Pergamon

PII: S0300–3256(97)00029–9

Sophisticated falsification and research cycles: Consequences for differential character weighting in phylogenetic systematics

ARNOLD G. KLUGE

Kluge, A. G. 1998. Sophisticated falsification and research cycles: Consequences for differential character weighting in phylogenetic systematics.—*Zool. Scr. 26:* 349–360.

Practicing phylogenetic systematics as a sophisticated falsification research program provides a basis for claiming increased knowledge of sister species relationships and synapomorphies as evidence for those cladistic propositions. Research in phylogenetic systematics is necessarily cyclic, and the place where the positive shift in understanding occurs is subsequent to discovering the most parsimonious cladogram(s). A priori differential character weighting is inconsistent with seeking the maximally corroborated cladogram (sensu Popper), because weighting adds to background knowledge, the evidence being then less improbable than it would be otherwise. Also, estimating weights from character state frequencies on a cladogram is inconsistent with the view that history is unique. Sophisticated falsification provides the place in the cycle of phylogenetic systematic research where weight of evidence can be evaluated and these inconsistencies do not apply. On balance, phylogenetic systematics appears to achieve greater coherence and generality as a result of focusing on the foundations for claiming increased knowledge and avoiding efforts to differentially weight characters. © 1998 The Norwegian Academy of Science and Letters

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

Introduction

Theories and methods of historical inference have been critically evaluated in the last 50 years, and by most accounts the coherence and generality of phylogenetic systematics have increased as the result of these debates. Current controversies in phylogenetic inference include the application of maximum likelihood (Swofford *et al.* 1996) and three-taxon analysis (Nelson & Platnick 1991); however, there appear to be few novel issues and arguments in either of these challenges. Is maximum likelihood anything more than syncretism (the evolutionary systematics of E. Mayr and G. G. Simpson; Mayr & Ashlock 1991), with its emphasis on deterministic (evolutionary process) models, special knowledge, and verificationism? Is three-taxon analysis anything more than a revised form of phenetics (Sokal & Sneath 1963), that is, theory independent operationalism? Are there no more substantive challenges to phylogenetic systematics?

In this paper, I expand on the idea that the day-to-day empirical science of cladistics is cyclic (Hennig 1966; Kluge 1991). The published studies of many taxa suggest an iterative process is at work (e.g., Amniota; Eernisse & Kluge 1993). Surely, no one doubts that there is traditional knowledge, that which comes from prior scientific study.[1] Popperian testability and sophisticated falsification, that

which I review and relate to phylogenetic systematics, formalize this perception. Further, the refutationist philosophy that all hypotheses are fallible, including observational propositions (so-called 'facts'), indicates why cladistic research is *necessarily* cyclic. Significant consequences may yet follow from this conclusion. For example, the particular methods used in phylogenetic systematics, and their accompanying arguments, must now be re-examined for logical consistency. Even widely favored methods, such as differential character weighting, cannot avoid being scrutinized. I expect my conclusions on weighting, which are to follow, to be controversial, and so there is at least the prospect that the content and limits of phylogenetic systematics may be critically evaluated, and changed if necessary.

Testability, cladistic parsimony, and sophisticated falsification

Popperian testability (Popper 1992; Farris 1995; see review by Kluge 1997) is a function of the logical interplay between evidence (**e**), hypothesis (**h**), and background knowledge (**b**). The logic plays out in two contexts, degree of corroboration and severity of test.

Degree of corroboration, **C**, the degree of support given to **h** by **e**, in light of **b**, is defined as

$$C(h,e,b) = p(e,hb) - p(e,b)/p(e,hb) - p(eh,b) + p(e,b),$$

---

[1] Traditional knowledge is not to be confused with 'conventional wisdom,' that which does not have a scientific basis.

which in the numerator reads, the probability of **e**, given **h** and **b**, minus the probability of **e** on **b** alone. **h** receives a higher degree of corroboration the smaller p(**e**,**b**), in particular when p(**e**,**b**) $\ll$ 1/2. Of course, **e** must be possible given **b**, but **e** should be improbable given **b** alone if **h** is to receive corroboration from **e**.

Severity of test, **S**, the severity of the test **e** interpreted as *supporting evidence* of the theory **h**, given the background knowledge **b**, is defined as

$$S(e,h,b) = p(e,hb) - p(e,b)/p(e,hb) + p(e,b).$$

S follows directly from the inverse relationship which exists between the logical probability of **h** and its degree of falsifiability. Expressed another way, there is a direct relationship between the logical improbability of **h** and its potential to be tested.[2] Explanatory power and degree of corroboration can be said to increase together, because **S** is also the power of the hypothesis to explain the evidence.

According to **C**, the maximum **h** can be corroborated is determined by the maximum it is testable. In turn, maximum testability is determined, in part, by the extent of the content of **h**, that is, by those qualities which determine the amount of empirical information conveyed by **h**. The extent of the content of **h** is some 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 informativeness is defined 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 all refer to degree of falsifiability, or testability, the potential to be tested. The amount of empirical content conveyed by a hypothesis increases with its degree of falsifiability. Informativeness is the failure to falsify that which is falsifiable, and that which is falsifiable relates to content. That a hypothesis must be tested by severe tests is the same as saying, given only **b**, that it must be subjected to those tests which are most likely to fail. Also, for **h** to be maximally corroborable, **h** must offer a variety of independent testable consequences. It is in this sense that independence relates to content. Still further, the maximum **h** is testable is inversely related to the number of assumptions made, in terms of the content of **h** relative to **b**.

In practice, hypotheses are to be tested in the severest manner possible, in relation to their empirical content, and the degree to which a hypothesis has withstood those tests constitutes its degree of corroboration. A hypothesis is falsified if it has so far failed relevant tests. 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. Falsification and corroboration comprise alternative results of testing. The number of tests performed does not relate directly to falsification or corroboration, because of the distinction between severe and weak tests.

A hypothesis can never be disproven, however, even where deductive logic applies, because the disconfirming

observational proposition itself is fallible. Thus, as a rule in a Popperian evaluation of scientific theories, the hypothesis which requires the ad hoc dismissal of the fewest falsifiers is preferred. In cladistics, the least refuted hypothesis is the most parsimonious cladogram that minimizes requirements for ad hoc hypotheses of homoplasy (Farris 1983). Consequently, the content of the hypothesis is maximized, empty statements are minimized. And, 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**). The potential to maximize explanatory power, that is, being able to provide an explanation for the 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 as due to inheritance. Again, in strictly Popperian terms, most parsimonious cladograms are most explanatory, because both **C** and **S** increase with p(**e**,**hb**), a term that occurs in their shared numerator. Thus, we have the logic of parsimony analysis in cladistics and its relationship to a Popperian evaluation of scientific theories.

There is, however, more to claiming increased knowledge than operationalizing Popperian testability with cladistic parsimony. Falsification must be "sophisticated" (Lakatos 1993: 116), in the sense that there is an interplay between evidence (**e**) and competing hypotheses $(h_1, h_2, h_3, \ldots h_n)$ that "leads to the discovery of novel facts." Such a progressive problem shift indicates a situation in which h_2 suggests something more than h_1 suggested, *as well* as suggesting more than is required by the data themselves. A progressive problem shift cannot come from maximizing **C**(**h**,**e**,**b**) alone.

The cyclic nature of cladistics research

Most taxa have been studied previously and with various kinds of data, and such published prior research cannot be omitted from new cladistic studies. To overlook prior hypotheses relevant to one's research shows a lack of scholarship; to purposely ignore them without cause is authoritarian. The number of possible cladistic hypotheses for a given set of taxa $(h_1, h_2, h_3, \ldots h_n)$ is not the problem, presuming the parsimony algorithm employed is effective at finding the maximally corroborated cladogram(s). The issue is how honestly the relevant data are surveyed for those synapomorphies that actually have the potential to refute a cladistic hypothesis, those synapomorphies that can count as independent ad hoc hypotheses of homoplasy. To test both cladograms and synapomorphies over and over again is consistent with Popperian testability, and it is in such cycles of research that hypotheses will be critically re-evaluted. Excluding prior evidence from an analysis reduces the content of **h** and severity of test, and these omissions are also expected to be revealed with repeated critical review.

Realistically, cladistics is a plexus of research cycles. The complexity is due to the fact that testable phylogenetic hypotheses occur at different levels of taxonomic inclusiveness, and because cladistic research often identifies par-

---

[2]Logical improbability = 1 − logical probability.

aphyletic assemblages, which then must be reclassified monophyletically. The many kinds of data and the recognition that the data used to test competing phylogenetic propositions are themselves fallible hypotheses requiring reanalysis, and possibly reinterpretation, also contribute to the complex nature of phylogenetic systematic research.

Figure 1 summarizes the more general aspects involved in one rotation of the cycle. Phase A is all of the scholarship which leads to the familiar taxon by character matrix. Identifying terminal taxa (species or more inclusive clades), selecting organisms representative of those taxa, and delimiting the synapomorphies which constitute potential evidence are the principle operations. Preliminarily testing the monophyly of each terminal taxon means that phase A is itself potentially cyclic. Monophyletic terminal taxa are critical, because they omit the possible patterns of character generality which can serve as evidence. The characters summarized in phase A are of two kinds: (1) newly discovered generalities (observed and recorded for the first time), and (2) those synapomorphies used in previous research. The latter characters to some degree have more corroborated content, to the extent they have been tested (phase B), reanalyzed and possibly reinterpreted (phase C). The organisms on which the observations are made are not chosen at random. Rather, they are selected for the unambiguous evidence they provide as to the plesiomorphic condition of a terminal taxon.

Some students of phylogenetic inference have argued that character compatibility, testing characters individually, against each other, for their self-consistency (Wilson 1965; Le Quesne 1969), should be used in the identification of potentially informative characters (phase A). However, character compatibility cannot be endorsed

at any place in the cycle of cladistics research, because: (1) excluding characters according to their incompatibilities is a form of verificationist weighting (see below); (2) compatibility/clique analysis (Le Quesne 1969) does not necessarily maximize explanatory power (Kluge 1976); (3) the pair-wise character tests of compatibility are not as severe as when all those tests are conducted simultaneously with cladistic parsimony (Kluge 1997).

Phase B (Fig. 1) is where all of the possible cladograms $(h_1, h_2, h_3, \dots h_n)$ are tested simultaneously with a set of potentially disconfirming synapomorphies (e), in light of background knowledge (b). The more the competing cladograms are severely tested and refuted the better. No phase B operation assesses the truth of a hypothesis of species relationships, or measures the accuracy of the evidence. It is of course phase B where cladistic parsimony is applied in the minimization of requirements for ad hoc hypotheses of homoplasy. "Descent with modification" as necessary and sufficient background knowledge peculiar to phylogenetic inference has been discussed elsewhere (Kluge 1997).

There are many activities in phylogenetic systematics that address the relationship between maximally corroborated hypotheses and evidence; however, only those scientific tests which lead to a progressive problem shift constitute phase C (Fig. 1). To be sure, increasing the empirical content of cladistic hypotheses and/or minimizing ad hoc hypotheses of homoplasy may justify character reanalysis. However, it is only in the sense that a maximally corroborated hypothesis (h_2) suggests something more than a competing proposition (h_1) suggested, as well as suggesting more than is required by the data themselves, that there can be a positive increase in the

selecting terminal taxa
and identifying synapomorphies

increased knowledge of
phylogeny and homologies

taxa by character
matrix

sophisticated
falsification

maximally
corroborated cladogram(s)
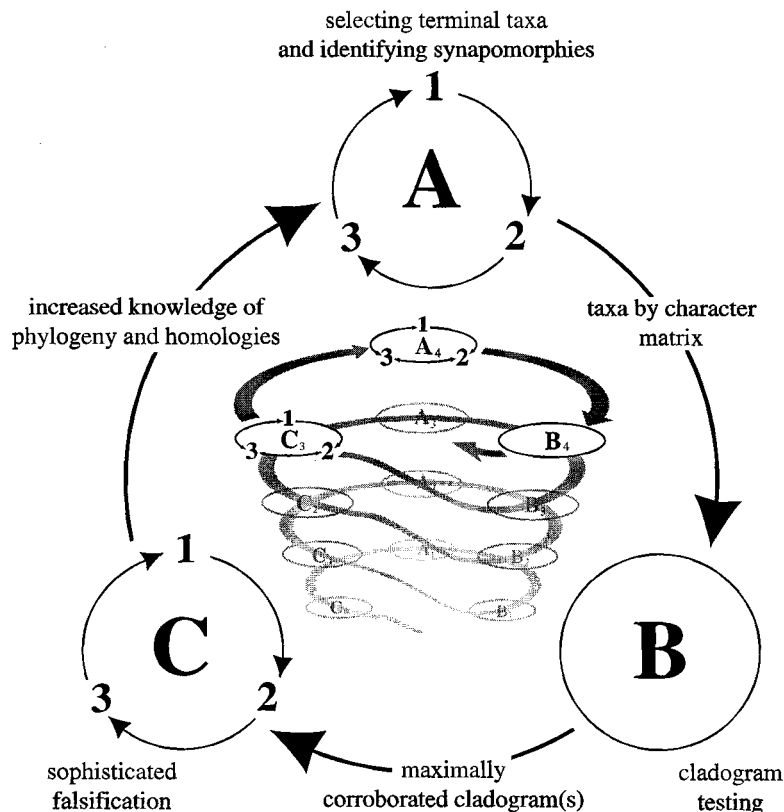
cladogram
testing

Fig. 1. A graphic outline of the cyclic nature of cladistics research. A complete round (A–C) is but one part of a more complicated pattern of testing hypotheses of sister group relationships and synapomorphy relations. See text for further explanation.

knowledge of phylogeny and synapomorphies. None of the phase C operations assesses the truth of a hypothesis of species relationships, or measures the accuracy of the evidence—all data are 'observational propositions' which are fallible in themselves.

At least tests of consilience and some kinds of character reanalysis meet the conditions of sophisticated falsification (Lakatos 1993). Consilience offers especially severe tests (sensu Popper 1992), and which therefore should be an important part of phylogenetic systematic research. Tests of consilience involve the search for incongruence between the maximally corroborated cladogram(s) discovered in phase B and the branching patterns provided by earth history and parasite (or host) phylogenetic hypotheses (Kluge 1983). According to Kluge (1997), whatever congruence is observed between the maximally corroborated cladogram(s) and earth history and/or parasite (or host) phylogenetic hypotheses it is more improbable than observing an equally large set of congruent synapomorphies. The greater improbability derives from the fact that vicariance biogeography and host-parasite coevolution allow for both temporally (vertically) and spatially (horizontally) evolved patterns, whereas cladistic analysis only presupposes the former (Sober 1988*a*). Thus, the reanalyses of those synapomorphies responsible for the incongruence between the maximally corroborated cladogram(s) and the patterns provided by earth history and parasite (or host) phylogenetic hypotheses are of special interest.

Hypotheses of character state polarity and similarity (including sequence alignment) and assumptions of character independence are the particular concerns of character reanalysis (e.g., see Jones *et al.* 1993). Many different strategies have been used to identify characters for reanalysis, including consilience, and it is unclear whether there is a logical order to their employment. One approach is to focus the reanalysis on the characters diagnostic of those clades which are relatively weakly corroborated, as determined by unweighted branch support (Bremer 1994; not the weighted support of Gustafsson & Bremer 1995). Another approach, although not commonly employed, involves the reanalysis of the disconfirming evidence, because minimizing requirements for ad hoc hypotheses of homoplasy is the criterion for identifying the most parsimonious cladogram(s). As Farris (1983: 10) emphasized, "[i]f external evidence favors the interpretation of homoplasy,... that hypothesis is not ad hoc." The external evidence, I conjecture, is that which comes from character reanalysis. "Scrutiny of a structure may indicate that what had been regarded as a single feature is instead the union of distinct qualities. Sometimes conflicts can be removed by changing hypotheses of plesiomorphy, and surveying the distribution of a feature among taxa may provide support for such a reinterpretation. If a conflicting character survives all attempts to remove it by searching for such evidence, then the conclusion of homoplasy in that character, required by selecting a placement, satisfies the usual definition of an ad hoc hypothesis" (Farris 1983: 10). An incongruence which can only be judged an investigator error cannot be counted as an ad hoc hypothesis. It only deserves to be recoded.

Other phase C activities may include manipulating lineages in a reticulate pattern, in order to determine if a hybrid origin explains the data more parsimoniously (Nelson 1983), and transformation series analysis (Mickevich 1982), where hypotheses of character state adjacency are tested (Lipscomb 1992; see also Mickevich & Lipscomb 1991).

It is certainly a truism that not all characters yield equally strong evidence of phylogenetic relationships (Farris 1983: 11); however, the issues of how those weights are to be formulated and where in the cycle of cladistic research they are to be applied remain hotly debated topics (e.g., Goloboff 1995). The following discussion indicates that there are logical and analytical problems when differential character weights are applied in the process of discovering the maximally corroborated cladogram (phase B); however, sophisticated falsification (phase C) seems to provide an appropriate 'how and where' for judging *weight of evidence*[3] (see also Turner & Zandee 1995). For example, the evidence responsible for equally most parsimonious hypotheses (secondary cladograms) would be evaluated in phase C (e.g., with character reanalysis), not in phase B (e.g., with a posterior [successive] character weighting).

Character weighting: an overview

Three kinds of differential character weighting are recognized according to when the weights are applied. There are the familiar a priori and a posteriori kinds of weighting, which are determined immediately before and immediately after phase B (Fig. 1), respectively. A priori applications include those based on character compatibility (e.g., Le Quesne 1969, 1972; Gauld & Underwood 1987; Moody & O'Nolan 1987; Sharkey 1989) and relative rate tests (e.g., Mindell & Thacker 1996). Successive weighting (Farris 1969; Carpenter 1988, 1994) exemplifies the most often used objective a posteriori application. In addition, there is the less well known character weighting within phase B, that which Goloboff (1993) has used as the basis for his new optimality criterion for choosing among competing cladograms.

All of the justifications for differential character weighting, whenever determined, immediately before or after, or during, phase B (Fig. 1), follow a verificationist agenda—the application of weights supposedly improve one's chances of discovering objective truth, *the* phylogeny (Mindell 1997). Weighting under any such guise negatively impacts C and S, either by adding to **b**, or by reducing the empirical content of **h**.

The details of the arguments for differential character weighting vary considerably. For example, there is the simple proposition that no one supposes that all characters deserve the same weight, that is, that they all provide equally strong evidence in phylogenetic inference. Then there is the notion of conservatism in systematics, the more conservative the character the better. In the older literature, character conservatism was simply related to speculations on homology and homoplasy, the more instances

---

[3] I believe this distinction deserves the different terminologies employed, 'differential character weighting' and 'weight of evidence.'

of independent evolution one imagined a character to have, the less 'reliable' it was considered to be.[4] The discontinuous nature of the historical record may also have given license to the idea that relative constancy is important in characters if they are to track that record accurately. Appeals to special knowledge concerning characters (e.g., Hecht & Edwards 1977), while common in the older literature, have been soundly rejected by phylogenetic systematists (e.g., Kluge 1994).

All of the more objective arguments for differential character weighting in phylogenetic systematics appear to be based on conservatism-constancy, which involve references to rate-related within population (or taxa)[5] variation or character reliability. Farris' (1966) within population basis for weighting has only rarely been put into practice (e.g., Kluge 1969), even though it was restated in terms of taxa (Kluge & Farris 1969). Sober's (1986) theoretical argument that frequencies of traits in populations are not good descriptors of taxa appears to be contradicted by the empirical findings of Kluge & Kerfoot (1973), which illustrate a reasonably good correspondence between amounts of within and between population variation. The within taxon approach has been applied most recently to molecular data (Mindell & Thacker 1996).

Character reliability methods have come to predominate in cladistics,[6] in spite of opposing points of view as to the relationship between parsimony and weighting. For example, Sober (1986: 28) argued that "[p]arsimony is not a device that tells biologists how to weight characters; rather, parsimony requires that the characters already should be weighted." The arguments advanced by many prominent phylogenetic systematists for believing there is a *direct* relationship between parsimony and weighting are summarized in the following historical review.

Reliability weighting: historical digressions

Old history

An evolutionary basis for character reliability weighting in classification seems to have originated with Darwin (1859: 415), who stated that the importance of organs "depends on their greater constancy throughout large groups of species; and this constancy depends on such organs having generally been subjected to less change in the adaptation of the species to their conditions of life."

The idea of a constancy relation to information content was introduced into phylogenetic systematics by Farris 1966, 1970*a*. It must be remembered that pheneticists (e.g.,

Sokal & Sneath 1963: 118–120) advocated weighting characters according to their constancy *between* operational taxonomic units (OTUs), which was termed 'equal weighting,' and they were otherwise critical of weighting. Farris 1966 responded effectively to their position with an argument for a priori differential character weighting according to a character's variation within discrete biological populations, as measured by the sample variance. The connection he developed between within-OTU variability, evolutionary rate, and weighting was summarized by Kluge & Farris (1969: 4): "The rate at which a character can change in evolution is necessarily limited by the variability of the character within populations. Selection, no matter how intense, cannot change the average character state of a population in some unit time by a greater amount than the range of values available in the population in that unit time. If a character has high variability within OTUs, then a large difference between two OTUs does not imply lack of close relationship, since the variable character could have changed rapidly. If on the other hand, a character has low variability within OTUs, then a large difference between OTUs is probably indicative of lack of close relationship, since the highly stable character probably could not evolve rapidly. Hence, in drawing taxonomic conclusions, we place greater weight on characters with low variability within OTUs." Farris (1966: 589) was careful to point out that the "problem of avoiding error due to convergence is quite distinct from that of finding conservative characters" at any rank, a convergent character being unreliable for taxonomic purposes "even though it is conservative according to the criterion of constancy within populations."

Successive (iterative, a posteriori) character weighting also appears to have connections to the constancy concept. For example, according to Farris (1966: 587), there is "little doubt that if a character is constant throughout a group that is natural in an evolutionary sense, then that character is a more reliable indicator of evolutionary relationship than is a character that varies widely in the same group." In any case, the concept of character reliability, as measured in terms of self-consistency, became central to his development of successive weighting (Farris 1969; see also Goloboff 1993: 85). For example, (p. 374), "[c]haracters that are reliable for cladistic inference are those that are consistent with the true phyletic relationships, that is, those that have little homoplasy." Further, in completing this theoretical basis for successive weighting, he assumed that "[s]everal unreliable characters will each vary from the phylogeny in its own random way, and chances are very slight that a series of random variables will by accident form a pseudo–hierarchic pattern of variation" (p. 376). His ingredients for operationalizing cladistic reliability were the unit character consistency index (ci$_i$; see Kluge & Farris 1969) and the hierarchic correlation of a set of characters. That a "set of variables with high hierarchic correlation will all be highly consistent with a single branching pattern" was the defining property of hierarchic correlation, and from which it followed that "cladistically reliable characters are hierarchically correlated with each other" (p. 376); "cladistically reliable characters [have] high unit character consistency with the true cladistic relationships of the group under study" (p. 375). And

---

[4]The terms precision, reliability, and accuracy are all too often used loosely in discussions of differential character weighting. In this paper I take precision to mean exactness, being definite, clear, and unambiguous (Kluge 1998) and reliability and accuracy to mean conformity to truth (Farris 1969: 374–376).

[5]These are not to be confused with the pheneticists' between OTU methods which formed the basis for their claims of equal character weighting.

[6]The current literature is rife with many different kinds of character reliability applications, such as character state transformation weights (Williams & Fitch 1989), combinatorial weights in molecular sequence data (Wheeler 1990), and character state weighting based on asymmetrical step-matrices (Albert *et al.* 1992).

finally, Farris concluded that (p. 377) "successive tree estimates and consistency estimates [the unbounded concave weight function being recommended] will become progressively better until the process terminates at the correct phylogenetic tree."

Farris (1988: 19; see also Carpenter 1988) also noted that successive weighting was useful in reducing "the ambiguity of complex data sets: there may be many trees of minimal length without weighting, but substantially fewer with successive weights." Of course, removing ambiguity in the results will also be sought in phase C (Fig. 1), with, for example, character reanalysis, and there appear to be some examples of successive weighting actually increasing the number of secondary cladograms (J. Wenzel, pers. comm.). Ambiguity in the individual, potentially disconfirming, synapomorphies is error which is expected to have been removed prior to phase A (Fig. 1).

New history

Carpenter (1988: 292; see also Carpenter *et al.* 1993: 137–138) appears to have been the first to argue that Farris' (Farris 1969, 1988) successive weighting method "*is* related to evidential support", because that "method is developed directly from the concept of 'cladistic reliability'". Curiously, Carpenter (1988: 292) cited only Farris 1983 (not Farris 1966, 1969, or 1970*a*) as the basis for his rationale for successive weighting, the salient quote being "No one supposes...that characters in general all deserve the same weight—that they all yield equally strong evidence" (Farris 1983: 11). Such a citation is especially curious, because the rationale in this quote provides no reason for endorsing cladistic reliability *in particular* as a basis for weighting. I believe Farris' more general statement should not have been so over-interpreted, because he does not appear to have been arguing the issues of weighting per se (pp. 10–12). Rather, he seems to have been making the case for parsimony being necessary operationally in assessing requirements for ad hoc hypotheses of homoplasy. For example, as Farris concluded (p. 11), even "the characters comprising a congruent suite are hardly *observed* to be free of homoplasy" (my italics).

A further questionable modification of Farris' (Farris 1966, 1969, 1970*a*) conceptualization of cladistic reliability is evident in a more recent paper by Carpenter (Carpenter 1994; see also Carpenter *et al.* 1993). To begin with, Carpenter's (Carpenter 1994: 216) reference to reliability being "*an extension of* parsimony" (my italics) is unjustified—as is his conclusion: "Successive weighting allows the characters of a given data set to judge themselves in terms of their reliability; that is best fit to the solution supported by all the characters. Homoplastic characters, being less reliable, are down-weighted, hence successive weighting employs evidence that would otherwise be discarded..." Actually, the modification seems to have originated in the earlier papers to which Carpenter 1994 referred, namely in Carpenter (1988: 292), and in Farris & Kluge 1979, which he cited therein. In fact, Farris & Kluge (1979: 402) stated unambiguously that Farris 1969' successive weighting techniques were used "*in conjunction with* parsimony methods" (my italics; see also Kluge & Farris 1969; Farris 1983: 10). Those authors did not, nor did Farris 1983,

claim that cladistic reliability is *an extension of* parsimony. As I see it, parsimony is simply the operation of optimizing on a criterion.

Unfortunately, Carpenter's error has been compounded. For example, Goloboff (1993: 83) has now concluded "that parsimony *does not preclude* weighting, but rather that it *requires* weighting." Such a novel position, in turn, has led Agosti *et al.* (1996: 79–80) to assert that "[c]onsequently, given the logical necessity or requirement of character weighting, methods for accomplishing objective character weights are highly desirable." There are substantial downsides to this misadventure. As Carpenter's (Carpenter 1994: 217–218) own review indicates, cladistic parsimony continues to be criticized unnecessarily for its presumed dependence on deterministic evolutionary models, like those justifying character reliability. That is, following from the discussion above, reliability weighting requires a model of conservatism or slow rates of change.

Goloboff 1993's required, "heaviest" tree, weighting criterion deserves careful review, because it significantly changes the basis for determining maximally corroborated cladistic hypotheses (phase B in Fig. 1). Like successive weighting, Goloboff's criterion uses hypotheses of homoplasy to estimate character reliability. In addition, it maximizes total fit, the sum of the fits of the characters among all possible trees, the cladogram with highest total fit being chosen. The heaviest tree criterion is not iterative (unlike successive weighting); it evaluates character consistency one cladogram at a time (not across multiple equally most parsimonious hypotheses, like successive weighting); relative weights are determined by the ratio of the fit slopes to a concave function, $f_i = (k+1)/(s_i + k + 1 - m_i)^7$ (not, as in successive weighting, by the ratio of the fits themselves); the fitting functions exhibit, at least in theory, precision to two or three significant digits (not the 0–10 integer scale employed by Farris 1988); and, most challenging to phylogenetic systematic principles, it produces trees which are maximally "reliable", but which are not necessarily most parsimonious (of minimum length; sensu Kluge & Farris 1969, Farris 1970*b*, and Farris *et al.* 1970). The latter departure from traditional cladistic practice is justified by Goloboff (1993: 88) on the grounds that the best hypothesis, for a given set of data, "is the one which confers the highest weight on the data. Searching for the "heaviest" trees, therefore, is in direct agreement with cladistic ideas, and provides what self-consistency alone did not." Thus, Goloboff has reinterpreted explanatory power (Farris 1983), the relationship between evidence and competing hypotheses, as a weighted explanatory power, but for which he has given no further rationale than the premise of fit increase itself.

Explanatory power is also weighted under Farris 1969's successive weighting criterion when it decides among equally most parsimonious cladograms, the chosen hypothesis(es) being longer (of increased weighted length). When successive weighting points to a different topology, as it does occasionally, explanatory power has also been transformed.

---

[7] k = constant for degree of concavity; s = minimum number of steps a character can exhibit on the cladogram; m = minimum number of steps a character can exhibit on any cladogram (the one where there are no requirements for ad hoc hypotheses of homoplasy).

Differential character weighting: criticisms

I believe all ordinary uses of differential character weighting in phylogenetic systematics, including all forms of reliability weighting, must be called into question. Of particular importance to this conclusion is the verificationist justification which accompanies differential character weighting. Also, the uniqueness of history is logically inconsistent with frequency as a basis for defining "good" and "bad" characters (Siddall & Kluge 1997; see however Lauder 1982: 65, Lauder 1990: 332). And, as will be discussed below, at least some forms of frequency weighting can affect the independence of tests.

Obviously, the terms a priori and a posteriori weighting belie the cyclic nature of cladistic research, and for which more appropriate, atemporal, labels are required. Further, what has been called a priori differential weighting cannot be employed if cladistics is to retain its refutationist character. According to Popperian testability, a cladistic hypothesis receives corroboration from synapomorphies only to the degree that the evidence is improbable given the background knowledge alone. A priori differential character weighting adds to the background knowledge and 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. That is any cladogram can be supported on the basis of some arbitrarily chosen a priori weighting scheme. Popperian testability only requires that each character in the data matrix provides an independent, potentially disconfirming, test. Independently evolved homoplasious characters are of equal weight in this sense of test.

Character reliability is measured as some function of the total number of steps a character exhibits over the entire cladogram. Those extra steps required to explain the data are ordinarily counted as requirements for ad hoc hypotheses of homoplasy, independent evolution due to some biological process, deterministic or stochastic. Investigator error can also result in the same apparent pattern and interpretation (Farris 1969: 374), and it may be argued that those incongruences should also be down-weighted, because they might confound the discovery of the true phylogeny. But fallibility does not specify why something is fallible. If error were distinguisable, a correction would have been undertaken prior to phase B (Fig. 1). As Kluge & Farris (1969) pointed out, so-called apparent homoplasy may result from an injudicious coding of characters.

A further complication is the fact that the cladogram over which the homoplasy is counted is limited to the terminal taxa in the data matrix, which does not take account of the character evolution which has occurred more globally or which has taken place within terminal taxa. Thus, at best, a suite of weights can constitute only a crude hypothesis on frequencies of incongruence due to supposed homoplasy and/or investigator error.

Figure 2 illustrates the calculation of self-consistency weights and how they affect hypothesis choice. Assume all parts of the cladogram in Fig. 2—A, except the relationships among terminal taxa A–C, are highly corroborated by unique and unreversed synapomorphies. The accompanying three characters (1–3) provide evidence for equally parsimonious hypotheses of relationships among

taxa A–C weighted by overall fit. Characters 1 and 2 exhibit a minimum of four steps each, and unit character consistency index (ci_{1-2}) of 0.25, whereas character 3 has a minimum of two steps, and a consistency index (ci_3) of 0.50. The weighted evidence supports equally well the alternative hypotheses (A,B) and (B,C), the strict consensus of which is the unresolved (A,B,C) clade. All of the minimum extra steps contributing to the ambiguity in the evidence come from the recovered (F(G(H(I(J,K))))) part of the cladogram, which is historically independent of those apomorphies in (((A,B,C)D)E). If the relationships of the (A,B,C,D,E) clade were to have been studied more narrowly, without the more global reference to the relationships in the (F,G,H,I,J,K) clade, then (((A,B)C)D,E) would have been the best supported, completely resolved, cladogram (Fig. 2—A′). This example, and any number of others (e.g., Fig. 3), demonstrates that self-consistency weighting presupposes a character that behaves badly in one part of the cladogram, for whatever reason, must do so as well elsewhere in the cladogram. But why must that be so? Aren't characters 1 and 2 maximally congruent when examined over terminal taxa A–E, and isn't there twice as much support for the (A,B) clade (Fig. 2—A′)?

As noted earlier, Goloboff (1993) used character self-consistency weighting within phase B of the cladistics research cycle as an optimality criterion for choosing among cladograms. His criterion gives greater weight to characters more consistent with tree(s), and it is the heaviest cladogram which is considered optimal. The basis for, and the problems associated with, this approach are illustrated with four sets of examples (Fig. 3—A–D). In Fig. 3—A–A′, assume that the ingroup and outgroup histories and the distribution of two characters, one highly variable, the other much more conservative. There are equally parsimonious ingroup clades, (A,B)C and (A,C)B, each with a minimum of ten steps for the two characters scored. However, according to Goloboff's criterion, the former hypothesis is preferred, because it confers more weight (1.111) than the latter (0.625). Hypothesis (A,C)B confers less weight (Fig. 3—A′), because its confirming character is more variable among the outgroups. In Fig. 3—B–B′, assume the ingroup and outgroup histories and the distribution of two characters, one being only slightly more conservative than the other. There are equally parsimonious ingroup clades, (AB)C and (AC)B, each with a minimum of four steps for the two characters scored. However, according to Goloboff's criterion, the former hypothesis is preferred, because it receives more weight (1.333) than the latter (1.000). Hypothesis (A,C)B receives less weight (Fig. 3—B′), because its confirming character is more variable among the outgroups.

Obviously, the conclusion in Fig. 3—C–C′ is quite different. Here, the longer, less parsimonious, cladogram, (A,B)C with 19 steps, is considered optimal under Goloboff's criterion, because it has greater character weight (1.222). Although the alternative hypothesis, (A,C)B, is more parsimonious, with 18 steps, it receives less weight (0.750), because its confirming character is more variable among the outgroups. Fig. 3—D–D′ is a similar example, where the less parsimonious ingroup hypothesis is preferred, because of variation in the outgroup taxa. How-

												Min. Steps:	Fit (ci):
Character 1:	1	1	0	0	0	1	0	1	0	0	1	4	0.25
Character 2:	1	1	0	0	0	0	1	0	1	0	1	4	0.25
Character 3:	0	1	1	0	0	0	0	0	0	1	0	2	0.50
Terminal Taxa:	A	B	C	D	E	F	G	H	I	J	K		

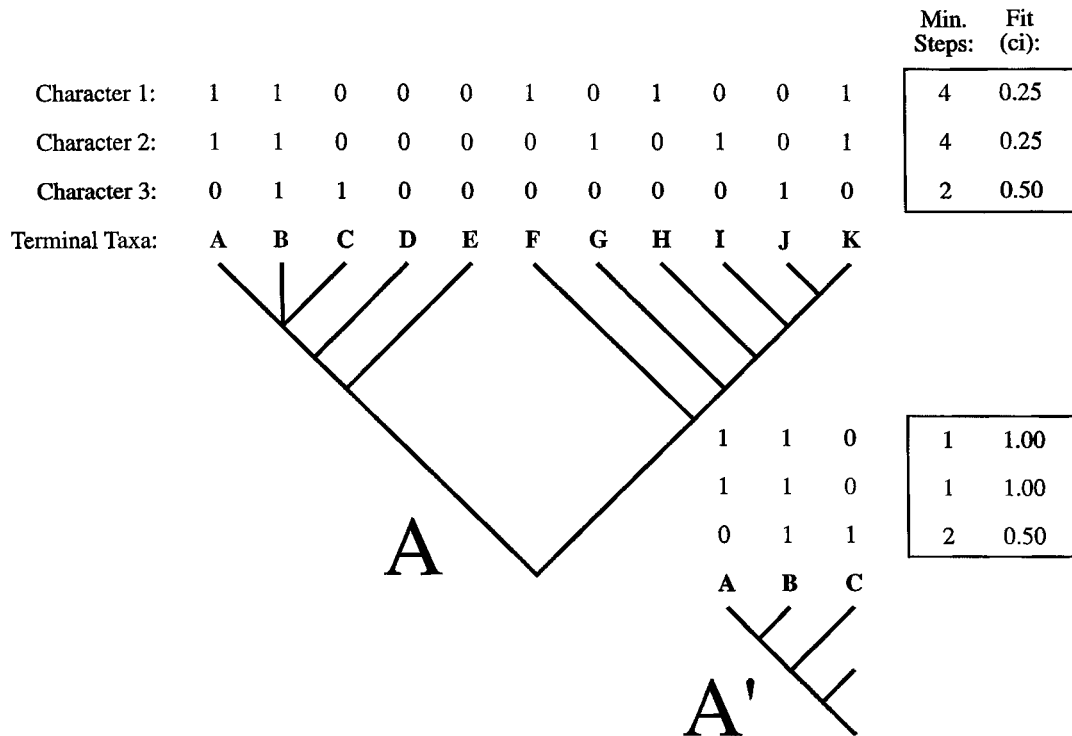|			Min. Steps:	Fit (ci):
1	1	0	1	1.00
1	1	0	1	1.00
0	1	1	2	0.50
A	B	C		

Fig. 2. An example of differential character weighting (modified after M. Siddall, pers. comm.). The cladogram in panel A is given. The self-consistency weights for the three characters in panel A are responsible for the equally parsimonious relationships, (A,B) and (B,C), which are summarized as a trichotomy. Panel A′ illustrates the resolved most parsimonious relationships based on the unweighted characters. See text for further explanation.

ever, this example is noteworthy, because there is nothing in the weights of the individual characters that would suggest the most parsimonous hypothesis (Fig. 3—D′).

Goloboff (1993: 88) acknowledged that longer cladograms can be optimal under his self-consistency weighting criterion (Fig. 1, phase B), and argued that "[s]earching for "heaviest" trees...is in direct agreement with cladistic ideas, and provides what self-consistency alone did not". Still, those less parsimonious weighted hypotheses are more ad hoc—they explain the original observations less well—an observation which Carpenter (1988: 291–292) appears to have used to reject Brooks *et al.* 1986's D measure as a weighting criterion. The bottom line is that if a weighted cladogram is different from the most parsimonious unweighted hypothesis, then the transformations on the former proposition are overall necessarily more ad hoc than those on the latter. That is, in the former, there will necessarily be more instances of change.

The frequentist interpretation of character state incongruence is an obvious problem for differential character weighting.[8] On the face of it, there is a logical contradiction; namely, states which are hypothesized as incongruent cannot then be counted as historically the same. The acquisitions, or losses, of character states which are hypothesized to have occurred within exclusive clades represent instances of independent evolution. In an apparent attempt to counter this problem for incongruence, Goloboff (1993: 84; see also Goloboff 1995) assumed that characters are differentially committed to homoplasy: "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." But, such an assumption presupposes a general underlying biological process, or a general character condition responsible for error in observation or coding, which like a priori weighting obviously adds to background knowledge and decreases degree of corroboration. Moreover, that general biological process is as yet unspecified, and as such can't be tested. In any case, where could the evidence for the process come from, if not from a cladistic analysis where self-consistency weighting is absent? I see no reason to pretend that a set of non-homologs forms a natural class justifying prediction.

Prediction and postdiction (retrodiction) are often confused in historical biology. Strictly speaking, the former extends only from intensionally defined classes, the latter being descriptive of ostensively or extensionally defined historical individuals (Frost & Kluge 1994). The importance of the confusion is exemplified by Goloboff (1993: 84), where there is no evidentiary justification for the "prediction" he makes concerning homoplastic characters. The subtlety in the distinction between prediction and postdiction in phylogenetic inference can be illustrated with a simple example (largely due to P. Goloboff, pers. comm.). Imagine encountering a live spider. You determine that it has all of the external features diagnostic of a group that is also known to possess venom which is lethal to humans. How do you handle the spider, without actually determining the toxicity of its venom and observing its biting behavior when it is disturbed? Does membership in such a group actually predict the spider in question is deadly, which then demands extreme care in handling the organism? Even assuming the spider has been correctly determined to be a part of a group all of whose others' bite is known to be lethal, there remain two possibilities: either

---

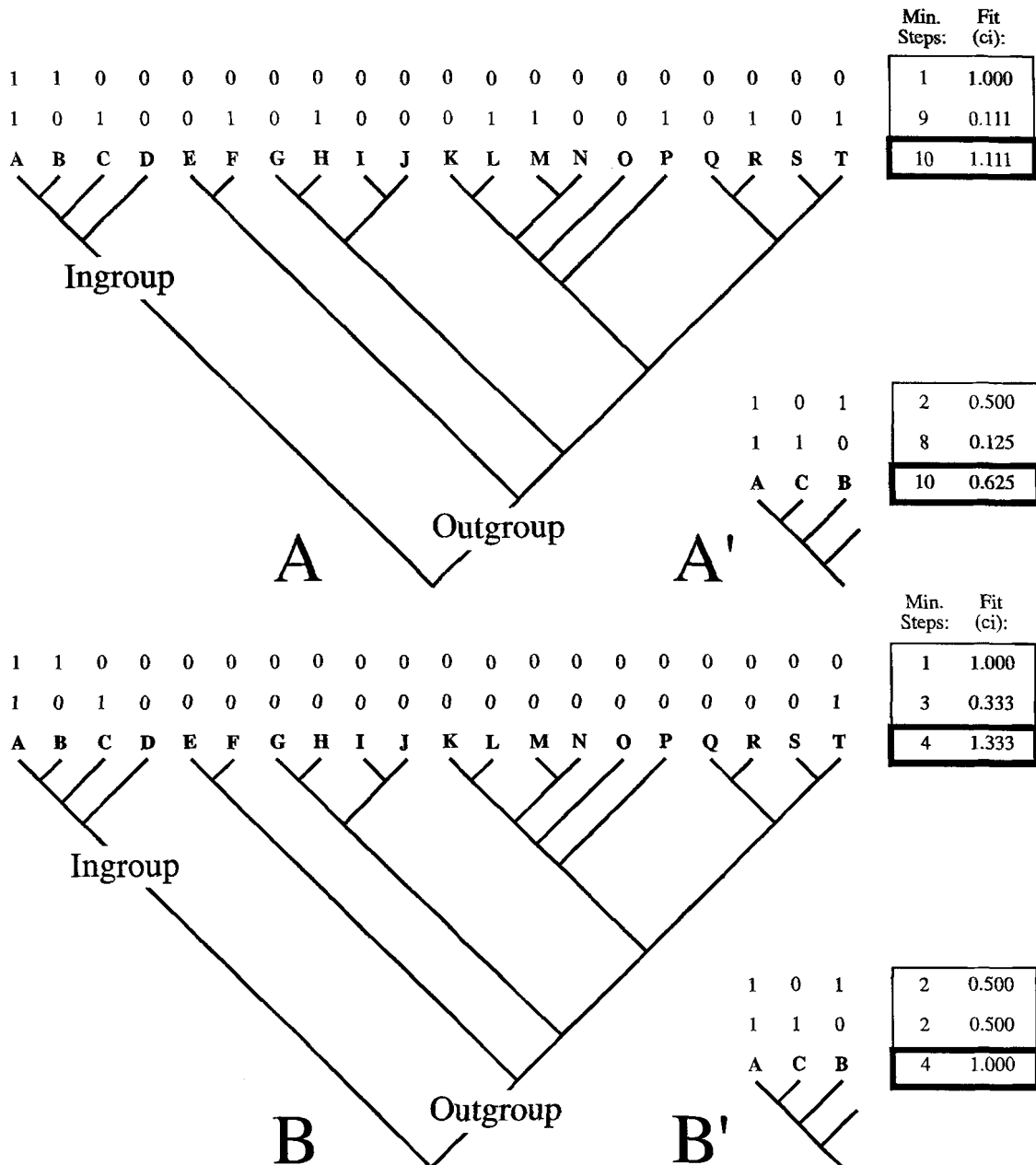[8] This also applies to Sober 1988b's Smith–Quackdoodle theorem.

Fig. 3. Four sets of alternative results (panels A–A'; B–B'; C–C'; D–D') illustrating the relationship between self-consistency weights and most parsimonious cladograms (modified after M. Sorenson, pers. comm.). In panels A–A' and B–B' the more heavily weighted cladogram is also one of the most parsimonious alternatives. In panels C–C' and D–D' the more heavily weighted cladogram is not a most parsimonious hypothesis. See text for further explanation.

the spider's venom is lethal, or it is not, extreme toxicity being absent (meaning secondarily reduced, or lost altogether, in this part of arachnid history). To ignore the latter possibility is to be anti-evolutionary, or to deny the possibility of teratology. Ignoring these alternatives may also be judged unscientific. Scientists are engaged in examining alternative possibilities, not presupposing the impossible. That you use long forceps to collect the spider, whether or not it is benign, does not validate the notion of phylogenetic classification being predictive. Rather, practically speaking, it costs you little, if anything, to play it safe.

Lastly, there is the issue that weighting can affect independence. If weights are based on distributional values (e.g., base compositions, TV:TI ratios) across all characters or across a character class (3rd positions) then independence of these potential falsifiers is lost (M. Siddall,

pers. comm.). If in successive weighting it is the consistency of other characters which determines the inconsistency of the down-weighted characters, then again there is loss of independence. Apparently, only if weights were to be applied arbitrarily would character independence be retained.

Discussion

Obviously, it is important for scientists to pursue increased knowledge. Surely in time, for example, interest in phylogenetics would decline without real progress in understanding species relationships and explaining synapomorphies as either homologous or homoplasious. Hennig 1966's principle of reciprocal clarification signaled the need to demonstrate progress in cladistics, but his ideas
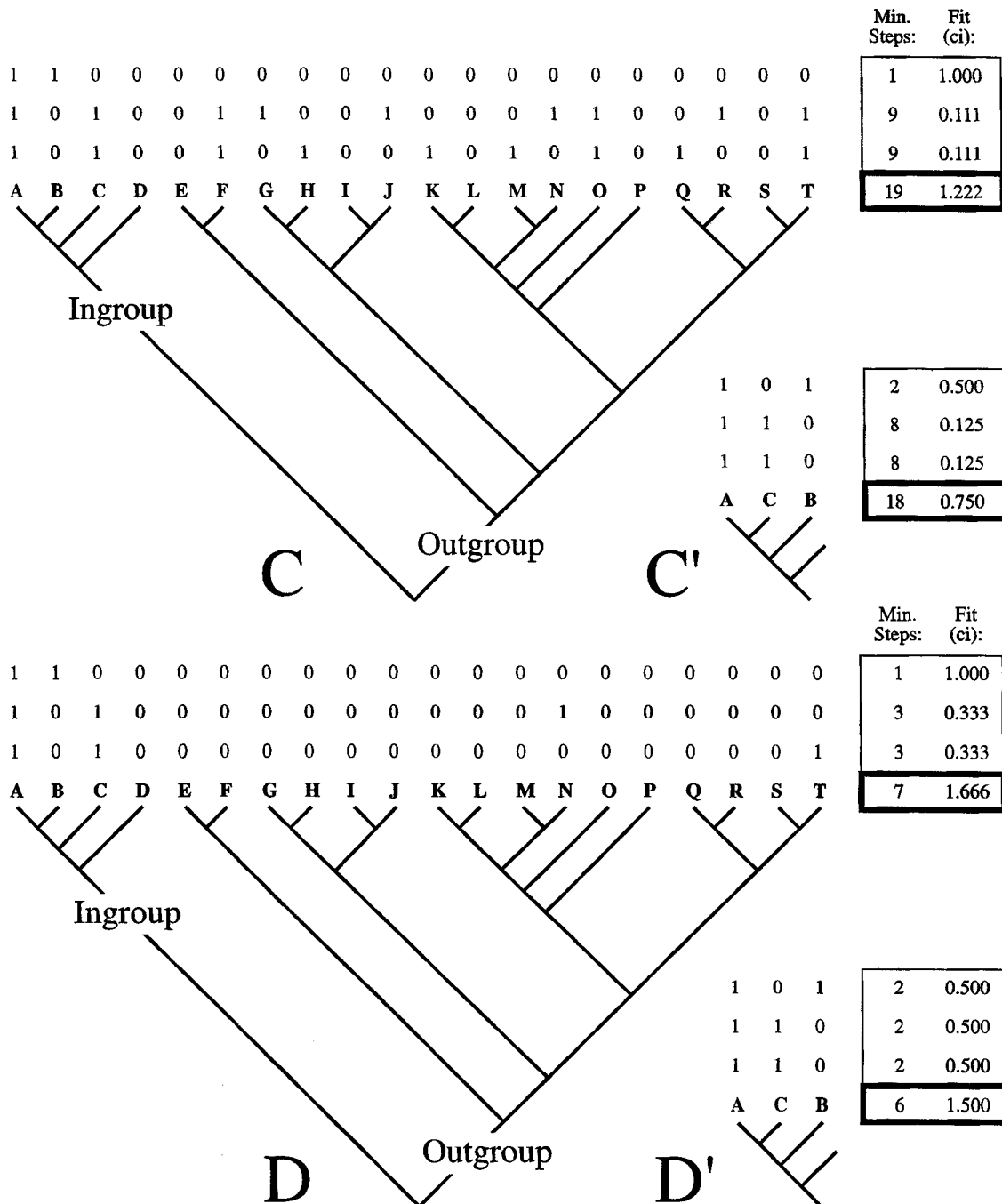
C

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Min. Steps	Fit (ci)
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1.000
1	0	1	0	0	1	1	0	0	1	0	0	0	1	1	0	0	1	0	1	9	0.111
1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	1	0	0	1	9	0.111
																				19	1.222

Ingroup

Outgroup

C'

A	C	B	Min. Steps	Fit (ci)
1	0	1	2	0.500
1	1	0	8	0.125
1	1	0	8	0.125
			18	0.750

D

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Min. Steps	Fit (ci)
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1.000
1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0.333
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0.333
																				7	1.666

Ingroup

Outgroup

D'

A	C	B	Min. Steps	Fit (ci)
1	0	1	2	0.500
1	1	0	2	0.500
1	1	0	2	0.500
			6	1.500

Fig. 3—continued.

have yet to be developed beyond the obvious (e.g., re-examining character states for their similarity and polarity; Kluge 1991). As well, attempts to identify important operations and order them one to another have yet to suggest a convincing basis for claiming increased knowledge in phylogenetic systematics (Neff 1986: figs 1–4). But matters are even worse in verificationist research programs, such as those employing maximum likelihood (Swofford *et al.* 1996), because there is no incentive to critically re-examine the resulting hypothesis, the original evidence, or the assumed process-model of evolution. In fact, the maximum likelihood method itself does not produce diagnoses or particular hypotheses of character transformation which can be used as the basis for character reanalysis. The inductionist agenda of maximum likelihood only seems to encourage the addition of more data of a similar kind,

and the formulation of more complex, assumption laden, models. Surely, this is not the nature of science.

I have attempted to show in this paper that practicing phylogenetic inference as a sophisticated falsificationist research program (Lakatos 1993) provides a foundation for claiming increased knowledge. Research is necessarily cyclic, and the severe testing of phylogenetic hypotheses and evidence leading to a positive shift in understanding actually occurs subsequent to discovering maximally corroborated cladogram(s) (sensu Popper 1992). However, sophisticated falsification and research cycles expose other problems, such as the inconsistency of differential character weighting. To be sure, weighting has had a long history in phylogenetic inference, but its justification has been argued, with few exceptions, in verificationist terms of increased accuracy. Moreover, the vast majority of weigh-

ting methods rely on frequentist assessments of independently evolved states, the more homoplasious a character is hypothesized to be the more it is down-weighted. But, independently evolved states by definition are historically unique, and there is no basis, aside from essentialism, for counting hypothesized instances of homoplasy as if they were the same.

Thus, there is a tradeoff. The student of phylogenetic inference has to decide between real increased knowledge or presumed accuracy. I am inclined to go with sophisticated falsification and research cycles, because of their scientific character, greater coherence and generality, and the expansive nature of the research program they offer.

Acknowledgements

References

Agosti, D., Jacobs, D. & DeSalle, R. 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case.—*Cladistics 12:* 65–82.

Albert, V. A., Mishler, B. D. & Chase, M. W. 1992. Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. In *Molecular systematics of plants* (eds P. S. Soltis, D. E. Soltis & J. J. Doyle): 369–403. Chapman and Hall, London.

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

Brooks, D. R., O'Grady, R. T. & Wiley, E. O. 1986. A measure of the information content of phylogenetic trees, and its use as an optimality criterion.—*Syst. Zool. 35:* 571–581.

Carpenter, J. M. 1988. Choosing among multiple equally parsimonious cladograms.—*Cladistics 4:* 291–296.

Carpenter, J. M. 1994. Successive weighting, reliability and evidence.—*Cladistics 10:* 215–220.

Carpenter, J. M., Strassmann, J. E., Turillazzi, S., Hughes, C. R., Solis, C. R. & Cervo, R. 1993. Phylogenetic relationships among paper wasp social parasites and their hosts (Hymenoptera: Vespidae; Polistinae).—*Cladistics 9:* 129–146.

Darwin, C. 1859. *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London. [1964. Facsimile of the First Edition. Harvard University Press, Cambridge, Massachusetts.

Eernisse, D. J. & Kluge, A. G. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology.—*Mol. Biol. Evol. 19:* 1170–1195.

Farris, J. S. 1966. Estimation of conservatism of characters by constancy within biological populations.—*Evol. 20:* 587–591.

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

Farris, J. S. 1970a. On the relationship between variation and conservatism.—*Evol. 24:* 825–827.

Farris, J. S. 1970b. Methods for computing Wagner Trees.—*Syst. Zool. 19:* 172–189.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. In *Advances in cladistics. Vol. 2. Proceedings of the Second Meeting of the Willi Hennig Society* (eds N. I. Platnick & V. A. Funk): 7–36. Columbia University Press, New York.

Farris, J. S. 1988. *Hennig. Hennig86 reference.* Version 1.5. Privately printed.

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

Farris, J. S. & Kluge, A. G. 1979. A botanical clique.—*Syst. Zool. 28:* 400–411.

Farris, J. S., Kluge, A. G. & Eckhardt, M. J. 1970. A numerical approach to Phylogenetic Systematics.—*Syst. Zool. 19:* 172–189.

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

Gauld, I. & Underwood, G. 1987. Some applications of the Le Quesne compatibility test.—*Biol. J. Linn. Soc. 29:* 191–222.

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:* 91–104.

Gustafsson, M. H. G. & Bremer, K. 1995. Morphology and phylogenetic interrelationships of the Asteraceae, Calyceraceae, Campanulaceae, Goodeniaceae, and related families (Asterales).—*Am. J. Bot. 82:* 250–265.

Hecht, M. K. & Edwards, J. L. 1977. The methodology of phylogenetic inference above the species level. In *Major patterns in vertebrate evolution* (eds M. K. Hecht, P. C. Goody & B. M. Hecht): 3–51. Plenum Press, New York.

Hennig, W. 1966. *Phylogenetic systematics.* University of Illinois Press, Chicago, Illinois.

Jones, T. R., Kluge, A. G. & 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.

Kluge, A. G. 1969. The evolution and geographic origin of the New World *Hemidactylus mabouia-brookii* complex (Gekkonidae, Sauria).—*Misc. Pubs., Mus. Zool., Univ. Mich. 138:* 1–78.

Kluge, A. G. 1976. Phylogenetic relationships in the lizard family Pygopodidae: an evaluation of theory, methods and data.—*Misc. Pubs., Mus. Zool., Univ. Mich. 152:* 1–72.

Kluge, A. G. 1983. Cladistics and the classification of the great apes. In: *New interpretations of ape and human ancestry* (eds R. L. Ciochon & R. S. Corruccini): 151–177. Plenum Press, New York.

Kluge, A. G. 1991. Boine snake phylogeny and research cycles.—*Misc. Pubs., Mus. Zool., Univ. Mich. 178:* 1–58.

Kluge, A. G. 1994. Principles of phylogenetic systematics and the informativeness of the karyotype in documenting gekkotan lizard relationships.—*Herpetol. 50:* 210–221.

Kluge, A. G. 1997. Testability and the refutation and corroboration of cladistic hypotheses.—*Cladistics 13:* 81–96.

Kluge, A. G. 1998. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics 14.*

Kluge, A. G. & Farris, J. S. 1969. Quantitative phyletics and the evolution of anurans.—*Syst. Zool. 18:* 1–32.

Kluge, A. G. & Kerfoot, W. C. 1973. The predictability and regularity of character divergence.—*Am. Nat. 107:* 426–442.

Lakatos, I. 1993. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge* (eds I. Lakatos & A. Musgrave): 91–196. Cambridge University Press, London.

Lauder, G. V. 1982. Historical biology and the problem of design.—*J. Theor. Biol. 97:* 57–67.

Lauder, G. V. 1990. Functional morphology and systematics: studying functional patterns in an historical context.—*Annu. Rev. Ecol. Syst. 21:* 317–340.

Le Quesne, W. J. 1969. A method of selection of characters in numerical taxonomy.—*Syst. Zool. 18:* 201–205.

Le Quesne, W. J. 1972. Further studies based on the uniquely derived character concept.—*Syst. Zool. 21:* 281–288.

Lipscomb, D. L. 1992. Parsimony, homology and the analysis of multistate characters.—*Cladistics 8:* 45–65.

Mayr, E. & Ashlock, P. D. 1991. *Principles of systematic zoology.* Second Edition. MacGraw-Hill, New York.

Mickevich, M. F. 1982. Transformation series analysis.—*Syst. Zool. 31:* 461–478.

Mickevich, M. F. & Lipscomb, D. L. 1991. Parsimony and the choice between different transformations for the same character set.—*Cladistics 7:* 111–139.

Mindell, D. P. 1997. 'Misleading' molecules?—*Sci. 276:* 1629.

Mindell, D. P. & Thacker, C. E. 1996. Rates of molecular evolution: phylogenetic issues and applications.—*Annu. Rev. Ecol. Syst. 27:* 279–303.

Moody, S. M. & O'Nolan, P. 1987. An a priori character weighting method for cladistic analysis.—*Am. Zool. 27:* 108A.

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

Nelson, G. 1983. Reticulation in cladograms. In *Advances in cladistics. Vol. 2. Proceedings of the Second Meeting of the Willi Hennig Society* (eds N. I. Platnick & V. A. Funk): 105–111. Columbia University Press, New York.

Nelson, G. & Platnick, N. 1991. Three-taxon statements: a more precise use of parsimony?—*Cladistics 7:* 351–366.

Popper, K. 1992. *Realism and the aim of science.* Routledge, London.

Sharkey, M. J. 1989. A hypothesis-independent method of character-weighting for cladistic analysis.—*Cladistics 5:* 63–86.

Siddall, M. E. & Kluge, A. G. 1997. Probabilism and phylogenetic inference.—*Cladistics 13:* 336.

Sober, E. 1986. Parsimony and character weighting.—*Cladistics 2:* 28–42.

Sober, E. 1988a. The conceptual relationship of cladistic phylogenetics and vicariance biogeography.—*Syst. Zool. 37:* 245–253.

Sober, E. 1988b. *Reconstructing the past: parsimony, evolution, and inference.* A Bradford Book, MIT Press, Cambridge, Massachusetts.

Sokal, R. R. & Sneath, P. H. A. 1963. *Principles of numerical taxonomy.* W. H. Freeman and Company, San Francisco, California.

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

Turner, H. & Zandee, R. 1995. The behaviour of Goloboff's tree fitness measure F.—*Cladistics 11:* 57–72.

Wheeler, W. 1990. Combinatorial weights in phylogenetic analysis: a statistical parsimony procedure.—*Cladistics 6:* 269–275.

Williams, P. L. & Fitch, W. M. 1989. Finding the minimal change in a given tree. In *The hierarchy of life: molecules and morphology in phylogenetic analysis: proceedings from Nobel Symposium 70 held at Alfred Nobel's Bjorkborn, Karlskoga, Sweden, August 29–September 2, 1988* (eds B. Fernholm, K. Bremer & H. Jornvall): 453–470. Elsevier Press, Amsterdam.

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