

## LETTERS

## Likelihood and Inconsistency

James S. Farris

*Molekylärsystematiska laboratoriet, Naturhistoriska riksmuseet, Box 50007 S-104 05 Stockholm, Sweden*

Accepted April 28, 1999

Parsimony can be inconsistent, but not maximum likelihood—likelihood advocates often say. This difference and conclusions drawn from it have provided the main reasons advanced by likelihoodists against the use of parsimony. Recent statistical research, however, shows that maximum likelihood estimation of phylogenetic trees can become inconsistent in all but the simplest cases, so that under realistic conditions the consistency of maximum likelihood cannot be assured. If likelihoodists wish to dispose of parsimony, they will have to find another argument. © 1999 The Willi Hennig Society

## INTRODUCTION

Felsenstein (1978) described a hypothetical case—now famous as the Felsenstein Zone—in which parsimony would be statistically inconsistent, that is, would not give the right tree even if the data comprised an indefinitely large random sample of characters. In contrast, he emphasized (p. 408, italics added)

Methods of phylogenetic inference which *entirely avoid the problem of statistical inconsistency* are already known. Maximum likelihood is one of them. I have outlined elsewhere (Felsenstein, 1973) how this may be done.

That advantage of maximum likelihood has been

widely cited as a reason to abandon parsimony. More recent statistical research, however, has shown that Felsenstein's claim was incorrect, as I will discuss here.

## PROOFS

To be sure, Felsenstein (1973: 246) reported a proof that maximum likelihood estimation of phylogenetic trees would entirely avoid inconsistency, but he omitted some details:

An estimate has the property of *consistency* if, as we sample more and more data, the estimate converges to the true value. Maximum likelihood estimates have this property under a wide variety of circumstances (Wald, 1949). Wald gives eight conditions which, if all are satisfied, guarantee that the maximum likelihood estimate is consistent. *These conditions are too complex to discuss here in detail* [italics added], but it can be shown that estimates based on the likelihood expressions (4), (5), and (6) satisfy them, so that these estimates are consistent. Expression (4) is used to estimate the tree topology.

Not all authors have found those conditions too complex to discuss. For example (Yang, 1996: 304)

Felsenstein (1973, 1978) referred to the regularity conditions of Wald (1949) for a proof of the consistency of the maximum likelihood method for estimating the tree topology. These conditions would include the continuity and differentiability of the

likelihood function with respect to the topology parameter. Such concepts are not defined.

Not defined, because topologies are nonnumerical and discrete rather than continuous. Maximum likelihood estimation of trees did *not* satisfy Wald's conditions.

This does not mean that maximum likelihood can never be consistent, but the method does not retain consistency under the "wide variety of circumstances" that Felsenstein suggested. Felsenstein's (1978) model included a restrictive homogeneity assumption: all characters (sites) were required to evolve according to the same stochastic process, in which case all would have the same substitution rate. The phylogenetic tree could be estimated consistently *if* that assumption (with some others) were true (Chang and Hartigan, 1991; Steel *et al.*, 1993; Rogers, 1997), but that is not enough to establish Felsenstein's claim that maximum likelihood will "entirely avoid the problem of statistical inconsistency." As Felsenstein (1978: 408) pointed out himself

The models employed here certainly have severe limitations. It will hardly ever be the case that we sample characters independently, with all of the characters following the same probability model of evolutionary change.

That departure from reality can have an unfavorable effect on maximum likelihood estimation (Swofford *et al.*, 1996: 442):

The maximum likelihood models described above all assume that every site evolves at the same rate. Violation of this assumption can have devastating consequences. For instance, Gaut and Lewis (1995) showed that maximum likelihood inference under the assumption of rate homogeneity can become inconsistent when the true evolutionary process exhibits site-to-site rate variation. . . . Thus, maximum likelihood can become "positively misleading" [Felsenstein, 1978] for exactly the same reasons as parsimony.

Swofford *et al.* (1996), like most other advocates of maximum likelihood, nonetheless thought that the method could still entirely avoid inconsistency, by employing more realistic models. They realized that achieving consistency requires an accurate model (p. 427):

Farris's [1983, 1986] point that a maximum likelihood method will guarantee consistency only if evolution proceeds according to the assumed model is of course true.

But otherwise they saw no difficulties in principle in

guaranteeing consistency: if present models are inadequate, they can always be improved.

As it turns out, however, there is a limit to such improvement. This was discovered by Steel *et al.* (1994: 157), whose Theorem 3 shows

If none of the conditions (i)–(iii) [to be listed shortly] hold, then [the tree]  $T$  may no longer be determined by its sequence spectrum—indeed each tree, with an associated  $\vartheta = \vartheta_T$ , can induce an identical sequence spectrum.

Here  $\vartheta$  denotes a distribution of rates: each site is considered to have a substitution rate drawn at random from the same distribution.  $\vartheta_T$  is a rate distribution chosen to fit tree  $T$ . A sequence spectrum is a probability distribution on patterns, a pattern being a possible assignment of character states to the terminals. In any evolutionary model, the sequence spectrum is a function of the tree (*inter alia*). That this relationship be *invertible* (Steel *et al.*, 1993)—that the tree be uniquely determined by the sequence spectrum—is a necessary condition for maximum likelihood estimation of the tree to be consistent. Consequently, unless one of the conditions (i)–(iii) is satisfied, consistency is not assured *even if the model is accurate*. Those conditions are (same page)

- (i)  $\vartheta$  is known.
- (ii)  $\vartheta$  is unknown, but it has positive measure [frequency] only on 0 and one other (unknown) value [i.e., rate].
- (iii)  $\vartheta$  is neither known nor constrained, but we assume the molecular clock hypothesis.

In practice  $\vartheta$  is not known, but must be estimated (fitted), and the molecular clock is hardly realistic, so that only condition (ii) applies to real cases. This result sharply restricts the variety of circumstances under which maximum likelihood can guarantee consistency. For example, Swofford *et al.* (1996: 443f), aware that substitution rates vary among sites, recommended a technique in which the fraction of invariable sites is estimated, while a discretized gamma distribution is fitted to the rates of the remaining sites. There are thus multiple non-0 rate values, and this is more than condition (ii) covers, so that the consistency of this procedure is *not* assured under Theorem 3.

Even this is not the full extent of the problem. The type of rate variation considered in Theorem 3 is relatively simple. While sites may have different substitution rates, the ratio in rates between two sites would be the same in all branches of the tree. Of course there

is no reason why that ratio cannot vary between branches, and Chang (1996) has demonstrated that such variation can cause maximum likelihood estimation of the tree to become inconsistent, even in cases that otherwise conform to Felsenstein's (1978) simple model.

Here is an example (cf. Chang, 1996, his Table 1). Like Felsenstein's (1978), it involves 0/1 characters, but similar conclusions apply to nucleotide sequences. The correct unrooted tree for terminals A–D is (AB)(CD). There are two classes of characters, with different suites of branch lengths:

	A	B	C	D	X
Class 1	0.5	2.0	0.5	2.0	0.5
Class 2	2.0	0.5	2.0	0.5	0.5

X denotes the interior branch. Within each class, characters evolve according to Felsenstein's (1978) reversible<sup>1</sup> model (F78). Note that Felsenstein's branch substitution probabilities  $P$  are related to Chang's branch lengths  $L$  by

$$P(L) = (1 - e^{-2L})/2.$$

The total population of characters is a 50/50 mixture of the two classes. Under these conditions, maximum likelihood estimation based on F78 is inconsistent: if the number of characters sampled is large enough, the method is certain to yield tree (AC)(BD) instead of the right one.

While the details of this example are artificial, there is nothing unrealistic in the idea that the relative substitution rates of different characters (sites) may vary among parts of the tree. Under the older name of mosaic evolution, that kind of rate variation has long been regarded as a widespread phenomenon. Yet there seems to have been no attempt to take such effects into account in any maximum likelihood method. Unless such a method can be devised and proved consistent, mosaic evolution will pose another barrier to guaranteeing the consistency of maximum likelihood estimation.

<sup>1</sup>Felsenstein (1978) also had a model that assumed irreversibility.

## MODELS

These results make it clear that attempting to guarantee consistency of maximum likelihood by further improvements in models will not work. Even relatively simple models have already gone beyond the rudimentary level of complexity that would allow consistency to be assured, while at the same time there are realistic departures from homogeneity that likelihood methods do not address. Contrary to Felsenstein's (1978) claim, maximum likelihood certainly does not "avoid the problem of statistical inconsistency."

Advocates of maximum likelihood have nonetheless often identified inconsistency with parsimony, and this premise has been carried over into discussions of other issues. Swofford *et al.* (1996: 426) used it to argue for the usefulness of models:

Considerable disagreement exists as to whether the "model-free" nature of parsimony [as opposed to the model-specific nature of maximum likelihood] is an advantage or a disadvantage. . . . Examination of the conditions under which parsimony is an inconsistent estimator will be helpful in understanding the usefulness of explicit evolutionary models.

By that reasoning, the inconsistency of maximum likelihood in the mosaic evolution example would presumably imply that explicit evolutionary models are not useful! What it does illustrate, of course, is the danger of relying on an oversimplified model, but for just that reason, the implication is true in a sense. The claimed advantage of models has always been that they provide valuable information to the analysis by describing the evolutionary process—whence the need for realism (Felsenstein, 1978: 409):

The weakness of the maximum likelihood approach is that it requires us to have a probabilistic model of character evolution which we can believe.

Yet the models actually used have hardly been selected for their accuracy (Felsenstein, 1993, in *dnaml.doc*):<sup>2</sup>

This rather disconcerting model is used because it has nice mathematical properties which make likelihood calculations far easier.

If models are made realistic, however, model-specific

<sup>2</sup>This is not an isolated case. Dietrich (1994), for example, has documented how the neutral theory was adopted primarily to simplify calculations.

maximum likelihood loses consistency. Arguing from inconsistency is exactly the wrong way to show that models are useful.

Swofford *et al.*'s (1996: 429f) view of branch lengths suffers a related weakness:

Whereas parsimony ignores information on branch length when evaluating a tree, maximum likelihood considers that changes are more likely along long branches than short ones, and estimation of branch lengths is an important part of the method. This difference explains the consistency of maximum likelihood under many situations in which parsimony is inconsistent.

They cited Felsenstein (1978) for the inconsistency of parsimony, and throughout this section of their paper ("Differences in Perspective between Parsimony and Likelihood," pp. 428ff) they employed his homogeneity assumption in their examples. Now *if* that assumption were true, there would be nothing much wrong with the idea of branch length. In that case all characters (or sites) would have the same probability of changing in a given branch, and the branch length would have a straightforward interpretation in terms of that probability, as in the formula above. The situation shifts, however, if mosaic evolution is taken into account. Then some characters may be most likely to change when change in others is least likely, and the idea that a branch has one particular substitution probability—one particular length—becomes suspect. It is precisely by insisting on fitting a single length to each branch that F78 maximum likelihood becomes inconsistent in the example above. One could well say that estimation of branch lengths explains cases in which maximum likelihood is inconsistent.

In their argument, Swofford *et al.* distinguished parsimony from maximum likelihood by effectively equating likelihood, not simply with models, but with the particular kind of model in which homogeneity is assumed. A different perspective on that distinction is gained by considering the consequences of entirely doing away with the homogeneity assumption, so that each character (or site) in effect is free to have its own suite of branch lengths.<sup>3</sup> This is the "no common mechanism" (NCM) model of Tuffley and Steel (1997), who proved (p. 599, italics in the original)

<sup>3</sup>Notice that this is a less restrictive case than that considered in Theorem 3, above, in which the ratio of branch lengths of different characters would be required to be the same in all branches of the tree.

THEOREM 5. *Maximum parsimony and maximum likelihood with no common mechanism are equivalent in the sense that both choose the same tree or trees.*

Disagreement with parsimony is not intrinsic to maximum likelihood as such, but instead depends on the choice of model. This raises the question of why the homogeneity assumption was adopted. Certainly not for its realism: Felsenstein (1978; quoted above) did not believe it himself. It could not even have been done for computational convenience, since NCM (i.e., parsimony) calculations are much more efficient. It would appear that homogeneity was assumed simply because it would provide the appearance of a basis for attributing just one substitution probability to each branch.

## HISTORY

Siddall (1998) has called attention to an interesting phenomenon, what might be called selective presentation. Advocates of likelihood (for example, Huelsenbeck, 1995) have frequently published results from simulations—always with homogeneous models—illustrating the advantages of maximum likelihood in the Felsenstein Zone, that is, when two nonsister branches are long while other branches are short. No one, however, had published investigations of the case in which the longer branches are sisters, and when Siddall did so, he obtained results less flattering to maximum likelihood. Other discussions of likelihood show a similar pattern. In their recent review, Huelsenbeck and Crandall (1997) cited a wide variety of papers on maximum likelihood estimation of trees, but they did not mention Steel *et al.*'s (1994) Theorem 3 or Chang's (1996) examples. Neither did Swofford *et al.* (1996).

A last argument from Swofford *et al.* (1996: 427) works in much the same way, and discussing it will help to put the present conclusions in context:

Felsenstein's [1978] results have often been criticized (e.g. Farris [1983, 1986]) because they are based on unrealistic and restrictive models of the evolutionary process. This criticism is unjustified, however, as the point could equally well be made with more general and believable models, but requiring more complex mathematics.

Compare this with my actual comments (Farris, 1986: 22, italics added):<sup>4</sup>

Since no one had claimed that parsimony (or any other scientific method) is incapable of error, the significance of [Felsenstein's, 1978] example might well be questioned. But Felsenstein argued for a different interpretation. *He contended that [maximum likelihood] estimation procedures could be devised so as to be consistent, so portraying the inconsistency of parsimony as a drawback by comparison to other methods.* Rather than regarding the admitted lack of realism of his model as a reason for qualifying his conclusions, he took the opposite position.

After some further discussion, I summarized (p. 25)

Any method for inferring genealogy that is consistent under one set of circumstances can be made inconsistent under others: it is only a matter of imagining the circumstances. . . . Felsenstein [1978] presented his example as if it demonstrated a weakness of parsimony analysis in particular. It does not do so. Since the kind of argument that he employed shows the same "fault" for every conceivable method, it in fact shows nothing.

As Swofford *et al.* presented the matter, extending Felsenstein's (1978) argument to realistic cases would involve more calculation, but would be otherwise unproblematic. The results discussed here make it obvious that their position is unfounded. Felsenstein's (1978) claim, that maximum likelihood will "entirely avoid the problem of statistical inconsistency," certainly *cannot* be established with "more general and believable models." Once models are made believable, maximum likelihood loses its guarantee of consistency, and then

Since the kind of argument that he employed shows the same "fault" for every conceivable method, it in fact shows nothing.

Guarantees of consistency were never more than a sham issue. They can be achieved only under absurdly oversimplified circumstances. In the real world, no method can guarantee consistency.

Nor should this be surprising. The idea that consistency can be guaranteed in real—as opposed to imaginary—cases is readily recognized as a version of a now-classical philosophical blunder, the belief that empirical inductions can be made infallible (for a discussion see Popper, 1972). That is why cladists, aware of this, have always been more interested in assessing

<sup>4</sup>In 1986 I had not checked Wald's (1949) conditions: it had never occurred to me that Felsenstein would base his position on a nonexistent proof.

the corroboration of phylogenetic hypotheses than in making homogeneity assumptions.

## ACKNOWLEDGMENTS

I thank M. Steel and J. Chang for helpful discussion, A. G. Kluge and D. Lipscomb for their kind assistance, and M. Siddall and Z. Goldstein for their comments and suggestions. This work was supported by NFR Grant 10204-303 to the author.

## REFERENCES

- Chang, J. T. (1996). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.* **134**, 189–215.
- Chang, J. T., and Hartigan, J. A. (1991). Reconstruction of evolutionary trees from pairwise distributions on current species. In "Computing Science and Statistics: Proceedings of the 23rd. Symposium on the Interface" (E. M. Keramidas, Ed.), pp. 254–257. Interface Foundation, Fairfax Station, VA.
- Dietrich, M. R. (1994). The origins of the neutral theory of molecular evolution. *J. Hist. Biol.* **27**, 21–59.
- Farris, J. S. (1983). The logical basis of phylogenetic analysis. In "Advances in Cladistics" (N. I. Platnick and V. A. Funk, Eds.), pp. 7–36. Columbia Univ. Press, New York.
- Farris, J. S. (1986). On the boundaries of phylogenetic systematics. *Cladistics* **2**, 14–27.
- Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**, 240–249.
- Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.
- Felsenstein, J. (1993). "Phylip." Version 3.5. Computer software and documentation. Dept. Genetics, Univ. Washington, Seattle.
- Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48.
- Huelsenbeck, J. P., and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Rev. Ecol. Syst.* **28**, 437–466.
- Popper, K. R. (1972). "Objective Knowledge." Oxford Univ. Press, Oxford.
- Rogers, J. S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* **46**, 354–357.
- Siddall, M. E. (1998). Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris Zone. *Cladistics* **14**, 209–220.

- Steel, M. A., Hendy, M. D., and Penny, D. (1993). Invertible models of sequence evolution. Mathematical and Information science Report 93/02. Massey University, Palmerston North, New Zealand.
- Steel, M. A., Székely, L. A., and Hendy, M. D. (1994). Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.* **1**, 153–163.
- Swofford, D. L., Olsen, G. L., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. In “Molecular Systematics” (D. M. Hillis, C. Morowitz and B. K. Mable, Eds.) Second ed., pp. 407–514. Sinauer, Sunderland, MA.
- Tuffley, C., and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Bio.* **59**, 581–607.
- Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294–307.
- Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595–601.