To Tree the Truth: Biological and Numerical Simulations of Phylogeny

David M. Hillis and John P. Huelsenbeck

Department of Zoology, The University of Texas at Austin, Austin, Texas 78712

The reconstruction of phylogenetic history has become an integral part of all comparative biological studies over the past few decades (e.g., see Brooks and McLennan, 1991; Harvey and Pagel, 1991; Hillis and Moritz, 1990; Maddison and Maddison, 1992). The range of applications of phylogenetic inference is immense: phylogenies are used for everything from tracking infections of viruses within human populations (e.g., Ou et al., 1992) to studying the evolution of sex determining mechanisms across hundreds of millions of years (e.g., Hillis and Green, 1990) to tracing the earliest lineages of life billions of years ago (e.g., Olsen, 1987). However, it is obviously not possible to go back in time and directly observe any of these phylogenies, so how can we know if phylogenetic methods are finding the correct phylogenies? As with most scientific theories and methods, there are two choices to evaluate the validity of phylogenetic techniques: empirical and theoretical experimentation (or to put it in other terms, biological and numerical simulation). The purpose of this review is to examine the results of both types of studies with regard to performance of phylogenetic methods, and then to make general recommendations about selecting a method for use.

Numerical Versus Biological Simulations

To date, most evaluations of phylogenetic methods have involved numerical simulations: an investigator defines a simple model of evolution, creates some sequences (perhaps at random), and specifies a tree or some rules for generating a tree (e.g., a Markov process of speciation). The investigator then applies the model of evolution to the given sequences and tree, and a computer program carries out the tasks of assigning mutations and recording successive generations of new sequences. After the sequences have "evolved" in computer memory, the various methods of phylogenetic inference can be tested to see which ones perform the most effectively. The advantages of simulating phylogenies in this manner include the ability to generate a sample of thousands or millions of phylogenies with great ease and the ability to generate any conceivable phylogeny. Thus, we can choose some aspect of trees to investigate, define the parameter space of interest, and then examine samples of trees from throughout this parameter space. The limitations of the approach lie only in our ability to identify relevant problems and in computational limitations in analyzing the simulated data sets.

Given the flexibility of the numerical simulation approach, why would we ever turn to experiments with real organisms? There are two principal reasons: we are painfully ignorant about the details of molecular evolution, and computer simulations, by necessity, incorporate gross simplifications of evolutionary processes. The

Molecular Evolution of Physiological Processes © 1994 by The Rockefeller University Press

most complex of computer simulations still make sweeping generalizations and simplifying assumptions about how organisms evolve. This will likely always be the case, because a computer simulation that did not simplify evolutionary processes would have to be as complex as a real functioning organism. As an example, the vast majority of simulations of molecular evolution to date have defined one or two parameters associated with mutation rates; typically, there is either a single mutation rate or two different rates for transitions and transversions, respectively. The most complex simulations may specify as many as twelve different mutation rates for the twelve possible changes that can occur among nucleotides. However, in a real sequence, these rates are likely to differ in ways we are yet to understand across every position in a given gene. Many simulations also assume a constant rate of change across all positions, a situation we know to be very different from what actually occurs in real sequences. There are also complexities that we can imagine but are difficult to model: there may be complex interactions among different nucleotide positions (e.g., having to do with RNA or protein secondary structure, or related to the binding of control sequences), or there may be fluctuating kinds and levels of selection at various developmental stages. Of course, given that we could thoroughly understand such complexities, we could incorporate them into simulations. Unfortunately, our knowledge of molecular evolution is far too rudimentary to develop any but the simplest models at present, and even if we had complete knowledge, it would be computationally intractable to develop such detailed models. Therefore, we need some check on the simulations to see to what extent our simplifications have led us astray, as well as to suggest ways in which the models need to be modified to make them more realistic. This is the role of experimental phylogenies, also known as biological simulations.

When we create an experimental phylogeny, we would like to control some aspects of evolution while we let the experimental organisms control the rest. For instance, if we are interested in the effects of differing branch lengths on the performance of phylogenetic methods, we might design a series of experimental trees in which we systematically vary branch lengths across trees for some experimental lineages, while we hold such factors as population size and environmental conditions constant. The biological constraints of the organisms are established by the organisms themselves, rather than modeled by an investigator. If we model the same trees and show that the results are consistent with the experimental lineages, then we can begin to conclude that the simplifications of the numerical simulations are not adversely affecting our conclusions. On the other hand, as will often be the case, we may see a difference between the simulated trees and the experimental trees. In this situation, we can evaluate the two data sets and determine why they are different. After doing so, not only will we now know more about the processes of molecular evolution, but we can also incorporate this information into new and better simulations. The new simulations may suggest new conditions to test experimentally, and the process can be repeated indefinitely, with the investigators learning more about the behavior of phylogenetic methods and the processes of molecular evolution with every cycle.

Of course, the scenario above assumes three things. First, we need to be able to define what we mean by "good performance" of a phylogenetic method. Second, we need to be able to identify and define relevant "parameter space" to explore in the numerical and biological simulations. Third, we need to identify organisms that evolve quickly enough that we can create experimental phylogenies in reasonable periods of time (hopefully measured in weeks or months rather than years).

How Do We Know a Good Method When We See One?

An ideal phylogenetic method would be fast, powerful, consistent, robust, discriminating, and versatile (see Penny, Hendy, and Steel, 1992). Unfortunately, there are trade-offs involved in optimizing these criteria, so that it is usually necessary to rank their importance in selecting a method for a given problem. Below we consider each criterion in turn.

Computational speed. If everything else were equal, computational speed would be very important. In general, however, the methods that rank the best for speed rank among the worst for some of the other criteria, and "quick-and-dirty" approaches are not often favored in science except as a way to get a first approximation. However, there is a wide spectrum of computational speeds among the methods from very fast single-tree clustering algorithms, to the character and distance-based methods that identify an optimality criterion, to methods such as maximum likelihood that require enormous computational efforts. Even though we may not want to select a method based on speed considerations alone, we still must select a method that is fast enough to give an answer without having to wait across geological time, and for some applications, a fast approximation may be appropriate.

Power. In the real world, we have a finite number of data that we can analyze for a given problem. If two methods are otherwise equal, but one correctly estimates a phylogeny from sequences 100 bp long whereas the other requires sequences 1,000 bp long to achieve the same success rate, then we would obviously prefer the one that requires fewer data to get the correct answer. Methods may differ in power because they consider different kinds of variation among the sequences to be informative, or because they give different weights to different kinds of variable characters. In the latter case, power may be a function of the model of evolution and the degree to which the assumptions of the method are matched.

Consistency. A method is consistent if it converges on the correct answer as more data are examined. All methods of phylogenetic analysis proposed to date are consistent under some conditions but are inconsistent under others. Some methods explicitly state a set of assumptions, which if violated may lead to inconsistency; for other methods, the assumptions are implicit and the conditions that lead to inconsistency have to be determined empirically. Many methods are based on a stated model of evolution, and as long as the organisms are evolving under the model conditions the method is consistent. Ideally, we would like to have a method that is consistent for the most general possible models of evolution.

Robustness. Even if we know the model conditions under which a method is consistent, it does not necessarily follow that deviations from the model will automatically lead to inconsistency. A robust method is insensitive to deviations from the ideal (model) conditions. This is obviously an important attribute, because it is unlikely that any real organisms ever evolve precisely in accord with any but the most general of models.

Discriminating ability. Methods should be able to return no answer if certain basic assumptions are violated (e.g., if there is no underlying tree), and they should be capable of comparing and ranking alternative hypotheses. Some methods (e.g., clustering algorithms like unweighted pair-group method of averages [UPGMA] and neighbor-joining) will always return one tree with no means of comparing alternatives, although tests can be applied to ask if any of the branch-lengths are significantly different from zero. Most other methods show limited ability to reject a tree-like structure but do specify an optimality criterion that can be used to compare and test alternative trees. Separate methods have been developed to identify data sets that contain no more structure than would be expected at random (e.g., Hillis and Huelsenbeck, 1992), and in principle such methods can be applied before deciding that it is appropriate to proceed to phylogenetic analysis. Once it has been determined that phylogenetic analysis is appropriate, a discriminating method should identify a range of potential solutions and provide a means of evaluating their optimality.

Versatility. The objectives of a phylogenetic analysis are usually more than simply finding the branching structure of a tree, although that is a universal first step. Some methods do little more than specify a branching structure, however. A versatile method would also provide such properties as estimates of branch lengths and estimates of the character states of the ancestral nodes in the tree. A method is also versatile if it is applicable to a wide range of character types (e.g., both molecular and morphological data); some methods are not versatile because they are applicable only to DNA sequences or incorporate only information about substitutional changes (e.g., information on insertion-deletion events is ignored).

Given that we accept the above criteria as important, how can we rate a given method for each criterion? Some of the rankings are straightforward, such as computational speed, but others (such as power and robustness) are harder to evaluate. Two main approaches can be used to compare the methods, namely numerical simulations and experimental phylogenies.

Numerical Simulations: Examining Conceivable Limits

A common objection to numerical simulations is that the conclusions of a typical simulation study invariably seem to support the investigator's a priori views on the relevant methods. For instance, one investigator who likes the neighbor-joining method (for whatever reason) may simulate phylogenies that indicate its superiority, whereas another investigator who prefers the UPGMA method may conduct simulations to show support for that approach. The reason for such discrepancies is that each of the methods has conditions under which it performs optimally, so a method looks best if trees are simulated under only those conditions. As an example, UPGMA performs best if rates of evolution are exactly the same in all lineages in the tree. Under such conditions, UPGMA can be shown to estimate the correct trees as well as or better than many other methods. However, the conclusions from a study that only includes such conditions are not very general and do not present a fair comparison of different methods. If simulations are to be used effectively, then, we need to define a specific problem for investigation and then examine the potential parameter space for that problem as exhaustively as possible.

As an example of defining a problem exhaustively, consider the simple and often-simulated "four-taxon tree with two rates" problem (Figs. 1 and 2). Felsenstein (1978) discussed this problem to demonstrate that some methods of phylogenetic analysis are inconsistent for some trees of this type; if the lineages represented by two opposing peripheral branches are evolving at a very high rate compared to the other three branches, then the parallel changes in the two long branches can overwhelm

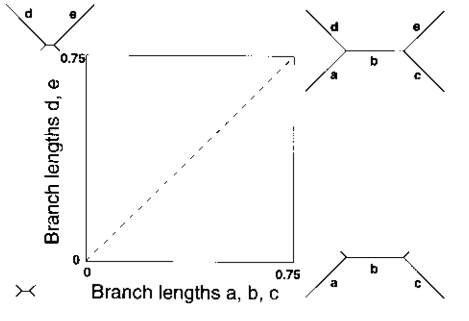


Figure 1. The four-taxon, two-rates problem. Consider a four-taxon tree in which two of the lineages (represented by the opposing branches *d* and *e*) have a different rate of change than the other three lineages (represented by branches *a*, *b*, and *c*). The instantaneous rate of change varies from 0 to infinity along the two axes of the graph, so that in the upper right corner of the graph all the characters of the taxa are as divergent as if chosen at random (e.g., the probability that there is a difference in any given nucleotide position in a DNA sequence is 0.75). Different areas of the resulting parameter space represent trees with different shapes. Along the diagonal (*dashed line*) the lengths of all five branches are equal. Near the top left corner of the graph, branches *a*–*c* are short and branches *d*–*e* are long, which produces conditions that are inconsistent for several methods of phylogenetic inference (see Fig. 2).

any signal in the small internal branch, leading one to be positively misled in estimating the tree (Fig. 2). Several methods are inconsistent under such conditions: the more data that are applied to the problem, the more likely the incorrect tree will be estimated. Because of this well-known behavior, and because of the simplicity of simulating and evaluating such trees, there have been numerous simulation studies of this problem (see summaries by Nei, 1991; and Huelsenbeck and Hillis, 1993). However, it is also simple to identify specific types of four-taxon trees that are particularly amenable to most of the common phylogenetic methods. Therefore, given that we have identified a particular problem (namely a four-taxon tree with two different rates), there is no reason not to examine the problem exhaustively for any given model of evolution. This is relatively easy to do in this case: we can graph out the two rates along two axes, and vary the instantaneous rate of evolution from zero to infinity along both axes (Fig. 1). We can then partition the graph space as finely as our computational limitations will permit, and simulate trees from throughout the entire possible parameter space. Using this approach, we can compare any set of methods for all potential conditions simultaneously, rather than only examining a biased set of trees that tends to support an a priori preference. If we are interested in consistency, we can calculate the expectations for infinitely large data sets; if we are interested in power, we can examine a regular series of finite data sets. A power analysis using this approach is illustrated in Fig. 3, with colors used to show the probability that a given method will find the correct tree in different areas of the parameter space.

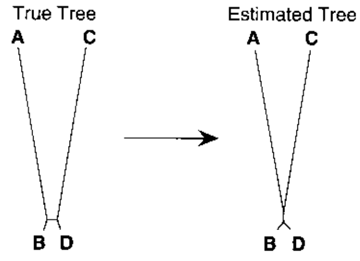


Figure 2. A tree from the Felsenstein zone: two of the opposing branches are long, and the other three branches (including the internal branch) are short. Parallel changes in the two long branches can be confused as phylogenetic signal, which leads some methods to estimate the incorrect tree on the right.

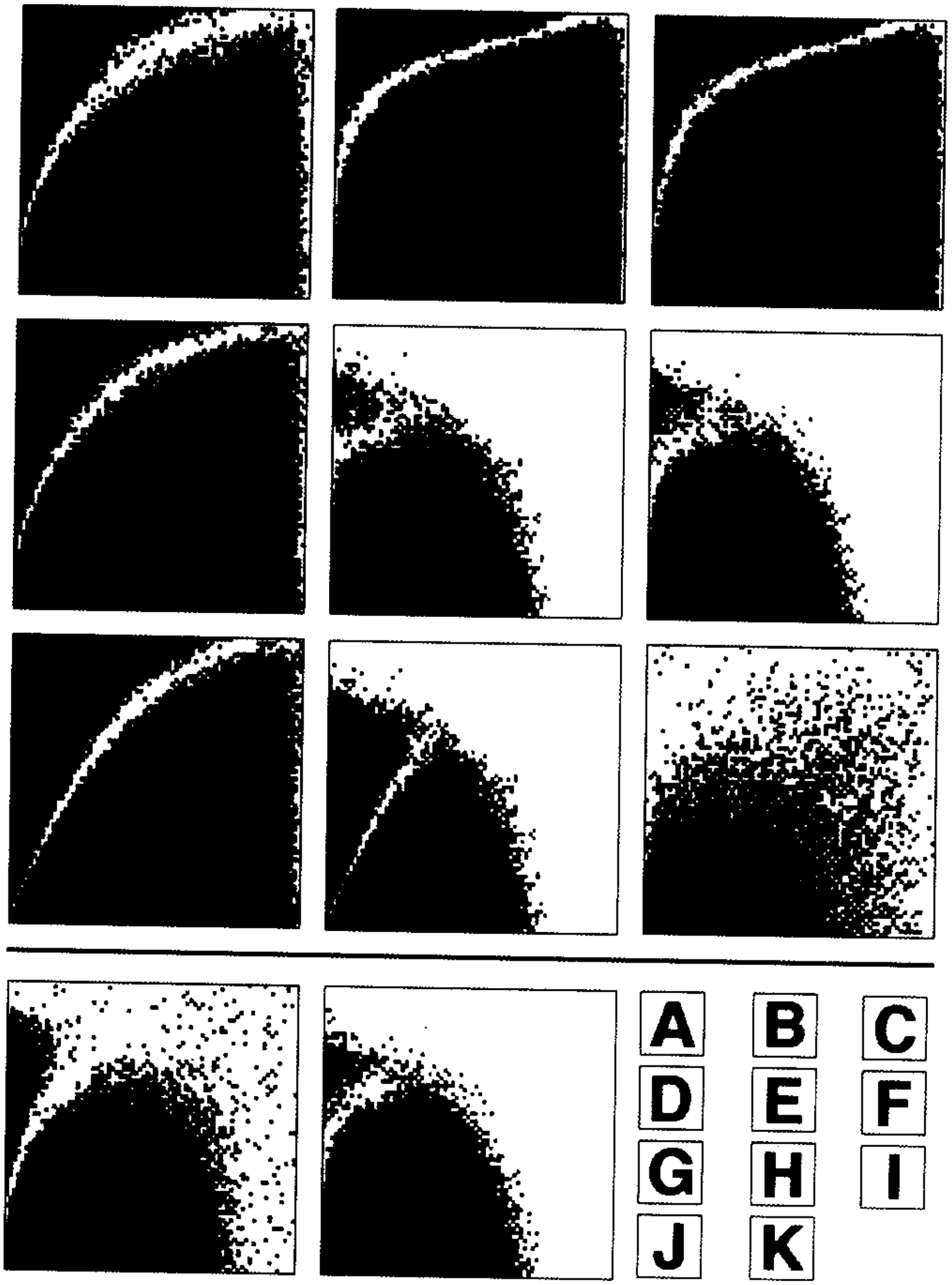


Figure 3. Comparison of the power of several methods of phylogenetic inference. The colors represent the probability of correctly estimating the phylogeny: green (>95%), dark blue (80–95%), light blue (60–80%), magenta (40–60%), yellow (20–40%), and red (<20%). White areas represent conditions in which over 90% of the data sets include undefined pairwise distances (so no tree can be constructed). In graphs *A–I*, DNA sequence evolution followed the Kimura model, with a 5:1 transition:transversion ratio. In graphs *J* and *K*, there was a 5:1 ratio of G–C and A–T changes compared to G–A, G–T, A–C, or C–T changes. 100

The model of evolution used to simulate the data sets analyzed in Fig. 3, *a–i,* is the simple Kimura model of nucleotide substitutions: there is one mutation rate for transitions and another for transversions, and in this case transitions are five times as common as transversions. This simple model is widely used because it approximates the pattern of evolution observed for many genes, in particular mammalian mitochondrial genes (e.g., see Brown, Prager, Wang, and Wilson, 1982). Parsimony (Fig. 3 *a*), UPGMA (Fig. 3 *g*), and neighbor-joining using uncorrected distances (Fig. 3 *d*) are inconsistent under these conditions for trees in the upper left-hand corner of the graph, a region sometimes called the Felsenstein zone (for details of the regions of inconsistency for most major methods under these conditions, see Huelsenbeck and Hillis, 1993). The performance of most of the methods shown falls off at high rates of change and near the Felsenstein zone. The performance of parsimony is increased dramatically at higher rates of evolution by weighting the transversions more heavily than the transitions, or by simply ignoring the transitions altogether (Fig. 3, *b* and *c*). Distance methods and parsimony can be made consistent by correcting the data in accord with the model of evolution. For this model, pairwise distance methods like neighbor-joining can be made consistent by correcting the distances as suggested by Kimura (1980); parsimony can be made consistent through a Hadamard transformation (see Penny et al., 1992). If the model of evolution matches the correction exactly, as shown in Fig. 3 *e* for neighbor-joining with Kimura distances, then many methods are consistent throughout the parameter space (Penny et al., 1992). However, note that these corrections may have a minimal effect on the power of the technique, except within the region of former inconsistency (e.g., compare Fig. 3, *d* with *e*). At high rates of change, the area in which neighbor-joining with Kimura-corrected distances finds the correct tree is still small compared to the weighted parsimony method (compare Fig. 3, *c* with *e*). Moreover, essentially the same level of performance can be achieved using the Kimura distances with the Fitch-Margoliash method as with neighbor-joining, with the added advantage of discriminating ability with the Fitch-Margoliash approach (compare Fig. 3, *e* with *f*). Using the Kimura corrections actually decreases the performance of the UPGMA method, which has little power in any case (Fig. 3, *g* and *h*). Lake's method of invariants (Fig. 3 *i*) shows an extreme trade-off between consistency and power: the method is consistent over the entire parameter space under these conditions, but has very low power. Interestingly, if we modify the model slightly by changing the classes of common versus rare substitutions, thereby violating the assumptions of Lake's method of invariants, the power of the method actually increases, even though it also become inconsistent in upper left corner of the graph (Fig. 3 *j*). In contrast, a similar change of models has a

variable nucleotide positions were included in all data sets. (*A*) Parsimony, all changes weighted equally; (*B*) transversion parsimony (transitions weighted zero); (*C*) weighted parsimony (transversions weighted five times more heavily than transitions); (*D*) neighbor-joining with uncorrected distances; (*E*) neighbor-joining with Kimura distances; (*F*) Fitch-Margoliash method with Kimura distances; (*G*) UPGMA with uncorrected distances; (*H*) UPGMA with Kimura distances; (*I*) Lake's method of invariants, all assumptions met; (*J*) Lake's method of invariants, mutation assumptions violated (see above); (*K*) neighbor-joining with Kimura distances, mutation assumptions violated (see above). For a description of all these methods, see Swofford and Olsen (1990).

comparatively little effect on the power of neighbor-joining (Fig. 3 *k;* although it too becomes inconsistent in part of the parameter space under these conditions).

The challenge in the field of numerical simulations is to identify ways that more complex problems can be explored in a similarly unbiased manner. This requires that we specify the purposes of a set of simulations explicitly: once the problem is defined, the relevant parameter space will often be obvious. Unfortunately, for many of the problems we face, the necessary computations will be much more difficult than in the simple four-taxon problem discussed above. As the number of taxa increases, the number of possible phylogenies increases rapidly, even if we ignore the lengths of branches. For rooted bifurcating trees, the number of distinct, labeled topologies for *n* species is equal to

$$\prod_{k=2}^{n}(2k-3)$$

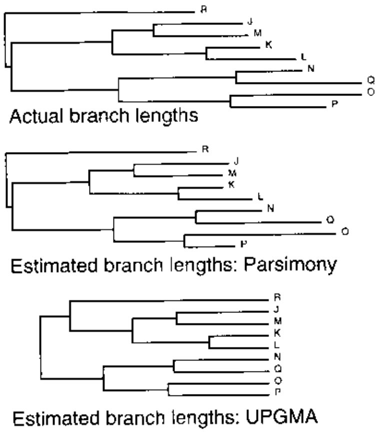
(Cavalli-Sforza and Edwards, 1967). Therefore, for just fifty taxa, there are 2.8×10^{76} possible rooted bifurcating trees. Obviously, we can not examine every possible solution in such cases; even with a computer that could examine 10^{50} trees a second, it would take much more time to examine all the trees than has existed in the history of the universe! We can get around this problem to a large extent by developing efficient heuristic searching methods (see Swofford and Olsen, 1990; Swofford, 1993), but the harder problems will still require major increases in computational power, such as those afforded by massively parallel computing (see W. D. Hillis and Boghosian, 1993).

Experimental Phylogenies: Observing Evolution in Action

The principal limitation in experimental phylogenies is producing the relevant levels of evolutionary change within a human life span (or more realistically, within the life span of a funded research project). This requires an organism with a short generation time and high mutation rate. Ideally, we would like to be able to control the spectrum of potential mutations, have the ability to examine a large fraction of the organism's genome (to avoid or explore problems of sampling bias), control population size over a wide spectrum, control and manipulate the organism's environment with ease (to examine the effects of selection), and culture the organism without great difficulty. All of these conditions are met most closely by viruses, and particularly bacteriophages, the viruses of bacteria.

Fig. 4 shows an actual and two estimated trees for an experimental phylogeny based on cultures of the bacteriophage T7 (see Hillis, Bull, White, Badgett, and Molineux, 1992). The T7 cultures were grown in the presence of a mutagen (in this case, nitrosoguanidine) to increase the rate of mutations. The mutational spectrum produced in such an environment is biased in favor of certain substitutions, but deletions occur as well. With nitrosoguanidine, many kinds of mutations occur, but the majority are transitions, especially G → A and C → T. This spectrum of mutations is similar to that observed in some natural systems that have been examined (e.g., mammalian genes and pseudogenes: Gojobori, Li, and Graur, 1982; Li, Wu, and Luo, 1984; human immunodeficiency viruses: Moriyama, Ina, Ikeo, Shimizu, and Gojobori, 1991). To create the phylogeny, an initial culture of bacteria plus bacteriophage is divided into three lineages (one of which will become the outgroup for purposes of rooting the tree), and then the lineages are redivided after a predetermined number of lytic cycles. Stocks of ancestral cultures are saved at each cycle for later analysis. After the phylogeny has been created, DNA is isolated from the mutated lineages of T7 and analyzed by restriction digestion, sequencing, or DNA-DNA hybridization. Different methods of phylogenetic analysis can then be used to estimate the phylogeny, the branch lengths, and the genotypes of the ancestors, and in each case evaluated against the true tree and actual ancestors.

What are the results of such experiments? An analysis of a set of just over 200 restriction sites that are variable among the lineages produces the correct phylogeny with almost every method tested to date. However, if one uses the presence or absence of restriction fragments as the primary data, instead of mapping out the sites for analysis, then every method returns the incorrect phylogeny. Restriction fragment analyses are troublesome because restriction fragments are not independent characters. There are two reasons for this. First, a single site gain causes one fragment to be lost and two other fragments to be gained, and second, a single deletion can cause changes to multiple fragments. However, it is often argued that mapping and aligning restriction sites may not be worth the extra effort (e.g., Bremer, 1991). Yet, in the seemingly ideal situation of the experimental T7 phylogeny, restriction sites always give the correct tree and restriction fragments always give the wrong tree, indicating that the problem with nonindependence is a severe one. This result has not been obvious from simulations, probably because simulations rarely include insertions or deletions in the models of nucleotide change. This result indicates that we need to improve our simulation models to include realistic frequencies of insertion-deletion events.



Actual branch lengths

Estimated branch lengths: Parsimony

Estimated branch lengths: UPGMA

Figure 4. An actual and two estimated trees for an experimental phylogeny derived from bacteriophage T7 (Hillis et al., 1992). The branch lengths are drawn proportional to the number of actual or estimated restriction-site changes. Both estimated trees produce the correct unrooted topology (the root is ambiguous, because the ancestral node is a trichotomy), but parsimony produces much better estimates of branch lengths than does UPGMA.

As discussed earlier, the branching structure is not the only aspect of the tree we would like to estimate. Branch lengths are also of interest, as are the estimation of ancestral genotypes. For this phylogeny, as well as for simulations based on the kinds and distributions of mutations observed across the tree (Bull, Cunningham, Molineaux, Badgett, and Hillis, 1993), the character-based parsimony method produces significantly better estimates of branch lengths (estimated number of restriction-site changes) than do pairwise distance methods such as neighbor-joining or the Fitch-Margoliash method, and these latter methods produce significantly better estimates than does UPGMA. Fig. 4 compares the actual branch lengths (in number of restriction-site changes) with the estimated branch lengths from parsimony and UPGMA.

Of the common methods of phylogenetic analysis, only parsimony provides estimates of the ancestral character states at internal nodes in the tree (see Maddison and Maddison [1993] for an excellent discussion of the reconstruction of ancestral character states). How well does this method fair in this example? For any given restriction site, the estimated state in an ancestor may be (*a*) correctly inferred, (*b*) incorrectly inferred, or (*c*) ambiguous (i.e., the presence or absence of the site is equally parsimonious). When we compared the inferred restriction maps to the actual restriction maps for the seven ancestral nodes in the model phylogeny, we found that the inferred restriction maps were >98% correct. This finding indicates that the method for reconstructing ancestral character states is working well under the conditions tested.

To what extent can we generalize these findings from a study of bacteriophage to other organisms? Phylogenetic methods are purported to be general for living organisms, so any organism should therefore provide a test of the fallibility of the methods. Some findings, such as the efficiency of reconstructing ancestral genotypes, will be dependent on the level of homoplasy (reversals, parallelisms, convergences) present across the phylogeny. In the case presented above, the level of homoplasy present in the tree was almost exactly the average of that seen in real examples in the literature for the same number of taxa and characters (Hillis, Bull, White, Badgett, and Molineux, 1993; data on other organisms in Sanderson and Donoghue, 1989). Obviously, there is a need to examine other tree topologies, the effects of highly unequal branch lengths on the performance of the various methods, the effects of parallel selection pressures on different parts of the tree, the effects of other mutational spectra, the effects of population size, and many other aspects of evolution that potentially affect our ability to reconstruct phylogenies. This is a strength of experimental phylogenies: all of these aspects of phylogenies can be studied, without detailed a priori knowledge of the underlying mechanisms. In fact, the studies also can provide first-hand information on the processes of molecular evolution. Experimental phylogenies can (and do) provide actual examples of the failure of certain methods, as in the use of restriction fragment analysis in the example above. Such findings can be used to improve the methods or to identify conditions under which they should not be applied.

How Do the Major Methods Rate?

It should be clear that no single method is optimal for all of the criteria we have identified. As with most things in life, there are trade-offs. Thus, the methods that are fastest produce just a single estimate of the tree, without any obvious way to compare or rank the alternatives (i.e., they have very low discriminating ability); fast methods also tend to be less powerful. Methods that are fully consistent under a specific model of evolution are usually less versatile (e.g., they only apply to certain kinds of data). We have attempted to rank the most commonly used methods for the assessment criteria identified earlier (see Table I). These rankings are somewhat subjective, and we recommend that interested readers refer to the primary literature on simulations and experimental phylogenies to make their own rankings. Also note that these methods are constantly being modified and refined, so the rankings may change over time. This is particularly true for the maximum likelihood methods, which have increased dramatically in speed and versatility in recent years (our rankings for maximum likelihood reflect the current common implementations rather than the full potential of the method).

Which method is finally selected will depend to a large extent on the goals of the study: neighbor-joining is perfectly adequate for a fast, initial estimate of a tree, even though it is useless for comparing and ranking alternative solutions. For most studies, one of three classes of methods will usually be most appropriate: Fitch-Margoliash and related methods (including the minimum evolution method), parsimony methods, or maximum likelihood. Fitch-Margoliash and related methods specify an optimality criterion for fitting observed pairwise distances to path-length distances on trees. The optimality criterion can be assessed for any tree, and the methods are relatively insensitive to variations in branch lengths across the tree. They are currently somewhat limited in their power and versatility by the simplistic models that are used to compute evolutionary distances (e.g., it is difficult to incorporate different weights for different characters or types of character change into the analyses, and it is difficult to combine analyses of different kinds of data into a single analysis). Parsimony methods correct these deficiencies, but at a cost of lower overall consistency for simple models of evolution. Parsimony methods are the most versatile approaches: they can be applied to all kinds of data, can easily incorporate differential weights for different character-state changes or for entire characters, are amenable to combination of results among studies, and provide accurate reconstructions of branch lengths and ancestral character states. As shown in Fig. 3, this versatility (e.g., character-state weighting) can lead to increased power. Although parsimony methods are consistent under more limited conditions than some other approaches, this limitation has been lifted to some extent by recent developments (for more information, see discussion of the Hadamard transformation in Penny et al., 1992). Finally, maximum likelihood methods are likely to undergo the most development in coming years. Their use has been limited because they currently are too slow to be applied to many real world data sets, and because they are not very versatile (for instance, they currently are limited primarily to substitutional changes

TABLE I
Comparison of the Most Commonly Used Methods of Phylogenetic Analysis

Method	Criterion					
	Speed	Power	Consistency	Robustness	Discrimination	Versatility
UPGMA	++	−	−	−	−	−
Neighbor-joining	++	+	++	+	−	+
Fitch-Margoliash (and related methods)	+	+	++	+	++	+
Parsimony methods	+	++	+	+	++	++
Methods of invariants	+	−	++	+	++	−
Maximum likelihood (as currently implemented)	−	++	++	±	++	±

Key: ++, excellent; +, good; and −, poor.

in DNA sequences). However, their versatility is being constantly improved, and they are useful methods when they can be applied because of their high power, consistency, and discriminating ability.

References

Bremer, B. 1991. Restriction data from chloroplast DNA for phylogenetic reconstruction: Is there only one accurate way of scoring? *Plant Systematics and Evolution.* 175:39–54.

Brooks, D. R., and D. A. McLennan. 1991. Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology. University of Chicago Press, Chicago. 434 pp.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences in primates: tempo and mode of evolution. *Journal of Molecular Evolution.* 18:225–239.

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

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution.* 21:550–570.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology.* 27:401–410.

Gojobori, T., W.-H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution.* 18:360–369.

Harvey, P. H., and M. D. Pagel. 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford. 239 pp.

Hillis, D. M., and D. M. Green. 1990. Evolutionary changes of heterogametic sex in the phylogenetic history of amphibians. *Journal of Evolutionary Biology.* 3:49–64.

Hillis, D. M., and C. Moritz, editor. 1990. Molecular Systematics. Sinauer Associates, Inc., Sunderland, MA. 588 pp.

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., and J. P. Huelsenbeck. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity.* 83:189–195.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1993. Experimental approaches to phylogenetic analysis. *Systematic Biology.* 42:90–92.

Hillis, W. D., and B. M. Boghosian. 1993. Parallel scientific computation. *Science.* 261:856–863.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology.* 42:247–264.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide, sequences. *Proceedings of the National Academy of Sciences, USA.* 78:454–458.

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. *Journal of Molecular Evolution.* 21:58–71.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: Analysis of Phylogeny and Character Evolution. Sinauer Associates, Inc., Sunderland, MA. 398 pp.

Moriyama, E. N., Y. Ina, K. Ikeo, N. Shimizu, and T. Gojobori. 1991. Mutation pattern of human immunodeficiency virus genes. *Journal of Molecular Evolution.* 32:360–363.

Nei, M. 1991. Relative efficiencies of different tree-making methods for molecular data. *In* Phylogenetic Analysis of DNA Sequences. M. M. Miyamoto and J. Cracraft, editors. Oxford University Press, Oxford.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred from various techniques. *Cold Spring Harbor Symposia in Quantitative Biology.* 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.

Penny, D., M. D. Hendy, and M. A. Steel. 1992. Progress with methods for constructing evolutionary trees. *Trends in Ecology and Evolution.* 7:73–79.

Sanderson, M. J., and M. J. Donoghue. 1989. Patterns of variation in levels of homoplasy. *Evolution.* 43:1781–1795.

Swofford, D. L. 1993. PAUP 3.1: Phylogenetic Analysis Using Parsimony. Illinois Natural History Survey, Champaign, IL.

Swofford, D. L., and G. J. Olsen. 1990. Phylogeny reconstruction. *In* Molecular Systematics. D . M. Hillis and C. Moritz, editors. Sinauer Associates, Inc., Sunderland, MA.