Some statistical aspects of the maximum parsimony method

Mike Steel

Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

Summary. The last three decades have seen considerable debate concerning the relative merits and problems associated with two competing approaches to phylogeny—approaches based on the parsimony principle *versus* maximum likelihood methodology. Although the two approaches may seem quite opposed, there are in fact some close relationships between them. For example, we describe a recent result that shows how maximum parsimony can be regarded as a type of maximum likelihood estimator when there is no common mechanism between sites (such as might occur with morphological data and certain forms of molecular data). Distinguishing between this and other implementations of maximum likelihood helps clarify some of the dispute that has surrounded the two methodologies. We also provide a brief overview of some mathematical and statistical properties of the maximum parsimony criterion.

Introduction

Many techniques now exist for reconstructing phylogenetic trees from genetic sequence data. Two of the most popular approaches are usually referred to as 'maximum parsimony and 'maximum likelihood'. We will abbreviate these approaches here as MP and ML respectively. Although MP continues to be widely used, it is often criticised as being statistically unsound and as failing to make explicit an underlying 'model' of evolution. Indeed 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 in his joint 1963 paper with Cavalli-Sforza (in the context of continuous characters) merely as a computational approximation for ML, and by not as a method of choice in its own right. The discussion is further complicated by claims that MP variously is, or is not, a form of ML, and by the discussion of 'zones' within which either method performs worse than the other in recovering the true tree.

Several authors (for example, Farris, Kluge and Eckardt, 1970; Sober, 1988) claim MP is the preferred method of tree reconstruction, citing Willi Hennig's writings on phylogenetic inference, or alternatively the *Principle of Parsimony*. The latter is a minimalist principle, also referred to as 'Ockham's razor'. It states that one should prefer simpler explanations, requiring fewer assumptions, over more complex, *ad hoc* ones. In phylogeny reconstruction,

about any underlying model or mechanism for evolution (however, this can also be used as an argument in favour of the more usual forms of ML), or (ii) to emphasise the feature that MP favours the tree requiring the fewest evolutionary events (such as mutations) to explain the observed data, and so is, in some sense, the 'simplest' or an 'optimal' description of the data.

Several authors (e.g., Farris, 1973; Sober, 1985, 1988) have also presented explicit statistical arguments in favour 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; Kishino and Hasegawa, 1989; Maddison and Slatkin, 1991; Archie and Felsenstein, 1993; Steel, Hendy and Penny, 1992; Steel, Lockhart and Penny, 1993b, 1995).

The simplicity of a method like MP and variations that allow weightings on characters and transition types, together with its apparent lack of assumption involving underlying models, has made it popular in phylogeny, particularly in the 1970s and 80 s. However, 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 (1978). This paper demonstrated that if sequence sites evolved under certain models and combinations of rates, then MP would favour 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 Section 4). The particular combination of short and long branches that Felsenstein used has become known as the 'Felsenstein Zone'.

Both Felsenstein (1973) and Yang (1994) informally claimed the nonexistence of any such zone within which ML would be statistically inconsistent (though this was questioned by Sober (1988, Ch. 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 recently by Chang (1996b). Note that the use of the 'correct' model (the same as the model used to generate the data) is essential to the proof that maximum likelihood is consistent, and ML can be inconsistent if the model used to analyse the data differs from that which generated it (see Chang, 1996a). Although one may seldom know the correct model of evolution, the more one knows about the evolutionary process, the more likely one is to avoid a zone of inconsistency by analysing the data correctly.

Nevertheless, objections to ML have arisen on a number of fronts, which we now describe. First, there is concern about the validity and exact form of any underlying stochastic model (for example, there is concern as to the choice of underlying parameters/distributions), and that by selecting the appropriate model one could perhaps reconstruct any favoured tree. There is also concern

that ML estimation of a tree (and statistical tests between different trees) that involves optimizing 'nuisance (supplementary) parameters' is statistically problematic. There are also suggestions that the Felsenstein zone rarely if ever arises for real data and claims for the existence of a 'Farris zone' where MP outperforms ML. Another factor is the increasing analysis of aspects of genome data that extend beyond site substitution—for example, gene order, SINEs (short interspersed nuclear elements) for which MP may be more appropriate. Finally there is some concern about the computational complexity of ML. Even on a *given* tree, optimising the likelihood can be problematic (unlike MP, where Fitch's algorithm (Fitch, 1971a) provides a linear time algorithm for computing the parsimony score).

In this chapter we will explore some 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. We also describe some statistical properties of the parsimony score function.

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, exhibits rate heterogeneity, etc.) and the initial conditions (for example, inter-speciation times or rates on each edge (branch) of the tree).

Often researchers will seek 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'.

Varieties of forms of ML in phylogenetics

According to Edwards (1972), 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. A ML method of inference selects the hypothesis H that maximises the likelihood function for the data D (given the specified mechanism). In the context of phylogeny reconstruction from sequences, the data 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; however, some authors have recently incorporated this also (for example,

Thorne, Goldman and Jones, 1996; Giribet and Wheeler, 1999b). 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).

What complicates matters is that $P(D|T)$, and hence the likelihood of T, requires more information to specify it than just the data D and the parameter T. More precisely, the probability of evolving D depends on further parameters, sometimes referred to as 'nuisance parameters'. In order to talk about $P(D|T)$ we either need to 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 (inter-speciation times and rates of mutation on the edges), parameters associated with the substitution matrix (for example, transition/transversion bias) and parameters that describe how rates vary across sites.

Nuisance parameters arise widely in many statistical settings and have been discussed in the phylogeny setting by several authors, for example Goldman (1990). Nuisance parameters may further be classified into 'structural' and 'incidental' parameters. The former are parameters that influence all (or nearly all) of the sites; incidental parameters influence only one or a few. Structural parameters typically correspond to the edge (branch) lengths and parameters that constrain the substitution process (for example, the transition/transversion bias). Typically, such parameters are either selected to maximize the likelihood or estimated directly from the data. Incidental nuisance parameters arise either if (i) we wish to hypothesise 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 if (ii) 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 the Greek letter θ.

Two frequent assumptions concerning substitution models are that aligned sites evolve *independently* and according to an *identical* process—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 one has 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 modelled 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, 1997a), even though the original covarion model (e.g., Fitch, 1971b) 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 analysed. These include models that allow a limited degree of non-independence between sites (for example pairwise interactions in stem regions, Schöniger and von Haeseler, 1994), and models that work with non-aligned 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 (MIL* versus *MRL)*

If the nuisance parameters θ and the phylogeny T are generated according to some known prior distribution (for example, 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 a tree T that maximizes $P(D|T)$ we will refer to as a *maximum (integrated) likelihood tree*. Maximum integrated likelihood (MIL), and, more generally, the assignment of posterior probabilities to trees based on sequence data (using Markov chain Monte Carlo technique 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. A natural question arises, namely, in what sense is MIL 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 the 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 propo-

sition describes precisely the method that maximizes the expected reconstruction probability:

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 $p(T)P(D|T)$.

For a proof of this last assertion, see Székely and Steel (1999). The tree(s) that maximizes $p(T)P(D|T)$ is sometimes referred to as the *maximum a posteriori* (MAP) estimate. This is precisely the maximum (integrated) likelihood 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 probability to all binary trees, MIL maximises one's average chance of recovering the correct tree. However, if the distribution on binary trees is not uniform—for example, if the trees are described by a Yule process —then the optimal selection criteria are slightly different. In any case, it is clearly a difficult problem to find (let alone agree upon!) a compelling and 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 data D depend 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 than $\max_\theta P(D|x_2,\theta)$, 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 different selections from MIL and it seems to have been a fundamental issue in the exchange between Felsenstein and Sober (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 later.

For the remainder of this chapter 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.

Three forms of maximum relative 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 maximum (relative) likelihood in molecular phylogenetics, one effectively averages over all possible assign-

ments 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 suggested explicitly by Barry and Hartigan (1987) who called it *most parsimonious likelihood*, to distinguish it from $M_{av}L$. They remarked that most parsimonious likelihood 'is therefore similar to the maximum parsimony fitting technique'. However, it differs slightly 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 (Koshi and Goldstein, 1996; Pagel, 1999) where the interest has been primarily in reconstructing, say, ancestral sequences of proteins (or other characters), rather than in selecting an optimal tree. Goldman (1990) described a link between MP and most parsimonious likelihood. He showed that, under a symmetric 2-state mutation model, and with the artificial constraint that all mutation probabilities on each edge of any binary tree are equal to some value p, then the MP tree(s) are exactly the most parsimonious likelihood trees.

Given the most parsimonious likelihood approach, it might seem natural to carry the approach of assigning ancestral sequences further. That is, one could select sequences for each time interval right through the tree (jointly with the other parameters) to maximise 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.

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 showed that the tree(s) that maximizes the likelihood in this sense are *exactly* the maximum parsimony trees. 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 optimised. 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's evolutionary pathway likelihood is crucial for reconciling the apparent paradox between Felsenstein's claim that ML (but not MP) is statistically consistent and with Farris's claim that MP is a ML method. Both claims are correct; they are simply referring to different forms of ML. Figure 1 illustrates the three forms of ML we have just discussed.

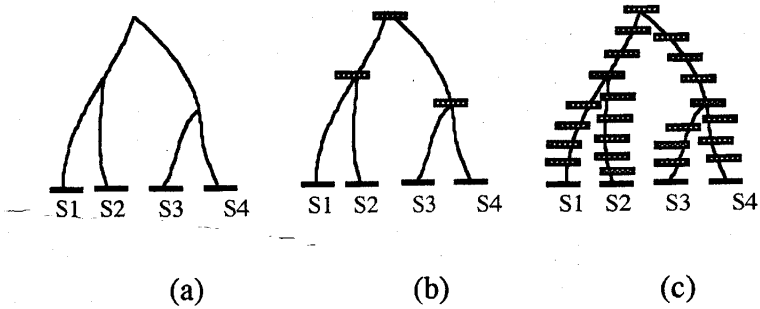

Figure 1. Three forms of ML. (a) *Maximum average likelihood* ($M_{av}L$): all possible sequences at the internal vertices contribute to the likelihood; (b) *Most parsimonious likelihood*: sequences to maximize the likelihood are placed at the internal vertices; (c) *Evolutionary pathway likelihood*: sequences to maximize the likelihood are placed at each position throughout the tree.

A model for which maximum parsimony is a maximum (average) likelihood estimator

Most parsimonious likelihood and evolutionary pathway likelihood both involve 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—the Jukes-Cantor type model—in which each of the possible substitutions at a site occurs with equal probability. 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 might 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; however, the ratios of the edge lengths are equal across the sites. In the Jukes-Cantor style model with no common mechanism (not even the same rates for different characters) the following theorem applies.

Under the model described (with no common mechanism) the $M_{av}L$ tree(s) are precisely the maximum parsimony tree(s).

A proof of this result is given by Tuffley and Steel (1997b) who generalised an earlier special case by Penny et al. (1994). The significance of the result should not be taken as any special justification of MP over usual implementations of ML; neither 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 is of a more philosophical nature, as it describes a model in which MP can be regarded as a ML method in the usual 'average' ML setting (that is, where one does not select particular sequences for the internal vertices as part of the optimisation step).

The argument used to establish the above theorem also shows that, under the Jukes-Cantor type 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 Jukes-Cantor style 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 (1997b).

Of course this type of underlying model (in the above theorem) 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 (for example, 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 towards the more usual implementations of maximum (average) likelihood 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, each sequence being many thousands of nucleotides long. 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 several thousand different mechanisms, one for each site. In such a case, the Principle of Parsimony would support the usual maximum (average) likelihood over using data uncorrected for multiple changes.

This conclusion should, however, be taken with care. Such a model may not apply to other sequence data and would not often apply to morphological data (for example, where the evolution of numbers of legs may differ from that of wing colour). 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.

Regions where MP may outperform ML

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. What is perhaps more surprising is that MP can perform better than $M_{av}L$, even when the underlying model matches the generating model. These regions of parameter space have been called the 'Farris Zone' (Siddall, 1998) and the 'anti-Felsenstein Zone' (Waddell, 1996); this phenomenon has been noted by others (for example Huelsenbeck, 1998; Yang, 1996).

Here the 'performance' of a tree reconstruction method M (on sequence data generated under a tree-indexed Markov model) is again taken to mean the reconstruction probability $\rho(M,T,\theta)$ described in Section 2 (the probability that the method will correctly return the true tree T). This depends not just on M but also on T and the parameters on the edges of the tree. Now there exist trees T and parameters where MP will have a higher probability of returning the 'true tree' T than $M_{av}L$. In more detail, consider a fully resolved tree T on four species a, b, c, d, with species a, b on one side of the central edge, and species c and d on the other. Consider the simple symmetric 2-state model with mutation probability $p(e) = \varepsilon$ on the two edges incident with leaves a,b; while $p(e) > 0.5 - \varepsilon$ on the other three edges, where ε is small but positive. Thus three edges have long interspeciation times (or, alternatively, high mutation rates) and so are near site saturation, while two sister taxa are recently separated (or, alternatively, have low mutation rates on their incident edges). Note that such a situation is entirely possible under a molecular clock, though 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. Then, as ε converges to 0 (with k fixed) we have:

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

A proof of this result is presented in Steel and Penny (2000) (a similar result was stated without proof in Székely and Steel (1999)). Notice that for ε very small (but positive), MP will recover T with 99% probability with just 16 sites, yet $M_{av}L$ could take potentially millions of sites to achieve the same probability of correctly reconstructing T. In that case, for realistic length sequences, other effects, for example deviations from the model, might have more effect on the reconstructed tree than the sequence data.

It is tempting to dismiss this example as a triviality by noting that one could also outperform $M_{av}L$ in this example by simply disregarding the data and always outputting the tree that groups species a and b together, and c and d together. However, there is a fundamental difference here, since MP will outperform $M_{av}L$ for *any* of the three possible underlying trees on four species,

when the parameters are in the right range. Clearly a trivial method, like the one described, cannot achieve this.

While the example described above is somewhat extreme, it still shows there are cases where we would expect $M_{av}L$ to require much longer sequences to recover the true tree than MP needs. In fact we actually only require $p(e) > 0.5 - \varepsilon$ 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 with four species; yet to be fair, if we want to impose a molecular clock, we should implement ML with a molecular clock, and then ML no longer behaves as described above.

The significance of this example should not be overstated—it does not mean that one 'should' be using MP—it may well be that 'on average' (under some prior distribution on trees and their parameters) $M_{av}L$ outperforms MP, but it does not globally outperform (in the sense described above) MP. This example also does not demonstrate statistical 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 MIL, 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,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 $p(e) \leq \varepsilon$ on two edges incident with two sister leaves and $p(e) > 0.5 - \varepsilon$, 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 statistics of parsimony under a null model

In order to carry out hypothesis tests using the parsimony score of a tree, one needs to know the distribution of this score on a given tree under a suitable null model for generating characters. This approach has been adopted by a number of authors, for example Archie and Felsenstein (1993), Maddison and Slatkin (1991), Goloboff (1991), Steel et al. (1992, 1993b, 1995), Kishino and Hasegawa (1989).

In this section we describe some exact formulae for this problem under certain simple null models. Although these results have been in the literature for several years now, they are not well known, yet they are surprisingly simple and explicit. We consider first the very simplest such null model. In this case there are just two character states and each leaf in the binary tree T has equal probability of being assigned either of the two states. The resulting parsimony score of the character on T is then a random variable, which we denote here as

$L(T)$. Let $P[L(T) = k]$ denote the probability that this parsimony score takes the value k. For example, for any binary tree with 4 leaves, we have $P[L(T) = 2] = {}^4/_{16}$ since there are $2^4 = 16$ binary characters and exactly four of them require two mutations on T. One would like to determine this probability distribution, as well as its mean $\mu(T)$ and variance $\sigma^2(T)$. Several authors (Maddison and Slatkin, 1991; Goloboff, 1991; Archie and Felsenstein, 1993) have constructed recursive formulae for $\mu(T)$. However, it is possible to give exact and explicit formulae, not just for the mean (and variance) but for the entire probability distribution, as we describe shortly. All of these formulae depend only on the number of leaves of the binary tree T, and not on its shape (this surprising, and pleasing property does not extend to characters with more than 2 states). The explicit formulae for the probability distribution and its mean (from Steel, 1993) are:

$$P[L(T) = k] = \frac{(2n-3k)(n-k-1)!2^{k-n}}{k!(n-2k)!}$$

$$\mu(T) = \frac{(3n-2-(-0.5)^{n-1})}{9}$$

where n is the number of leaves of the binary tree T.

Notice that for at least modest-sized binary trees (that is, when $n > 6$) we have the close approximation $\mu(T) \sim n/3$ (here and below 'close' means that the difference between the true value and its approximation goes to zero exponentially fast with n). It is instructive to contrast this with the expected value of $L(T)$ when T is star-shaped (fully unresolved). In that case it can be shown that $\mu(T) \sim n/2$ (Steel, 1993). Thus, the additional edges present in a binary tree allow one to reduce the expected number of mutations required to fit random data from approximately $n/2$ per character (for an unresolved tree) to $n/3$ (for a binary tree), a difference of $n/6$ mutations per character. There is also a slightly more complicated but exact formula for the variance $\sigma^2(T)$ (see Steel, 1993), from which one obtains the close approximation:

$$\sigma^2(T) \sim 2n/27.$$

One can extend this very simple null model in three ways—(i) by allowing more than two character states (ii) by allowing the probability distribution of the states to be non-uniform and (iii) by allowing the probability distribution of the states at the leaves to vary between leaves. Extension (ii) recognises that some states may be more frequent than others, while extension (iii) allows for phenomena such as GC-variation between different genetic sequences.

Even if we allow all of these three extensions ((i)–(iii)) simultaneously, one can still efficiently compute the probability distribution of $L(T)$. An algorithm to do this is described in Steel et al. (1996) and this paper also shows that the limiting distribution of $L(T)$ converges to a normal distribution as the number

of leaves in the binary tree T becomes large. This again applies under the extended null model (allowing (i)–(iii)) subject to a mild technical condition[1]). Of course if we take the cumulative sum of a large number of characters generated independently under this (extended) null model, then the parsimony score of these data on T will also be normally distributed (by the central limit theorem) regardless of whether T has few or many leaves (though if the tree is large the approximation should be much better for a small number of characters).

One can also consider the statistics of the "dual" setting where a character is given, and we wish to find the probability that a binary tree chosen uniformly at random has a given parsimony score for that character. Determining these probabilities provides, for example, a simple formula for the average parsimony score of a collection of characters over all binary trees (Hamel and Steel, 1997). Once again, in the case of binary characters there is, surprisingly, an exact formula for these probabilities, which we now describe.

First, it is easily seen that the number of binary trees having a given parsimony score on a given character depends only on the numbers of species assigned the two states. Consequently, if the number of species assigned the two states is a and b we can denote the probability that a randomly selected binary tree has parsimony length k by $p_k(a,b)$. For example, $p_2(2,2) = 2/3$, since two of the three binary trees on 4 leaves require exactly two mutations to fit a character of type 0011. The *bichromatic binary tree theorem* gives an exact formula for $p_k(a,b)$ as follows.

$$p_k(a,b) = \frac{(k-1)!(2n-3k)N(a,k)N(b,k)}{B(n-k+2)}.$$

In this formula, $n = a + b$ is the total number of species, and $N(m,k)$ is the number of forests consisting of k rooted trees on a total of m leaves, and this is given exactly by the formula:

$$N(m,k) = \frac{(2m-k-1)!}{(m-k)!(k-1)!2^{m-k}}$$

(for $1 \leq k \leq m$), while

$$B(m) = \frac{(2m-4)!}{(m-2)!2^{m-2}}$$

is the number of binary trees on m labelled leaves (and so $B(m) = N(m-1,1)$).

This remarkable formula for $p_k(a,b)$ was first established by Carter et al. (1990) using complicated generating function techniques. However, the combinatorial nature of the quantities in the above formula suggested there should

[1] For some fixed $\varepsilon > 0$, and each state α, each leaf has probability at least ε of being assigned state α.

be a constructive proof of this theorem based on matching up forests of trees. Such a constructive proof was given by Steel (1993) and subsequently simplified by Erdös and Székely (1993). These proofs, and some others involving parsimony lean heavily on a fundamental property of 2-state parsimony, that follows from Menger's theorem in graph theory. For completeness we state this property (established formally as Lemma 1 in Tuffley and Steel, 1997b) as follows:

The parsimony score of a 2-state character on a tree T equals the maximum number of paths that can be placed in T so that (i) each path joins leaves that are assigned different states by the character, and (ii) no two paths share any edge of T.

An example of such a path packing is illustrated in Figure 2. This result is an example of a "min-max" theorem since it relates a quantity we seek to minimize (the number of mutations) to a quantity we maximize (the number of allowed paths in a packing). A clever extension of this theorem to r-state characters has been obtained by Erdös and Székely (1992).

Furthermore, the distribution of trees according to their parsimony score on a fixed character becomes normally distributed as n (the number of species) becomes large, at least for 2-state characters (it is likely also to hold for r-state characters, though this has not yet been rigorously established). More precisely Moon and Steel (1993) show that as n grows, $p_k(a,b)$ becomes normally distributed (as k varies) with mean μn and variance $s^2 n$ where

$$\mu = \frac{2}{3}\left\{1-\sqrt{1-3\frac{ab}{n^2}}\right\}, s = \frac{\mu\sqrt{1-\mu}}{2-3\mu}.$$

There has been only limited success in generalising the bichromatic binary tree theorem to non-binary characters. However, one noteworthy and pleasing result is an exact formula for the probability that a randomly selected binary

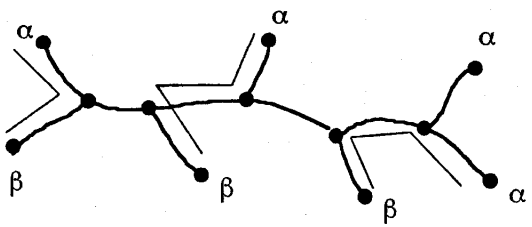

Figure 2. An illustration of the min-max theorem for parsimony score, provided by a tree and a 2-state character with parsimony score 3. Indicated is a maximal system of three edge-disjoint paths, each of which joins a pair of leaves assigned different states by the character.

tree displays no homoplasy for a given r-state character (that is, the character has parsimony score $r-1$ on the tree). For details see Carter et al. (1990) or Steel (1993).

Conclusion

This chapter highlights two contrasting points: First, the parsimonious approach suggested by Ockham's razor can, given information on a common mechanism, support the usual forms of ML over MP for sequence data. Second, when we generalise 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. Indeed, as models become increasingly sophisticated and parameter-rich, one risks losing the ability to discriminate between different underlying trees. Essentially, this is because the data may be able to be described perfectly by any underlying tree, by adjusting the other parameters appropriately. This is a real possibility for site-substitution models that allow a distribution of rates across sites. Indeed there are situations where all trees could perfectly describe the same data, provided one can select, for each tree, a corresponding distribution of rates across sites (Steel, Székely and Hendy, 1994). The model we described earlier (no common mechanism), where MP can be regarded as a ML method, clearly would also have this non-identifiability problem. An interesting problem for future investigation would be to determine the extent to which a stochastic model needs to be constrained in order that the underlying tree can be recovered from sufficient data.

Acknowledgments
I thank the New Zealand Marsden Fund for supporting this research and the Isaac Newton Institute (Cambridge, UK) for its hospitality during the 1998 BFG program. I also thank David Penny, Peter Lockhart and Charles Semple for helpful comments on an earlier version of this manuscript.