The measurement of test severity, significance tests for resolution, and a unified philosophy of phylogenetic inference

Kevin de Queiroz

Accepted: 25 November 2003

Queiroz, K. de (2004). The measurement of test severity, significance tests for resolution, and a unified philosophy of phylogenetic inference. — *Zoologica Scripta*, *33*, 463–473.

The philosopher Karl Popper described a concept termed *degree of corroboration*, *C*, for evaluating and comparing hypotheses according to the results of their tests. *C* is, fundamentally, a comparison of two likelihoods: $p(e \mid hb)$, the likelihood of the hypothesis (h) in conjunction with the background knowledge (b), and $p(e \mid b)$, the likelihood of b alone. *C* is closely related to the likelihood ratio of nested hypotheses. When phylogenetic analysis is interpreted as an attempt to assess *C* for a phylogenetic tree (the hypothesis, h), several interpretations have been given for $p(e \mid b)$. Here I describe a new interpretation that equates $p(e \mid b)$ with the probability of the data in the absence of a hypothesis of phylogenetic resolution, that is with the likelihood of an unresolved or polytomous tree. Under this interpretation, *C* for a fully or partially resolved phylogenetic tree is the likelihood of that tree minus the likelihood of the corresponding unresolved tree. These same two likelihoods can be used in a likelihood ratio test (LRT) to assess the significance of the degree of corroboration of the hypothesis of phylogenetic resolution. This LRT for resolution is closely related to permutation tests for phylogenetic structure in the data, because data that evolved on a true polytomous tree are expected to be phylogenetically randomized. It therefore reconciles the interpretation of the evidence (e) as the distribution of character states among taxa (rather than the score of the optimal tree) with the interpretation of permutation tests as methods for assessing *C*. Likelihood methods are (contrary to the views of some commentators) central to understanding how Popper's *C* applies to phylogenetic hypotheses, and they form the foundation of a unified and inclusive philosophy of phylogenetic inference.

Kevin de Queiroz, Department of Systematic Biology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560–0162, U.S.A. E-mail: dequeiroz.kevin@nmnh.si.edu

Introduction

Phylogenetic methods have been interpreted (e.g. Faith 1992, 1999; Kluge 1997a,b; de Queiroz & Poe 2001) as methods for evaluating alternative phylogenetic hypotheses in terms of their *degree of corroboration*, *C*, a general measure developed by the philosopher Karl Popper (1959, 1962, 1983) for evaluating and comparing hypotheses according to the results of their tests. The assessment of test severity is central to assessing *C*, and constitutes the most important difference between that concept and Fisher's (1921, 1970) concept of *likelihood*, *L*, upon which *C* is otherwise based (e.g. Popper 1959: 388, 410, 414; 1983: 252).

In Popper's (1983) definition of *C* (p. 240), test severity is measured by the term $p(e \mid b)$: the probability of the evidence e given the background knowledge b (assumptions) alone, that is, in the absence of the hypothesis of interest h (p. 238). When phylogenetic analysis has been interpreted as the evaluation of alternative phylogenetic hypotheses in terms of *C*, the assessment of test severity — that is the calculation of $p(e \mid b)$ — has received several conflicting interpretations (e.g. Faith 1992, 1999; de Queiroz & Poe 2001; Kluge 2001), none of which is entirely satisfactory (e.g. Farris 1995; de Queiroz & Poe 2001, 2003; Faith & Trueman 2001; Farris *et al.* 2001).

In this paper, I propose a direct method for calculating $p(e \mid b)$, which forms the basis of a previously proposed likelihood ratio test for phylogenetic resolution (e.g. Felsenstein 1988; Ota *et al.* 1999, 2000). The close relationship between this test and the permutation tail probability (PTP) test of Faith & Cranston (1991) reconciles two seemingly conflicting interpretations of $p(e \mid b)$, resulting in a unified philosophy of phylogenetic inference.

Degree of corroboration

Popper (1983: 240, 242) presented two definitions of *C*, the simpler of which is $C(h,e,b) = p(e \mid hb) - p(e \mid b)$. This formula states that the degree of corroboration (*C*) of the hypothesis

of interest (h) by the evidence (e) in the presence of certain background knowledge or assumptions (b) is equal to the probability of e given h in conjunction with b minus the probability of e given b alone. Popper also provided a more complex definition that divides $p(e \mid hb) - p(e \mid b)$ by the 'normalization factor' $p(e \mid hb) - p(eh \mid b) + p(e \mid b)$ so that $C = -1$, rather than ≈ 0 in the presence of falsifying evidence. In the present discussion, I will (for the most part) follow the simpler definition, which corresponds with the numerator of the more complex definition and thus forms the crux of the concept.

In both of these definitions, the two central terms, $p(e \mid hb)$ and $p(e \mid b)$, are likelihoods (Popper 1959: 388, 410–414; 1983: 252; de Queiroz & Poe 2001, 2003). The term $p(e \mid hb)$, the probability of e given h in conjunction with b, is what Fisher (1921; 1970) called the *likelihood* (L), of the hypothesis. In contrast to Popper's terminology, the standard definition, $L(h \mid e) = p(e \mid h)$, does not contain the separate term b but treats it (including those components commonly termed *the model*) as part of h. Therefore, h in the definition of L corresponds with hb in the definition of C (de Queiroz & Poe 2001), which, defined as $p(e \mid hb) - p(e \mid b)$, is the likelihood of h minus the likelihood of b; or, in Popper's terminology, the likelihood of h in conjunction with b minus the likelihood of b alone.

In Popper's philosophy (Popper 1959, 1962, 1983), the comparison of these two probabilities is critically important. Evidence (e) that has a high probability given a hypothesis (h) in conjunction with particular background knowledge (b) can only be considered to provide meaningful support for h if it does not have a similarly high probability given b alone, that is in the absence of h (Popper 1983: 237). In other words, h can only be considered significantly corroborated if $p(e \mid hb) \gg p(e \mid b)$, and therefore an assessment of $p(e \mid b)$ is crucial to assessing the degree of corroboration of h. This proposition bears an obvious similarity to the idea from classical statistics that the probability of the data must differ to some accepted degree from an expectation under a null hypothesis if the data are to be considered to provide significant support for a mutually exclusive alternative hypothesis.

Degree of corroboration and statistical inference

In fact, Popper's C bears much more than this general similarity to methods of statistical inference. In this section, I will describe the relationship between C and three statistical concepts and/or methods: Fisher's (1921, 1970) *likelihood*, L, Edwards' (1972) *support*, S, and the *significance tests* of classical statistical inference. Although Popper noted in several places that C is based on L (e.g. Popper 1959: 388, 410, 414; 1983: 252) and discussed the testing of statistical hypotheses in general (Popper 1959: chapter VIII, appendix *ix), he does not seem to have addressed the relationship of C to S (i.e. likelihood ratios) or to significance tests.

Popper stated very explicitly that C is based on L. For example, '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" (in light of the evidence y, or given y …)' (Popper 1959: 388). Similarly, 'these two definitions [of *degree of corroboration*] … are based on $p(b,a)$ [i.e. $p(e \mid h)$] — called the likelihood of a with respect to b by R. A. Fisher — rather than upon $p(a,b)$ [i.e. rather than upon the probability of the hypothesis given the evidence, $p(h \mid e)$]' (Popper 1983: 252). Thus, as noted in the previous section, both of the conditional probabilities that define C are likelihoods: the likelihood of h in conjunction with b, $p(e \mid hb)$, and the likelihood of b alone, $p(e \mid b)$.

Although C is based on L, there is an important difference between the degree of corroboration of a hypothesis $C(h,e,b)$ and the likelihood of that hypothesis, $L(h \mid e)$. As noted above, for Popper $L(h \mid e)$ (represented by the term $p(e \mid hb)$ in the definition of C) is, by itself, inadequate for evaluating $C(h, e, b)$. The reason is that a high value of $p(e \mid hb)$ may have little or nothing to do with h. In other words, the probability of e may be high even without h, that is in the presence of b alone. According to Popper (1983: 237), 'if e should be *probable*, in the presence of b alone … , then its occurrence can hardly be considered as significant support of h'. Consequently, $p(e \mid hb)$ can only be considered an adequate measure of C in cases in which the probability of e given the background knowledge alone, $p(e \mid b)$, is very small (Popper 1959: 413–414). More generally, h can only be considered well-corroborated by e if $p(e \mid hb) \gg p(e \mid b)$. For this reason, the central idea in Popper's C is a comparison of the likelihood of h in conjunction with b, $p(e \mid hb)$, with the likelihood of b in the absence of h, $p(e \mid b)$.

Although the term h represents the hypothesis of interest, both hb and b in the definition of C represent hypotheses. The background knowledge b consists of 'assumptions' or 'theories not under test' (Popper 1962: 238, 1983: 252), which are themselves hypotheses. So b represents a hypothesis (or several hypotheses in conjunction) not including h. Moreover, according to Popper (1983: 236), 'b must be consistent with h'. Given that b is consistent with h and differs from hb in lacking the hypothesis of interest h, it represents a special case of hb. Thus, hb and b represent more and less general (i.e. nested) hypotheses. It follows that the central idea of Popper's C, $p(e \mid hb) - p(e \mid b)$, is a comparison of the likelihoods of nested hypotheses.

Statisticians have also developed a method for comparing two hypotheses in terms of their likelihoods. It is called by Edwards (1972: 31) *support*, S, 'defined as the natural logarithm of the likelihood ratio' of the two hypotheses. In the case of nested hypotheses, the likelihood ratio is, using Popper's symbols, $p(e \mid hb)/p(e \mid b)$. Thus, C and S applied to nested hypotheses are both comparisons of $p(e \mid hb)$ and $p(e \mid b)$. The main differences are that C is the *difference* between $p(e \mid hb)$

and $p(e|b)$ normalized (in the complex definition) so that it varies between −1 and +1, while S is the *ratio* of $p(e|hb)$ to $p(e|b)$ transformed using the natural logarithm. This fundamental similarity confirms the belief of Edwards (1972: 211), who stated 'On the [wide subject of the philosophy of science] I have the impression that the Method of Support 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 the wider field.'

The likelihood ratio (or its natural logarithm) is commonly used as a test statistic in a significance test, called the *likelihood ratio test* (e.g. Neyman & Pearson 1928a,b; Kendall & Stuart 1979) in the context of the classical approach to statistical inference — that is, in the context of the ideas of repeated sampling and type I and type II errors described by Neyman & Pearson (1933). C varies between 0 and +1, or between −1 and +1 (normalized), but does not specify what range of values corresponds to a significant degree of corroboration. However, because of the close relationship between C and the likelihood ratio of nested hypotheses, the likelihood ratio test (LRT) applied to nested hypotheses is basically a test of the significance of $C(h,e,b)$ (i.e. of the significance of the value of C obtained in a particular test of h).

In light of the above discussion, three points seem worth emphasizing concerning the relationship between Popper's C and various statistical concepts and tests. First, C is based on (i.e. defined in terms of) likelihood, L. Second, C is an alternative metric for measuring the support, S, of a more general hypothesis relative to a less general one. Third, if C were to be used as a test statistic in the context of the ideas of repeated sampling and type I and type II errors, then the resulting significance test would be analogous to a likelihood ratio test. Thus, Popper's method for assessing the degree of corroboration of a hypothesis is basically a method of statistical inference. It is based on the concept of likelihood (as are most approaches to statistical inference), it is fundamentally similar to the likelihood ratio of nested hypotheses, and it is therefore also closely related to likelihood ratio tests.

Because Popper's method does not incorporate the ideas of repeated sampling and type I and type II errors, or prior probabilities, it bears a particularly close resemblance to an approach to statistical inference based more or less exclusively on the comparison of the likelihoods of alternative hypotheses (e.g. Fisher 1925, 1970; Edwards 1972; Royall 1997), an approach that has been called *likelihood inference* to distinguish it from the classical and Bayesian approaches (e.g. Barnett 1999). On the other hand, the significance tests of classical statistics, which are based on the ideas of repeated sampling and type I and type II errors, represent explicit attempts to reject a null hypothesis and are therefore highly congruent with Popper's general falsificationist philosophy (Gillies 2000: 145–150).

The measurement of $p(e|b)$ in phylogenetics

When Popper's C has been used as a context for understanding methods of phylogenetic analysis, several conflicting interpretations of $p(e|b)$ have been proposed (e.g. Faith 1992, 1999; de Queiroz & Poe 2001; Kluge 2001). Both de Queiroz & Poe (2001) and Faith & Trueman (2001) discuss C and test severity in the context of several types of phylogenetic analyses, including tests of hypotheses about evolutionary processes described by model parameters (de Queiroz & Poe 2001) and tests of monophyly (Faith & Trueman 2001). In this paper, I will restrict my considerations to the case that de Queiroz & Poe (2001) called a *standard phylogenetic analysis*, defined as an attempt to identify an optimal tree using a single data set (in the case of character data, a single taxon × character matrix) and a single phylogenetic method and its associated implicit and explicit assumptions. In particular, I will address the assessment of C for the optimal tree identified in such an analysis. Most of the conflicts among alternative interpretations $p(e|b)$ relate to this type of analysis.

In this discussion, I will adopt the interpretation of de Queiroz & Poe (2001) concerning the correspondence between the components of a standard phylogenetic analysis and the terms in the definition of C. In particular, I will equate a tree (commonly the optimal tree) with h, the taxon × character matrix with e, and the analytical method and any associated models with part of b. Although this interpretation differs from that of Faith and collaborators (Faith 1992, 1999; Faith & Cranston 1992; Faith & Trueman 2001), in which e is interpreted not as the taxon × character data matrix but as the score of the optimal tree, I will argue that despite these differences, the general conclusions about assessing C for a phylogenetic hypothesis proposed by those authors are highly compatible with the methods that I discuss in this paper.

For the case of a standard phylogenetic analysis, under the interpretation of h, e, and b just described, the first term in the definition of C, $p(e|hb)$, is the probability of the data (distribution of character states among taxa, or the different character patterns and their numbers) given a particular tree and the assumptions of the phylogenetic method used. This is the same thing as the likelihood of the tree, which can therefore be calculated using a phylogenetic likelihood method with its intrinsic probabilistic model (de Queiroz & Poe 2001). If a nonprobabilistic phylogenetic method is used, then $p(e|hb)$ must be calculated using a probabilistic equivalent. For example, Tuffley & Steel (1997) proposed a probabilistic model that is equivalent to (i.e. gives the same results as) a Fitch (unordered) parsimony method with equally weighted transformations both within and among characters (see also Goldman 1990). Regardless of whether it is based on likelihood or some other optimality criterion, a standard phylogenetic analysis, by itself, contains no component that corresponds

with $p(e|h)$ (Faith & Cranston 1992; de Queiroz & Poe 2001; Faith & Trueman 2001). Consequently, such an analysis assesses only the *relative* degree of corroboration of alternative phylogenetic hypotheses (de Queiroz & Poe 2001). It is possible, however, to supplement a standard phylogenetic analysis with an additional analysis that assesses $p(e|h)$ and thus the full degree of corroboration as described by Popper.

Most authors agree that h in a standard phylogenetic analysis is a tree (e.g. de Queiroz & Poe 2001; Faith & Trueman 2001; Farris *et al.* 2001). If we interpret that tree as any tree whatsoever, then $p(e|h)$ is the probability of the data given the assumptions of the phylogenetic method in the complete absence of a tree (de Queiroz & Poe 2001). This interpretation presents a problem regarding the calculation of a value for $p(e|h)$, because a phylogenetic model cannot be used to calculate the probability of a particular distribution of character states among taxa (i.e. a set of character patterns) in the complete absence of a tree. For models used in phylogenetic analysis, the probability of the data is based on estimates of the probabilities of character state changes along the branches of a tree, so the complete absence of a tree precludes the use of those models. Conversely, although it is possible to calculate the probability of the data in the absence of a tree, doing so requires an entirely different (i.e. nonphylogenetic) model. For example, Goldman's (1993) unconstrained model, in which the probability of a given character pattern is simply the frequency of that pattern in the entire set of characters, does not require, or even use, a tree. This model, however, is not the same model used to calculate $p(e|hb)$, the probability of the data given a particular tree, in a phylogenetic analysis, so it cannot be used to calculate $p(e|h)$ for such an analysis.

Alternatively, because a common goal of a standard phylogenetic analysis is to identify a fully or partially resolved tree, it is possible to interpret h not as any tree whatsoever, which might be completely unresolved, but as a fully or partially resolved tree — a hypothesis of phylogenetic resolution. In this context, $p(e|h)$ is the probability of the data given the assumptions of the phylogenetic method and an unresolved tree, that is, the absence of the hypothesis of resolution. The probability $p(e|h)$ can therefore be calculated by constraining the relevant internal branches of the tree to have zero length (Fig. 1). A completely unresolved or polytomous (star) tree constraint is appropriate for cases involving unrooted trees and no assumptions beyond those associated with the phylogenetic method and its underlying probabilistic model — that is, under no assumptions whatsoever about relationships (Fig. 1A). Under an assumption of ingroup monophyly, the unrooted tree would be allowed to have a single internal branch with nonzero length separating the ingroup and outgroup taxa (Fig. 1B). If the tree is rooted and ingroup monophyly is assumed, then the tree would have a single resolved monophyletic group (the ingroup) with the branch subtending it allowed to have nonzero length (Fig. 1C). Any additional assumptions about ingroup or outgroup relationships allow additional internal branches to have nonzero length. In any case, under the interpretation of h as a hypothesis of resolution, it is possible to calculate a value for $p(e|h)$. The probability $p(e|h)$ is simply the likelihood of the constrained (unresolved or polytomous) tree.

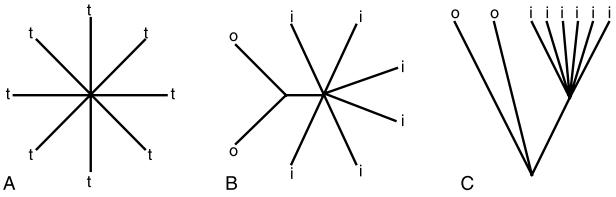

Fig. 1 A–C. Trees corresponding to different assumptions in a standard phylogenetic analysis. —A. Completely unresolved and unrooted (star) tree corresponding to no assumptions about phylogenetic relationships. —B. Partially resolved unrooted tree corresponding to an assumption of ingroup monophyly. —C. Partially resolved rooted tree corresponding to an assumption of ingroup monophyly. Abbreviations: t = taxon (outgroup and ingroup taxa not distinguished), o = outgroup taxon, i = ingroup taxon.

The interpretation of $p(e|h)$ as the likelihood of a polytomous tree is counterintuitive in that h is supposed to consist of assumptions that are made when testing h, but one does not assume that the optimal tree is polytomous when estimating that tree. Nevertheless, a polytomous tree does indeed represent the absence of the hypothesis of resolution represented (for example) by the optimal tree. The likelihood of this unresolved tree is the probability of the data under a given probabilistic model assuming nothing more about relationships than that the taxa are related according to a general diverging or tree-like model of evolution — that is, in the absence of a hypothesis about the specific pattern of relationships among those taxa. So a polytomous tree is effectively assumed when estimating an optimal tree, not in the sense that the optimal tree is assumed to be unresolved, but in the sense that the analysis starts from a point that assumes that the relationships to be inferred conform to a tree-like model while assuming nothing about the specific relationships among the taxa in question. This situation is manifested in heuristic methods based on star decomposition, in which the optimal tree is estimated through the successive uniting of taxa starting from an initial star tree (reviewed by Swofford *et al.* 1996). Other heuristic methods build up a tree by the successive addition of taxa and therefore start from a point (a small subset of the taxa) that has no relevance to assessing the probability of the evidence in the absence of h.

Tests for phylogenetic resolution

Now that a method for calculating a value for $p(e|h)$ has been identified, it is a straightforward matter to calculate a value

for the degree of corroboration, $C = p(e \mid hb) - p(e \mid b)$, of a resolved phylogenetic tree, h. C is simply the likelihood of the resolved tree (normally the optimal tree), $p(e \mid hb)$, minus the likelihood of the unresolved (polytomous) tree, $p(e \mid b)$. This value, however, will merely be a number greater than zero and less than one, and it is both more useful and more in the spirit of Popper's general falsificationist philosophy to devise a significance test based on the two probabilities. A significance test provides an explicit criterion (α) for assessing whether the resolved tree (h) is to be considered significantly corroborated and the unresolved tree effectively falsified (Popper 1959: 189–205). C itself could be used as the test statistic; however, because both of its defining probabilities are likelihoods, a more commonly adopted test statistic is the likelihood ratio $\lambda = L(h_1)/L(h_2)$ or $\delta = 2\ln[L(h_1)/L(h_2)] = 2(\ln[L(h_1)] - \ln[L(h_2)])$ (e.g. Neyman & Pearson 1928a,b; Kendall & Stewart 1979; see also Goldman 1993).

When the likelihood ratio is used as the basis of a significance test for phylogenetic resolution, the constrained (unresolved or polytomous) tree represents the null hypothesis, h_0, in that it posits no resolution among the taxa — or zero length for the internal branches. The unconstrained (usually resolved) tree represents the alternative hypothesis, h_A. Recall that in the definition of C, $p(e \mid b)$ corresponds to the likelihood of the constrained (unresolved) tree, and $p(e \mid hb)$ to the likelihood of the unconstrained (resolved) tree or hypothesis of resolution. Therefore, $L(h_A) = p(e \mid hb)$ and $L(h_0) = p(e \mid b)$ (de Queiroz & Poe 2003). If we set $h_1 = h_A$ and $h_2 = h_0$, then the test statistic $\lambda = L(h_1)/L(h_2) = L(h_A)/L(h_0) = p(e \mid hb)/p(e \mid b) =$ the likelihood of the unconstrained (resolved) tree divided by the likelihood of the constrained (unresolved) tree.

A likelihood ratio test comparing an unconstrained tree with a tree in which a single branch is constrained to have zero length was initially proposed by Felsenstein (1987). This test has been discussed subsequently (e.g. Felsenstein 1988; Gaut & Lewis 1995; Antezana & Hudson 1999; Slowinski 2001) and generalized to trees with multiple zero-length branches (e.g. Swofford *et al.* 1996: 506; Ota *et al.* 1999, 2000). The constrained (less resolved) tree is a special case of the unconstrained (more resolved) tree in that certain branch lengths are fixed at one of their many possible values (this situation may be counterintuitive if one is accustomed to thinking of the resolved tree as one of many possible resolutions of the unresolved tree). Consequently, the hypotheses are nested.

For LRTs involving nested hypotheses, it is common to assess significance using the χ^2 distribution with degrees of freedom equal to the difference in the number of parameters (k) between the two hypotheses (Neyman & Pearson 1928a,b; Silvey 1975; see also Swofford *et al.* 1996). In this case, that number is the number of branches constrained to have zero length in the constrained tree, which is the same as the total number of internal branches in the unconstrained tree (= $t - 3$, where t is the number of taxa) if the constrained tree is completely unresolved. However, the constraints used in the null hypothesis are branch lengths of zero, which correspond with boundary conditions for the k constrained branch-length parameters, and therefore the χ^2 distribution with k degrees of freedom is inappropriate (Self & Liang 1987; Ota *et al.* 1999, 2000; Whelan & Goldman 1999). To solve this problem, appropriate values can be calculated by averaging the values of the standard χ^2 distribution for different degrees of freedom (Self & Liang 1987; Ota *et al.* 1999, 2000; Goldman & Whelan 2000), assuming that the internal branch length estimates are uncorrelated under h_A. Alternatively, a probability distribution can be generated using simulations (Huelsenbeck & Rannala 1997; Whelan & Goldman 1999), though this approach is susceptible to type I (rejection) errors (Buckley 2002).

Analogous tests for resolution can be used with likelihood as well as other optimality criteria. Several tests that have been developed for comparing alternative topologies in terms of differences in their ability to explain the data under various optimality criteria can be used to compare a resolved tree (e.g. the optimal tree) with an unresolved (polytomous) tree. These tests include those proposed by Prager & Wilson (1988) based on binomial probabilities, by Templeton (1983) based on the Wilcoxon signed ranks test, and by Kishino & Hasegawa (1989) based on the paired t-test, as well as the modification of the Kishino–Hasegawa test proposed by Shimodaira & Hasegawa (1999). Although use of these alternative methods to test for significant resolution does not involve calculating a value for $p(e \mid b)$, the results can nonetheless be considered to provide a measure of the degree of corroboration of a phylogenetic hypothesis (i.e. of a hypothesis of resolution). If the test result is significant, then the resolved tree, representing h, can be considered to have a significant degree of corroboration in the sense that it explains the data significantly better, under a specified optimality criterion and model (and α value), than does the unresolved tree that corresponds to part of h.

A couple of things should be noted about resolution tests, whether based on likelihood ratios or other test statistics. First, these tests are susceptible to type I (rejection) errors if the analytical method or model is inappropriate (Gaut & Lewis 1995; Ota *et al.* 2000). In particular, if the method or model is susceptible to long-branch attraction (Felsenstein 1978; Hendy & Penny 1989) and branch lengths are sufficiently unequal, the test will tend to reject a true null hypothesis (polytomous tree) as the result of false resolution resulting from long-branch attraction (Fig. 2). Second, rejection of the null hypothesis constitutes evidence only that the resolved tree is significantly better corroborated (i.e. explains the data significantly better) than an unresolved tree. Thus, when the resolved tree is the optimal tree, rejection of the

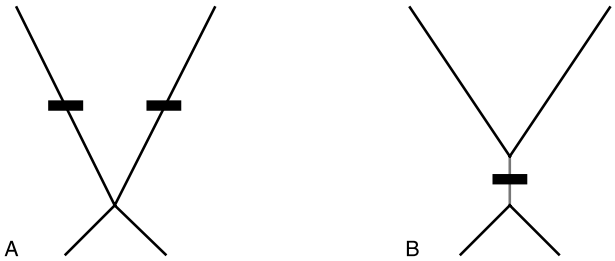

Fig. 2 A, B. An inappropriate method (model) can cause a resolution test to reject a true null hypothesis (polytomous tree), representing a type I error. —A. True polytomous tree with two long and two short branches and parallel state changes (horizontal bars) in a particular character on the long branches. —B. Incorrect reconstructed tree with parallel state changes interpreted as a single change (horizontal bar) on the internal branch. If the number of such characters is sufficiently great, a resolution test will yield a significant difference between the resolved and unresolved trees, leading to rejection of the true null hypothesis (type I error).

null hypothesis does not constitute evidence that the optimal tree is significantly better corroborated than other resolved trees, which can only be established by direct comparisons with those alternative trees. In most cases, the optimal tree will not be significantly better corroborated than at least some of the other resolved trees (those that are most similar to it). This situation does not, however, call into question the status of the optimal tree as the best estimate of the phylogeny. An analysis that attempts to identify an optimal tree, regardless of whether it is used as part of a significance test for resolution, is an example of estimation rather than hypothesis testing. Therefore, in the terminology of statistics, the optimal tree can be considered the best 'point' estimate of the phylogeny; in Popper's terminology, it can be considered to have the highest *relative* degree of corroboration.

It should also be noted that in a significance test for resolution, the hypothesis being tested is not the optimal tree h but the null hypothesis of a polytomous tree h_0, and the mutually exclusive alternative hypothesis h_A is not the optimal tree itself but a general hypothesis of resolution — the hypothesis that the internal branches do *not* all have zero length. The test is a standard LRT that asks whether the maximum probability that can be assigned to the data is significantly higher in the absence (h_A) vs. the presence (h_0) of a constraint corresponding with the null hypothesis (e.g. Huelsenbeck & Rannala 1997). Thus, the optimal tree, rather than representing h_A per se, represents the estimated tree for which the likelihood is maximal in the absence of the constraint — that is, the tree corresponding with $\max[L(h_A|e)]$. This minor difference is simply a consequence of adapting Popper's C (or Edwards' S) to the framework of significance testing. For the purpose of assessing C (or S) itself (i.e. outside of the significance testing framework), there is no need to identify a mutually exclusive alternative hypothesis h_A, and although systematists are often most interested in the optimal tree, h can be any fully or partially resolved tree.

Different interpretations of $p(e|h)$ and their reconciliation

The LRT for resolution discussed in the previous section provides a means for reconciling alternative interpretations of $p(e|h)$ in Popper's definition of C as it relates to standard phylogenetic analysis. One of these interpretations was proposed by Faith, Cranston, and Trueman (Faith & Cranston 1992; Faith 1992, 1999; Faith & Trueman 2001). Under the Faith–Cranston–Trueman (FCT) interpretation, $p(e|h)$ of C is equated with the permutation tail probability (PTP) value of Faith & Cranston's (1991) PTP test. The PTP test is used to test a null hypothesis of no hierarchical structure (no phylogenetic signal) in the data. It works by comparing the score of the optimal tree(s) for the observed data with a frequency distribution of optimal tree scores generated under the null hypothesis by randomly permuting the data — specifically, by randomly reassigning the observed character states within each character to the taxa in which they occur. The PTP value is the probability of obtaining an optimal tree score as good as or better than the score obtained for the observed data under the null hypothesis of randomized data.

The PTP test captures the spirit of Popper's C in attempting to assess test severity. It does so by asking whether the data contain sufficient hierarchical information (i.e. whether the test is sufficiently severe) to provide a meaningful degree of corroboration for the optimal tree. The reason that such a test is important for assessing C is that a standard phylogenetic analysis will almost always yield a resolved optimal tree, or a set of such trees the consensus of which is at least partially resolved, even if the data are random. Therefore, given that random data are not normally considered to confer a significant degree of corroboration on any tree, simply obtaining a resolved tree in a phylogenetic analysis does not, by itself, indicate that the hypothesis (tree) is well-corroborated. The PTP test addresses this concern by assessing whether the score of the optimal tree(s) represents a significant degree of corroboration. Under this test, the optimal tree(s) is only considered significantly corroborated if obtaining a score as good or better than its score is highly improbable given the null hypothesis of randomized data.

Despite the intuitive appeal of the FCT interpretation, there are several problems with equating $p(e|h)$ with PTP. For one thing, $p(e|h)$ is a point probability while PTP is a cumulative tail probability (Farris 1995; Farris *et al.* 2001). For another, no satisfactory explanation has been given for treating the null hypothesis of randomized data as part of h, as it is under the FCT interpretation. It is claimed that the 'null model [of the PTP test] provides the set of accepted

facts or assumptions [for an analysis]' (Faith & Cranston 1992: 254), even though those accepted assumptions (i.e. randomized data) are commonly rejected by the test. The main justification for equating PTP with $p(e|b)$ seems to be that a low value of PTP is associated with rejection of the null hypothesis, implying a significant degree of corroboration for the optimal tree, and similarly, a low value of $p(e|b)$ is associated with a high degree of corroboration C (e.g. Faith & Cranston 1992: 254). Thus, a convincing reason for associating PTP with $p(e|b)$ has been lacking, and the intuitive appeal of the FCT interpretation seems to result mainly from the general conformity of significance tests with Popper's falsificationist philosophy (see 'Degree of corroboration and statistical inference', above).

An alternative interpretation of $p(e|b)$ was proposed by de Queiroz & Poe (2001). Under the de Queiroz-Poe (QP) interpretation, a standard phylogenetic analysis contains no component corresponding with the assessment of test severity, $p(e|b)$, nor is it possible to calculate this particular probability in the context of such an analysis. The reason, described above (see 'The measurement of $p(e|b)$ in phylogenetics'), is that a phylogenetic method or model cannot be used to calculate the probability of the evidence e (the distribution of character states among taxa) in the absence of a tree. Nevertheless, de Queiroz & Poe (2001) described methods for measuring probabilities analogous to $p(e|b)$. In addition, they interpreted PTP not as a direct measure of test severity, $p(e|b)$, but as the (cumulative) tail probability used to test an assumption (a component of b) adopted in a standard phylogenetic analysis — specifically, the assumption that the data exhibit nonrandom hierarchical structure.

The QP interpretation avoids the problems with the FCT interpretation concerning the confusion of point vs. cumulative probabilities. On the other hand, the QP interpretation denies the possibility of identifying a value corresponding precisely with $p(e|b)$ and thus of ever fully reconciling standard phylogenetic analysis with Popper's method for assessing C. This situation creates no logical inconsistencies in that standard phylogenetic analysis is a method for estimation, while C is a method for hypothesis testing. Nevertheless, these two aspects of statistical inference (estimation and hypothesis testing) are not entirely separate from one another, and the QP interpretation seems to preclude fully integrating standard phylogenetic analysis into the framework of hypothesis testing and thus also of Popper's C.

Similarities between resolution and PTP tests

The method for measuring $p(e|b)$ described in the present paper (i.e. calculating the likelihood of an unresolved tree) and the LRT of resolution based on it, overcome the problems with both the FCT and the QP interpretations of test severity, $p(e|b)$, effectively merging the two interpretations into a single, unified view. Unification results because the methods in question reconcile the fundamental incompatibilities between the two interpretations, namely, the FCT proposition that the PTP test assesses C for the optimal tree (QP interpret the PTP test as a test of an assumption adopted in the analysis that identifies the optimal tree) and the QP proposition that e is a taxon × character matrix of character states (FCT interpret e as the score of the optimal tree). More specifically, these methods eliminate the inconsistencies that result (under the FCT interpretation) from equating PTP with $p(e|b)$. They provide the missing connection (under the FCT interpretation) between the null hypothesis of randomized data and the absence of b, and they overcome the limitation (under the QP interpretation) of not being able to calculate a value for $p(e|b)$.

The key to reconciling these alternative interpretations of $p(e|b)$ is provided by a close relationship between resolution and PTP tests, which is based, in turn, on a close relationship between the null hypothesis of a polytomous (unresolved) tree and that of randomized data. This relationship derives from the fact that characters that have evolved on a true polytomous tree are expected to exhibit phylogenetically uninformative, and in this sense effectively randomized, state distributions. On a tree with no (or all zero length) internal branches, all character state changes must occur on terminal branches, and all changes that produce shared derived states must result from homoplasy. Therefore, provided that homoplasy is randomly distributed among the terminal branches, derived character states should be shared randomly with respect to taxa. Moreover, although randomized data can evolve on a tree with internal branches, when the data are effectively randomized, that tree should be statistically indistinguishable from a polytomous tree. That is, the two trees should not differ significantly in terms of their scores under a given optimality criterion (provided that the phylogenetic method is appropriate — see below).

This situation can occur when all (variable) characters have high probabilities of change on all of the terminal branches, which might result from high rates of change, long temporal duration, or a combination of both factors. In any case, circumstances that result in failure to reject the null hypothesis of randomized data under a PTP test should also result in failure to reject the null hypothesis of an unresolved tree under a resolution test. Conversely, circumstances that lead to rejection of the null hypothesis under a PTP test should also lead to rejection of the null hypothesis under a resolution test. (I am here assuming that both tests employ the same phylogenetic model and ignoring possible differences in power. I am therefore interpreting PTP as a general statistic that can be used in conjunction with any phylogenetic optimality criterion, as opposed to a restricted one based on parsimony length, as in the original proposal.)

The relationship between resolution and PTP tests provides the missing justification for interpreting the null hypothesis of randomized character state distributions as part of *b* (e.g. Faith & Cranston 1992; Faith 1992; Faith & Trueman 2001). According to the interpretation proposed in this paper, the absence of *b* (e.g. the optimal tree) corresponds to an unresolved tree for the same taxa, which therefore forms part of *b*. This unresolved tree can also be treated as a constraint corresponding to a null hypothesis in a significance test (see 'Tests for phylogenetic resolution'), thus establishing the connection between *b* and a null hypothesis. Moreover, because data that evolved on a polytomous tree are expected to be phylogenetically randomized, a specific connection is established between a component of *b* (the unresolved tree) and the null hypothesis of randomized data adopted in the PTP test.

Differences between resolution and PTP tests

Despite the close relationship between PTP and resolution tests, the two types of tests evaluate different null hypotheses. Consequently, although they are generally expected to give congruent assessments of *C*, similar test results can have somewhat different implications. Table 1 summarizes the relationship between PTP and resolution tests for the four possible cases (combinations) regarding the truth or falsity of their respective null hypotheses under a stochastic process of character state change. The first and fourth cases are unproblematic in that both tests are expected to yield congruent and appropriate results. When both null hypotheses are true (case 1), both tests should fail to reject their respective null hypotheses; similarly, when both null hypotheses are false (case 4), both tests should reject their respective null hypotheses.

It should be noted that for both types of tests, appropriate test results when the null hypothesis is true depend on the appropriateness of the phylogenetic method and its associated assumptions. Thus, in case 1, where the data are effectively randomized and the true tree is polytomous, appropriate assumptions should result in the failure of both tests to reject their respective null hypotheses. Inappropriate assumptions, however, can cause both tests to reject true null hypotheses, representing type I errors. For example, if homoplasy is more common on long branches and some branches are sufficiently long, then methods that are susceptible to long-branch attraction (Felsenstein 1978; Hendy & Penny 1989) could cause a PTP test to yield a result indicating that phylogenetically uninformative data generated on a polytomous tree are informative (exhibit significant, nonrandom, hierarchical structure). For the same reason, such a method could cause a resolution test to yield a result indicating that a false resolved tree explains the data significantly better than does the true polytomous one.

Case 2, in which the null hypothesis of the PTP test is true and that of the resolution test is false, reveals the most important difference between the tests. When the data are effectively randomized, both tests should fail to reject their respective null hypotheses. In the case of the PTP test, this result is appropriate in that the null hypothesis is true — the data are randomized. In the case of the resolution test, however, the null hypothesis is false (the true tree has internal structure), so failure to reject it represents a type II error. As noted above (see 'Similarities between resolution and PTP tests'), this situation is expected to occur when the true tree has internal structure but the terminal branches are sufficiently long that the data are effectively randomized. Despite this difference between PTP and resolution tests, when either test fails to reject the null hypothesis, it is appropriate to interpret the optimal tree as not being significantly corroborated. Regardless of whether the true tree has internal structure, randomized data cannot confer a significant degree of corroboration on the optimal (or any other) resolved tree.

Case 3, the combination of a false null hypothesis for the PTP test (nonrandomized data) and a true null hypothesis for the resolution test (polytomous tree), should not occur. If the true tree is polytomous, the data should be phylogenetically

Table 1 The relationship between PTP and resolution tests. The cell entries describe the expected test results under four cases (1–4) corresponding to the four possible combinations concerning the truth or falsity of the null hypotheses of the two different types of tests. A stochastic process of character state change is assumed, and differences in power are not considered. Case 3 should not occur for reasons discussed in the text, though data that evolved on a polytomous tree may appear nonrandomized if branch lengths are sufficiently unequal and the phylogenetic method is misled by long-branch attraction.

	H_0 (PTP; Resolution)			
Test	(1) True; True (Randomized data; Polytomous tree)	(2) True; False (Randomized data; Non-polytomous tree)	(3) False; True (Non-randomized data; Polytomous tree)	(4) False; False (Non-randomized data; Non-polytomous tree)
PTP	Fail to reject*	Fail to reject	—	Reject
Resolution	Fail to reject*	Fail to reject (Type II error)	—	Reject

*This result assumes an appropriate phylogenetic method/model. If the phylogenetic method/model is inappropriate, the test may tend to reject a true null hypothesis (type I error).

uninformative (effectively randomized) for the reasons described above (see 'Similarities between resolution and PTP tests'), thus corresponding with case 1 (i.e. both null hypotheses are true). However, if the phylogenetic method is inappropriate, then the data may appear to exhibit significant nonrandom structure (and the optimal tree may appear to exhibit significant resolution) even when the true tree is polytomous. As noted above (see 'Tests for phylogenetic resolution'), the appearance of nonrandom hierarchical structure (phylogenetic information) in data that evolved on a polytomous tree can occur when there is variation in branch lengths and the phylogenetic method or model incorrectly interprets convergent or parallel changes on long branches as single changes on an internal branch.

The true tree is not normally expected to lack internal structure entirely. For this reason, it seems reasonable to interpret failure to reject the null hypothesis in a resolution test (at least when many branches are involved) as suggesting that the data are phylogenetically uninformative. This possibility could be investigated further by estimating the terminal branch lengths to determine if they exceed an expected randomization threshold, or by examining the power of the test to determine if sufficient data have been collected to be able to reject the null hypothesis under the assumption that the lengths of the internal branches exceed some specified minimum value (e.g. Poe & Chubb, 2004). In any case, and setting aside problems with the interpretation of negative results, the possibility of interpreting failure to reject the null hypothesis in a resolution test as indicative of phylogenetically uninformative (or insufficient) data highlights the relationship between resolution and PTP tests, as this conclusion is precisely the implication of failure to reject the null hypothesis in a PTP test.

The preceding discussion has assumed a stochastic process of evolutionary change in the context of a real phylogeny. For cases in which data are manufactured by human contrivance, resolution and PTP tests can yield conflicting results. For example, the PTP test has been criticized for returning a significant result when the data consist of characters that are highly incongruent with one another in a highly regular (i.e. nonrandom) pattern (e.g. Källersjö *et al.* 1992; Carpenter *et al.* 1998). Although the PTP test appropriately rejects the null hypothesis of randomized data (the data set in question, designated *Matrix Three* in the previously cited papers, clearly is not random), it seems inappropriate to interpret this result as indicating a significant degree of corroboration for the optimal trees, given that their strict consensus is completely unresolved. If these same data are evaluated with parsimony-based resolution tests that use the strict consensus of the optimal trees as the hypothesis of interest h, the tests will not, of course, reject the null polytomous tree in favour of the similarly unresolved strict consensus tree (though they will reject the null polytomous tree in favour of each individual highly resolved optimal tree).

Reconciliation

Because of the close relationship between PTP and resolution tests, the former can legitimately be considered to assess degree of corroboration through a consideration of test severity. Although PTP is not the same quantity as $p(e \mid h)$, it nonetheless serves as the basis for a significance test that is closely related to a test based on a direct estimate of $p(e \mid h)$. The close relationship between resolution and PTP tests thus provides an at least partial reconciliation of the FCT and QP interpretations of $p(e \mid h)$. Full reconciliation requires only minor changes to each of those interpretations. (I consider the necessary changes minor in that they compromise neither the fundamental FCT proposition that the PTP test, or something like it, is necessary to assess the degree of corroboration of the optimal tree identified in a standard phylogenetic analysis, nor the fundamental QP proposition that, in the context of a standard phylogenetic analysis, $p(e \mid hb)$ is the likelihood of a particular tree.) The FCT interpretation must be modified so that PTP is not considered strictly equivalent to $p(e \mid h)$ but instead is considered the tail probability in a significance test (the PTP test) that is analogous to a significance test for resolution based on $p(e \mid h)$. Likewise, the QP interpretation must be modified so that the absence of h, represented by the optimal tree, is interpreted not as the complete absence of a tree but instead as the absence of phylogenetic resolution.

A unified and inclusive philosophy of phylogenetic inference

Both de Queiroz & Poe (2001) and Faith & Trueman (2001) described what can be considered inclusive philosophies of phylogenetic inference in the terminology of Faith & Trueman (2001). Both views are inclusive in considering diverse phylogenetic methods to be philosophically justified and thus scientifically legitimate. The close correspondence between resolution and PTP tests described in this paper provides the basis for unifying these views into a single, inclusive philosophy of phylogenetic inference. This inclusive philosophy stands in contrast to the exclusive philosophy of authors (e.g. Siddall & Kluge 1997; Kluge 2001) who consider only parsimony-based methods to be scientifically legitimate. Although commonly defended in the context of C, the exclusive philosophy turns out to be inferior in terms of exemplifying Popper's method. In particular, it has so far provided no concrete method for calculating $p(e \mid h)$ or for assessing whether a particular data set provides a significant degree of corroboration for a particular phylogenetic tree. Instead, it equates test severity with a vague, nonprobabilistic, nonquantitative notion of 'traditional character reanalysis' (e.g.

Kluge 2001: 327), which implies the dubious proposition that simply reanalysing characters ought to result in a systematic decrease in the value of $p(e|h)$, thus indicating a more severe test (see de Queiroz & Poe 2003).

In contrast, the unified inclusive philosophy provides an explicit and direct method for calculating $p(e|h)$, which is simply the likelihood of an unresolved tree. It also provides explicit methods for assessing whether particular data provide a significant degree of corroboration for the optimal tree estimated from them, including the likelihood ratio and other tests for phylogenetic resolution discussed in this paper, as well as the PTP and other tests, such as that based on skewness of the distribution of tree scores (Hillis 1991; Hillis & Huelsenbeck 1992), for phylogenetically informative data. Moreover, in agreement with Popper's statements that $p(e|h)$ is inversely related to test severity and sample size (Popper 1959: 411, 413, 1983: 238), the likelihood of the unresolved tree is expected to decrease with increasing sample size — that is, with increasing numbers of characters (see de Queiroz & Poe 2001). Similarly, in significance tests for both resolution and phylogenetically informative data, test severity, as manifested in the related concept of statistical power, is expected to increase with increasing sample size (de Queiroz & Poe 2001, 2003).

Given this situation, it is ironic that proponents of the exclusive philosophy (e.g. Siddall & Kluge 1997; Kluge 2001) have argued that probabilistic approaches to phylogenetics in general, and maximum likelihood methods in particular, are incompatible with the philosophy of science described by Popper (1959, 1962, 1983). On the contrary, the probabilistic concept of likelihood provides the foundation for Popper's concept of degree of corroboration (de Queiroz & Poe 2001, 2003; see above, 'Degree of corroboration and statistical inference') and therefore also for any philosophy of phylogenetic inference based on that concept. Both of the central terms in Popper's definition of C — $p(e|hb)$ and $p(e|b)$ — are likelihoods (de Queiroz & Poe 2001, 2003). In the case of a phylogenetic analysis, $p(e|hb)$ is the likelihood of a particular resolved tree, representing h, and $p(e|b)$ is the likelihood of an unresolved (polytomous) tree for the same taxa. Thus, far from being inconsistent with Popper's philosophy, likelihood methods are central to understanding how his method for assessing C applies to phylogenetic hypotheses, and they provide the foundation for a unified and inclusive philosophy of phylogenetic inference.

Acknowledgements

This paper is dedicated to the memory of Joseph B. Slowinski (1962–2001), a friend with whom I shared interests in both reptile systematics and phylogenetic theory, including tests for phylogenetic resolution (Slowinski 2001). I wish to thank Carole Baldwin for asking a question that prompted me to reconsider how to calculate $p(e|b)$, Teri Peterson for information on statistical symbols, Rosario Castañeda for reminding me about a property of PTP tests, and Steve Poe, Dan Faith, and John Trueman for comments on earlier drafts of this paper.

References

Antezana, M. A. & Hudson, R. R. (1999). Type I error and the power of the s-test: Old lessons from a new, analytically justified statistical test for phylogenies. *Systematic Biology*, *48*, 300–316.

Barnett, V. (1999). *Comparative Statistical Inference*, 3rd edn. Chichester: John Wiley & Sons.

Buckley, T. R. (2002). Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. *Systematic Biology*, *51*, 509–523.

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

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press, [Expanded edition published in 1992; Baltimore: Johns Hopkins University Press].

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

Faith, D. P. (1999). [Review of] Error and the Growth of Experimental Knowledge. *Systematic Biology*, *48*, 675–679.

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

Faith, D. P. & Cranston, P. S. (1992). Probability, parsimony, and Popper. *Systematic Biology*, *41*, 252–257.

Faith, D. P. & Trueman, J. W. H. (2001). Towards an inclusive philosophy of phylogenetic inference. *Systematic Biology*, *50*, 331–350.

Farris, J. S. (1995). Conjectures and refutations. *Cladistics*, *11*, 105–118.

Farris, J. S., Kluge, A. G. & Carpenter, J. M. (2001). Popper and likelihood versus 'Popper*'. *Systematic Biology*, *50*, 438–444.

Felsenstein, J. (1978). Cases in which parsimony and compatibility will be positively misleading. *Systematic Zoology*, *27*, 401–410.

Felsenstein, J. (1987). Estimation of hominoid phylogeny from a DNA hybridization data set. *Journal of Molecular Evolution*, *26*, 123–131.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Ecology and Systematics*, *22*, 521–565.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Fisher, R. A. (1970). *Statistical Methods for Research Workers*, 14th edn [1st edn. published in 1925]. New York: Hafner Publishing Co.

Gaut, B. S. & Lewis, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution*, *12*, 152–162.

Gillies, D. (2000). *Philosophical Theories of Probability*. London and New York: Routledge.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. *Systematic Zoology*, *39*, 345–361.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, *36*, 182–198.

Goldman, N. & Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, *17*, 975–978.

Hendy, M. D. & Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, *38*, 297–309.

Hillis, D. M. (1991). Discriminating between phylogenetic signal and random noise in DNA sequences. In M. M. Miyamoto & J. Cracraft (Eds) *Phylogenetic Analysis of DNA Sequences* (pp. 278–294). Oxford: Oxford University Press.

Hillis, D. M. & Huelsenbeck, J. P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity*, *83*, 189–195.

Huelsenbeck, J. P. & Rannala, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science*, *276*, 227–232.

Källersjö, M., Farris, J. S., Kluge, A. G. & Bult, C. (1992). Skewness and permutation. *Cladistics*, *8*, 275–287.

Kendall, M. G. & Stuart, A. (1979). *The Advanced Theory of Statistics*, 4th edn. New York: MacMillan.

Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. *Journal of Molecular Evolution*, *29*, 170–179.

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. *Zoologica Scripta*, *26*, 349–360.

Kluge, A. G. (2001). Philosophical conjectures and their refutation. *Systematic Biology*, *50*, 322–330.

Neyman, J. & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, *20A*, 175–240.

Neyman, J. & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika*, *20A*, 263–294.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H. & Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, *17*, 798–803.

Ota, R., Waddell, P. J. & Kishino, H. (1999). Statistical distribution for testing the resolved tree against [the] star tree. In *Proceedings of the Annual Joint Conference of the Japanese Biometrics and Applied Statistics Societies* (pp. 15–20). Minato-ku, Tokyo: Sinfonica.

Poe, S. & Chubb, A. L. (2004). Birds in a bush: Five genes indicate explosive evolution of avian orders. *Evolution*, *58*, 404–415.

Popper, K. R. (1959). *The Logic of Scientific Discovery*, 1st edn. New York: Basic Books.

Popper, K. R. (1962). *Conjectures and Refutations*. New York: Basic Books, [2nd edn. published in 1968, New York: Harper & Row].

Popper, K. R. (1983). *Realism and the Aim of Science*. London: Routledge.

Prager, E. M. & Wilson, A. C. (1988). Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. *Journal of Molecular Evolution*, *27*, 326–335.

de Queiroz, K. & Poe, S. (2001). Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology*, *50*, 305–321.

de Queiroz, K. & Poe, S. (2003). Failed refutations: Further comments on parsimony and likelihood methods and their relationship to Popper's *Degree of Corroboration*. *Systematic Biology*, *52*, 352–367.

Royall, R. M. (1997). *Statistical Evidence. A Likelihood Paradigm*. London: Chapman & Hall.

Self, S. G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Shimodaira, H. & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, *16*, 1114–1116.

Siddall, M. E. & Kluge, A. G. (1997). Probabilism and phylogenetic inference. *Cladistics*, *13*, 313–336.

Silvey, S. D. (1975). *Statistical Inference*. London: Chapman & Hall.

Slowinski, J. B. (2001). Molecular polytomies. *Molecular Phylogenetics and Evolution*, *19*, 114–120.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). Phylogenetic inference. In D. M. Hillis, C. Moritz & B. K. Mable (Eds) *Molecular Systematics* (pp. 407–514). Sunderland, Massachusetts: Sinauer.

Templeton, A. R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution*, *37*, 221–244.

Tuffley, C. & Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, *59*, 581–607.

Whelan, S. & Goldman, N. (1999). Distribution of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, *16*, 1292–1299.