Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics

Mike Steel[*] *and David Penny*[†]

[*]Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand; and [†]Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

Methods such as maximum parsimony (MP) are frequently criticized as being statistically unsound and not being based on any ‘‘model.’’ On the other hand, advocates of MP claim that maximum likelihood (ML) has some fundamental problems. Here, we explore the connection between the different versions of MP and ML methods, particularly in light of recent theoretical results. We describe links between the two methods—for example, we describe how MP can be regarded as an ML method when there is no common mechanism between sites (such as might occur with morphological data and certain forms of molecular data). In the process, we clarify certain historical points of disagreement between proponents of the two methodologies, including a discussion of several forms of the ML optimality criterion. We also describe some additional results that shed light on how much needs to be assumed about underling models of sequence evolution in order to successfully reconstruct evolutionary trees.

Introduction

Maximum parsimony (MP) is a popular technique for phylogeny reconstruction. However, MP is often criticized as being a statistically unsound method and one that fails to make explicit an underlying ‘‘model’’ of evolution. Discussion is further clouded by claims that MP variously is, or is not, a form of maximum likelihood (ML) and the promotion of ‘‘zones’’ within which either method performs worse than the other in recovering the true tree. There is little agreement on how, or even whether, MP should be justified. According to Edwards (1996), who prefers to call MP the ‘‘method of minimum evolution,’’ the method was introduced by himself and Cavalli-Sforza in 1963 (in the context of continuous characters) merely as a computational approximation for ML, and not as a method of choice in its own right.

However, others (e.g., Farris, Kluge, and Eckardt 1970; Sober 1988) claim that MP is the preferred method of tree reconstruction. Advocates of this viewpoint sometimes appeal to Willi Hennig’s writings on phylogenetic inference or, alternatively, to the Principle of Parsimony. The latter is a minimalist principle, sometimes also referred to as ‘‘Ockham’s razor,’’ and states that one should prefer simpler explanations (requiring fewer assumptions) over more complex, ad hoc ones. In phylogeny reconstruction, this principle has been applied in two ways. One emphasizes the feature that MP favors the tree requiring the fewest evolutionary events (such as mutations) to explain the observed data and thus is, in some sense, the ‘‘simplest,’’ or an ‘‘optimal’’ description of the data. A second appeal to the Principle of Parsimony is to assume as little as possible about any underlying model or mechanism for evolution. Actually, we will see that this second application of the Principle of Parsimony can also be used, instead, as an argument in favor of the more usual forms of ML.

Some authors (e.g., Farris 1973; Sober 1985, 1988) have also presented explicit statistical arguments in favor of MP based on underlying evolutionary models. Still others have undertaken the more modest task of providing a statistical framework for using MP (Cavender 1978, 1981; Kishino and Hasegawa 1989; Maddison and Slatkin 1991; Steel, Hendy, and Penny 1992; Archie and Felsenstein 1993).

The simplicity of a method like MP (and its embellishments that allow weightings on characters and transition types), together with its apparent lack of assumption involving underlying models, made it popular in phylogeny, particularly in the 1970s and the 1980s. Furthermore, it is possible to state sufficient conditions on the process by which characters evolve so that MP will recover the true tree. Essentially, these conditions amount to requiring that convergent evolution and reversals occur in (sufficiently) low numbers in comparison with the characters that identify edges of the tree (a more precise formulation is given by the lemma given in section (a) of the appendix). The main problem with such simple criteria is that they are very unlikely to be satisfied for most real data sets, and even when they are, it may be impossible to tell this directly from the data (without knowing in advance the true tree).

MP is still widely used, but model-based approaches have come to rival, and even dominate, phylogenetic methodology, particularly over the last decade. While ML is the leading alternative, other approaches include distance-based methods that use transformed or inferred distances, for example, logdet/paralinear distances (see Swofford et al. [1996] for a review of distance methods which are outside the scope of this overview of parsimony and likelihood). One justification for model-based approaches was the classic and much-cited statistical inconsistency of MP due to Felsenstein’s paper (1978), which demonstrated that if sequence sites evolved under certain models and combinations of rates, then MP would favor an incorrect tree. Furthermore, the probability of selecting an incorrect tree would tend to 1 as the sequence length grew (this phenomenon of statistical inconsistency will be discussed further in *When Is MP Statistically Consistent?* below). The conditions that Felsenstein used—a particular combination of short and

Key words: phylogeny, maximum likelihood, maximum parsimony, site substitution models.

Address for correspondence and reprints: Mike Steel, Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: m.steel@math.canterbury.ac.nz.

Mol. Biol. Evol. 17(6):839–850. 2000

long branches—have become the deliciously sinister-sounding ‘‘Felsenstein Zone.’’

Both Felsenstein (1973) and Yang (1994) informally claimed that the nonexistence of any such zone within which ML would be statistically inconsistent (although this assertion was questioned by Sober [1988, chapter 5]). Indeed, the statistical consistency for ML (when the underlying model had no rate distribution across sites and this same model was then also used in the ML method to reconstruct the tree) was rigorously established only recently by Chang (1996*a*) (and, for more special types of models, by Rogers [1997]). Note that the use of the ‘‘correct’’ model (the same as the model used to generate the data) is essential to the proof that ML is consistent. For the biologist, this is a mixed blessing: although one may seldom know the correct model of evolution, the more one knows about the evolutionary process, the better we would expect the chances to be of avoiding a zone of inconsistency by analyzing the data correctly.

Nevertheless, ML methodology enjoys far from universal acceptance. Objections to ML include the following: Concern about the validity and exact form of any underlying stochastic model (e.g., there is concern as to the choice of underlying parameters/distributions and as to the idea that by selecting the appropriate model, perhaps one could reconstruct any favored tree);

The concern that ML estimation of a tree (and statistical tests between different trees) that involves optimizing ‘‘nuisance (supplementary) parameters’’ is statistically problematic.

Suggestions that the Felsenstein Zone rarely, if ever, arises for real data.

The existence of a ‘‘Farris Zone,’’ where MP outperforms ML.

The analysis of new types of genome data, e.g., gene order and short interspersed nuclear elements (SINEs), for which MP may be more appropriate.

Concern about the computational complexity of ML. Even on a given tree, optimizing the likelihood can be problematic (unlike with MP, for which Fitch’s [1971*b*] algorithm provides a linear time algorithm for computing the parsimony score).

In this paper we will explore most of these objections and survey some recent theoretical results that shed light on the interplay between the two methodologies and on the limits of what one can hope to achieve in phylogeny reconstruction.

Before proceeding, it is necessary to clarify some terminology. We have already pointed out that the Principle of Parsimony (Ockham’s razor) has two general applications, one as justification for an attempt to analyze data without reference to an underlying model, the other as a tree selection process (MP) to minimize mutations. However, this latter usage combines the two aspects of selecting a tree with a minimal number of mutations and using only observed data (not corrected for any multiple changes). However, these are independent concepts and can be used in different combinations. For example, minimization of the number of mutations can be applied after correction for multiple changes (corrected parsimony; Penny et al. 1996); also, distance methods (such as neighbor joining) can be used on the observed (uncorrected) distances. This separates the optimality criterion for selecting a tree from assumptions about the mechanisms of sequence evolution. Some forms of weighted parsimony make certain assumptions about the mechanism of evolution, for example, giving transversions more weight than transitions (see Swofford et al. 1996).

In general, we prefer to treat a ‘‘method’’ for inferring evolutionary trees as being composed of three largely independent parts: the choice of optimality criterion, the search strategy over the space of trees, and assumptions about the model of evolution. It is useful to make a three-way division of the model of evolution. This consists of a tree *T* (or, more generally, a graph, when median networks or splits graphs are considered), a stochastic mechanism of evolution (such as whether or not it is neutral, Kimura 3ST, whether it exhibits rate heterogeneity), and the initial conditions (e.g., interspeciation times or rates on each edge [branch] of the tree).

An additional factor is that the researcher may be hoping to recover different aspects of the model. Most frequently, perhaps it is just the unweighted tree, regardless of the amount of mutation on each edge of the tree. In addition, the tree will usually be unrooted unless an outgroup or an assumption about a molecular clock is used. Frequently, however, the rates of mutation will be required in order to estimate times of divergence. Others will also wish to estimate the character states at the internal nodes. It is thus too simple just to compare ‘‘parsimony’’ and ‘‘likelihood.’’ Indeed, likelihood itself comes in many flavors, and these will be discussed next. The usual form of ML is ‘‘maximum average likelihood,’’ an example of ‘‘maximum relative likelihood.’’ These and other distinctions we discuss below have also been noted by others, in particular, Goldman (1990), Felsenstein (1973), and Barry and Hartigan (1987).

What Is ML, and What Does it Maximize?

The likelihood of the hypothesis *H*, given data *D* and a specific model, is proportional to $P(D|H)$, the conditional probability of observing *D* given that *H* is correct (Edwards 1972). An ML method of inference selects the hypothesis *H* that maximizes the likelihood function for the data *D* (given the specified mechanism).

In the context of phylogeny reconstruction from sequences, *D* typically counts the number of ‘‘site patterns’’ that occur in a collection of aligned sequences. The order in which these patterns occur (and the phylogenetic information that this might convey) is usually discarded, although some authors have explicitly incorporated this into their analysis (e.g., Felsenstein and Churchill 1996; Thorne, Goldman, and Jones 1996; Yang 1996*b*; Halpern and Bruno 1998). The hypothesis *H* is usually the discrete phylogeny (unweighted tree) *T*, and the model is some stochastic process for site substitution (or, more generally, genome transformation, if insertions and deletions are allowed).

Unfortunately, $P(D|T)$, and hence the likelihood of T, requires more information to specify it, since the probability of evolving D depends on further parameters, sometimes referred to as ''nuisance parameters.'' In order to talk about $P(D|T)$, we need to either specify these parameters or place some prior distribution on them. The word ''nuisance'' is a little misleading. It does not imply that these parameters are of no interest, but rather that they need to be considered even if all one wants to know about is the tree T. Examples of such parameters in molecular phylogenetics are the edge lengths (interspeciation times and rates of mutation on the edges), parameters associated with the substitution matrix (e.g., transition/transversion bias), and parameters that describe how rates vary across sites.

Nuisance parameters (and the associated problems they cause) arise widely in many statistical settings. They have been discussed in the phylogeny setting by several authors, perhaps most lucidly by Goldman (1990). Nuisance parameters may further be classified into ''structural'' and ''incidental'' parameters. Structural parameters influence all (or nearly all) of the sites, while incidental parameters influence only one or a few. Structural parameters typically correspond to the edge (branch) lengths and parameters that constrain the substitution process (e.g., the transition/transversion bias). Typically, such parameters are either selected to maximize the likelihood or estimated directly from the data. Incidental nuisance parameters arise (1) if we wish to hypothesize a particular choice of sequences to appear at internal vertices of the tree, in which case we need to specify states for each site, or (2) if the process varies from site to site. We will discuss both these situations below. In any case, for a model of sequence evolution, we will represent nuisance parameters collectively by θ.

Two frequent assumptions concerning substitution models are that aligned sites evolve independently and according to identical processes—the so-called ''i.i.d.'' assumption. Note that the i.i.d. assumption still allows sites to evolve at different rates by regarding the rate of a site as being randomly and independently selected from an appropriate distribution (such as a gamma distribution). Of course, in real sequences, there is clustering of ''conserved'' and ''hypervariable'' sites (so the real process is definitely not i.i.d. across sites), but when one passes to the frequencies of site patterns (i.e., the data D), the process can be modeled by an i.i.d. process. Similarly, certain covarion-style mechanisms (where sites can alternate between invariable and variable during evolution) can be modeled using an i.i.d. process (Tuffley and Steel 1997*b*), even though the original covarion model (e.g., Fitch 1971*a*) implied explicit dependency between sites.

The i.i.d. assumption allows one to readily compute $P(D|T, \theta)$ by identifying this with the product of the probabilities of evolving each particular site. Occasionally, more intricate models have been proposed and analyzed. These include models that allow a limited degree of nonindependence between sites (e.g., pairwise interactions in stem regions; Schöniger and von Haeseler 1994) and models that work with nonaligned sequences and explicitly model the insertion-deletion process as well as the site substitution process (Thorne, Kishino, and Felsenstein 1992).

Maximum Integrated Likelihood Versus Maximum Relative Likelihood

If the nuisance parameters θ and the phylogeny T are generated according to some known prior distribution (e.g., a Yule pure-birth process) one can formally integrate out these nuisance parameters, and thereby take $P(D|T)$ to be this average value. That is, if $\Phi(\theta|T)$ denotes the distribution function of the nuisance parameters conditional on the underlying tree T, then

$$P(D|T) = \int P(D|T, \theta)\, d\Phi(\theta|T).$$

This approach is sometimes referred to as ''integrated likelihood,'' and we will refer to a tree T that maximizes $P(D|T)$ as a maximum integrated likelihood (MIL) tree. MIL, and, more generally, the assignment of posterior probabilities to trees based on sequence data (using Markov chain Monte Carlo techniques to approximate the integral in the above equation), has been independently developed by several authors recently, in particular, Yang and Rannala (1997) and Mau, Newton, and Larget (1999).

Assume for the moment that one possesses such a prior distribution (e.g., based on a Yule process). A natural question arises: namely, in what sense is maximum integrated likelihood an optimal method for selecting a tree? In particular, is it the method that is most likely (on average) to return us the true tree? In order to formalize this question, suppose we have a tree reconstruction method, and we apply it to sequences that have been generated by a model with underlying parameters T and θ. The reconstruction probability, denoted $\rho(M, T, \theta)$, is the probability that the sequences so generated return the correct tree T when method M is applied. Since we have a distribution on trees and the nuisance parameters, let $\rho(M)$ denote the expected reconstruction probability of method M, obtained by integrating $\rho(M, T, \theta)$ over the joint parameter space. That is,

$$\rho(M) = E[\rho(M, T, \theta)] = \sum_T p(T) \int \rho(M, T, \theta)\, d\Phi(\theta|T),$$

where $p(T)$ is the probability of the tree T under the prior distribution (we will assume that only binary trees have positive probability). The following theorem precisely describes the method that maximizes the expected reconstruction probability (for a proof of this, see Székely and Steel 1999).

THEOREM 1. *Under the conditions described, the method M that maximizes the expected reconstruction probability* $\rho(M)$ *is precisely that method that selects, for any data D, the tree(s) T that maximizes* $p(T)P(D|T)$.

This tree(s) that maximizes $p(T)P(D|T)$ is sometimes referred to as the maximum a posterior probability (MAP) estimate. It is precisely the MIL tree(s) whenever the prior distribution on binary trees is uniform (i.e., when all binary trees are equally likely). Consequently, assuming that the prior distribution assigns equal probabilities to all binary trees, MIL maximizes one's average chance of recovering the correct tree. However, if the distribution on binary trees is not uniform—for example, if it is described by a Yule process—then the optimal selection criteria are slightly different. In any case, an obvious question is that of how to agree upon a biologically reasonable distribution on trees and parameters.

The alternative approach, which is more widely adopted, is sometimes called maximum relative likelihood (MRL). One simply assumes that the nuisance parameters take values that, simultaneously with an optimal tree T, maximize $P(D|T, \theta)$. Usually, one then discards θ and outputs just the tree(s) T. Such an approach can be problematic in general statistical settings where D depends on both continuous (nuisance) parameters and a discrete parameter x of interest. In this situation, there may be one ''unlikely'' value of θ that, for $x = x_1$, gives a higher $P(D|x, \theta)$ value that $\max_\theta P(D|x_2, \theta)$, yet for most ''likely'' values of θ, the probability $P(D|x_1, \theta)$ is less than $P(D|x_2, \theta)$. This property means that MRL may make selections different from those of MIL, and this seems to have been a fundamental issue in the exchange between Felsenstein and Sober (1986) on the relative merits of MP and ML. Moreover, in the phylogenetic setting, MRL may select different trees from the MIL method described above, even when all binary trees are equally likely (at least for certain distributions on the edge parameters of the tree). An example of this is described at the end of *Can MP Outperform M$_{av}$L?* below.

For the remainder of this paper, we will generally assume there is no prior distribution given for trees and edge parameters, and so all forms of ML involve MRL. With this in mind, we review some further distinctions.

Maximum Average Likelihood, Most-Parsimonious Likelihood, and Evolutionary Pathway Likelihood

In fitting sequence data to a tree, the sequences at the leaves (tips) of the tree are given, but those at the internal vertices (speciation or branching points) of the tree are not. In the usual implementation of MRL in molecular phylogenetics, one effectively averages over all possible assignments of sequences to these internal vertices. Following Barry and Hartigan (1987), we call this maximum average likelihood, and we denote it as M$_{av}$L.

However, one could also assign sequences to the internal vertices (along with the other parameters) so as to maximize the likelihood. Such an approach was explicitly suggested by Barry and Hartigan (1987), who called it ''most-parsimonious likelihood'' to distinguish it from maximum average likelihood (see fig. 1*a* and *b*). They remarked that most-parsimonious likelihood ''is therefore similar to the maximum parsimony fitting technique.'' However, it differs from MP in that the other parameters (e,g., edge lengths) must be fixed across all the characters. Likelihood calculations that place sequences at the internal vertices of a fixed tree have also been explored by other authors (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996; Pagel 1999) for whom the interest has been primarily in reconstructing, say, ancestral sequences of proteins (or other characters) on a given tree, rather than in selecting an optimal tree.

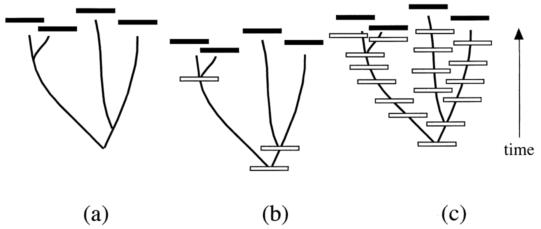

FIG. 1.—Schematic representations of three forms of maximum likelihood. *a,* Maximum average likelihood (M$_{av}$L), where all possible sequences at the internal vertices contribute to the likelihood. *b,* Most-parsimonious likelihood, in which sequences are placed at the internal vertices to maximize the likelihood. *c,* Evolutionary pathway likelihood, in which sequences are placed at each position throughout the tree to maximize the likelihood.

We pause here to note that Goldman (1990) has already noted one link between MP and most-parsimonious likelihood. He showed that under a symmetric two-state mutation model, if one imposed the rather artificial constraint that the mutation probability associated with each edge of any binary tree is set equal to some value p, then the MP tree(s) were exactly the most- parsimonious likelihood trees. This result applies either with p fixed or allowing p to be optimized.

Given the most-parsimonious likelihood approach, one might ask, what is so special about the sequences at the internal vertices of the tree? That is, perhaps one might carry the approach further and select sequences for each time interval right through the tree (jointly with the other parameters) to maximize the probability of observing the given sequences at the leaves. Thus, one would associate along each edge of the tree a series of sequences, corresponding to their evolution at frequently sampled time intervals (see fig. 1*c*).

Such an approach was suggested by Farris (1973), and it was subsequently referred to as an ''evolutionary pathway'' approach, since it is a complete specification of the sequences through time. Farris (1973) showed that the tree(s) that maximizes the likelihood in this sense is exactly the MP tree. Indeed, the argument is straightforward and requires few assumptions regarding the underlying model—in particular, it does not require any assumption about mutations occurring at a slow rate (only that they occur at a continuous rate) or edge lengths that are constrained in any way. Also, the equivalence with MP holds with the edge lengths either specified or allowed to be optimized. Of course, there will generally be a huge (potentially infinite) choice of possible evolutionary pathways of maximal probability—however, this is not a problem if the value of this maximal probability is all that is being used to select trees. As noted by Felsenstein (1978) (see also Sober 1988, p. 160), the distinction between M$_{av}$L and Farris' evolutionary pathway likelihood is crucial for reconciling the apparent paradox between Felsenstein's claim that ML (but not MP) is statistically consistent with Farris' claim that MP is an ML method. Both claims are correct; they are simply referring to different forms of ML.

Does MP = $M_{av}L$ Under Some Model?

Most-parsimonious likelihood and evolutionary pathway likelihood both entail the specification of a choice of sequences to points inside the tree. Although a particular selection of sequences may be the most probable, the attraction of $M_{av}L$ is that it effectively allows all possible assignments of sequences to the interior of the tree. These are weighted according to their probability and then summed up to give the marginal probability of evolving the sequences observed at the leaves. The question arises then as to whether MP can be regarded as a $M_{av}L$ method under some model.

Suppose we take the simplest type of substitution model at a particular site, a Poisson model in which each of the possible substitutions at that site occurs with equal probability. This model, sometimes called the Neyman model (or the Jukes-Cantor model, when dealing with exactly four states) will be referred to here simply as the Poisson model. Now suppose the rates of evolution on each branch of the tree can vary freely from site to site. In this case, we have some constraints on the underlying type of substitution model (i.e., Jukes-Cantor type) but no constraints on the edge parameters from site to site. We refer to this as ''no common mechanism.'' This is even more general than the type of approach considered by Olsen (see Swofford et al. 1996, p. 443) in which the rate at which a site evolves can vary freely from site to site, but the ratios of the edge lengths are equal across the sites. For the Poisson model with no common mechanism (not even the same rates for different characters) the following result applies.

THEOREM 2. *Under the model described* (*with no common mechanism*), *the maximum average likelihood tree*(*s*) *is precisely the maximum parsimony tree*(*s*).

This result, by Tuffley and Steel (1997*a*), generalizes an earlier special case by Penny et al. (1994) . The significance of this theorem should not be taken as any special justification for MP over usual implementations of ML, nor does it imply that MP trees are the same as those that ML would produce under the ''usual'' models (e.g., Jukes-Cantor with fixed edge lengths). Rather, the significance of the theorem is of a more philosophical nature, as it describes a model in which MP can be regarded as an ML method in the usual ''average'' ML setting (i.e., where one does not select particular sequences for the internal vertices as part of the optimization step).

The argument used to establish Theorem 2 also shows that, under a Poisson model, if we are given just a tree and a single character (and no information as to the edge lengths), the ML estimate of the state at any internal vertex of the tree (given the states at the leaves of the tree) is precisely the MP estimate. For a further link between ML and MP, suppose we take any sequence data and add a sufficiently large number of unvaried sites. Then, under a Poisson model, the ML tree of this extended data set is always an MP tree. For details and justification of these last two results see Tuffley and Steel (1997*a*).

Of course, this type of underlying model (in Theorem 2) is almost certainly too flexible, since it allows many new parameters for each edge. It might be regarded as the model one might start with if one knew virtually nothing about any common underlying mechanism linking the evolution of different characters on a tree (e.g., as with some morphological characters).

For processes like nucleotide substitution, as one learns more about the common mechanisms involved, it would seem desirable to use this information. This would lead to the more usual implementations of $M_{av}L$ where the model parameters (such as edge lengths) are constant across sites. Indeed, advocates of Ockham's razor (the Principle of Parsimony) might well invoke the principle at this point, as illustrated by the following example. Consider sequences of a pseudogene, with each sequence being over 10,000 nt long (Miyamoto et al. 1988). As a first approximation, there is no selection at any of the sites, and therefore it is more ''parsimonious'' to assume one common mechanism for all sites rather than 10,000 different mechanisms, one for each site. In such a case, the Principle of Parsimony would support the usual $M_{av}L$ over using data uncorrected for multiple changes.

Again, this conclusion must be taken with care. Such a model may not apply to other sequence data and would not often apply to morphological data (e.g., where the evolution of numbers of legs may differ from that of wing color). It is clear that we still need to learn more about the processes leading to different types of insertion and deletion events in sequence data to postulate a common mechanism.

In summary, this subsection suggests two ironies: first, that the parsimonious approach suggested by Ockham's razor can, given information of a common mechanism, support the usual forms of ML over MP for sequence data. Second, by Theorem 2, when we generalize traditional substitution models (like Jukes-Cantor) sufficiently far—namely, to allow different edge parameters at different sites—the usual ML approach arrives back at MP.

When is MP Statistically Consistent?

Given a model of site substitution, a tree reconstruction method is said to be statistically consistent if the probability of its reconstructing the true tree converges to certainty as the sequence length tends to infinity, regardless of what value the structural nuisance parameters take. Note that the reconstructed tree is considered correct if it matches the generating tree up to the position of any root vertex in the latter tree, since the root generally cannot be determined without additional assumptions (e,g., a molecular clock). The concept of statistical consistency is always relative to the model in question, and methods that are consistent for one class of models may be inconsistent for others (Chang 1996*b*).

Statistical consistency is often seen as a desirable, if not essential, property of an estimator in most statistical settings. However this viewpoint is sometimes questioned in phylogenetics (e.g., by Sober 1985, 1988), where sequences are of a pregiven finite length, and so the concept of collecting more data may not be applicable. Some methods (e.g., linear invariant methods) that are statistically consistent can perform very poorly on realistic-length sequences compared with methods that can be statistically inconsistent. Thus, more recent studies (e.g., Charleston, Hendy, and Penny 1994; Hillis 1996; Kim 1996; Rice and Warnow 1997) have instead tended to concentrate on the comparison of different methods and their corresponding reconstruction probabilities (the probability that the method reconstructs the true tree from sequences of a given length that evolve according to the model) or the related probabilities of reconstructing the true tree up to a given measure of accuracy.

Nevertheless, the issue of consistency has tended to dominate much of the discussion concerning the relative merits of ML over MP, particularly since Felsenstein's (1978) classic paper showing that MP (and the related maximum-compatibility method) can be inconsistent. However, distance methods applied to uncorrected data can also be inconsistent; indeed, under the symmetric two-state model, the conditions for inconsistency of some standard distance methods (applied to uncorrected distances) are identical to those for MP on four-taxon trees (Penny, Hendy, and Steel 1991).

A seductive, but erroneous, belief is that if the mutation probabilities on the edges are all sufficiently small, then MP is statistically consistent under simple models. However, Felsenstein's (1978) counterexample allows arbitrarily low mutation probabilities. Nevertheless, if one fixes the relative branch lengths on any tree, one can easily show that if the rate of substitution is sufficiently small, then MP is statistically consistent. For four sequences, it is possible to say exactly when MP will be statistically consistent (in terms of the edge parameters), at least for simple models such as the symmetric two-state model (Penny, Hendy, and Steel 1991). An explicit sufficient condition for the statistical consistency of MP under a Poisson model with any number of states, and for general numbers of sequences, is described by Steel (1999).

If the branch lengths satisfy a molecular clock, then the Felsenstein Zone disappears for four-taxon trees, at least for symmetric models like the Kimura 3ST and Jukes-Cantor models (see Hendy and Penny 1989; Steel, Hendy, and Penny 1998). Unfortunately, the molecular clock does not rescue MP, since it can fail on five-taxon trees and, more dramatically, on six-taxon trees (in this latter case, the edge lengths can be made arbitrarily small), as shown by Hendy and Penny (1989).

A curious consequence of these results arises when a molecular clock applies. If one uses MP on the entire data set, the method may be statistically inconsistent, yet if one had used MP to reconstruct trees on quartets of taxa and then combined these quartet trees, the method would be statistically consistent. Note that, just as with distances, it is possible for some models (e.g., the Kimura 3ST model) to transform sequence data so that MP applied to this new data will always be consistent (this approach, called ''corrected parsimony,'' is described in Steel, Penny, and Hendy [1993] and Penny et al. [1996]).

Felsenstein's (1978) original demonstration of the statistical inconsistency of MP involved the interplay of long and short edges, where the edge ''length'' refers to the expected number of mutations on the edge (i.e., the product, for each edge, of the mutation rate with the corresponding timescale). However, one can also construct zones of inconsistency for MP for other reasons—for example, when the process of substitution exhibits nonstationarity across the tree (Lockhart et al. 1994). In this case, the mutation rates may be constant and low across the tree. Of course such nonstationarity may also be a problem for ML if the model used in the ML analysis is stationary across the tree.

An unresolved issue is to what extent such inconsistency occurs with biological (as distinct from computer-simulated) sequence data. Suggested examples of tree-building inconsistency arising from the use of inappropriate analysis models include those of Lockhart et al. (1996), Van de Peer et al. (1996), Penny and Hasegawa (1997), and Huelsenbeck (1998).

A further relevant factor is the size of the state space of characters. With site substitutions, one generally has a state space of size 2 (purines/pyrimidines) or, more usually, 4 (the 4 nt), while for amino acid and codon data, the state space has size 20 or 64, respectively. With other types of genomic data—for example, gene order (Blanchette, Kunisawa, and Sankoff 1999), SINEs (Nikaido, Rooney, and Okada 1999)—there is a much larger state space. In this case, if the states evolve by a simple Markov model, then one might expect MP (and related methods like maximum compatibility) to behave better, since there is less likelihood of returning to the same state that was present earlier in the tree. We formalize this as follows.

Suppose, for example, we generate characters independently and by an identical process according to a tree-based Markov model, in which there are r states that evolve on a tree T with n leaves. We will suppose that the probability of a mutation on an edge e of the tree, conditional on there having been any given number of mutations earlier in the tree, lies strictly between a and b, where $0 < a \leq b < 1$. We will also suppose that, conditional on (1) a mutation occurring on edge $e = (u, v)$ and (2) given the state at u, the probability that the state at v is any one of the particular $r - 1$ alternative states is at most $c/(r - 1)$ for some constant c. For example, in a Poisson model, where each of the $r - 1$ different states is equally likely to be selected if a mutation occurs, we have $c = 1$. This model allows some transition events to have very low (or zero) probability, since we only require $c/(r - 1)$ to be an upper bound to these conditional transition probabilities.

We summarize the relevant constraints on this model by the quadruple (n, a, b, c), although other parameters may also be involved in specifying the model. We have the following result, proved in section (a) of the appendix.

THEOREM 3. *If the number of states* (*r*) *is large enough* (*relative to the other constraints n, a, b, c*), *then MP is statistically consistent for all binary trees with n leaves.*

Thus, for simple mutation models with bounded mutation probabilities, if the state space is large enough, then there is hope of escaping the Felsenstein Zone. However, this claim needs qualifying: it does not imply that any simple enlargement of the state space will automatically make MP statistically consistent. For example, suppose one enlarges the state space by considering pairs (2-tuples) or triples (3-tuples) or, more generally, *k*-tuples of sites (in which case the size of the state space *r* is 4^k if we have four-state sites). We suppose that changing one *k*-tuple of states into a different pair of states costs 1 unit regardless of the number of site changes involved. Note that MP applied to pairs (or, more generally, to *k*-tuples) of sites may lead to different trees than MP applied to single sites, even for four sequences. Nevertheless, for four sequences, MP will be consistent when applied to pairs of sites if and only if it is statistically consistent on the original single site data. Formally we have:

THEOREM 3A. *For four sequences and any i.i.d. model of sequence evolution, MP is statistically consistent on k-tuple–site data if and only if MP is statistically consistent on single-site data.*

A proof of Theorem 3a is given in section (b) of the appendix. Note that Theorem 3a does not contradict Theorem 3, since if we take *k*-tuples of sites, then the effective mutation probability increases toward 1 as *k* increases, so *b* is not fixed as *r* grows (i.e., as we put the sites together, the effective rate of mutation increases). We note in passing that a simple corollary of Theorem 3 is the consistency of MP under the type of ‘‘infinite-sites’’ model employed in population genetics.

Leaving MP briefly, one can also consider the consequences of a molecular clock on tree reconstruction methods that use uncorrected distances (i.e., the distance between each pair of sequences is taken to be the proportion of sites at which there is a substitution). In this case, under most models, even those that allow an (unknown!) distribution of rates across sites, the uncorrected distances will, in expectation, already be treelike. Thus, there is no need to correct them, and to do so can be problematic since (1) the correction depends on the (unknown) distribution of rates across sites, and (2) the corrected distances typically will have higher variance (and be biased upward) compared with the uncorrected distances. Formally stated (a proof is given in section (c) of the appendix), we have the following result, where by a ‘‘standard’’ site substitution model we mean a model that satisfies two conditions, namely, that it is stationary (unvaried across the tree) and reversible (the process appears the same whether viewed into the past or into the future).

THEOREM 4. *For standard site substitution models with a distribution of rates across sites, the expected uncorrected Hamming* (*observed*) *distances between pairs of sequences are additive on the underlying tree.*

Thus, if a molecular clock applies, then as far as reconstructing the tree is concerned (without regard to branch lengths), it may be preferable to work with uncorrected distances. Once the tree is reconstructed, it is clearly preferable for the estimation of the branch lengths to use the corrected distances (or ML estimation) instead of the uncorrected distances.

Can MP Outperform $M_{av}L$?

It is easy to construct examples where $M_{av}L$ will be inconsistent if the model used in the ML analysis differs from the model that generated the sequences. However, some investigators have noted that MP can perform better than $M_{av}L$, even when the underlying model matches the generating model (Waddell 1996; Yang 1996*a*; Huelsenbeck 1998; Siddall 1998).

To make this idea more precise, by the ‘‘performance’’ of a tree reconstruction method *M* (on sequence data generated under a tree-indexed Markov model) we again mean the reconstruction probability $\rho(M, T, \theta)$ described in *What Is ML, and What Does it Maximize?* (the probability the method will correctly return the true tree *T*). This quantity depends not just on *M* but also on *T* and the parameters on the edges of the tree. Now, for each tree *T*, there exist parameters for which MP will have a higher probability of returning the ‘‘true tree’’ *T* than $M_{av}L$. Of course, it is trivial to construct a method that can have a higher reconstruction probability than $M_{av}L$ for a given underlying tree: simply ignore the data, and always output a fixed (favorite) tree. This ‘‘method’’ performs splendidly if the favored tree is the true tree, but otherwise it performs very badly. So why is the construction we discuss here any less trivial? The crucial difference is that MP has a higher reconstruction probability than $M_{av}L$ not just on one four-species tree, but on any of the underlying trees (provided the other associated parameters are chosen appropriately)—and this is something a trivial method like the one described clearly cannot achieve.

Again, this should not be overinterpreted—it does not mean that we should be using MP—it may well be that on average (under some prior on trees and their parameters) $M_{av}L$ outperforms MP, but it does not globally outperform (in the sense described above) MP.

In more detail, consider a fully resolved tree *T* on four species—say, *a*, *b*, *c*, and *d*—with the topology $ab|cd$ and the simple symmetric two-state model with mutation probability $p(e) = \epsilon$ on the two edges incident with leaves *a*, *b*, while $p(e) > 0.5 - \epsilon$ on the other three edges, where ϵ is small but positive. Thus, three edges involve long interspeciation times (and/or high mutation rates) and so are near site saturation, while two sister taxa are recently separated (and/or have low mutation rates on their incident edges). Note that such a situation is entirely possible under a molecular clock (see fig. 1*a*), although we need not insist on this.

Suppose we evolve *k* sites independently on this tree. Let $P_1(k)$ be the probability that MP recovers the true tree T, and let $P_2(k)$ be the probability that $M_{av}L$ recovers T from the k sites.

THEOREM 5. *As* ϵ *converges to* 0 (*with the number of sites k fixed*),

$$P_1(k) \cong 1 - \left(\frac{3}{4}\right)^k; \qquad P_2(k) \leq \frac{2}{3}.$$

In particular, the probability that MP correctly reconstructs T can be higher than the corresponding probability for $M_{av}L$ for any fixed sequence length $k \geq 4$.

A similar result was stated without proof in Székely and Steel (1999); we outline a proof here in section (d) of the appendix. Note that for ϵ very small (but positive), MP will recover T with probability 0.99 with just 16 sites, yet $M_{av}L$ could potentially take 10^{10} sites to achieve the same probability of correctly reconstructing T (in which case, for realistic length sequences, other effects, e.g., deviations from the model, might have more effect on the reconstructed tree than the sequence data). This is of course an extreme situation; nevertheless, it shows that there are situations in which we would expect $M_{av}L$ to require much longer sequences than MP needs to recover the true tree.

Note that we actually only require $p(e) > 0.5 - \epsilon$ on two of the three edges, but we have opted to allow three edges to be near site saturation, since then the example can arise under a molecular clock. In contrast, the Felsenstein Zone cannot arise under a molecular clock, yet, to be fair, if we want to impose a molecular clock, we should implement ML with a molecular clock, and if we did, ML would no longer behave as described above.

Also, this example does not demonstrate any inconsistency of $M_{av}L$, since if the edge mutation probabilities are fixed (and strictly between 0 and 0.5), then $M_{av}L$ will eventually recover the true tree with probability converging to certainty as k tends to infinity.

This example can also be modified to demonstrate that $M_{av}L$ can differ from maximum integrated likelihood, even when all trees have equal prior probabilities (provided the prior distribution on the edge lengths is sufficiently contrived). Specifically, suppose that each of the three binary trees on sequences a, b, c, and d has equal probability and that the prior distribution on the edge lengths allows all possible values for the mutation probabilities, but with probability $1 - \delta$, we have $p(e) \leq \epsilon$ on two edges incident with two sister leaves and $p(e) > 0.5 - \epsilon$ on the other three edges. Then it can be shown that for ϵ, δ sufficiently small (but positive), MIL can select a different tree than $M_{av}L$ on certain data.

The Limits to Models: Recent Developments and Future Directions

As models become increasingly sophisticated and parameter-rich, one risks losing the ability to discriminate between different underlying trees (Yang, Goldman, and Friday 1995). Essentially, this is because when the parameters are twiddled appropriately, the data may be able to be described perfectly by any underlying tree. This is a real possibility for site substitution models that allow a distribution of rates across sites, as demonstrated in Steel, Székely, and Hendy (1994). This paper showed that there are situations in which all trees could perfectly describe the same data, provided one can select for each tree a corresponding distribution of rates across sites. The model we described earlier (no common mechanism) in which MP can be regarded as a M_{av} method clearly would also have this nonidentifiability problem.

Even if one knows the distribution of rates across sites, nonstationarity can also lead to a similar nonidentifiability phenomenon, at least for pairwise comparisons, as Baake (1998) has shown. Baake's example was particularly simple—exactly half the sites are invariable, while the other half evolve according to the same Markov process. It is an open question whether this nonidentifiability of the tree is also true if one simultaneously uses all the sequence information. There are other related problems where reducing data to pairwise information destroys information about the underlying structural parameters. For example, Chang (1996*a*) showed that for a nonstationary model (and without rates across sites), triplewise comparisons of sequences generally suffice to determine all of the edge parameters (i.e., relative rates of substitution between the different nucleotides), but pairwise comparisons generally do not. Independently, Lake (1997) also described a triplewise technique for reconstructing these edge parameters.

The question of phylogeny reconstruction can also be viewed from an information-theoretic perspective. One such approach (based on the concept of Fisher information) has been presented by Goldman (1998) and developed as a tool for experimental design. In phylogeny reconstruction, it is helpful to regard each site as containing some information concerning the underlying tree and note that this signal depends on the other underlying structural nuisance parameters, such as the edge lengths. For example, very many sites will be required in order to reliably recover a very short internal edge, while a very long external edge (i.e., leading to a distant outgroup sequence) will need very many sites in order to be correctly placed in the tree. A fundamental question in terms of these edge lengths and the number of sequences is, how many sites are required to accurately reconstruct the underlying tree? Recently, this question has been shown to have a rather surprising answer. Namely, for simple models, if the underlying structural parameters are sufficiently constrained, then the sequence length required to reconstruct the true tree can grow even slower than the number of sequences (even though the number of possible trees grows exponentially with the number of sequences) (for details, see Erdös et al. 1999). These theoretical results are relevant to recent simulation studies (and the surrounding controversy) suggesting that trees on large numbers of sequences can sometimes be reconstructed from surprisingly short sequences (Hillis 1996; Purvis and Quicke 1997; Yang and Goldman 1997; Graybeal 1998). The theoretical results suggest that one should be able to reconstruct large trees from short sequences, at least for some choices of the underlying parameters.

One of the useful tools in the theoretical analyses has been explicit sufficient conditions for distance-based methods to correctly reconstruct the underlying tree. For example, suppose one's estimate of the evolutionary distance between each pair of sequences comes within an error bound of x of the true evolutionary distance. Provided each edge in the underlying (true) evolutionary tree has an evolutionary distance of at least $2x$, several standard distance-based tree reconstruction methods will correctly reconstruct the underling tree. This was established recently for the neighbor- joining method by Atteson (1997).

Of course, one may not know in advance that the internal edges are sufficiently long to be recovered. An alternative approach is to reconstruct an edge-weighted tree and regard two such trees as ''close'' if the difference in their edge lengths (both shared and missing edges) is small (a metric on edge-weighted trees along these lines was suggested by Robinson and Foulds [1979]). From this perspective, a phylogeny reconstruction corresponds to a point in a continuous (rather than discrete) tree space. Probabilistic bounds on the distance in continuous tree space between the reconstructed tree and the underlying tree can then be stated in terms of the sequence length and the diameter of the underlying tree. This approach has been developed by Farach and Kannan (1999). One limitation in such an approach is that the short edges are often the ones of greatest phylogenetic interest, since they are generally the spots exhibiting the uncertainty concerning the exact order of speciation.

Other recent approaches for displaying conflicting or uncertain phylogenetic information have included the construction of networks (rather than trees), such as median networks (Bandelt et al. 1995), the split decomposition methodology (Dress, Huson, and Moulton 1996), stochastic networks (Strimmer and Moulton 2000; von Haeseler and Churchill 1993), and networks that allow for genetic events such as recombination and horizontal gene transfer (Hein 1993; Fitch 1997).

Another approach to this problem has been to construct confidence sets of phylogenetic trees (analogous to confidence intervals). One can then apply consensus tree methods to obtain a semiresolved tree that represents a conservative single-tree summary of this confidence set (alternatively, one might apply maximum agreement subtree techniques to obtain a more resolved tree on a subset of the species). A different strategy for constructing a semiresolved tree that is based on statistics (rather than combinatorics) is the minimum model-based complexity approach (Tanaka et al. 1999), which introduces additional edges into a tree only if this leads to a simpler statistical description of the data.

In summary, a variety of techniques are likely to be of use, particularly in analyzing new types of data. Such techniques will include various forms of ML and MP along with other methods. Clearly, one must be careful in making special claims about the 'special status' of either ML or MP. The latter method may be considered a type of ML and can be appropriate for certain types of data. Conversely, the usual (average) form of ML can in certain settings be justified by the parsimonious arguments usually reserved for MP. Even under simple models, neither model always outperforms the other in terms of the probability of reconstructing the correct tree. Although MP may fail to be statistically consistent, more is now known about when this will occur, although there still remain several unanswered questions.

Acknowledgments

We thank the New Zealand Marsden Fund for supporting this research, and the Isaac Newton Institute (Cambridge, England) for its hospitality during the 1998 BFG program. We also thank Peter Lockhart and Charles Semple for helpful comments on an earlier version of this manuscript, and a referee for numerous helpful suggestions.

LITERATURE CITED

ARCHIE, J. W., and J. FELSENSTEIN. 1993. The number of evolutionary steps on random and minimum length trees for random evolutionary data. Theor. Popul. Biol. **43**:52–79.

ATTESON, K. 1997. The performance of the neighbor-joining method of phylogeny reconstruction. Pp. 133–147 *in* B. MIRKIN, F. R. MCMORRIS, F. S. ROBERTS, and A. RZHETSKY, eds. Mathematical hierarchies and biology, DIMACS series in discrete mathematics and theoretical computer science. Vol. 37. American Mathematical Society, Providence, RI.

BAAKE, E. 1998. What can and cannot be inferred from pairwise sequence comparisons? Math. Biosci. **154**:1–21.

BANDELT, H. J., P. FOSTER, B. C. SYKES, and M. B. RICHARDS. 1995. Mitochondrial portraits of human populations using median networks. Genetics **141**:743–753.

BARRY, D., and J. A. HARTIGAN. 1987. Statistical analysis of hominoid molecular evolution. Stat. Sci. **2**:191–210.

BLANCHETTE, M., T. KUNISAWA, and D. SANKOFF. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. **49**:193–203.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. **40**:271–280.

CAVENDER, J. A. 1981. Tests of phylogenetic hypotheses under generalized models. Math. Biosci. **54**:217–229.

CHANG, J. 1996*a*. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. **137**:51–73.

———. 1996*b*. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Mol. Biosci. **134**:189–215.

CHARLESTON, M. A., M. D. HENDY, and D. PENNY. 1994. The effects of sequence length, tree topology and number of taxa on the performance of phylogenetic methods. J. Comp. Biol. **1**:133–151.

DRESS, A., D. HUSON, and V. MOULTON. 1996. Analysing and visualizing sequence and distance data using SPLITSTREE. Discr. Appl. Math. **71**:95–109.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, England.

———. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. Syst. Biol. **45**:79–91.

EDWARDS, A. W. F., and L. L. CAVALLI-SFORZA. 1963. The reconstruction of evolution. Heredity **18**:533; Ann. Hum. Genet. **27**:104–105.

Erdös, P., M. A. Steel, L. a Székely, and T. Warnow. 1999. A few logs suffice to build (almost) all trees (part 1) Random Struct. Algorithms **14**:153–184.

Farach, M., and S. Kannan. 1999. Efficient algorithms for inverting evolution. J. Assoc. Comput. Mach. **46**:437–449.

Farris, J. S. 1973. A probability model for inferring evolutionary trees. Syst. Zool. **22**:250–256.

Farris, J. S., A. G. Kluge, and M. J. Eckardt. 1970. A numerical approach to phylogenetic systematics. Syst. Zool. **19**:172–189.

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

———. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

Felsenstein, J., and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

Felsenstein, J., and E. Sober. 1986. Parsimony and likelihood: an exchange. Syst. Zool. **35**:617–626.

Fitch, W. M. 1971*a*. Rate of change of concomitantly variable codons. J. Mol. Evol. **1**:84–96.

———. 1971*b*. Towards defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**: 406–416.

———. 1997. Networks and viral evolution. J. Mol. Evol. **44**(Suppl.):S65–S75.

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

———. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B Biol. Sci. **265**:1779–1786.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. **47**:9–17.

Halpern, A. L., and W. B. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **20**:402–411.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

Hillis, D. M. 1996. Inferring complex phylogenies. Nature **383**:130–131.

Huelsenbeck, J. P. 1998. Systematic bias in phylogenetic analysis: is the strepsiptera problem solved? Syst. Biol. **47**: 519–537.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. **45**:363–374.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. J. Mol. Evol. **29**:170–179.

Koshi, J. M., and R. A. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. **42**: 313–320.

Lake, J. A. 1997. Phylogenetic inference: how much evolutionary history is knowable? Mol. Biol. Evol. **14**:213–219.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA **93**:1930–1934.

Lockhart, P. J., M. A. Steel, D. Penny, and M. D. Hendy. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

Maddison, W. P., and M. Slatkin. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution **45**:1184–1197.

Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

Miyamoto, M. M., B. F. Koop, J. F. Slightom, M. Goodman, and M. R. Tennant. 1988. Molecular systematics of higher primates: genealogical relations and classification. Proc. Natl. Acad. Sci. USA **85**:7627–7631.

Nikaido, M., A. P. Rooney, and N. Okada. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotomuses are the closest extant relatives of whales. Proc. Natl. Acad. Sci. USA **96**:10261–10266.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature **401**:877–884.

Penny, D., and M. Hasegawa. 1997. Platypus put in its place. Nature **387**:549–550.

Penny, D., M. D. Hendy, P. J. Lockhart, and M. A. Steel. 1996. Corrected parsimony, minimum evolution and Hadamard conjugations. Syst. Biol. **45**:593–603.

Penny, D., M. D. Hendy, and M. A. Steel. 1991. Testing the theory of descent. Pp. 155–183 *in* M. Miyamoto and J. Cracraft, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, Oxford, England.

Penny, D., M. A. Steel, P. J. Lockhart, and M. D. Hendy. 1994. The role of models in reconstructing evolutionary trees. Pp. 211–230 *in* R. W. Scotland, D. J. Siebert, and D. M. Williams, eds. Models in phylogeny reconstruction. Oxford University Press, Oxford, England.

Purvis, A., and D. L. J. Quicke. 1997. Building phylogenies: are big trees easy? Trends Ecol. Evol. **12**:49–50.

Rice, K., and T. Warnow. 1997. Parsimony is hard to beat. Pp. 124–133 *in* T. Jiang and D. T. Lee, eds. Lecture notes in computer science, Vol. 1276. Springer, Berlin.

Robinson, D., and L. R. Foulds. 1979. Comparison of weighted labeled trees. Pp. 119–126 *in* A. Dold and B. Eckmann, eds. Lecture notes in mathematics. Vol. 748. Springer-Verlag, Berlin.

Rogers, J. S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. **46**:354–357.

Schöniger, M., and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated sequences. Mol. Phylogenet. Evol. **3**:240–247.

Siddall, M. E. 1998. Success of parsimony in the four-taxon case: Long branch repulsion by likelihood in the Farris Zone. Cladistics **14**:209–220.

Sober, E. 1985. A likelihood justification of parsimony. Cladistics **1**:209–233.

———. 1988. Reconstructing the past: parsimony, evolution and inference, MIT Press, Cambridge, Mass.

Steel, M. 1999. Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. Research Report NI 98025-BFG. Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

Steel, M., M. D. Hendy, and D. Penny. 1992. Significance of the length of the shortest tree. J. Classif. **9**:71–90.

———. 1998. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. Discr. Appl. Math. **88**:367–396.

Steel, M. A., D. Penny, and M. D. Hendy. 1993. Parsimony can be consistent! Syst. Biol. **42**:581–587.

Steel, M. A., L. A. Székely, and M. D. Hendy. 1994. Reconstructing trees from sequences whose sites evolve at variable rates. J. Comp. Biol. **1**:153–163.

Strimmer, K., and V. Moulton. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. **17**:875–881.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Marble, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

Székely, L. A., and M. Steel. 1999. Inverting random functions. Ann. Combin. **3**:103–113.

Tanaka, H., F. Ren, T. Okayama, and T. Gojobori. 1999. Topology selection in unrooted molecular phylogenetic tree by minimum model-based complexity. Pacif. Symp. Biocomput. **4**:326–337.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

Thorne, J. L., H. Kishino, and J. Felsenstein. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. **34**:3–16.

Tuffley, C., and M. Steel. 1997*a*. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. **59**:581–607.

———. 1997*b*. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. **147**:63–91.

Van de Peer, Y., S. A. Rensing, U.-G. Maier, and R. Dewachter. 1996. Substitution rate calibration of small subunit ribosomal subunit RNA identifies Chlorarachnida nucleomorphs as remnants of green algae. Proc. Natl. Acad. Sci. USA **93**:7732–7736.

von Haeseler, A., and G. A. Churchill. 1993. Network models for sequence evolution. J. Mol. Evol. **37**:77–85.

Waddell, P. J. 1996. Statistical methods of phylogenetic analysis. Ph.D. thesis, Massey University, Palmerston North, New Zealand.

Yang, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. **43**:329–342.

———. 1996*a*. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

———. 1996*b*. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. **42**:587–596.

Yang, Z., and N. Goldman. 1997. Are big trees indeed easy? Trends Ecol. Evol. **12**:357.

Yang, Z., N. Goldman, and A. E. Friday. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.

Appendix

(a) Proof of Theorem 3

Suppose the state space S has size r. For a character f taking values in S, let $c(f)$ denote the number of states that are actually mentioned by f. Thus, $1 \leq c(f) \leq r$. Now, for any tree T, the parsimony score of f on T, which we will denote as $l(f, T)$, is at least $c(f) - 1$. Let $i(f, T)$ denote the difference between $l(f, T)$ and $c(f) - 1$. That is,

$$i(f, T) = l(f, T) - c(f) + 1.$$

Note that $i(f, T) = 0$ precisely if f could have evolved on T without any parallel or convergent mutations, so $i(f, T)$ is a measure of homoplasy of the character f with respect to T. For a collection $F = f_1, \ldots, f_k$ of characters, let

$$I(F, T) = \sum_{j=1}^{k} i(f_j, T),$$

and for an internal edge e of T, let $n(f, e) = 1$ provided $c(f) = 2$ and f can be extended to an assignment of states to all the vertices of T by a single mutation on edge e; otherwise, set $n(f, e) = 0$. Let $n_{\min}(F, T)$ be the minimal value of $\Sigma_j\, n(f_j, e)$ across all internal edges of T. Thus, $n_{\min}(F, T)$ is the smallest number of two-state characters from F that support some internal edge. We first establish the following:

Lemma. *If $n_{\min}(F, T) > I(F, T)$, then MP reconstructs T from F.*

Proof. Let $L(F, T_1)$ denote the parsimony score of F relative to a tree T_1. Then,

$$L(F, T_1) = \sum_{j=1}^{k} i(f_j, T_1) + \sum_{j=1}^{k} (c(f_j) - 1),$$

and if T_1 differs from T, there is some edge e in T that induces a bipartition of the taxa not found in T_1. For this edge e, let $K = \Sigma_j\, n(f_j, e)$. Then, $K \geq n_{\min}(F, T) > I(F, T)$ (by our assumption in the lemma). Now, for the K characters with $n(f_j, e) = 1$, we have $i(f_j, T_1) \geq 1$, and so

$$\begin{aligned} L(F, T_1) &\geq K + \sum_{j=1}^{k} (c(f_j) - 1) \\ &> I(F, T) + \sum_{j=1}^{k} (c(f_j) - 1) = L(F, T), \end{aligned}$$

which establishes the lemma.

By this lemma, MP is statistically consistent if for each internal edge e, the expected value of $n(f, e)$ is larger than the expected value of $i(f, T)$ for a character f generated according to the model described. Let us (totally) order the vertices of T as $v_0, v_1, \ldots, v_t$ in any way that respects ancestry in the tree—that is, if v_j is a descendant of vertex v_i, then $i < j$. Note that T has $t - 2$ edges and that $t = 2n - 2$, where n is the number of leaves of T (we are assuming T is binary). Suppose a character evolves on the tree from vertex v_0 according to the model specified. The probability that there is one mutation on any particular edge and no mutations on the remaining $t - 3$ edges of T is at least $a(1 - b)^{t-3}$ (by the generalized product rule, and the restriction placed on the model). Thus, since $n(f, e)$ is a 0/1 random variable,

$$E[n(f, e)] = P[n(f, e) = 1] > a(1 - b)^{t-3}.$$

For $j \geq 1$, let U_j be the 0/1 random variable that takes the value 1 precisely when the state at vertex v_j is different from the state at its immediate ancestor vertex but equals the state at some vertex v_i with $i < j$. Thus, $U_j = 1$ precisely if the mutation on the edge leading to this vertex describes a return to an 'earlier' state. If f is the resulting character at the leaves of the tree, we have (by induction)

$$i(f, T) \leq \sum_{j=1}^{t} U_j.$$

Consequently,

$$E[i(f, T)] \leq \sum_{j=1}^{t} E[U_j] = \sum_{j=1}^{t} P[U_j = 1].$$

Now, by the assumptions of the model,

$$P[U_j = 1] \leq c(j - 1)/(r - 1),$$

and so

$$E[i(f, T)] \leq \sum_{j=1}^{t} c(j - 1)/(r - 1)$$
$$= ct(t - 1)/2(r - 1).$$

By selecting r sufficiently large, we have $E[n(f, e)] > E[i(f, T)]$, as required.

(b) Proof of Theorem 3a

Let p_0 denote the probability of generating an unvaried site pattern (type *xxxx*, for some nucleotide x). Let us say a pattern is of type 1 if it has lower parsimony score on the true tree than on the other trees (in which case the difference in parsimony score between the trees is exactly one mutation). Thus, the type 1 patterns are precisely those of the form *xxyy* (for different nucleotides x, y). The two site patterns *xyxy* and *xyyx* will be called type 2 and type 3, respectively. Note that these are the two types of patterns that favor the other two binary trees with respect to parsimony score.

For $i = 1, 2, 3$, let p_i denote the probability of generating a type i site pattern. Thus, MP is statistically consistent on single-site data precisely if $p_1 > \max\{p_2, p_3\}$. Now, the only k-tuples of sites that distinguish between the possible trees as far as MP is concerned are k-tuples that consist of at least one type i site and for which any remaining sites that are not of type i are unvaried. Once again, in this case, the difference in parsimony scores between the true tree and an alternative tree is 1. The probability of generating such a k-tuple of sites is $(p_i + p_0)^k - p_0^k$. Thus, MP is statistically consistent if and only if

$$(p_1 + p_0)^k - p_0^k$$
$$> \max\{(p_2 + p_0)^k - p_0^k, (p_3 + p_0)^k - p_0^k\}.$$

However, this inequality applies if and only if $p_1 > \max\{p_2, p_3\}$, which was precisely the condition described above for the statistical consistency of MP.

(c) Proof of Theorem 4

Suppose the underlying substitution process can be described by a reversible, stationary Markov process coupled with a distribution of rates across sites (this is broad enough to encompass most of the models in current use). In this case, the expected Hamming distance d_{ij} between two sequences can be written as $d_{ij} = \mu(K_{ij})$, where K_{ij} is the expected evolutionary distance (number of mutations that occurred on the path in the tree connecting the sequences) and μ is a monotone increasing function (Tuffley and Steel 1997*a*, eq. 4). Now, if a molecular clock applies, then $K = [K_{ij}]$ satisfies the ultrametric criterion (i.e., $K_{ij} \leq \max\{K_{ij}, K_{il}, K_{jl}\}$ for all i, j, k), and so does any montone increasing function of K, in particular, $d = [d_{ij}]$ satisfies the ultrametric criterion and thus is additive on the true tree.

(d) Proof of Theorem 5

For k fixed, $P_1(k)$ and $P_2(k)$ are both continuous functions of ϵ. Consequently, their limiting values as ϵ converges to 0 are identical to their values if we put $\epsilon = 0$, and we will assume $\epsilon = 0$ in the calculations that follow. Then, the symmetric two-state model produces just four (of the eight possible) types of patterns, each with equal probability. If we order the taxa as *abcd* (and recall that the tree groups a and b vs. c and d), these four patterns are *xxxx*, *xxxy*, *xxyx*, and *xxyy*. Note that only the last one contributes unequally to the parsimony score of the three trees—it favors the true tree and penalizes the other two. Consequently, $P_1(k)$ is the probability that this pattern occurs at least once, which, by the independence assumption between characters and elementary probability, is simply $1 - (3/4)^k$.

Regarding $P_2(k)$, note that when $\epsilon = 0$, the *xxxx*, *xxxy*, *xxyx*, and *xxyy* patterns occur at any site with equal probability, namely, 1/4. On the true tree ($ab|cd$), and for such randomly generated data D, let $p_a(D)$, $p_b(D)$, $p_c(D)$, $p_d(D)$, and p_5 denote an assignment of mutation probabilities to the edges incident with taxa a, b, c, and d and the central edge, respectively, so as to maximize the probability of generating D. For any data D consisting of just the four patterns described, $p_a(D) = p_b(D) = 0$. Let E be the event that D, generated as described, is such that $\max\{p_c(D), p_d(D)\} = 0.5$ (note that these probabilities are constrained to lie in the interval [0, 0.5]), and let p denote the probability of event E. Then it can be shown that

$$p \geq 0.5.$$

Now, when E occurs, we could place appropriate mutation probabilities on the edges of either of the two alternative trees to obtain the same likelihood score, and so our probability of recovering the true tree would be 1/3 (assuming ties are broken randomly). So, when $\epsilon = 0$,

$$P_2(k) \leq (1 - p) \times 1 + p \times \frac{1}{3} \leq \frac{2}{3},$$

since $p \geq 0.5$, as required.

SIMON EASTEAL, reviewing editor

Accepted March 2, 2000