Syst. Biol. 53(4):644–653, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490468657

The Contest Between Parsimony and Likelihood

ELLIOTT SOBER

Philosophy Department, Stanford University, Stanford, CA 94305, USA; E-mail: esober@stanford.edu; and Philosophy Department, University of Wisconsin, Madison, WI 53706, USA; E-mail: ersober@wisc.edu

In a "classic" phylogenetic inference problem, the observed taxa are assumed to be the leaves of a bifurcating tree and the goal is to infer just the "topology" of the tree (i.e., the formal tree structure linking the extant taxa at the tips), not amount of time between branching events, or amount of evolution that has taken place on branches, or character states of interior vertices. Two of the main methods that biologists now use to solve such problems are *maximum likelihood* (*ML*) and *maximum parsimony* (*MP*); distance methods constitute a third approach, which will not be discussed here. *ML* seeks to find the tree topology that confers the highest probability on the observed characteristics of tip species. *MP* seeks to find the tree topology that requires the fewest changes in character state to produce the characteristics of those tip species. Besides saying what the "best" tree is for a given data set, both methods also provide an *ordering* of trees, from best to worst. The two methods sometimes disagree about this ordering—most vividly, when they disagree about which tree is best supported by the evidence. For this reason, biologists have had to address this methodological dispute head on, rather than setting it aside as a merely "philosophical" dispute of dubious relevance to scientists "in the trenches."

The main objection that has been made against *ML* is that it requires the adoption of a model of the evolutionary process that one has scant reason to think is true. *ML* requires a process model because hypotheses that specify a tree topology (and nothing more) do not, by themselves, confer probabilities on the observations. The situation here is familiar to philosophers as an instance of "Duhem's Thesis." Pierre Duhem was a French philosopher of science who contended that physical theories do not entail[1] claims about observations unless they are supplemented with auxiliary assumptions (Duhem, 1914). The American philosopher W.V. Quine (1953) later generalized Duhem's thesis, claiming that *all* hypotheses fail to entail observational predictions all by themselves. The present point about genealogical hypotheses gives a probabilistic twist to the Duhem/Quine thesis. 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 those observations. 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 *MP* 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. Because *MP* is standardly formulated as a rule for choosing among phylogenies that contain no reticulations, it is natural to suspect that the method assumes that evolution is a branching process in which no reticulations occur; however, users of parsimony need not postulate a tree structure as an additional assumption, because it is a result in the theory of Steiner trees (a Steiner tree is also a topology, though it may contain reticulations) that the most parsimonious graph will have a tree structure (Semple and Steel, 2003: 97–98). One also might wonder whether *MP* assumes that evolution proceeds parsimoniously—does *MP* assume 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 (see, e.g., Farris, 1983), some of whom maintain that parsimony assumes only that there has been descent with modification.

Which is better—using a method that explicitly makes unrealistic assumptions or a method whose assumptions are unknown? I will argue that this unhappy dilemma misrepresents the dialectical situation. Likelihood methods do *not* require one to adopt a single process model. And something substantive *is* known about what parsimony assumes, though much less than some have suggested. These are the two topics I'll address in what follows.

The debate about *ML* and *MP* may appear to some biologists to be settled by the type of data one wishes to analyze, the thought being that aligned sequences require *ML* and phenotypes require *MP*. To be sure, *ML* is often applied to sequences and rarely to phenotypes (see Lewis, 2001 for an exception), whereas *MP* is often applied to morphological data and with increasing

[1]To say that *X* entails (or implies) *Y* means that if *X* is true, then *Y* must be true. If *X* entails *Y*, then *X* suffices for *Y* to be true, and *Y* is a necessary condition for *X* to be true. Some simple facts may give the reader a feeling for what this logical notion involves: Every proposition entails itself, the conjunction "*A and B*" entails *A*, and *A* entails the disjunction "*A or B*." Entailment (implication) is a formal mathematical relation; it has nothing to do with psychological questions about what anyone assumes.

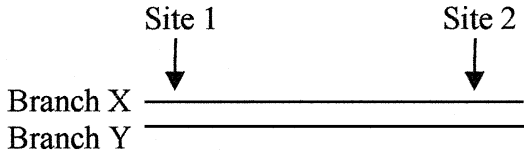

FIGURE 1. Two sites in each of two aligned sequences drawn from different branches of a phylogeny.

reluctance to sequences. However, this is a sociological fact, not a logical inevitability. In what follows I will describe a set of questions that must be addressed if *ML* is to be used on sequence data; the very same questions also are central to the task of applying *ML* to phenotypic data. Symmetrically, *MP* can be applied to sequence data just as it can be applied to morphology. In addition, *ML* and *MP* are sometimes *equivalent* (more on this below), so it is hard to see how *MP* can be tied essentially to phenotypic data and *ML* to sequence data. Choice of a type of data leaves open what method one should use in interpreting that data—that is, an argument is required for opting for one approach rather than the other. The data do not have written on their sleeves which method one ought to use.[2]

LIKELIHOOD METHODS WITH MULTIPLE PROCESS MODELS

Biologists have used *ML* to compare different genealogical hypotheses under different models of the evolutionary process. Because this methodology is most often encountered in studies that use sequence data, I will discuss *ML* in that context. To get a feeling for the different process models that have been used, consider two lineages and the aligned sequences of *G*'s, *A*'s, *T*'s, and *C*'s present in each, as depicted in Figure 1. I refer to each location in a sequence as a "site." To model the evolutionary process at work in the two lineages, we need to answer the following questions:

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 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 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 at the same site in the same branch?

The Jukes-Cantor (1969) model answers all three questions in the affirmative. This model contains a single parameter, which characterizes the probability per unit time of all possible changes at all sites in all lineages. The Kimura (1980) model says *yes* to the first and second question but *no* to the third; it is a two-parameter model in which transversions and transitions are allowed to have different probabilities. Both these models are pretty simple, in terms of the number of adjustable parameters they contain. 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. However, this model requires that all changes at a site on a branch have the same probability. If there are *m* branches in the tree one is considering and the sequence of nucleotides in one's data set contains *n* sites, then there are *mn* parameters in this model. There is an even more complex model than the one explored by Tuffley and Steel; it drops the requirement that all changes at a site on a branch have the same probability, thus answering *yes* to all three questions. I will call this the 12*mn* model, for the number of parameters it contains.[3] I mention this range of possibilities, not because I think they are equally good (see below) but to give an indication of the choices that are available. For additional models, see Page and Holmes (1998) and Swofford et al. (1996).

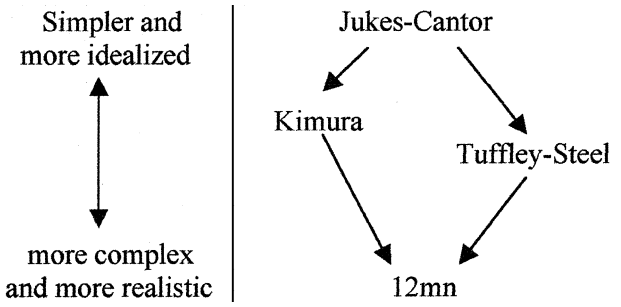

FIGURE 2. Models of the evolutionary process are simpler or more complex according to how many adjustable parameters they contain. The arrow represents deductive implication; "$M_1 \rightarrow M_2$" means that if M_1 is true, M_2 must be true.

Notice that the three questions just listed 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 *C* and a change from *G* to *T* must have the same probability, whereas the Kimura model leaves open whether these changes have the same or different probabilities. The logical relationships that obtain among the four models discussed thus far are shown in Figure 2. The Jukes-Cantor model entails (is a special case of) the three other models. As one follows a chain of arrows from top to bottom, models become less restrictive, more realistic, and more complex. It is an idealization to say that transitions and transversions must have precisely the same probability, and it can hardly fail to be true to say that they may or may not have the same probability. The most complex model described—the 12*mn* model—is the least idealized model mentioned; it can hardly fail to be true that

[2] I also should mention that I will not discuss *ML* or *MP* in connection with data on gene order or on SINEs.

[3] As an expository convenience, I will ignore the number of parameters that these models require for the state of the root of the tree.

each change in each site in each branch may have its own unique probability of occurring, though it need not. But even this very complex model contains an idealization—like the others mentioned, it assumes that each site evolves *independently* of all the others. A model that allowed for the possibility of probabilistic dependencies across sites 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 described 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 possible rooted trees—*(HC)G*, *H(CG)*, and *(HG)C*. As noted earlier, none of these confers a probability on the characteristics we observe. 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 at least one adjustable parameter. Until values for adjustable 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.[4]

The parameters that describe the probabilities of different changes are examples of what statisticians call *nuisance parameters*. The reason for this name is not that biologists never take an interest in these probabilities; rather, the point is that when we are interested in comparing the likelihoods of different tree topologies, we are forced to deal with questions about the evolutionary process even though these are not the focus of our interest. Indeed, Edwards (1972) remarks that the process model itself, and not just the values of the parameters it contains, may be viewed as a nuisance parameter. Of course, what is a nuisance parameter in one problem may be the subject of interest in another; for example, in testing hypotheses about natural selection, the tree topology will be a nuisance parameter.

There are different statistical philosophies that provide guidance concerning how nuisance parameters should be handled. To clarify how they differ, I want to consider a much simpler problem as an example. Suppose you observe (*O*) that an organism is a heterozygote at a given locus and you wish to compare the likelihoods of three hypotheses about the genotype of the organism's mother:

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

[4]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 in another, depends on the "instantaneous" probabilities of change *and* on the branch's duration. In the end, what we need to know are values for the "branch transition probabilities" that the conjunction of a topology and a process model postulates.

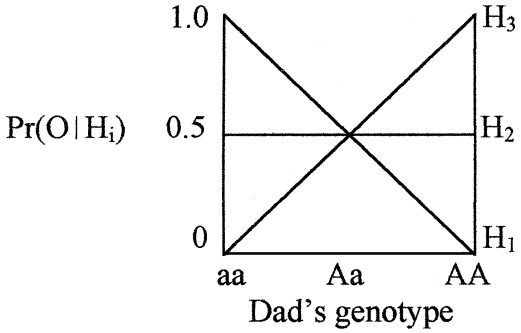

FIGURE 3. The likelihoods of three hypotheses about Mom's genotype, relative to the observation (O) that the offspring is a heterozygote. For two of them (H_1 and H_3), Dad's genotype is a nuisance parameter.

The inferential situation is depicted in Figure 3. 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 said to be "statistically simple," whereas H_1 and H_3 are composite. For H_1 and H_3, the father's genotype is a nuisance parameter.

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?

This brings us to the second procedure for dealing with nuisance parameters. According to Bayesianism, you need to estimate the values of 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, 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 in the population. In the absence of such frequency data, a Bayesian may suggest adopting a set of assumptions that allows one to assign values to the nuisance parameters. These assumptions may or may not be plausible; the point I want to emphasize is that they address a question of a certain form—what are the *probabilities* of the different

genotypes that Dad may have, conditional on Mom's having this or that genotype? Bayesians take 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 conditional probabilities of those settings. For example,

$$\begin{aligned}
&\Pr(\text{Offspring is } Aa \mid \text{Mom is } AA) = \\
&\quad \Pr(\text{Offspring is } Aa \mid \text{Mom is } AA \,\&\, \text{Dad is } AA) \\
&\qquad \times \Pr(\text{Dad is } AA \mid \text{Mom is } AA) \\
&\quad + \Pr(\text{Offspring is } Aa \mid \text{Mom is } AA \,\&\, \text{Dad is } Aa) \\
&\qquad \times \Pr(\text{Dad is } Aa \mid \text{Mom is } AA) \\
&\quad + \Pr(\text{Offspring is } Aa \mid \text{Mom is } AA \,\&\, \text{Dad is } aa) \\
&\qquad \times \Pr(\text{Dad is } aa \mid \text{Mom is } AA) \\
&= \sum_i \Pr(\text{Offspring is } Aa \mid \text{Mom is } AA \\
&\quad \&\ \text{Dad has genotype } i) \\
&\quad \times \Pr(\text{Dad has genotype } i \mid \text{Mom is } AA).
\end{aligned}$$

Notice that 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 the one used in frequentist statistics. Instead of trying to figure out what Dad's genotype *probably* is, conditional on different hypotheses about Mom's genotype, one simply assigns a genotype to Dad that maximizes the likelihood of the hypothesis about Mom. For H_1, your procedure is to

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

whereas for H_3, you should

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

The assignment licensed by (Freq-1) is that Dad is *aa*, whereas 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 $\frac{1}{2}$. Notice that (Freq-1) and (Freq-3) tailor their recommendations about the setting of the nuisance parameter to the hypothesis under consideration, just as (Bayes-1) and (Bayes-3) do. The difference is that Bayesians seek to determine the *probabilities* of different settings of the nuisance parameter, given the hypotheses under consideration, whereas frequentists assign values that maximize the *likelihoods* of the hypotheses under consideration.

Of the three strategies described, the first is the best. However, if we don't know what Dad's genotype is, 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 relevant conditional probabilities. But if we lack this sort of information, what should we do? 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. In our running example, we want to compare the likelihoods of three genealogical hypotheses about Humans, Chimps, and Gorillas. We have sequence data from each species, and we can use one or more process models—Jukes-Cantor, Kimura, 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 just described for handling nuisance parameters is to find out which process model is true and what the true values are for the parameters in that model. The problem with this strategy is that all the models described so far contain idealizations and so all are false. A fully realistic model would need even more parameters than those contained in the $12mn$ model because it would have to replace the assumption of independence between sites with a suite of parameters that allow for possible failures of independence. Full realism would require so many parameters that it would be impossible to estimate their values. The situation is analogous to the following coin-tossing problem. Suppose you take 50 pennies and toss each of them once. One possible model is very simple; it says that the coins have identical probabilities of landing heads. You can obtain a *ML* estimate of the value of this single parameter by seeing what proportion of the 50 tosses landed heads. An alternative model is both more realistic and more complex; it takes seriously the possibility that each of the coins may have its own unique probability of landing heads, and so this model contains 50 parameters, one for each coin. If you use *ML* to estimate the values of those 50 parameters, you will 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 landing tails. These estimates will be subject to huge error, because you have only one observation that pertains to each. Although the 50-parameter model is more realistic than the 1-parameter model, it is far from clear that you should use the more realistic model if you want 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 now turn to the third strategy described above for handling nuisance parameters—the frequentist procedure. Here, for each conjunction that has the form "genealogical hypothesis & process model," we must find the setting of the parameters that maximizes the likelihood of the conjunction. An example is depicted in Table 1. Notice that the settings of parameter values for a given process model can, in principle, change when we shift from one genealogical hypothesis to another; that

TABLE 1. Conjunctions of the form "process model & tree topology" contain adjustable parameters; these are nuisance parameters in the context of making inferences about topologies. Frequentists set these at their maximum likelihood values, denoted by "L[process model & tree topology]."

Model	(HC)G	H(CG)
12mn	L[(HC)G & 12mn]	L[H(CG) & 12mn]
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]

is, entries in the same row can differ. That is why I've written "L[genealogical hypothesis & model]" in the different cells, not "genealogical hypothesis & L[model]."

Now let's consider the second, Bayesian, strategy for dealing with nuisance parameters. Here there are two tasks that need to be discharged, because

$$\begin{aligned}&\Pr[\text{Data} \mid (HC)G] \\ &= \sum_i \sum_j \Pr[\text{Data} \mid (HC)G \ \&\ \text{process model } i \\ &\quad \&\ \text{values } j \text{ for the parameters in model } i] \\ &\quad \times \Pr[\text{process model } i \ \&\ \text{values } j \text{ for the} \\ &\quad \text{parameters in model } i \mid (HC)G].\end{aligned}$$

I state this as a (discrete) summation rather than as a (continuous) integration to make the concepts more transparent. 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 is under evaluation. It is hard to know what to say about the first of these, except that simpler models can't have higher probabilities than the more complex models in which they are nested. For example, it is a consequence of the axioms of probability theory that Pr(Jukes-Cantor | *(HC)G*) ≤ Pr(Kimura | *(HC)G*). Although Bayesians may one day take on the problem of assigning probabilities to process models, in the phylogenetic literature they so far have declined to do so; instead, they have chosen a single process model and estimated the values of the conditional probabilities found in one of the rows in Table 2. It is important to realize that the conjunction of a process model (like Jukes-Cantor) and a genealogical hypothesis does not provide instructions about how one should figure out how probable this or that setting of the parameter values is; this requires additional assumptions.

TABLE 2. Bayesians handle the nuisance parameters in a conjunction of the form "tree topology & process model" by seeking to discover the probabilities of the various settings that the nuisance parameters may have, conditional on the conjunction.

Model	(H)CG	H(CG)
12 mn	Pr[parameter values | HC(G) & 12 mn]	Pr[parameter values | H(CG) & 12 mn]
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 earlier 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 a difference between the two problems. There is an objective basis 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. And even if the Bayesian declines to assign probabilities to process models (a refusal represented in Table 2), it still is hard to see how objective probabilities are to be assigned to the parameter values in a process model. Readers will have to decide for themselves how much subjectivity they are willing to tolerate in methods of phylogenetic inference.

Let's go back to the frequentist approach to the problem of nuisance parameters, whose solution is depicted in Table 1. We have found the likeliest setting of the parameters in each conjunction of the form "genealogical hypothesis & model." We now must ask how the likelihoods of these different conjunctions compare. Likelihoods tend to increase as we ascend each column, because the presence of a larger number of parameters allows a model to achieve greater fit to data. However, the likelihoods in rows also tend to get closer together as we ascend (Lewis, 1998). And once we reach a sufficiently complex process model—say, the 12 *mn*-parameter model—the likelihoods are identical; each has achieved the maximum value of unity, no matter what the data happen to be. Does this mean that we reduce, and ultimately obliterate, our ability to discriminate between genealogical hypotheses as we make our models more realistic?

Whether this depressing thought is correct depends on the method used to compare the conjunctions depicted in Table 1. If we retain only those conjunctions whose likelihoods are maximal, we will have 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 procedure. Rather, they use the *likelihood ratio test*. According to this method, the question is whether the likelihood of a more complex model is *sufficiently* greater than the likelihood of a simpler model to justify rejecting the simpler model. Thus, 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 some of the "vertical" comparisons in Table 1, but not for "horizontal" or "diagonal" comparisons (Felsenstein, 2004: 318–319). See Burnham and Anderson (2002: 337–339) for additional criticisms.

Because of this, biologists should consider setting likelihood ratio tests to one side and shifting to a model

selection criterion such as the one proposed by Akaike (1973). AIC—the Akaike information criterion—applies to nested and non-nested models alike; it is based on a theorem that Akaike (1973) proved (see also Sakamoto et al. [1986] and Burnham and Anderson [1998]):

> An unbiased estimate of the predictive accuracy of model $M \approx \log[\Pr(\text{Data} \mid \text{L}(M)] - k$.

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 Table 1, so does the value of k. For this reason, it is not inevitable that more complex models will receive better AIC scores than simpler ones; that depends on the data. Forster and Sober (1994) suggest 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 fitted to old data. Kishino and Hasegawa (1990) have applied AIC to choice between tree topologies; see Posada and Crandall (2001) and Posada and Buckley (forthcoming) for further discussion.

In using AIC, one obtains 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—in the first instance, what one is testing 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 first 5, or 15, or 50 conjunctions at the top of the list. The larger this group of conjunctions 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, and the more robust the better. But suppose that *(HC)G* appears in the first, but not the second, of the conjunctions on this list, and then appears in the third through twentieth entries. Because AIC provides a quantitative score for each conjunction, and not just an ordering of conjunctions, one can ask what the *average effect* is of shifting from one tree topology to another. For example, in our running example, perhaps AIC scores are on average improved by moving from *H(CG)* to *(HC)G* across a range of process models. The suggestion is to treat AIC scores (or perhaps the *Akaike weights* discussed in Burnham and Anderson 1998) rather like the data that figure in applications of the analysis of variance.

Bayesians face a similar Duhemian problem if they decline to say how probable this or that process model is, conditional on a tree topology. This refusal prevents them from computing the likelihoods of tree topologies; instead, they restrict themselves to computing the average likelihood of conjunctions that have the familiar form "genealogical hypothesis & process model." However, the question can still be addressed of whether, say, *(HC)G* has a higher likelihood than *H(CG)* for each of several process models. The larger the range of process models for which this holds, the more robust *(HC)G* is. But suppose *(HC)G* is more likely than *H(CG)* for some process models but that the reverse is true for others. What can a Bayesian say about such cases?

One solution would be to embrace the idea of assigning values to conditional probabilities of the form Pr(process model | tree topology). As noted earlier, when models are nested, there is a constraint on those probabilities; for example, Pr(Jukes-Cantor | —) ≤ Pr(Kimura | —). Once this constraint is satisfied, it is unclear what, beyond the subjective convictions of the investigator, can be used to justify such assignments. Bayesians also may want to consider using the Bayesian information criterion (BIC) first derived by Schwarz (1978):

$$\Pr(\text{Data} \mid \text{Model } M) \approx \log[\Pr(\text{Data} \mid \text{L}(M)] - (k/2)\log(n),$$

where n is the size of the data. Before pressing this result into service, biologists should consider the assumptions that enter into its proof; the same is true, of course, for frequentists who contemplate using AIC. 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 criteria are in conflict. In fact, they are not, because AIC and BIC have different goals—the former estimates predictive accuracy whereas 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, one question would be whether one genealogical hypothesis dominates the others in all the conjunctions considered. However, if no genealogical hypothesis is robust in this sense, one can see what the *average effect* is on BIC scores of shifting from one tree topology to another across a range of process models.

Regardless of this difference between frequentist and Bayesian approaches to the problem of nuisance parameters, the use of model selection criteria such as AIC and BIC leads to the following conclusion—*statistical inference of genealogical hypotheses does not require one to choose a single process model and assume that it is true*. In fact, one needn't even assume that there is a true model on the list of process models considered; all may contain idealizations. 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. However, the hope of using a fully realistic process model must be abandoned, since it will contain too many parameters.

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 some guidance. Consider the two sentences

(P) Jones is poor but honest

and

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

I hope it is clear that (P) *assumes* (presupposes) that (A) is true, but that (A) does not assume that (P) is true. Notice that (P) *entails* (A)—that is, if (P) is true, then (A) must also be true. However, (A) does not entail (P); if there is a conflict between poverty and honesty, this says nothing about Jones and the characteristics he happens to have. This points to a general fact about what it means to talk about the assumptions of a proposition:

If P assumes A, then P entails A.

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).

Given this clarification of what an assumption is, we can turn to the question of what it means to ask what parsimony assumes. 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 demand 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? Those 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 has the consequence that both likelihood and parsimony are statistically inconsistent—there are settings of the parameters in that model that lead both methods to converge on a false genealogical hypothesis. Biologists who think that the Tuffley-Steel model is a legitimate model to consider—perhaps as one of a range of candidate process models—will not want to embrace the requirement of statistical consistency.

The interpretation I want to explore here maintains that parsimony is a legitimate method precisely when it is *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

(OE) 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 $\Pr_M(D \mid H_1) > \Pr_M(D \mid H_2)$.

The subscript "M" in the likelihood terms is there to remind us of the Duhemian point that phylogenetic hypotheses do not confer probabilities on data, save in the context of a process model. In fact, it is misleading to talk of parsimony and likelihood being, or failing to be, ordinally equivalent. Rather, the question is whether likelihood-when-implemented-by-an-assumed-process-model-M is or is not ordinally equivalent with parsimony. I am interested in (OE) 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, (OE) 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 about the evolutionary process comes to this—which propositions about evolution must be true, if (OE) is correct? For example, does parsimony presuppose the no common mechanism model described by Tuffley and Steel (1997)? What Tuffley and Steel demonstrated is that the no common mechanism model *suffices* for ordinal equivalence. This means that someone using that model to estimate the likelihoods of tree topologies (while using the frequentist strategy for handling nuisance parameters) will single out as the likeliest tree the very same tree as someone using parsimony on the same data. However, the Tuffley-Steel result does not show that parsimony *assumes* that the no common mechanism is true; no one has established that this model is *necessary* for ordinal equivalence to obtain (and it isn't—see below).

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 model → ordinal equivalence
→ assumptions of parsimony

Any proposition that is entailed by ordinal equivalence also must be entailed by the no common mechanism model. However, the fact that a proposition is entailed by the no common mechanism 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 significant, in view of Felsenstein's (1973, 1979) argument that a low probability of change on branches suffices for parsimony and likelihood to coincide. The Tuffley-Steel model also does not entail 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 significant, in view of Goldman's (1990) presentation of a model that makes this independence assumption and which, according to Goldman, suffices to ensure ordinal equivalence.

I have run this test for 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 *ML* and *MP* 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 interior nodes. The same question arises in connection with Farris's (1973) analysis of a model that he claims induces ordinal equivalence. Farris' model assumes very little about the evolutionary process—in particular, there is no assumption that all changes at a site have the same probability. The problem is that Farris (1973) conceives of both *ML* and *MP* as outputting a tree topology *and* an assignment of character states to interior nodes *and* an assignment of character states to all the other time slices on branches in the tree's interior. However, if *MP* outputs only a tree topology, and considers interior character states only as a means for deciding which topology is best, the arguments by Farris and Goldman do not identify models that induce ordinal equivalence (Sober, 1988; Steel and Penny, 2000).

But what of Felsenstein's (1973, 1979) arguments, which understand *ML* and *MP* as outputting only tree topologies? How are the sufficient conditions that Felsenstein identified for ordinal equivalence related to the no common mechanism model? Unlike the no common mechanism model, Felsenstein's (1973) and (1979) models do not assume that all changes at a site within a branch have the same probability. Felsenstein shows that when rates are sufficiently small, the most parsimonious tree will be the tree of maximum likelihood. This is a sufficient condition for ordinal equivalence that is disjoint from the one that Tuffley and Steel established.[5] If we use Felsenstein's (1973, 1979) results to implement the partial test described above, we can answer a question that naturally arises concerning 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 (including a change in the opposite direction), then there is no selection or any other directional process favoring one character state over any 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 on an assumed tree to assign character states to ancestors. Maddison (1991) demonstrated that likelihood and parsimony are ordinally equivalent (for quantitative characters where parsimony is interpreted to mean minimizing squared-change) in this problem if a neutral model of evolution is assumed. Sober (2002) showed that parsimony can fail to coincide with likelihood in this problem if there is directional selection. Given neutralism's connection with ordinal equivalence in the context of inferring ancestral character states, perhaps neutralism is critical for ordinal equivalence in the context of inferring tree topologies. This conjecture is refuted by Felsenstein's results. Felsenstein's (1973, 1979) models don't entail neutralism; hence, neutralism is not an assumption that parsimony makes.

Much remains to be learned about parsimony's presuppositions. The sufficient conditions for ordinal equivalence derived by Tuffley and Steel (1997) and by Felsenstein (1973, 1979) do not tell us what parsimony assumes. The partial test described here can demonstrate that this or that proposition is *not* an assumption of parsimony's, but it cannot demonstrate that a given proposition *is* one of parsimony's presuppositions. However, the criterion of ordinal equivalence allows us to describe a second test procedure that goes beyond the one just described. If a model entails that parsimony and likelihood are *not* ordinally equivalent, then parsimony assumes that that model is false. Unfortunately, this test procedure will be limited in its analytic power. A model that induces a failure of ordinal equivalence will inevitably involve a number of postulates; what one can conclude is that parsimony presupposes that at least one of them is false, but the test procedure does not say which of them parsimony assumes is false.

Another avenue of inquiry that is worth exploring derives from the fact that Tuffley and Steel (1997) and Felsenstein (1973, 1979) both use the frequentist procedure for handling nuisance parameters. What connection can be established between parsimony and likelihood when nuisance parameters are handled in a Bayesian fashion?

And finally, the quest to discover parsimony's presuppositions could be set to one side if a model of the evolutionary process could be presented that everyone grants is plausible and that suffices to induce ordinal equivalence. The demonstration of sufficiency would not, of course, show that parsimony assumes this model is true. However, people prepared to grant that the model is true will thereby have reason to conclude that parsimony is legitimate (when judged by the criterion of ordinal equivalence). *They* assume the model is true, even

[5]Felsenstein (1981) derived of ordinal equivalence in the context of a symmetrical clock model in which rates can vary among characters. This model also could be used to implement the partial test concerning what parsimony presupposes.

if *parsimony* does not, and that will suffice to justify parsimony in their eyes.

Concluding Comments

Since the early 1970s, a dispute has raged between defenders of *ML* and defenders of *MP*. 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 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 *MP* as the common enemy. But, as in other domains of human conflict, once a common enemy is perceived to be less threatening, one-time allies may turn their attention to the issues that divide them. Frequentists and Bayesians need to discuss their disagreements; it is to be hoped that they will do so without the acrimony that has characterized the debate between cladists and frequentists.

One way to begin the comparison of frequentism and Bayesianism is to see how often the two approaches license different conclusions. In the toy example presented earlier of assessing hypotheses about Mom's genotype, the decision about how to handle nuisance parameters makes an enormous difference. Is there reason to think that this difference disappears in the data sets that systematists analyze? The comparison is worth developing by using both real data sets and sets that are invented for exploratory purposes.

Frequentists and Bayesians also need to explore the use of model selection criteria such as AIC and BIC. The likelihood ratio test is limited to nested models (i.e., to some of the vertical comparisons in Table 1), whereas AIC has no such limitation. Similarly, Bayesians have so far limited themselves to considering alternative genealogical hypotheses within the context of a single process model, but BIC permits one to consider a range of process models.

The statistical analysis of *MP* is also worth developing further. This is an interesting project, both for those who are already sold on the correctness of a statistical approach and for those who think that parsimony makes sense in ways that 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 results of Tuffley and Steel (1997) and of Felsenstein (1973, 1979) are illuminating, but other models (see, for example, Felsenstein [1981] and Steel and Penny [2004]) need to be explored before the conceptual terrain can be said to be well understood.

Acknowledgments

My thanks to Thomas Buckley, Kenneth Burnham, Benoit Dayrat, Bret Larget, Paul Lewis, David Posada, Michael Steel, and an anonymous referee of this journal for useful comments. The research reported here was partially supported by National Science Foundation grant SES-9906997.

References

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* Second International Symposium on Information Theory (B. Petrov and F. Csaki, eds.). Akademiai Kiado, Budapest.

Burnham, K., and D. Anderson. 2002. Model selection and inference—A practical information-theoretic approach, 2nd Edition. Springer, New York.

Duhem, P. 1914. The aim and structure of physical theory. Princeton University Press, Princeton, New Jersey.

Edwards, A. 1972. Likelihood. Cambridge University Press, Cambridge, United Kingdom.

Farris, J. 1973. On the use of the parsimony criterion for inferring phylogenetic trees. Syst. Zool. 22:250–256.

Farris, J. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics—proceedings of the 2nd Annual Meeting of the Willi Hennig Society (N. Platnick and V. Funk, eds.). Columbia University Press, New York. Reprinted in Conceptual issues in evolutionary biology (E. Sober, ed.). MIT Press, Cambridge, Massachusetts, 1994.

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

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods can be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1979. Alternative methods of phylogenetic inference and their interrelationships. Syst. Biol. 28:49–62.

Felsenstein, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linnaean Soc. 16:183–196.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer, Sunderland, Massachusetts.

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. Br. J. Phil. Sci. 45:1–36.

Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. Syst. Biol. 39:345–361.

Huelsenbeck, J., F. Ronquist, R. Nielsen, and J. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jukes, T., and C. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

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

Kishino, H., and M. Hasegawa. 1990. Converting distance to time: Application to human evolution. Methods Enzymol. 183:550–570.

Lewis, P. 1998. Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. Pages 132–166 *in* The molecular systematics of plants II (D. Soltis and J. Doyle, eds.). Chapman and Hall, New York.

Lewis, P. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Maddison, W. 1991. Squared-change parsimony reconstructions of ancestral States for continuous-valued characters on a phylogenetic tree. Syst. Zool. 40:304–314.

Page, R., and E. Holmes. 1998. Molecular evolution—A phylogenetic approach. Blackwell, Oxford.

Posada, D., and T. Buckley. (Forthcoming). Model selection and model averaging in phylogenetics—Advantages of AIC and Bayesian approaches over likelihood ratio tests. Syst. Biol.

Posada, D., and T. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Quine, W. 1953. Two dogmas of empiricism. Pages 20–46 *in* From a logical point of view. Harvard University Press, Cambridge, Massachusetts.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. Springer, New York.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–465.
Semple, C., and M. Steel. 2003. Phylogenetics. Oxford University Press, Oxford.
Sober, E. 1988. Reconstructing the past—Parsimony, evolution, and inference. MIT Press, Cambridge, Massachusetts.
Sober, E. 2002. Reconstructing ancestral character states—A likelihood perspective on cladistic parsimony. The Monist 85:156–176.
Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.
Steel, M., and D. Penny. 2004. Two further links between *MP* and *ML* under the poisson model. Appl. Math. Lett. in press.
Swofford, D., G. Olsen, P. Paddell, and D. Hills. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. Hillis, C. Moritz, and B. Mable, eds.). Sinauer, Sunderland, Massachusetts.
Tuffley, C., and Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.

First submitted 29 July 2003; revisions returned 17 November 2003; final acceptance 4 January 2004
Associate Editor: Mike Steel

Syst. Biol. 53(4):653–661, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490472959

The Role of Morphological Data in Phylogeny Reconstruction

JOHN J. WIENS

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA; E-mail: wiensj@life.bio.sunysb.edu

We live in the age of comparative genomics, and it may seem that there is not much point in reconstructing phylogenies using morphological data anymore. As more and more genes and genomes are being sequenced, the possibility that thousands or even millions of informative, independently evolving molecular characters can be brought to bear on a given phylogenetic problem is quickly becoming a reality (e.g., Rokas et al., 2003). Given the rate that new sequence data are being added, and the rate at which new innovations continue to accelerate this process, it seems possible that in the not-too-distant future we will be able to have a perfectly accurate and well-supported phylogeny of most living species on earth using molecular data alone. So why bother with morphology?

A recent paper by Scotland et al. (2003; SEA hereafter) offered a reappraisal of the role of morphology in phylogeny reconstruction. This is certainly an important and timely topic to discuss, and their main thesis is bold and controversial. They state that "We view any attempt to include more morphological data in phylogeny reconstruction as inherently problematic" (p. 545). Unfortunately, most of their arguments are based on unsupported speculation, and they fail to mention numerous studies that clearly contradict their conclusions. Given that many of their comments are written as responses to book chapters written by my collaborators and myself (e.g., Hillis and Wiens, 2000; Poe and Wiens, 2000; Wiens, 2000a), I feel obligated to elucidate some of these problems. Many of the issues raised are central to how systematics is done and will be conducted in the future. I will argue that, despite many undeniable advantages of molecular data, it is still absolutely necessary that we continue to collect additional morphological data for phylogenetic analysis, and continue to improve our methods for morphology-based phylogenetics. Note that Jenner (2004) has provided an independent rebuttal of the SEA paper, and he describes a large number of substantive criticisms which show only limited overlap with my own.

WHY WE STILL NEED TO COLLECT MORE MORPHOLOGICAL DATA

The Future

There are many reasons to continue to do morphological phylogenetics. But given the incredible rate of advances in molecular systematics, it may be useful to divide these reasons into those pertaining to the present and future. Those pertaining to the future may actually be the most relevant, because many present-day limitations of molecular phylogenetics seem likely to be overcome very soon. I will focus on the putative future first, and then deal with present-day issues.

The most compelling reason to continue to collect morphological data long into the future is to resolve the phylogenetic relationships of fossil taxa and their relationships to living taxa (e.g., Maddison, 1996; Hillis and Wiens, 2000; Jenner, 2004). The reconstructed Tree of Life must include fossil taxa. Considering all the species that have ever evolved, most are now extinct (>99% according to some estimates; Novacek and Wheeler, 1992), and many extinct groups were diverse, ecologically important, and very distinct from their closest living relatives. For now and the immediate future, the relationships of most fossil taxa can only be determined through phylogenetic analysis of morphological data (despite impressive molecular studies of very recent fossil taxa). Contrary to what SEA imply (p. 543), fossils are not merely important for their potential to help resolve relationships of living taxa, and Hillis and Wiens (2000) did not advocate thorough taxon sampling solely because of its potential benefits for phylogeny estimation.

Our understanding of the rate and timing of macroevolutionary processes in both living and fossil taxa also