these generalizations.

There are two major types of known phylogenies: agricultural or laboratory lineages for which records have been kept (e.g., Baum, 1984; Fitch and Atchley, 1985, 1987; Atchley and Fitch, 1991; Hillis and Bull, 1991) and experimental phylogenies generated for the purpose of testing phylogenetic methods (e.g., Hillis et al., 1992, 1994a; Bull et al., 1993a). Neither of these approaches begins to cover the diversity of phylogenies that are estimated from the real world, but they do involve real, evolving biological organisms and situations for which phylogenetic methods are supposed to be applicable. Thus, known phylogenies provide an opportunity to test predictions made from simulations in systems where the evolutionary processes are constrained by biological organisms rather than by the mind of the investigator. If predictions made from simulations are falsified with known phylogenies, then it is clear that the predictions are not generally applicable to all of life. However, if predictions from simulations are supported by studies of known phylogenies, then we know that the predictions from the idealized model conditions apply to at least part of the real world.

Moreover, differences in the results between simulations and experimental phylogenies may suggest ways that the simulations can be made more realistic, just as simulations can suggest conditions of interest for testing with known phylogenies.

Experimental or other known phylogenies clearly have major limitations. Historical records of cultivated organisms are severely limited, and such organisms typically have undergone many reticulations and relatively little genetic divergence. This situation provides a testing ground for methods designed to reconstruct networks (i.e., graphs with cycles) of closely related organisms (e.g., Templeton et al., 1987, 1992; Hein, 1990, 1993; Bandelt and Dress, 1992; Crandall, 1994; Crandall et al., 1994), but it is very limiting for studies of phylogenetic trees. Experimentally generated phylogenies, however, are limited primarily by mutation rates. For an experimental phylogeny to be useful, the lineages must undergo a significant amount of divergence in a short period of time (preferably measured in months rather than millennia). Most systematists are accustomed to the very slow mutation rates of typical eukaryotes, which for gene sequences are often on the order of 10^{-9} mutations/site/year. However, all of life is not evolving so slowly, and the genes of some RNA viruses are evolving up to tens of millions of times faster (Domingo and Holland, 1994). Most DNA viruses evolve at a somewhat slower pace, but their mutation rate can be controlled to some degree by manipulating their mutagenic environment (e.g., Studier, 1980) or through selection for a novel environment. These manipulations allow the controlled study of the effects of differing mutation biases and selection pressures, but under the constraints imposed by biological function. In a simulation, investigators may impose a particular mutation model, but then they have no way of knowing which substitutions would be tolerated by a real organism or how substitutions in different parts of a gene might interact. Experimental phylogenies overcome this limitation.

Many of the interesting problems in phylogenetic reconstruction concern organisms that differ by a large percentage of their genome. To date, most experimental phylogenies have involved comparisons of organisms that are relatively closely related. However, many viral genomes do not seem to be greatly constrained in the amount of divergence that is tolerated, and in the future we can expect to see experimentally generated viral lineages that differ for given genes by as much as ribosomal RNA genes differ across all of life.

Unfortunately, it will be difficult to construct experimental phylogenies of cellular organisms that incorporate much genetic divergence, and it will be particularly difficult to extend the approach to a eukaryotic system. The generation times of many viruses are measured in minutes, rather than hours, days, or years, and with their small genomes they seem to require less time to accommodate new substitutions than do cellular organisms. Bacteria hold considerable promise for experimental phylogenies, but assessing the genetic variation across their comparatively large genomes will be much harder than with viruses. Among the eukaryotes, small organisms with rapid generation times (e.g., *Saccharomyces*) have potential for experimental phylogenies, but it is unlikely that highly divergent lineages can be created without considerable time and effort.

An additional role for experimental phylogenies is in providing information about molecular evolutionary processes and how these processes can affect phylogenetic analyses. For instance, molecular systematists sometimes claim that molecular data are immune to selective convergence (e.g., Sibley and Ahlquist, 1987), despite some evidence to the contrary (e.g., Stewart and Wilson, 1987). Experimental phylogenies provide an opportunity to test this assertion directly by manipulating viral environments in parallel and then comparing phenotypic convergence (e.g., growth characteristics) with molecular convergence. Such studies have considerable potential for identifying the limits and possibilities of phylogenetic analyses.

Statistical Approaches

> As DNA sequences accumulate, there will be an increasing demand for statistical methods to estimate evolutionary trees from them, and to test hypotheses about the evolutionary process. (Felsenstein, 1981:368)

> It is remarkable that, in a century which has seen such a large growth in the application of statistics to the natural sciences, the fundamental issues of statistical inference have not been resolved. There are not many more statisticians than opinions as to how to assess rival hypotheses in the light of data. (Edwards, 1969:1233)

Statistical approaches typically are used to address phylogenetic accuracy in a particular case rather than to identify general conditions where methods perform well or poorly. There has been rapid development of statistical tests for phylogenetic analyses over the past decade (see reviews by Felsenstein, 1988; Li and Guoy, 1991; Hillis et al., 1993a; Li and Zharkikh, 1995). This development has not been without controversy, and debates about statistical approaches in systematics have in many ways paralleled more general debates within the field of statistics. One topic of debate is whether phylogenetic results should be evaluated in probabilistic or relativistic terms. In the probabilist framework (which tends to dominate the field), statistical tests seek to provide a probability of a particular hypothesis being true given the observed data (accuracy), or at least the probability that a given result would be supported by a particular method upon repeated trials (repeatability). In the relativist framework (discussed in various forms by Fisher [1956], Birnbaum [1962], Hacking [1965], and Edwards [1969, 1992]), tests examine the relative support (e.g., the likelihood ratio) of a given data set for one hypothesis versus another, without any statement about the absolute probability of either. Both approaches have their advocates, among statisticians in general as well as among phylogeneticists. Some approaches, such as likelihood, have been used in both frameworks (Edwards and Cavalli-Sforza, 1964; Edwards, 1969, 1992; Felsenstein, 1973a, 1973b, 1981).

Despite (or perhaps because of) the considerable recent development of statistical methods in phylogenetics, statistical tests and terminology are widely misapplied in systematics. It is very common to read that a result is statistically significant without a clear indication of what the author means by this statement. Some statistical methods are designed to test whether a given data set is more structured than would be expected from random data, whereas others test the strength of a particular result; clearly, "statistically significant" will mean very different things in these two cases. The results of bootstrap (and other) analyses are often called confidence limits, even though no range of results is presented, as would be expected for the limits of a confidence interval, and no one has suggested how phylogenies could be described in terms of a continuous variable. The precision of a test statistic is sometimes confused with its accuracy, and authors rarely indicate whether they are interpreting a given statistic as a measure of phylogenetic accuracy or repeatability or simply as a comparative but otherwise undefined heuristic (Hillis and Bull, 1993). The confusion over precision versus accuracy has led some workers to recommend very large numbers of bootstrap replications (e.g., Hedges, 1992), which does nothing to reduce the bias of the estimates (it merely makes them highly precise biased estimates). The computation time would be used much more effectively by conducting either iterated bootstraps (Hall and Martin, 1988; Rodrigo, 1993) or preferably the complete-and-partial bootstrap technique (Li and Zharkikh, 1995; Zharkikh and Li, in press).

During the past decade there has been considerable development of resampling methods (e.g., Felsenstein, 1985a, 1988; Lanyon, 1985; Penny and Hendy, 1985a, 1986; Zharkikh and Li, 1992a, 1992b, in press; Hillis and Bull, 1993; Rodrigo, 1993; Steel et al., 1993; Li and Zharkikh, 1995), randomization methods (e.g., Archie, 1989; Faith, 1991; Faith and Cranston, 1991; Hillis, 1991; Huelsenbeck, 1991; Hillis and Huelsenbeck, 1992; Källersjö et al., 1992), analytical methods (e.g., Felsenstein, 1981, 1983a, 1985b, 1987; Templeton, 1983a, 1983b; Lake, 1987; Prager and Wilson, 1988; Kishino and Hasegawa, 1989; Li, 1989; Williams and Goodman, 1989; Bull et al., 1993a), and relative support approaches (e.g., Bremer, 1988; Donoghue et al., 1992; Davis, 1993). Goldman (1993) and Yang et al. (1994) recently presented another approach that involves testing the model of the phylogenetic method against the observed data to ask if the model is adequate to explain the observations. Other workers (e.g., Sanderson, 1989) have suggested the construction of "confidence sets" of trees, which would be analogous to confidence limits of continuous variables. This concept may prove especially useful in comparing the results from multiple data sets and forms a logical link between statistical approaches and congruence studies.

Congruence Studies

> Extensive congruence among branching patterns derived from independent data sets and by different methods of analysis is unlikely to occur for any reason other than phylogeny. (Sheldon and Bledsoe, 1993:256–257)

> There may indeed be substantial congruence between the two data sets, but that "congruence" is not quite what we had hoped it would be. (Swofford, 1991:326)

Congruence studies seek out common phylogenetic patterns in multiple, independent data sets. If multiple trees inferred from independent data sets all show the same pattern of relationships, this is usually taken as strong evidence for the veracity of the shared components and the accuracy of the phylogenetic method or methods (Mickevich and Johnson, 1976; Prager and Wilson, 1976; Mickevich, 1978; McKitrick, 1985; Hillis, 1987; Miyamoto and Cracraft, 1991). Even if the trees are not identical, if their similarities are greater than is expected from chance alone, the level of congruence may be taken as a measure of phylogenetic accuracy (Mickevich and Farris, 1981; Penny et al., 1982, 1991; Penny and Hendy, 1985b, 1986; Guyer and Slowinski, 1991; Page, 1991; Swofford, 1991; Miyamoto et al., 1994). Although other sources of spurious congruence have been explored (e.g., long branches; Allard and Miyamoto, 1992), phylogeny is the only obvious explanation in the vast majority of cases.

Quantitative congruence studies are a type of meta-analysis, a term coined to describe "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976). Although the basic concept of meta-analysis dates back to the early part of this century (Pearson, 1904; see Olkin, 1990), meta-analyses have become increasingly commonplace during the past decade (Hedges and Olkin, 1985; Dickersin and Berlin, 1992). Many meta-analyses have two basic parts: an analysis of the combined data from across studies (e.g., a mean estimate of some parameter) and a comparison of the combined findings with those of the individual studies (e.g., does the mean estimate fall within the confidence limits of the individual estimates?). It is not unusual to find that a grand mean from many independent studies intersects the confidence limits of the results for each of the individual investigations, lending support to the combined result as a general explanation (for examples, see Mann, 1990; Dickersin and Berlin, 1992).

Recent discussions of combined versus separate analyses of multiple, independent phylogenetic data sets have covered much of the same ground that has centered around the debate over meta-analysis in general (Hillis, 1987; Kluge, 1989; Swofford, 1991; Bull et al., 1993b; de Queiroz, 1993; Chippindale and Wiens, 1994; Huelsenbeck et al., 1994). The combination of multiple data sets into a single analysis carries with it assumptions that the same underlying tree is being reconstructed in each of the studies and that the methods of analysis are appropriate for each data set. Significant differences in the results are an indication that one or both of these assumptions have been violated for at least one of the data sets (Bull et al., 1993b; de Queiroz, 1993). The current problem is to determine whether two phylogenetic data sets are significantly heterogeneous or if the differences can be attributed to stochastic variation; several tests are under development (see Swofford, 1991; Rodrigo et al., 1993).

Consensus techniques (Adams, 1972; Nelson, 1979; Hillis, 1987; Bremer, 1990; Swofford, 1991) are sometimes viewed as methods for combining data from separate analyses to generate a mean phylogenetic estimate (Kluge, 1989). However, combination of phylogenetic results across studies through the use of consensus techniques tends to discard information about the strength of individual results, resulting in the loss of information on one hand and overemphasis of weakly supported results on the other (Miyamoto, 1985; Barrett et al., 1991). Thus, it is important not to interpret consensus trees as estimates of phylogenies but rather simply as statements about areas of agreement among trees (Swofford, 1991).

The biggest obstacle to the successful implementation of something that resembles a meta-analysis in phylogenetics is development of reasonable methods for generating confidence sets of phylogenetic trees (see Sanderson, 1989). Although a combined analysis of several data sets (assuming that they are appropriate for combining) may give the single best estimate of phylogeny (Hillis, 1987; Kluge, 1989), the conclusion would be greatly strengthened if it were compatible with that of each of the individual data sets as well (even if it were not equivalent to the best point estimate from each analysis; Swofford, 1991). If some of the individual studies are incompatible, then alternative explanations (e.g., different gene trees, inappropriate methods of analysis) can be sought (see Bull et al., 1993b; de Queiroz, 1993). To implement this approach, a systematist would not just present a single tree from an analysis but instead would give the optimal estimate as well as describe a set of trees considered to be consistent with the data at a given level of confidence (i.e., a confidence set of trees). In comparing multiple data sets, an investigator synthesizing the results could look for the common intersection of the independent studies (e.g., Lanyon, 1993; Miyamoto et al., 1994) or ask if the result from the combined data sets is within the confidence sets of the individual studies.

To date, most congruence studies of phylogenetic analyses have been much less forgiving of the individual investigations than in the approach suggested above. Such studies have instead looked for exact matches in the phylogenetic results across studies. Nonetheless, the fact that congruence studies typically have found a high degree of correspondence among independent phylogenetic estimates is strong evidence that the individual studies are doing a remarkably good job of recovering the underlying historical information. The other possibility is that the separate studies are in agreement because of some common, spurious, nonphylogenetic signal, but to date no one has offered a convincing general alternative to phylogeny.

What Are the Prospects for an Accurate Tree of Life?

> On that happy day when molecular systematists achieve the goal of adequate sampling in terms of both taxa and sequence length ..., and when the computer and the program capable of analysing the alignment of life exist, there are two possible extremes: "one tree," or "10^{999} equally parsimonious trees." (Patterson et al., 1993:180)

> "I checked it quite thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is." (Adams, 1979:181)

Most studies of phylogenetic accuracy indicate that existing methods should be highly successful for many classes of phylogenetic problems, given data sets within the range that reasonably can be expected to be obtained. However, these studies also indicate that certain classes of phylogenetic problems may simply be too difficult to expect a well-supported resolution, given the limits on organismal genome sizes (and hence the number of independent characters) (Hillis et al., 1994b). Assuming there are about 30 million living species on Earth, then there are approximately $10^{300,000,000}$ possible bifurcating trees that could depict the relationships among these species, give or take a few million orders of magnitude (Hillis et al., 1994a). Therefore, even the worse of the two "extremes" suggested by Patterson et al. (quoted above) would represent outstanding resolution. Simulations, known phylogenies, statistical analyses, and congruence studies all indicate that methods of phylogenetic analysis can be (and often are) highly accurate for problems as diverse as life itself, given sufficient sampling, sufficient attention to rigorous analysis, and sufficient computational power. Just don't expect the final answer anytime soon.

Acknowledgments

Jim Bull, Mike Charleston, Paul Chippindale, Keith Crandall, Tim Crowe, Cliff Cunningham, A. W. F. Edwards, Jotun Hein, John Huelsenbeck, Mike Miyamoto, Barbara Mable, and an anonymous reviewer read this manuscript and offered useful suggestions. I thank Mike Miyamoto for inviting me to prepare an introduction to this series of papers on phylogenetic accuracy. My studies on phylogenetic accuracy have been supported by grants from the National Science Foundation.

References

Adams, D. 1979. The hitchhiker's guide to the galaxy. Crown, New York.

Adams, E. N., III. 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21:390–397.

Allard, M. W., and M. M. Miyamoto. 1992. Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. Mol. Biol. Evol. 9:778–786.

Archie, J. W. 1989. Phylogenies of plant families: A demonstration of phylogenetic randomness in DNA sequence data derived from proteins. Evolution 43:1796–1800.

Atchley, W. R., and W. M. Fitch. 1991. Gene trees and the origins of inbred strains of mice. Science 254:554–558.

Bandelt, H.-J., and A. W. M. Dress. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1:242–252.

Barrett, M., M. J. Donoghue, and E. Sober. 1991. Against consensus. Syst. Zool. 40:486–493.

Baum, B. R. 1984. Application of compatibility and parsimony methods at the infraspecific, specific, and generic levels in Poaceae. Pages 192–220 *in* Cladistics: Perspectives on the reconstruction of evolutionary history. Columbia Univ. Press, New York.

Birnbaum, A. 1962. On the foundations of statistical inference. J. Am. Stat. Assoc. 57:269–326.

Blanken, R. L., L. C. Klotz, and A. G. Hinnebusch. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. 19:9–19.

Bremer, K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803.

Bremer, K. 1990. Combinable component consensus. Cladistics 6:369–372.

Bull, J. J., C. W. Cunningham, I. J. Molineux, M. R. Badgett, and D. M. Hillis. 1993a. Experimental molecular evolution of bacteriophage T7. Evolution 47:993–1007.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993b. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

Charleston, M. A., M. D. Hendy, and D. Penny. 1994. The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. J. Comput. Biol. 1:133–151.

Chippindale, P. T., and J. J. Wiens. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. Syst. Biol. 43:278–287.

Crandall, K. A. 1994. Intraspecific cladogram estimation: Accuracy at higher levels of divergence. Syst. Biol. 43:222–235.

Crandall, K. A., A. R. Templeton, and C. F. Sing. 1994. Intraspecific phylogenetics: Problems and solutions. Pages 273–297 *in* Models in phylogeny reconstruction (R. W. Scotland, D. J. Siebert, and D. M. Williams, eds.). Clarendon Press, Oxford, England.

Davis, J. I. 1993. Character removal as a means for assessing stability of clades. Cladistics 9:201–210.

DeBry, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. 9:537–551.

de Queiroz, A. 1993. For consensus (sometimes). Syst. Biol. 42:368–372.

Dickersin, K., and J. A. Berlin. 1992. Meta-analysis: State-of-the-science. Epidemiol. Rev. 14:154–176.

Dixon, M. T., and D. M. Hillis. 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10:256–267.

Domingo, E., and J. J. Holland. 1994. Mutation rates and rapid evolution of RNA viruses. Pages 161–184 *in* The evolutionary biology of viruses (S. S. Morse, ed.). Raven Press, New York.

Donoghue, M. J., R. G. Olmstead, J. F. Smith, and J. D. Palmer. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Mo. Bot. Gard. 79:249–265.

Edwards, A. W. F. 1969. Statistical methods in scientific inference. Nature 222:1233–1237.

Edwards, A. W. F. 1992. Likelihood: An account of the statistical concept of likelihood and its application to scientific inference, 2nd edition. Johns Hopkins Univ. Press, Baltimore, Maryland.

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. Syst. Assoc. Publ. 6:67–76.

EFRON, B. 1979. Bootstrapping methods: Another look at the jackknife. Ann. Stat. 7:1–26.

EFRON, B. 1985. Bootstrap confidence intervals for a class of parametric problems. Biometrika 72:45–58.

EFRON, B. 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82:171–185.

FAITH, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40:366–375.

FAITH, D. P., AND P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics 7:1–28.

FELSENSTEIN, J. 1973a. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

FELSENSTEIN, J. 1973b. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

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

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

FELSENSTEIN, J. 1983a. Inferring evolutionary trees from DNA sequences. Pages 133–150 *in* Statistical analysis of DNA sequence data (B. S. Weir, ed.). Marcel Dekker, New York.

FELSENSTEIN, J. 1983b. Parsimony in systematics: Biological and statistical issues. Annu. Rev. Ecol. Syst. 14:313–333.

FELSENSTEIN, J. 1985a. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

FELSENSTEIN, J. 1985b. Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34:152–161.

FELSENSTEIN, J. 1987. Estimation of hominoid phylogeny from a DNA hybridization data set. J. Mol. Evol. 26:123–131.

FELSENSTEIN, J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521–565.

FISHER, R. A. 1956. Statistical methods and scientific inference. Oliver and Boyd, Edinburgh.

FITCH, W. M., AND W. R. ATCHLEY. 1985. Evolution in inbred strains of mice appears rapid. Science 228:1169–1175.

FITCH, W. M., AND W. R. ATCHLEY. 1987. Divergence in inbred strains of mice: A comparison of three different types of data. Pages 203–216 *in* Molecules and morphology in evolution: Conflict or compromise? (C. Patterson, ed.). Cambridge Univ. Press, Cambridge, England.

GLASS, G. V. 1976. Primary, secondary and meta-analysis of research. Ed. Res. 5:3–8.

GOJOBORI, T., W.-H. LI, AND D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360–369.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

GUYER, C., AND J. B. SLOWINSKI. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. Evolution 45:340–350.

HACKING, I. 1965. Logic of statistical inference. Cambridge Univ. Press, Cambridge, England.

HALL, P., AND M. A. MARTIN. 1988. On bootstrap resampling and iteration. Biometrika 75:661–671.

HEDGES, L. V., AND I. OLKIN. 1985. Statistical methods for meta-analysis. Academic Press, Orlando, Florida.

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. Mol. Biol. Evol. 9:366–369.

HEIJERMAN, T. 1991. Adequacy of numerical taxonomic methods: Further experiments using simulated data. Z. Zool. Syst. Evolutionsforsch. 31:81–97.

HEIN, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98:185–200.

HEIN, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. 36:396–405.

HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

HILLIS, D. M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18:23–42.

HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. Pages 278–294 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

HILLIS, D. M., M. W. ALLARD, AND M. M. MIYAMOTO. 1993a. Analysis of DNA sequence data: Phylogenetic inference. Methods Enzymol. 224:456–487.

HILLIS, D. M., AND J. J. BULL. 1991. Of genes and genomes. Science 254:528.

HILLIS, D. M., AND J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analyses. Syst. Biol. 42:182–192.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT, AND I. J. MOLINEUX. 1992. Experimental phylogenetics: Generation of a known phylogeny. Science 255:589–592.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT, AND I. J. MOLINEUX. 1993b. Experimental approaches to phylogenetic analysis. Syst. Biol. 42:90–92.

HILLIS, D. M., AND J. P. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83:189–195.

HILLIS, D. M., AND J. P. HUELSENBECK. 1994. To tree the truth: Biological and numerical simulations of phylogeny. Pages 55–67 *in* Molecular evolution of physiological processes (D. M. Fambrough, ed.). Rockefeller Univ. Press, New York.

HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994a. Application and accuracy of molecular phylogenies. Science 264:671–677.

HILLIS, D. M., J. P. HUELSENBECK, AND D. L. SWOFFORD. 1994b. Hobgoblin of phylogenetics? Nature 369:363–364.

HUELSENBECK, J. P. 1991. Tree-length distribution skewness: An indicator of phylogenetic information. Syst. Zool. 40:257–270.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

HUELSENBECK, J. P., D. L. SWOFFORD, C. W. CUNNINGHAM, J. J. BULL, AND P. W. WADDELL. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? Syst. Biol. 43:288–291.

JIN, L., AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82–102.

KÄLLERSJÖ, M., J. S. FARRIS, A. G. KLUGE, AND C. BULT. 1992. Skewness and permutation. Cladistics 8:275–287.

KIM, J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. 42:331–340.

KIM, J., F. J. ROHLF, AND R. R. SOKAL. 1993. The accuracy of phylogenetic estimation using the neighbor-joining method. Evolution 47:471–486.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38:7–25.

KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

LAKE, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. 4:167–191.

LANYON, S. M. 1985. Detecting internal inconsistencies in distance data. Syst. Zool. 34:397–403.

LANYON, S. M. 1993. Phylogenetic frameworks: Towards a firmer foundation for the comparative approach. Biol. J. Linn. Soc. 49:45–61.

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. 6:424–435.

LI, W.-H., AND M. GUOY. 1991. Statistical methods for testing phylogenies. Pages 249–277 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

LI, W.-H., C.-I. WU, AND C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. 21:58–71.

LI, W.-H., AND A. ZHARKIKH. 1995. Statistical tests of DNA phylogenies. Syst. Biol. 44:49–63.

MANN, C. 1990. Meta-analysis in the breech. Science 249:476–480.

MCKITRICK, M. C. 1985. Monophyly of the Tyrannidae (Aves): Comparison of morphology and DNA. Syst. Zool. 34:35–45.

MICKEVICH, M. F. 1978. Taxonomic congruence. Syst. Zool. 27:143–158.

MICKEVICH, M. F., AND J. S. FARRIS. 1981. The implications of congruence in *Menidia*. Syst. Zool. 30:351–370.

MICKEVICH, M. F., AND M. S. JOHNSON. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. Syst. Zool. 25:260–270.

MIYAMOTO, M. M. 1985. Consensus cladograms and general classifications. Cladistics 1:186–189.

MIYAMOTO, M. M., AND W. M. FITCH. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44:64–76.

MIYAMOTO, M. M., M. W. ALLARD, R. M. ADKINS, L. L. JANECEK, AND R. L. HONEYCUTT. 1994. A congruence test of reliability using linked mitochondrial DNA sequences. Syst. Biol. 43:236–249.

MIYAMOTO, M. M., AND J. CRACRAFT. 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. Pages 3–17 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

MORIYAMA, E. N., Y. INA, K. IKEO, N. SHIMIZU, AND T. GOJOBORI. 1991. Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. 32:360–363.

NEI, M. 1991. Relative efficiencies of different tree making methods for molecular data. Pages 90–128 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

NELSON, G. J. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Familles des Plantes* (1763–1764). Syst. Zool. 28:1–21.

OLKIN, I. 1990. History and goals. Pages 3–10 *in* The future of meta-analysis (K. W. Wachter and M. L. Straf, eds.). Russell Sage Foundation, New York.

OLSEN, G. J. 1987. Earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred from various techniques. Cold Spring Harbor Symp. Quant. Biol. 52:825–837.

OU, C.-Y., C. A. CIESIELSKI, G. MYERS, C. I. BANDEA, C.-C. LUO, B. T. M. KORBER, J. I. MULLINS, G. SCHOCHETMAN, R. L. BERKELMAN, A. N. ECONOMOU, J. J. WITTE, L. J. FURMAN, G. A. SATTEN, K. A. MACINNES, J. W. CURRAN, AND H. W. JAFFE. 1992. Molecular epidemiology of HIV transmission in a dental practice. Science 256:1165–1171.

PAGE, R. D. M. 1991. Clocks, clades, and cospeciation: Comparing rates of evolution and timing of cospeciation events in host–parasite assemblages. Syst. Zool. 40:188–198.

PATTERSON, C., D. M. WILLIAMS, AND C. J. HUMPHRIES. 1993. Congruence between molecular and morphological phylogenies. Annu. Rev. Ecol. Syst. 24:153–188.

PEACOCK, D., AND D. BOULTER. 1975. Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. 95:513–527.

PEARSON, K. 1904. Report on certain enteric fever inoculation statistics. Br. Med. J. 3:1243–1246.

PENNY , D., L. R. FOULDS, AND M. D. HENDY. 1982. Testing the theory of evolution by comparing evolutionary trees constructed from five different protein sequences. Nature 297:197–200.

PENNY, D., AND M. D. HENDY. 1985a. Testing methods of evolutionary tree construction. Cladistics 1:266–278.

PENNY, D., AND M. D. HENDY. 1985b. The use of tree comparison metrics. Syst. Zool. 34:75–82.

PENNY, D., AND M. D. HENDY. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3:403–417.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1991. Testing the theory of descent. Pages 155–183 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73–79.

PRAGER, E. M., AND A. C. WILSON. 1976. Congruency of phylogenies derived from different proteins. J. Mol. Evol. 9:45–57.

PRAGER, E. M., AND A. C. WILSON. 1988. Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. J. Mol. Evol. 27:326–335.

RODRIGO, A. G. 1993. Calibrating the bootstrap test of monophyly. Int. J. Parasitol. 23:507–514.

RODRIGO, A. G., M. KELLY-BORGES, P. R. BERGQUIST, AND P. L. BERGQUIST. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. N.Z. J. Bot. 31:257–268.

ROHLF, F. J., W. S. CHANG, R. R. SOKAL, AND J. KIM. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution 44:1671–1684.

SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SANDERSON, M. J. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5:113–129.

SCHÖNIGER, M., AND A. VON HAESELER. 1993. A simple method to improve the reliability of tree reconstructions. Mol. Biol. Evol. 10:471–483.

SHELDON, F. H., AND A. H. BLEDSOE. 1993. Avian molecular systematics, 1970s to 1990s. Annu. Rev. Ecol. Syst. 24:243–278.

SIBLEY, C. G., AND J. E. AHLQUIST. 1987. Avian phylogeny reconstructed from comparisons of the genetic material, DNA. Pages 95–121 *in* Molecules and morphology in evolution: Conflict or compromise? (C. Patterson, ed.). Cambridge Univ. Press, Cambridge, England.

SIDOW, A. 1993. Parsimony or statistics? Nature 367:26.

SOBER, E. 1993. Experimental tests of phylogenetic inference methods. Syst. Biol. 42:85–89.

SOURDIS, J., AND M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. 5:298–311.

STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

STEWART, C.-B., AND A. C. WILSON. 1987. Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. Cold Spring Harbor Symp. Quant. Biol. 52:891–899.

STUDIER, W. F. 1980. The last of the T phages. Pages 72–78 *in* Genes, cells, and behavior: A view of biology fifty years later (N. H. Horowitz and E. Hutchings, Jr., eds.). W. H. Freeman, San Francisco.

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

TATENO, Y., M. NEI, AND F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. 18:387–404.

TATENO, Y., N. TAKEZAKI, AND M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11:261–277.

TEMPLETON, A. R. 1983a. Convergent evolution and nonparametric inferences from restriction data and DNA sequences. Pages 411–501 *in* Statistical analysis of DNA sequence data (B. S. Weir, ed.). Marcel Dekker, New York.

TEMPLETON, A. R. 1983b. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to humans and the apes. Evolution 37:221–244.

TEMPLETON, A. R., E. BOERWINKLE, AND C. F. SING. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351.

TEMPLETON, A. R., K. A. CRANDALL, AND C. F. SING. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. III. Cladogram estimation. Genetics 132:619–633.

WHEELER, W. C., AND R. L. HONEYCUTT. 1988. Paired sequence differences in ribosomal RNAs: Evolutionary and phylogenetic implications. Mol. Biol. Evol. 5:90–96.

WILLIAMS, S. A., AND M. GOODMAN. 1989. A statistical test that supports a human/chimpanzee clade based on noncoding DNA sequence data. Mol. Biol. Evol. 6:325–330.

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.

ZHARKIKH, A., AND W.-H. LI. 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119–1147.

ZHARKIKH, A., AND W.-H. LI. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J. Mol. Evol. 35:356–366.

ZHARKIKH, A., AND W.-H. LI. In press. Estimation of confidence in phylogeny: The full-and-partial bootstrap technique. Mol. Phylogenet. Evol.

Received 11 August 1994; accepted 5 October 1994