Syst. Biol. 50(3):305–321, 2001

Philosophy and Phylogenetic Inference: A Comparison of Likelihood and Parsimony Methods in the Context of Karl Popper's Writings on Corroboration

Kevin de Queiroz[1] and Steven Poe[1,2,3]

[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] *Department of Zoology and Texas Memorial Museum, University of Texas, Austin, Texas 78712-1064, USA; E-mail: stevepoe@mail.utexas.edu*

[3] *Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720-3160; E-mail: stevepoe@uclink.berkeley.edu*

Abstract.—Advocates of cladistic parsimony methods have invoked the philosophy of Karl Popper in an attempt to argue for the superiority of those methods over phylogenetic methods based on Ronald Fisher's statistical principle of likelihood. We argue that the concept of likelihood in general, and its application to problems of phylogenetic inference in particular, are highly compatible with Popper's philosophy. Examination of Popper's writings reveals that his concept of corroboration is, in fact, based on likelihood. Moreover, because probabilistic assumptions are necessary for calculating the probabilities that define Popper's corroboration, likelihood methods of phylogenetic inference—with their explicit probabilistic basis—are easily reconciled with his concept. In contrast, cladistic parsimony methods, at least as described by certain advocates of those methods, are less easily reconciled with Popper's concept of corroboration. If those methods are interpreted as lacking probabilistic assumptions, then they are incompatible with corroboration. Conversely, if parsimony methods are to be considered compatible with corroboration, then they must be interpreted as carrying implicit probabilistic assumptions. Thus, the non-probabilistic interpretation of cladistic parsimony favored by some advocates of those methods is contradicted by an attempt by the same authors to justify parsimony methods in terms of Popper's concept of corroboration. In addition to being compatible with Popperian corroboration, the likelihood approach to phylogenetic inference permits researchers to test the assumptions of their analytical methods (models) in a way that is consistent with Popper's ideas about the provisional nature of background knowledge. [Assumptions; corroboration; Karl Popper; likelihood; parsimony; philosophy; phylogeny; probability.]

> Thus our analysis shows that statistical methods are essentially hypothetical-deductive, and that they proceed by the elimination of inadequate hypotheses—as do all other methods of science.
>
> Popper 1959: 413

> We can interpret . . . our measure of degree of corroboration as *a generalization of Fisher's likelihood function.*
>
> Popper 1959: 413–414

Methods of phylogenetic inference based on the statistical principle of likelihood (e.g., Fisher, 1946; Edwards, 1972) are being used with increasing frequency by systematic biologists. Advocates view these methods as powerful tools for analyzing phylogenetic relationships, particularly when the evolutionary history of a group exhibits characteristics that are expected to cause problems for other methods (e.g., Felsenstein, 1978; Hillis et al., 1994). Other advantages of phylogenetic likelihood methods include explicitness and thus testability of assumptions (e.g., Goldman, 1993a), robustness to violations of assumptions (e.g., Huelsenbeck, 1995), and consistency and efficiency under diverse conditions (e.g., Huelsenbeck, 1995). Nevertheless, likelihood methods have come under criticism from advocates of cladistic parsimony (e.g., Siddall and Kluge, 1997), an alternative class of methods that minimize the number of character state transformations necessary to account for the observed distributions of character states among taxa.

Preferences for different analytical methods sometimes reflect alternative criteria for evaluating those methods. For example, some systematists favor criteria based on performance, using simulated, laboratory-generated, and well-supported real phylogenies to evaluate which methods reconstruct phylogenies most accurately under different conditions (reviewed by Hillis, 1995). Others favor criteria based on philosophy, evaluating methods based on their consistency with a theory of epistemology (e.g., Wiley, 1975; Gaffney, 1979; Kluge, 1997a). Performance-oriented studies have shown that both likelihood and parsimony methods perform well under diverse phylogenetic conditions, though likelihood methods that model the probability of character change as a function of branch length perform better under certain conditions involving branch-length inequalities (e.g., Huelsenbeck, 1995; Huelsenbeck and Crandall, 1997; and references therein). In contrast, authors of a recent philosophy-oriented treatment (Siddall and Kluge, 1997) have argued that parsimony conforms to a theory of epistemology developed by Karl Popper (e.g., 1959, 1962, 1983), whereas likelihood supposedly does not.

In this paper we evaluate likelihood methods of phylogenetic inference in the context of Popper's writings on corroboration. We argue that Popper's corroboration is based on the general principle of likelihood and that likelihood methods of phylogenetic inference are thoroughly consistent with corroboration. We also evaluate cladistic parsimony in the same context and argue that parsimony methods are compatible with Popper's corroboration (see also Carpenter, 1992; Farris, 1995; Carpenter et al., 1998) only if they are interpreted as incorporating implicit probabilistic assumptions. Our conclusions contradict the views of authors (e.g., Siddall and Kluge, 1997) who have attempted to justify a preference for parsimony over likelihood on the basis of Popper's concept of corroboration yet deny that parsimony methods carry probabilistic assumptions. We also argue that the likelihood approach to phylogenetic inference, which permits evaluation of the assumptions inherent in its models, is consistent with Popper's views on the provisional nature of background knowledge. We do not attempt to address the entire spectrum of objections to likelihood raised by advocates of cladistic parsimony. Instead, we emphasize the relationship between Popper's corroboration and different methods of phylogenetic inference. Nevertheless, our analysis addresses some other objections to likelihood methods that are tied to Popper's philosophy, including the problem of induction, different interpretations of probability, and the nature of background knowledge.

Parsimony and Likelihood

Before analyzing parsimony and likelihood methods of phylogenetic inference in the context of Popper's ideas about corroboration, let us first describe the methods briefly and clarify some terminological issues. The *principle of parsimony,* also known as Ockham's razor, is a general philosophical principle commonly attributed to the English Franciscan William Ockham (1285–1347). This principle states that entities are not to be multiplied beyond necessity, which is often interpreted as implying that when alternative hypotheses explain the data equally well, the simplest one is to be preferred (Sober, 1994). Although frequently invoked as a general scientific virtue, the justification for this practice has been questioned (e.g., Sober, 1994).

The principle of parsimony should not be confused with the *method of parsimony* used in phylogeny reconstruction, also known as *cladistic parsimony,* which ranks alternative phylogenetic trees on the basis of the minimum number of character transformations needed to account for the observed occurrences of character states among taxa (e.g., Camin and Sokal, 1965; Farris, 1970; Farris et al., 1970). The method of cladistic parsimony is really a set of methods, because transformations both within and among characters can be weighted (assigned costs) in various ways. More importantly, it is an optimality criterion—that is, a measure for establishing a preference among alternative trees (as opposed to a method for finding the preferred tree or trees). The method of cladistic parsimony conforms to the general principle of parsimony in that hypothesized character transformations (and thus hypothesized homoplasies) are not multiplied beyond necessity, though what counts as a necessary hypothesis of character transformation depends on the costs assigned to different classes of transformations. Nevertheless, the method of cladistic parsimony should not be confused with the principle of parsimony. The principle is a very general one that can be applied in the context of many different methods, including those based on likelihood (see below).

Likelihood, developed by Ronald A. Fisher (1890–1962), is a general statistical concept based on the probability of the observed data (see Edwards, 1972, for a discussion of the history of likelihood). As stated by Edwards (1972:9), "The *likelihood,* $L(H \mid R)$, of the hypothesis H given data R, and a specific model, is proportional to $P(R \mid H)$ [the probability of obtaining results R

given the hypothesis H], the constant of proportionality being arbitrary." If the constant of proportionality is taken to equal one (e.g., Goldman, 1990:346), then $L(H \mid R) = P(R \mid H)$, or, as restated by Popper (1983:243; Popper's lowercase "l" has been capitalized for consistency):

$$L(h,e) = p(e,h) \tag{1}$$

where h = hypothesis, e = evidence, p = probability, and the term on the right is a conditional probability. Thus, the expression is to be read "the likelihood of the hypothesis given the evidence is the probability of the evidence given the hypothesis." Most authors use "|" in place of Popper's commas, a convention that we adopt in the remainder of this paper.

The concept of likelihood forms the basis of the *law of likelihood*, which is used to assess "the relative merits of rival hypotheses in the light of observational or experimental data that bear upon them" (Edwards, 1972:1). According to the law of likelihood, "a particular set of data *supports* one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis" (Edwards, 1972:30). In this sense, likelihood, like cladistic parsimony, is an optimality criterion.

Calculating the probability of the observed data given a hypothesis requires that probabilistic predictions about the data can be derived from the hypothesis (h). For commonly cited examples involving coin flips, card games, and the like, the hypothesis itself can often be described as a probabilistic model (e.g., the coin is unbiased, or $p_{\text{heads}} = p_{\text{tails}}$), in which case probabilistic predictions about the data can be derived directly from the hypothesis. Many scientific hypotheses do not, by themselves, yield probabilistic predictions about the data. Consequently, their evaluation under likelihood requires additional probabilistic assumptions, which are collectively termed *the model*.

In the case of phylogenetic inference under likelihood, the hypothesis is a tree, the specific topology of which is a hypothesis about the relationships among taxa (in addition, its general branching form is a model of the evolutionary process). The model is a probabilistic description of the evolutionary process or processes that generated the data, including both a set of parameters and estimates of their specific values. Likelihood itself should not be confused with any specific probabilistic model or set of such models. The former is a general statistical concept that can be used to evaluate many different methods and their underlying models, including cladistic parsimony (see below).

The generality of the principle of parsimony, as well as the distinction between the principle of parsimony and the method of cladistic parsimony, can be illustrated by applying the principle of parsimony in the context of likelihood—specifically, to the evaluation of alternative phylogenetic models. In a phylogenetic likelihood analysis, a researcher often wishes to know whether incorporating a specific parameter increases the explanatory power of the probabilistic model. For example, one might wish to evaluate the explanatory power of a model in which substitutions among all classes of DNA base pairs have equal probabilities—the Jukes–Cantor (one-parameter) model (Jukes and Cantor, 1969)—relative to that of a model in which transitions and transversions are allowed to have different probabilities—the Kimura (two-parameter) model (Kimura, 1980). If the different models explain the data equally well (i.e., yield identical likelihood scores), the principle of parsimony dictates that the simpler model (i.e., the one with fewer parameters) is to be preferred because it does not multiply parameters beyond necessity. In practice, likelihood scores are rarely identical, so the more complex model is usually adopted when it results in a significant improvement in the likelihood score, as judged by a test of statistical significance (e.g., Navidi et al., 1991; Swofford et al., 1996). In any case, the principle of parsimony is sufficiently general that it can be applied in the context of likelihood.

The generality of likelihood can likewise be illustrated by applying its general statistical perspective to other methods, including those of cladistic parsimony. Any use of likelihood, including its application to the problem of phylogeny reconstruction, is necessarily based on a probabilistic model. Phylogenetic methods that were not developed in the context of likelihood—such as parsimony (but see Edwards, 1996), compatibility, and phenetic clustering—are not based on such models, at least not explicitly. Nevertheless, any method can be interpreted in the context of likelihood by determining the conditions under which it will correspond to, in the sense of giving the same results as, a maximum likelihood method (e.g., Farris, 1973; Felsenstein, 1981). The reason for performing this exercise is to gain insight into the assumptions implicit in the use of methods that were not developed, but which one wishes to interpret, in a statistical context. This understanding in turn provides insight into the conditions under which a method may fail to reconstruct phylogeny correctly. One of the great benefits of adopting the general statistical perspective of likelihood is that it forces the researcher to consider analytical methods in terms of explicit models and thus to confront the implicit assumptions as well as the limitations of those methods.

The Philosophy of Karl Popper

Philosophically oriented criticisms of likelihood as an approach to phylogenetic inference (e.g., Siddall and Kluge, 1997; see also Kluge, 1997a,b) have adopted the characterization of science developed by Karl R. Popper (1902–1994) as an epistemological context. Specifically, they have proposed that likelihood is incompatible with Popper's (e.g., 1959, 1962, 1983) ideas about scientific corroboration. Thus, according to Siddall and Kluge (1997:329), "likelihood denies corroboration." In this section we argue that these criticisms are misguided. Likelihood is not only consistent with Popper's concept of corroboration, it is also, as is evident in Popper's own writings, the foundation of Popper's concept. Nevertheless, we note that Popper did not end the centuries-old debate over epistemology, and that not all subsequent philosophers have endorsed his characterization of science (for examples of critiques and alternatives to corroboration see Putnam, 1974; Salmon, 1988; Howson and Urbach, 1989). Our purpose is not to promote Popper's philosophy as the basis for evaluating phylogenetic methods. Instead, we intend merely to show that Popper's philosophy has been misrepresented in attempts to criticize likelihood and that likelihood is well-justified even according to the philosophical criteria adopted by its critics who favor cladistic parsimony.

Popper's Corroboration

Popper (1959, 1962, 1983) developed a concept that he termed *degree of corroboration*, the purpose of which was to compare rival theories in light of empirical evidence—that is, "to grade hypotheses according to the tests passed by them"(Popper, 1983:220). He defined that concept using the following expression:

$$C(h, e, b) = \frac{p(e, hb) - p(e, b)}{p(e, hb) - p(eh, b) + p(e, b)} \tag{2}$$

where p = probability, h = hypothesis (the hypothesis being evaluated), e = evidence (the results of a particular test or tests), b = background knowledge (those theories that one accepts for the purpose of conducting a test), hb refers to the conjunction of h and b, and the term $p(e, hb)$, for example, is a conditional probability read as "the probability of the evidence given the hypothesis and the background knowledge." Again, we will hereafter use "|" in place of Popper's commas to denote "given" in these conditional probabilities.

According to Popper (1983:240), the numerator in this expression "has a clear and simple intuitive significance"—that is, the probability of the evidence given the hypothesis and the background knowledge minus the probability of the evidence given the background knowledge alone. In contrast, "The denominator . . . has no such significance; it is chosen merely because it leads to satisfactory results—it removes . . . blemishes . . . —and because it seems to be the simplest normalization factor to lead to these results" (Popper, 1983:240). Specifically, it was chosen so that $C(h, e, b) = -1$ if e falsifies h in the presence of b (Popper, 1983:242). With or without the normalization factor, values for C are positive (with a maximum of $+1$) when the evidence supports the hypothesis; they are negative when the evidence undermines the hypothesis; and when the evidence has no bearing on the hypothesis, $C = 0$ (Popper, 1983:241).

Corroboration and Likelihood

Popper's views on likelihood.—Although Siddall and Kluge (1997) stated that likelihood is incompatible with Popper's corroboration, they did not mention that Popper himself made explicit statements about the relationship between corroboration and likelihood. Unlike Siddall and Kluge, Popper did not perceive any fundamental incompatibility between corroboration and likelihood. On the contrary, he viewed the two concepts as being very similar. Thus, according to Popper (1959:388; see also 1983:252), "both, my 'corroboration' and Fisher's likelihood, are intended to measure the acceptability of the hypothesis." Moreover, corroboration, according to Popper (1959:414), is "*a generalization of Fisher's likelihood function.*" A similar view was expressed by Edwards (1972:211), an advocate of likelihood, who stated that "the Method of [likelihood] is not greatly at variance with the views of Popper [1959], whose book would be a starting point in any attempt to relate the Method to a wider field."

The discrepancy between Popper's views on likelihood and those of Siddall and Kluge (1997) centers around what Popper (1983:217) called a "*mistaken solution to the problem of induction.*" According to Popper (1983:217), who followed Hume, the "problem of induction . . . arises from the fact that inductive inferences are *not valid*," which is to say that no number of specific observations can establish the truth of a general hypothesis. And a mistaken solution to the problem of induction is "the view that although induction is unable to establish an induced hypothesis with *certainty*, it is able to do the next best thing: it can attribute to the induced hypothesis some degree of *probability*" (Popper, 1983:217).

Siddall and Kluge (1997:314) hold the erroneous belief that likelihood suffers from this mistaken solution to the problem of induction. On the contrary, likelihood has no such problem. In Popper's own words:

> the problem of induction . . . consists in determining the value of r in $p(h,e) = r$: that is to say, the value of the probability of the induced hypothesis h given the evidence e (Popper, 1983:218).

However:

> Degree of corroboration is not a probability; that is to say, it does not satisfy the rules of the calculus of probability.[8] [8]The same holds true even for $l(h,e)$, the 'likelihood of h' in Fisher's sense, defined by $l(h,e) = p(e,h)$; for even though it is a probability, it is not one of h (Popper, 1983:243; the second sentence is a footnote).

The point is that the mistaken solution to the problem of induction involves assigning probabilities to hypotheses, but likelihood does not assign probabilities to hypotheses. Likelihood is not the probability of the hypothesis given the evidence but the probability of the evidence given the hypothesis.

The view that the appropriate criterion for judging the relative merits of rival hypotheses is something other than the probability of those hypotheses is reflected in the terms chosen by both Popper and Fisher to describe that criterion—that is, "corroboration" and "likelihood," respectively. Thus, as stated by Popper:

> I introduced the terms '*corroboration*' . . . and . . . '*degree of corroboration*' . . . because I wanted a *neutral* term to describe the degree to which a hypothesis has stood up to severe tests, and thus 'proved its mettle'. By 'neutral' I mean a term not prejudging the issue whether, by standing up to tests, the hypothesis becomes 'more probable,' in the sense of the probability calculus (Popper, 1959:251).

And in the case of Fisher:

> What has now appeared is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term '*Likelihood*' to designate this quantity (Fisher, 1946:10).

Thus, Fisher's likelihood and Popper's corroboration reflect similar views on the evaluation of rival scientific hypotheses. Siddall and Kluge's (1997) erroneous implication that likelihood is incompatible with corroboration stems from their failure to distinguish between probability in general and the probability of a hypothesis. Only methods that attempt to assign probabilities to hypotheses are compromise by what Popper called a mistaken solution to the problem of induction, and likelihood is not one of those methods.

The relationship between corroboration and likelihood.—The similarity between Fisher's likelihood and Popper's corroboration extends well beyond the fact that neither attempts to assign probabilities to hypotheses. Another similarity is that both likelihood and corroboration are based on the general concept of probability (e.g., Edwards, 1972:9; Popper, 1959:388; 1983:233, 238; see also Carpenter et al., 1998). The probabilistic basis of both concepts can readily be seen by examining the expressions used to describe those concepts (Eqs. 1 and 2 above): In both expressions, every term in the defining formula (definiens) is a conditional probability.

In the case of likelihood, a probabilistic basis is not disputed, but corroboration also has an explicit probabilistic basis (see also Farris, 1995; Carpenter et al., 1998). Thus, in his derivation of the corroboration expression, Popper (1983) stated, "our definition of degree of corroboration will be put in the more familiar terms of probability ... the definiens will be a function of probabilities" (p. 233), and "[w]e therefore need the calculus of probability [to define the concept of corroboration]; and we shall write $p(a, b) = r$, which is to be read: 'The probability of a, given b, equals r'" (p. 238). That degree of corroboration is not itself a probability refers to the fact that, despite being based on probabilities, it does not obey the laws of probability (analogously, the consistency index [Kluge and Farris, 1969], despite being based on length, is not itself a length).

Siddall and Kluge (1997) suggest that the probability term in the likelihood expression is of a different kind than are those in the corroboration expression. Specifically, they distinguish between "the calculus of frequency probability typified by Bayes' Theorem [which they associate with likelihood] ... and ... logical probability exemplified in Popper's (1983) degree of corroboration" (Siddall and Kluge, 1997: 313). This statement misleadingly associates Popper's corroboration with a particular interpretation of the concept of probability.

As noted by Popper (1959, 1983), probabilities can be interpreted in several ways: as numerical frequencies (number x out of y in the limit of y; e.g., Von Mises, 1928), as subjective degrees of belief (e.g., De Finetti, 1931), as propensities (inherent tendency of an event to occur; e.g., Popper, 1983), or as degrees of a logical relationship between statements (e.g., Keynes, 1921). Popper explicitly intended for any of these interpretations to be applicable to the probability terms in his corroboration equation. Thus, "It is desirable to construct a system of axioms ... in which 'p(x_1, x_2)' [read 'the probability of x_1 with regard to x_2'] ... is constructed in such a way that it can be equally interpreted by any of the proposed interpretations" (Popper, 1959:320). "In *L.Sc.D.* [*The Logic of Scientific Discovery*] (Appendix *iv) I gave a number of axiom systems for the formal calculus of probability (one of whose interpretations is the logical interpretation)" (Popper, 1983:218). And, "I propose to use the word 'probability' here, and in other places, for all and only those meanings that satisfy the well known *mathematical calculus of probabilities*" (Popper, 1983:282). Moreover, Popper (1959:119) noted that "numerical probability can be linked with logical probability ... It is possible to interpret numerical probability as applying to a subsequence (picked out from the logical probability relation) for which a system of measurement can be defined, on the basis of frequency estimates." Thus, contrary to the view of Siddall and Kluge (1997), according to Popper himself, the frequency interpretation of probabilities is fully compatible with his corroboration equation.

The similarities between likelihood and corroboration extend further still, for not only are both concepts based on probabilities (however interpreted), both are based on the probability of the same thing—that is, the probability of the evidence (e) rather than that of the hypothesis (h). In other words, the conceptual basis of likelihood and corroboration is identical, or more to the point, corroboration is based on likelihood. In Popper's own words:

> I soon found that, in order to define $C(x,y)$—the degree of corroboration of the theory x by the evidence y—I had to operate with some converse $p(y,x)$, called by Fisher the '*likelihood* of x' (Popper, 1959:388; see also 1983:252).

Comparing the defining formulas of the likelihood (eq. 1) and corroboration (eq. 2) expressions reveals how corroboration is based on likelihood. The defining formula in the likelihood expression is $p(e \mid h)$, the probability of the evidence given the hypothesis. The defining formula in the corroboration expression is more complex, consisting of five probabilities, two in the numerator and three in the denominator. As noted above, however, the denominator in the corroboration expression is merely a "normalization factor," which was chosen so that $C = -1$, rather than 0, when e falsifies h (Popper, 1983:242). Thus, the numerator, $p(e \mid hb) - p(e \mid b)$, is the crux of corroboration (Popper, 1983:240). It differs from likelihood in the recognition of a distinct term for what Popper called the "background knowledge," b. b refers to "assumptions" (Popper, 1962:238), to "theories not under test" (1983:252), to other hypotheses that we treat as "unproblematic" (1962:238; 1983:188)—to hypotheses that we "accept—perhaps only tentatively—while we are testing h" (1983:236). However, if b designates a hypothesis (theory) or hypotheses, even if treated as unproblematic, then both of the terms in the numerator of the corroboration expression—$p(e \mid hb)$ and $p(e \mid b)$—are likelihoods. Corroboration, then, is the normalized difference between two likelihoods.

Considering the differences between the corroboration and likelihood expressions provides further insights into the relationship between the two concepts. The corroboration expression includes separate terms for the hypothesis being tested (h) and the other propositions (hypotheses, assumptions) needed for calculating the probability of the evidence, that is, the background knowledge, b. In contrast, the likelihood expression does not include a term for the background knowledge, which corresponds with the model of likelihood. As evidence of this correspondence, compare Popper's (1983:244) characterization of the background knowledge— "knowledge which, by common agreement, is not questioned while testing the theory under investigation"—with Edwards's (1972:3) characterization of the model as "that part of the description [of the phenomenon responsible for generating the observations] which is not at present in question, and may be regarded as given." Because the likelihood expression does not contain a term for the model (background knowledge), but the model is necessary for calculating the probability of the evidence given the hypothesis, the term $p(e \mid hb)$ of the corroboration expression is equivalent to $p(e \mid h)$ of the likelihood expression. This conclusion is confirmed by Popper's (1959:413) statement that when $p(e \mid b)$ is very small, and thus $p(e \mid hb) - p(e \mid b) \approx p(e \mid hb)$, "it will . . . be possible to accept Fisher's likelihood function as an adequate measure of degree of corroboration." Therefore, the main difference between the two expressions is the presence of the term $p(e \mid b)$ in the corroboration expression.

Popper introduced the term $p(e \mid b)$ in the context of his observation that "if e should be *probable*, in the presence of b alone ('probable' in the sense of the probability calculus), then its occurrence can hardly be considered as significant support of h" (1983:237). For a hypothesis to be corroborated by a particular body of evidence, $p(e \mid hb)$, the probability of the evidence given the hypothesis and the background knowledge, must be greater than $p(e \mid b)$, the probability of the evidence given the background knowledge alone. The likelihood expression does not explicitly address this issue—that is, the probability of the evidence given the hypothesis and the background knowledge in relation to the probability of the evidence given the background knowledge alone; however, we will argue in the next section that for standard phylogenetic analyses, whether under parsimony or likelihood, the term $p(e \mid b)$ can effectively be ignored (see also Appendix).

Corroboration and Phylogenetic Inference

In this section we apply Popper's concept of corroboration to the problem of phylogenetic inference. First, we take the basic components of the corroboration expression, as well as the conditional probabilities based on them, and identify the corresponding components in a phylogenetic study. We then examine parsimony and likelihood methods of phylogenetic inference in the context of this analysis. We show that for the evaluation of rival phylogenetic hypotheses, likelihood methods fit easily into the context of corroboration. We argue that parsimony methods fit into that context as well, but only if they are interpreted as carrying implicit probabilistic assumptions. We then analyze Kluge's (1997a) claims about the assumptions of cladistic parsimony and explain why descent with modification is insufficient background knowledge for phylogenetic inference as an example of Popper's corroboration. Finally, we extend our analysis of corroboration to consider the evaluation of models (assumptions), showing how the likelihood approach to phylogenetic inference is concordant with Popper's views on the provisional nature of background knowledge.

Evaluation of Alternative Phylogenetic Trees

Let us first consider the evaluation of alternative phylogenetic trees using a single phylogenetic method—for example, likelihood under a particular probabilistic model or parsimony under a particular weighting scheme (including equal weights). For the sake of simplicity, we will restrict our considerations to data taking the form of discrete characters. The basic components of the corroboration expression (Eq. 2) are the evidence (e), the hypothesis (h), and the background knowledge

(b). In a phylogenetic analysis, whether under parsimony or likelihood, the evidence (e) consists of the observed character states and their occurrences in particular taxa. The hypothesis (h) is a tree—that is, a topology. Popper's corroboration is used to compare rival hypotheses according to the results of their tests. Therefore, when performing a phylogenetic analysis in this context, each tree is a rival hypothesis, and the problem is to determine the degree of corroboration, C, for each tree—that is, C_1, C_2, $C_3, \ldots C_n$ for h_1, h_2, $h_3, \ldots h_n$.

The background knowledge (b) consists of hypotheses that are necessary for a particular test or analysis but are not questioned while conducting that analysis (i.e., assumptions; see *The relationship between corroboration and likelihood*). In a phylogenetic analysis, these hypotheses include the basic axiom of descent with modification as well as any other propositions that are held constant in the analysis, including the assumption that the relationships in question conform to a tree-like pattern (given that the analysis is so constrained) and whatever assumptions are implied by using a particular method to evaluate alternative trees (minimally, the assumption that the method of choice provides a suitable means for reconstructing phylogeny). The method includes both an optimality criterion (e.g., parsimony, likelihood) and various propositions concerning character transformation (e.g., character state order, character and state weights, transformation probabilities, among-site rate variation) and forms a critical part of the background knowledge. The reason is that the method provides the basis for selecting a preferred tree or trees and implies that this tree is the most highly corroborated by the data (i.e., has the highest positive value of C).

When alternative trees are compared using a single analytical method—whether parsimony under a particular weighting scheme or likelihood under a particular probabilistic model—all trees are evaluated under the same set of assumptions. In other words, the background knowledge (b) is held constant. Because alternative trees are evaluated by using the same data, the evidence (e) is also constant. Therefore, $p(e \mid b)$ is constant, and the problem of determining the relative degree of corroboration, C, for each member of a set of alternatives trees (h_1, h_2, h_3, $\ldots h_n$) reduces to determining the value of $p(e \mid hb)$ for each alternative tree (see Appendix for a more detailed discussion of why $p(e \mid b)$ is ignored). However, as we argued above, $p(e \mid hb)$ of the corroboration expression is equivalent to $p(e \mid h)$ of the likelihood expression. Therefore, when evaluating alternative trees with a single phylogenetic method, the problem of comparing alternative trees in terms of their degree of corroboration reduces to comparing the likelihoods of the alternative trees.

Thus, contrary to the view of Siddall and Kluge (1997), the application of likelihood to problems of phylogenetic inference is fully compatible with Popper's concept of corroboration. Indeed, there seems to be no difference between determining the degree of corroboration (C) of alternative trees and determining their likelihoods (L). Moreover, under likelihood, comparing alternative trees in terms of their degree of corroboration has a clear meaning and an explicit basis. Because Popper's concept of corroboration is based on probabilities—most notably $p(e \mid hb)$, the probability of the data given a particular tree and phylogenetic method—probabilistic assumptions are needed to calculate the degree of corroboration (C). This demand is met by the explicit probabilistic models integral to any phylogenetic analysis that uses likelihood.

In contrast with the clear conformity of phylogenetic likelihood methods with Popper's concept of corroboration, the compatibility of cladistic parsimony methods with Popper's corroboration is not obvious. Unlike likelihood methods, parsimony methods are not based on explicit probabilistic models, and thus they provide no basis for translating the minimum number of character transformations required by a tree into the probability of the observed distribution of character states among taxa given that tree. Therefore, demonstrating a connection between cladistic parsimony and Popper's corroboration simply by identifying the components of a parsimony analysis that correspond with e (the observed distribution of character states among taxa), h (a tree topology) and b (a parsimony method), as we have done for likelihood methods, is not possible.

Siddall and Kluge (1997), who assert that "cladistic parsimony does not assume a process model" (p. 326), attempt to solve this problem by interpreting the probabilities of the corroboration expression as logical probabilities. They state, "Cladistic parsimony denies frequency probabilism" (p. 329) and propose instead that the parsimony method makes use of "logical probability regarding corroboration in historical inference" (p. 333). Their statements imply that cladistic parsimony is associated with the logical interpretation of probabilities, which somehow enables them to calculate C without postulating explicit probabilities. Siddall and Kluge do not explain how this calculation is to be accomplished; regardless, their position is contradicted by Popper's views. Popper was clear in stating that "*there cannot be a metric of logical probability which is based upon purely logical considerations*" (Popper, 1959:404) and "Only from probability estimates can probabilities be calculated" (1959:247).

There is nothing special about cladistic parsimony that excuses it from this requirement. Without invoking probabilistic assumptions, a cladistic parsimony analysis contains insufficient information to determine $p(e \mid hb)$, the probability of the observed character state distributions given a tree and the parsimony method. One can interpret the lengths of alternative trees as indicative of relative degree of corroboration—that is, one can consider the most-parsimonious tree to have the greatest value of C and thus also of $p(e \mid hb)$. For example, according to Kluge (1997b:350; see also Farris, 1995), "in strictly Popperian terms, most parsimonious cladograms are most explanatory because both C and S [severity of a test] increase with $p(e, hb)$, a term that occurs in their shared numerator." Kluge's statement is ironic given his antipathy toward likelihood methods of phylogenetic inference, for as we argued above, $p(e \mid hb)$ of the corroboration expression is equivalent to $p(e \mid h)$ of the likelihood expression. Therefore, according to Kluge, most-parsimonious cladograms are most explanatory because they maximize $p(e \mid h)$—that is, because they maximize likelihood! In any case, without invoking probabilistic assumptions, equating tree length with degree of corroboration begs the question by avoiding the actual calculation of $p(e \mid hb)$ and assuming a one-to-one correspondence between that quantity and tree length. To assign actual values to $p(e \mid hb)$ and thus C, one must identify explicit probabilistic propositions—be they logical, statistical, or otherwise—associated with the parsimony method (see *Critique of Kluge [1997a]*, below). If parsimony analyses are not at least implicitly based on such probabilistic assumptions, then they cannot be examples of Popperian corroboration.

In sum, probabilistic assumptions are necessary to determine $p(e \mid hb)$, and any attempt to justify cladistic parsimony in terms of Popperian corroboration requires that those methods be interpreted as invoking implicit probabilistic assumptions (see below). An analysis of the precise nature of the probabilistic assumptions inherent to the interpretation of parsimony as a method for assessing the degree of corroboration (i.e., likelihood) of alternative trees is beyond the scope of the present paper (for such analyses see Farris, 1973; Felsenstein, 1973, 1981; Goldman, 1990; Tuffley and Steel, 1997). But regardless of the precise nature of those assumptions, parsimony methods can be reconciled with Popper's concept of corroboration only by invoking probabilistic assumptions.

Critique of Kluge (1997a)

Our conclusions about the implications of Popperian corroboration for cladistic parsimony contradict those of Kluge (1997a; see also Siddall and Kluge, 1997), who claimed that parsimony methods are justified by Popperian corroboration but argued that descent with modification is sufficient background knowledge for phylogenetic inference under parsimony—in other words, that additional probabilistic assumptions are unnecessary. According to Kluge (1997a:88):

> Given only descent with modification as the background knowledge, synapomorphies characteristic of (A,B), (A,C), and (B,C) should be equally likely . . . However, if a large majority of one class of those possible synapomorphies were to be discovered, say that which characterizes hypothesis (A,B), then this is unlikely given the background knowledge alone, but not under the background knowledge plus the postulated rooted (A,B)C cladogram. The (A,B)C cladogram is said to be corroborated to the degree to which those (A,B) synapomorphies are observed.

In other words, according to Kluge, a large majority of characters exhibiting the pattern 110 for taxa ABC has a low probability given the background knowledge of descent with modification alone, so $p(e \mid b)$ is small; that same preponderance of characters has the highest probability given the same background knowledge plus the rooted topology (AB)C, so $p(e \mid hb)$ is maximally large (i.e., relative to alternative topologies). If so, then $p(e \mid hb) - p(e \mid b)$ has the largest positive value for the topology (AB)C, which is therefore the most highly corroborated hypothesis.

Kluge's reasoning is flawed. Given descent with modification alone, nothing can be inferred about the probabilities of the different possible character patterns for three taxa (i.e., 000, 100, 010, 001, 110, 101, 011, 111). Determining the probabilities of different character patterns requires postulation of a probability distribution or a process for generating such a distribution (see Popper, 1959:208, 247). However, the assumption of descent with modification alone provides neither. If no probability distribution or generating process is specified, then no distribution of states is any more or less likely than any other, and this is true even with the additional assumption of a tree-like model of descent. Consequently, Kluge's statement that certain character patterns should be equally likely, given only the assumption of descent with modification, has no basis. Instead, the statement itself is an additional and unjustified probabilistic assumption.

Let us consider biological assumptions that could be incorporated into the background knowledge to allow us to conclude that the character patterns 110, 101, and 011 are equally probable. One possibility is the assumption of a rooted star topology (trichotomy) combined with that of equal (and non-zero) probabilities of change for every character on every branch. Alternatively, equal probabilities for the alternative character patterns can be inferred without assuming a particular topology by assuming that the probabilities of change for every character on every branch are sufficiently high that the data are effectively randomized. An assumption of this sort underlies the permutation tail probability (PTP) test of Faith and Cranston (1991), which creates permuted data sets with equal probabilities for the different character patterns by randomization and uses these to test a null hypothesis of no hierarchical structure in the observed (unpermuted) data. In any case, additional assumptions (beyond descent with modification) are needed to reach the conclusion that certain character patterns are equally probable, and those assumptions are necessarily probabilistic. As stated by Popper (1959:247), "Frequency statements . . . need their own assumptions which must be specifically statistical."

As a consequence, and contrary to Kluge's (1997a) position, additional probabilistic assumptions are also necessary to infer that a preponderance of characters exhibiting a particular pattern (e.g., 110) has the highest probability on one of the three, rooted, bifurcating, three-taxon topologies—in other words, that $p(e \mid hb)$ (and thus C) is maximal for one of the three topologies. Following Kluge (1997a:87), we will use 0 to designate an ancestral state and 1 to designate a derived state, so that the hypothetical ancestor at the root of the tree has all characters with state 0. Suppose also that the evidence (e) exhibits a preponderance of characters with the 110 pattern in taxa ABC relative to those with the patterns 101 and 011, as in Kluge's example. Without invoking additional probabilistic assumptions, it would be incorrect to conclude that characters with the pattern 110 have the highest probability for the topology (AB)C. The reason is that the probability of the pattern 110 is greater on one (or both) of the alternative topologies than on the (AB)C topology for certain patterns of inequality in the probabilities of change among branches (Fig. 1).

Given Kluge's (e.g., 1997b) preference for cladistic parsimony with all characters equally weighted, the probabilistic assumptions necessary to arrive at his conclusion are, not surprisingly, those that have been identified in likelihood models approximating his preferred method. One possibility is that the probability of change is the same for every character (i.e., equal weighting) on every branch (e.g., Farris, 1973; Felsenstein, 1983; Goldman, 1990). It also seems necessary to assume that the probability (rate) of character change does not exceed a certain limit. If the probability of change is sufficiently great that the data are effectively randomized, a large majority of 110 characters has a low and equal probability on all of the alternative trees. Another possibility is that the probabilities of change on the various branches are estimated separately for each character—in other words, an assumption that the probability of change for a given character on a given branch is not related to the probability of change for other characters on that branch (Tuffley and Steel, 1997). Under either set of probabilistic assumptions, characters exhibiting the pattern 110 have the highest probability for the (AB)C topology.

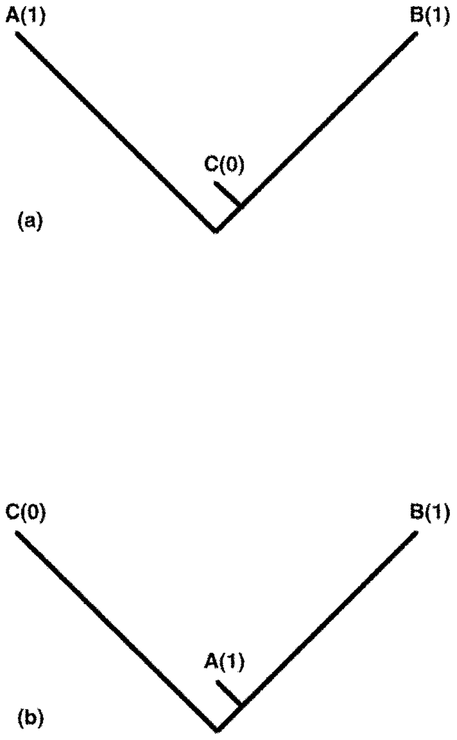

FIGURE 1. Example illustrating that the occurrence of a derived state (1) in two taxa (A and B) but not a third (C) does not necessarily have the highest probability on a tree in which the two taxa that share the derived state are sister groups (after Felsenstein, 1978). (a) If taxa B and C are sister groups, and the probabilities of change on the common ancestral BC branch and the C branch are low but the probabilities of change on the A and B branches are high, then the probability of observing the derived state in taxa A and B (as the result of convergence) but not in C is relatively high because the probability of the 0 → 1 change occurring is high for both the A and B branches and low for both the BC and C branches. (b) If taxa A and B are sister groups, and the probabilities of change on the common ancestral AB branch and the A branch are low but the probabilities of change on the C and B branches are high, then the probability of observing a derived state in taxa A and B (as the result of inheritance from a common ancestor) is relatively low because the probability of the 0 → 1 change occurring is low for the common ancestral AB branch, and even if it happens, the probability of reversal to 0 is high for the B branch. Thus, the character pattern 110 for taxa ABC has a higher probability on at least some trees in which taxa C and B are sisters (a) than it does on at least some trees in which A and B are sisters (b). Lengths of branches are proportional to the probabilities of change.

The point is, assumptions in addition to that of descent with modification are required to reach the conclusion that a particular character pattern has the highest probability on a particular topology, and those assumptions are necessarily probabilistic.

Thus, contrary to the conclusion of Kluge (1997a; see also Siddall and Kluge, 1997), descent with modification by itself does not constitute sufficient background knowledge for phylogeny reconstruction as an example of Popperian corroboration. Assessing the degree of corroboration for alternative trees requires calculating the probability of the evidence given each tree and the background knowledge, $p(e \mid hb)$, but the assumption of descent with modification (together with a tree) contains no information from which this probability can be calculated. If no probabilistic assumptions are invoked, then no set of character patterns (e) can be inferred to be more or less probable on any topology, and thus there is no basis for selecting a most corroborated tree. Assessing the degree of corroboration for alternative trees requires additional assumptions, and those assumptions must be probabilistic in nature.

Evaluation of Alternative Phylogenetic Methods or Models

So far, we have considered phylogenetic analysis under a single method or model. In such cases, the method or model and its implicit or explicit assumptions are held constant and thus form part of the background knowledge. As described by Siddall and Kluge (1997:23), “Background knowledge is, by definition (Popper, 1963), unproblematic. It is something we can assume as holding ‘true’ while we conduct our test.” But according to Popper (1962:238), “Few parts of the background knowledge will appear to us in all contexts 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” (see also Popper, 1983:188). The provisional nature of background knowledge described by Popper allows phylogeneticists to evaluate not only alternative topologies but also alternative phylogenetic methods or models in terms of degree of corroboration. Once again, in terms of compatibility with Popper’s views, the likelihood approach to phylogenetic inference compares favorably with cladistic parsimony as interpreted and practiced by advocates of those methods.

When evaluating alternative phylogenetic methods or models according to their degree of corroboration, the methods or models are treated as part of the hypothesis (h) rather than the background knowledge (b). The topology can be allowed to vary, in which case it also forms part of the hypothesis. Alternatively, the methods or models can be evaluated on a single topology, in which case the topology forms part of the background knowledge. In either case, the background knowledge (b) consists of whichever propositions (components, parameters) are common to the alternative methods or models and is therefore held constant. Given that the methods or models are to be evaluated in terms of their ability to explain the same set of observations, the evidence (e) is also constant. If b and e are constant, then $p(e \mid b)$ is constant, and the problem reduces once again to determining the value of $p(e \mid hb)$ for each member of a set of alternative hypotheses ($h_1, h_2, h_3, \ldots h_n$). In this case, however, the alternative hypotheses are (or at least include) the components or parameters that differ among the alternative methods or models.

The likelihood approach to phylogenetic inference is fully compatible with the use of Popper's degree of corroboration to evaluate various components of the background knowledge. As just noted, comparing different phylogenetic models in terms of their degree of corroboration reduces to determining the value of $p(e \mid hb)$ for each model, but $p(e \mid hb)$ of the corroboration expression is the same thing as $p(e \mid h)$ of the likelihood expression (see *The relationship between corroboration and likelihood*). Therefore, determining the degree of corroboration for different models is the same thing as determining their likelihoods. Moreover, evaluating alternative models—that is, evaluating hypotheses that in other situations form part of the background knowledge—is consonant with Popper's view that any part of the background knowledge may be challenged at any time. Similar views have been expressed by advocates of likelihood. For example, compare the quotation from Popper (1962) in the first paragraph of this section with the following statement by Edwards (1972:4): "There is no absolute distinction between the two parts of a statistical description, for what is on one occasion regarded as given, and hence part of the model, may, on another occasion, be a matter for dispute, and hence part of a hypothesis." Finally, different likelihood models—for example, those that differ with respect to incorporation of a specific model parameter—are commonly evaluated in terms of their degree of corroboration (likelihood) (e.g., Goldman, 1993a,b; Yang, 1994, 1996; Cunningham et al., 1998). The ability to evaluate alternative models enables investigators to tailor their analyses to individual data sets as well as test hypotheses about evolutionary processes by incorporating them as model parameters (reviewed by Huelsenbeck and Rannala, 1997).

Parsimony methods are more difficult to reconcile with Popper's views on corroboration and background knowledge. Even if interpreted as invoking implicit probabilistic assumptions, so that they are consistent with Popper's corroboration, different parsimony methods/models (e.g., Wagner, Fitch, Dollo; different weighting schemes) are rarely, if ever, compared in terms of their degree of corroboration. Instead, the alternative methods are most commonly compared in terms of whether they yield different optimal trees, without any attempt to determine which method is associated with the highest probability for the observed data. Consequently, either no choice is made among alternative parsimony methods, or the choice is based on a criterion other than the degree of corroboration.

Only Kluge (1997b) has attempted to justify a preference for one parsimony method over another in the terms of Popperian corroboration by arguing that equal weighting is to be preferred because differential weighting "adds to background knowledge" (p. 349). The implication is that adding to background knowledge increases $p(e \mid b)$, which decreases the difference between $p(e \mid hb)$ and $p(e \mid b)$, thus lowering C. However, if the weighting scheme is at issue, then it is part of the hypothesis (h), not of the background knowledge (b). In other words, if character weights are permanently relegated to the background knowledge, the possibility of testing different weighting schemes in terms of their degree of corroboration is denied. This practice not only goes against Popper's views but is decidedly unscientific. Just as with alternative tree topologies, the relative merits of alternative weighting schemes can and should be evaluated empirically. Under Popperian corroboration, this evaluation involves weighting according to probabilities of change and determining which scheme is associated with the highest probability of the data. Thus, although cladistic parsimony can be interpreted in a way that is compatible with Popper's views on corroboration and background knowledge, that interpretation is not adopted by the very people who claim the compatibility.

Conclusion

Contrary to the views of authors who have criticized the likelihood approach to phylogenetic inference as being incompatible with Karl Popper's degree of corroboration, an examination of Popper's own writings reveals that the general concept of likelihood forms the very basis of his degree of corroboration. Consequently, it is not surprising that likelihood methods of phylogenetic inference are fully compatible with Popperian corroboration and that cladistic parsimony methods are compatible with corroboration only if they are interpreted as incorporating probabilistic assumptions. But if parsimony methods are interpreted as incorporating probabilistic assumptions, then those assumptions constitute models that can be used in the context of likelihood, and non-probabilistic implementations of those methods are simply proxies for their probabilistic counterparts. Interpreted this way, there is no conflict between parsimony and likelihood, because the general statistical perspective of likelihood—and of Popperian corroboration—subsumes all of the individual methods and models that can be applied within the context of that perspective, including those of cladistic parsimony. One of the primary advantages of adopting this perspective is that all of the various phylogenetic methods/models are unified under a single, general, theoretical framework that allows phylogeneticists to compare those methods/models directly in terms of ability to explain data. In this context, all phylogenetic methods/models are legitimate philosophically, though all have limitations, and some may explain the data better than others in particular cases. But regardless of the relationship between cladistic parsimony methods and either Popper's degree of corroboration or Fisher's likelihood, likelihood forms the basis of Popper's degree of corroboration, and likelihood methods of phylogenetic inference are fully compatible with that concept.

Acknowledgments

We thank G. Yaffe and a discussion group composed of F. Anderson, R. Brumfield, L. Buckley, P. Foster, J. Huelsenbeck, J. McGuire, S. Steppan, D. Swofford, and J. Wilgenbusch for helpful discussions. D. Faith, J. Trueman, and M. Siddall provided comments on an earlier version of the manuscript. Part of this work was supported by a Smithsonian Predoctoral Fellowship awarded to Steven Poe.

References

Camin, J. H., and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311–326.

Carpenter, J. M. 1992. Random cladistics. Cladistics 8:147–153.

Carpenter, J. M., P. A. Goloboff, and J. S. Farris. 1998. PTP is meaningless, T-PTP is contradictory: A reply to Trueman. Cladistics 14:105–116.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

De Finetti, B. 1931. Sul significato soggettivo della probabilita. Fundam. Math. 17:298–329.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge. [Expanded edition published by Johns Hopkins University Press, Baltimore, 1992.]

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

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

Faith, D. P., and P. S. Cranston. 1991. Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. Cladistics 7:1–28.

Faith, D. P., and P. S. Cranston. 1992. Probability, parsimony, and Popper. Syst. Biol. 41:252–257.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

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

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

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

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

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

Felsenstein, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linn. Soc. 16:183–196.

Felsenstein, J. 1983. Parsimony in systematics: Biological and statistical issues. Annu. Rev. Ecol. Syst. 14:313–333.

Fisher, R. A. 1946. Statistical methods for research workers, 10th edition [1st edition published in 1925]. Oliver and Boyd, Edinburgh.

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

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

GOLDMAN, N. 1993b. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. 37:650–661.

HILLIS, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44:3–16.

HILLIS, D. M., J. P. HUELSENBECK, AND D. L. SWOFFORD. 1994. Hobgoblin of phylogenetics? Nature 369:363–364.

HOWSON, C., AND P. URBACH. 1989. Scientific reasoning: The Bayesian approach. Open Court, La Salle, IL.

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK, J. P., AND K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28:437–466.

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

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KEYNES, J. M. 1921. A treatise on probability. Macmillan, London.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

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

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

KLUGE, A. G., AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

NAVIDI, W. C., G. A. CHURCHILL, AND A. V. VON HAESELER. 1991. Methods for inferring phylogenies from nucleic acid sequence data using maximum likelihood and linear invariants. Mol. Biol. Evol. 8:128–143.

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 in 1968 by Harper and Row, New York.]

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

PUTNAM, H. 1974. The "corroboration" of theories. Pages 221–240 *in* The philosophy of Karl Popper, Book 1 (P. E. Schlipp, ed.). Open Court, La Salle, Illinois.

SALMON, W. 1988. Rational prediction. Pages 47–60 *in* The limitations of deductivism (A. Grunbaum and W. C. Salmon, eds.). Univ. of California, Berkeley.

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

SOBER, E. 1994. From a biological point of view. Cambridge Univ. Press, Cambridge, England.

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

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.

VON MISES, R. 1928. Wahrscheinlichkeit, Statistik und Wahrheit. J. Springer, Wien. [English translation, Probablity, statistics, and truth, published in 1939 by Macmillan, New York.]

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. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

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

Received 30 April 1999; accepted 29 September 1999
Associate Editor: D. Faith

APPENDIX: ADDITIONAL CONSIDERATIONS CONCERNING P (E | B)

Role of p(e | b) *in Different Kinds of Analyses*

Our conclusion that the term $p(e \mid b)$ can be ignored in the case of standard phylogenetic analyses may seem inconsistent with the importance that Popper attached to this term. For Popper, $p(e \mid b)$ provided the critical difference between his concept of corroboration and Fisher's likelihood (e.g., Popper, 1959:413–414). It was also crucial for assessing the severity of a test. C, with the numerator of $p(e \mid hb) - p(e \mid b)$, will be largest when $p(e \mid hb)$ is large and $p(e \mid b)$ is small. Therefore, provided that $p(e \mid hb) > p(e \mid b)$, then "*the smaller* $p(e \mid b)$ [i.e., the more improbable the evidence given the background knowledge alone] *the stronger* will be the support which e renders to h" and the more severe will be the test (Popper, 1983:238). Our conclusion that $p(e \mid b)$ can be ignored in the case of standard phylogenetic analyses is in no way contradicted by these propositions. Instead, the seeming discrepancy results from the fact that different kinds of analyses emphasize different aspects or components of corroboration: (1) the evaluation of rival hypotheses in terms of the results of a test—which is based primarily on $p(e \mid hb)$, and (2) the evaluation of different tests in terms of their severity—which is based primarily on $p(e \mid b)$. Considering these components of corroboration separately makes it clear why $p(e \mid b)$ can be (and is) ignored in standard phylogenetic analyses.

Standard phylogenetic analyses (i.e., those evaluating alternative trees in the context of a single data set—even if it results from combining separately collected bodies of data—and a single phylogenetic method), are examples of the first aspect of corroboration—the comparison of rival hypotheses (alternative trees) in terms of the results of a test. Because both e (the data) and b (the analytical method) are constant across hypotheses, $p(e \mid b)$ is also constant; therefore, $p(e \mid b)$ does not affect the order of preference among rival trees determined by $p(e \mid hb)$. Moreover, no component of such an analysis corresponds with the second aspect of corroboration, that is, evaluating the test in terms of its severity. Thus, $p(e \mid b)$, which measures the severity of a test, is effectively ignored.

Popper (e.g., 1959, 1962, 1983) was concerned with developing a general framework for evaluating diverse scientific theories according to the results of their tests, but he did not consider the case of standard phylogenetic analyses (most of his examples are from physics and astronomy). In other cases, the evaluation of different tests in terms of their severity—the second aspect of corroboration—is important, such as when multiple tests have been performed, particularly when those tests involve methods based on very different assumptions (b) and yield very different classes of observations (e). For example, two famous tests of Einstein's general theory of relativity (one of Popper's favorite examples) involved measurements of (1) the angular deflection of starlight passing close to the sun (measured during an eclipse) and (2) the reduction in the wavelength of light emitted by massive stars. When different tests involve very different assumptions and classes of observations, combining the results of those tests and thus calculating a single value for C may be impossible. Therefore, evaluating the relative severity of the different tests becomes important, which is accomplished using $p(e \mid b)$.

Although $p(e \mid b)$ is ignored in what we have called standard phylogenetic analyses, there are other kinds of phylogenetic tests in which $p(e \mid b)$ can be used to evaluate severity. The most obvious example involves the comparison of more and less general evolutionary models, such as those describing the evolution of nucleotide sequences (reviewed by Swofford et al., 1996). In this case, the hypothesis being tested corresponds with the parameter that distinguishes the more general model from the less general one (e.g., base frequencies, rates for different classes of substitutions). Therefore, the less general model (i.e., the one lacking the parameter corresponding with the hypothesis) constitutes the background knowledge and can itself be used to calculate $p(e \mid b)$ for evaluating the severity of the test. For example, the model of Jukes and Cantor (1969), in which the substitution probabilities among all four classes of nucleotide bases are assumed to be equal, can be used to calculate $p(e \mid b)$ for a test of the more general model of Kimura (1980) and thus of the hypothesis that transitions and transversions have different substitution probabilities.

In such cases, that which constitutes a severe test, as measured by $p(e \mid b)$, corresponds with the generally accepted proposition that, all else being equal, an analysis based on a larger body of evidence (e.g., more characters) constitutes a more severe test than one based on a smaller body of evidence (e.g., fewer characters). This conclusion is supported by Popper's (1959:413) statement that, in the case of statistical hypotheses, very small $p(e \mid b)$ is possible only for large samples. Thus, just as the probability of obtaining a particular fraction of heads in a series of coin tosses under any given hypothesis about the coin's bias (or lack thereof) is lower for larger numbers of tosses, similarly the probability of sampling particular fractions of the different possible character patterns (i.e., state distributions among taxa) under a given evolutionary model is lower for larger numbers of sampled characters (Table 1). Although systematists do not typically use $p(e \mid b)$ to evaluate the severity of their tests of evolutionary models, similar concerns are taken into consideration using statistical significance tests, such as the likelihood ratio test (e.g., Goldman, 1993a, b). D. Faith (pers. comm.) interprets likelihood ratio tests as direct measures of severity (i.e., $p(e \mid b)$), an interpretation with which we disagree. In any case, significance tests permit rejection of the null hypothesis only when data that deviate from the expectation under the null by as much or more than the observed data have a very low probability, given the null, which can be achieved only with a reasonable amount of data (i.e., a sufficient severe test). Moreover, the power of such tests (i.e., the probability of rejecting the null hypothesis when it is false, which is also related to severity) increases with increasing sample size.

TABLE 1. Probabilities of proportionally similar bodies of evidence for different amounts of data given particular hypotheses. Larger amounts of data are associated with lower probabilities of the evidence given the hypothesis, which indicates a more severe test when the hypothesis in question constitutes the background knowledge. In the first example, involving a coin, the probabilities are for data consisting of equal numbers of heads and tails, given that the coin is unbiased (the hypothesis for which the data have the highest probability). In the second example, involving a phylogeny, the probabilities are for data consisting entirely of two-state characters exhibiting the pattern 1100 in taxa ABCD, given the tree ((A,B),(C,D)) (the tree for which the data have the highest probability) and a model that assumes equal frequencies of states 0 and 1 and equal rates of change among all characters. Probability values in the phylogenetic example were calculated from the negative ln likelihoods obtained using PAUP* version 4.0b4a (Swofford, in prep.).

Hypothesis	Sample Size (N)	Evidence	Probability of the Evidence
$P_{heads} = P_{tails}$	2	1 Head: 1 Tail	0.500
	4	2 Heads: 2 Tails	0.375
	20	10 Heads: 10 Tails	0.176
Equal freqs.,	2	Both 1100	6.25×10^{-2}
Equal rates,	4	All 1100	3.91×10^{-3}
((A,B),(C,D))	20	All 1100	9.09×10^{-13}

One might also wish to evaluate the severity of tests in which the rival hypothesis are alternative phylogenetic trees. In this case, however, $p(e \mid b)$ cannot be used to evaluate severity, because the background knowledge (method/model) cannot be separated completely from the hypothesis (tree). Although the methods/models do not assume any particular tree, an assumption that the relationships of interest take the basic form of a tree is integral to any of these methods/models. Therefore, calculating either a parsimony or a likelihood score that corresponds with the probability of the evidence given a *phylogenetic* parsimony or likelihood method/model in the absence of a tree is impossible, lending further credence to our conclusion that $p(e \mid b)$ is to be ignored in the case of standard phylogenetic analyses. Nevertheless, one can evaluate the severity of tests using a probability analogous to $p(e \mid b)$. This evaluation can be accomplished with a probability based on an unconstrained likelihood model (e.g., Goldman, 1993a) or on the model used in the analysis in conjunction with a particular tree or set of trees. These probabilities are not to be equated with $p(e \mid b)$, meaning that they should not be incorporated into the corroboration expression (Eq. 2), but they can nevertheless be used to evaluate the severity of tests. As in the case of $p(e \mid b)$, these probabilities are expected to decrease (indicating a more severe test) with increasing amounts of data.

In summary, our conclusion that the term $p(e \mid b)$ can be ignored in the case of standard phylogenetic analyses is fully consistent with the role of that term in Popper's corroboration expression. The term $p(e \mid b)$ measures one component of corroboration, the severity of tests. Standard phylogenetic analyses are concerned with another component of corroboration, the evaluation of rival hypotheses according to the results of a single test. No attempt is made to evaluate the severity of the test, which is the purpose of $p(e \mid b)$. Moreover, because the same test is applied to all of the alternative hypotheses, $p(e \mid b)$ is constant. Finally, because the assumption of a tree is integral to phylogenetic methods/models, $p(e \mid b)$—the probability of the evidence given a phylogenetic method/model but without a tree—cannot be calculated. Nevertheless, the severity of tests of alternative trees can be evaluated by using probabilities other than $p(e \mid b)$.

Incidentally, the observation that $p(e \mid b)$ and analogous quantities decrease with increasing sample size indicates that a researcher who uses likelihood methods is not simply following the "rule 'Obtain high probabilities!'" (Popper, 1983:223), a view misleadingly associated with likelihood by Siddall and Kluge (1997:314) when they characterized likelihood as seeking the hypothesis with the highest probability and labeled it "verificationist." According to Popper (1959:399): "*Science does not aim, primarily, at high probabilities. It aims at a high informative content, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or very little.* A high degree of probability is therefore not an indication of 'goodness'—it may be merely a symptom of low informative content."

Popper's statement is not at odds with the law of likelihood. For one thing, his statement (as well as Siddall and Kluge's misleading criticism) concerns the probability of a hypothesis, whereas likelihood is the probability of the evidence (see *Popper's views on likelihood*). Moreover, the practitioner of likelihood seeks high probabilities only in a relative, not an absolute, sense. The goal is not to obtain the highest possible probability but rather to determine the hypothesis for which a given body of data has the highest probability. Consequently, the practitioner of likelihood may actually strive to obtain *low* probabilities in an absolute sense. Because the probability of the evidence tends to decrease with increasing amounts of data, any researcher who advocates collecting more rather than less data is effectively seeking to obtain lower rather than higher absolute probabilities.

p(e | b) *and PTP*

Our conclusions about $p(e \mid b)$ run counter to those of Faith and Cranston (1992; see also Faith, 1992), who equated $p(e \mid b)$ with the PTP value—that is, the probability of an optimal tree length as short as or shorter than the length of the optimal tree for the observed data given the null hypothesis of random data. We dispute neither the usefulness of the PTP for testing hypotheses about structure in a data matrix nor the compatibility of such tests with Popper's concept of corroboration. Nevertheless, we take issue with Faith and Cranston's interpretation of PTP as representing $p(e \mid b)$ in Popper's corroboration expression (see also Farris, 1995; Carpenter et al., 1998), an interpretation that confuses the hypothesis and background knowledge of different tests.

Faith and Cranston devised the PTP to test the null hypothesis that the states within characters are distributed randomly among taxa, and they equated the assumptions of this null hypothesis with the background knowledge (b) in Popper's corroboration expression. These two positions are logically incompatible. According to Popper, the background knowledge consists of hypotheses not currently being tested (see *The relationship between corroboration and likelihood*). Consequently, a hypothesis—including a null hypothesis—cannot both be the subject of a test (in this case, the PTP test) and at the same time form part of the background knowledge. Contrary to Faith and Cranston's interpretation, the PTP value does not correspond with $p(e \mid b)$ but with $p(e \mid hb)$. The correspondence, however, is not exact; the PTP value is not the probability of the evidence (e) itself—that is, the probability of the score of the optimal tree—but the cumulative probability of all scores as good or better than that of the optimal tree (Farris, 1995).

On the other hand, the PTP is used to test an assumption—that is, part of the background knowledge (b)—adopted in a test that we have called a standard phylogenetic analysis. Such an analysis is used to evaluate a different hypothesis or set of hypotheses than the null hypothesis of the PTP test. Specifically, it is used to evaluate a set of alternative trees. As Faith and Cranston noted, a standard phylogenetic analysis can yield an optimal tree or trees even from random data. Therefore, if such an analysis is to be considered a meaningful test of the hypotheses (trees), one must assume that the data are not random, that they exhibit significant non-random structure. An assumption of non-random data thus forms part of the background knowledge for the test of the alternative trees, and in accordance with Popper's views (see *Evaluation of Alternative Phylogenetic Methods or Models*), this assumption can itself be tested. Testing this assumption is the role of the PTP test. A low PTP value means a low probability of obtaining a score as good as or better than that of the optimal tree for the observed data, given the null hypothesis that the character state distributions among taxa are random. This low probability translates to a low degree of corroboration for the null hypothesis, which implies a high degree of corroboration for the mutually exclusive alternative hypothesis that the character state distributions are *not* random. And this high degree of corroboration for the alternative hypothesis in turn justifies use of that hypothesis as an assumption—that is, part of the background knowledge—in an evaluation of alternative trees.

Thus, contrary to Faith and Cranston's view, a low PTP value is desirable not because it represents a low $p(e \mid b)$ for a test of alternative trees, but because it represents a low $p(e \mid hb)$ for a different test in which the hypothesis (in this case a null hypothesis) must be rejected to justify use of a mutually exclusive alternative hypothesis as an assumption (i.e., part of b) in the test of alternative trees. Although the PTP is used to test part of the background knowledge required by the test of alternative trees, it cannot be equated with $p(e \mid b)$ in the latter test because two entirely different tests are involved with regard to both hypotheses and evidence. In a test of alternative trees, the hypothesis (h) is a tree, and the evidence (e) for which the probability is calculated consists of an observed distribution of character states among taxa. In a PTP test, the hypothesis (h') is a null hypothesis of random character state distributions, and the evidence (e') for which the probability (i.e., PTP) is calculated is the score of the optimal tree for the observed (unpermuted) data. Therefore, the permutation tail probability cannot possibly be equated with $p(e \mid b)$ in a test of alternative trees, because it is the probability of entirely different evidence (e'). In sum, the PTP test fits squarely into Popper's corroboration, but not in the manner described by Faith and Cranston.