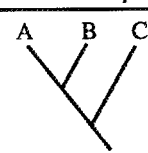
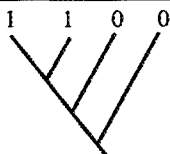
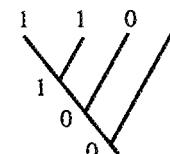


RECONSTRUCTING THE CHARACTER STATES OF ANCESTORS: A LIKELIHOOD PERSPECTIVE ON CLADISTIC PARSIMONY

Introduction

Although the justification for using cladistic parsimony to infer phylogenetic trees has been extensively discussed, much less attention has been paid to the use of cladistic parsimony to reconstruct the character states of the ancestral species postulated by an inferred phylogenetic tree. These two problems differ in terms of both their inputs and their outputs, as shown in the following table. In the former, one begins with data on the character states of extant species and tries to find the best supported phylogenetic tree. In the latter, one begins with the tree that is most firmly supported by characteristics C_1, C_2, \dots, C_{n-1} ; one then takes some *new* characteristic C_n and records the character states of C_n that attach to the tree's interior nodes, which represent common ancestors. In both cases, cladistic parsimony solves the problem by finding the hypothesis that requires the fewest changes in character state that are needed to explain the observations.

Table

	Input	Output
Problem 1: Inferring a tree topology	Data on species A, B, and C	
Problem 2: Inferring character states of ancestors		

Perhaps the reason the second question has been addressed much less often is the belief that the justification for using cladistic parsimony must be the same in both problems, so there is no point in examining the second question on its own. This may be true, but at present it can only be regarded as a conjecture, given how fragmentary our understanding of the two inference problems is. Another reason the first problem has received more attention is the principle of *first things first*. The first problem is more fundamental—in the first, one infers a tree; in the second, one uses an inferred tree to solve a further problem.² In the present paper, I propose to address the less fundamental problem. The results, I think, are interesting.

To discuss the justification for using cladistic parsimony in either inference problem, one must consider what the use of that methodology presupposes about the evolutionary process. Parsimony is fundamentally a comparative principle. It doesn't tell you whether to accept or reject a given hypothesis; rather, it says whether one hypothesis is better supported than another. For this reason, the basic question about parsimony's justification concerns what must be true of the evolutionary process for the following biconditional to be correct:

(OE) For any hypotheses H_1 and H_2 and any data set D , H_1 is more parsimonious than H_2 (relative to D) if and only if H_1 is better supported by D than H_2 is.

(OE) says that a parsimony ordering of a set of hypotheses and a support ordering of that set will be *ordinally equivalent*. It is hard to know how to evaluate (OE) until some clarification is provided of what "support" means. One possibility that has been of interest is that support should be understood in terms of the technical concept of *likelihood*. The likelihood of a hypothesis H , relative to the observations O , is the probability $\Pr(O | H)$ that H confers on O . Don't confuse the likelihood of H with the probability $\Pr(H | O)$ that O confers on H . The *likelihood principle* (Edwards 1971, Royall 1997) provides one way to make (OE) more precise:

(LP) For any hypotheses H_1 and H_2 and any data set D , D supports H_1 more than D supports H_2 if and only if $\Pr(D | H_1) > \Pr(D | H_2)$.

If we accept the likelihood principle, the question of whether we should use parsimony to interpret the evidence reduces to the question of whether more parsimonious explanations are more likely.

This latter question can be answered only if we are prepared to describe a model of the evolutionary process on which we will base our likelihood assessments. For example, if we are considering the first problem mentioned above—that of inferring a phylogenetic tree—we must recognize that a tree topology does not, by itself, confer a probability on the character states of tip species. One must say, in addition, what the probabilities are of different types of change and stasis in the tree's interior. This point does not change if we merely ask which of two topologies confers a higher probability on the observations. Whether we want to know the point value of a topology's likelihood, or, more modestly, whether one topology is more likely than another, the relevant principle is this: *no model, no inference*.

If there were a model of the evolutionary process that we could accept, we could use that model to determine whether parsimony and likelihood must coincide. By 'model' I don't mean some vague statement like "natural selection has been important." Rather, I mean a specification of the quantitative rules of change that govern different traits in different branches of a phylogenetic tree. Unfortunately, biologists who don't already know what the true topology is for a group of taxa are also unlikely to know which process model they should accept.

What is actually known about the relationship of parsimony and likelihood in the context of the first problem mentioned above—that of evaluating tree topologies in the light of data on tip taxa? Penny *et al.* (1994) and Tuffley and Steel (1997) identified a specific model of the evolutionary process that suffices to render parsimony and likelihood ordinally equivalent.³ One important feature of this model is that it says that each character obeys a symmetrical rule of evolution; if the character has just two states, this means that the probability of the character's evolving from state *i* to state *j* is the same as its probability of evolving from *j* to *i*. This means that natural selection and other directional forces are entirely absent; the process described is one of pure drift. Does this mean that cladistic parsimony makes sense only in connection with data on characters that evolve by drift?

The answer to this question is NO. The fact that a set of assumptions *suffices* for parsimony and likelihood to go hand-in-hand does not show that those assumptions are *necessary* for that relation to obtain (Sober 1988). This, I think, is the most important *caveat* to bear in mind when considering questions about the justification of cladistic parsimony. The fact that an *investigator* makes an assumption in analyzing the logic of

parsimony arguments does not mean that *parsimony* depends for its validity on that assumption.

Rather than specifying a model that suffices for parsimony and likelihood to *agree*, the investigator might try to construct a model that suffices for them to *disagree*. If a model ensures that the most parsimonious explanation of the data will *not* be the one with maximum likelihood, what should we conclude? If we accept the likelihood principle (LP), we face a choice. Either we must reject the process model, or we must reject the use of cladistic parsimony. That is, the use of parsimony presupposes that the specified process model is false. In the search for parsimony's presuppositions, it is models that make parsimony *fail* that are significant, not models that make it *succeed*.

Something like this strategy is the one pursued by Felsenstein (1978). Felsenstein is not mainly concerned in that paper to show that parsimony and likelihood disagree, although he does briefly address that question at the end. His main objective is to describe an example in which parsimony is *statistically inconsistent*—the accumulation of more and more data guarantees that parsimony will converge on a false topology. What is true in this example is (roughly) that there is a model *M* of the evolutionary process that has the following property: as the data set is made large, it becomes increasingly certain that *T* will be less parsimonious than *T'* even though $\Pr(\text{Data} \mid T \ \& \ M) > \Pr(\text{Data} \mid T' \ \& \ M)$, where *T* is the true topology and *T'* is a false topology. In the limit of large data, parsimony converges on *T'* while likelihood converges on *T*; parsimony and likelihood disagree about which topology is best, and likelihood gets the answer right (provided that it uses the right process model *M*). This suggests that parsimony assumes that the process model *M* is false.

One feature of Felsenstein's model is that it assumes that branches differ dramatically in their transition probabilities. On one branch, the probability that a character will evolve from state 0 to state 1 is high, but on another branch that has the same duration, the probability is low. Does it follow that parsimony assumes that branches cannot differ in their transition probabilities? Again, the answer is NO. Felsenstein's model *M* includes several assumptions. For example, he also assumes that all characters on the same branch evolve according to exactly the same rules. What is true is that the *conjunction* of these process assumptions suffices for parsimony and likelihood to part ways. If we accept the likelihood

principle (LP), then the use of parsimony in the problem that Felsenstein addresses presupposes that this conjunction must be false. However, we should not lose sight of the simple logical point that a conjunction can be false without each of its conjunct's being false. Not all of Felsenstein's process assumptions can be true, if parsimony and likelihood are to coincide, but this doesn't mean that all must be false. So it does not follow from Felsenstein's example that parsimony assumes that simultaneous branches must follow the same rules of evolution.

In summary, the problem of justifying parsimony in likelihood terms and the problem of using likelihood to uncover the presuppositions of parsimony can be clarified by considering the following two conditionals:

- If X, then parsimony and likelihood are ordinally equivalent.
- If parsimony and likelihood are ordinally equivalent, then Y.

Two comments are relevant to the first conditional. First, a *plausible* process model X that makes the first conditional true will provide parsimony with a likelihood *justification*. Second, an *implausible* model X that makes the first conditional true does nothing to undermine parsimony. Moving on to the second conditional, we can say that any process claim Y that makes this conditional true is a presupposition of parsimony (assuming that the likelihood principle is correct). If Y is *implausible*, this second conditional provides a likelihood *criticism* of parsimony.

The *presuppositions* of parsimony, then, are the conditions (Y) that are *necessary* for parsimony and likelihood to coincide. The fact that condition (X) *suffices* for this relation to obtain does not show that parsimony assumes that X is true. Still, these two considerations, of necessity and sufficiency, come together in an interesting way: *If X suffices for parsimony and likelihood to be ordinally equivalent, then Z is not a presupposition of parsimony, if X does not entail Z*. If X entails ordinal equivalence, X must entail whatever ordinal equivalence entails. This logical point means that the result of Penny *et al.* (1994) and Tuffley and Steel (1997) has great significance. The model they construct does *not* entail that change is rare. Hence it is *false* that parsimony assumes that change is rare. The sufficient condition that these investigators have identified provides a litmus test for claims about parsimony's presuppositions, but the test is one-sided. If their sufficient condition *does not* entail Z, then Z is not an assumption of

parsimony; but if their sufficient condition *does* entail Z , Z may or may not be an assumption.

I won't pursue the problem of how likelihood and parsimony are related as methods for inferring tree topologies any further, but will now turn to the main topic of this paper. How are parsimony and likelihood related as methods for inferring the character states of ancestors in a tree that one already has reason to accept? The bare bones of this problem are depicted in Figure 1. Species C is the common ancestor of A and B ; these descendants are observed to be in character states α and β , respectively. Given these observations, which assignment of character state to C is most parsimonious, and which has the highest likelihood? We want to see when parsimony and likelihood agree and when they disagree, both for dichotomous and for quantitative characters. As before, the issue is ordinal equivalence.

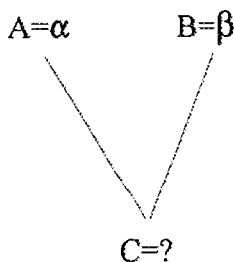


Figure 1

I should mention that this skeleton problem—two descendants and one ancestor—does not capture the full scope of the problem of assigning character states to ancestors. Polytomies aside, the general problem concerns a fully bifurcating tree that has n terminal taxa—which character states should one assign to its $(n-1)$ interior nodes? I won't try to address the general problem here, although I hope it is intuitive that the simple problem I'll address is a "building block" in the more general setting. Still, one has to be cautious. For example, consider our two-descendant problem when the character comes in two states, 0 and 1. Parsimony favors setting $C = 1$, if $A = 1$ and $B = 1$. However, it isn't always true in the wider setting of a tree with n terminal taxa that this is the assignment that is globally most parsimonious. For consider the tree depicted in

Figure 2. In this tree, A and B are the two terminal taxa in state 1; the (globally) most parsimonious assignment of character state to their most recent common ancestor is 0. Thus, our simple two-descendant/one-ancestor problem may be a “building block” for the more general problem, but there are subtleties that arise when one moves from the simple problem to problems that are more complex.

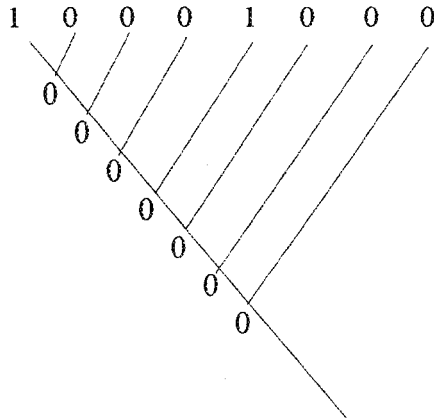


Figure 2

Dichotomous Characters

There are two problems to consider. When $A = 1$ and $B = 1$, the most parsimonious assignment to C is $C = 1$. And when $A = 1$ and $B = 0$, the two assignments to C , $C = 0$ and $C = 1$, are equally parsimonious. The question is—what must be true of the processes in the lineages leading to A and to B , if $C = 1$ is to be the maximum likelihood assignment in the first problem, and what must be true for the two assignments to C in the second problem to have equal likelihoods? These problems, it turns out, have different solutions.

Before we get to these solutions, however, I have to say how I'll understand the processes of character change and stasis that may occur in a lineage. The slogan *no model, no inference* applies to the present problem of inferring ancestral character states no less than it applies to the problem of inferring phylogenetic trees. And in conformity with my previous admonitions to not confuse the assumptions of an *investigator* with the

assumptions of a *method*, let me emphasize that the following model is something I will use to frame the problem: I do not claim that *parsimony* presupposes that this model is correct.

What I'll use here comes from the standard theory of stochastic processes (Parzen 1962). We divide a lineage into a large number of brief temporal intervals. In each, there is a probability u that the lineage will change from state 0 to state 1; and there is a (possibly different) probability v that the lineage will change from state 1 to state 0. Each of these instantaneous probabilities are assumed to be small (at least less than 0.5). They allow us to describe the probability that a lineage will end in state j , given that it starts in state i ($i, j = 0, 1$), when the lineage is N units of time in duration.⁴ I will use the notation $\text{Pr}_N(i \rightarrow j)$ to represent these *lineage transition probabilities*:

$$\text{Pr}_N(0 \rightarrow 1) = u/(u + v) - [u/(u + v)](1 - u - v)^N$$

$$\text{Pr}_N(1 \rightarrow 1) = u/(u + v) + [v/(u + v)](1 - u - v)^N$$

$$\text{Pr}_N(1 \rightarrow 0) = v/(u + v) - [v/(u + v)](1 - u - v)^N$$

$$\text{Pr}_N(0 \rightarrow 0) = v/(u + v) + [u/(u + v)](1 - u - v)^N$$

There is no assumption here as to whether $u = v$. If $u = v$, the lineage undergoes an unbiased process of drift. If $u > v$, there is a directionality or bias in the evolutionary process, favoring state 1 over state 0. One possible source of this bias is natural selection; however, mutation asymmetry and migration also can induce a directional bias.

When N is very small, the two probabilities of stasis $\text{Pr}_N(i \rightarrow i)$ are close to unity and the two probabilities of change $\text{Pr}_N(i \rightarrow j)$ are close to 0. When N is infinite $\text{Pr}_N(i \rightarrow j) = \text{Pr}_N(j \rightarrow j)$; the lineage has the same probability of ending in state j , regardless of what the state was in which the lineage began. Thus, when lineages have very short duration, their initial conditions virtually determine their final state and the relationship of u and v doesn't matter; when lineages are very old, it is the processes that occur during the lineage's duration (represented by u and v) that matter; the initial condition is forgotten.

In summary, the model I am using makes no assumption about drift versus selection, and it also takes no stand on whether the lineage is young or old. It also is neutral on whether the process in the lineage leading from common ancestor C to descendant A is the same as the process in the

lineage leading from C to B. Each lineage has its own values for u and v . This model, I think, is pretty permissive.

It is a property of this model that a certain *backwards inequality* (Sober 1988) obtains: $\Pr_N(j \rightarrow j) > \Pr_N(i \rightarrow j)$ if N is finite; an equality obtains when N is infinite. This inequality holds for all values of u , v , and (finite) N . The backwards inequality does not say that stasis is more probable than change; don't confuse the backwards inequality with the *forwards inequality* $\Pr_N(j \rightarrow j) > \Pr_N(j \rightarrow i)$. An instance of this forwards inequality [e.g., that $\Pr_N(1 \rightarrow 1) > \Pr_N(1 \rightarrow 0)$] will be true for some values of u , v , and N , but not for others. The backwards inequality says that if a descendant is in state j , that outcome is made more probable by the hypothesis that its ancestor was in state j than by the hypothesis that the ancestor was in state i . The backwards inequality provides a solution to the first of our problems about dichotomous characters:

Theorem 1: If $A = i$ and $B = i$ ($i = 0, 1$), then $C = i$ is the assignment of maximum likelihood.⁵

As noted before, $\Pr(i \rightarrow j)$ and $\Pr(j \rightarrow j)$ get closer together, the more time there is in a lineage. This means that when $A = 1$ and $B = 1$, the likelihoods of $C = 1$ and $C = 0$ get closer together, the more ancient their most recent common ancestor is. $C = 1$ is always more likely, but the degree of its superiority depends on time. This quantitative effect is not reflected in the parsimony analysis, which registers only the qualitative point that if $A = 1$ and $B = 1$, then $C = 1$ is more parsimonious than $C = 0$, no matter how long ago their most recent common ancestor existed.⁶

Matters become more complicated when we move to our second problem. When $A = 1$ and $B = 0$, the two possible assignments of character state to C are equally parsimonious. When will these two assignments have the same likelihood? That is, when will it be true that $\Pr_A(0 \rightarrow 1)\Pr_B(0 \rightarrow 0) = \Pr_A(1 \rightarrow 1)\Pr_B(1 \rightarrow 0)$? The subscripts A and B represent which of the two lineages the transition probability describes. It is helpful to rewrite this equality as

$$\frac{\Pr_A(0 \rightarrow 1)}{\Pr_A(1 \rightarrow 1)} = \frac{\Pr_B(1 \rightarrow 0)}{\Pr_B(0 \rightarrow 0)} .$$

Two sufficient conditions for this equality can be stated:

Theorem 2a: If $A = 1$ and $B = 0$, then $C = 1$ and $C = 0$ are equally likely if the two lineages are infinitely old or if $u_A = v_B$ and $v_A = u_B$.

The second disjunct means that if one lineage experiences a process that is biased in one direction, the other must experience a process whose bias is equal and opposite. For example, if the lineage leading to A experiences a selection process that favors trait 1 over trait 0 by a certain degree (because $u_A > v_A$), then the lineage leading to B must have the same quantitative bias favoring trait 0 over trait 1 (because $u_B < v_B$).

Another sufficient condition for likelihood and parsimony to coincide can be obtained if we assume that the process at work in the lineage leading to A is the same as the process in the lineage leading to B:

Theorem 2b: If $A = 1$ and $B = 0$ and the lineages are characterized by the same pair of values for u and v , then $C = 1$ and $C = 0$ are equally likely if and only if the lineages are infinitely old or $u = v$.

What one can't have, if likelihood and parsimony are to agree, is a single selection process (or any other biased process in which $u \neq v$) that occurs in both lineages, where those lineages have finite duration.

Quantitative Characters

Suppose that the descendant species A and B are each scored for some quantitative character, with the result that $A = 40$ and $B = 40$. The most parsimonious estimate of the state of their common ancestor C is $C = 40$. What must be true of the evolutionary processes in the two lineages for $C = 40$ to be the assignment of maximum likelihood?

To answer this question, we must provide a model of the evolutionary process, analogous to the one used in the previous section, but suitable for describing quantitative characters. Let's begin by setting limits on the values of the character in question; suppose it can't go below zero or above 100. We can think of u as the probability of the lineage's increasing its character state by a very small amount during a brief interval of

time, and v as the probability of the lineage's reducing its value during that instant. However, in contrast to what is true in the case of dichotomous characters, it is obvious that u and v cannot remain constant over the full range of the lineage's possible states; for example, u must have a value of zero when the lineage is in state 100, though of course it can have a nonzero value when the lineage has a value less than 100. In addition, consider the possibility that the lineage is evolving towards a stable equilibrium; perhaps a trait value of 75 is optimal, and selection is pushing the lineage towards that value. This means, first of all, that $u > v$ when the lineage's trait value is less than 75, but that $u < v$ when the population has a value greater than 75. In addition, the degree to which $u > v$ must decline as the population approaches 75 from below. Thus we need to think of the stochastic model for a lineage as having different values of u and v attaching to the different character states the lineage might occupy.

Despite these complications, the theory for continuous characters is similar to the theory for dichotomous characters. For example, suppose a biased process (like natural selection) is pushing a lineage towards a single attractor state—for example, a value of 75. Then the lineage's probability of reaching that equilibrium is greater, the closer its initial state is to 75. This is an analog of the backwards inequality for continuous characters.⁷ Similarly, this equilibrium has a higher probability of being attained, the more time there is in a lineage. When the lineage has a very short duration, stasis is almost certain; as the lineage is given a longer duration, the evolutionary process takes over and the initial condition recedes in its impact on the lineage's final state. In the limit of infinite time, the initial condition is entirely forgotten and the lineage's probability of attaining a given end state is the same, regardless of what the state was in which the lineage began.

How should we conceptualize a pure drift process for continuous characters? With very little time, the expected value of the end state is tightly peaked around the lineage's initial condition. As time goes on, this low variance bell curve is squashed down. With infinite time, there is a flat distribution—each character state has the same probability.

As was true in the model for dichotomous characters, the present model is not particularly restrictive. The processes at work may be biased or unbiased; time may be short or long; and the two lineages may evolve according to the same rules, or according to different rules.

Let us now consider our first problem: if $A = 40$ and $B = 40$, what character state should we assign to the common ancestor C ? The most parsimonious assignment is $C = 40$. Under what conditions is this also the assignment of maximum likelihood? Since the backwards inequality for dichotomous characters sufficed to entail Theorem 1, one might expect the model for continuous characters to have the unconditional consequence that $C = 40$ has maximum likelihood. This is not correct. For example, if directional selection is pushing both lineages towards the attractor value of 50, then the maximum likelihood assignment to C will be *less* than 40; how much less than 40 the maximum likelihood value is depends on how long the lineage has been evolving and on how strong the directional force is. Nonetheless, two simple sufficient conditions can be specified for when $C = 40$ is the assignment of maximum likelihood:

Theorem 3a: If $A = 40$ and $B = 40$ and the same evolutionary process is at work in the two lineages, then $C = 40$ is the maximum likelihood value if and only if the process is one of pure drift or 40 is the single attractor state towards which the directional process at work in the two lineages is pushing.

If we drop the assumption of "lineage homogeneity," a further sufficient condition can be specified:

Theorem 3b: If $A = 40$ and $B = 40$, then $C = 40$ is the maximum likelihood value if the one lineage has a directional force pushing it towards an attractor whose trait value is $40 + x$, while the other lineage has a directional force of equal strength pushing it towards an attractor whose trait value is $40 - x$.

For example, if selection favors a trait value of 50 in the lineage leading to A and a value of 30 in the lineage leading to B , then (assuming that the two forces are of equal magnitude) the most likely state for C is $C = 40$.

The next problem to consider arises when A and B are observed to have different values. For example, if $A = 30$ and $B = 50$, what value should we assign to C ? In the previous questions we have investigated, it was obvious what parsimony recommended. The present problem is a bit different. It is true that assigning $C = 40$ is one way to minimize the *total*

amount of change that took place in the two lineages, but the same is true for any other assignment of value to C , so long as it is between 30 and 50. The idea of cladistic parsimony is generally taken to recommend the assignment of $C = 40$, but why should this halfway point be regarded as the best estimate? One answer to consider is that it minimizes the amount of *squared* change; $10^2 + 10^2 = 200$, while $(10 - x)^2 + (10 + x)^2 = 200 + 2x^2$. However, the question then needs to be faced of why parsimony should minimize *squared* change rather than minimize *total* change. It is hard to see how this question can be answered without considering a process model and its probabilistic properties.⁸

As one might expect, the maximum likelihood assignment to C isn't always $C = 40$, but it is under certain process assumptions:

Theorem 4a: If $A = 30$ and $B = 50$ and the same evolutionary process is at work in the two lineages, then $C = 40$ is the maximum likelihood value if and only if the process is one of pure drift or 40 is the single attractor towards which the directional process at work in the two lineages is pushing.

As before, if we drop the assumption of lineage homogeneity, a further sufficient condition can be specified:

Theorem 4b: If $A = 30$ and $B = 50$, then $C = 40$ is the maximum likelihood value if the lineage leading to A has a directional force pushing it towards an attractor whose trait value is $30 - x$ while the B lineage has a directional force of equal strength pushing it towards an attractor whose trait value is $50 + x$.

Notice that Theorems 4a and 4b leave room for the fact that $C = 40$ may not be the maximum likelihood solution. For example, if both lineages experience a strong directional force pushing them towards a global attractor of 60, the maximum likelihood value of C will be less than 40; indeed, if the force is strong enough, it may be less than 30.⁹

Testing Adaptive Hypotheses

Hypotheses about adaptation, and about other evolutionary processes, make claims about the causal processes at work in lineages. This suggests

that testing these hypotheses requires one first to ascertain which changes have occurred in lineages, and then see whether they are of the sort predicted by the adaptive hypothesis. For example, if the adaptive hypothesis says that character state 1 is selectively advantageous over character state 0 in some clade, one wants to know whether changes from 0 to 1 have occurred more frequently than changes in the opposite direction. The same question, differently formulated, is whether lineages maintain the 1 state more often than they maintain the 0 state.

The problem is that we do not observe changes in lineages. Rather, what we observe, in the first instance, are the character states of extant species. Once we infer a tree topology for these species, we can say that what we observe are the character states of tip species. The point is that we do not observe the events that occur in the tree's interior. How, then, are we to test hypotheses about the processes that occurred in that interior?

A natural suggestion is to use cladistic parsimony to reconstruct the character states of the ancestors in the tree. These inferred character states then have implications about the pattern of stasis and change that occurred in the tree's interior; one then can ask whether these events confirm or disconfirm the adaptive hypothesis. If cladistic parsimony as a method for reconstructing ancestral character states made no assumptions about the evolutionary process, this procedure would be unobjectionable. But the take-home message of our previous discussion is that this is not the case.¹⁰

As an example of the present problem, suppose we know that the *X* values of the species we are considering evolved by a random drift process. The question is whether species change their *Y* values as adaptive responses to those *X* values. The adaptive hypothesis we wish to test specifies an optimality line, which represents the optimal *Y* value for each *X* value; the hypothesis asserts that once species have their *X* values shifted at random, they then evolve in the direction of the optimality line. Notice that the adaptive hypothesis does not assert that extant species are optimal or even that they are close to optimal. Rather, the modest claim is that natural selection has acted to increase the adaptive fit of *Y* values to *X* values.

We observe the *X* and *Y* values of species A and B. What *X* and *Y* values should we assign to their most recent common ancestor C? Once we have made this assignment, we can assess whether the two lineages have moved in the direction of the optimality line. Figure 3 represents three possible assignments of *X* and *Y* values to the common ancestor.

They agree that the common ancestor should be assigned an X value that is intermediate between the observed X values of A and B. This is the assignment of maximum parsimony, and it also is the assignment of maximum likelihood, given that the X values evolved by random drift (Theorem 4a).

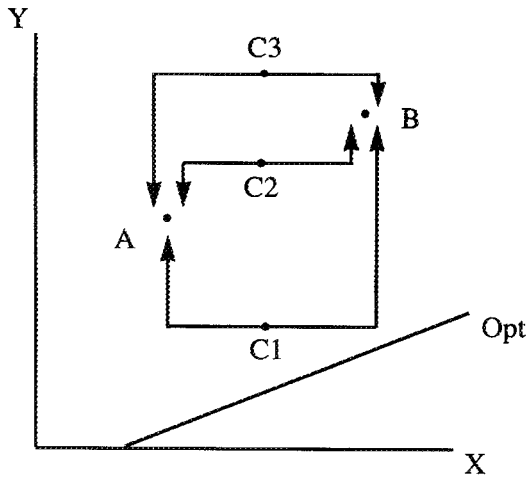


Figure 3

Which Y value should we assign to the common ancestor? Figure 3 represents three possible choices—C1, C2, and C3. If C1 is true, both lineages evolved away from the optimality line. If C2 is correct, one line evolved towards the optimality line and the other away. If C3 is right, both lineages evolved towards the optimality line. These three possibilities correspond to three different verdicts concerning the adaptive hypothesis. If C1 is correct, we have two pieces of evidence against the adaptive hypothesis. If C2 is correct, we have one pro and one con. And if C3 is right, we have two pieces of evidence that confirm the adaptive hypothesis. C2 happens to be the most parsimonious assignment. But why does this show that C2 is the best estimate? After all, if the adaptive hypothesis is true, C3 is the most likely assignment of the three, whereas if Y evolved by random drift C2 would be the likeliest estimate. Thus the maximum likelihood estimate of the common ancestor's Y-value cannot be ascertained independently of saying whether the adaptive hypothesis is correct. This

illustrates the point that parsimony cannot be regarded as a neutral and pre-suppositionless device for inferring ancestral character states (Cunningham 1999).

The situation would be different, if A and B were sitting on the optimality line. As Figure 4 illustrates, the lineage leading to A and the lineage leading to B both move in the direction of the optimality line, regardless of whether C1, C2, or C3 is taken to be the best estimate of the common ancestor's character states. It happens that C2 is the most parsimonious assignment. But from the point of view of testing the adaptive hypothesis, it doesn't matter whether one uses this or one of the other assignments.

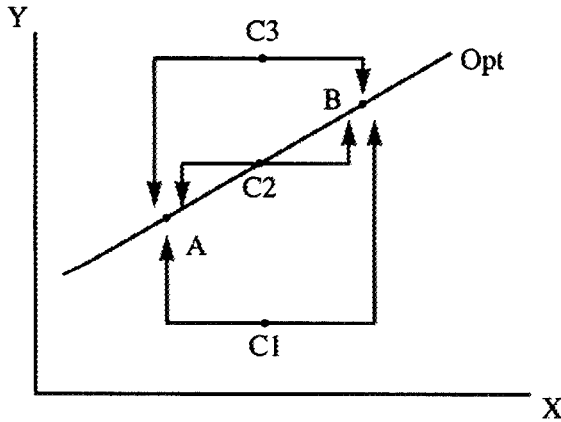


Figure 4

We have arrived at a dilemma. Whether parsimony and likelihood agree about which character states should be assigned to ancestors depends on which evolutionary processes are under way. For those who regard parsimony as a first principle that requires no justification, this result does not matter. But for those who think that parsimony is justified only to the extent that it coincides with likelihood, it must. From a likelihood point of view, we can't test adaptive hypotheses by using parsimony to reconstruct the character states of ancestors. The parsimonious reconstruction embodies assumptions about the evolutionary process, and so it cannot be viewed as a neutral vehicle for testing hypotheses about that process. What would be

highly desirable is a procedure for testing adaptive hypotheses that does not require the estimation of ancestral character states; that, however, is a topic for another occasion (Orzack and Sober 2001).

Elliott Sober

*University of Wisconsin at Madison
London School of Economics and Political Science*

NOTES

1. My thanks to Thomas Hansen and Mike Steel for useful comments and to the National Science Foundation (SES—9906997) for financial support.

2. If using parsimony to choose a tree topology required one to infer the character states of the ancestors postulated by that tree, the two problems would not be separable, in that the first would subsume the second. Although the use of parsimony to choose a tree is sometimes described in this way (see, for example Lewis, 1998, p. 138), this is a mistake. Finding a tree topology involves calculating what the minimum number of changes would be if the tree were correct; it is not required that one think that this *minimum* is the *actual* number of changes. This point is relevant to discussions of the circumstances under which parsimony will be statistically consistent. It is sometimes claimed that parsimony (as a method for inferring a tree topology) can be expected to be inconsistent because the number of parameters one must estimate grows with the number of characters in one's data set, since for each new character, one has to infer the state of that character that each ancestor has.

The mistake is analogous to the following. It is known that maximum likelihood estimation is statistically consistent in the context of estimating the mean in a normal population. The sample mean is the maximum likelihood estimate of the population mean, and the sample mean approaches the population mean as the sample is increased in size. But suppose someone doubted this claim of convergence on the ground that each new observation requires a new computation—each time you sample a new individual you have to recompute the sample mean. The point about computation is correct, but this does not mean that the parameters one is estimating grow in number; there is just one of them—the population mean.

3. As far as I know, this is the only model yet discovered that renders parsimony and likelihood ordinarily equivalent as methods for inferring tree topologies. Goldman (1990) describes a model that renders parsimony and likelihood equivalent, but for a slightly different problem. The hypotheses he evaluates are tree topologies with character states specified for all interior nodes. For discussion of the difference between these two problems, see Sober (1988, pp. 150–66) and Steel and Penny (2000).

4. This is the model used in Sober (1988) and in Pagel (1994).

5. Pagel (1999) uses a method for deriving maximum likelihood estimates of ancestral character states that conflicts with this result. To evaluate the likelihoods of $C = 0$ and $C = 1$, given the observation that $A = 1$ and $B = 1$, Pagel compares the likelihoods of two con-

junctive hypotheses: ($C = 0$ and X_0) and ($C = 1$ and X_1), where X_i specifies the values of the branch transition probabilities that maximize the likelihood of $C = i$. X_0 says that $\text{Pr}(0 \rightarrow 1) = 1$ and X_1 asserts that $\text{Pr}(1 \rightarrow 1) = 1$. Since the two conjunctions have the same likelihood ($= 1.0$), Pagel's procedure concludes that the data do not discriminate between the two hypotheses about the ancestor's character state. His analysis would be the same if there were twenty descendants, and not just two, all in the same state.

Pagel's analysis and my own differ because I am not estimating branch transition probabilities; rather, I am identifying the range of values in parameter space that makes $C = 1$ the assignment of maximum likelihood. Assuming that the two branches follow the same rules of evolution, we can locate Pagel's analysis and my own in the unit square depicted in Figure 5. Given the process model I am using (which entails the backwards inequality), only the lower-right half of this square is possible. Every point in that region has the effect of making $C = 1$ more likely than $C = 0$. There is no need to say which of these points is most likely.

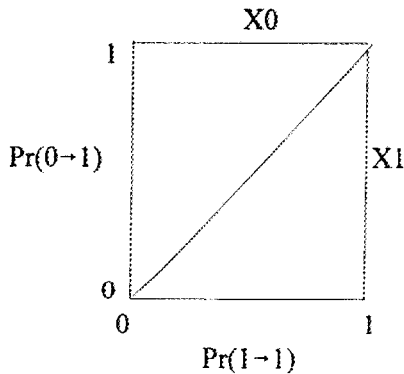


Figure 5

Rather than surveying this triangular region, Pagel's procedure is to focus on a point and a line that it contains. There is a point where $\text{Pr}(0 \rightarrow 1) = 1$; and there is a line where $\text{Pr}(1 \rightarrow 1) = 1$. This leads Pagel's procedure to conclude that $C = 0$ and $C = 1$ are equally likely. (Incidentally, if one assumed that $u = v$, and took account of the fact that N is finite, Pagel's procedure and my own would deliver the same verdict.)

Schluter *et al.* (1997), like Pagel, estimate ancestral character states by evaluating conjunctions of the sort just described, though they endorse a procedure that differs from Pagel's. Both use a "best case expedient" for dealing with nuisance parameters; for criticisms, see Sober (1988, pp. 150–65) and Schultz and Churchill (1999, p. 652).

6. This effect of time on the likelihood reconstruction of ancestral character states is reflected in the simulations of Martins (1999).

7. However, it is possible that the lineage has a higher probability of reaching 75 if it starts at 65 than if it starts at 80; suppose there is a strong directional force pushing the lineage towards a value of 90.

8. Maddison (1991) showed that squared change is the appropriate criterion under a Brownian motion-drift model.

9. These theorems about quantitative characters agree with the simulations performed by Martins (1999), who found that squared parsimony and likelihood both do well (and agree with each other) when there is a Brownian motion process, but tend to diverge under evolution towards a stable attractor, as in an Ornstein-Uhlenbeck process.

10. Ridley (1983) recommends the use of cladistic parsimony to reconstruct character states of ancestors; he further recommends that the implied changes in character state that occur in the tree's interior be regarded as the *only* evidence that allows one to test an adaptive hypothesis; the fact that character states are sometimes retained in a lineage should be viewed as evidentially meaningless. Ridley says that this methodology biases the case against the adaptive hypothesis; he defends it on the ground that if an adaptive hypothesis can receive confirmation even when the deck is stacked against it, that this is strong evidence indeed in favor of that hypothesis.

In fact, Ridley's procedure does not *always* have the effect of biasing the case against the adaptive hypothesis. Consider, for example, the adaptive hypothesis that says that tooth shape (sharp or flat) is an adaptive response to dietary regime (carnivore or herbivore). I take it that this hypothesis predicts that changes from CF to CS should occur more often than changes in the opposite direction, and that changes from HS to HF should occur more often than changes in the opposite direction. Now consider the tree depicted in Figure 6. The most parsimonious assignment of character states to ancestors is 1 = CS, 2 = CF, and 3 = CF. This has the consequence that there was one change from CF to CS and no changes in the opposite direction, a result that conforms to the predictions of the adaptive hypothesis. However, consider the unparsimonious assignment 1 = CS, 2 = CS, and 3 = CS. This entails that there were two changes from CS to CF and no changes in the opposite direction, a result that goes contrary to what the adaptive hypothesis predicts. Thus, it is false that Ridley's procedure always introduces a bias against the adaptive hypothesis.

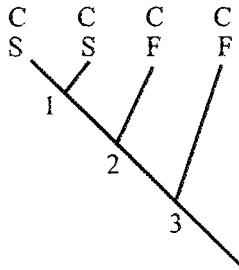


Figure 6

Similar problems attach to Ridley's decision to regard stasis as having no evidential bearing on the adaptive hypothesis. He claims that this biases the case against the adaptive hypothesis, but this, too, is not always the case. Consider, for example, the tree in Figure 7. Ridley's procedure of first using parsimony to reconstruct ancestral character states and then counting changes as the sole bearers of evidential meaning has the result that one regards this clade as evidence for the adaptive hypothesis—after all, parsimony says that there was one change from CF to CS and none in the opposite direction, and changes in

character state are the only events that count. However, the question may be asked as to why CF was retained by so many lineages in the tree. Why isn't this evidence *against* the adaptive hypothesis (on which see Hansen 1997 and Orzack and Sober 2001)? Ridley's decision to ignore stasis, in this instance, helps the adaptive hypothesis.

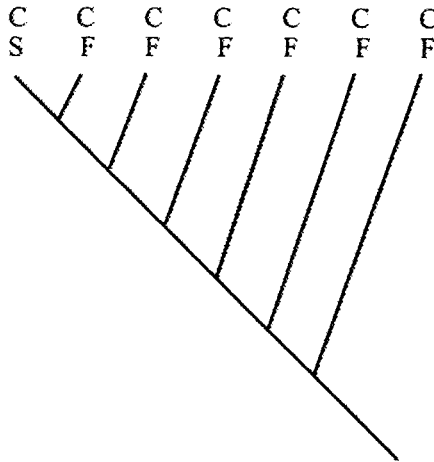


Figure 7

My conclusion here is not that Ridley's two-part procedure is incorrect, but just that it cannot be justified in the way he suggests.

REFERENCES

- Cunningham, C. (1999): "Some Limitations of Ancestral Character-State Reconstruction When Testing Evolutionary Hypotheses." *Systematic Biology* 48: 665-74.
- Edwards, A. (1971): *Likelihood*. Cambridge: Cambridge University Press.
- J. Felsenstein, "Cases in Which Parsimony and Compatibility Methods Can Be Positively Misleading." *Systematic Zoology* 1978, 27: 401-10.
- Goldman, N. (1990): "Maximum Likelihood Inference of Phylogenetic Trees, with Special Reference to a Poisson Process Model of DNA Substitution and to Parsimony Analyses." *Systematic Zoology* 39: 345-61.
- Hansen, T., 1997. "Stabilizing Selection and the Comparative Analysis of Adaptation." *Evolution* 51: 1341-51.
- Lewis, P.(1998): "Maximum Likelihood as an Alternative to Parsimony for Inferring Phylogeny using Nucleotide Sequence Data." In D. Soltis, P. Soltis, and J. Doyle (eds.), *Molecular Systematics of Plants II—Data Sequencing*. Kluwer, pp. 132-63.
- Maddison, W. (1991): "Squared-Change Parsimony Reconstructions of Ancestral States for Continuous-Valued Characters on a Phylogenetic Tree." *Systematic Zoology* 40: 304-14.

- Martins, E. (1999): "Estimation of Ancestral States of Continuous Characters—A Computer Simulation Study." *Systematic Biology* **48**: 642–50.
- Orzack, S. and Sober, E. (2001): "Adaptation, Phylogenetic Inertia, and the Method of Controlled Comparisons." In S. Orzack and E. Sober (eds.), *Adaptationism and Optimality*. Cambridge: Cambridge University Press.
- Pagel, M. (1994): "Detecting Correlated Evolution on Phylogenies—a General Method for Comparative Analysis." *Proceedings of the Royal Society* **B255**: 37–45.
- (1999): "The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies." *Systematic Biology* **48**: 612–22.
- Penny, D., Lockhart, P., Steel, M., Henny, M. (1994): "The Role of Models in Reconstructing Evolutionary Trees." In R. Scotland, D. Siebert, and D. Williams (eds.), *Models of Phylogeny Reconstruction*. Systematics Association Special vol. **52**, pp. 211–30. Oxford: Clarendon Press.
- Parzen, E. (1962): *Stochastic processes*. San Francisco: Holden-Day.
- Ridley, M. (1983): *The Explanation of Organic Diversity*. Oxford: Oxford University Press.
- Royall, R. (1997): *Statistical Evidence—a Likelihood Paradigm*. Boca Raton: Chapman and Hall.
- Schluter, D., Price, T., Mooers, A., and Ludwig, D. (1997): "Likelihood of Ancestor States in Adaptive Radiation." *Evolution* **51**: 1699–1711.
- Schultz, T. and Churchill, G. (1999): "The Role of Subjectivity in Reconstructing Ancestral Character States—A Bayesian Approach to Unknown Rates, States, and Transformation Asymmetries." *Systematic Biology* **48**: 651–64.
- Sober, E. (1988): *Reconstructing the Past—Parsimony, Evolution, and Inference*. Cambridge: M.I.T. Press.
- Steel, M. and Penny, D. (2000): "Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics." *Molecular Biology and Evolution* **17**: 839–50.
- Tuffley, C. and Steel, M. (1997): "Links Between Maximum Likelihood and Maximum Parsimony Under a Simple Model of Site Substitution." *Bulletin of Mathematical Biology* **59**: 581–607.

A vertical bar on the left side of the page, consisting of a series of horizontal segments in shades of yellow and orange, with a small red diamond at the top.

COPYRIGHT INFORMATION

TITLE: Reconstructing the character states of ancestors: a likelihood perspective on cladistic parsimony

SOURCE: The Monist 85 no1 Ja 2002

WN: 0200101621008

The magazine publisher is the copyright holder of this article and it is reproduced with permission. Further reproduction of this article in violation of the copyright is prohibited..

Copyright 1982-2002 The H.W. Wilson Company. All rights reserved.