Available online at www.sciencedirect.com

SCIENCE DIRECT®

ACADEMIC PRESS

Cladistics 19 (2003) 91–103

Cladistics

www.elsevier.com/locate/yclad

Parsimony, likelihood, and simplicity

Pablo A. Goloboff

Consejo Nacional de Investigaciones Cientificas y Técnicas, Miguel Lillo 205, 4000 San Miguel de Tucumán, Argentina

Accepted 15 February 2003

Abstract

The latest charge against parsimony in phylogenetic inference is that it involves estimating too many parameters. The charge is derived from the fact that, when each character is allowed a branch length vector of its own (instead of the homogeneous branch lengths assumed in current likelihood models), the results for likelihood and parsimony are identical. Parsimony, however, can also be derived from simpler models, involving fewer parameters. Therefore, parsimony provides (as many authors had argued before) the simplest explanation of the data, or the most realistic, depending on one's views. If (as argued by likelihoodists) phylogenetic inference is to use the simplest model that provides sufficient explanation of the data, the starting point of phylogenetic analyses should be parsimony, not maximum likelihood. If the addition of new parameters (which increase the likelihood) to a parsimony estimation is seen as desirable, this may lead to a preference for results based on current likelihood models. If the addition of parameters is continued, however, the results will eventually come back to the same place where they had started, since allowing each character a branch length of its own also produces parsimony. Parsimony can be justified by very different types of models—either very complex or very simple. This suggests that parsimony does have a unique place among methods of phylogenetic estimation.

The two most widely used criteria for phylogenetic inference are parsimony and maximum likelihood. Usually, parsimony is defended by recourse to realism, generality, and economy of assumptions, while maximum likelihood is defended by its explicit use of evolutionary models and the idea that phylogenetic inference must be viewed exclusively as a problem in statistical inference. Parsimony, under some specific models, is also a maximum likelihood estimator; as noted by Farris (1986, p. 24), the "method of maximum-likelihood is not a technique for estimating anything in particular, but a way of deriving estimation procedures from models."

The most widely used maximum likelihood methods are now the ones based on the work of Felsenstein (1973, 1981), reviewed in Swofford et al. (1996), which assume stochastic, Markovian models of evolution, where all the sites have the same probability of change along a branch (a limited amount of rate variation is allowed in some models; see Yang, 1993, 1994). This probability depends on the "length" of the branch (time and mutation rates combined). These models are derived from neutral theories of evolution, which assume that only time and mutation rate are the forces behind most of molecular evolution. Throughout this paper, the term "likelihood" (or "likelihoodist") is used to denote this type of method (or the people espousing its use).

Parsimony was not originally justified by means of an explicit probabilistic model. In the belief that only methods based on explicit probabilistic models are defensible, the likelihoodists have tried to discover (starting with Felsenstein, 1978) "the model" implicit in parsimony; the resulting findings have been used to criticize the assumptions supposedly required to justify parsimony.

Some authors have defended parsimony from a philosophical perspective (Kluge, 1997, 2001; Siddall, 1997, 2002), but most likelihoodists (with few exceptions, such as de Queiroz and Poe, 2001) have ignored these philosophical issues. The purpose of the present paper is to revise some aspects of the parsimony vs likelihood controversy, from a more statistical perspective. Recent criticisms of parsimony accuse it of relying

E-mail address: instlillo@infovia.com.ar.

0748-3007/03/$ - see front matter © 2003 The Willi Hennig Society. Published by Elsevier Science (USA). All rights reserved.
doi:10.1016/S0748-3007(03)00017-3

on an implicit model that is too complex and therefore over-fits the data (this charge has now replaced earlier criticisms that accused it of being too simplistic—e.g., Felsenstein, 1982, p. 388; Felsenstein, 1988, p. 535). My main conclusion is that, since most parsimonious trees are maximum likelihood estimates under different models (either very simple or very complex), parsimony can be seen as providing either the simplest explanation of the data or the most realistic. Current likelihood methods lie in between. More importantly, the fact that parsimony can be derived under very different types of models also casts doubt on the notion that one can evaluate a method justified on logical grounds by simply evaluating statistical models that happen to produce similar results.

Philosophy

Many likelihoodists (e.g., Edwards, 1996; Felsenstein, 1973, 1978, 1988; Goldman, 1990; Yang et al., 1995) have claimed that methods of phylogenetic inference can be properly justified only by recourse to statistical reasoning, instead of the logical and philosophical arguments often advanced in favor of parsimony (e.g., Farris, 1983, 1986). They often portray advocates of parsimony as people who are unaware of the sound statistical principles behind maximum likelihood and misunderstand the technical aspects of maximum likelihood. Felsenstein's comment on the work of two statisticians (Barry and Hartigan, 1987a) is a good example of this attitude:

> after coping with taxonomists, who tend to dismiss statistical inference and adopt arbitrary and bizarre "hypothetico-deductive" philosophical frameworks, it is refreshing to deal with statisticians, who are not tempted to replace the hard work of inference by philosophical quotation-mongering (Felsenstein, 1987, p. 208).

Writing for the general public Edwards (1992) appeared to have a very open attitude toward science and philosophy; the introductory remarks for his book on likelihood could well have been written by a philosophically inclined pattern cladist:

> The incentive for contemplating a scientific hypothesis is that through it we may achieve an economy of thought in the description of events, enabling us to enunciate laws and relations of more than immediate validity and relevance. The classical concepts of probability allow us to extend our activities into the realm of uncertainty, for it appears that even the most random of events, such as the results of a penny-tossing experiment, exhibit, in the aggregate, certain regularities. The greater the regularity of pattern in a sequence of events, the more we feel compelled to seek an 'explanation' in terms of a law ... It is our task to detect regularity in the presence of confusion, order in the presence of chaos. It will not be sufficient, when faced with a mass of observations, to plead special creation, even though, as we shall see, such a hypothesis commands a higher numerical likelihood than any other. We prefer more general and more simple hypotheses (Edwards, 1992, p. 1).

But when discussing the ideas of those who attempt to justify parsimony on general philosophical principles, Edwards (1996) had a rather different attitude:

> It is surprising to find the philosophy of systematics ensnared in prestatistical arguments, as though Darwin had lived before Pascal, Bernoulli, Gauss, and Laplace, and that therefore the implications of Darwin's revolutionary hypothesis have to be studied without reference to modern theories of scientific inference (p. 81).

> Any approach that attempts to grapple with inference under uncertainty without using ideas from the theory of probability is unlikely to command scientific respect (p. 89).

Since "inference under certainty" cannot exist in real empirical research (from physics to anthropology), Edwards' statement amounts to saying that no conclusion established without strict statistical reasoning is scientific.

Admittedly, not all defenders of likelihood will present the problem in terms as extreme as Felsenstein or Edwards. Although they discuss the problem of phylogenetic inference as a purely statistical problem, Swofford et al. (1996) present their views from what seems a very moderate and eclectic perspective:

> It is often argued that it is circular to model character change for the purpose of estimating a phylogeny because we cannot begin to understand the processes of character change without first knowing the tree. We prefer, instead, to think of the problem as one of "reciprocal illumination" (Hennig, 1966): having some idea of the phylogeny is relevant to the development of good models, but ever-improving models can also lead to better phylogenetic inferences. Thus, both classes of methods are useful and important (p. 409).

Their argument for "reciprocal illumination" is appealing at first,[1] but not so much when one considers that practice among likelihoodists always falls short of this scenario. In fact, one can seriously doubt that Swofford et al. (1996) truly believe that parsimony is so "useful and important," because they very clearly point out what they perceive as the general advantages of maximum likelihood but never actually inform the reader what the advantages of parsimony are supposed to be (this asymmetry is also quite obvious in the concluding remarks of Swofford et al., 2001, p. 538). Their abstract praising of parsimony seems more intended to

[1] Note that Hennig (1966) actually meant by "reciprocal illumination" the consideration of independent evidence (e.g., reexamination of characters in case of conflict, agreement with other sources of evidence) that could bear on a phylogenetic hypothesis. By "reciprocal illumination" Swofford et al. (1996) mean instead the theoretical improvement of existing models using the implications of accepted phylogenies (which is in fact very rarely done).

create the impression of open-mindedness and the impression that they avoid unfair criticisms of other authors or methods.[2] The impression of open-mindedness and fairness of judgment makes superficially convincing their almost casual dismissal of arguments against likelihood (e.g., arguments by Farris (1983, 1986), are not fully discussed in the text, but just misrepresented and rejected in a footnote (p. 427)).

Defendants of parsimony are often more concerned with epistemology than with statistics. Likelihoodists would make us believe that the controversy, once the superior value of statistical reasoning is accepted, can be easily or automatically resolved. However, accepting statistical reasoning as supreme under all circumstances requires itself philosophical considerations. On top of this, whether a given model of evolution is considered valid cannot be decided *only* on a statistical basis; the decision will "also be influenced by the simplicity of the hypothesis, by their relevance to other situations, and by a multitude of subtle considerations that defy explicit statement" (Edwards, 1992, p. 34). Therefore, to practitioners of parsimony, the increase in precision gained by applying a rigorous statistical methodology is—in the face of so many imponderables—entirely illusory. Rather than trying to attribute a degree of statistical confidence to phylogenetic hypotheses, it seems more honest to acknowledge that phylogeny estimation has a strong element of irreducible uncertainty. On the other hand, even if a statistical approach is adopted, it does not follow automatically that current likelihood methods are the best estimation procedure. For example, Sanderson and Kim (2000), advocates themselves of the statistical approach, argue against the very use of parametric models. If one adopts their point of view, much of the present discussion (cast in terms of parametric estimations) is rendered irrelevant.

Contrary to most likelihoodists, I do not consider that philosophy is irrelevant to the controversy; many aspects of the likelihood vs parsimony controversy do involve philosophical points. The so-called "statistical viewpoint" of phylogeny estimation, taken to the extreme, means that no method can be considered justified *in general* ; each individual case will require use of just that method most likely to recover the true tree in that specific case. Those who approach the problem from the more philosophical side are trying to decide whether some general method or principle can guide phylogenetic inference in all its applications—a deeper justification, in some sense. Likelihoodists, however, have repeatedly made it clear that they are not willing to listen to this kind of argument, and therefore my general discussion is cast in more statistical terms.

Consistency and simplicity

In earlier literature, the property of statistical consistency (i.e., the property to converge on the true tree when the underlying model generated the data and many characters are sampled) was loudly voiced as the main advantage of likelihood methods. The emphasis on consistency on the part of likelihoodists has gradually decreased. It decreased with the realization that maximum likelihood can be inconsistent even with minor violations of the model (Chang, 1996a). It decreased with the realization that, given some evolutionary models, even maximum likelihood estimators could suffer inconsistency (Farris, 1999; Steel et al., 1994). It decreased with the realization that parsimony can be consistent (Steel et al., 1993). It decreased with the realization that, even if likelihood was a more accurate method in principle, inferences based on trees suboptimal under likelihood could be less reliable than inferences based on trees actually optimal under otherwise inferior but faster criteria (Sanderson and Kim, 2000). It decreased with the realization that (under some models) parsimony may be more likely than maximum likelihood to find the correct tree, given finite amounts of data (Pol and Siddall, 2001; Siddall, 1998; Yang, 1997).

As the emphasis shifted away from consistency, advocates of parsimony (e.g., Farris, 1999, 2000; Siddall and Kluge, 1999) often cited Tuffley and Steel (1997) in their support. Tuffley and Steel (1997) demonstrated that parsimony is a maximum likelihood estimation when each site can have its own branch length, and this (according to those who defend parsimony) is an indication that parsimony is simply a more realistic model: it does not force uniform probabilities of change onto all characters. Defenders of likelihood (e.g., Lewis, 2001, p. 914; Steel, 2002, p. 133; Steel and Penny, 2000, p. 843; D. Swofford, pers. comm.)[3] reacted by pointing out that the model that assumes uniform probabilities for all sites is simpler (i.e., has fewer parameters) than the model of Tuffley and Steel, where each site has its own branch length. Simplicity has always been recognized as desirable in scientific inference:

> If we are to pursue the fundamental idea that similar circumstances have similar consequences, then we must formulate a law which embodies the similarities. ... It follows that the law will be simpler than the observations if it is to achieve anything. The wider the circumstances to which it is to apply, the simpler it will be; and since our natural interest is in laws which express the similarity in a wide variety of circumstances, our natural interest is in simple laws. A law of wide applicability contains few 'ifs'

[2] To be fair, Swofford et al. (1996) may well believe this themselves. I have no way to know.

[3] Swofford maintained this position during the discussion—in which I took part—of a seminar given by James S. Farris at the Smithsonian Institution (May 2001). He also made the same charge against parsimony in a seminar that he gave at Columbia University (November 1999; I did not attend this seminar but heard accounts of it from D. Pol, J. Faivovich, and G. Giribet).

> and 'buts' to cover special circumstances, and a law with few 'ifs' and 'buts' is what we call simple (Edwards, 1992, p. 200).

Likelihoodists point out that statisticians eschew the use of too complex a model, because, although this increases fit, the results become less predictive and explanatory, more computationally demanding, and more prone to errors. Phylogenetic inference, the likelihoodists then claim, should always use the simplest possible model, using more parameter-rich models if (and only if) the likelihood (fit) is significantly higher. The starting point of a phylogenetic analysis should therefore be a model like Felsenstein's (1981), which (in using uniform branch lengths for all sites) is much simpler than parsimony. What parsimony does, according to this line of reasoning, is like fitting a curve that passes through each and every one of the data points, instead of using a straight line. Such a curve has a perfect fit, but no predictive value; any one new data point will certainly fall outside the previously specified trajectory, which (if fit is to be preserved) will have to be modified every time a new data point is added.

While the preference for simpler hypotheses is hardly objectionable, it is far from obvious that parsimony really requires such a complex formulation. For example, "perfect fit" is certainly never obtained by applying parsimony to real data sets of some size; otherwise, phylogeneticists would not have been struggling for decades with ways to deal with homoplasy. If parsimony is so complex a model (and given that estimation procedures with more parameters always involve more computations), it is surprising (as noted by M. Steel, in Sanderson and Kim (2000)) that parsimony requires so little computational work, as compared to a likelihood method like Felsenstein's (1981). All this suggests that a closer look at the actual simplicity of parsimony and likelihood is required.

Integrated and maximum relative likelihood

In phylogenetics, likelihoodists have used only what is known as maximum relative likelihood, even if admitting that the ideal estimation is integrated likelihood. Calculating the actual integrated likelihood of a tree would require a probabilistic model of branch lengths for the given tree topology (as early recognized by Felsenstein (1973); the same point was made by Farris (1973)):

> To evaluate the likelihood of a topology τ [for data D and model M], we would calculate
>
> $$P_M(D|\tau) = \int_s P[\text{times}|\ \tau]\ P[D|\text{times}, \tau], \tag{4}$$
>
> where we integrate over the set S of all branch point times compatible with the topology τ ... The methods for doing this have not been developed ... An easier approach would be simply to estimate both topology and branch point times, and then to ignore the branch point time estimates. Such a procedure does not make fully efficient use of the data, but it will have to suffice until methods for calculating (4) [the integrated likelihood] have been devised (Felsenstein, 1973, p. 243).

Thus, for tractability, the estimation of the branch point times and topology is done under the assumption that the parameters (branch lengths, tree) take values such that $P(D|\tau)$ is highest.[4] The parameters corresponding to branch lengths become a type of nuisance parameter, in that (even if the only interest is in the topology τ) they have to be estimated to determine the value of $P(D|\tau)$. Assuming that the parameters maximize the likelihood may be problematic (as noted by E. Sober, in Felsenstein and Sober, 1987; Goldman, 1990; Steel and Penny, 2000), since some of the values that maximize $P(D|\tau)$ may themselves be very improbable. The problems become worse when there are many nuisance parameters (i.e., many branch lengths).

In the usual likelihood formulation, a single set of branch lengths is chosen, but all the possible pathways (reconstructions of ancestral states) that could have led to the data are considered. This formulation, according to likelihoodists, integrates ancestral reconstructions into the model. In this way, the probability assigned to a given site does not depend on a given reconstruction, and ancestral states are actually not estimated (Felsenstein, 1973, 1978). According to likelihoodists, specific assignments of ancestral states to internal nodes are (just like branch lengths) nuisance parameters, which (unlike branch lengths) increase more and more as new characters are added. The "pruning" algorithm of Felsenstein (1981) allows calculating more or less efficiently the individual conditional likelihoods at each node of the tree, in a postorder traversal of the tree (see Swofford et al. (1996), for review; see also Fig. 2). If branch lengths are changed one at a time (under the "pulley" principle of Felsenstein (1981)), the change in likelihood for the total tree is easy to derive (using only the conditional likelihoods of the nodes delimiting the branch); thus the optimal branch lengths are found in practice by iteratively adjusting each of the branch lengths, until the likelihood cannot be further improved by adjusting branches one at a time. This branch-length fitting procedure is the most time-consuming part in searches under maximum likelihood (and it is likely to always keep likelihood well below

[4] In practice, other parameters (e.g., probability ratios for all possible nucleotide transformations) are often also estimated, using the same approach; throughout, my discussion holds regardless of whether those parameters are given or estimated.

parsimony with regard to speed; see Sanderson and Kim (2000, pp. 822, 823)).[5]

Once the optimal branch lengths have been found, they are discarded, and only the tree topology is retained. However, since the maximization of $P(D|\tau)$ involves fixing branch lengths, this in effect amounts to considering trees of identical topology but different branch lengths (Fig. 1) as different trees. Since we are interested only in choosing topologies (i.e., hypotheses of monophyly), then it seems more logical to evaluate trees regardless of branch lengths. For a single tree, this requires summing the probabilities of the data, given the topology and all possible combinations of branch lengths. In the absence of more detailed models of branch lengths, this is not exactly equivalent to the integrated likelihood of Felsenstein's (1973, his formula 4), which multiplies each combination of branch lengths by its (prior) probability. Here, the values of $P(D|\tau, \lambda)$ for different branch lengths λ are simply summed. The approach, however, seems more consistent than just integrating reconstructions; after all, just as different assignments of ancestral states form plausible reconstructions, different branch lengths are also plausible.[6]

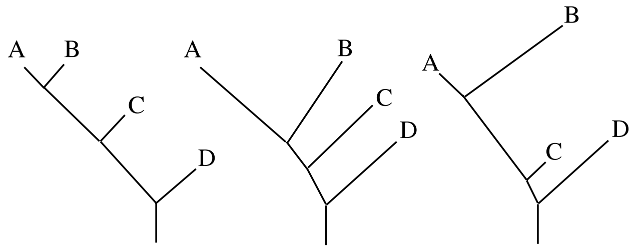

Fig. 1. Three trees with identical topologies, but different branch lengths. Current models of maximum likelihood will consider one of these trees better than the others, even if their topologies are identical.

The integration of branch lengths could be done in two different ways. One could calculate first the sum of likelihoods (or, equivalent for tree selection, their average) for each of the individual sites, under the same range of branch lengths for all sites, and then multiply the site likelihoods. This is easily done. Consider the postorder likelihood calculations for a fixed set of branch lengths, for a given site in the tree in Fig. 2 (using, for simplicity, only two states). In that tree, node **X** gives rise to **A** and **B**, and node **Y** gives rise to **C** and **X**; $\mathbf{b_i}$ is the length of the branch leading to node **i**. If only the terminals descended from **X** are considered (and branch lengths are fixed), the probability of obtaining the observed data for **A** and **B** if node **X** had state **i** is $\mathbf{L_{iX}}$. The probability of a state remaining unchanged along a branch **i** ($\mathbf{P_{00i}}$ and $\mathbf{P_{11i}}$), or changing ($\mathbf{P_{01i}}$ and $\mathbf{P_{10i}}$), is a function $\mathbf{f_{(b)}}$ of the length $\mathbf{b_i}$ of the branch (in a Neyman, 1971 two-state model, this would be $0.5 - 0.5e^{-b}$ for a different state and $0.5 + 0.5e^{-b}$ for the same state; other models would use different formulae, but this makes no difference for the present argument). Once the conditional likelihoods at node **X** have been calculated, it is possible to calculate the conditional likelihoods at node **Y**. The conditional likelihoods at the root node are final, for the given branch lengths. There is an infinite number of combinations of lengths for the branches $\mathbf{b_A}$, $\mathbf{b_B}$, $\mathbf{b_C}$, $\mathbf{b_X}$, and $\mathbf{b_Y}$. Note, however, that the conditional likelihoods at node **Y** depend (by multipli-

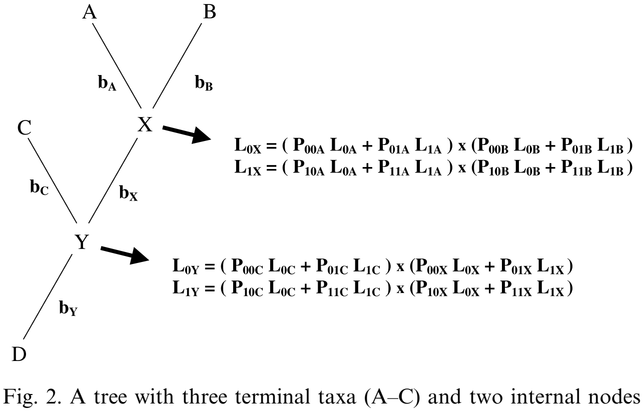
A
B
b_A
b_B
C
X
$L_{0X} = (P_{00A} L_{0A} + P_{01A} L_{1A}) \times (P_{00B} L_{0B} + P_{01B} L_{1B})$
$L_{1X} = (P_{10A} L_{0A} + P_{11A} L_{1A}) \times (P_{10B} L_{0B} + P_{11B} L_{1B})$
b_C
b_X
Y
$L_{0Y} = (P_{00C} L_{0C} + P_{01C} L_{1C}) \times (P_{00X} L_{0X} + P_{01X} L_{1X})$
$L_{1Y} = (P_{10C} L_{0C} + P_{11C} L_{1C}) \times (P_{10X} L_{0X} + P_{11X} L_{1X})$
b_Y
D

Fig. 2. A tree with three terminal taxa (A–C) and two internal nodes (X, Y), showing how the conditional likelihoods are determined (after Felsenstein's (1981) pruning algorithm; for simplicity, the example assumes a two-state character). L_{jK} is the likelihood of the data observed for the descendants of node K, given that the node K has state j; P_{ijK} is the probability of ending in state j at node K if the ancestor had state i (this is a function of the length, b_K, of the branch leading to node K).

[5] Sanderson and Kim (2000) note that parsimony has a significant advantage over likelihood in that the evaluation of a candidate tree during a search "can be accomplished very efficiently in time linearly proportional to the number of taxa … by way of the Fitch–Hartigan algorithm." Sanderson and Kim, however, grossly underestimate the speed of actual parsimony calculations under branch-swapping. For T terminal taxa, the algorithms described by Goloboff (1996, 1999) use 6T times fewer operations than direct application of the Fitch–Hartigan algorithms. Therefore parsimony can evaluate trees at the same speed for any number of taxa (or even use less time to evaluate trees with more taxa). To be fair to likelihood, however, the optimization of branch lengths from scratch for each tree examined during branch-swapping (as done in current programs; see Rogers and Swofford (1998)) is also unnecessary; Barry and Hartigan (1987a) had already proposed to calculate the likelihood (when adding a terminal or a group to a subtree) by optimizing only the three branches subtending the newly created node. For real data sets, this produces so much error (pers. observ.) that the approximate evaluation becomes almost meaningless; a modified procedure based on the same idea, however, may produce more meaningful evaluations, by optimizing the length of only a reduced number of branches around the new node, for each tree examined during branch swapping (P. Goloboff, unpublished). With this, the evaluation of a tree during branch-swapping with maximum likelihood could be (depending on the desired precision and/or the greediness of the data) from T/10 to T/20 times faster than current implementations.

[6] Note that Bayesian analysis (Huelsenbeck and Ronquist, 2001; Larget and Simon, 1999; Yang and Rannala, 1997) treats this integration of branch lengths as an essential desideratum of the analysis, since the branch lengths themselves are part of the parameter space to be explored with the Monte Carlo Markov Chain.

cation) on the conditional likelihoods at node **X**, and those of node **X** correspond to the multiplication of a factor corresponding to the left branch, $\mathbf{b_A}$, and a factor corresponding to the right branch, $\mathbf{b_B}$. Thus, if the length of the branch $\mathbf{b_A}$ is fixed, it is possible to change the lengths of the other branches (the results will be carried over by the multiplication); likewise for the other branches. Thus, the average probabilities can be calculated one branch at a time. Therefore, to calculate $\mathbf{L_{0X}}$ changing branch $\mathbf{b_A}$, it is necessary to first calculate the average value of $\mathbf{P_{00A}L_{0A}} + \mathbf{P_{01A}L_{1A}}$. Since $\mathbf{L_{0A}}$ and $\mathbf{L_{1A}}$ are fixed at this point, this is equivalent to

$$\overline{\mathbf{P}}_{\mathbf{00A}}\mathbf{L_{0A}} + \overline{\mathbf{P}}_{\mathbf{01A}}\mathbf{L_{1A}}.$$

The average probabilities of change and stasis for a given range ($\mathbf{t_0}$ to $\mathbf{t_1}$) of branch lengths are easily determined with

$$\overline{\mathbf{P}} = \frac{\int_{t_0}^{t_1} \mathbf{f_{(t)}dt}}{\mathbf{t_1 - t_0}}.$$

For a two-state Neyman or Jukes–Cantor type of model, the probability $\boldsymbol{\alpha}$ of stasis is

$$\overline{\mathbf{P}}_{\mathbf{00}} = \overline{\mathbf{P}}_{\mathbf{11}} = \frac{\int_{t_0}^{t_1} (\mathbf{0.5 + 0.5e^{-t}})\mathbf{dt}}{\mathbf{t_1 - t_0}}$$

and the probability $\boldsymbol{\beta}$ of change is

$$\overline{\mathbf{P}}_{\mathbf{01}} = \overline{\mathbf{P}}_{\mathbf{10}} = \frac{\int_{t_0}^{t_1} (\mathbf{0.5 - 0.5e^{-t}})\mathbf{dt}}{\mathbf{t_1 - t_0}}.$$

If the range of branch lengths is from 0 to infinity, change and stasis are equiprobable, and no tree choice is possible. The range of branch lengths can be logically bounded between 0 and some positive number (since both time and mutation rate are bounded). In real maximum likelihood analyses, a branch length of 2 is considered very long. For a range 0–2, $\boldsymbol{\alpha} \approx 0.72$, and $\boldsymbol{\beta} \approx 0.28$.

The postorder traversal of the tree (applying the pruning algorithm of Felsenstein (1981)) now can be done with regard to these fixed average probabilities of stasis, $\boldsymbol{\alpha}$, and change, $\boldsymbol{\beta}$. This is a simpler method, because it avoids estimation of a host of nuisance parameters; the branch lengths become instead incorporated into the model. Accordingly, the computational cost of calculating this integrated likelihood is much lower than that for the maximum relative likelihood proposed by Felsenstein (1981).[7] Interestingly, Goldman (1990) considered that this type of model, where probabilities of change and stasis are the same across all branches, lacked a "time structure." However, it is precisely the model that results from considering alternative branch lengths in the likelihood calculations.

The most remarkable result of integrating out branch lengths is that it produces probabilities of change fixed along all branches. Branch lengths become therefore irrelevant to choosing trees (Fig. 3). This is similar to the model proposed by Sober (1985), which he intended as a model for parsimony. For unrooted trees of four taxa (with no missing entries), this always produces the same results as parsimony (see Fig. 3), but this need not be so for more taxa or rooted trees. Goldman (1990) already showed that a method with fixed probabilities of stasis and change is not generally equivalent to parsimony, using four taxa and one additional (root) node. Additional examples showing differences between this method and parsimony are illustrated in Fig. 4. This method (just like parsimony for four taxa) can be inconsistent. The inconsistency, however, comes only from considering that different combinations of branch lengths are plausible alternatives.

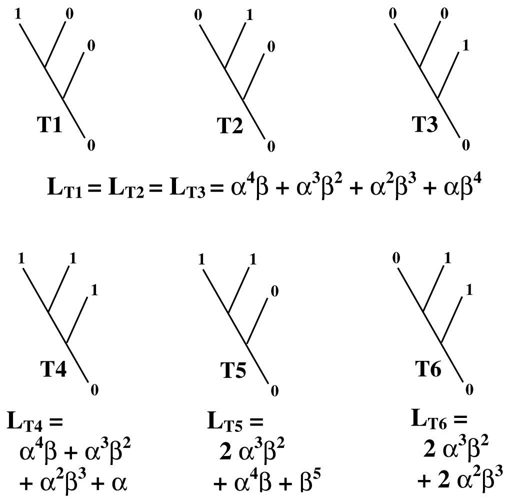
1 0 0 0
T1
0 1 0 0
T2
0 0 1 0
T3
$L_{T1} = L_{T2} = L_{T3} = \alpha^4\beta + \alpha^3\beta^2 + \alpha^2\beta^3 + \alpha\beta^4$
1 1 1 0
T4
1 1 0 0
T5
0 1 1 0
T6
$L_{T4} = \alpha^4\beta + \alpha^3\beta^2 + \alpha^2\beta^3 + \alpha$
$L_{T5} = 2\,\alpha^3\beta^2 + \alpha^4\beta + \beta^5$
$L_{T6} = 2\,\alpha^3\beta^2 + 2\,\alpha^2\beta^3$

Fig. 3. Example showing the implications of considering alternative branch lengths when calculating the likelihood of each site, which leads to probabilities of stasis (α) and change (β) uniform for all branches (see text). For simplicity, only two states are considered. There are six nontrivial (i.e., nonuniform) types of character distributions for a tree of four taxa (where one taxon is considered as the ancestral node). The individual likelihood contribution of each type, given values of α and β, can be calculated either by enumerating possible reconstructions or by applying the postorder conditional calculations and operating algebraically. Note that types 1, 2, and 3 interconvert when switching between trees, so that these types of characters (with identical individual likelihoods) do not not influence tree choice. Since the same taxon is always used to root the tree, type 4 remains the same for any tree (and has a likelihood identical to those of types 1–3). The only types relevant for tree choice are therefore types 5 and 6, as in parsimony. For any value of $\alpha \neq \beta$, type 5 ("synapomorphy") has a higher likelihood than type 6 ("parallelism"), so that the tree that invokes the fewest parallelisms is always preferred.

[7] Shortcuts similar to those used by Goloboff (1998) for Sankoff characters could be used here during branch-swapping (e.g., deriving the likelihood of a tree produced by joining two subtrees from the conditional likelihoods of the nodes delimiting the branches to be joined). This would produce tree searches about as fast as those for Sankoff parsimony (i.e., hundreds of times faster than current branch-swapping for maximum relative likelihood).

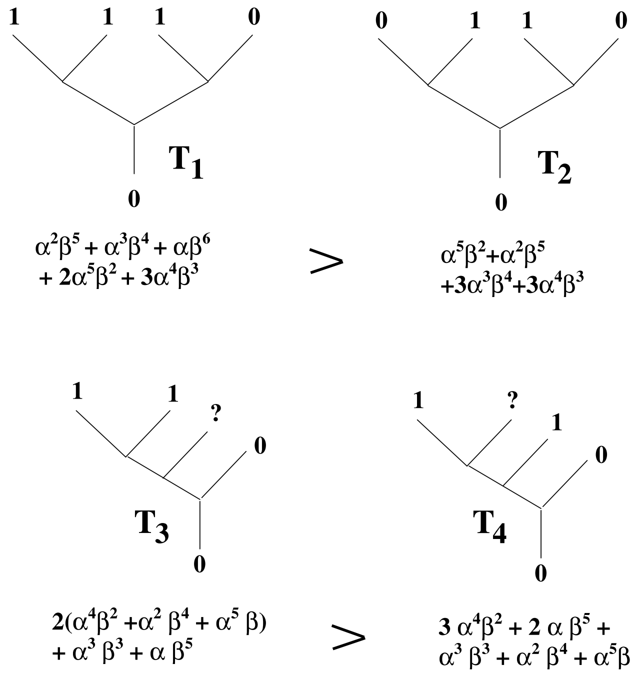
1 1 1 0
T_1
0
$\alpha^2\beta^5 + \alpha^3\beta^4 + \alpha\beta^6 + 2\alpha^5\beta^2 + 3\alpha^4\beta^3$
>
0 1 1 0
T_2
0
$\alpha^5\beta^2 + \alpha^2\beta^5 + 3\alpha^3\beta^4 + 3\alpha^4\beta^3$
1 1 ? 0
T_3
0
$2(\alpha^4\beta^2 + \alpha^2\beta^4 + \alpha^5\beta) + \alpha^3\beta^3 + \alpha\beta^5$
>
1 ? 1 0
T_4
0
$3\alpha^4\beta^2 + 2\alpha\beta^5 + \alpha^3\beta^3 + \alpha^2\beta^4 + \alpha^5\beta$

Fig. 4. Two cases where the integrated likelihood (produced by considering alternative branch lengths, with average probability of stasis α and average probability of change β) considers trees with the same numbers of steps as having a different likelihood. Tree 1 is better than 2, and 3 is better than 4. In both cases, the best tree is the one that has an ambiguous optimization (and therefore has more reconstructions with minimum transformations; these contribute more to the total likelihood score). Note that trees 3 and 4 are identical, except for the placement of the taxon with a missing entry.

An alternative integration of branch lengths is possible, although it is computationally more difficult. Instead of calculating the likelihood for each site under a range of branch lengths, one could calculate the likelihood for all the data (i.e., the product of the individual site likelihoods), for different combinations of branch lengths. This has the drawback that the average likelihood of individual sites cannot be obtained from the calculations, but (even if considering $P[\text{times}|\ \tau]$ the same for all branch length combinations) it may be closer to the intent of Felsenstein (1973). As he noted, the integrals are here much more complicated, and there seems to be no general form to solve them. However, computers are now much faster than in 1973, and this allowed me to examine the behavior of this method with brute force, by considering (for the unrooted four-taxon, two-state case, under a Neyman model) a large number of different branch length combinations. The range of lengths considered was 0–1 for the five branches in the four-taxon network; 100 different branch lengths were considered for each branch (from 0 to 1, increasing by 0.01); this requires considering $100^5 = 10^{10}$ different combinations of branch lenghts. The fourth taxon was considered as an all-0 ancestor (the results are simply mirrored if the state 1 for the ancestor is considered also). The probabilities for each of the possible four reconstructions at the internal nodes were then evaluated for each combination of branch lengths (summing them up). The frequency (= probability) of each one of the eight possible types of characters (the eight different combinations of 0/1 in three taxa A, B, and C) was calculated for a fixed combination of branch lengths in the model tree. These probabilities were used to estimate the average likelihood, under different branch lengths for the model tree, for the three possible trees for four taxa (e.g., if a given character type has a probability **X** under the branch lengths of the model tree, and the character has an individual likelihood of **L** under a given set of branch lengths in the estimated tree, the likelihood contribution of that character was calculated as $\mathbf{L}^{\mathbf{x}}$). Although long branches are less strongly attracted in this method than in parsimony, the resulting method is not statistically consistent, showing some attraction of long branches. For example (using for the model tree the same notation as that in Fig. 2), when branch lengths $\mathbf{b_A} = \mathbf{b_X} = \mathbf{b_Y} = 0.02$ and $\mathbf{b_B} = \mathbf{b_C} = 0.5$, the model tree (((AB)C)D) has a lower likelihood than the wrong tree (((BC)A)D) (the estimated average likelihoods are, respectively, 0.0822177 and 0.0835179, or 1.5% lower for the model tree).

The inconsistency produced by either type of integration makes it unlikely that likelihoodists will accept any of them, even if it seems a more proper procedure than choosing the parameters that maximize the likelihood. In the words of Felsenstein (1973):

> my estimates of the tree topology are obtained by first estimating more than the topology, then dropping some of that information. This is not the same as making a maximum likelihood estimate of the topology. Only the expression based on (4) is the maximum likelihood estimate of the topology. If we cannot use (4), either because we have no model of branching to give us $P[\text{times}|\ \tau]$ or because we cannot evaluate the integrals, my procedure would seem to have at least one major advantage, consistency (p. 246).

It is now seen that, in the absence of a model for $P[\text{times}|\ \tau]$, the advantage of consistency cannot be claimed for formulations that do not depend on estimating specific values for branch lengths. This raises an interesting question. As pointed out by Yang (1996, p. 304), there are some significant differences between the conventional maximum likelihood estimation and the maximum likelihood estimation of a tree topology as in Felsenstein's formulation; given these differences, consistent estimations of tree topology are not guaranteed by Wald's (1949) conditions (which Felsenstein (1973), had improperly cited as providing proof of the consistency of his method; see Farris (1999)). Rogers (1997) proved, however, that topology estimation under Felsenstein's formulation is consistent. As shown above, if Felsenstein's formulation is changed so that only the topology is estimated—integrating branch lengths—the method becomes inconsistent. Rogers's proof was based on demonstrating that sums of branch lengths along the path between different taxa are consistently estimated (Rogers, 1997, p. 357). Once all the branch lengths between pairs of taxa are consistently estimated, a tree topology is automatically determined—as a by-product of the branch length estimations, so to speak.[8] Given Felsenstein's model, perhaps no method can estimate the tree topology alone consistently.

Other formulations

Aside from consistency, the most likely argument against considering alternative branch lengths is that the data themselves make some branch lengths more likely (or better) than others. If that line of reasoning is accepted, defending the formulation that maximizes branch lengths on those grounds also implies that one should also choose the individual reconstructions (considered by likelihoodists as a nuisance parameter) to maximize the likelihood. Choosing values to maximize likelihood for one type of parameter, but not for the other, seems logically inconsistent. Barry and Hartigan (1987a) were the first to formally propose the idea that one can choose the parameters that maximize the likelihood for both branch lengths and reconstructions. They called their procedure "most parsimonious likelihood" and used both optimal branch lengths and individual reconstructions to maximize likelihood. It is not entirely clear how they chose state assignment to the interior nodes; they state that "the values of the internal nodes are usually assigned to agree as much as possible with neighboring nodes" (p. 200). They estimated the complete transition matrices at each node (they were aware that the resulting probability model is not identifiable; see 1987a, p. 201), but a fixed one could be used for all the branches.

As noted by Barry and Hartigan themselves (1987b), this method may produce inconsistent estimations. As respectable as Felsenstein may have considered statisticians, they are obviously also capable of producing seriously inconsistent methods. For the four-taxon, two-state case (under a Neyman model), using as the likelihood of a tree the best branch lengths and the best (unconstrained) reconstructions produces results which are hardly defensible. Consider the case where the model tree is (((AB)C)D), with branches leading to B and C long and all other branches short. Since change is seen as more likely along long branches than along short branches, all the changes are pushed toward the long branches, and no change is implied along the intermediate branch. This method therefore tends to distort branch lengths (shortening short branches, lengthening long ones). Even in cases of small differences in branch length (where parsimony still performs consistently), this method may lead to prefer the wrong trees. When the underlying branch lengths in the model tree are $\mathbf{b_A} = \mathbf{b_X} = \mathbf{b_Y} = 0.6$ and $\mathbf{b_B} = \mathbf{b_C} = 0.8$, the tree with the highest likelihood (calculated as before) is ((BC)(AD)) (0.131487, with branch lengths set to $\mathbf{b_A} = \mathbf{b_B} = 0$, $\mathbf{b_X} = 1.4$, $\mathbf{b_C} = 2.2$, and $\mathbf{b_Y} = 1.8$; the model tree has a likelihood of 0.113415). With these branch lengths in the model tree, parsimony is consistent. Under the same model tree, if the reconstructions are restricted to be "parsimonious,"[9] Barry and Hartigan's procedure leads to preference for the correct tree; restricting assignments based on parsimony considerations improves the situation. Even under such restriction, however, other combinations of branch lengths in the model tree produce branch length repulsion or translocation (apparently, never attraction).

Perhaps the most interesting aspect of Barry and Hartigan's formulation is that it is precisely a simplification of this method that produces the derivation of parsimony proposed by Goldman (1990). Goldman showed that parsimony is a maximum likelihood estimator when the probabilities of change and stasis are fixed (at any given value) across all the branches of the tree and across all characters, as long as the probability of change, β, is less than the probability of stasis, α. A reconstruction with a step along **n** branches (and no change along **m**) implies a probability $\alpha^m\beta^n$. As long as $\alpha > \beta$, the expression $\alpha^m\beta^n$ increases as **n** (= steps) decreases, and thus reconstructions (or trees) with fewer steps imply a higher probability. Goldman's is a simpler method than Barry and Hartigan's, since it does not rely on estimating branch lengths.

Assumptions and specious arguments

Goldman (1990) claimed that his model showed certain assumptions implicit in parsimony. He criticized (p. 356) the idea that probabilities of change could be constant over time, the idea that "an event ... is as likely

[8] An earlier, more abstract, proof of consistency was provided by Chang (1996b). If I understand his proof correctly, the same comments apply to it, since it is based on the probability of joint distributions among pairs of taxa.

[9] This was done by assigning to interior nodes the state present in two of the three neighboring nodes (the cases examined had only two states and no missing entries). Reconstructions were ignored when an interior node had the state present in only one neighbor node. When there are more than four taxa this is "parsimonious" in a general sense, but may not imply most parsimonious reconstructions. Consider the tree ((((0 1) 1) 1) 0), where all the branches leading to the 1's are very long and the other branches are very short. In such a situation, assigning state 0 to the three internal nodes satisfies the requirement and produces a higher likelihood than assigning state 1 (the most parsimonious assignment).

to occur during a short period of time as during a long one." He continued that in such a model the "probabilities depend only on the structure of a hypothesized tree; more precisely… only on the branching events that have occurred." He considered this a weakness of the model, as it "implies we should consider all of the lineages representing descendants of the "root" species", which we would fail to do "if, for example, we studied a subset of all the mammals."

Goldman's idea that parsimony trees lack a "time structure" but display only "tree structure" is surprising, in that "tree structure" means precisely *a temporal sequence of branching events*. The implied probabilities depend simply on whether (or how well) the branching sequence hypothesized in the tree matches the one implicit in the observations. If the estimation procedure indeed required that one had hypothesized correctly all the branching events in the history of the group analyzed, the results of parsimony analysis would never be as stable to the addition of new taxa as they are in practice (shown by decades of phylogenetic analysis; for a nice example of how parsimony is more stable than likelihood to the addition of new taxa, see Siddall and Whiting (1999)).

The type of argument advanced by Goldman (1990) against parsimony seems somewhat specious. By the same line of reasoning that Goldman uses, one could criticize current likelihood models (at least those which sum probabilities for reconstructions, such as Felsenstein's methods) for being totally insensitive to branching events (as admitted by Lewis (2001, p. 916)). The probability of changing from one state to another between any two nodes does not depend at all on the number of branching events in between but depends instead simply on the sum of the lengths of the intervening branches; the path between any two nodes of the tree is simply seen as a smooth continuum (Fig. 5). This is not a side product of the likelihood calculations; it is instead purposefully built into the model. Likelihood, in essence, models phylogeny as a series of populations undergoing genetic drift. Most current models of speciation postulate that speciation events bring about sudden change and/or disruption in the genetic makeup of populations. Therefore, we could argue (following Goldman's line of reasoning) that what is in conflict with established knowledge is current likelihood models, which treat speciation events as virtually nonexistent.

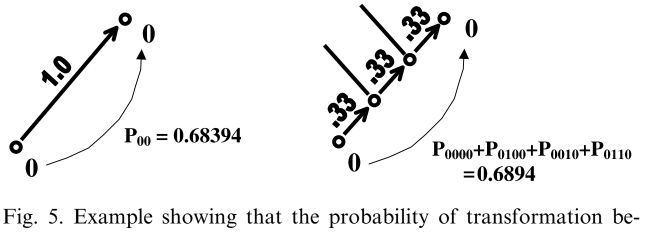

Fig. 5. Example showing that the probability of transformation between any two states, for two nodes in the tree, depends only on the length of the connecting branches and not on the branching events between the two nodes. The segment at the left, of length 1, connects two nodes; the probability of stasis is exactly the same if the segment is divided into three parts and the probabilities of the four alternative pathways from 0 to 0 are considered.

Steel and Penny (2000) also provided a misleading treatment of the assumptions entailed by parsimony. In reference to Tuffley and Steel's (1997) model, they claimed that Ockham's Razor (the principle of parsimony, in a philosophical sense) could be better applied to Felsenstein's formulation, because it is more "parsimonious to assume one common mechanism for all sites rather than 10,000 different mechanisms, one for each site" (Steel and Penny, 2000, p. 843). It would be more proper to say that (cladistic) parsimony does not assume that each site evolves according to the same mechanism; there is an important difference between not requiring existence of a common mechanism and requiring that no common mechanism exists. Likewise, Yang (1996, p. 305) claimed to have "been unable to see any connection between the parsimony method of phylogenetic tree reconstruction and the parsimony or simplicity principle of science and philosophy, or any scientific merit of discussions that claim such a connection." However, the only paper cited by Yang (1996) that attempted to establish such a connection is an early paper of Wiley (1975), which concerned a very abstract discussion not directly related to parsimony; Yang's inability to see the connection stems only from not being aware of relevant literature. Farris (1983) did connect homoplasies with the (philosophical) principle of parsimony in a very specific sense: they both concern disregarding nonconforming observations for the sole purpose of protecting a theory from rejection. Farris even discussed (1983, pp. 23, 24) how the departure from parsimonious arrangements might be justified by covering assumptions (using the common mechanism postulated by Felsenstein as an example of a covering assumption; p. 24). As Farris noted, using the covering assumption could be justified if it is supported by empirical evidence.

Ancestral states

Under the view that ancestral states are parameters, the usual likelihood formulation (which sums over all possible reconstructions but chooses optimal branch lengths) seem logically inconsistent. Barry and Hartigan's (1987a) formulation, or Goldman's (1990) simplification of it to produce parsimony, seems preferable. Likelihoodists object to both on the grounds of the number of incidental parameters estimated.

Goldman (1990) provided a discussion of nuisance parameters in likelihood estimation; he distinguished between incidental and structural nuisance parameters. Structural parameters apply to all of the observations, and thus the precision with which they are estimated increases as more data are added; branch lengths are structural nuisance parameters. Incidental parameters are those for which the precision is not increased as more data are added; likelihoodists have suggested that ancestral states are incidental nuisance parameters. However, it is much less than obvious that the ancestral states are really a parameter. As Goldman (1990) admitted,

> the values [of each state at each ancestral node] are not parameters of the evolutionary process, but random variables: particular realizations of parts of the process. [A possible approach] is to estimate the random variables as though they were parameters of the model. However, in this case they will be incidental parameters: as the amount of data (i.e., the number of characters) increases, the number of parameters also increases. For each additional character (labelled *a*, for instance), there are additional data x_a, consisting of the states for all n species, and additional parameters y_a, consisting of the states for all the internal nodes $\{N_l : \ l = 1, 2, \ldots, m\}$ (Goldman, 1990, p. 350).

Goldman (1990) decided that, even if the ancestral reconstructions are not parameters, they "could be treated as if they were." But they could also be treated (much more properly) as if they were *not* a parameter. The ancestral states are more like a kind of inferred observation (Farris, 1986). Parameters are instead those variables of the process that determine the conditions of the problem—the variables that determine the outcome of evolution, that is. Even if not observed, the ancestral states are (just like observed states) part of that outcome.

Felsenstein (1978) had similarly concluded that the estimation of ancestral states was the cause of the inconsistency for parsimony, but some years later (talking then to two respectable statisticians such as Barry and Hartigan, instead of taxonomists) admitted that "it is not obvious whether [assigning a given state to internal nodes in the tree sequences] amounts to estimating a host of new parameters, one at each site at each internal node of the tree" (Felsenstein, 1987, p. 208). Farris (1986) argued that the inconsistency in determining ancestral states need not determine inconsistency in estimating the tree topology:

> In one of the standard estimation problems the aim is to estimate the mean m of a normally distributed population on the basis of a random sample $x(1), x(2), \ldots, x(n)$ of n independent observations. It is well known that the maximum-likelihood estimator is the sample mean, and that the error of the estimation vanishes as n increases without limit. But any estimate of m also provides estimates of n independent quantities $x(1) - m, x(2) - m, \ldots, x(n) - m$. The sampling error of those several estimates does not vanish as n increases, so that those estimates could hardly be said to be consistent. This plainly does not imply, however, that the mean is not consistently estimated. That some parameters are not consistently estimated, then, does not imply inconsistency of every estimate (Farris, 1986, p. 22).

Farris (1986) also suggested that the behavior of parsimony is determined only by the frequency with which each type of character appears. For four taxa, if partitions (AB)(CD) are more frequent than (AC)(BD) or (BC)(AD), parsimony will choose tree (AB)(CD). Under the hypothetical situation posed by Felsenstein (1978), the frequencies of each type of character are correctly (and consistently) estimated, regardless of the method used to infer the phylogeny. For the four-taxon unrooted case, only eight types of character distributions are possible in the remaining taxa. The likelihood contribution of each type can be calculated beforehand. Once this is done, calculating the likelihood for any set of characters (no matter how numerous), does not involve repeating calculations of ancestral states over and over, but simply involves calculating the products of these basic likelihoods (elevated to their respective exponents). The difference between likelihood and parsimony is only in how the likelihood contribution of each type of character is calculated.

If all the branches in the model tree have the same length (probability of change, that is), then the probability of evolving each possible type of state distribution in the terminal taxa follows; all the information required is already contained in the tree and the length of the branches. Since the model is Markovian, the true probability with which each individual type occurs is determined by summing the probabilities of all possible reconstructions or pathways. Goldman's formulation of parsimony assumes that each character type occurs with a probability equal to the pathway with highest probability, among all the pathways that lead to that character type. If the probability of change in each branch is low, this estimation produces probabilities that are roughly proportional to the actual probabilities (i.e., the ones obtained by summing); that is, all the resulting character types are ranked in the same order of increasing probability by both criteria. This, however, does not convert the calculations under Goldman's model into estimations of a parameter; if a reconstruction was indeed a parameter, there would be one of them which would confer to the corresponding character type its true probability of occurrence under the model, and there is none. Only the sum of all reconstructions provides the true value for a given type.

Farris (1986) was entirely correct that the estimation of ancestral states is not itself the cause of the inconsistency of parsimony, but the inconsistency does come from using just one reconstruction to estimate the probability of each character type: different types are simply expected to occur with the wrong frequencies.[10] The advantage, however, is that then the calculations

[10] A consequence of this is that, for a given starting point at the root of the tree, the sum of the probabilities attributed to all possible types of character distributions in the terminals by their most likely pathways does not sum up to one—as it does when all possible pathways are considered.

can be done much more easily than under a more proper estimation method, and the error introduced is significant only if the probability of change is very high.[11] It is then perfectly justified to consider (with Edwards (1996)) that parsimony is appropriate because it produces results expected to mimic those "of a proper method for that probabilistic model," i.e., those that would be obtained by using the exact, but harder to calculate, probabilities with which each character type occurs under the model. What is more important, perhaps, is that parsimony is also a reasonable estimator given other models (such as the models of Farris (1973)), the model of Tuffley and Steel (1997), and possibly other models).[12]

Considering what happens when the number of character states increases, we can also see that the estimation of ancestral reconstructions is not in itself the cause of the inconsistency of parsimony. It is well known (see Steel and Penny (2000)) that the parsimony estimation then becomes consistent. But according to Goldman (1990) this also implies that the "additional parameters y_a, consisting of the states for all the internal nodes" must now be selected from a much larger number of possibilities. Yet, even if estimating the ancestral states now becomes more complex (and a given most parsimonious reconstruction is less likely to be correct), parsimony becomes consistent. Again, parsimony becoming consistent is not what is expected by considering that ancestral states are incidental parameters but rather what is expected from considering that (under the model used by Felsenstein (1978, 1981)), equating the probability of each character type with the most likely pathway to that type produces much more accurate estimations of the true probability when there are more states.

Note that, since ancestral reconstructions are not a parameter while branch lengths are, the usual likelihood model is not logically inconsistent (as suggested at the beginning of this section) in integrating one but not the other. A likelihoodist might at this point accuse parsimony (under Goldman's derivation) of using approximate calculations of probabilities, instead of the actual ones. However, to calculate the probability of the data (given the model and tree plus branch lengths), current likelihood methods also use some approximations instead of the actual probabilities. Consider for example the assumption that base frequencies remain constant over time. The base frequencies are necessary to determine the probability of transformation between two different states along a given branch. The base frequencies are in practice assumed equal or estimated from the set of terminal taxa (alternatively, they can be chosen so as to maximize the likelihood). The usual likelihood calculations consider all reconstructions for each individual site and then multiply the likelihoods of the individual sites. However, a given reconstruction for a certain site may make some reconstructions for other site more (or less) likely. Consider a four-taxon tree where two sites, 1 and 2, have terminal states ((AG)(AG)); the observed base frequencies are 50:50 for A:G. The likelihoods for all reconstructions for each site are considered independently, but they are not strictly independent, if the base frequencies are to remain constant over time. Suppose that for site 1 the ancestors of the two groups in the tree are assigned state G; in that case, only the reconstruction that assigns (in site 2) state A to both groups is going to preserve the 50:50 ratio for A:G. The reconstruction that assigns (for site 2) state G implies that the base frequencies are A:G = 0:100, not 50:50; it is not an impossible combination of reconstructions, certainly, but if base frequencies are in equilibrium, it is less likely than the double A/G reconstruction (that the ancestors of the two groups in the tree need not be contemporaneous merely compounds the problem). Considering that the two reconstructions are—a priori—equally likely amounts to saying that the base frequencies may change over time, but if so, the substitution probabilities as a function of time cannot be determined. Alternatively, if the base frequencies do not change, considering both sets of reconstructions equivalent is incorrect. The final probability of observing the data calculated with the usual approach, therefore, is not the actual probability under the stipulated model. Taking this into account would require evaluation of combinations of reconstructions for different sites—which is computationally impossible. This is not intended as a criticism of likelihood, but rather as an example showing that, even without violations of the model, the probabilities calculated under usual likelihood methods are only approximations—just as in parsimony.

Simplicity and realism

Once ancestral reconstructions are reconsidered, it is seen that the difference in behavior between parsimony and likelihood stems only from the model used and that the estimation method involving more parameters is likelihood, not parsimony. In parsimony (i.e., Goldman's model), all that determines the fit of a tree to the data is its topology (and the probabilities of stasis and change).

[11] If data are generated from a model tree, and the probabilities of change along each branch are up to three or four times the average amounts of change observed in real DNA data sets with large numbers of taxa, parsimony still recovers easily the model tree.

[12] For example, it seems that a modified two-rate model (where each branch can have different sets of sites in each of the two rate categories, instead of having each site in a fixed category across all branches) would also produce exactly the same results as parsimony, with fewer parameters than Tuffley and Steel's model.

If we decide to integrate branch lengths out of the model (with the site by site approach described above, producing Sober's (1985) formulation), all that determines the fit of a tree to the data is also its topology (and the range of branch lengths considered). Sober's method is as simple as Goldman's but produces different results. In at least some cases, however, Sober's formulation implies differences in the likelihood of trees, which seem hardly justifiable. As shown in Fig. 4, that formulation implies that failure to observe the state in one taxon is "more likely" if the taxon is the sister group to two taxa with identical states. The parsimony method implies that both trees confer exactly the same probability on the observations, and this seems more logical.

In the likelihood methods derived from Felsenstein (1981) the branch lengths are allowed to vary to obtain a higher likelihood. Whether the increase in likelihood, by allowing branch lengths to vary, is considered significant will often be subject to discussion, of course:

> We like explanations which will fit the facts, and we like simple explanations. The question is: how much simplicity are we prepared to lose for a given increase in the excellence of the fit? What increase in support do we require to justify an increase in complexity in the model, say the addition of a new parameter? What, in other words, is the rate of exchange between support and simplicity? I ... offer no specific guidance on the 'rate-of-exchange' problem, but only a general warning to eschew dogmatism (Edwards, 1992, p. 200).

That parsimony (i.e., Goldman's formulation) is an estimation procedure with fewer parameters than likelihood agrees perfectly well with the historical perception of phylogeneticists (e.g., Farris, 1982, 1983) that parsimony is to be preferred on the grounds of simplicity of explanation. If one is willing to introduce more parameters into the estimation, one may allow for differences in branch lengths, which will improve the likelihood. If one is willing to further improve the likelihood, one will eventually arrive at a model where each character can have its own branch length.[13] At this point, one will have come full circle, by necessity arriving at the same conclusion that one had started with, a conclusion that can be defended by recourse either to realism or to simplicity. Parsimony is therefore at both ends of the spectrum from simplicity to realism.

This leads to an additional question that some likelihoodists have posed: is there a method of phylogenetic estimation that, given different models and sets of parameters, is in general the one with the highest probability of recovering the true tree? So far, parsimony has been compared to a very reduced set of quite similar models, but the fact that it is derivable from very different circumstances suggests that it is perhaps justifiable under other types of models also. In being at both ends of the spectrum of complexity, and in being derivable from very different types of models, parsimony does seem to have a unique place among methods of phylogenetic estimation.

Acknowledgments

Parts of the present paper were presented at the XXI Annual Meeting of the Willi Hennig Society, in Finland, 2002. I thank the favorable comments and encouragement from many members of the audience (and organizer Jyrkimuona for the invitation). I also thank M. Allard, J. Farris, N. Giannini, M. Giannini, A. Kluge, D. Lipscomb, M. Ramírez, and C. Szumik for discussions on the subject and/or assistance with literature. I am especially thankful to D. Pol and T. Grant, who shared with me many useful insights and helped shape my views of likelihood. J. Faivovich, T. Grant, R. Page, D. Pol, W. Wheeler, and an anonymous reviewer provided helpful comments on the ms. I thank A. Kluge's editorial zeal and efficiency. The Instituto Superior de Entomología "Dr. Abraham Willink" (Facultad de Ciencias Naturales, Universidad de Tucumán) provided workspace and equipment. The ANPCyT (PICT 01-04347) and A-TOL (NSF) provided financial support.

References

Barry, D., Hartigan, J., 1987a. Statistical analysis of hominoid molecular evolution. Stat. Sci. 2, 191–207.

Barry, D., Hartigan, J., 1987b. Rejoinder [on Statistical analysis of hominoid molecular evolution]. Stat. Sci. 2, 209–210.

Chang, J., 1996a. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134, 189–215.

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

de Queiroz, K., Poe, S., 2001. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writing on corroboration. Syst. Biol. 50, 305–321.

Edwards, A., 1992. Likelihood. Hopkins University Press, Baltimore (expanded edition).

Edwards, A., 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. Syst. Biol. 45, 79–91.

[13] Tuffley and Steel's result, that allowing each site to have its optimal branch length is equivalent to parsimony, also holds when single reconstructions—as in Barry and Hartigan's formulation—are considered. More properly, for a given set of optimal branch lengths under Tuffley and Steel's method, only a single reconstruction is possible, and all the others have zero likelihood, so that both methods become equivalent. There may be different sets of branch lengths that are optimal, of course, but each one determines a single reconstruction. This seems to have caused some confusion in Lewis (2001, p. 917), who apparently considered that the alternative sets of optimal branch lengths could occur only for multistate characters; the alternative sets of branch lengths occur whenever most parsimonious optimization is ambiguous (which can happen also for binary characters).

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

Farris, J., 1982. Simplicity and informativeness in systematics and phylogeny. Syst. Zool. 31, 413–444.

Farris, J., 1983. The logical basis of phylogenetic analysis. In: Platnick, N., Funk, V. (Eds.), Advances in Cladistics, vol. 2: Proceedings of the Second Meeting of the Willi Hennig Society. Columbia University Press, New York, pp. 7–36.

Farris, J., 1986. On the boundaries of phylogenetic systematics. Cladistics 2, 14–27.

Farris, J., 1999. Likelihood and inconsistency. Cladistics 15, 199–204.

Farris, J., 2000. Corroboration versus "strongest evidence". Cladistics 16, 385–393.

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

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

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1982. Numerical methods for inferring evolutionary trees. Q. Rev. Biol. 57, 379–404.

Felsenstein, J., 1987. Comment [on Statistical analysis of hominoid molecular evolution]. Stat. Sci. 2, 208–209.

Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 2, 521–565.

Felsenstein, J., Sober, E., 1987. Parsimony and likelihood: an exchange. Syst. Zool. 35, 617–626.

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

Goloboff, P., 1996. Methods for faster parsimony analysis. Cladistics 12, 199–220.

Goloboff, P., 1998. Tree searches under Sankoff parsimony. Cladistics 14, 229–237.

Goloboff, P., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Hennig, W., 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.

Huelsenbeck, J., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

Kluge, A., 1997. Testability and the refutation of and corroboration of cladistic hypotheses. Cladistics 13, 81–96.

Kluge, A., 2001. Philosophical conjectures and their refutation. Syst. Biol. 50, 322–330.

Larget, B., Simon, D., 1999. Markov chain Monte Carlo algorithms for the bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.

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

Neyman, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, pp. 1–17.

Pol, D., Siddall, M., 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17, 266–281.

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

Rogers, J., Swofford, D., 1998. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. Syst. Biol. 47, 77–89.

Sanderson, M., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817–829.

Siddall, M., 1997. Prior agreement: arbitration or arbitrary? Syst. Biol. 46, 765–769.

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

Siddall, M., 2002. Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. Cladistics 17, 395–399.

Siddall, M., Kluge, A., 1999. Letter to the editor. Cladistics 15, 439–440.

Siddall, M., Whiting, M., 1999. Long branch abstractions. Cladistics 15, 9–24.

Sober, E., 1985. A likelihood justification for parsimony. Cladistics 1, 209–233.

Steel, M., 2002. Some statistical aspects of the maximum parsimony method. In: De Salle, R., Giribet, G., Wheeler, W. (Eds.), Molecular Systematics and Evolution: Theory and Practice. Birkhäuser Verlag, Basel, Switzerland, pp. 125–139.

Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.

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

Steel, M., Szekely, L., Hendy, M., 1994. Reconstructing trees from sequences whose sites evolve at variable rates. J. Comp. Biol. 1, 153–163.

Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular systematics, second ed.. Sinauer, Sunderland, MA.

Swofford, D., Waddell, P., Huelsenbeck, J., Foster, P., Lewis, P., Rogers, J., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50, 525–539.

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

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20, 595–601.

Wiley, E., 1975. Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists.. Syst. Zool. 24, 233–243.

Yang, Z., 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.

Yang, Z., 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42, 294–307.

Yang, Z., 1997. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 414, 105–108.

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

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