The Contest Between Parsimony and Likelihood

Elliott Sober*

Two of the main methods that biologists now use to infer phylogenetic relationships are *maximum likelihood* and *maximum parsimony*. The method of maximum likelihood seeks to find the tree topology that confers the highest probability on the observed characteristics of tip species. The method of maximum parsimony seeks to find the tree topology that requires the fewest changes in character states to produce the characteristics of those tip species.[1] This conflict between methods makes a practical difference to working biologists, in that the methods sometimes disagree about which tree is best supported by the observations. Biologists interested in the evolutionary history of this or that group of organisms may find this methodological dispute an irritating distraction from what they really care about. But for philosophers, it is gratifying to find here a philosophical dispute that actually matters to the practice of science.

The main objection that has been made against likelihood methods is that they require one to adopt a model of the evolutionary process that one has scant reason to think is true. The reason likelihood methods require the adoption of a model is that genealogical hypotheses that describe how various species are related to each other do not, by themselves, confer probabilities on the observations. The situation here will be familiar to philosophers under the heading of Duhem's Thesis. Pierre Duhem (1914) was a French physicist, historian, and philosopher of science who contended that physical theories do not entail observations unless they are supplemented with auxiliary assumptions. Forty-some years later, the American philosopher W.V. Quine (1953) generalized Duhem's thesis, claiming that *all* theories (indeed, all hypotheses), whether or not they are in physics, fail to entail observational predictions. What I am saying about genealogical hypotheses involves giving the Duhem/Quine thesis a probabilistic twist. From a likelihood point of view, it isn't essential that the hypotheses we wish to evaluate *deductively entail* observational claims about the characteristics of species. What is required is that they *confer probabilities* on these statements. The problem is that they do not. In the language of statistics, genealogical hypotheses are composite, not simple.

The main objection that has been made against parsimony is that parsimony implicitly assumes this or that questionable proposition about the evolutionary process. The difficulty here is that it is far from clear which propositions the method in fact assumes. Since parsimony is standardly formulated as a rule for choosing among phylogenetic trees that contain no reticulations, it is natural to think that the method assumes that evolution is a branching process in which no reticulations occur. But beyond that, it isn't clear what the method assumes. Does it assume that evolution proceeds parsimoniously – that if a lineage starts with one character state and ends with another, that it got there via a trajectory that involved the smallest possible number of evolutionary changes? This allegation has been strenuously denied by proponents of parsimony, some of whom maintain that the only thing that parsimony assumes is that there has been descent with modification.[2]

Which is better – using a method that explicitly makes unrealistic assumptions or a method whose assumptions are unknown? I will not try to answer this question, not just because I have no idea how to do so, but because I think it misrepresents the dialectical situation. Likelihood methods do not require one to adopt an unrealistic process model. And something

substantive *is* known about what parsimony assumes. These are the two topics I'll address in what follows.

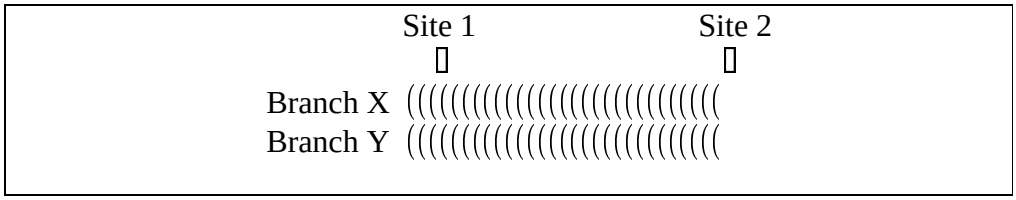
1. Likelihood Methods with Multiple Process Models

Biologists have been using the *likelihood ratio test* to compare different genealogical hypotheses under different models of the evolutionary process. This methodology is most often encountered in studies that use molecular data, but it has also been applied to morphological data. To get a feeling for what the different process models are like, let's consider the two lineages – different branches that are part of a single phylogenetic tree -- depicted in Figure 1. If we drew sequence data from a time slice of one lineage and a time slice of the other, we would find a series of G's, A's, T's, and C's. I will refer to each of the locations in a sequence as a "site." Each letter found at a site has a probability of changing in a given small unit of time. Here are some decisions we need to make about how to model these possible changes:

> *Across branches within sites*: must a change from one letter to another at a site in a branch have the same probability per unit time as the same change when it occurs in a different branch at the same site?
>
> *Across sites within branches*: must a change from one letter to another at one site in a branch have the same probability per unit time as the same change when it occurs at a different site in the same branch?
>
> *Within sites within branches*: must a change from one letter to another at one site in a branch have the same probability per unit time as any other change that might occur at the same site in the same branch?

The Jukes-Cantor (1969) model answers all these questions in the affirmative. It is a one-parameter model – there is a single parameter that characterizes the probability per unit time of all possible evolutionary changes at all sites and in all lineages. The Kimura (1980) model answers the first and second question in the affirmative, but says no to the third; it is a two-parameter model in which there is one parameter for changes between A and C ("transversions") and a second parameter for changes between G and T ("transitions"), where these two parameters apply to all sites on all branches. Both of these models are pretty simple, in terms of the number of adjustable parameters they contain.[3] Far more complex is the Tuffley and Steel (1997) "no common mechanism" model, which says no to the first two questions but yes to the third. It allows different branches to follow different rules, and different sites in the same branch to do so as well. The only constraint this model places on the evolutionary process is that all changes at a site on a branch are assumed to have the same probability. If there are n branches in the tree one is considering and the sequence of nucleotides in one's data set contains m sites, then there are (n)(m) parameters in this model. There is an even more complex model than the one explored by Tuffley and Steel; it simply drops the requirement that all changes at a site on a branch have the same probability. This model has (n)(m)(12) independent parameters.[4]

Figure 1



Notice that the three questions listed above ask about constraints that must be obeyed. Negative answers simply leave matters open. For example, the Jukes-Cantor model assumes that a change from A to G and a change from G to A must have the same probability, whereas the Kimura model leaves it open whether these two changes have the same or different probabilities. It follows that more permissive models are more probable and also more realistic. The most complex model is a near tautology -- it says no more than that each change may have its own unique probability of occurring, though it need not.[5]This model assumes that each site evolves independently of all the others. A model that allowed for the possibility of probabilistic dependencies would be even more complicated.

How are these different process models put to work in a likelihood assessment of phylogenetic hypotheses? Here I must return to the Duhemian point I made before. Suppose we are interested in the genealogical relationship of Humans, Chimps, and Gorillas. Assuming that the tree must be strictly bifurcating (i.e., that it contains no reticulations or polytomies), there are three possibilities – (HC)G, H(CG), and (HG)C. As noted earlier, none of these tree topologies confers a probability on the characteristics we observe the three species to have. However, the same is true if we conjoin these genealogical hypotheses with one or another of the process models just described. The reason is that each process model contains one or more adjustable parameters. Until values for these parameters are specified, we cannot talk about the probability of the data under different hypotheses. In short, the propositions that have well-defined likelihoods take the form of a conjunction:

Tree topology & process model & specified values for the parameters in the model.[6]

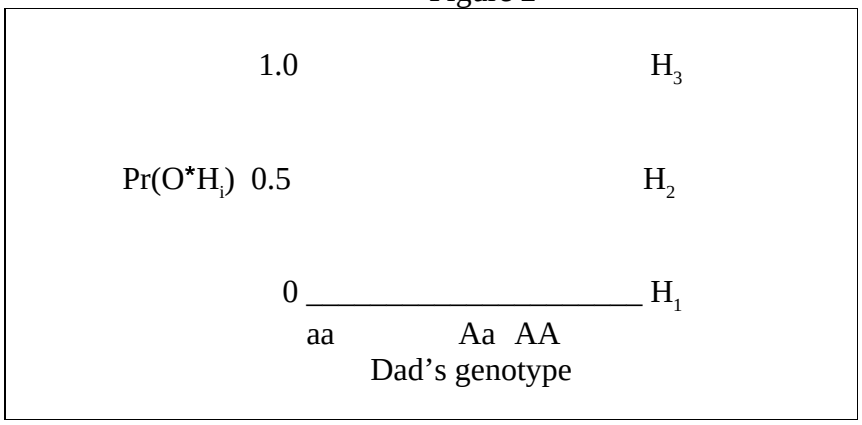
The parameters that describe the probabilities of different changes are examples of what statisticians call *nuisance parameters*. The reason for this name is that they aren't what we are really interested in – we want to compare the likelihoods of different genealogical hypotheses and are forced to deal with these questions about the evolutionary process only because of the Duhemian situation in which we find ourselves. Indeed, the process model itself, and not just the values of the parameters it contains, may be viewed as a nuisance parameter (Edwards 1972).

There are different statistical philosophies that provide guidance for dealing with this problem. To make it clear how they differ, I want to change to a much simpler problem. Suppose you observe that an organism is a heterozygote at a given locus, and you wish to compare different hypotheses about the genotype of the organism's mother. In particular, you wish to evaluate the likelihoods of the following three hypotheses:

(H_1) Mom is AA. (H_2) Mom is Aa. (H_3) Mom is aa.

The inferential situation is depicted in Figure 2. Notice that H_2 has the same likelihood, regardless of what the father's genotype is, whereas H_1 and H_3 have different likelihoods, depending on what the father's genotype is taken to be. H_2 is statistically simple, whereas H_1 and H_3 are composite. For H_1 and H_3, the father's genotype is a nuisance parameter.

Figure 2



.

There are three ways to solve this problem. The first is entirely noncontroversial. If you *know* the father's genotype, you should use that information to compare the likelihoods of the three hypotheses. This knowledge will settle which of three likelihood orderings is the right one to consider; if Dad is AA, the likelihood ordering is $H_3 > H_2 > H_1$, if Dad is Aa, the three hypotheses are equally likely, and if Dad is aa, the ordering is $H_1 > H_2 > H_3$. However, if you don't know Dad's genotype, what should you do?

The second procedure is Bayesian. What you need to do is estimate different conditional probabilities. For H_1, you need to

(Bayes-1) Estimate Pr(Dad is AA * Mom is AA), Pr(Dad is Aa * Mom is AA), and Pr(Dad is aa * Mom is AA),

whereas for H_3, what is required is that you

(Bayes-3) Estimate Pr(Dad is AA * Mom is aa), Pr(Dad is Aa * Mom is aa), and Pr(Dad is aa * Mom is aa).

Notice that it is perfectly possible for these two triplets of numbers to have different values. For example, if there is strong positive assortative mating, then Pr(Dad is AA * Mom is AA) will be large whereas Pr(Dad is AA * Mom is aa) will be small. Only in the special case where there is random mating will the two triplets be the same. The obvious way to estimate the values of these nuisance parameters is to observe the frequencies of different matings. However, if these are not observed, a Bayesian may suggest adopting a set of assumptions that allow one to assign values to the nuisance parameters. The suggested assumptions may or may not be plausible; the point I want to emphasize is that they will address the question – what are the *probabilities* of the different genotypes Dad may have, conditional on assumptions about Mom's genotype? The Bayesian is taking seriously the fact that the likelihood of a composite hypothesis is a *weighted average* of the likelihoods that arise from different settings of the nuisance parameters, where the weights are supplied by the probabilities of those settings. For example,

Pr(Offspring is Aa * Mom is AA) =

Pr(Offspring is Aa * Mom is AA & Dad is AA)Pr(Dad is AA * Mom is AA) +
Pr(Offspring is Aa * Mom is AA & Dad is Aa)Pr(Dad is Aa * Mom is AA) +
Pr(Offspring is Aa * Mom is AA & Dad is aa)Pr(Dad is aa * Mom is AA)

= $\sum$Pr(Offspring is Aa * Mom is AA & Dad has genotype i)Pr(Dad has genotype i * Mom is AA).

Pr(Offspring is Aa * Mom is AA) will be strictly between 0 and 1 if the three weighting terms are all nonzero.

The third strategy for handling nuisance parameters is used in frequentist statistics. Instead of trying to figure out what Dad's genotype *probably* is, conditional on the different hypotheses about Mom's genotype, one simply assigns a genotype to Dad that makes the hypothesis about Mom as likely as possible. For H_1, your procedure is to

(Freq-1) Assign to Dad a genotype that maximizes the value of Pr(Offspring is Aa * Mom is AA and Dad is __),

whereas for H_3, you should

(Freq-3) Assign to Dad a genotype that maximizes the value of Pr(Offspring is Aa * Mom is aa and Dad is __).

The assignment licenced by (Freq-1) is that Dad is aa, while that endorsed by (Freq-3) is that Dad is AA. The result is that H_1 and H_3 both have likelihoods of unity, whereas H_2 (which, recall, contains no nuisance parameters) has a likelihood of ½. Notice that (Freq-1) and (Freq-3) tailor their recommendations about the setting of the nuisance parameter to the hypotheses under consideration, just as (Bayes-1) and (Bayes-3) do. The difference is that Bayesians seek to determine the *probability* of different settings of the nuisance parameters, given the hypothesis under consideration, whereas frequentists assign values that maximize the *likelihoods* of the hypotheses under consideration.

Of the three strategies described, the first strategy is the best. However, if we don't know

what Dad's genotype was, should we be Bayesians or frequentists? Here it seems clear to me that we should be Bayesians if we can obtain good estimates from frequency data of the different conditional probabilities that are relevant to dealing with the nuisance parameters. But if we lack this sort of information, what should we do then? I will not attempt to answer this question here. Instead, I want to explain how these three strategies apply to the problem of dealing with the nuisance parameters that arise in phylogenetic inference. We are interested in comparing the likelihoods of three genealogical hypotheses about Humans, Chimps, and Gorillas. We have sequence data from each species, and we could use one or more process models – Jukes-Cantor, Kimura-2, Tuffley-Steel, and so on. In each instance, we need to assign values to the nuisance parameters if the genealogical hypotheses are to have determinate likelihoods.

The first strategy is to find out which process model is true and what the true values are for the parameters in that model. Half of this procedure can be carried out, but the second cannot. As noted earlier, the maximally complex model – which has a separate parameter for every change in every site in every branch – has to be true. The problem is that the values of these parameters are unknowable. The situation is analogous to the following coin-tossing problem. Suppose I take 50 pennies and toss each of them once. A very simple model would say that they have identical probabilities of landing heads. I can estimate the value of this single parameter by seeing what percentage of those 50 tosses came up heads. A more realistic model is also more complex – it takes seriously the possibility that each of the coins may have its own unique probability of landing heads, and so it contains fifty parameters, one for each coin. If I use maximum likelihood estimation to estimate the values of those fifty parameters, I'll infer that the coins that landed heads had a probability of unity of landing heads and that the coins that landed tails had a probability of unity of doing that. However, these estimates will be subject to huge error, since I have only one observation that pertains to each. In what sense is this more realistic model "better" than the one-parameter model that stipulates that all coins have the same probability of landing heads? Which model would you use to predict a second round of tosses? In coin tossing as in evolution, the realism of a model can be increased by increasing the number of parameters. However, the price of making a model more realistic is that it becomes more difficult to accurately estimate parameter values.

I'll next consider the third strategy described above -- the frequentist procedure. Consider a range of models and assign values to the parameters that maximize the likelihoods of the different genealogical hypotheses. An example of this procedure is depicted in Figure 3. The goal is to find the likeliest setting of the parameters in conjunctions that have the form "genealogical hypothesis & model."

Figure 3

	(HC)G	H(CG)
max complex	L[(HC)G & max complex]	L[(H(CG) & max complex]
Tuffley-Steel	L[(HC)G & Tuffley-Steel]	L[(H(CG) & Tuffley-Steel]
Kimura	L[(HC)G & Kimura]	L[H(CG) & Kimura]
Jukes-Cantor	L[(HC)G & Jukes-Cantor]	L[H(CG) & Jukes-Cantor]

Notice that the setting of parameter values for a given process model can change when we shift from one genealogical hypothesis to another; that is, entries in the same row can differ. That's why I've written "L[genealogical hypothesis & model]" in the different cells, not "genealogical

hypothesis & L[model]."

How would the second, Bayesian, strategy for dealing with nuisance parameters apply to this problem? If the goal is to estimate the likelihood of a tree topology (e.g., (HC)G), then there are two tasks that need to discharge, since

Pr[Data * (HC)G] =
$\sum_j$ Pr[Data * (HC)G & process model i & values j for the parameters in model i] ×
Pr[process model i & values j for the parameters in model i * (HC)G].[7]

The second product term on the right-hand side is where the problems arise. One needs to compute the probabilities of process models and of the values of parameters in those process models, conditional on the tree topology whose likelihood we seek to evaluate. It is hard to know what to say about the former problem, except that simpler models can't have higher probabilities than the more complex models in which they are nested. That is, it is a consequence of the axioms of probability theory that Pr(Jukes-Cantor* (HC)G) # Pr(Kimura * (HC)G) # Pr(Tuffley-Steel* (HC)G). Beyond this, it is hard to say how one might go about assigning probabilities to process models.[8]Pr[Jukes-Cantor * (HC)G] = Pr[Jukes-Cantor * H(CG)] = Pr[Jukes-Cantor * (HG)C]. The values of these three conditional probabilities would still be unknown, but one would at least know that they are equal, a point that might be useful in comparing the likelihoods of the genealogical hypotheses. However, it is not obvious that this independence relation obtains. What Bayesians typically do is choose one or more process models and estimate the values of conditional probabilities of the form depicted in Figure 4. It is important to realize that the process models and the genealogical hypotheses do not provide instructions about how one is to figure out how probable this or that setting of the parameters values in a model is. This requires additional assumptions about prior probabilities – before you look at the data, how probable is it that the parameters will take this value or that?

Figure 4

	(HC)G	H(CG)
max complex	Pr[parameter values * (HC)G & max complex]	Pr[parameter values * H(CG) & max complex]
Tuffley-Steel	Pr[parameter values * (HC)G & Tuffley-Steel]	Pr[parameter values * H(CG) & Tuffley-Steel]
Kimura	Pr[parameter values * (HC)G & Kimura]	Pr[parameter values * H(CG) & Kimura]
Jukes-Cantor	Pr[parameter values * (HC)G & Jukes-Cantor]	Pr[parameter values * H(CG) & Jukes-Cantor]

Although I used the problem of comparing the likelihoods of hypotheses about Mom's genotype as a heuristic for explaining the problem of nuisance parameters in phylogenetic inference, it is important to recognize one difference between the two problems. There is an objective procedure for using Bayesian methods in the genetics problem. The system of mating that a population obeys is a biological property of the population that can be ascertained by looking at frequency data. However, it is very hard to see how the Bayesian approach to nuisance parameters in phylogenetic inference can be put on an objective footing Readers will have to decide for themselves how much subjectivity they are willing to introduce into methods of phylogenetic

inference.

Let's go back to the frequentist approach to nuisance parameters, whose solution is depicted in Figure 3. We have found the likeliest setting of the parameters in each conjunction of the form "genealogical hypothesis & model," which we denote as "L[genealogical hypothesis & model]." We now can ask how the likelihoods of these different conjunctions compare with each other. Since the models are nested, we know that likelihoods increase as we ascend each column -- you get better fit-to-data as you increase the number of adjustable parameters. However, the likelihoods in rows tend to get closer together as we ascend. In the top row, the likelihoods are identical; each has the maximum value of unity. As we make our process model more realistic, do we reduce our ability to discriminate between genealogical hypotheses?

That depends on the method used to compare the conjunctions depicted in Figure 3. If we use the criterion of maximum likelihood, we will ave to conclude that the three conjunctions in the top row are best, but that we are unable to discriminate among them. However, frequentists do not endorse this method. Rather, they use the *likelihood ratio test.* This test is designed to apply to nested models; according to this method, the question is whether the likelihood of the more complex model is sufficiently greater than the likelihood of the simpler model to justify rejecting the simpler model. So it is not inevitable that one will reject all conjunctions in a column, save for the one at the top. However, there is a problem with this approach – likelihood ratio tests are well-grounded only for nested models. The procedure makes sense for "vertical" comparisons in Figure 3, not for "horizontal" or "diagonal" comparisons.

Because of this, it might make sense to leave likelihood ratio tests behind, and shift to a model selection criterion such as the one proposed by Akaike (1973).[9] AIC – the Akaike information criterion – applies to nested and non-nested models alike. AIC is based on a theorem that Akaike (1973) proved. He was able to show that

An unbiased estimate of the predictive accuracy of model M . log[Pr(Data * L(M)] - k.

Here M is a model that contains adjustable parameters – in the case at hand this would be a conjunction of a genealogical hypothesis and a process model. L(M) is the likeliest setting of the parameters in the model. One merely computes the log-likelihood of this fitted model and then subtracts k, which is the number of adjustable parameters in the model. Although likelihood increases as we ascend the columns in Figure 3, so does the value of k. For this reason, it is not inevitable that more complex models will receive higher AIC scores than simpler ones; that depends on the data. Forster and Sober (1994) suggested the term "predictive accuracy" for the quantity that AIC attempts to estimate. The predictive accuracy of a model is how well, on average, it will predict new data when it is fitted to old data.

In using AIC, one will obtain an ordered list, from best to worse, of conjunctive hypotheses, each of which has the form [genealogical hypothesis & process model]. The Duhemian point continues to apply -- what one is testing here, in the first instance, are the different conjunctions, not the genealogical hypotheses taken on their own. Still, one can reach inside the conjunctions and examine the conjuncts of interest in the following way. Suppose (HC)G is the genealogical hypothesis that figures in the n conjunctions that receive the best AIC scores. The larger n is, the more we are entitled to conclude that the data favor (HC)G. In this case, (HC)G is *robust* across variation in process model.

Bayesians face a similar Duhemian problem. If they decline to say how probable this or that

process model is, conditional on a tree topology, they will be unable to compute the likelihoods of tree topologies. Instead they will be able to compute the average likelihood of conjunctions that have the familiar form "genealogical hypothesis & tree topology." The question can then be addressed of whether, say, (HC)G has a higher likelihood than H(CG), across some range of process models. The larger the range, the more robust (HC)G is.[10]Notice that BIC imposes a penalty on models for complexity, though the penalty differs from the one that AIC deploys. This may give the impression that the two methods are in conflict. In fact, they are not, since AIC and BIC have different goals – the former estimates predictive accuracy while the latter estimates average likelihood. BIC scores for conjunctions of genealogical hypotheses and process models can be ordered from best to worst; as with AIC, the question would be whether one genealogical hypothesis dominates the others in the n conjunctions that receive the best BIC scores. It would be interesting to compare the results of using BIC with those of using the Markov Chain Monte Carlo methods now used in Bayesian phylogenetic inference (Huelsenbeck *et al.* 2001).

Regardless of this difference between the Bayesian and the frequentist approach, there is an important feature that they have in common – statistical treatments of genealogical hypotheses do not require one to choose a single process model and assume that it is true. Statistical methods permit one to explore multiple models. One can consider both relatively simple models that impose substantial idealizations and more complex models that are more realistic. The question is whether the ordering of genealogical hypothesis is robust across variation in process model.

2. What does parsimony assume about the evolutionary process?

What does the word "assume" mean in the question that forms the title of this section? An example from outside science provides the necessary guidance. The sentence

(P) Jones is poor but honest

assumes (presupposes) that

(A) There is a conflict between being poor and being honest.

This is why the meaning of statement (P) would change if we replaced "but" with "and." The relation of (P) and (A) illustrates the following simple point about what assumptions are:

If P assumes A, then P entails A.

What is manifestly false is the opposite claim:

If P assumes A, then A entails P.

To say that there is a conflict between poverty and honesty implies nothing about the characteristics that Jones happens to have. To find out what a proposition assumes, you must look for conditions that are *necessary* for the proposition to be true, not for conditions that *suffice* for the proposition's truth (Sober 1988).

Having said something about what "assumptions" involve, we can turn to the question of what it means to ask what parsimony assumes. The trouble is that the word "parsimony" does not express a proposition. What is involved here, of course, is the proposition that "parsimony is a legitimate method of phylogenetic inference." What parsimony assumes about the evolutionary process are the propositions that must be true if parsimony is to be a legitimate method of phylogenetic inference. But what does "legitimate" mean?

There are a number of choices to consider. For example, one might require that a legitimate phylogenetic method be statistically consistent – that it converge on the true phylogeny as the number of observations is made large without limit. This is the approach taken in Felsenstein (1978). The question about parsimony's assumptions then becomes – what features must the evolutionary process have if parsimony is to be statistically consistent? People who reject the requirement of statistical consistency will not accept this line of argument. For example, Sober (1988) argues that likelihood methods can be legitimate even when they fail to be statistically consistent. And it turns out that Tuffley and Steel's (1997) "no common mechanism" model fails to conform to the sufficient condition for statistical consistency presented by Wald (1949), in that the number of parameters that the model requires one to estimate grows as one draws more and more data. If you reject the Tuffley-Steel model, this point has not bite. However, rejecting the model involves rejecting all the models nested within it, for example, the Jukes-Cantor model. Biologists who think that the Tuffley-Steel model is a legitimate model will not want to embrace the requirement of statistical consistency.

The interpretation I want to explore here is the idea that parsimony's legitimacy consists in its being *ordinally equivalent* with likelihood. This idea is easy to understand by considering the Fahrenheit and Centigrade scales of temperature. These are ordinally equivalent, meaning that for any two objects, the first has a higher temperature-in-Fahrenheit than the second precisely when the first has a higher temperature-in-Centigrade than the second. The two scales induce the same ordering of objects. For parsimony and likelihood to be ordinally equivalent, the requirement is that

(P) For any phylogenetic hypotheses H_1 and H_2, and for any data set D, H_1 provides a more parsimonious explanation of D than H_2 does precisely when H_1 has a higher likelihood than H_2, relative to D.

I am interested in (P) as a device for exploring the legitimacy of parsimony because I already think that likelihood is a good measure of the degree to which evidence favors one hypothesis over another. However, (P) could be employed in the opposite direction -- by someone who already believes that cladistic parsimony is legitimate, and who wants to see whether likelihood can be justified in terms of parsimony.

So our question about the assumptions that parsimony makes comes to this -- which propositions about the evolutionary process does proposition (P) entail? For example, does parsimony presuppose the no common mechanism model described by Tuffley and Steel (1997)?[11] Well, what Tuffley and Steel demonstrated is that

The no common mechanism model entails ordinal equivalence.

But this does not show that parsimony assumes ordinal equivalence. This model *suffices* for ordinal equivalence; there is no proof here that the assumptions of the model are *necessary*.

Still, the Tuffley-Steel result has great significance for the question of what parsimony assumes, in virtue of the fact that logical entailment is transitive:

no common mechanism ⊃ ordinal equivalence ⊃ assumptions of parsimony

In virtue of the Tuffley-Steel result, any proposition that is entailed by ordinal equivalence also must be entailed by the no common mechanism model. But the fact that a proposition is entailed by the model does not ensure that it is entailed by ordinal equivalence. This provides the following partial test for whether a proposition is assumed by parsimony (Sober 2002):

– If a proposition is entailed by the no common mechanism model, it *may or may not be* an assumption that parsimony makes.

– If a proposition is not entailed by the no common mechanism model, it is *not* an assumption that parsimony makes.

This test for what parsimony assumes has some interesting consequences. First, the no common mechanism model does not entail that homoplasies are rare or that the probability of change on branches is very low. Hence parsimony does not assume that homoplasies are rare or that change is very improbable. This result is interesting in view of Felsenstein's (1979, 1981) argument that a low probability of change on branches suffices for parsimony and likelihood to coincide The Tuffley-Steel model also does not assume that the probability of a change's occurring on a branch is independent of the branch's duration. Hence, this (implausible) independence assumption is not an assumption of parsimony's. This is interesting, in view of Goldman's (1990) presentation of a model that makes this independence assumption and which Goldman claims suffices to ensure ordinal equivalence.

I have run this argument about parsimony's presuppositions by using the Tuffley-Steel model as the basis for the test, but, in principle, any other model that induces ordinal equivalence could be used in the same way. Does this mean that we could use Felsenstein's or Goldman's models to evaluate which entailments of the Tuffley-Steel model are assumptions that parsimony makes? There is a complication here. The Tuffley-Steel model understands both likelihood and parsimony as outputting tree topologies; but no assignments of states to interior nodes. Goldman (1990) conceives of both procedures differently – each outputs a tree topology *and* an assignment of character states to all branch points in the tree's interior. Just for the record, it is worth recalling that Farris (1973) showed that the most parsimonious tree is the tree of maximum likelihood when both methods are taken to output a tree topology *and* an assignment of character states to interior branch points *and* an assignment of character states to all other time slices on branches in the tree's interior. The model Farris used to establish this connection between likelihood and parsimony assumes very little about the evolutionary process – in particular, there is no assumption that all changes at a site have the same probability. If the method of maximum parsimony simply outputs a tree topology, and talks about interior character states merely as a means for deciding which topology is best, then the arguments by Farris, Felsenstein, and Goldman do not identify models that induce ordinal equivalence (Steel and Penny 2000).

There is a third proposition about evolution that we can consider from the point of view of the test procedure that uses the Tuffley-Steel result. The no common mechanism model assumes neutral evolution. If the probability of a site's evolving from one character state to another is the

same as the probability of any other change that might occur at the site, then there is no selection favoring one character state over the other. The question is whether neutralism is entailed by ordinal equivalence. That there may be some plausibility to this conjecture is suggested by some findings about parsimony in another inferential context. Instead of using that method to reconstruct an evolutionary tree, parsimony can be used to take an assumed tree and assign character states to ancestors. Maddison (1991) was able to show that likelihood and parsimony are ordinally equivalent in this problem if a neutral model of evolution is assumed.[12] Sober (2002) showed that parsimony can fail to coincide with likelihood in this problem if there is directional selection. Given neutralism's intimate connection with ordinal equivalence in the context of inferring ancestral character states, perhaps neutralism is also Sober (2002) showed that parsimony can fail to coincide with likelihood in this problem if there is directional selection. Given neutralism's intimate connection with ordinal equivalence in the context of inferring ancestral character states, perhaps neutralism is also critical for ordinal equivalence in the context of inferring tree topologies. This merits investigation

It also would be worth returning to the contrast drawn earlier in this paper between Bayesian and frequentist methods for handling nuisance parameters. Tuffley and Steel use the frequentist procedure. What connection can be established between parsimony and (average) likelihood when nuisance parameters are handled in a Bayesian fashion? Much remains to be learned about parsimony's presuppositions.

3. Concluding Comments

Since the early 1970's, there has been a raging dispute between defenders of maximum likelihood and defenders of maximum parsimony. The former group has followed a frequentist statistical philosophy, using the likelihood ratio test. The entry of Bayesian methods into the arena of phylogenetic inference is much more recent. There has been little discussion in the literature of the difference between frequentist and Bayesian approaches. Perhaps the reason is that frequentists and Bayesians agree that phylogenetic inference should be understood as a statistical problem, and so they see maximum parsimony as the common enemy. But, as in other domains of human conflict, as a common enemy is perceived to be less threatening, one-time allies turn their attention to the issues that divide them. When frequentists and Bayesians start discussing their disagreements, it it to be hoped that the discussion will be less acrimonious than debate that has been underway between cladists and frequentists.

A place to begin the comparison of frequentism and Bayesianism is to see how often the two approaches makes a practical difference. Are there data sets for which the two approaches generate different answers? How often do such data sets arise in biological practice? This question is worth pursuing by using both real data sets and sets that are invented for exploratory purposes. In the toy example about assessing the likelihoods of hypotheses about Mom's genotype, the decision about how nuisance parameters should be handled makes an enormous difference. Is there reason to think that this difference disappears as more data are brought to bear?

A statistical analysis of the properties of the method of maximum parsimony is also worth developing further. This is an interesting question, both for those who are already sold on the correctness of some statistical approach and on those who think that parsimony makes sense in ways that explicitly statistical methods do not. As noted earlier in connection with the idea of ordinal equivalence, finding models in which parsimony and likelihood agree throws light in both

directions. The Tuffley-Steel result is one step in this direction, but other models need to be developed before the conceptual terrain can be said to be well understood.

References

Akaike, H. (1973): "Information Theory as an Extension of the Maximum Likelihood Principle." In B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267-281.

Burnham, K. and Anderson, D. (1998): *Model Selection and Inference – a Practical Information-Theoretic Approach*. New York: Springer.

Edwards, A. (1972): *Likelihood*. Cambridge: Cambridge University Press.

Duhem, P. (1914): *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press. 1954.

Farris, J. (1973): "On the Use of the Parsimony Criterion for Inferring Phylogenetic Trees." *Systematic Zoology* 22: 250-256.

Felsenstein, J. (1978): "Cases in which Parsimony and Compatibility Methods can be Positively Misleading." *Systematic Zoology* **27**: 401-410.

Felsenstein, J. (1979): "Alternative Methods of Phylogenetic Inference and their Interrelationships." *Systematic Biology* 28: 49-62.

Felsenstein, J. (1981): "A Likelihood Approach to Character Weighting and What It Tells Us About Parsimony and Compatibility." *Biological Journal of the Linnaean Society* 16: 183-196.

Forster, M. and Sober, E. (1994): "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45: 1-36. Also available at the following URL:

Goldman, N. (1990): "Maximum Likelihood Inference of Phylogenetic Trees, with Special Reference to a Poisson Process Model of DNA Substitution and to Parsimony Analyses." *Systematic Zoology* 39: 345-361.

Huelsenbeck, J., Ronquist, F., Nielsen, R., and Bollback, J. (2001): "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology." *Science* 294: 2310-2314.

Jukes, T., and Cantor, C. (1969): "Evolution of Protein Molecules." In H. Munro (ed.), *Mammalian Protein Metabolism*. New York: Academic Press, pp. 21-132.

Kimura, M. (1980): "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences." *Journal of Molecular Evolution* 16: 111-

120.
States for Continuous-Valued Characters on a Phylogenetic Tree." *Systematic Zoology* 40: 304-314.

Page, R. and Holmes, E. (1998): *Molecular Evolution – A Phylogenetic Approach*. Oxford: Blackwell.

Quine, W. (1953): "Two Dogmas of Empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press. 20-46.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986): *Akaike Information Criterion Statistics*. New York: Springer.

Schwarz, G. (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6**: 461-465.

Sober, E. (1988): *Reconstructing the Past – Parsimony, Evolution, and Inference*. Cambridge: MIT Press.

Sober, E. (2002): "Reconstructing Ancestral Character States – A Likelihood Perspective on Cladistic Parsimony." *The Monist* 85: 156-176.

Steel, M. and Penny, D. (2000): "Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics." *Molecular Biology and Evolution* 17: 839-850.

Tuffley, C. and Steel, M. (1997): "Links Between Maximum Likelihood and Maximum Parsimony under a Simple Model of Site Substitution." *Bulletin of Mathematical Biology* 59: 581-607.

Wald, A. (1949): Note on the Consistency of the Maximum Likelihood Estimate. *Annals of Mathematical Statistics* 20: 595-601.

Notes

*. My thanks to David Posada and Michael Steel for useful comments.

1 This description of the two methods is a little misleading, in that both methods indicate, not just which phylogenetic hypothesis is best supported by the data, but an *ordering* of hypotheses, from best to worst.

2 Michael Steel has pointed out to me that parsimony need not assume a tree structure. It is a result in the theory of Steiner trees that the most parsimonious graph will have a tree structure, so the requirement of parsimony leads to trees, without presupposing them.

3 For other models of molecular evolution now used in phylogenetic inference, see Page and Holmes (1998).

4 Notice that in the Tuffley and Steel model and in this more complex model as well, the number of parameters grows as one draws more sequence data. Thus Wald's (1949) sufficient condition for statistical consistency is not satisfied.

5 Actually, there are models more complicated than the most complicated model just described..

6 Although these models specify the probabilities of different changes per unit time, they also describe how the probability of a branch's ending in some state, given that it begins ,n another, depends on the "instantaneous" probabilities of change *and* on the lineage's duration. In the end what we need to know are values for the "branch transition probabilities" that a given topology and process model require.

7 I state this as a (discrete) summation rather than as a (continuous) integration just to make the basic concepts more transparent.

8 If process model were *independent* of genealogical hypothesis, then, for example,

9 See also Sakamoto *et al.* (1986) and Burnham and Anderson (1998).

10 Instead of restricting one's self to a series of within-row comparisons in Figure 4, Bayesians may want to consider using the Bayesian Information Criterion (BIC) first derived by Schwarz (1978):

$$\Pr(\text{Data} \mid \text{Model } M) \cdot \log[\Pr(\text{Data} \mid L(M)] - (k/2)\text{Log}(N).$$

11 I should note here that Tuffley and Steel handle the nuisance parameters of branch transition probabilities by using the frequentist method described earlier, not by a Bayesian averaging over different possible values. It would be interesting to see whether and how the result changes if a Bayesian method for handling nuisance parameters is adopted.

12 Maddison (1991) was discussing quantitative characters and he interpreted parsimony to mean the minimizing of squared character change.