Cladistics 13, 81–96 (1997)
WWW http://www.apnet.com

Testability and the Refutation and Corroboration of Cladistic Hypotheses

Arnold G. Kluge

Museum of Zoology, University of Michigan, Ann Arbor, MI 48109, U.S.A.

Accepted 6 February 1997

Both refutationist and verificationist philosophies and practices are becoming increasingly evident in phylogenetic inference. Refutation and verification are fundamentally different epistemologies, and it seems unlikely that they can continue to coexist as the basis for inferring species history. The refutationist nature of cladistics is explored in terms of Popperian testability, in order to understand better the alternatives and to predict the outcome of the expected revolution. Testability concerns the logical relationship between a hypothesis (*h*, such as a cladogram), evidence (*e*, such as synapomorphy), and background knowledge (*b*). Of particular importance is the direct relationship between the logical improbability of *h* and its potential to be tested, because for *e* to corroborate *h*, *e* should be improbable given *b* alone. Simplicity and boldness, amount of empirical content, and logical improbability all refer to the potential to be tested. That *h* must be testable by severe tests is the same as saying that those tests have greater probability of failing, given only *b*. Descent with modification is sufficient as background knowledge (*b*) in phylogenetic inference, and such a minimal assumption explains the generality of cladistics. Also of interest to the refutationist position is total evidence. In terms of testability, a statement describing the results of multiple tests is less probable than a statement describing only some of the tests, the multiple test result being more improbable, and accordingly more severe, than its component tests. All other cladistic principles and practices considered in this review are also understandable in terms of Popperian testability, refutation and corroboration. These include minimizing ad hoc hypotheses of homoplasy and minimizing explanatory power, and choosing tentatively among cladograms according to their degree of corroboration (support). Differential character weighting is determined to be unacceptable in terms of testability. Also, testability does not provide a basis for assessing the accuracy of hypotheses, but then that is of no consequence to cladists, because they are not preoccupied with knowing the absolute truth, unlike verificationists.

© 1997 The Willi Hennig Society

> "Everybody had taken it uncritically for granted that a hypothesis high in probability is something good, something we ought to aim at. But the highest probability will be that of a hypothesis which says nothing (like a tautology) or next to nothing (like certain purely existential statements), or which goes as little as possible beyond the facts it is expected to explain (that is to say, a hypothesis which is *ad hoc*). Not only has the alleged aim of obtaining high probabilities never been critically examined, but the intuitive principle that high probabilities are something good can be shown to clash with another intuitive principle: the principle that *ad hoc* hypotheses are something bad. And it is the latter principle that is adopted in actual critical discussions of scientific theories as well as in scientific practice, not the former" (Karl Popper, 1992: 232).

INTRODUCTION

Alternative *refutationist* and *verificationist* (frequency probabilist, neojustificationist) philosophies and practices are evident in the current phylogenetic inference

0748-3007/97/010081+16 /$25.00/0/ cl970033
Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

literature. The cladist's use of synapomorphies to falsify cladograms is consistent with the logic of Popperian *testability* and its practice of refutation and corroboration (Popper, 1968; 1992). In contrast, many molecular biologists use probabilistic models as the basis for estimating the unknowable phylogeny. Here, the emphasis is on verifying phylogenetic hypotheses with taxonomic congruence and consensus or according to statistical estimates or likelihood, which assume one or more parameters relating to evolutionary processes (e.g. Lanyon, 1993; Miyamoto and Fitch, 1995). The refutationist and verificationist alternatives require further study, because they underlie substantively different attitudes and actions that appear to be dividing the field of phylogenetic inference (Felsenstein, 1993).

Few papers in the cladistics literature discuss the basic tenets of testability (see however Cracraft, 1978; Gaffney, 1979; Rieppel, 1979), and much that has been written erroneously equates testability with the falsifiability of universal statements (e.g. Gaffney, 1975; Wiley, 1975; Lovtrup, 1977; Platnick and Gaffney, 1977, 1978a,b; Patterson, 1978; Rieppel, 1979; Panchen, 1982, 1992), or misrepresents testability's foundation as a calculus of probability (e.g. Faith, 1992; see rebuttal by Farris, 1995). The following brief review of testability, or falsifiability as it is often labeled, is meant to clarify these particular issues. My discussion of refutation and corroboration in phylogenetic systematics, in terms of testability, is intended to provide a more complete understanding of the basis for knowledge-claims made by cladists. The problems associated with the truth-claims of verificationists, by contrast, should then be more evident. Certain aspects of verification in phylogenetic inference, such as independence, have been considered elsewhere (Kluge, 1977a).

My understanding of testability and more general notions of refutation and corroboration come largely from Popper's "The Logic of Scientific Discovery" (1968) and "Realism and the Aim of Science" (1992), and especially Chapter 4 of Part I from the latter book, which is entitled "Corroboration". The 1992 contribution is one of three volumes from the "Postscript to the Logic of Scientific Discovery". Popper has gone to extraordinary lengths in the "Postscript" to answer his critics in the clearest possible terms, and those who are studying testability for the first time may find it easier to start with this work before reading the "Logic of Scientific Discovery". Farris' 1983 classic, "The Logic of Phylogenetic Analysis", has profoundly affected the way I see cladistics in terms of testability (see also Lakatos, 1993; Farris, 1995). Yet, I believe the review and synthesis to follow are much needed if we are to understand fully the scientific character of the refutationist and verificationist alternatives at work in phylogenetic inference.

TESTABILITY: SOME BASICS

The centerpiece of testability is the potential to be tested, and it is in this important sense that testability is an instrument of rational criticism. Testability also provides a system for rationally and critically evaluating the nature of assumptions that accompany scientific tests. Testability does *not* provide an argument for proving conclusions. That is, no amount of observational propositions can either prove or disprove a theoretical proposition.

Testability stands in sharp contrast to induction, which is operationally "the collection and (statistical) tabulation of instances, especially confirming instances" (Popper, 1992: 256). Induction does not have the decisive power found in refutation. Naive induction, by simple enumeration of observational propositions, is not persuasive, because repeated observations of the same phenomenon in similar circumstances provide no reason to make an inference beyond what is observed (see however Howson and Urbach, 1993). Even strong induction, where the quality (e.g. independence) and number of observations is high, does not add to the power (Kluge, 1997a). In contrast to refutation, the practice of verification is seen as too easy, and consequently of less scientific merit.

The power of testability is a function of the logical interplay between evidence (e), theory, or hypothesis (h), and background knowledge (b). That these are separate terms is clear from the formulae to follow; however, e must pertain to h, and b represents those initial conditions which *must be* consistent with h. h cannot be supposed without b, and b is accepted (even if only tentatively) while h is being tested. b should not be merely supposed, but if false it is only irrelevant to testing h (Siddall and Kluge, in press). Given just these

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

parameters, *degree of corroboration*[1], *C*, the degree of support given to *h* by *e* in light of *b*, was defined by Popper (1968, 1972a: 288, 1992: 240) as

$$C(h, e, b) = p(e, hb) - p(e, b) \;/\; p(e, hb) - p(eh, b) + p(e, b) \quad (1)$$

which in the numerator reads, the probability of *e*, given *h* and *b*, minus the probability of *e* on *b* alone[2]. Although it may appear counter-intuitive, *h* receives a higher degree of corroboration the smaller $p(e, b)$ is, in particular when $p(e, b) < ½$. Of course, *e* must be possible given *b*, but *e* should be improbable given *b* alone if *h* is to receive corroboration from *e*. Bear in mind that the realm of verification is entered when $p(e, b)$ becomes high. The importance of the demand that *h* cannot be expected, or probable, that *h* and *e* must be independent, can be illustrated with tautology, as a hypothesis (*T*), which by definition has perfect probability. Under those conditions, where we are dealing with an empirically (observationally–experimentally) untestable statement, $p(e, T) = p(e, \text{non-}T) = 1$, and consequently $C(T, e, b) = 0$.

Testability cannot be understood completely without consideration of *severity of test*, *qua* supporting evidence. According to Popper (1972a: 391), *S*, "the severity of the test *e* interpreted as *supporting evidence* of the theory *h*, given the background knowledge *b*", is defined as

$$S(e, h, b) = p(e, hb) - p(e, b) \;/\; p(e, hb) + p(e, b) \quad (2)$$

again where *e* cannot be impossible given *b*. *S* follows directly from the inverse relationship that exists between the logical probability of *h* and its degree of falsifiability (Popper, 1968). Expressed another way, there is a direct relationship between the logical improbability of *h* and its potential to be tested[3]. Explanatory power and degree of corroboration can be said to increase together, because *S* is also the power of hypothesis to explain the evidence (Popper, 1968: 401; Farris, 1983). Obviously, *S* is identical to *C*, except for the absence of $p(eh, b)$ in the denominator of *S* (the probability of both *e* and *h* given *b*), and *C* and *S* are equal when $h = 0$. The numerator shared by *C* and *S* is simply the difference in the probability of the evidence *with and without* the hypothesis, in light of the background knowledge.

The numerator determines the sign of *C* or *S*, because the denominators of those expressions cannot be negative. If *e* neither supports nor undermines *h* then $C(S) = 0$; $C(S)$ is negative when *e* undermines *h*; and if *e* supports *h*, given *b*, then $C(S)$ is positive. $C(S) = -1$ only when *e* absolutely contradicts *h* (on *b*). If, and only if, $p(e, hb) = 1$, $p(e, b) = 0$, and $p(h, b) = 0$ can $C(S)$ reach +1. Such consequences set the limits for degree of corroboration and severity of test, and they give formal meaning to favorable (positive), unfavorable (negative), and irrelevant (zero) as regards the difference in the probability of the evidence with and without the hypothesis, in light of the background knowledge.

According to *C*, the maximum to which *h* can be corroborated is determined by the maximum to which it is testable. In turn, maximum testability is determined, in part, by the *extent* of the *content* of *h*, i.e. by those qualities which determine the amount of empirical information conveyed by *h*. According to Popper (1972b: 81; 1992: 225), the extent of the content of *h* is merely a function of the *simplicity* and *clarity* with which *h* can be described, and the *higher* the content of *h* the *bolder h* is said to be. Simpler hypotheses can also be considered to be objectively more informative, where Sober (1975) defined informativeness as the measurable extent to which the hypothesis alone answers questions about individuals in its domain. Thus, simplicity and boldness, amount of empirical content, and logical improbability (the complement of logical probability) all refer to degree of falsifiability, or testability, the potential to be tested. The amount of empirical information conveyed by a hypothesis increases with its degree of falsifiability (Popper, 1968: 133). That a hypothesis must be testable by severe tests is the same as saying that those tests have a greater probability of failing, given only the background knowledge (*b*). Also, for *h* to be maximally corroborable, *h* must offer a variety of *independent* testable consequences, that is, independent of *b* or any particular *e*. Still further, the maximum that *h* is testable is

[1]Also referred to as the index of corroboration or confirmation.

[2]The denominator of *C(h, e, b)*, and *S(e, h, b)* to follow, is only a normalization factor, intended to remove "blemishes" from the intuitively significant numerator, and is therefore of relatively little consequence to this discussion. For example, according to Popper (1992: 242), the denominator of *C(h, e, b)* "makes, for every *h* (provided it is consistent with *b*) minimal and maximal degrees of corroboration equal to -1 and to the content or degree of testability of *h* (whole maximum is +1)".

[3]Logical improbability = 1 - logical probability

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

inversely related to the number of assumptions made, in terms of the content of *h* relative to *b*.

In practice, hypotheses are tested in the severest manner possible, in relation to their empirical content, and the degree to which a hypothesis has withstood these tests constitutes its degree of corroboration.

> "[F]alsification and corroboration comprise alternative results of testing. A theory is falsified [though not false] if it has been refuted by empirical tests, and it is corroborated [though not proven] if it has so far passed relevant tests" (Farris, 1995: 106).

A hypothesis is accepted, but only tentatively, when it has withstood the most severe tests available, and when it has done better in that regard than any competing hypothesis (*sensu* Lakatos, 1993). The tentativeness of acceptance relates to our inability to know the truth. Corroboration says nothing of proof, only of the relative acceptability of competing hypotheses.

TESTABILITY: SOME CLARIFICATIONS

The epigraph to this paper, and the section immediately above, allude to there being two usages of probability, the familiar calculus probability and the less well-known logical probability of testability, *the probability of a hypothesis according to its tests* (Popper, 1972a: 59; Siddall and Kluge, 1997). The tests passed successfully by a hypothesis are fundamental to establishing *C*, and given this relationship to support, corroboration might be viewed strictly as an exercise in calculus probability. However, consider, for example, that the axiom of monotony (i.e. the probability of a statement describing an event decreases with the increasing logical content of the statement) is fundamental to calculus probability, but does not apply to corroborability (Popper, 1992). There are other reasons for seeing logical and calculus probabilities as different. For instance, as noted above, tautology has a degree of corroboration of zero, whereas that same proposition has a calculus probability of one. The logical relationship between testability and simplicity, and explanatory power (*S*), described by Popper (1968: 112–145; 1992: 225) provides an additional distinction—the simplicity and explanatory power of a hypothesis is related to its logical improbability, not its probability. As should be evident from these distinctions, logical probability and calculus probability are significantly different, and confusing the two must be avoided in future discussions of refutation and verification (Popper, 1968, 1992: 323).

As an aside, the fact that the study of phylogeny is concerned with the discovery of historical singularities means that calculus probability and standard (Neyman–Pearson) statistics *cannot* apply to that historical science (*contra* Felsenstein, 1983; Frost and Kluge, 1994; Depew and Weber, 1995; Huelsenbeck, Bull and Cunningham, 1996; Siddall and Kluge, 1997). Ordinarily, in frequency based probabilistic reasoning, a relevant "population" must be sampled, and it is essential that the sampling be repeated in a random, unbiased, and precise manner, because that is the source of the empirical information necessary to characterize a parameter's probability space. A precisely sampled parameter is the set of definably the same things that can be counted (Bartlett, 1962: 10–36). Obvious parameters in phylogenetic inference are the sister group relationships and the evolutionary transformations which serve as evidence for group relationships. However, the individuality of phylogeny and the uniqueness of each of its constituent patterns and individual processes means that there are no parameter spaces to characterize beyond the singular. Further, all historical sciences, phylogenetics included, are limited in their discovery to ostensively defined entities (Frost and Kluge, 1994). Those entities lack the precision of intensionally defined sets, which is demanded in statistical applications (Bartlett, 1962). Thus, it is simply meaningless to assess, in frequentist probability terms, the parameter phylogenetic history (*contra* Felsenstein and Kishino, 1993). That same conclusion applies to character transformation (Coddington, 1994), and explains why the "homoplasy" approach advocated by Harvey and Pagel (1991) for studying adaptation fails. It fails because it depends on the examination of a predefined class of supposedly repeated historical events (Wenzel and Carpenter, 1994; Siddall and Kluge, 1997).

Much has been attributed to *C* or *S* that is incorrect (see e.g. Bryant, 1989; Faith, 1992). For example, it should be obvious that no part of either formula directly references ad hoc hypotheses (Popper, 1972a: 288). That "*[a]d hoc* hypotheses do have low corroboration" does not make them a part of logical probability (Farris, 1995: 115). Therefore, a reason for minimizing ad hoc hypotheses must come from conjectures other than testability (Farris, 1983; see also below). Farris'

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

(1983: 17) argument was that ad hoc hypotheses are to be avoided, because

> "[s]cience requires that choice among theories be decided by evidence, and the effect of an ad hoc hypothesis is precisely to dispose of an observation that otherwise would provide evidence against a theory. If such disposals were allowed freely, there could be no effective connection between theory and evidence, and the concept of evidence would be meaningless."

Further, there is no term in either *C* or *S* which specifies a particular number for the logical probability of *h*. The

> "logical improbability of *h* is the maximum possible value of the entire expression *C*. That maximum is not determined by the available evidence *e*, but by the most favourable evidence conceivable" (Farris, 1995: 115).

Consequently, it is possible to reject a logically better, more improbable hypothesis, when it has been refuted successfully. The fact that *C* depends on presently available *e*, whereas the logical improbability of *h* does not, further explains how another better-corroborated, though logically less improbable, hypothesis can be preferred.

Other points requiring clarification concern the relationship between content and corroboration, and simplicity and ad hoc hypotheses. As underscored by Farris (1995: 117),

> "Popper's content and corroboration are distinct quantities; that content is the capacity to be tested, while corroboration is the result of testing; that corroboration is thus assessed from presently available evidence, while content is not; or that content sets the upper limit of corroboration."

Simplicity is related to the empirical content of a hypothesis (Popper, 1968: 140–144), but not to cladistic parsimony, i.e. the number of ad hoc hypotheses that may be minimally required when choosing among competing propositions (*contra* Johnson, 1982; see also Bryant, 1989: 218). As noted above, simplicity and explanatory power (*S*) are directly related by virtue of their formal relationship to logical improbability. The number of ad hoc hypotheses is also connected to explanatory power; however, that relationship is complementary (see below).

At least in his most recent writings, Popper (1992; see also Popper, 1968: 100) does not limit testability to the falsification of universal statements, nor does he overly emphasize the importance of testing *predictions*, as they derive from natural laws (universals). To be sure, the level of universality and the degree of precision of a hypothesis can be related to a hypothesis' empirical content, and in turn to its testability; however, Popper made it clear that neither falsification (*sensu stricto* in the naive sense) or prediction is a virtue in itself (see for example, Popper, 1992: 276). Further, as the following quotation established unambiguously, Popper acknowledged the relevance of testability to phylogenetic inference:

> "…some people think that I have denied scientific character to the historical sciences, such as paleontology, or the history of the evolution of life on Earth; or to say, the history of literature, or of technology, or of science itself. This is a mistake, and I here wish to affirm that these and other historical sciences have in my opinion scientific character: their hypotheses can in many cases be tested. It appears as if some people would think that the historical sciences are untestable because they describe unique events. However, the description of unique events can very often be tested by deriving from them testable predictions or *retrodictions*" (Popper, 1980: 611; my italics).

That predictions (*sensu* universals) cannot be derived from historical statements continues to be overlooked by some students of phylogenetic inference (e.g. Penny, Hendy and Steel, 1991).

Unfortunately, a good deal of the early debate concerning the testability of cladistic hypotheses was confounded by interpreting falsification and prediction too narrowly (e.g. Bonde, 1975: 563; Patterson, 1978: 221). For example, there was Cartmill's (1981) extreme deductivist position, that a cladistic hypothesis falsified by an observed incongruent synapomorphy must be judged false, which as Sober (1983: 339) pointed out would require "the preposterous assumption that a true tree must require no homoplasies". Cartmill conflated rejection and disproof.

Bock's (1977: 868) defense of evolutionary classification is an example where the importance of prediction is stressed unreasonably. In fact, historical propositions cannot be judged in terms of bold and highly improbable predictions, because they offer only postdictions (retrodictions). That postdiction does not go beyond the evidence which was used to test (or formulate) the hypothesis in the first place might be seen as counting against the testability of phylogenetic hypotheses. However, as Farris (1979: 512–514) argued, prediction, at least in the sense of *extrapolation*, does have meaning in cladistic propositions. He pointed out that the predictor for

> "the group united from all terminal taxa is just the set of states associated with that group in forming the non-redundant encoding of the data".

From this perspective, the most parsimonious hierarchy, the one that minimizes requirements for ad hoc hypotheses of homoplasy within a clade, is the most predictive, and it is in this sense that the predictiveness of cladistic hypotheses may be maximized. Farris'

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

formulation of prediction is to be preferred over that of Platnick (1978), because it includes all levels of character generality, not just apomorphy.

In testability, Popper's (1968: 31) focus is on the logical analysis of scientific theories, and his lack of concern for the "initial stage, the act of conceiving or inventing a theory", has been criticized (e.g. Panchen, 1992: 306). The assumed deficiency might even be considered particularly serious for cladistics, because observations made from organisms constitute the empirical evidence, while testable cladistic hypotheses involve relationships among taxa, i.e. groups of organisms. Upon reflection, this does not appear to be a real problem for phylogenetic systematics, because of the part-whole relation that exists between organisms and natural groups of organisms, and which can be operationalized with the taxonomic rule of monophyly (Frost and Kluge, 1994). Arguably, it may even be impossible to have absolutely no notion of relationships, because of the background knowledge condition, descent with modification, which specifies minimally "life", the genetic code, or nucleic acids (M. Siddall, pers. comm.). In any case, there are many primitive hypotheses formulated on purely phenetic grounds for cultural reasons (e.g. Bulmer, Menzies and Parker, 1975), and these are testable with cladistic methods. Even in those few instances where one might find it difficult to claim the existence of any kind of hypothesized hierarchical branching order, there remains the completely unresolved proposition, the trichotomous cladogram in the potentially informative simplest case. Such a polytomy would be analogous to the statistical "null hypothesis", and that proposition can be justified with only the background knowledge of descent with modification.

> "The null and alternative hypotheses can be regarded as competing theories, preference between them to be established according to which is better corroborated" (Farris, 1995: 113).

Another criticism that might be leveled at testability is that it is too narrowly focused on the empirical, not enough attention being paid to the conceptual aspects of science. I believe such a criticism lacks force, at least in the context of phylogenetic inference, because the science of cladistics is concerned with the discovery of historical individuals, where the essentialist interpretation of concept has no meaning (Frost and Kluge, 1994).

Lastly, let it be understood that the critical nature of evidence in testability (empirical content) pertains to conceivable outcomes of experiments (conceivable observations—propositions in light of b). Whether strongly decisive evidence (refutation/corroboration) is actually observed is another matter.

THE REFUTATION AND CORROBORATION OF CLADISTIC HYPOTHESES

A Popperian evaluation of scientific theories (Popper, 1968, 1972a,b, 1992) requires a test statement, which involves bringing evidence to bear on a hypothesis, in light of the background knowledge. Synapomorphies, not autapomorphies or symplesiomorphies, constitute evidence in phylogenetic systematics, because only those empirical generalities have the *potential to refute* (falsify) a particular cladistic hypothesis (Hennig, 1966; see also below). Earlier, with the example of tautology, I presented Popper's logical argument for why e must be completely independent of h, because $p(e, b)=p(e,\text{non-}b)=1$, if it is not, and consequently, $C(h, e,b)=0$. The same outcome appears to apply to autapomorphies, and that consequence may illustrate formally why those observations cannot test cladograms. Thus, autapomorphies are observational propositions, but they do not bear on the theoretical propositions (or hypothesis) of taxonomic relationships.

Refutation (falsification) resides in incongruent synapomorphies, because those shared-derived traits imply evidence for a different cladogram. Homoplasy (synapomorphic similarity which is not due to inheritance) and homology (synapomorphic similarity which is due to inheritance) are the familiar process explanations given to incongruent and congruent synapomorphies, respectively, in light of a phylogenetic hypothesis (Frost and Kluge, 1994). Observationally speaking, homoplasy and homology cannot be facts, because those interpretations of process are contingent on a historical hypothesis.

The test of a cladistic hypothesis cannot be a matter of deduction (*contra* Cartmill, 1981), in the sense that an empirical statement falsifies a hypothesis if it can be deduced that the hypothesis is false from the truth of the empirical statement. Deduction is impossible, because a cladogram is logically consistent with all synapomorphy distributions, congruent and

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

incongruent (Farris, 1983; Sober, 1983). Said another way, homoplasy has no decisive power in a deductive sense because it can be used to explain patterns and non-patterns alike. Although there may be good reason to minimize such a universal explanation as homoplasy, for otherwise knowledge claims could not be made at all, that reason does not make a case for deduction in evaluating phylogenetic hypotheses.

There remains, however, a sound logical basis for homoplasy as a test of cladistic hypotheses (*contra* Bryant, 1989: 219–220). Assume a rooted three-taxon cladogram, (A,B)C, and synapomorphies distributed as 110, 101 and 011. The parenthetic taxonomic notation describes the relative recency of common ancestry, taxa A and B share a more recent common ancestor than either does with C, and the character states in each synapomorphy apply to the taxa in the order in which they appear. Throughout this paper, states 0 and 1 are plesiomorphic and apomorphic, respectively. Thus, the congruent synapomorphy has the taxonomic distribution A_1, B_1, C_0, while the incongruent synapomorphies have the distributions A_1, B_0, C_1 and A_0, B_1, C_1. Now, consider the alternative conclusions which follow from simply conjoining the cladogram (A,B)C with the congruent and incongruent synapomorphies:

1. a cladogram alone does *not* imply the derived states of a congruent synapomorphy are homologous (are of a common origin);
2. a cladogram *by itself* does imply the derived states of an incongruent synapomorphy are homoplasious (are of independent origin) (Farris, 1983: 13; Kluge, 1995).

It is only in this sense that synapomorphy constitutes a test, and that homoplasy can be said to *count* evidentially against a particular cladistic hypothesis (Sober, 1988a; see discussion of character compatibility analysis below). This asymmetrical relationship between homology and homoplasy and phylogeny is considered to be *fundamental*, because without it there would appear to be no decisive evidence with which to test cladistic hypotheses. That such a logical basis for refutation must be understood, strictly speaking, as "*non*-deductive" (Sober, 1993) has no real bearing on the refutation of cladograms, and the relevance of testability to phylogenetic systematics. A hypothetico-deductive system may apply perfectly to universals (or perhaps not; see Stamos, 1996); however, that does not deny testability to a science devoted to the discovery of historical individuals (Popper, 1980).

The falsehood of a hypothesis can never be proven, even where deductive logic applies, because the falsifying observational proposition may itself be false (Cracraft, 1978: 215). Thus, as a rule in a Popperian evaluation of scientific theories, a preference is shown for the hypothesis that requires the ad hoc dismissal of the fewest falsifiers (Gaffney, 1979: 98). In cladistics, the least refuted hypothesis is the most parsimonious cladogram which minimizes requirements for ad hoc hypotheses of homoplasy, thereby minimizing empty statements and consequently maximizing content. It is the most parsimonious cladogram that achieves the highest degree of corroboration (*C*), because of the inverse relationship between ad hoc hypotheses and explanatory power (*S*). As Farris (1983: 18) emphasized, the potential to maximize explanatory power, that is, being able to provide an explanation for the similarity of congruent shared derived traits as due to inheritance, is a consequence of minimizing requirements for ad hoc hypotheses of homoplasy. Precisely, the more homoplasies required of a phylogenetic hypothesis, the more evidence it fails to explain. Again, in strictly Popperian terms, most parsimonious cladograms are most explanatory, because both *C* and *S* increase with p(*e*, *hb*), a term which occurs in their shared numerator (Farris, 1995: 116). Thus, we have the logic of parsimony analysis in cladistics and its relationship to a Popperian evaluation of scientific theories (Farris, 1983: 17–19).

The practice of minimizing ad hoc hypotheses is accepted widely in the empirical sciences (see however Howson and Urbach, 1993), for otherwise every proposition could be protected from criticism, and as set forth above for cladistics; it is the *requirement* for an explanation of homoplasy that is considered ad hoc, and which is minimized with cladistic parsimony. Under these circumstances, the *independence* of required ad hoc hypotheses of homoplasy is of special concern, because the total number of instances of homoplasy serves as the basis for choosing among cladograms (h_n). As Farris (1983: 19–20) argued,

> "[i]f two characters were logically or functionally related so that homoplasy in one would imply homoplasy in the other, then homoplasy in both would be implied by a single ad hoc hypothesis. The "other' homoplasy does not require a further

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

> hypothesis, as it is subsumed by the relationship between the characters. This is the principle underlying such common observations as that only independent lines of evidence should be used in evaluating genealogies…"

As I suggested above, this particular concern for independence cannot be traced to logical probability, because no part of either *C* or *S* actually refers to ad hoc hypotheses (Popper, 1972a: 288).

It continues to be asserted that the use of cladistic parsimony in phylogenetic inference assumes that homoplasies are rare in nature (e.g. Pritchard, 1994; see also Crisci and Stuessy, 1980; Cartmill, 1981). However, as Farris (1983: 13; see also Sober, 1988a: 136) countered long ago, the method of minimizing requirements for ad hoc hypotheses of homoplasy does not necessarily presume minimality, i.e. rarity of homoplasy. A most parsimonious cladogram places only a lower bound (but not an upper bound) on the number of homoplastic events required of the evidence. Consider, for example, that it is possible to imagine an even less parsimonious history for synapomorphy 011 in relation to the rooted cladogram (A,B)C than the independent origin of state 1 in taxon B and in taxon C. Very many origins of state 1, followed by reversals to state 0, might have occurred in either lineage (or both), and cladistic parsimony does not place a limit on those possibilities.

Also, the commonness of homoplasy in gene sequence data is becoming an increasingly popular basis on which to criticize cladistic parsimony, and to recommend other methods of inference, statistical ones, such as maximum likelihood (Felsenstein, 1993; Swofford et al., 1996). Such criticisms usually follow from specious arguments having to do with the potential statistical inconsistency of unweighted parsimony methods, where entering the dreaded statistically inconsistent Felsenstein zone

> "the only hope [one has] of getting the correct tree is by sampling *few enough characters* that we may be lucky enough to obtain more of the character patterns favoring the true tree than of the more probable character patterns favoring the wrong tree!" (my italics; Swofford et al., 1996: 427)

Of course, this criticism follows only from a verificationist philosophy. Refutationist molecular systematists are more inclined to indite the gene sequences, because of arbitrariness and ambiguity in alignment (Gatesby, De Salle and Wheeler, 1993) and the noise exposed in total evidence parsimony analyses (Wheeler, 1995). These are not the qualities of convincing disconfirming evidence.

In the case of phylogenetic hypotheses, the assumed background knowledge[4] so critical to testability is descent with modification (Darwin, 1859: 420). Although such a simple proposition describing evolution is sufficient, background knowledge might be anything else one claims to know (but not merely suppose!) concerning species history, but not the phylogeny in question. As noted earlier, in discussing severity of test, there is the demand that *h* should be improbable on *b*. As an example of how background knowledge relates to severity of test, that is the improbability of *h*, consider the following simple cladistic example, where the hypothesis of relationships of three terminal taxa is determined with synapomorphic evidence. Given only descent with modification as the background knowledge, synapomorphies characteristic of (A,B), (A,C) and (B,C) should be equally likely (*contra* Penny, Hendy and Steel, 1991: 156–157). 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 hypothesis is said to be corroborated to the degree to which those (A,B) synapomorphies are observed. The severity of a test can be increased, made more critical, by increasing the number of independent characters, because the probability ratio [p(*e*,*hb*) : p(*e*, *b*)] increases with the number of independent tests. Obviously, severity of test decreases with the more background knowledge that is included which favors one class of synapomorphies, as is the case with a priori weighting where a preference is shown for some form of congruence (see below).

By including only descent with modification as background knowledge, by avoiding other conditions relating to pattern and process, relatively more of the general features of phylogeny and evolution can be critically evaluated. For example, including only descent with modification, the hierarchical nature of phylogeny itself can be judged from how the observed synapomorphies are distributed. Similarly, differential rates and patterns of character evolution can be inferred from those phylogenetic hypotheses which

[4]These are often auxiliary hypotheses, the already well-tested assumptions which carry a high degree of corroboration.

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

have the highest degree of corroboration, when minimal background knowledge is included. In other words, such basic knowledge claims as these can flow from the most severely tested hypotheses, but only if the background knowledge is kept to a minimum, which is, I believe, descent with modification. Generally, verificationists appear to hold the opposite perspective on pattern and process (e.g. Huelsenbeck, Bull and Cunningham, 1996).

Nelson (1989), among others (e.g. Brady, 1983), has claimed that the empirical science of cladistics is a discovery procedure completely, or nearly completely, unbounded by theory and assumptions. The question is how much theorizing and assuming does it take to change cladistics into a non-empirical science (Rieppel, 1991). For example, Nelson and Platnick (1991; see also Nelson and Platnick, 1981) do not appear to appeal to any relevant background knowledge in their three-taxon methodology, which they offered as a possible improvement to cladistic parsimony. The apparent absence of any condition relating to background knowledge, and the fact that their procedure decreases explanatory power (Kluge, 1994), must be taken to mean that the usefulness of the three-taxon methodology cannot be argued in terms of a Popperian evaluation of scientific theories. Surely, there can be nothing less in phylogenetic inference than an appeal to descent with modification as background knowledge.

There are important presuppositions concerning evidence, which are not to be confused with background knowledge. These other initial conditions bear on the genuineness of the test, that the evidence actually has the potential to refute a hypothesis and that the result is convincing (Cracraft, 1978: 215). Given the rooted cladistic hypothesis (A,B)C, refutation includes synapomorphies with state distributions 101 and 011, evidence which supports alternative rooted hypotheses (A,C)B and A(B,C), respectively. Two conditions must be obtained for this synapomorphy test to be considered genuine:

1. the 1 states must be sufficiently similar to be called the same (Owen, 1866) at some level of taxonomic generality (Riedel, 1978: 52); but

2. the generality of that distribution must be limited to two of the three taxa, that it is not 111, a symplesiomorphy.

Four perspectives are usually recognized as bases for evaluating the similarity of apomorphic states, i.e. composition, conjunction, ontogeny, and topography (see e.g. Patterson, 1982). The importance of evaluating similarity is almost always emphasized in light of its relationship to homology (e.g. Patterson, 1982; Rieppel, 1992), and some students of historical inference have even claimed that a proportionate relationship between relative recency of common ancestry and organismal similarity is assumed *necessarily* in cladistics (e.g. Laws and Fastovsky, 1987: 2–3). The latter seems unlikely, however, given that cladists criticized pheneticists for appealing to that same proportionality. In any case, if evidence is sought for its ability to refute a hypothesis, then the similarity of the homoplasious states should also be considered important in cladistic analyses. As a rule, it seems, there is a direct relationship between the similarity of homoplasious states and how convincing they are as refutations of a cladistic hypothesis—compare the similarities of the several bird–mammal synapomorphies summarized by Gauthier, Kluge and Rowe (1988). The outgroup method (*sensu* Clark and Curran, 1986; see also Farris, 1982) is recommended as the basis for judging the generality of character states (polarity), because that method achieves a more globally most parsimonious hypothesis and tests the individuality of the ingroup. Neff's (1986: 116; see also Bryant, 1989) claim that a rigorous form of hypothetico-deductive testing may apply at the level of character analysis, prior to testing a cladogram, is outside the focus of this paper, which is the testability of cladistic hypotheses. However, that topic is addressed elsewhere (Kluge, 1997b).

According to testability (not statistical consistency, *contra* Huelsenbeck, Bull and Cunningham, 1996: 152), cladists should use all of the relevant available synapomorphies, the total evidence, when testing a phylogenetic hypothesis (Kluge, 1989, 1997; Eernisse and Kluge, 1993; Jones, Kluge and Wolf, 1993; Kluge and Wolf, 1993). That is so, because a statement describing the results of *multiple* tests (if the tests are independent) "will be less probable than a statement describing only some of the tests" (Popper, 1992: 247–248)—a multiple test result being more improbable, and accordingly *more severe*, than its component tests. Further, the taxa whose historical relationships are to be tested, including the outgroups employed, determine the organisms on which the traits are actually

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

observed, and over which the generality of the apomorphies are determined. Thus, surveying taxa as broadly as possible, including fossils, which is a goal of total evidence, also increases severity of test (Kluge and Wolf, 1993) because it offers greater opportunity to discover incongruent (Sanderson and Donoghue, 1989: figure 4) and disconfirming evidence from a larger sample of organisms, which are potentially more diverse.

Currently, there is considerable debate concerning taxonomic congruence and total evidence methodologies in phylogenetic inference, and at the heart of the controversy is the nature of consensus hypotheses. Taxonomic congruence is being promoted because that approach *necessarily* leads to a consensus hypothesis of fundamental cladograms as the result of partitioning and analysing separately the relevant available evidence (for review see Kluge and Wolfe, 1993; Kluge, 1997a). However, consensus hypotheses so formed, as perfectly consistent phylogenetic propositions, have zero degree of corroboration, $C\ (h, e, b)=0$; they are devoid of empirical content and explanatory power. Such hypotheses are no better in these regards than the "trivial" two-taxon statement, or any unresolved cladogram. For a given data matrix, a total evidence phylogenetic hypothesis by its nature entails more empirical content and explanatory power than does a consensus of two or more different cladograms derived from partitions of those data.

Pairwise character comparisons (character compatibility analysis) certainly exemplify the fundamental notion of test in phylogenetic inference (Wilson, 1965; Patterson, 1982: 74; see also above). Logically, one or the other, or both, of two incompatible characters must be homoplasious. However, each such pairwise comparison constitutes the weakest possible test (as would be the case with *complete* partitioning in taxonomic congruence), because of the piecemeal testing of the characters. Compatibility analysis fell from grace in phylogenetic inference because of the failure of the largest cliques of compatible characters to maximize explanatory power (S) (Kluge, 1976), not because compatibility analysis is inconsistent with refutation.

As a rule, cladists search for disconfirming synapomorphies when testing a particular hypothesis of sister group relationships. For example, those shared-derived traits that relate taxon A with C, and B with C, would be sought when testing the particular rooted hypothesis (A,B)C. Of course, additional synapomorphies confirming the (A,B) clade may be discovered coincidentally when searching for disconfirming evidence, and the question arises as to what to do with them. It may be recalled that C is the maximum h can be corroborated and that is determined by the maximum h is testable. And, in turn, maximum testability is determined by the extent of the content of h, and that is the amount of empirical information conveyed by h. Thus, those additional confirming synapomorphies should be recorded as having been observed, because they increase the empirical content of the phylogenetic hypothesis that is expected to be retested at some future date (Kluge, 1991; Kluge, 1997b).

Long-held phylogenetic hypotheses might be interpreted as especially worthy of testing, because those propositions may be assumed to have considerable explanatory power; at least more power than alternative theories. Such worthiness is probably justified, because maximum testability is determined by the amount of empirical information conveyed by h. The further testing of a very highly corroborated cladistic hypothesis of relationships might even be viewed as certainly justified, because of the improbability that accompanies h, that which has led to the large number of observed congruent synapomorphies. The archosaur relationships of birds exhibit both of these qualifications. The hypothesis has been around a long time, is assumed to be highly explanatory, and in recent years it has achieved a high degree of corroboration as a result of total evidence testing (see review by Eernisse and Kluge, 1993). However, not all tests of cladograms are what they seem to be, and some of the recent studies of archosaur–bird relationships provide examples of specious tests. For instance, Hedges, Moberg and Maxson (1990; see also Huelsenbeck, Bull and Cunningham, 1996: 155–157) emphasized the importance of taxonomic congruence, the separate analysis of different genes and their consensus, in their evaluation of the sister-group relationships of birds. In effect, this study really sought verification of Haematothermia, a group consisting of birds and mammals, excluding testudines and saurians. In another recent study, Gardiner's (1993) review of amniote relationships had more to do with justifying his previous attempt to deny importance to fossils in phylogenetic inference (Gardiner, 1982; see also Patterson, Williams

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

and Humphries, 1993; Patterson, 1994). In that regard, Gardiner also sought verification of Haematothermia. The bottom line is that not all evaluations of phylogenetic hypotheses involve valid tests, and considerable care must be exercised in judging degree of corroboration and explanatory power.

Not all agree that a cladistic hypothesis is tested with synapomorphies and in turn provides the context in which those particular synapomorphies are explained historically as homoplasious or homologous. Consider, for example, Brooks and McLennan's (1991: 63)

> "cardinal rule: *never use the characters that are part of the evolutionary hypothesis under investigation to build your phylogenetic tree.* Rather, these characters should be mapped onto an existing tree."

Apparently, circularity of reasoning is the issue underlying this alternative perspective (Kluge and Wolf, 1993). As Coddington (1988: 7) put it, "to avoid circularity, [the cladogram] should not be inferred from characters involved in the hypothesis of adaptation". I see no basis for this concern, from the perspective of testability. Of course, the cladogram should not be based on *only* the characters involved in the hypothesis of adaptation, but rather should be refuted or corroborated by additional characters. *Independent synapomorphies* are the only historical test of adaptation, and requirements for ad hoc hypotheses of homoplasy are minimized when exercising those tests. That the explanatory power of such evidence is maximized consequently follows directly from the logic of cladistic parsimony analysis (Farris, 1983: 17–19). It has nothing to do with an individual investigator "reasoning" in a circular manner.

Differential character weighting in phylogenetic inference remains a hotly debated topic (Kluge, 1997b). Much of the disagreement concerns the justification for weighting and when weighting methods might be reasonably applied in a given round of phylogenetic research. There are the familiar a priori and a posteriori types of weighting, weights applied before or after cladistic parsimony has been exercized, respectively (Farris, 1969, 1988). In addition, Goloboff (1993, 1995) has proposed an optimality criterion for choosing among competing cladograms, which is based on differential character weighting. Unlike cladistic parsimony, where equally weighted steps are minimized, Goloboff's criterion calls for maximizing self-consistent character weights. The concept of hierarchical character correlation forms the basis for the justifications for Goloboff and a posteriori weighting. The concept is also obvious in most justifications for a priori weighting (for other arguments see Mindell and Thacker, 1996).

> "The *defining property* of hierarchic correlation is that a set of variables with high hierarchic correlation will all be highly consistent with a single branching pattern. Characters that are hierarchically correlated may *or may not* be correlated in any other apparent way" (Farris, 1969: 376).

Almost all arguments for character weighting based on the concept of character correlation (self-consistency) are, in one way or another, exercises in verificationism. For example, Goloboff (1993: 84) argued that

> "characters which have failed repeatedly to adjust to the expectation of hierarchic correlation are more likely to fail again in the future, and so they are less likely to predict accurately the distribution of as yet unobserved characters."

In general, so go the arguments, homoplasious characters, those which are noisy for whatever reason (investigator coding error, biological process, or accident), are less reliable and must be downweighted, because they confound the discovery of the true phylogeny (e.g. Mindell and Thacker, 1996; Wakeley, 1996).

According to Popperian testability, as emphasized above, a cladistic hypothesis receives corroboration from synapomorphies only to the degree that the evidence is improbable given the background knowledge alone. Testability only requires that each character in the data matrix provides an independent, potentially disconfirming, test. Independent synapomorphies may be considered of equal weight in this sense. However, a priori and Goloboff differential character weighting cannot be recommended, because in their application they add to background knowledge, which decreases the improbability of a hypothesis in light of its tests. Adding to background knowledge is a verificationist slippery slope, which ultimately ends in tautology.

The character "reliability" justification for weighting can be criticized for other reasons. For example, the reliability argument, counting instances of homoplasy across a cladogram, assumes the historical dependence among lineages, a position which is contradicted by the historical independence of exclusive clades (Wenzel and Carpenter, 1994). Also, reliability weighting suggests that, in some intrinsic biological sense, characters are differentially committed to homoplasy, past, present and future (see Goloboff quote above). This I believe involves an extra assumption about a biological

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

process, which, like a priori weighting, adds to background knowledge and decreases degree of corroboration. Moreover, that general biological process is as yet unspecified, and as such is untestable. Thus, I see no alternative but to reject all forms of character weighting in a refutationist program of cladistical research, where the maximally corroborated (*C*) cladogram is sought. However, weighting might still play a practical role in character reanalysis (Kluge, 1997b). As Farris (1969: 374) pointed out long ago, weighting may provide an algorithmically efficient way to explore the sensitivity of the best-fitting cladogram(s) in light of the evidence.

Missing data and polymorphisms can have a negative impact on severity of test, when their uncertainty contributes to imprecision and ambiguity (Platnick, Griswold and Coddington, 1991). A fully bifurcating pattern of species relationships and unambiguous optimization of all character states at the internodes of a cladogram add to the severity of test. In terms of testability, the more parsimonious the hypothesized species relationships and character state optimizations, the better (Farris, 1970: 92). "[F]inding the correct tree" when there are missing data (*sensu* Huelsenbeck, 1991; Wiens and Reeder, 1995) or polymorphisms is not an issue relevant to testability.

I have suggested elsewhere that phylogenetic hypotheses can, and should, be tested with empirical evidence other than synapomorphies (e.g. Kluge, 1983). Two obvious sources of these observations are the most parsimonious biotic area cladogram resulting from vicariance biogeographical studies (Kluge, 1988) and the maximally congruent hypothesis sought in the study of host/parasite coevolution. The basic idea is that these other sources provide generalities in their own right, temporal and ancestor–descendent, and which are sought for their power to refute the cladogram being tested. The importance of these different classes of evidence was first espoused by Whewell (1847, Vol. 2: 469) but, as an inductionist-verificationist, it is not surprising that he emphasized the importance of their *consilience*, the coincidence of inductions being "a test of the *truth* of the theory in which it occurs" (my italics). The widely accepted theories that the organic and inorganic worlds (Nelson and Platnick, 1981) and hosts and their parasites (Ehrlich and Raven, 1964) evolve together justify the use of observations from these other histories as tests of cladistic hypotheses. That these tests are obviously logically independent suggests that they may be judged as more critical than are sets of synapomorphies at refuting cladograms. Also, biogeographical and coevolutionary tests may be judged more critical than sets of synapomorphy, because it is generally assumed that apomorphies are vertically transmitted[5], whereas both vertical and horizontal (dispersal) transmission are assumed when interpreting biogeographical and host-parasite patterns (Sober, 1988b). Thus, synapomorphies are expected to be consistent historically, whereas there is a greater improbability of observing a congruent set of biogeographical or host-parasite patterns. Although earth-history and coevolution types of evidence are usually sought when equally strongly competing, equally most parsimonious cladograms have been discovered, those sources of evidence should be used more generally given the special powers they add to severity of test.

The age of clades also provides an additional source of evidence with which to test phylogenetic hypotheses (Gauthier, Kluge and Rowe, 1988: 188–190; see also Norell and Novacek, 1992), the question of how to measure parsimony debt aside (e.g. Fisher, 1994). The truism that ancestors must precede their descendants in time justifies the use of this source of observations, the minimum age of clades, as a test of phylogenetic hypotheses. However, it is the level of certainty of the justification that makes the test rather ordinary, unlike geographical or host–parasite patterns.

Scientists do not actually seek the truth, because truth is unknowable. Scientists do, however, attempt to approach some unattainable objective truth, and do so by critically evaluating different explanations. Hypotheses can never be proven true, as inductivists seek to do, nor be proven false, as deductivists claim to be able to do; however, they can be found to be more or less corroborated. Those that persist continue to be more corroborated; the others do not. This is science according to Popper (1972b; 1992).

Such a connection between truth and explanation is perfectly clear when it comes to testability and phylogenetic systematics. The most severely testable cladistic hypotheses are those which have minimal

[5]This assumption fails, although apparently only rarely, when horizontal (between species) transmission occurs (e.g. gene capture and movement by viruses).

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

requirements for ad hoc hypotheses of homoplasy, and as Farris (1983) argued convincingly it is that minimization which effectively maximizes explanatory power, that is where the largest number of synapomorphies are interpretable as homologues. Of course, an explanation of homology, as a mark of history, is never proven; it is only a tentative hypothesis awaiting further critical tests (Kluge, 1997b).

METRICS

Cladograms are often accompanied by metrics which, in one way or another, are meant to assess the fit of the hypothesis of sister group relationships to the data analysed. The ensemble consistency index (CI) is one of the oldest such metrics (Kluge and Farris, 1969). The importance of the ensemble CI lies in the fact that it counts requirements for ad hoc hypotheses of homoplasy, and it is of particular value when comparing different cladograms, in terms of a *given* data set (Farris, 1989). The higher the ensemble CI the fewer ad hoc hypotheses required to explain the data. Also, the cladogram of minimum length continues to be the preferred context in which to examine the results of maximizing explanatory power, the hypothesized histories of individual characters (e.g. Wenzel, 1992).

Calculating group and total support metrics is becoming increasingly common in cladistics (e.g. Eernisse and Kluge, 1993), but the relationship of those metrics to degree of corroboration remains to be explored fully. According to Källersjö et al. (1992: 284),

> "a group on a considered most parsimonious tree is supported by strong evidence when a large increase in length of included trees is required before that group is lost in the consensus",

or according to Bremer (1994: 295), it is "the extra length needed to lose a branch in the consensus of near-most-parsimonious trees". The total support index was defined by Källersjö et al. (1992: 284) as the sum of group supports, and Bremer (1994: 295) provided a rescaled measure of total support, the "sum of all branch support values over the tree divided by the length of the most parsimonious tree[s]". These support metrics are meant to assess degree of corroboration, relative to a consensus, a hypothesis lacking empirical content. Unlike Bremer (1994), I do not see these metrics' relevance to cladistics as measures of tree *stability*. Indeed, phylogenetic hypotheses may be stable, but stability *per se* is not one of the goals of cladistics (Kluge, 1989: 7–8; *contra* Siddall, 1995). Even the stablest hypothesis' degree of corroboration goes down when new disconfirming characters are discovered (Sober, 1988). As a consequence, a once-stable hypothesis can lose its acceptability.

Popper (1972b: 58f; 1992: 220) never visualized a calculus of probability for degree of corroboration, and such an expectation seems especially misplaced when summarizing the results of phylogenetic tests. As enumerated in the previous section, severity of test is conditional on many qualities, such as independence, ambiguity, and weights, which are not quantifiable with any reasonable degree of precision. Also, the simple fact that phylogenetic inference is concerned with historical individuals, ostensively, but not intensionally, defined entities (Kluge, 1990; Frost and Kluge, 1994), must be taken to mean that an exact value of degree of corroboration *cannot* be determined. As J. S. Farris has commented (pers. comm.; see also Carpenter, 1992),

> "Exact values could be obtained from a detailed model of evolution, but phylogeneticists have mostly avoided that approach. Imprecision has generally seemed preferable to the specious precision obtained from ludicrous premises."

Further, it must be borne in mind that it is unlikely that any measure of corroboration (support) can be formulated which actually takes account of how honestly and diligently a scientist has sought critical evidence. Degree of corroboration will always be more than a score, more than a simple summing of congruent synapomorphies, more than measures of group and total support. Phylogenetic hypotheses are not provable or disprovable in any absolute sense. The more genuine tests that refute a cladistic proposition the less prominence it can be expected to have as an explanatory hypothesis. The most parsimonious cladogram, the one least refuted, is only the focus of the next round of testing, and so it goes (Kluge, 1997b).

ACKNOWLEDGEMENTS

My views on testability and refutation and corroboration, as they relate to phylogenetic systematics, were presented initially at the University of Georgia, Athens, June, 1994, in a Society of Systematic Biology Symposium, "Separate Versus Simultaneous Phylogenetic Analysis". I appreciate Kirk Fitzhugh's encouragement, which convinced me to publish an updated version of those perspectives. The insights provided by

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

James S. Farris and James Carpenter on the nature of background knowledge and severity of test, respectively, gave me additional impetus to go ahead with this publication. Kirk Fitzhugh, Sharon Jansa, Barb Lundrigan, Dave Mindell, Olivier Rieppel, Mark Siddall, and John Wenzel read an early, somewhat shorter, draft of the manuscript, and their critical comments helped greatly to sharpen many of the connections I have identified between testability and cladistics. Kirk Fitzhugh and Mark Siddall were especially critical and inspiring. Nonetheless, what is published finally remains my responsibility entirely. Much of this work was completed at the Cladistics Institute, Harbor Springs, Michigan.

REFERENCES

Bartlett, M. S. (1962). "Essays on Probability and Statistics". John Wiley, New York.

Bock, W. J. (1977). Foundations and methods of evolutionary classification. *In* "Major Patterns of Vertebrate Evolution" (M. K. Hecht, P. C. Goody, and B. M. Hecht, Eds), pp. 851–895. Plenum Press, New York.

Bonde, N. (1975). Review of "Interrelationships of fishes". *Syst. Zool.* **23**, 562–569.

Brady, R. H. (1983). Parsimony, hierarchy, and biological implications. *In* "Advances in Cladistics" (N. I. Platnick, and V. A. Funk, Eds), Vol. 2., pp. 49–60. Columbia University Press, New York.

Bremer, K. (1994). Branch support and tree stability. *Cladistics* **10**, 295–304.

Brooks, D. R., and McLennan, D. A. (1991). "Phylogeny, Ecology and Behaviour: A Research Program in Comparative Biology". University of Chicago Press, Chicago.

Bryant, H. N. (1989). An evaluation of cladistic and character analyses as hypothetico-deductive procedures, and the consequences for character weighting. *Syst. Zool.* **38**, 214–227.

Bulmer, R. N., Menzies, J. I., and Parker, F. (1975). Kalam classification of reptiles and fishes. *J. Polynesian Soc.* **84**, 267–308.

Carpenter, J. M. (1992). Random cladistics. *Cladistics* **8**, 147–153.

Cartmill, M. (1981). Hypothesis testing and phylogenetic reconstruction. *Z. Zool. Syst. Evol.* **19**, 73–96.

Clark, C., and Curran, D. J. (1986). Outgroup analysis, homoplasy, and global parsimony: A response to Maddison, Donoghue, and Maddison. *Syst. Zool.* **35**, 422–426.

Coddington, J. A. (1988). Cladistic tests of adaptational hypotheses. *Cladistics* **4**, 3–22.

Coddington, J. A. (1994). The roles of homology and convergence in studies of adaptation. *In* "Phylogenetics and Ecology" (P. Eggleton, and R. I. Vane-Wright, Eds), pp. 53–78. Linnean Society Symposium Series, **17**. Academic Press, London.

Cracraft, J. (1978). Science, philosophy, and systematics. *Syst. Zool.* **27**, 213–216.

Crisci, J. V., and Stuessy, T. F. (1980). Determining primitive character states for phylogenetic reconstruction. *Syst. Bot.* **5**, 112–135.

Darwin, C. (1859). "On the Origin of Species". 1964 facsimile reprint of first edition. Harvard University Press, Cambridge, Massachusetts.

Depew, D. J., and Weber, B. H. (1995). "Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection". Bradford Book, MIT Press, Cambridge, Massachusetts.

Eernisse, D., and Kluge, A. G. (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.* **10**, 1170–1195.

Ehrlich, P. R., and Raven, P. H. (1964). Butterflies and plants: A study in coevolution. *Evolution* **18**, 586–608.

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

Farris, J. S. (1969). A successive approximations approach to character weighting. *Syst. Zool.* **18**, 374–385.

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

Farris, J. S. (1979). The information content of the phylogenetic system. *Syst. Zool.* **28**, 483–519.

Farris, J. S. (1982). Outgroups and parsimony. *Syst. Zool.* **31**, 328–334.

Farris, J. S. (1983). The logical basis of phylogenetic analysis. *In* "Advances in Cladistics" (N. I. Platnick, and V. A. Funk, Eds), Vol. 2, pp. 7–36. Columbia University Press, New York.

Farris, J. S. (1988). Hennig. Hennig86 Reference. Version 1.5. Privately printed.

Farris, J. S. (1989). The retention index and the rescaled consistency index. *Cladistics* **5**, 417–419.

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

Felsenstein, J. (1983). Statistical inference of phylogenies. *J. Roy. Stat. Soc.* **146**, 246–272.

Felsenstein, J. (1993). Unspoken disagreements. Review of "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto, and J. Cracraft, Eds) (1991). Oxford University Press, New York. *Cladistics* **9**, 119–126.

Felsenstein, J., and Kishino, H. (1993). Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.* **42**, 193–200.

Fisher, D. C. (1994). Stratocladistics: Morphological and temporal patterns and their relation to phylogenetic process. *In* "Interpreting the Hierarchy of Nature" (L. Grand, and O. Rieppel, Eds), pp. 133–171. Academic Press, New York.

Frost, D. R., and Kluge, A. G. (1994). A consideration of epistemology in systematic biology, with special reference to species. *Cladistics* **10**, 259–294.

Gaffney, E. S. (1975). A phylogeny and classification of higher categories of turtles. *Bull. Am. Mus. Nat. Hist.* **155**, 387–436.

Gaffney, E. S. (1979). An introduction to the logic of phylogeny reconstructions. *In* "Phylogenetic Analysis and Paleontology" (J. Cracraft, and N. Eldredge, Eds), pp. 79–111. Columbia University Press, New York.

Gardiner, B. G. (1982). Tetrapod classification. *Zool. J. Linn. Soc.* **74**, 207–232.

Gardiner, B. G. (1993). Haematothermia: Warm-blooded amniotes. *Cladistics* **9**, 369–395.

Gatesy, J., De Salle, R., and Wheeler, W. (1993). Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogen. Evol.* **2**, 152–157.

Gauthier, J., Kluge, A. G., and Rowe, T. (1988). Amniote phylogeny and the importance of fossils. *Cladistics* **4**, 105–209.

Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics* **9**, 83–91.

Goloboff, P. A. (1995). Parsimony and weighting: A reply to Turner and Zandee. *Cladistics* **11**, 95–104.

Harvey, P. H., and Pagel, M. D. (1991). "The Comparative Method in Evolutionary Biology". Oxford University Press, New York.

Hedges, S. B., Moberg, K. D., and Maxson, L. R. (1990). Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

and a review of the evidence for amniote relationships. *Mol. Biol. Evol.* **7**, 607–633.

Hennig, W. (1966). "Phylogenetic Systematics". University of Illinois Press, Chicago.

Howson, C., and Urbach, P. (1993). "Scientific Reasoning: The Bayesian Approach". Open Court, Chicago and La Salle, Illinois.

Huelsenbeck, J. P. (1991). When are fossils better than extant taxa in phylogenetic analysis? *Syst. Zool.* **40**, 458–469.

Huelsenbeck, J. P., Bull, J. J., and Cunningham, C. W. (1996). Combining data in phylogenetic analysis. *Trends Ecol. Evol.* **11**, 152–158.

Johnson, R. (1982). Parsimony principles in phylogenetic systematics: A critical re-appraisal. *Evol. Theory* **6**, 79–90.

Jones, T. R., Kluge, A. G., and Wolf, A. J. (1993). When theories and methodologies clash: A phylogenetic reanalysis of the North American ambystomatid salamanders (Caudata: Ambystomatidae). *Syst. Biol.* **42**, 92–102.

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

Kluge, A. G. (1976). Phylogenetic relationships in the lizard family Pygopodidae: An evaluation of theory, methods and data. Miscellaneous Publications, Museum of Zoology, University of Michigan **152**: 1–72.

Kluge, A. G. (1983). Cladistics and the classification of the great apes. *In* "New Interpretations of Ape and Human Ancestry", (R. L. Ciochon, and R. S. Corruccini, Eds), pp. 151–177. Plenum Publishing Corporation, New York.

Kluge, A. G. (1988). Parsimony in vicariance biogeography: A quantitative method and a Greater Antillean example. *Syst. Zool.* **37**, 315–328.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38**, 7–25.

Kluge, A. G. (1990). Species as historical individuals. *Biol. Philos.* **5**, 417–431.

Kluge, A. G. (1991). Boine snake phylogeny and research cycles. Miscellaneous Publications, Museum of Zoology, University of Michigan **178**: 1–58.

Kluge, A. G. (1994). Moving targets and shell games. *Cladistics* **10**, 403–413.

Kluge, A. G. (1995). Parsimony. *Herpetol. Rev.* **2**, 76–78.

Kluge, A. G. (1997a). Total evidence or taxonomic congruence: Cladistics or consensus classification. *Cladistics* (in press).

Kluge, A. G. (1997b). Old and new challenges to phylogenetic systematics: Consequences for character weighting. *Zool. Scripta* (in press).

Kluge, A. G., and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Syst. Zool.* **8**, 1–32.

Kluge, A. G., and Wolf, A. J. (1993). Cladistics: What's in a word? *Cladistics* **9**, 183–199.

Lakatos, I. (1993). Falsification and the methodology of scientific research programmes. *In* "Criticism and the Growth of Knowledge", (I. Lakatos, and A. Musgrave, Eds), pp. 91–196. Cambridge University Press, London.

Lanyon, S. M. (1993). Phylogenetic frameworks: Towards a firmer foundation for the comparative approach. *Biol. J. Linn. Soc.* **49**, 45–61.

Laws, R. A., and Fastovsky, D. E. (1987). Characters, stratigraphy, and "depopperate" logic: An essay on phylogenetic reconstruction. *PaleoBios* **44**, 1–9.

Lovtrup, S. (1977). "The Phylogeny of Vertebrata". John Wiley and Sons, New York.

Mindell, D., and Thacker, C. (1996). Rates of molecular evolution: Phylogenetic issues and applications. *Ann. Rev. Ecol. Syst.* **27**, 279–303.

Miyamoto, M. M., and Fitch, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**, 64–76.

Neff, N. A. (1986). A rational basis for a priori character weighting. *Syst. Zool.* **35**, 110–123.

Nelson, G. (1989). Cladistics and evolutionary models. *Cladistics* **5**, 275–289.

Nelson, G., and Platnick, N. I. (1981). "Systematics and Biogeography: Cladistics and Vicariance". Columbia University Press, New York.

Nelson, G., and Platnick, N. I. (1991). Three-taxon statements: A more precise use of parsimony. *Cladistics* **7**: 351–366.

Norell, M. A., and Novacek, M. J. (1992). The fossil record and evolution: Comparing cladistic and paleontologic evidence for vertebrate history. *Science* **255**, 1690–1693.

Owen, R. (1866). "On the Anatomy of Vertebrates". Vol. 1. Fishes and reptiles. Longmans, Green and Company, London.

Panchen, A. L. (1982). The use of parsimony in testing phylogenetic hypotheses. *Zool. J. Linn. Soc.* **74**, 305–328.

Panchen, A. L. (1992). "Classification, Evolution, and the Nature of Biology". Cambridge University Press, New York.

Patterson, C. (1978). Verifiability in systematics. *Syst. Zool.* **27**, 218–222.

Patterson, C. (1982). Morphological characters and homology. *In* "Problems of Phylogenetic Reconstruction", (K. A. Joysey, and A. E. Friday, Eds), pp. 21–74. Academic Press, New York.

Patterson, C. (1994). Bony fishes. *In* "Major Features of Vertebrate Evolution" (R. S. Spencer, Ed.). Short Courses in Paleontology, no. 7, pp. 57–84, convened by D. R. Prothero and R. M. Schoch. Publication of The Paleontological Society.

Patterson, C., Williams, D. M., and Humphries, C. J. (1993). Congruence between molecular and morphological phylogenies. *Ann. Rev. Ecol. Syst.* **24**, 153–188.

Penny, D., Hendy, M. D., and Steel, M. A. (1991). Testing the theory of descent. *In* "Phylogenetic Analysis of DNA Sequences", (M. M. Miyamoto, and J. Cracraft, Eds), pp. 155–183. Oxford University Press, New York.

Platnick, N. I. (1978). Gaps and prediction in classification. *Syst. Zool.* **27**, 472–474.

Platnick, N. I., and Gaffney, E. S. (1977). Systematics: A Popperian perspective. *Syst. Zool.* **26**, 360–365.

Platnick, N. I., and Gaffney, E. S. (1978a). Evolutionary biology: A Popperian perspective. *Syst. Zool.* **27**, 137–141.

Platnick, N. I., and Gaffney, E. S. (1978b). Systematics and the Popperion paradigm. *Syst. Zool.* **27**, 381–388.

Platnick, N. I., Griswold, C. E., and Coddington, J. A. (1991). On missing entries in cladistic analysis. *Cladistics* **7**, 337–343.

Popper, K. (1968). "The Logic of Scientific Discovery". Harper and Row, New York.

Popper, K. (1972a). "Conjectures and Refutations: The Growth of Scientific Knowledge". Routledge and Kegan Paul, London.

Popper, K. (1972b). "Objective Knowledge: An Evolutionary Approach". Oxford University Press, Oxford.

Popper, K. (1980). Evolution. *New Scientist* **87**, 611.

Popper, K. (1992). "Realism and the Aim of Science". Routledge, London.

Pritchard, P C. H. (1994). Cladism: The great delusion. *Herpetol. Rev.* **25**, 103–110.

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved

Riedl, R. (1978). "Order in Living Organisms". John Wiley and Sons, Chichester.

Rieppel, O. (1979). The classification of primitive snakes and the testability of phylogenetic theories. *Biol. Zentral.* **98**, 537–552.

Rieppel, O. (1991). Things, taxa and relationships. *Cladistics* **7**, 93–100.

Rieppel, O. (1992). Homology and logical fallacy. *J. Evol. Biol.* **5**, 701–715.

Sanderson, M. J., and Donoghue, M. J. (1989). Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795.

Siddall, M. E. (1995). Another monophyly index: Revisiting the jackknife. *Cladistics* **11**, 33–56.

Siddall, M. E., and Kluge, A. G. (1997). Probabilism in phylogenetic inference. *Cladistics* (in press).

Sober, E. (1975). "Simplicity". Clarendon Press, Oxford.

Sober, E. (1983). Parsimony in systematics: Philosophical issues. *Ann. Rev. Ecol. Syst.* **14**, 335–357.

Sober, E. (1988a). "Reconstructing the Past: Parsimony, Evolution, and Inference". MIT Press, Cambridge, Massachusetts.

Sober, E. (1988b). The conceptual relationship of cladistic phylogenetics and vicariance biogeography. *Syst. Zool.* **37**, 245–253.

Sober, E. (1993). "Philosophy of Biology". Westview Press, San Francisco.

Stamos, D. N. (1996). Popper, falsifiability, and evolutionary biology. *Biol. Philos.* **11**, 161–191.

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

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **11**, 158–163.

Wheeler, W. C. (1995). Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.* **44**, 321–331.

Wenzel, J. W. (1992). Behavioral homology and phylogeny. *Ann. Rev. Ecol. Syst.* **23**, 361–381.

Wenzel, J. W., and Carpenter, J. M. (1994). Comparing methods: Adaptive traits and tests of adaptation. *In* "Phylogenetics and Ecology", (P. Eggleton, and R. I. Vane-Wright, Eds), pp. 79–101. Linnean Society Symposium Series **17**. Academic Press, London.

Whewell, W. (1847). "The Philosophy of the Inductive Sciences, Founded upon their History". Parker, London.

Wiens, J. J., and Reeder, T. W. (1995). Combining data sets with different numbers of taxa for phylogenetic analysis. *Syst. Biol.* **44**, 548–558.

Wiley, E. O. (1975). Karl R. Popper, systematics, and classification: A reply to Walter Bock and other evolutionary taxonomists. *Syst. Zool.* **24**, 233–242.

Wilson, E. O. (1965). A consistency test for phylogenies based on contemporaneous species. *Syst. Zool.* **14**, 214–220.

Copyright © 1997 by The Willi Hennig Society
All rights of reproduction in any form reserved