

# Points of View

*Syst. Biol.* 42(1):85–89, 1993

## Experimental Tests of Phylogenetic Inference Methods

ELLIOTT SOBER

*Philosophy Department, University of Wisconsin,  
Madison, Wisconsin 53706, USA*

Evolutionary biologists have described several methods for inferring the genealogical relationships that obtain within a set of taxa. Because these methods often disagree about which phylogenetic hypothesis is best supported by a given data set, a lively controversy has ensued concerning the merits of alternative methods. The debate has been difficult to resolve, in part because scientists cannot independently determine which method of inference has identified the true tree in a given problem. For this reason, it has been impossible to assess the overall "track records" of the competing methods.

Hillis et al. (1992) tried to cut this Gordian knot; they ran an experiment in which the reliability of different inference methods could be checked against an independently known phylogeny. In their experiment, they created a phylogeny of lineages derived from bacteriophage T7 and obtained restriction maps of genomes in different lineages. Restriction site data on the terminal taxa were then used with different methods of phylogenetic inference to see what trees would result. The methods tested were parsimony (Swofford, 1990), the method of Fitch and Margoliash (1967), the method of Cavalli-Sforza and Edwards (1967), neighbor-joining (Saitou and Nei, 1987), and UPGMA—the unweighted pair-group method of arithmetic averages (Sokal and Michener, 1958).

Hillis et al. reported that all the methods correctly retrieved the true topology, al-

though they varied in how well they estimated branch lengths. The authors suggested (Hillis et al., 1992:591) that their experiment points

the direction to a new field of research. . . . Experimental phylogenetics is not a substitute for numerical studies, nor is it likely that laboratory phylogenies will ever display the full complexity of phylogenies produced over long-term evolution, but such studies will fill an important void in the science of phylogenetic reconstruction.

The purpose of this paper is to estimate the size and character of this void.

I begin with a discussion of what their experiment means for the problem of inferring tree topologies; then I consider the problem of inferring branch lengths and the wider issue of the research program of experimental phylogenetics.

The experimental result provides no help in choosing among different methods for inferring tree topologies. Although the methods agreed on which topology is best in the experiment, this agreement is hardly typical of inference problems presented by nature.

If methods frequently disagree when applied to nature, how should one interpret the assertion that "the results of this study directly support the legitimacy of methods for phylogenetic estimation" (Hillis et al., 1992:591)? Although the authors did not indicate what they meant by this positive assessment, one possibility is that the experiment provides evidence that these

methods retrieve true phylogenies in nature when the methods all agree.

To determine whether the methods' joint success in the experiment indicates that they will probably succeed when they agree on a problem drawn from nature, one must discover whether the reason for their experimental success can be expected to apply to a problem drawn from nature. To do this, a model of the evolutionary process must be developed and shown to have three properties. First, the model's assumptions must apply to the processes in the experiment. Second, the model must make it probable that the methods all will retrieve the true phylogeny when supplied with a data set that is sufficiently large. This second requirement relates to the asymptotic property that statisticians term consistency (Felsenstein, 1978; Sober, 1988). A model with these two properties explains why the methods were able to retrieve the true phylogeny in the experiment. Third, the model also must be a realistic description of evolutionary processes in nature. If all three criteria can be met, one may conclude that the success of the methods in the experiment is evidence that they will succeed when applied to nature.

A reasonable model of the processes at work in the experiment can be extracted from what Hillis et al. (1992) reported. The authors noted that they "chose to construct a symmetric phylogeny with equal distances among nodes"; they pointed out that "the number of changes per ingroup branch is not significantly heterogeneous from the expectation under a Poisson distribution" (Hillis et al., 1992:590).

Figure 1 locates the experimental process in a wider space of possibilities. The tree in Figure 1a obeys a uniform rate but not a constant rate model; although the probability of change<sup>1</sup> is the same across contemporaneous lineages, it is not the same among lineages at different times that have the same duration. The trees in Fig-

ures 1b and 1c both obey the constant rate assumption but differ in another respect. Though the probability of change per unit time is constant in Figure 1b, the amount of change that can be expected to accumulate differs among the various ingroup branches because the branches have unequal durations. The branch from D to its immediate ancestor H is "longer" than the branch from A to its immediate ancestor I. In contrast, the ingroup branches in Figure 1c have equal duration and share a constant probability of change per unit time; this means that the ingroup branches have a constant probability of change. The experiment of Hillis et al. (1992) appears to conform to the pattern in Figure 1c, not to the patterns in Figures 1a or 1b, because the branches were of equal duration and evolved under identical experimental conditions.

How do the issues of constant rates and equal durations between nodes affect the question of whether a method will be statistically consistent? With four or fewer taxa, a uniform rates model guarantees that parsimony and distance methods will be statistically consistent (Sober, 1988; Hendy and Penny, 1989). When rates are nonuniform, the methods can fail to be consistent (Felsenstein, 1978; Hendy and Penny, 1989), and parsimony often is more sensitive to unequal rates than are many distance methods. When rates are constant and there are more than four taxa, constant rates do not suffice for parsimony to be consistent (Hendy and Penny, 1989). Hendy and Penny (1989) conjectured that with more than four taxa, the combination of uniform rates and equal branch durations in the ingroup suffices to guarantee consistency. Supposing that this conjecture is reasonable, we may have an explanation for why the methods all converged on the true phylogeny in the experiment. Because the problem involved nine terminal objects, the methods converged because the true phylogeny conformed (to a high degree, if not exactly) to the pattern in Figure 1c.

The next question is whether this pattern frequently carries over to inference problems drawn from nature. Systematists

<sup>1</sup> The probability of change is not the probability of various numbers of changes within a branch. Rather, it is the probability that a branch will end in state *i*, given that it begins in state *j* (where *i* ≠ *j*).

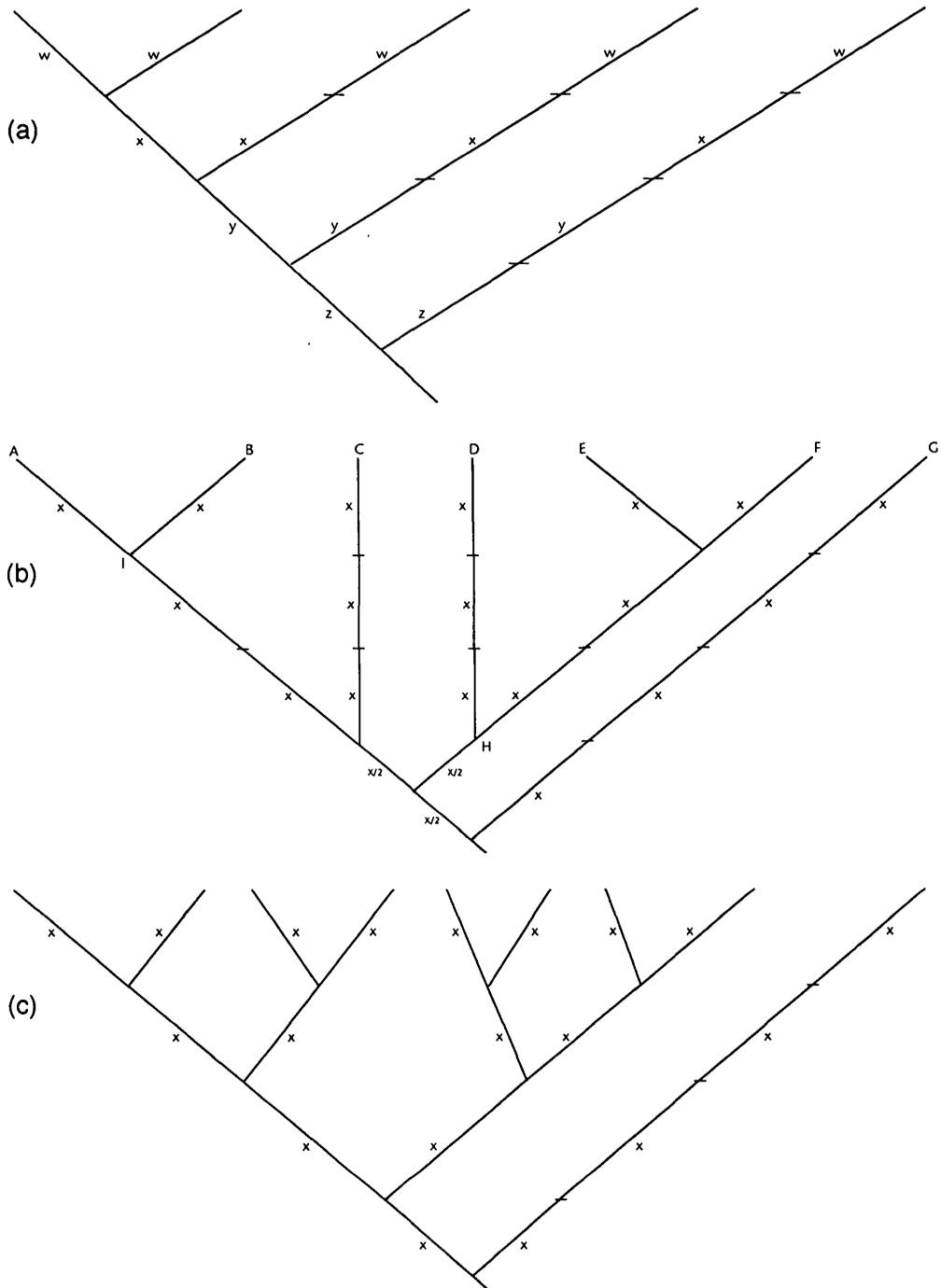


FIGURE 1. (a) Tree exhibiting uniform but not constant rates of evolution because  $w$ ,  $x$ ,  $y$ , and  $z$  are different. (b) Tree exhibiting both uniform and constant rates. Unequal branch durations and a constant probability of change per unit time together entail that branches have different probabilities of change. (c) Tree exhibiting both uniform and constant rates. Ingroup branches have the same probabilities of change because they are of equal duration and have the same probability of change per unit time.

have often been reluctant to regard uniform (or constant) rates as a model with general validity; the same can be said of the idea of constant ingroup branch durations. It is not clear why the success of these inference methods in the laboratory experiment provides much assurance that they can be expected to retrieve the true phylogeny in nature.

My argument has relied on the idea that one needs a model of the evolutionary process to assess whether laboratory success strongly supports the idea that a method will succeed in the wild. However, some may reply that systematists do not need a model of natural or laboratory processes to be confident that laboratory experiments are similar to natural processes. If this were correct, it would make sense to run an experiment and to extrapolate the result to nature without addressing the questions I have posed.

The problem with this reply is that it is difficult to determine whether natural processes and laboratory processes are relevantly "the same" without knowing which points of similarity and difference actually influence the performance characteristics of various phylogenetic inference methods.<sup>2</sup> There are many similarities and many differences between the laboratory and the wild; to see which of these matters to the epistemological problem at hand requires that one pose the questions I have enumerated. Systematists who reject the idea that organisms should be grouped atheoretically also should reject the idea that experiments can be interpreted atheoretically.

An experiment can be seen to refute certain unconditional claims (e.g., the claim that a method never retrieves the true phylogeny) without involving an explicit model. But more measured claims about the strengths and weaknesses of various methods cannot be assessed in this way.

---

<sup>2</sup> In the experiment, all the methods agree about which topology is best; however, for data sets drawn from nature, methods often disagree. Perhaps this is *prima facie* evidence that the processes in the two settings are relevantly different.

Hillis et al. (1992) did not explicitly advocate an atheoretical interpretation of their experiment; yet they drew conclusions about the legitimacy of methods for phylogenetic estimation even though they did not theorize about why the methods tested retrieved the true phylogeny.

Mathematical arguments have shown that parsimony and phenetic methods are statistically consistent in some settings but not in others. For this reason, an experiment that could empirically determine whether or under what circumstances rates are constant and branches have equal durations would be quite useful. However, an experiment designed to produce constant rates and equal branch durations does not do this. Furthermore, an experiment that detects rate inequalities or differences in branch durations need not compare how different methods of inference handle the data set obtained.

My conclusion, then, is not that laboratory experiments are useless in terms of studying the problem of phylogenetic inference applied to nature. These experiments may play a heuristic role, suggesting kinds of natural processes that make it probable (or improbable) that various methods will retrieve true phylogenies. A model may show that the success of an inference method depends on some feature of the evolutionary process, which one may seek to measure experimentally.

The problem of determining whether an inference method will probably retrieve true phylogenies from natural data sets has two parts. First, there is the empirical task of finding a realistic model of natural processes. Second, there is the mathematical problem of discovering whether the model makes it probable that the method will retrieve the true phylogeny when supplied with enough data. Neither of these problems is solved by observing that the method happens to retrieve the true phylogeny in an experiment.

I have concentrated so far on the question of how one should interpret the fact that all the methods tested in the experiment retrieve the same true tree topology. However, Hillis et al. (1992) also discov-

ered that parsimony did better than other methods in reconstructing branch lengths and that parsimony did quite well with the problem of inferring the character states of ancestors. The question of whether this result should inspire confidence in parsimony's ability to perform these tasks on data drawn from nature leads to the same problems I have discussed. Some conclusions about the meaning of the experiment can be drawn without an explicit process model. It takes no model to see that parsimony does not always fail; nor is a model needed for one to see that parsimony is not always inferior to the methods with which it competes. But a model is needed to decide whether parsimony's superior performance in the laboratory strongly supports the expectation that it will also excel in nature.

Hillis et al. (1992) not only provided an interpretation of a single experiment; they also announced the beginning of a research program. Although I have criticized the interpretation, these criticisms do not automatically entail that the research program will be unproductive. I make no such claim. Research programs are flexible things; they do not stand or fall with the result of a single experiment.<sup>3</sup> Although I am skeptical that the results of this experiment "directly support the legitimacy of methods for phylogenetic estimation," it

<sup>3</sup> This point is developed in detail in connection with the adaptationist research program in Orzack and Sober (unpubl. manuscript) and in Sober (1993).

remains to be seen what experimental phylogenetics can teach us about the problem of phylogenetic inference.

#### ACKNOWLEDGMENTS

I thank James Crow, Joe Felsenstein, Arnold Kluge, Gregory C. Mayer, Michael Miyamoto, Robert J. O'Hara, and an anonymous referee of this journal for useful suggestions.

#### REFERENCES

- CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.* 19:233-257.
- FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401-410.
- FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. *Science* 155:279-284.
- HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38:297-309.
- HILLIS, D. M., J. BULL, M. E. WHITE, M. R. BADGETT, AND I. J. MOLINEUX. 1992. Experimental phylogenetics: Generation of a known phylogeny. *Science* 255:589-592.
- SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
- SOBER, E. 1988. *Reconstructing the past: Parsimony, evolution, and inference.* Massachusetts Institute of Technology Press, Cambridge.
- SOBER, E. 1993. *Philosophy of biology.* Westview Press, Boulder, Colorado.
- SOKAL, R. R., AND C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 28:1409-1438.
- SWOFFORD, D. L. 1990. PAUP: Phylogenetic analysis using parsimony, version 3.0. Computer program distributed by the Illinois Natural History Survey, Champaign.

*Received 12 May 1992; accepted 19 September 1992*