Syst. Biol. 52(3):352–367, 2003

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390196984

Failed Refutations: Further Comments on Parsimony and Likelihood Methods and Their Relationship to Popper's Degree of Corroboration

KEVIN DE QUEIROZ[1] AND STEVEN POE[2,3]

[1]*Department of Systematic Biology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0162, USA; E-mail: dequeiroz.kevin@nmnh.si.edu*
[2]*Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720-3160, USA*

Abstract.— Kluge's (2001, Syst. Biol. 50:322–330) continued arguments that phylogenetic methods based on the statistical principle of likelihood are incompatible with the philosophy of science described by Karl Popper are based on false premises related to Kluge's misrepresentations of Popper's philosophy. Contrary to Kluge's conjectures, likelihood methods are not inherently verificationist; they do not treat every instance of a hypothesis as confirmation of that hypothesis. The historical nature of phylogeny does not preclude phylogenetic hypotheses from being evaluated using the probability of evidence. The low absolute probabilities of hypotheses are irrelevant to the correct interpretation of Popper's concept termed *degree of corroboration*, which is defined entirely in terms of relative probabilities. Popper did not advocate minimizing background knowledge; in any case, the background knowledge of both parsimony and likelihood methods consists of the general assumption of descent with modification and additional assumptions that are deterministic, concerning which tree is considered most highly corroborated. Although parsimony methods do not assume (in the sense of entailing) that homoplasy is rare, they do assume (in the sense of requiring to obtain a correct phylogenetic inference) certain things about patterns of homoplasy. Both parsimony and likelihood methods assume (in the sense of implying by the manner in which they operate) various things about evolutionary processes, although violation of those assumptions does not always cause the methods to yield incorrect phylogenetic inferences. Test severity is increased by sampling additional relevant characters rather than by character reanalysis, although either interpretation is compatible with the use of phylogenetic likelihood methods. Neither parsimony nor likelihood methods assess test severity (critical evidence) when used to identify a most highly corroborated tree(s) based on a single method or model and a single body of data; however, both classes of methods can be used to perform severe tests. The assumption of descent with modification is insufficient background knowledge to justify cladistic parsimony as a method for assessing degree of corroboration. Invoking equivalency between parsimony methods and likelihood models that assume no common mechanism emphasizes the necessity of additional assumptions, at least some of which are probabilistic in nature. Incongruent characters do not qualify as falsifiers of phylogenetic hypotheses except under extremely unrealistic evolutionary models; therefore, justifications of parsimony methods as falsificationist based on the idea that they minimize the ad hoc dismissal of falsifiers are questionable. Probabilistic concepts such as degree of corroboration and likelihood provide a more appropriate framework for understanding how phylogenetics conforms with Popper's philosophy of science. Likelihood ratio tests do not assume what is at issue but instead are methods for testing hypotheses according to an accepted standard of statistical significance and for incorporating considerations about test severity. These tests are fundamentally similar to Popper's degree of corroboration in being based on the relationship between the probability of the evidence e in the presence versus absence of the hypothesis h, i.e., between $p(e|hb)$ and $p(e|b)$, where b is the background knowledge. Both parsimony and likelihood methods are inductive in that their inferences (particular trees) contain more information than (and therefore do not follow necessarily from) the observations upon which they are based; however, both are deductive in that their conclusions (tree lengths and likelihoods) follow necessarily from their premises (particular trees, observed character state distributions, and evolutionary models). For these and other reasons, phylogenetic likelihood methods are highly compatible with Karl Popper's philosophy of science and offer several advantages over parsimony methods in this context. [Assumptions; corroboration; Karl Popper; likelihood; parsimony; philosophy; phylogenetics; probability.]

The conformity of phylogenetic parsimony and likelihood methods with the philosophy of science developed by Karl R. Popper (1902–1994) has become the focus of debate. Some authors (e.g., Siddall and Kluge, 1997) have asserted that parsimony methods, but not those based on the statistical principle of likelihood, can be interpreted as examples of Popper's method for assessing the degree of corroboration of scientific hypotheses. Contrary to those views, we previously presented evidence that Popper's concept termed *degree of corroboration* is based on likelihood, and we argued that phylogenetic likelihood methods are straightforwardly interpreted as methods for assessing degree of corroboration (de Queiroz and Poe, 2001). We argued that parsimony methods can also be interpreted as methods for assessing degree of corroboration, but only if they are considered to incorporate implicit probabilistic assumptions. Kluge (2001) claimed to refute our propositions. Here, we show that Kluge has explicitly or implicitly accepted all of our main propositions (although not always consistently) and that his purported refutations are uniformly unsuccessful. We address those purported refutations point by point and show that they rest on misrepresentations of Popper's views and other false premises, several of which we had already dealt with in our earlier paper. Far from refuting our propositions about philosophy and methods of phylogenetic inference, Kluge's arguments represent failed refutations. In exposing their flaws, we address diverse issues about the philosophy of phylogenetic inference, including several that we did not cover in our previous paper.

[3]Present address: Department of Biology, 167A Castetter Hall, University of New Mexico, Albuquerque, NM 87131-1091, USA; E-mail: anolis@unm.edu.

Falsificationism and Verificationism

In his section with the same title, Kluge (2001) presented his characterization of falsificationism and verificationism and some of their variant forms (e.g., sophisticated falsificationism, classical verificationism). Most of Kluge's summary of these concepts has no direct bearing on our disagreements with him, except for his misleading statement that "neo-verificationism . . . is concerned with the relative truthfulness (verisimilitude) of hypotheses, as determined by their degrees of probability" (Kluge, 2001:323), which relates to his incorrect assertion that our position on the relationship between corroboration and likelihood in phylogenetics "render[s] Popper a verificationist" (Kluge, 2001:322).

Kluge's statement is misleading in implying a necessary connection between probabilistic methods and verificationism and thus a fundamental incompatibility between those methods and Popper's falsificationist philosophy. On the contrary, as we argued previously (de Queiroz and Poe, 2001), Popper was not opposed to probabilistic methods themselves; his own degree of corroboration is one of them. Instead, Popper was critical of methods that attempt to assign probabilities to hypotheses as a substitute for establishing their certain truth, which he viewed as a mistaken solution to the problem of induction (Popper, 1983:217). He was also critical of the idea that scientists seek hypotheses that have high probabilities (i.e., of being true). The reason is that such hypotheses generally have low informative content; they posit little or nothing that is not already accepted in the background knowledge or the evidence that they are supposed to explain. Tautologies, for example, have high probabilities (e.g., Popper, 1959:399, 1983:232). According to Popper (e.g., 1959:118–119), the content and therefore the testability of a hypothesis h is inversely related to its probability; when the content and testability of a hypothesis h is maximal, the probability of the hypothesis $p(h)$ or $p(h|b)$ is equal to 0, where b is the background knowledge (Popper, 1983:241). However, as Popper himself pointed out (1983:243), and as we noted (de Queiroz and Poe, 2001:309), likelihood methods do not attempt to assign probabilities (high or low) to hypotheses. Likelihood is not the probability of the hypothesis $p(h)$ or $p(h|b)$ but the probability of the evidence (e) given the hypothesis $p(e|h)$ or $p(e|hb)$. Therefore, the probabilistic nature of likelihood is not incompatible with Popper's philosophy, nor is likelihood an example of (neo)verificationism, even according to Kluge's characterization of that approach.

Kluge's continued failure to distinguish consistently between the probability of hypotheses and the probability of evidence given particular hypotheses seems to underlie his erroneous proposition that our views on the relationship between corroboration and likelihood in phylogenetics render Popper a verificationist. Although Kluge presented no explicit argument to substantiate this proposition, he implied that his objection has to do with our argument that standard phylogenetic analyses (i.e., those that attempt to identify an optimal tree or trees for a particular data set given a single phylogenetic method or model) make no attempt to assess test severity (and thus critical evidence), represented by $p(e|b)$ in Popper's corroboration equation, and therefore that the relative degree of corroboration C of the alternative trees (representing the hypothesis h) is determined entirely by $p(e|hb)$, i.e., by their likelihoods. Thus, according to Kluge (2001:322) "de Queiroz and Poe interpret degree of corroboration *solely* as a likelihood argument—no importance is attributed to *critical* evidence—and so tacitly they render Popper a verificationist." Kluge is wrong in stating that we attribute no importance to critical evidence (see "Critical Evidence and Test Severity," below). Regardless, his implicit argument seems to be that likelihood is inherently verificationist, and therefore interpreting Popper's C solely as a likelihood argument effectively renders Popper a verificationist. This conclusion is incorrect because it rests on the false premise that likelihood is inherently verificationist, a view of Kluge's that does not come from Popper.

Popper (1983:234) characterized verificationism and falsificationism as approaches or attitudes:

> (a) The uncritical or verificationist attitude: one looks out for "verification" or "confirmation" or "instantiation," and ones finds it, as a rule. Every observed "instance" of the theory is thought to "confirm" the theory.
>
> (b) The critical attitude, or falsificationist attitude: one looks for falsification, or for counter-instances. Only if the most conscientious search for counter-instances does not succeed may we speak of a corroboration of the theory.

He continued (p. 234):

> It may be asked whether it is really so uncritical to look upon all instances of a theory as confirmations of it. But it can be shown that, for logical reasons, to do so amounts to the belief that *everything is a confirmation*—with the sole exception of a counter-instance. Thus not only white swans but also black ravens and red shoes "confirm" the theory that all swans are white.

According to these characterizations, neither parsimony nor likelihood methods of phylogenetic inference are inherently either falsificationist or verificationist; and adopting one attitude or the other is logically independent of a researcher's preferred optimality criterion. Thus, a verificationist would consider the hypothesis that A and B are sister taxa to be confirmed as long as the results of an analysis, using either parsimony or likelihood, did not contradict that relationship. For example, an analysis that failed to resolve the relationships of taxa A and B would be considered to confirm the hypothesized AB relationship. To use an example analogous to Popper's black ravens and red shoes, a resolved XY relationship would be considered to confirm the hypothesized AB relationship. A verificationist might also search for data that support the AB relationship without looking for data that might contradict it. In contrast, a falsificationist would only consider the hypothesized AB relationship corroborated if the AB group were present and well-supported by an analysis, using either parsimony or

likelihood, or by several such analyses based on different assumptions and/or data sets. A falsificationist would also collect data that had the potential to contradict the hypothesized AB relationship. Neither the verificationist nor the falsificationist attitude has any necessary connection with a particular phylogenetic optimality criterion.

In the context of the distinction between verificationist and falsificationist attitudes, parsimony and likelihood methods are, in fact, fundamentally similar regarding the general manner in which they are used to evaluate alternative hypotheses. Both classes of methods assign a score to each member of a set of alternative trees, and both use those scores to establish a preference among those trees, that is, "to grade hypotheses according to the tests passed by them" (Popper, 1983:220). The fact that one optimality criterion is explicitly probabilistic and the other is not hardly constitutes evidence that the former is inherently verificationist and the latter inherently falsificationist according to Popper's characterization of those attitudes. On the contrary, the most important intrinsic difference between standard parsimony and likelihood methods has nothing to do with falsificationist versus verificationist attitudes but instead concerns the nature of the evidence that the different methods treat as corroborating versus contradicting particular hypothesized relationships.

Consider the simple case of three taxa, A, B, and C, and a hypothesized sister-group relationship between taxa A and B. Under standard parsimony methods, a hypothesized AB relationship is considered corroborated if A and B share more derived character states with one another than either does with C; it is considered contradicted if either A or B shares more derived states with C. (This example considers only equally weighted characters and transformations, both for the sake of simplicity and because that is the model advocated by Kluge, 1997a.) In contrast, under standard likelihood methods, sharing more derived states does not necessarily constitute evidence of a closer relationship (though it may in many instances). Standard phylogenetic likelihood methods were developed to deal with the phenomenon of long-branch attraction (Felsenstein, 1978; Hendy and Penny, 1989), which results from derived states that are shared because of homoplasy. Likelihood models accomplish this by modeling the probability of character transformation, and therefore of homoplasy (given constraints on the number of possible states), as a function of branch length (Swofford et al., 1996; Lewis, 1998). As a consequence, the numbers of derived states expected to be shared as the result of homoplasy are related to the lengths of the various branches of the tree. In this context, for a hypothesized AB relationship to be considered corroborated, taxa A and B must share more derived states with one another than expected (as the result of homoplasy) given the lengths of their subtending branches. Similarly, to be considered contradicted, either A or B must share more derived states with C than expected given the lengths of the various subtending branches. Neither the criterion of corroboration adopted by parsimony nor that adopted by likelihood constitutes taking every instance of a hypothesis as confirming that hypothesis; therefore, contrary to Kluge's view, there is no basis for thinking that likelihood is inherently verificationist.

Probability and Historical Inference

In his section titled "Probabilism," Kluge (2001:323) defined probabilism as "the doctrine that the reasonableness of hypotheses is to be judged with degrees of probability, the position of 'reasonableness' being flanked by extreme optimist and skeptic positions:—'certainty can be achieved', and 'probabilities cannot even be assigned'." Kluge apparently considers himself a skeptic, at least with respect to phylogenetic hypotheses, for he seems to believe that there are problems with applying probabilities to historical events (see also Siddall and Kluge, 1997). He asked (2001:323) rhetorically "If it is unrealistic to assign degrees of probability to events in history, because they are necessarily unique, then what is the basis for choosing among competing cladograms?"

If the unique historical events to which Kluge referred are represented by his competing cladograms (alternative trees), then his position, even if correct, is not at odds with phylogenetic likelihood methods, which treat the alternative trees as hypotheses. The likelihood of a hypothesis is not the probability of the hypothesis h but the probability of the evidence e given the hypothesis. Phylogenetic likelihood methods make no attempt to assess the probabilities of the unique historical events represented by trees. Therefore, if Kluge's probabilism equates the reasonableness of a hypothesis with the probability of that hypothesis, with one possibility being that certainty about its truth can be achieved, then this concept has nothing to do with likelihood (see "Falsificationism and Verificationism").

The evolutionary character transformations underlying the evidence e, the distributions of the states of a set of characters among the members of a set of taxa, are also unique historical events. However, contrary to Kluge's suggestion, there is no reason to think that those events are any less amenable to probabilistic analysis than are events that are typically analyzed probabilistically, such as coin tosses, which are also unique historical events. Just as an individual character transformation (or set of transformations) cannot occur more than once in phylogeny (the tape of life cannot be replayed), an individual coin toss (or set of tosses) cannot occur more than once in history. If the possibility of additional coin tosses (replicates) is considered an important difference, the coin can be melted down to eliminate that possibility. Therefore, if it is permissible to evaluate hypotheses about the degree of bias (or lack thereof) of a coin that no longer exists based on the probabilities of the outcomes of the unique historical events represented by coin tosses, then evaluating hypotheses about phylogenetic relationships based on the probabilities of evolutionary character transformations should not be prohibited by the historical nature of those events (see also Sanderson, 1995).

Relative and Absolute Probabilities

Contrary to the views of Siddall and Kluge (1997), we previously presented quotations (de Queiroz and Poe, 2001) demonstrating that Popper intended for his degree of corroboration to be compatible with different interpretations of the probabilities upon which it is based, including both logical and frequency interpretations. Kluge presented no evidence against this proposition but instead changed his position, asserting that the important distinction is between absolute and relative, rather than logical and frequency, probabilities (although it is not clear that he recognized a difference). Thus, according to Kluge (2001:323), "de Queiroz and Poe's failure to correctly interpret Popper and his falsificationist philosophy of science stems from not distinguishing their use of relative probability from his use of absolute probability." Although Kluge never substantiated this statement, in his section titled "Probabilism," he reviewed various kinds of probability, noting that "probability can . . . be stated in non-relative (non-frequentist) terms, as an absolute (logical) probability" (Kluge, 2001:323) and concluding with the points, quoted or paraphrased from Popper, that the greater the content of a statement, the lower its absolute logical probability and that "absolute probability cannot be interpreted as a frequency, except in the most trivial sense" (Kluge, 2001: 323).

Contrary to Kluge's implication, these points are entirely compatible with our interpretation of Popper's philosophy. First, consider the distinction between relative and absolute probabilities and the fact that absolute probabilities cannot usefully be interpreted as frequencies. Kluge's summary of these points appears to have been taken directly from Popper (1983:283–284):

> There are only two formulae which have to be referred to in a discussion of the interpretations of the probability calculus; the first is
>
> (R) $$p(a, b) = r,$$
>
> or in words "the (relative) probability of a given b is equal to r," where r is some fraction between 0 and 1 (these limits included).
>
> The second formula is
>
> (A) $$p(a) = r,$$
>
> in words "the (absolute) probability of a is equal to r."
>
> The relative probability of a given b has sometimes also been called "conditional probability of a under the condition b"; and the absolute probability of a has sometimes been called the "prior" or the "initial" or the *a priori* probability of a.
>
> If we interpret (R) in the sense of the *frequency interpretation*, discussed at length in *L.Sc.D.* [*The Logic of Scientific Discovery*], then
>
> $$p(a, b) = r$$
>
> becomes "the relative frequency of a within the reference class b (or the reference sequence b) is equal to r."
>
> On the other hand,
>
> $$p(a) = r$$
>
> can hardly be interpreted in frequency terms (except in a trivial way—by taking the reference class b as "understood"). For there is no point in saying "sparrows occur frequently" unless "among European birds," or something like this, is taken as understood. But
>
> $$p(a) = r$$
>
> can be interpreted without difficulty in other than frequency terms, for example within what I call the "logical interpretation" of probability. Here we take the letters "a," "b," etc., to be names of statements. The absolute logical probability of a—i.e., $p(a)$—is then what I described in *L.Sc.D.* as "logical probability": its value r is the greater the less the statement a says.

Contrary to Kluge's assertion that we interpreted Popper incorrectly by not distinguishing between relative and absolute probabilities, this distinction in fact supports our interpretation of Popper's degree of corroboration in terms of relative probabilities. Kluge did not explain how our use of relative probabilities led to a supposed incorrect interpretation. In any case, our interpretation is perfectly consistent with Popper's definition of *degree of corroboration C*, which is based entirely on relative probabilities. Consider the expression used by Popper (1983:240) to define that concept:

$$C(h, e, b) = \frac{p(e, hb) - p(e, b)}{p(e, hb) - p(eh, b) + p(e, b)}.$$

Relative probabilities are conditional probabilities, and every term in the defining formula of the corroboration expression is a conditional or relative probability. Thus, the correct interpretation of Popper's degree of corroboration must use relative probabilities. As a consequence, Kluge's other point, that absolute probabilities cannot be usefully interpreted as frequencies, is irrelevant to the correct interpretation of Popper's definition of degree of corroboration. Relative probabilities can be interpreted as frequencies, and Popper's degree of corroboration C is defined solely in terms of relative probabilities.

Next, consider Popper's proposition that the content of a statement is inversely related to its absolute logical probability. As noted here ("Falsificationism and Verificationism") and previously (de Queiroz and Poe, 2001), Popper used this proposition to argue against the idea that scientists seek hypotheses with high probabilities, that is, as opposed to hypotheses for which the evidence has a high probability relative to its probability given competing hypotheses. If Kluge meant to imply that phylogenetic likelihood methods seek hypotheses with high probabilities, then he is simply incorrect. However, Kluge coauthored a paper (Farris et al., 2001) that agreed with our interpretation of phylogenetic analysis as an example of Popperian corroboration in equating the hypothesis h with a tree and included a statement that phylogenetic hypotheses have high content, implying that they have low absolute logical probabilities (Popper, 1983:231, 284). The absolute or prior probability of a particular tree (h) is not dependent on the phylogenetic method that is later used to evaluate that tree; if it were, it would be a relative rather than an absolute probability. Therefore, a tree that has a low absolute (prior) probability $p(h)$ will have such a probability regardless

of whether parsimony or likelihood is used to evaluate its degree of corroboration, and Popper's proposition that hypotheses of high content have low absolute logical probabilities $p(h)$ is in no way at odds with the use of methods (i.e., those based on likelihood) that associate a high relative probability of the evidence $p(e|hb)$ with a high degree of corroboration.

In sum, Kluge is correct in his belief that the distinction between relative and absolute probabilities is important to the correct interpretation of Popper's philosophy. That distinction supports our interpretation of Popper's degree of corroboration in terms of relative probabilities (and is therefore consistent with a frequency interpretation of probability). It is also consistent with the use of likelihood to evaluate hypotheses, including phylogenetic trees, that have low absolute logical probabilities. Kluge's incorrect belief that the use of phylogenetic likelihood methods is at odds with Popper's view that useful hypotheses have low absolute (i.e., prior) probabilities stems from Kluge's own failure to distinguish consistently between absolute and relative probabilities, specifically, between the absolute probability of a hypothesis $p(h)$ and the relative probability of the evidence given a hypothesis $p(e|hb)$.

Content, Severity, and Corroboration

In roughly the first page and a half of his section titled "Deriving Popperian Testability: An Exercise in Falsificationism" (i.e., up to his discussion of background knowledge), Kluge (2001) summarized Popper's views on the content of hypotheses, the severity of tests, and some general points about the evaluation of alternative hypotheses using degree of corroboration. This summary of Popper's views presented no arguments against any of our propositions and consisted mostly of statements that are not in dispute. The main points seem to be as follows: 1) the probability of a hypothesis decreases with increasing content so that highly improbable, rather than highly probable, hypotheses are valued by scientists; 2) the difference between the probability of the evidence with and without the hypothesis, $p(e|hb) - p(e|b)$, is central to Popper's notions of the severity S of a test and the explanatory power E and degree of corroboration C of a hypothesis; and 3) for a hypothesis to be maximally corroborated, the probability of the evidence given the background knowledge alone $p(e|b)$ must be zero. None of these points conflict with our propositions about the relationship between Popper's philosophy and methods of phylogenetic analysis.

In contrast with his earlier papers (e.g., Kluge, 1997a, 1997b; Siddall and Kluge, 1997), Kluge (2001:324) explicitly acknowledged that the term $p(e|hb)$ in Popper's equations defining *severity* S, *explanatory power* E, and *degree of corroboration* C is likelihood, as we argued previously (de Queiroz and Poe, 2001). One thing in this summary of Kluge's that is potentially misleading is this statement: "As Popper went on to argue, if growth of knowledge means increasing content, then surely high probability cannot be the goal of science" (Kluge, 2001:324). This statement and similar ones immediately preceding and following it refer to Popper's views about the probabilities of hypotheses, $p(h)$ or $p(h|b)$. Those statements are not at odds with the law of likelihood (e.g., Edwards, 1972:30), which says nothing about *hypotheses* of high probability being a goal of science but instead states that the best-supported hypothesis is the one for which the probability of the *evidence*, $p(e|h)$ or $p(e|hb)$, is maximal (see also de Queiroz and Poe, 2001, and "Falsificationism and Verificationism," above).

Background Knowledge

Following his summary of Popper's views on severity, explanatory power, and some general points about degree of corroboration, Kluge quoted four passages from Popper on background knowledge, represented by the term b in Popper's definition of degree of corroboration. Kluge then offered his interpretation of Popper's b, which he used to criticize the inclusion of likelihood models in b and thus the usefulness of phylogenetic likelihood methods. Kluge's interpretation of b is seriously flawed, which undermines his criticisms of likelihood models as components of b. In addition, his criticisms of likelihood methods apply equally to those of cladistic parsimony.

The Nature of Background Knowledge

Kluge (2001:325; see also Farris et al., 2001) described his interpretation of Popper's background knowledge as follows: "I consider Popper's concept of background knowledge b to comprise only currently accepted (well-corroborated) theories and experimental results.... Background knowledge does not include falsified theories, nor otherwise false assumptions. Thus, once a theory has been falsified, it can no longer serve as background knowledge." This statement misrepresents Popper's concept of background knowledge b, which was not intended to include only well-corroborated theories or those that have never been purported to have been falsified. In Popper's own words (1962:238),

> While discussing a problem we always accept (if only temporarily) all kinds of things as *unproblematic*: they constitute for the time being, and for the discussion of this particular problem, what I call our *background knowledge*. Few parts of this background knowledge will appear to us in all contexts as absolutely unproblematic, and any particular part of it *may* be challenged at any time, especially if we suspect that its uncritical acceptance may be responsible for some of our difficulties.

This statement, with its use of the phrases "if only temporarily," "for the time being," and "for the discussion of this particular problem," emphasizes the tentative nature of many aspects of b. Although Kluge quoted parts of the same passage, he omitted these phrases. Moreover, Popper's statement explicitly acknowledges that b may include components that have been accepted uncritically, which can hardly be equated with well-corroborated theories. If this is not sufficient evidence to contradict Kluge's interpretation, consider the following statement by Popper (1962:238):

> The fact that, as a rule, we are at any given moment taking a vast amount of traditional knowledge for granted . . . creates no difficulty for the falsificationist or fallibilist. For he does not *accept* this background knowledge; neither as established nor as fairly certain, nor yet as probable. He knows that even its tentative acceptance is risky, and stresses that every bit of it is open to criticism, even though only in a piecemeal way.

This statement explicitly rejects the equation of b with accepted or established (well-corroborated) theories. Furthermore, it emphasizes that any tentative acceptance (i.e., for the purpose of a particular test) is risky, a caution that reinforces his earlier point that certain components of b may have been uncritically accepted and therefore are anything but well corroborated. Thus, Kluge's interpretation of b clearly misrepresents Popper's views. Moreover, Kluge's statement that once a theory has been falsified it can no longer serve as background knowledge is an example of naïve falsificationism (Lakatos, 1970; see also Popper, 1983:xxii–xxv).

Kluge (2001:326) was also wrong in claiming that "Popper declared his minimalist philosophy when it came to auxiliary propositions," which, for Kluge (2001:326), includes background knowledge and models. For example, in the second passage from Popper (1962:238) quoted above, Popper equated background knowledge with a "vast amount of traditional knowledge." Similarly, according to Popper (1962:239), "One fact which is characteristic of the situation in which the scientist finds himself is that we constantly add to our background knowledge." These statements can hardly be considered indicative of a minimalist philosophy. Kluge's advocacy of minimizing assumptions is based on the fact that $p(e|hb) - p(e|b)$ and thus corroboration C is maximized when $p(e|b)$ is minimized (Kluge, 2001:326). However, his erroneous conclusion implies that the role of the term $p(e|b)$ in Popper's corroboration expression is to minimize assumptions (b). On the contrary, the quotations from Popper and his discussions concerning $p(e|b)$ (e.g., Popper, 1983:237–241) indicate that the role of that term is to identify severe tests and hypotheses (h) with high degrees of corroboration (C) and explanatory power (E) given whatever assumptions an investigator has provisionally accepted for the purpose of testing those hypotheses (see also Faith and Trueman, 2001). The minimalist view of b is Kluge's, not Popper's.

Background Knowledge and Models

Kluge's erroneous interpretation of background knowledge b undermines his criticism of likelihood models as parts of b. He argued that "to assume a common mechanism (Steel and Penny, 2000), such as a homogeneity assumption of evolution in a likelihood model . . . is likely to be false," which "is not what Popper had in mind for background knowledge" (Kluge, 2001:325). Because Popper did not prohibit b from containing components that might be false (see "The Nature of Background Knowledge"), Kluge's argument against the inclusion of likelihood models in b is nullified by its false premise. Moreover, parsimony's implicit assumption of no common mechanism is no less likely to be false (see next section).

Another argument presented by Kluge (see also Siddall and Kluge, 1997) against adopting likelihood models as part of the background knowledge is that such models are "deterministic to the inference," whereas according to Kluge (2001:325), background knowledge is not. To illustrate this supposed distinction, he proposed the assumption of descent with modification as an example of background knowledge, and he argued (2001:325) that "should 'descent, with modification' prove false, minimizing unweighted steps with the parsimony algorithm would still lead to the shortest length cladogram, and the character generalities to be explained as something other than homologues."

There are several flaws in this argument. First, although likelihood models are deterministic to the inference, the same is true for parsimony methods. However, this property hardly constitutes a problem for either approach in that being deterministic is what makes these methods useful for identifying optimal trees. Both parsimony and likelihood are optimality criteria (Swofford and Olsen, 1990; Swofford et al., 1996). For a phylogenetic method based on an optimality criterion, the method consists of the optimality criterion itself—in this case, parsimony or likelihood—and any additional factors inherent in its implementation, such as evaluating only hypotheses taking the form of trees under both optimality criteria, character state order (including unordered) and character weights (including equal weights) under parsimony, and transformation probabilities and among-site rate variation under likelihood. All of these additional factors are deterministic to the inference in that they determine which tree is considered optimal, so likelihood models are no different than parsimony methods with respect to this property. Contrary to Kluge's view, being deterministic provides a reason neither for rejecting the use of likelihood models as part of the background knowledge nor for preferring parsimony methods.

Second, Kluge's invocation of descent with modification as an example of background knowledge reinforces our conclusions rather than contradicting them. The assumption of descent with modification is a very general assumption that forms part of the background knowledge of both parsimony and likelihood methods and thus has no bearing on the differences between them. Instead, it influences both classes of methods similarly and only on a general methodological level. For example, the assumption of descent with modification presumably accounts for the fact that both parsimony and likelihood methods model the process that generated the data as character state transformations (modification) along the branches of trees (descent). Therefore, should the assumption of descent with modification be judged false, the consequences would be similar for both parsimony and likelihood methods. Both methods could still be used to obtain optimal trees, but in both cases, alternative (i.e., nonevolutionary) explanations would have to be invoked for shared character states.

Third, Kluge's discussion of assumptions is misleading in implying that descent with modification is the only assumption associated with parsimony methods, that nothing more need be invoked than the general philosophical principle of parsimony (see de Queiroz and Poe, 2001:306–307, for a discussion of the distinction between cladistic parsimony methods and the principle of parsimony). Additional assumptions associated with parsimony methods will be described below; here we comment only on Kluge's proposition that "there are two classes of auxiliary assumptions: background knowledge and models" (2001:326) and "the model *M* is deterministic to the inference" (2001:325).

Kluge's example of background knowledge (the assumption of descent with modification) is common to phylogenetic methods based on both parsimony and likelihood, and both classes of methods include components that are deterministic to the phylogenetic inference and therefore qualify as models according to his view. Therefore, Kluge's distinction between background knowledge and models has no bearing on the differences between parsimony and likelihood methods. Furthermore, in contrast with Kluge's concept of background knowledge, Popper's concept has nothing to do with whether an assumption is deterministic but instead requires only that it "is not questioned while testing the theory under investigation" (Popper, 1983:244). Because the deterministic components of a phylogenetic method, including both parsimony and likelihood models, are held constant in a given phylogenetic analysis, they represent part of the background knowledge according to Popper's (as opposed to Kluge's) definition of that concept.

Assumptions of Parsimony and Likelihood Methods

Kluge's (2001:325) proposal that background knowledge *b* does not include "admitted false assumptions," even if correct, would be just as damaging to parsimony as to likelihood methods. The reason is that both classes of methods incorporate assumptions that are admitted to be false, at least in some cases. For example, the exclusively diverging model of evolution assumed in both parsimony and likelihood methods that yield results only in the form of trees (as opposed to more extensively connected graphs) is such an assumption, which is to say that the existence of reticulation (e.g., hybridization, lateral gene transfer) is a well-corroborated hypothesis (e.g., Cole et al., 1988; McDade, 1990; Lawrence, 2001). Moreover, Kluge (1997a) advocated a method—parsimony with equally weighted characters—that would treat all sites in nucleotide sequence data identically, despite strong corroborating evidence for the existence of rate variation among sites (e.g., Yang, 1993, 1996b; Yang et al., 1994). In addition, a homogeneity (stationarity) assumption, which Kluge (see also Farris et al., 2001) viewed as damaging to current likelihood models, is not avoided by parsimony methods. A substitution process that exhibits nonstationarity across the tree can cause both likelihood (under stationary models) and parsimony methods to yield incorrect inferences (Lockhart et al., 1994; Steel and Penny, 2000).

A specific assumption of standard likelihood methods that Kluge criticized as "likely to be false" (2001:325) is the assumption of a "common mechanism," and he accepted (2001:327) the equivalency of parsimony with Tuffley and Steel's (1997) maximum likelihood model that assumes "no common mechanism." These assumptions are indeed among the most fundamental differences between standard likelihood and parsimony methods. Under standard likelihood methods (i.e., those that assume a common mechanism), different characters are assumed to evolve under the same evolutionary processes. Consequently, the probability of character change on a given branch is proportional among characters and thus can be modeled as a function of the branch's length. Under standard parsimony methods and likelihood models that assume no common mechanism, different characters are assumed not to evolve under the same evolutionary processes. Consequently, the inference of character state transformation on a branch is not influenced by the branch's length but only by the states of the terminal taxa (i.e., the branch lengths are different for each character). As far as we are aware, no explicit test of these alternative hypotheses has been preformed, so neither can be said to be well-corroborated. Moreover, at least for some kinds of data, the assumption of no common mechanism seems more likely to be false. For example, for pseudogenes and most third-codon positions in protein-coding genes, where natural selection presumably does not affect the rate of substitution, it seems reasonable to expect that transformations in all characters have higher probabilities on branches of long temporal duration than on those of short temporal duration rather than the probabilities of change being entirely independent of the temporal duration of branches.

Although both parsimony and likelihood methods may incorporate false assumptions, that does not mean that they will necessarily favor an incorrect tree. Both classes of methods can yield correct results when one or more of their assumptions are violated (e.g., Yang, 1996a). When discussing the assumptions of a method, it is important to distinguish among (1) propositions that are implied by the manner in which the method operates (e.g., certain parsimony and likelihood models treat transitions as if they were equivalent to transversions), (2) propositions that are entailed (i.e., as necessary conclusions) by the method's use (e.g., phenetic clustering methods necessarily produce trees in which all terminal taxa are equidistant from the root, implying equal rates of change among lineages), and (3) propositions that must be true (i.e., conditions that must be met) for the method to yield correct results. All three classes of propositions can be considered assumptions in the general sense that they are not questioned when using the method to evaluate alternative hypotheses; they are also related to one another. Of particular relevance to the present case is that the manner in which a method operates generally defines a set of conditions that includes as a subset the conditions under which the

method will fail to produce a correct result, that is, it defines conditions that are necessary but not sufficient for failure.

In the case of likelihood methods, the assumptions inherent in how the method operates are more or less explicit in the model and include such things as the rate matrix (assumptions about the relative rates of different classes of substitutions), base frequencies, and among-site rate variation. For parsimony methods, the assumptions are implicit rather than explicit; in any case, these methods operate by minimizing and thus systematically underestimating homoplasy (Swofford et al., 2001). Siddall and Kluge (1997), following Farris (1983), argued that parsimony methods do not assume, in the sense of entailing, that homoplasy is rare or minimal. This conclusion seems uncontroversial given that in many analyses of real data the most-parsimonious tree requires considerable homoplasy. However, it does not demonstrate the general independence of parsimony methods from assumptions about homoplasy. Specifically, it does not provide a reason for thinking that parsimony analyses will invariably yield correct phylogenetic inferences regardless of the pattern and/or frequency of homoplasy. Because parsimony methods operate by minimizing homoplasy, the conditions under which those methods will fail to yield correct results must involve homoplasy, though even large amounts of homoplasy will not always lead to failure. Homoplasy is a necessary condition, but not a sufficient one, for the failure of parsimony methods.

Sufficient conditions for parsimony methods to yield incorrect results are also easily identified, at least for equally weighted characters and transformations. Under this model, parsimony should lead to an incorrect inference whenever taxa that are not sister groups share more derived states with one another as the result of homoplasy than either does with its true sister group as the result of both inheritance from common ancestors and homoplasy. These conditions are necessary and sufficient for the failure of parsimony with equally weighted characters and transformations, and in this sense, parsimony methods *do* involve assumptions about homoplasy. Moreover, under certain patterns of branch length inequality (e.g., Felsenstein, 1978; Hendy and Penny, 1989), those assumptions are likely to be false.

Parsimony methods operate by minimizing and thus systematically underestimating homoplasy. Although homoplasy does not always cause parsimony methods to yield incorrect results, the cases in which those methods yield incorrect results necessarily involve homoplasy. Similar considerations apply to likelihood methods. For example, a likelihood model that does not incorporate a parameter describing rate variation among sites effectively assumes equal rates among sites. The existence of rate variation among sites will not always cause that model to yield incorrect results; however, those cases in which the model tends to give an incorrect result necessarily involve violation of one of its assumptions, such as that of equal rates among sites. Contrary to Kluge's implications, the charge of incorporating potentially false assumptions is no more damaging to likelihood methods than to methods of cladistic parsimony. Both classes of methods incorporate assumptions that can be violated, although violation may or may not lead to an incorrect phylogenetic inference.

Corroboration and Likelihood

At the end of his section "Deriving Popperian Testability: An Exercise in Falsificationism," Kluge questioned our conclusion about the relationship between corroboration and likelihood in phylogenetic analysis. According to Kluge (2001:326), "de Queiroz and Poe's conjecture that $p(e, hb) = C(h,e,b)$ is false, because there is no reference to critical evidence," which reiterated his earlier statement (2001:322) that "de Queiroz and Poe interpret degree of corroboration *solely* as a likelihood argument—no importance is attributed to *critical* evidence." Kluge's criticism confuses the general concept of corroboration with its specific application to the problem of identifying the most highly corroborated phylogenetic tree for a given data set and phylogenetic method, here and previously (de Queiroz and Poe, 2001: appendix) referred to as a *standard phylogenetic analysis*.

Kluge is incorrect in stating that we equated degree of corroboration $C(h, e, b)$ with likelihood $p(e|hb)$. As we noted previously (de Queiroz and Poe, 2001), Popper considered his degree of corroboration $C(h, e, b)$ to be a generalization of likelihood. We also noted that degree of corroboration differs from likelihood in containing both a normalization factor (the denominator) and the additional term $p(e|b)$. Our proposition was not that degree of corroboration $C(h, e, b)$ is generally equivalent to likelihood $p(e|hb)$ but only that no component of a standard phylogenetic analysis, whether based on parsimony or likelihood, corresponds with the term $p(e|b)$ (see also Faith and Cranston, 1992; Faith and Trueman, 2001). Therefore, in such an analysis (as opposed to science or even phylogenetics generally), the relative degree of corroboration of the alternative hypotheses (trees) is determined solely by $p(e|hb)$, which corresponds with likelihood, as Kluge himself acknowledged (see next section).

This conclusion is consistent with Popper's views on the differences between degree of corroboration and likelihood. Thus, as we noted previously (de Queiroz and Poe, 2001), Popper (1959:413) stated that under certain circumstances—the example he considered involved very small $p(e|b)$—likelihood would be approximately equivalent to and thus an adequate measure of degree of corroboration. We proposed that standard phylogenetic analysis represents another case in which likelihood is equivalent to (at least relative) degree of corroboration, though not because $p(e|b)$ is necessarily small but simply because it is not assessed. Kluge himself (2001:327) accepted this proposition, for he acknowledged that "for any *particular* matrix of evidence e, maximizing the likelihood $p(e, hb)$ also maximizes corroboration $C(h, e, b)$, and likelihood can be used to select among cladograms (de Queiroz and Poe, 2001)." In any case, we did not propose

a general equivalency between degree of corroboration and likelihood, and therefore Kluge criticized a misrepresentation of our views.

Critical Evidence and Test Severity

Kluge's erroneous interpretation of our statement concerning the relationship between corroboration and likelihood appears to have been based on his belief that under our interpretation, "no importance is attributed to *critical* evidence" (Kluge, 2001:322). Critical evidence is related to test severity, which is assessed using the term $p(e|b)$, the probability of the evidence given the background knowledge alone, in Popper's corroboration equation. Kluge is wrong in stating that we attribute no importance to critical evidence. We devoted a lengthy discussion to this topic (de Queiroz and Poe, 2001: appendix), including suggestions about how $p(e|b)$ or analogous probabilities could be assessed. Moreover, regardless of what one thinks about the importance of $p(e|b)$, standard phylogenetic analyses do not assess the value of that term.

Contrary to the implications of Kluge's criticism, this statement is not a proposition about the importance of critical evidence; it is only a description of the properties of standard phylogenetic analyses. That is, we did not propose that one ought to ignore $p(e|b)$ or that critical evidence is unimportant; we proposed only that there is no component of a standard phylogenetic analysis, whether based on parsimony or likelihood, that corresponds with the determination of $p(e|b)$. Such analyses simply do not assess the critical nature of the evidence. Kluge himself (2001:327) accepted this proposition when he acknowledged that "for any *particular* matrix of evidence e, maximizing the likelihood $p(e, hb)$ [i.e., as opposed to maximizing $p(e|hb) - p(e|b)$] also maximizes corroboration $C(h, e, b)$, and likelihood can be used to select among cladograms (de Queiroz and Poe, 2001)." An analysis based on a particular matrix of evidence e and a particular phylogenetic method is precisely what we were referring to when we proposed that standard phylogenetic analyses contain no component that corresponds with $p(e|b)$. If this means that standard phylogenetic analyses do not conform precisely to Popper's degree of corroboration, then this conclusion applies equally to both parsimony and likelihood methods.

We argued previously (de Queiroz and Poe, 2001: appendix) not only that standard phylogenetic analyses contain no component corresponding with $p(e|b)$ but also that for technical reasons $p(e|b)$ cannot be calculated in such analyses. Nevertheless, we also argued that it is possible (1) to calculate $p(e|b)$ in other sorts of phylogenetic analyses (specifically, tests of model assumptions) and (2) to calculate probabilities analogous to $p(e|b)$ in standard phylogenetic analyses. We also argued that a consideration of $p(e|b)$ and analogous probabilities leads to the common sense conclusion that the severity of a test, which is inversely related to $p(e|b)$, increases with increasing amounts of relevant data (evidence e). Kluge presented no argument contradicting these propositions; instead, he merely asserted (2001:327) that "traditional character reanalysis is the most common basis for increasing the severity of test. This is what Hennig (1966) referred to as reciprocal clarification, and which elsewhere has been conceptualized as a never-ending cycle of research (Kluge, 1997[a])."

Kluge's interpretation of critical evidence is dubious. Regardless of whether one adopts the optimality criterion of parsimony or likelihood, traditional character reanalysis has limited potential for increasing the severity of a test. Reinterpreting and rescoring some fraction of the original characters is unlikely to have a substantial effect on the value of $p(e|b)$, particularly if it does not greatly increase or decrease the number of characters. In contrast, character reanalysis could have a substantial effect on the values of $p(e|hb)$ if it were to greatly increase or decrease congruence between the taxon × character matrix e and particular trees h; these values, however, do not assess test severity or critical evidence. Moreover, it is not at all clear that the value of $p(e|b)$ would be expected to decrease as the result of character reanalysis, as it should if reanalysis led to a more severe test. If character reanalysis resulted in an increase in the number of characters, test severity should increase, but the number of characters could just as well decrease or remain the same.

In contrast, as we argued previously (de Queiroz and Poe, 2001), the addition of data is expected to have a direct and predictable effect on test severity: it should result in a decrease in the value of $p(e|b)$ or analogous probabilities, thus indicating a more severe test. This interpretation is supported by Popper's (e.g., 1959:413) statement that small $p(e|b)$ "is possible only for large samples" and the interpretation of Farris et al. (2001:442), who stated that $p(e|b)$ "depends only on the sample size n." Thus, although neither parsimony nor likelihood methods include a component that assesses test severity (critical evidence) when used in standard phylogenetic analyses, both classes of methods are compatible with the use of severe tests.

Kluge's equation of critical evidence with character reanalysis undermines his own proposition that there is a fundamental difference between parsimony and likelihood methods regarding their conformity with Popper's concept of corroboration, because characters can be reanalyzed under likelihood as well as they can under parsimony. The same conclusion holds if critical evidence is represented by a large body of relevant data, which can be collected regardless of the optimality criterion under which it is to be analyzed. Thus, regardless of whether one accepts our interpretation of critical evidence or Kluge's, that concept lends no credence to his proposition of a fundamental difference between parsimony and likelihood methods in terms of their conformity with Popper's degree of corroboration.

Corroboration and Parsimony

In his section titled a "Popperian Testability, Phylogenetic Systematics, and Parsimony," Kluge (2001)

reiterated his previous (Kluge, 1997b) interpretation of how cladistic parsimony methods conform to Popper's degree of corroboration. That interpretation consists of two main propositions: (1) that for data in which taxa A and B share more derived character states with one another than either does with taxon C, degree of corroboration *C* and therefore $p(e|hb)$ is maximized for the hypothesis (A,B)C and (2) that the first proposition follows "given only descent with modification as background knowledge" (Kluge, 1997b:88). We do not dispute that parsimony methods conform to Popper's *C*, a conclusion we reached previously (de Queiroz and Poe, 2001) under the provision that those methods are interpreted as incorporating (implicit) probabilistic assumptions. That is, we do not dispute Kluge's first proposition, given the appropriate probabilistic assumptions. What we dispute is his second proposition, the view that the first proposition can be justified without invoking probabilistic assumptions.

We previously presented an explicit refutation of the second proposition, explaining why probabilistic assumptions are necessary to reach the conclusion in question (de Queiroz and Poe, 2001:313–315). We argued (following Felsenstein, 1978) that depending on the probabilities of character transformation on the branches subtending taxa A, B, and C, data for which A and B share more derived states with one another than either does with C have a higher probability $p(e|hb)$ on some trees of topology (A,C)B and (B,C)A than on some trees of topology (A,B)C. Therefore, assumptions about the probabilities of character changes on the branches of the trees are necessary to conclude that $p(e|hb)$, and thus corroboration, is maximized for topology (A,B)C. Kluge presented no counterargument to our explicit refutation of his proposition; instead, he simply restated his proposition (2001:326).

However, Kluge has been inconsistent on this issue. In contrast with his earlier papers (Kluge, 1997a, 1997b; Siddall and Kluge, 1997), Kluge acknowledged that the term $p(e|hb)$ in the corroboration expression corresponds to likelihood (2001:324, 326, 327) and that for a particular body of evidence *e*, maximizing likelihood $p(e|hb)$ also maximizes corroboration *C* (2001:327). Moreover, he acknowledged that "parsimony does not directly evaluate the likelihood probability $p(e, hb)$" (2001:327). He attempted to solve this problem in two ways: (1) by appealing to Farris's (1989:107) statement that "a postulate of homology explains similarities among taxa as inheritance, while one of homoplasy requires that similarities be dismissed as coincidental, so that most parsimonious arrangements have greatest explanatory power" and (2) by invoking Tuffley and Steel's (1997) conclusion that the parsimony method is equivalent to a maximum likelihood model with "no common mechanism," which can then be used to calculate an actual value for $p(e|hb)$ and therefore *C*.

Appealing to Farris's (1989) statement does not solve but merely puts off Kluge's problem of how to calculate a value for $p(e|hb)$ from a parsimony method. The reason is that Farris's justification for parsimony is based on explanatory power, Popper's *E*, which is itself defined in terms of likelihood $p(e|hb)$ (see Kluge, 2001:324). Moreover, Farris's conclusion rests on the questionable proposition that attributing similarities to inheritance explains those similarities whereas attributing them to homoplasy does not count as an explanation but only as a dismissal. On the contrary, both postulated homologies and homoplasies are explanatory in accounting for the occurrences of character states in taxa. Hypotheses of homoplasy may be unparsimonious and ad hoc (under parsimony models), but that does not make them nonexplanatory.

Invoking Tuffly and Steel's (1997) likelihood model as equivalent to parsimony does indeed solve Kluge's problem of how to calculate a value for $p(e|hb)$; however, in so doing it contradicts his proposition that assumptions in addition to descent with modification—in particular, probabilistic assumptions—are not needed to interpret parsimony as a method for assessing Popper's *C*. Tuffley and Steel's model consists of several assumptions in addition to descent with modification, at least some of which are explicitly probabilistic. In particular, their model incorporates a mutation probability p_e, the probability of a net change of state occurring along an edge (branch), and it assumes that this probability is equal for all possible state changes. By invoking this model as equivalent to cladistic parsimony so that he can calculate a value for *C*, Kluge effectively refuted his own proposition that descent with modification is sufficient background knowledge for interpreting cladistic parsimony as example of Popperian corroboration.

Moreover, given that Kluge criticized particular assumptions of standard likelihood models, the assumptions of Tuffley and Steel's (1997) parsimony-equivalent model are not immune to criticism. (Tuffley and Steel were well aware of the limitations of this model, which they developed not to justify parsimony methods but to explore the relationship between those methods and standard likelihood methods.) One such criticism voiced from the perspective of likelihood exemplifies a view expressed by Popper and adopted by advocates of cladistic parsimony (see next section). According to Popper (1962:61), "the methodology of science (and the history of science also), becomes understandable in its details if we assume that the aim of science is to get explanatory theories which are as little *ad hoc* as possible: a 'good' theory is not *ad hoc*, while a 'bad' theory is." Similarly, "*ad hoc* hypotheses are disliked by scientists: they are, at best, stop-gaps, not real aims" (Popper, 1962:287), and with specific reference to *C*, "my definition [of *C*] can be shown to give most reasonable results if combined with a rule excluding *ad hoc* hypotheses" (Popper, 1962:288).

A fundamental difference between standard phylogenetic likelihood models and the parsimony-equivalent model of Tuffley and Steel (1997) is the assumption in the latter that Tuffley and Steel referred to as "no common mechanism." Under this assumption each character is allowed to have "a different vector of mutation probabilities" (Tuffley and Steel, 1997:597), as opposed

to all characters having "a single vector of mutation probabilities" (Tuffley and steel, 1997:597), as in standard likelihood models. That is, instead of the change probabilities for characters on a given branch being relatively high or low across characters (i.e., relative to the change probabilities for those same characters on other branches), the change probabilities can be relatively high for one character but low for another on a given branch. In effect, each character has its own ad hoc set of branch lengths. The Tuffley and Steel model with no common mechanism is therefore highly ad hoc, and given that this model is equivalent to cladistic parsimony, the parsimony model may not give the best results as a method for assessing degree of corroboration. This criticism of parsimony methods is related to one raised from the statistical perspective of likelihood, in which parsimony methods are characterized as overparameterized models because each different character vector constitutes a separate parameter (e.g., Tuffley and Steel, 1997; Lewis, 1998; Steel and Penny, 2000).

By accepting the equivalency between cladistic parsimony and a likelihood model with no common mechanism, Kluge contradicted several of his own views in addition to the idea that parsimony does not involve implicit probabilistic assumptions. Those propositions include (1) the view that parsimony does not include an evolutionary model, (2) the idea that parsimony but not likelihood can be interpreted as a method for assessing degree of corroboration, (3) the proposition that corroboration does not use numerical probabilities, and (4) the view that likelihood methods are verificationist (unless that criticism also applies to parsimony).

Falsification, Corroboration, and Phylogenetic Analysis

Popper's proscription against ad hoc hypotheses has also been used by advocates of cladistic parsimony as the basis of a dubious justification for those methods. Thus, in his section "Popperian Testability, Phylogenetic Systematics, and Parsimony," Kluge stated "as a rule of methodological falsification in the evaluation of scientific hypotheses, the cladogram(s) h that requires the ad hoc dismissal of the fewest falsifiers is preferred" (2001:326) and that "most-parsimonious cladograms . . . are least refuted" (2001:327). These statements are holdovers from an older interpretation (e.g., Wiley, 1975; Gaffney, 1979; Farris, 1983) of how cladistic parsimony methods conform with Popper's falsificationist philosophy, an interpretation that we argue is seriously flawed because it rests on a questionable assumption not required by more recent interpretations based on Popper's degree of corroboration (e.g., Faith, 1992; Faith and Cranston, 1992; Farris, 1995; Kluge, 1997a, 1997b, 2001; de Queiroz and Poe, 2001; Faith and Trueman, 2001; Farris et al., 2001). Under the older interpretation, characters that are incongruent with a particular phylogenetic hypothesis are viewed as falsifiers of that hypothesis, and cladistic parsimony methods are said to be justified on the grounds that they minimize the ad hoc dismissal of falsifiers.

The problem with this interpretation is that under standard parsimony (and likelihood) models incongruent characters do not qualify as falsifiers of phylogenetic hypotheses (see also Sober, 1988). To understand why this is true, consider the logical interpretation of probabilities, according to which probability is defined as the degree of a logical relation between statements (Popper, 1959:148–149, 320; 1962:59). Under this interpretation, the conditional probability $p(a|b)$ equals 1 if a is a logical consequence of b (i.e., if a follows from b), and it equals 0 if the negation of a is a logical consequence of b (i.e., if a is prohibited by b) (Popper 1959:320, 405). To give a simple example, the logical probability of obtaining heads in a coin toss given a two-headed coin is 1, and the logical probability of obtaining tails given such a coin is 0. The second proposition bears directly on the concept of falsification; as Popper (1983:242) noted in describing the consequences of his definition of degree of corroboration, "empirical evidence e which *falsifies* h in the presence of b . . . will make $p(e, hb)$ equal to zero." In nonnumerical terms, "natural laws might be compared to 'proscriptions' or 'prohibitions'. . . . They insist on the non-existence of certain things or states of affairs: they rule them out. And it is precisely because they do this that they are *falsifiable*" (Popper, 1959:69). Thus, the hypothesis that a particular coin has two heads rules out the possibility of tails and is falsified if tails is obtained.

The reason that incongruent characters do not qualify as falsifiers of phylogenetic hypotheses under standard parsimony and likelihood models is that there is no character for which $p(e|hb) = 0$. Under these models, there is no distribution of states among taxa (e) that is logically prohibited (i.e., ruled out) by any given tree topology (h). In the case of parsimony, this conclusion derives from the fact that although some characters (state distribution patterns) may be incongruent with particular topologies in that they require extra steps (ad hoc hypotheses of homoplasy), those characters can nonetheless be accounted for on the topologies in question by invoking those extra steps. Consequently, no topology rules out the possibility that even the most incongruent character could have evolved on that topology. In the case of likelihood, the same conclusion holds in an explicitly probabilistic form. Thus, although some characters may have very low probabilities on particular topologies, those characters nonetheless have finite, positive probabilities on those topologies. No character has a probability of 0 on any topology. The only way to legitimately interpret incongruent characters as falsifiers of phylogenetic hypotheses is to adopt an extremely unrealistic evolutionary model that prohibits the occurrence of homoplasy.

The same conclusion applies to sets of characters. There is no tree for which it is impossible to account for the state distributions of any set of characters under either parsimony or likelihood (which is not to deny that some trees account for the data better than others), and therefore there is no phylogenetic hypothesis for which any relevant taxon × character matrix has a probability of 0. Under methods that permit homoplasy, phylogenetic hypotheses cannot be falsified by character data, and

this fact undermines Kluge's attempt to justify cladistic parsimony as Popperian (falsificationist) on the grounds that it minimizes the ad hoc dismissal of falsifiers.

The conclusion that phylogenetic hypotheses (trees) cannot be falsified by characters should not be taken to imply that those hypotheses are unscientific according to Popper's general falsificationist philosophy. Popper (1959:189–205) pointed out that most probability hypotheses are not falsifiable in a strict logical sense. For example, no fraction of tails is ruled out by the hypothesis $p_{\text{heads}} = 0.9$ (although some fractions are highly improbable given that hypothesis). Nevertheless, Popper also noted that science (specifically physics) has achieved great success with probability hypotheses (1959:190), that scientists are well able to decide whether a particular probability hypothesis ought to be rejected as "practically falsified" (1959:191), and that this practice can be defended philosophically (1959:199–205). Thus, as long as a hypothesis has the potential to be tested empirically and to be considered contradicted if certain results are obtained, it is scientific according the Popper's general refutationist or falsificationist philosophy.

Phylogenetic hypotheses are unfalsifiable in the same logical sense that most probability hypotheses are unfalsifiable, that is, they do not logically forbid any particular test result (body of relevant data). In such cases, probabilistic concepts such as likelihood, degree of corroboration, and statistical significance play central roles in the evaluation of rival hypotheses. In this context, the shift from interpretations of phylogenetic analysis as a method for identifying the least falsified hypotheses to those that view phylogenetic analysis as a method for identifying the most highly corroborated hypotheses is very appropriate. Because of its probabilistic basis, Popper's degree of corroboration is applicable to hypotheses that are not logically falsifiable, and it therefore provides a more appropriate context for understanding how methods of phylogenetic analysis, including those based on both parsimony and likelihood, conform to Popper's philosophy of science.

Likelihood Ratio Tests

In his section titled "Likelihood Ratio Test," Kluge (2001:327) stated that "de Queiroz and Poe . . . argue the *statistical* credibility of maximum likelihood in terms of the likelihood ratio test" and that "the likelihood ratio test continues to be cited as the ampliative basis for maximum likelihood." These statements are false. We said very little about likelihood ratio tests, except that they can be used to test hypotheses that can be represented as parameters of evolutionary models (such as those used in phylogenetic likelihood methods) and that, as significance tests, they incorporate considerations about test severity. We certainly did not propose that the law of likelihood, the proposition that the hypothesis with the highest likelihood is best corroborated by the data (e.g., Edwards, 1972), derives its credibility from likelihood ratio tests. Nor are we aware of anyone who has championed likelihood ratio tests as the ampliative basis for maximum likelihood. On the contrary, likelihood itself is a general statistical principle upon which significance tests (including likelihood ratio tests) are based, and likelihood ratio tests derive their justification from the law of likelihood rather than vice versa (see Edwards, 1972:176).

The remainder of Kluge's criticism of likelihood ratio tests is equally misguided. For example, he criticized those tests on the grounds that they do "not actually test for goodness of fit" but only "how much better the fit is among alternative models," which he considered a problem because "the better-fitting hypothesis does not necessarily provide a significantly good fit" (2001:327). Although Kluge is correct if he means that rejection of the null hypothesis in a likelihood ratio test of a model assumption does not demonstrate the correctness or adequacy of the alternative model (which is only one of many possible models), this observation does nothing to bolster his case for the superiority of parsimony methods. For one thing, rejection of the null hypothesis provides effective falsification of the corresponding model assumption (e.g., equal transition [Ti] and transversion [Tv] probabilities) and a compelling reason for choosing a more realistic model in the context of likelihood. In contrast, parsimony methods provide no means for assessing the superiority of one assumption relative to another (e.g., equal versus different Ti and Tv costs), let alone for demonstrating that one of them is correct or adequate. For another, the simple parsimony model preferred by Kluge (e.g., 1997a) involves highly restrictive assumptions (e.g., equal Ti and Tv costs) similar to those (e.g., equal Ti and Tv probabilities) that are commonly rejected by likelihood ratio tests. Furthermore, simulation studies have shown that this simple parsimony model does a poor job of reconstructing the correct tree relative to likelihood under more complex models as the assumed model becomes more complex and realistic (e.g., Yang, 1996a).

Kluge (2001:327) also complained that "the [likelihood ratio] test can never be anything but an indefensible optimality criterion, because the assumptions of the model are contingencies that require testing outside the model itself." Thus, he apparently failed to appreciate the fact that when particular parameters are tested using likelihood ratio tests, they are indeed tested outside of the model itself. That is, when any proposition is the subject of a test, it is not part of the model but rather it is part of the hypothesis h of both corroboration C and likelihood L (de Querioz and Poe, 2001:"Evaluation of Alternative Phylogenetic Methods or Models").

Another of Kluge's criticisms is that "the likelihood ratio test is not empirical and must not be confused with the nature of critical evidence" (2001:327). Why Kluge thinks there is a danger of confusing a type of significance test with a property of the data is unclear. In any case, if he believes that likelihood ratio tests ignore the issues of test severity and critical evidence, he is wrong. As we argued previously (de Queiroz and Poe, 2001:319), significance tests (including likelihood ratio tests) incorporate considerations about test severity and critical evidence through the concept of power, which like test severity

increases with increasing sample size. Rejection of a null hypothesis is possible only with an adequate sample size, that is, with critical evidence and a sufficiently severe (powerful) test.

Kluge considered likelihood ratio tests to "lack empirical independence" (2001:327) and thus "to have nothing to say about causal hypotheses that is not confounded by assuming what is at issue in the argument" (2001:328). These statements reflect a simplistic view of how hypotheses about evolutionary processes are tested in a phylogenetic context. Such hypotheses are not tested, at least not severely, by performing a phylogenetic analysis and character optimization under minimal assumptions and subsequently making an inference about an evolutionary process from reconstructed character transformations on the optimal tree. As we have argued here (see "Assumptions of Parsimony and Likelihood Methods") and previously (de Queiroz and Poe, 2001), the notion that parsimony methods (whether for tree reconstruction or character optimization) are based on minimal assumptions (i.e., only descent with modification) is erroneous. Moreover, although such an approach can identify (under its assumptions) a best-fit hypothesis of character evolution, it offers limited potential for evaluating whether that hypothesis provides, in Kluge's own words, a "significantly good fit," that is, whether it explains the data significantly better than does an alternative hypothesis.

A more rigorous approach is to compare evolutionary models that differ only in the presence versus absence of a parameter corresponding with the hypothesis of interest in terms of their ability to account for the data. In such a comparison, the parameter that distinguishes the more general model from the less general one corresponds to Popper's h, the less general model (the one lacking the parameter corresponding to h) corresponds to Popper's b, and the more general model (the one incorporating the parameter corresponding with h) corresponds to hb (de Queiroz and Poe, 2001:319). Thus, for a given body of evidence e, $p(e|hb)$ of Popper's corroboration expression is the likelihood of the more general model (i.e., the probability of the evidence given that model), and $p(e|b)$ is the likelihood of the less general model. Degree of corroboration C is based on the difference between $p(e|hb)$ and $p(e|b)$, so these probabilities can be used to calculate the degree of corroboration C of the hypothesis h from Popper's equation. The value of C, however, will simply be a number between -1 and $+1$, and it is not clear what (positive) values (other than 1) constitute a significant degree of corroboration.

Significance tests provide a general method for this kind of evaluation. A likelihood ratio test is a significance test that uses as its test statistic the ratio of the likelihoods of the two models (Edwards, 1972), which in the case of nested (more and less general) models correspond to null and alternative hypotheses. Thus, $\Lambda = L(h_{\text{null}}|e)/L(h_{\text{alternative}}|e)$ (Huelsenbeck and Rannala, 1997). Here, the null hypothesis corresponds to the less general model or Popper's b, and the alternative hypothesis corresponds to the more general model or Popper's hb. Therefore $\Lambda = p(e|b)/p(e|hb)$. The likelihood ratio Λ is thus fundamentally similar to degree of corroboration C. It is based on the same two probabilities, $p(e|hb)$ and $p(e|b)$, that form the basis of C and differs primarily in being calculated as the ratio rather than the difference between these two probabilities.

For nested models, $-2 \log \Lambda$ is approximately χ^2 distributed under the null hypothesis with degrees of freedom equal to the difference in the number of free parameters between the models (Huelsenbeck and Rannala, 1997). Therefore, to assess the statistical significance of a particular result, the probability can be looked up in a standard table. Alternatively, the probability of a given value of Λ under the null hypothesis can be calculated using Monte Carlo simulations (Goldman, 1993; Huelsenbeck and Rannala, 1997). Thus, likelihood ratio tests of nested evolutionary models are simply a means of evaluating the significance, under an accepted statistical criterion, of the degree of corroboration of a hypothesis h. Contrary to Kluge's view, the use of such a test does not represent "assuming what is at issue in the argument," because there is no a priori reason to believe that the null hypothesis will (or will not) be rejected by the test.

Kluge "Replies" to de Queiroz and Poe

In his section "Popper Replies" to de Queiroz and Poe, Kluge (2001:328) quoted a long passage from Popper (1959) "to let Popper speak more fully on the merits of de Queiroz and Poe's claim that $p(e, hb) = C(e, h, b)$" and to show that Popper was "neither a verificationist nor an inductionist." There is little point in commenting extensively on this passage, which contradicts neither our interpretations of Popper's philosophy nor our propositions about its relevance to methods of phylogenetic inference. Concerning the points that Kluge used it to demonstrate, we did not propose a general equivalence between corroboration $C(h, e, b)$ and likelihood $p(e|hb)$, nor did we consider Popper a verificationist or an inductionist. Kluge's first point is a straw man that was not proposed by us but instead stems from his own failure to distinguish between the general concepts of corroboration and likelihood and the specific application of those concepts to the problem of identifying a most highly corroborated phylogenetic tree using a single analytical method and body of evidence. Kluge's second point also misrepresents our views and stems from his false premise that likelihood, because it is probabilistic, is necessarily a form of verificationism.

In his "Summary," Kluge reiterated these points as his two main "disagreements with de Queiroz and Poe" (2001:328). He proposed that "$L(h, e) \neq C(h, e, b)$" (2001:328) and that "likelihood inference . . . is, by definition, verificationist" (2001:329). We will not reiterate our arguments against these points (see "Corroboration and Likelihood" and "Falsificationism and Verificationism") or against his view that part of the reason that $L \neq C$ is that "M includes counterfactual assumptions, which are unlikely to be true and, which are excluded from b"

(see "Background Knowledge"). Instead, we only comment on two related points that we have not already refuted.

First, Kluge (2001:329) proposed that the other reason that $L \neq C$ is "because $L(h, e)$ maximizes the likelihood function with *non*-critical evidence, just $p(e|M, h)$, whereas $C(h, e, b)$ maximizes the likelihood function with *critical* evidence, $p(e, hb) - p(e, b)$." It is incorrect to suggest that likelihood uses noncritical evidence and corroboration uses critical evidence; analyses conducted under both approaches can use either critical or noncritical evidence. The difference is that corroboration attempts to assess whether the evidence is critical (i.e., the severity of the test), whereas maximum likelihood by itself does not. However, likelihood ratios of nested models and significance tests based on them do incorporate assessments of critical evidence.

In addition, this difference has no bearing on the comparison of phylogenetic parsimony and likelihood methods. The evidence contained in a given body of data is either critical or noncritical, and its status as such has more to do with its quality and quantity than with the methods by which it is analyzed (which is not to deny that those methods may differ with regard to statistical power or the hypotheses that they identify as most highly corroborated). Critical evidence does not become noncritical just because it is analyzed with likelihood methods, nor does noncritical evidence become critical just because it is analyzed with parsimony methods. Moreover, in standard phylogenetic analyses, neither parsimony nor likelihood methods assess whether the evidence is critical, though such assessments can be made for data to be analyzed using either class of methods. Thus, phylogenetic parsimony and likelihood methods are equally consistent and/or inconsistent with Popper's degree of corroboration with regard to assessing the critical nature of the evidence.

Second, Kluge proposed that the reason that likelihood is verificationist is that "the truth is sought inductively with degrees of probability," and he suggested that Popper shared this view (Kluge, 2001:329). On the contrary, Popper never said that likelihood is verificationist; this view is entirely Kluge's. Popper used likelihood as the basis of his own degree of corroboration (e.g., Popper, 1959:388; 1983:252); he explicitly acknowledged that likelihood, like degree of corroboration, does not satisfy the rules of the calculus of probability (Popper, 1983:243; see also Edwards, 1972:12), and he noted that in certain cases likelihood approximates degree of corroboration (Popper, 1959:413). Moreover, likelihood methods are no more inductive than those based on parsimony. Both classes of methods are inductive not only in the general sense that they are used to make inferences from specific observations but also in the specific sense that those inferences (particular trees) go beyond (and thus do not follow necessarily from) the observations upon which they are based (see also Bryant, 1989). On the other hand, both parsimony and likelihood are deductive in their calculation of tree scores, that is, both parsimony lengths and likelihood probabilities follow necessarily from the observations (data), the hypothesis (a particular tree), and the background knowledge (properties of the method/model). In addition, both classes of methods use these deduced tree scores, whether lengths or probabilities, to evaluate the relative degrees of corroboration of alternative trees. Kluge's idea that likelihood differs from parsimony in being inductive is groundless and seems to be based on his continued failure to distinguish between probabilistic methods in general, some of which are used to evaluate hypotheses using deductively established probabilities of the evidence $p(e|h)$ or $p(e|hb)$, and those that are used to assign probabilities to hypotheses $p(h)$ or $p(h|e)$ in an attempt to circumvent the impossibility of inductively establishing their certain truth (see Popper, 1983).

Conclusion

Contrary to Kluge's claims, he failed to refute any of our propositions either about the relationship between Karl Popper's concept of corroboration and Ronald Fisher's statistical principle of likelihood or about the relationship between Popper's C and parsimony and likelihood methods of phylogenetic analysis. In particular, Kluge presented no evidence or counterarguments to refute any of the main propositions of our original paper: (1) that Popper's degree of corroboration is based on Fisher's likelihood, (2) that the relative degrees of corroboration for alternative phylogenetic trees analyzed using a single phylogenetic method and a single body of evidence is determined solely by their likelihoods, and (3) that parsimony methods are interpretable as methods for assessing degree of corroboration only if they are viewed as incorporating implicit probabilistic assumptions. Instead, despite purporting to refute these propositions, Kluge implicitly endorsed all three of them (albeit inconsistently) by accepting (1) that the term $p(e|hb)$ in Popper's corroboration equation is equivalent to likelihood, (2) that for any particular character matrix e, maximizing likelihood $p(e|hb)$ maximizes corroboration $C(h, e, b)$, and (3) that likelihood under Tuffley and Steel's (1997) probabilistic model with no common mechanism is equivalent to parsimony. Moreover, by accepting these propositions, Kluge contradicted his own propositions that likelihood methods are incompatible with Popper's degree of corroboration and that descent with modification is sufficient background knowledge for justifying cladistic parsimony as a method for assessing degree of corroboration.

Kluge's explicit arguments also fail to refute any of our propositions, either because these arguments rest on false premises (e.g., that likelihood is verificationist), because they rely on misrepresentations of Popper's views (e.g., that background knowledge consists only of well-corroborated hypotheses and thus should be minimized), or because they refute propositions that we did not make (e.g., that likelihood is identical to degree of corroboration). Kluge's critique seems to rest largely on two

erroneous conjectures that we had previously refuted (de Queiroz and Poe, 2001) and for which Kluge provided no counterarguments: (1) the erroneous proposition that descent with modification is sufficient background knowledge for parsimony analysis as a method for assessing degree of corroboration and (2) the false assumption that likelihood is intrinsically inductive and/or verificationist simply because it is based on probabilities.

In sum, Kluge failed to refute any of our propositions concerning the relationship between Popper's degree of corroboration and alternative methods of phylogenetic analysis. Both parsimony and likelihood methods can be used to assess degree of corroboration; that is, both are consistent with the philosophical criteria that Kluge supposes to justify only parsimony methods. Likelihood methods conform with Popper's degree of corroboration, which is based on likelihood, without violating any fundamental views held by advocates of those methods. In contrast, contrary to the views of Kluge and others, parsimony methods conform with Popper's degree of corroboration only if they are interpreted as carrying implicit probabilistic assumptions. Likelihood methods have several advantages over their parsimony-based counterparts in the context of Popper's philosophy, including the potential for testing their assumptions and their avoidance of ad hoc hypotheses. Phylogenetic likelihood methods are well justified in terms of both performance and philosophy, which accounts for their increasing use by systematic biologists.

ACKNOWLEDGMENTS

We thank D. Swofford for helpful discussions and J. Huelsenbeck for pointing out the historical nature of coin tosses. D. Faith, M. Sanderson, J. Trueman, and an anonymous reviewer provided insightful and constructive criticisms on an earlier version of this paper. We also thank M. Thollesson, whom we forgot to acknowledge previously, for comments on our previous paper.

REFERENCES

BRYANT, H. 1989. An evaluation of cladistic and character analyses as hypothetico-deductive procedures, and the consequences for character weighting. Syst. Zool. 38:214–227.

COLE, C. J., H. C. DESSAUER, AND G. BARROWCLOUGH. 1988. Hybrid origin of a unisexual species of whiptail lizard, *Cnemidophorus neomexicanus*, in western North America: New evidence and a review. Am. Mus. Novit. 2905:1–38.

DE QUEIROZ, K., AND S. POE. 2001. Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. Syst. Biol. 50:305–321.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge Univ. Press, Cambridge, U.K. [Expanded edition published by Johns Hopkins Univ. Press, Baltimore, Maryland, 1992.]

FAITH, D. P. 1992. On corroboration: A reply to Carpenter. Cladistics 8:265–273.

FAITH, D. P., AND P. S. CRANSTON. 1992. Probability, parsimony, and Popper. Syst. Biol. 41:252–257.

FAITH, D. P., AND J. W. H. TRUEMAN. 2001. Towards an inclusive philosophy of phylogenetic inference. Syst. Biol. 50:331–350.

FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics, Volume 2 (N. I. Platnick and V. A. Funk, eds.). Columbia Univ. Press, New York.

FARRIS, J. S. 1989. Entropy and fruit flies. Cladistics 5:103–108.

FARRIS, J. S. 1995. Conjectures and refutations. Cladistics 11: 105–118.

FARRIS, J. S., A. G. KLUGE, AND J. M. CARPENTER. 2001. Popper and likelihood versus "Popper.*" Syst. Biol. 50:438–444.

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

GAFFNEY, E. S. 1979. An introduction to the logic of phylogeny reconstruction. Pages 79–111 *in* Phylogenetic analysis and paleontology (J. Cracraft and N. Eldredge, eds.). Columbia Univ. Press, New York.

GOLDMAN, N. 1993. Statistical tests of models of DNA sequence evolution. J. Mol. Evol. 36:182–198.

HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

HENNIG, W. 1966. Phylogenetic systematics. Univ. Illinois Press, Urbana.

HUELSENBECK, J. P., AND B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232.

KLUGE, A. G. 1997a. Sophisticated falsification and research cycles: Consequences for differential character weighting in phylogenetic systematics. Zool. Scr. 26:349–360.

KLUGE, A. G. 1997b. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13:81–96.

KLUGE, A. G. 2001. Philosophical conjectures and their refutation. Syst. Biol. 50:322–330.

LAKATOS, I. 1970. Falsification and the methodology of scientific research programmes. Pages 91–196 *in* Criticism and the growth of knowledge (I. Lakatos and A. Musgrave, eds.). Cambridge Univ. Press, Cambridge, U.K.

LAWRENCE, J. G. 2001. Catalyzing bacterial speciation: Correlating lateral transfer with genetic headroom. Syst. Biol. 50:479–496.

LEWIS, P. O. 1998. Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. Pages 132–163 *in* Molecular systematics of plants. II. DNA sequencing (D. E. Soltis, P. S. Soltis, and J. J. Doyle, eds.). Kluwer Academic, Boston.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, AND D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

MCDADE, L. 1990. Hybrids and phylogenetic systematics. I. Patterns of character expression in hybrids and their implications for cladistic analysis. Evolution 44:1685–1700.

POPPER, K. R. 1959. The logic of scientific discovery. Basic Books, New York.

POPPER, K. R. 1962. Conjectures and refutations. Basic Books, New York. [2nd edition published by Harper and Row, New York, 1968.]

POPPER, K. R. 1983. Postscript to the logic of scientific discovery, Volume 1. Realism and the aim of science, Routledge, London.

SANDERSON, M. J. 1995. Objections to bootstrapping phylogenies: A critique. Syst. Biol. 44:299–320.

SIDDALL, M. E., AND A. G. KLUGE. 1997. Probabilism and phylogenetic inference. Cladistics 13:313–336.

SOBER, E. 1988. Reconstructing the past: Parsimony, evolution, and inference. MIT Press, Cambridge, Massachusetts.

STEEL, M., AND D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

SWOFFORD, D. L., AND G. J. OLSEN. 1990. Phylogeny reconstruction. Pages 411–501 *in* Molecular systematics, 1st edition (D. M. Hillis and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, AND J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

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

WILEY, E. O. 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. 1996a. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

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

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.

First submitted 17 June 2002; reviews returned 4 September 2002; final acceptance 3 February 2003

Associate Editor: Mike Sanderson