Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cladistics 19 (2003) 379–418

Cladistics

www.elsevier.com/locate/yclad

Data exploration in phylogenetic inference: scientific, heuristic, or neither

Taran Grant[a,b,*] and Arnold G. Kluge[c,*]

[a] *Division of Vertebrate Zoology, Herpetology, American Museum of Natural History, New York, NY 10024, USA*
[b] *Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY 10027, USA*
[c] *Museum of Zoology, University of Michigan, Ann Arbor, MI 48109, USA*

Accepted 16 June 2003

Abstract

The methods of data exploration have become the centerpiece of phylogenetic inference, but without the scientific importance of those methods having been identified. We examine in some detail the procedures and justifications of Wheeler's sensitivity analysis and relative rate comparison (saturation analysis). In addition, we review methods designed to explore evidential decisiveness, clade stability, transformation series additivity, methodological concordance, sensitivity to prior probabilities (Bayesian analysis), skewness, computer-intensive tests, long-branch attraction, model assumptions (likelihood ratio test), sensitivity to amount of data, polymorphism, clade concordance index, character compatibility, partitioned analysis, spectral analysis, relative apparent synapomorphy analysis, and congruence with a "known" phylogeny. In our review, we consider a method to be scientific if it performs empirical tests, i.e., if it applies empirical data that could potentially refute the hypothesis of interest. Methods that do not perform tests, and therefore are not scientific, may nonetheless be heuristic in the scientific enterprise if they point to more weakly or ambiguously corroborated hypotheses, such propositions being more easily refuted than those that have been more severely tested and are more strongly corroborated. Based on common usage, data exploration in phylogenetics is accomplished by any method that performs sensitivity or quality analysis. Sensitivity analysis evaluates the responsiveness of results to variation or errors in parameter values and assumptions. Sensitivity analysis is generally interpreted as providing a measure of support, where conclusions that are insensitive (robust, stable) to perturbations are judged to be accurate, probable, or reliable. As an alternative to that verificationist concept, we define support objectively as the degree to which critical evidence refutes competing hypotheses. As such, degree of support is secondary to the scientific optimality criterion of maximizing explanatory power. Quality analyses purport to distinguish good, reliable, accurate data from bad, misleading, erroneous data, thereby assessing the ability of data to indicate the true phylogeny. Only the quality analysis of character compatibility can be judged scientific—and a weak test at that compared to character congruence. Methods judged to be heuristic include Bremer support, long-branch extraction, and safe taxonomic reduction, and we underscore the great heuristic potential of a posteriori analysis of patterns of transformations on the total-evidence cladogram. However, of the more than 20 kinds of data exploration methods evaluated, the vast majority is neither scientific nor heuristic. Given so little demonstrated cognitive worth, we conclude that undue emphasis has been placed on data exploration in phylogenetic inference, and we urge phylogeneticists to consider more carefully the relevance of the methods that they employ.

> [T]he cult of impressive technicalities or the cult of precision may get the better of us, and interfere with our search for clarity, simplicity, and truth [Popper, 1983, p. 60].

> Empirical papers chosen for publication are judged to be of interest to a broad systematics audience because they represent exemplary case studies involving some important contemporary issue or issues. These may be unusually thorough *explorations of data*, applications of new methodology, illustrations of fundamental principles, and/or investigations of interesting evolutionary questions. [Systematic Biology: Instructions for authors, 2002; italics added]

[*] Corresponding authors.
E-mail addresses: grant@amnh.org (T. Grant), akluge@umich.edu (A.G. Kluge).

0748-3007/$ - see front matter © 2003 The Willi Hennig Society. Published by Elsevier Inc. All rights reserved.
doi:10.1016/S0748-3007(03)00074-4

As phylogenetic inference has become increasingly important to comparative biology, methods of data exploration have achieved greater prominence—so much so that empirical phylogenetic studies are judged less than "cutting edge" when data exploration is absent or insufficient and are even denied publication unless they meet the criterion of being unusually thorough explorations of the data. Likewise, data exploration is often considered paramount when evaluating the merits of research proposals, with funding agencies denying support if data exploration is judged deficient. As a result, most empirical phylogenetic investigations now devote considerable resources to the methods and results of data exploration. However, despite its perceived importance, the scientific relevance of data exploration to phylogenetic inference has yet to be identified, and we are concerned that it has achieved the status of a cult of impressive technicalities (see epigraph). The lack of an explicit definition of data exploration suggests that it may serve more as a popular slogan than a scientifically relevant set of procedures.

What is data exploration and what role does it play in the science of phylogenetic inference? Our reason for seeking answers to these questions is that for phylogenetic inference to be scientific it must be logically consistent—methodologically, theoretically, and philosophically. Without that consistency, there can be no logical basis for phylogenetic inference.

Our goal in this paper is to outline a consistent position from which to evaluate the relevance of data exploration methods. The many methods currently in use have vastly different applications and justifications, making it impossible to pick out one or two exemplars to represent the entire field. On the other hand, an exhaustive logical analysis of all available data exploration methods lies beyond the scope of this paper. We have therefore attempted to strike a balance between breadth and depth by dividing this paper into three parts, each standing more or less on its own.

Part I provides theoretical background and summarizes the logical basis for our views on data exploration. **Part II** applies those views in detailed evaluations of Wheeler's (1995) sensitivity analysis and Mindell and Thacker's (1996) relative rate comparisons. We focus initially on these two methods because (1) they are currently among the most detailed and widely cited methods of data exploration and (2) they employ several of the same general assumptions and procedures as other methods of data exploration. **Part III** surveys briefly a broad representation of data exploration methods and their justifications as of 2002. The treatment of each method covered in this section is not exhaustive, and we are aware that this may draw criticism; rather, we aim only to apply our views as generally as possible to illustrate their consistency and to provide a starting point for further, more detailed debate. To facilitate reference to particular methods of interest, the different kinds of data exploration methods that we examine are listed in Table 1, along with the relevant page numbers.

Part I: Theoretical background

Preview of data exploration

An explicit definition of data exploration has yet to be offered in systematics, which has led to the proliferation of a bewildering number and variety of methods that purport to explore phylogenetic data. The lack of an explicit definition also hinders attempts to delimit what is, and what is not, data exploration. Nevertheless, common usage indicates that data exploration is accomplished by any method that performs either *sensitivity analysis*, defined broadly as the investigation of "the responsiveness of conclusions to changes or errors in parameter values and assumptions" (Baird, 1989, p. 358), or *quality analysis*, which purports to distinguish good, reliable data from bad, unreliable data, thereby assessing the ability of data to indicate the true phylogeny. Common usage also implies that methods of discovery such as maximum likelihood, parsimony, and neighbor-joining are not in themselves considered to be methods of data exploration (although application of multiple discovery operations is) nor are reports on the optimality criteria employed by those operations, such as the ensemble consistency (**CI**; Kluge and Farris, 1969) and retention (**RI**; Farris, 1989b) indices. Simulation is not included either, because data are generated from an abstract model, not from observation.

Sensitivity analysis (also known as scenario modeling, stability analysis, tolerance analysis, and deterministic modeling) is important in statistics and decision theory and is a useful tool in applied sciences such as economics, meteorology, engineering, and medicine. Likewise, its value has been demonstrated in the nomothetic sciences of physics and chemistry. However, in our paper we evaluate its use in phylogenetics, which is a strictly ideographic science (Carpenter, 1992; Frost and Kluge, 1994; Wenzel and Carpenter, 1994; Farris, 1995; Kluge, 1997, 2002; Siddall and Kluge, 1997; Wenzel, 1997; Grant, 2002). Sensitivity analysis began in systematics shortly after the advent of numerical methods (e.g., Fisher and Rohlf, 1969; Adams, 1972; Sneath and Sokal, 1973). A partial list of general uses of sensitivity analysis (following the outline provided by Pannell, 1997, p. 246) and corresponding phylogenetic examples are provided in Table 2.

Several different approaches to evaluate sensitivity to assumptions have been employed. For example, Wheeler (1995) examined sensitivity to assumptions of transversion–transition and indel–substitution cost

Table 1
List of data exploration methods assessed in this paper. Approaches that involve especially diverse methods are divided accordingly. See text for details.

Kind of method	Data exploration method(s)	Page(s)
Sensitivity analysis	Wheeler's sensitivity analysis	384, 388
	Decisiveness/ambiguity	388
	Bremer support	
	Double decay	
	Total support	
	Clade stability index	389
	Transformation series additivity	390
	Methodological concordance	391
	Sensitivity to prior probabilities (Bayesian phylogenetic inference)	393
	Skewness test	394
	Computer-intensive sampling	395
	Bootstrap	
	Jackknife	
	PTP	
	T-PTP	
	RT-PTP	
	HER	
	Long-branch attraction	398
	Likelihood ratio test for model selection	398
	Amount of evidence (missing data)	400
	Safe taxonomic reduction	
	Phylogenetic trunk	
	RILD test	
	Multiple regression analysis	
	Polymorphism	402
	Clade concordance index	403
Quality analysis	Relative rate comparison (saturation analysis)	386, 388
	Character compatibility	403
	Spectral analysis	404
	Relative apparent synapomorphy analysis (RASA)	405
	Data partition methods (taxonomic congruence)	406
	Topological incongruence test	
	Global congruence	
	χ^2 test	
	Mickevich–Farris incongruence index	
	Miyamoto incongruence index	
	ILD test	
	Partitioned Bremer support	
	Congruence with an empirically "known" phylogeny	410

ratios with regard to taxonomic congruence (see also Ballard et al., 1998; Barker and Lanyon, 2000; Flores-Villela et al., 2000; McGuire and Bong Heang, 2001). An equivalent implementation of sensitivity analysis is methodological concordance, which assesses robustness to choice of method of phylogenetic analysis by comparing the optimal hypotheses obtained from different phylogenetic discovery operations, such as parsimony, maximum likelihood, and neighbor-joining (e.g., Kim, 1993; Flores-Villela et al., 2000; McGuire and Bong Heang, 2001). Donoghue and Ackerly (1996, p. 1241) proposed "a variety of sensitivity tests to explore the robustness of comparative conclusions to changes in underlying assumptions."

Sensitivity to data has been considered a measure of how decisively a hypothesis is corroborated. By focusing on data, not assumptions, these methods aim to assess the objective support of data for a hypothesis. The most commonly employed sensitivity analyses performing this function are the bootstrap (Felsenstein, 1985b) and jackknife (e.g., Mueller and Ayala, 1982; Lanyon, 1985; Penny and Hendy, 1986; Siddall, 1995; Farris et al., 1996; Farris, 2002b), Monte Carlo routines that assess sensitivity by resampling the data (characters or taxa) at random, thereby creating multiple pseudoreplicates from the same underlying distribution. Another common indicator of the decisiveness of evidence is Bremer support (Bremer, 1988, 1994), which evaluates sensitivity by exploring suboptimal solutions and determining how much worse a solution must be for a hypothesized clade not to be recovered.

Examples of quality analysis include simple exploration of codon position and base composition to inform a priori character weighting (e.g., Chippindale and Wiens,

Table 2
General uses of sensitivity analysis and examples of corresponding procedures in phylogenetic systematics (see text for references)

General use	Phylogenetic example
Testing robustness of an optimal solution	Wheeler's sensitivity analysis Bootstrap Jackknife Bremer support Methodological concordance
Identifying critical values or thresholds where optimal solution changes	Wheeler's sensitivity analysis Bremer support Bootstrap Jackknife Clade stability index Safe taxonomic reduction
Identifying sensitive or important variables	Wheeler's sensitivity analysis Jackknife Clade stability index Long-branch extraction Partitioned Bremer support
Investigating suboptimal solutions	Wheeler's sensitivity analysis Bremer support
Estimating and understanding relationships between input and output variables	Wheeler's sensitivity analysis Long-branch extraction Partitioned Bremer support
Developing hypotheses for testing	All heuristic methods
Testing the model for validity or accuracy	Wheeler's sensitivity analysis Likelihood ratio test
Simplifying and/or calibrating the model	Likelihood ratio test
Coping with poor or missing data	Safe taxonomic reduction
Prioritizing acquisition of information	All heuristic methods

1994). A more technical procedure makes pairwise comparisons between taxa to assess substitution saturation, which, in turn, provides a basis for excluding or down-weighting certain classes of positions or transformations (e.g., Mindell and Thacker, 1996). A number of methods assess congruence among the results of separate analyses of partitioned data sets under the assumption that data sets of high quality will be mutually congruent (e.g., Miyamoto and Fitch, 1995; Huelsenbeck et al., 1996a). In a similar approach, data quality is evaluated on the basis of congruence with a "known" phylogeny (e.g., Naylor and Brown, 1997, 1998; Ballard et al., 1998; Miya and Nishida, 2000). In all these examples of quality analysis, the results of data exploration provide a basis for combining, excluding, differentially weighting, or otherwise manipulating data sets.

Epistemology: test and heurism in science

We adhere to an explicitly objective, realist view of science whereby cognitive progress is achieved by testing competing explanatory hypotheses with empirical evidence. Some biologists have defined a test as simply "a procedure that leads to a choice between hypotheses" by being "coupled with a decision rule" (Sanderson and Wojciechowski, 2000, p. 675), but such vague definitions fail to provide a rational justification for that choice, thus permitting arbitrary and subjective preferences. Although the details of the formalisms involved in effecting tests in principle and practice remain a subject of philosophical debate, it is consensually understood by both philosophers[1] and scientists that a scientific test involves applying empirical data that could potentially refute the hypothesis of interest. In this system, an observation is relevant or evidentially significant only if it has the potential to objectively disconfirm a specified hypothesis, i.e., if it is able to *test* a hypothesis; the greater that potential, the more critical the evidence and, accordingly, the severer the test. The hypothesis that is best able to explain the most objectively critical evidence is preferred rationally as both the most strongly

[1] Although our arguments are made throughout in terms of tests and refutations, verificationism can also be consistent with our position with regard to evidential significance and data exploration. For example, to avoid the paradoxes of confirmation, an observation may be considered relevant or evidentially significant to the extent that it has the potential to objectively alter a probability, and a hypothesis is statistically meaningful only insofar as it is empirically verifiable in this sense (Salmon, 1966, p. 91; von Wright, 1984; Bunge, 1998).

supported knowledge claim and the best proposition for additional testing. That is, "science aims at ever better *explanations*, and. . .choices between competing explanatory theories are controlled by *corroborations*," where "better explanations" are both "deeper and wider" than competitors (Watkins, 1997, p. 5, 9, italics in original).

Contrary to hypothesis choice, which may be explicated solely on the basis of the logic of scientific discovery, the invention or selection of which hypotheses to test or, more generally, which problems to investigate is nonscientific, often relying on idiosyncratic preferences and noncognitive concerns such as availability of funding or other social pressures (e.g., Kuhn, 1962, 1977; Fuller, 1993; Braun, 1998; Resnik, 2001). Nevertheless, problem selection may also be guided by cognitive considerations (Lakatos, 1978). Kluge (1997) suggested that long-held, highly corroborated hypotheses may be of special interest because their empirical content may be more simply and clearly described, making them more easily tested. Alternatively, problem selection can be based on degree of corroboration by focusing on bold, highly improbable hypotheses that have never been tested or are only weakly or ambiguously corroborated. Such hypotheses may be more easily disconfirmed than those that have been severely tested and are more strongly corroborated, and any procedure that identifies such hypotheses provides a heuristic shortcut to refutation and increased knowledge. This approach to problem selection is analogous to the heuristic strategies commonly employed in tree searching (e.g., Goloboff, 1999), where all possible hypotheses of phylogeny are worthy of testing, but algorithmic shortcuts direct attention toward the subset of hypotheses that are most likely to be fruitful (for a general discussion see Nickles, 2000).

Pursuit of problems on the basis of degree of corroboration is defensible only insofar as the emphasis is placed on weakly corroborated hypotheses being more easily refuted and not on strongly corroborated hypotheses being more accurate or certain or less worthy of testing, as such a verificationist perspective would be contrary to the necessarily critical nature of science. Moreover, even though the approach to problem selection may be so rationalized, this does not mean that testing has actually been achieved or that the approach is scientific.

Sensitivity and support

In systematics, sensitivity analysis is generally interpreted as providing a measure of support, where results that are insensitive (robust, stable) are considered well supported. Support, in turn, is almost universally taken to mean certainty, confidence, probability, or reliability (e.g., Farris, 1969, 1998, 2002b; Felsenstein, 1985b, 1988; Carpenter, 1988, 1994; Hillis and Huelsenbeck, 1992; Steel et al., 1993b; Brown, 1994; Sanderson, 1995; Wheeler, 1995; Donoghue and Ackerly, 1996; Efron et al., 1996; Buckley and Cunningham, 2002; Siddall, 2002a, p. 96; *contra* Siddall, 2002a, p. 96). Given such a verificationist interpretation, it would seem that the concept of support could play no role whatsoever in the science of phylogenetic inference. Even within the verificationist framework, the statistical reliability of an inductive generalization can be inferred only under the assumption that available data are representative of the universe of data, such as when data are drawn from a population at random, but this assumption is counterfactual in phylogenetic analysis (see Computer-intensive sampling, below).

Farris et al. (1996, 109; see also Källersjö and Farris, 1998) rejected the verificationist interpretation of support and offered an alternative interpretation, considering parsimony jackknifing "simply as a way of discovering ambiguities in data." We agree, and we submit that the concept of support can be salvaged generally as an indicator of evidential ambiguity and a report on the decisiveness of tests, where support is defined objectively as **the degree to which critical evidence refutes competing hypotheses**. A hypothesis is unsupported if it is either (1) decisively refuted by the critical evidence or (2) contradicted by other, equally optimal hypotheses (i.e., evidence is ambiguous, such as when multiple most-parsimonious cladograms obtain); otherwise it is supported. That is, rational hypothesis preference is based on the *relative* degree of corroboration of competing hypotheses, where the hypothesis that is least refuted by critical evidence is preferred (Popper, 1959). A hypothesis is therefore supported if the critical evidence confers a greater degree of corroboration on it than on any competing hypothesis, even if the *absolute* degree of corroboration of the optimal hypothesis is disturbingly low (Lakatos, 1978), such as when the most-parsimonious cladogram has a low **CI** (Farris, 1983).

Under this concept of support, there can be no basis for preferring a less parsimonious hypothesis of species relationships (Farris, 1983), nor is there any basis for attributing more confidence or reliability to more strongly supported clades. Most corroborated hypotheses are preferred, "if only from a theoretical point of view which makes them *theoretically most interesting objects for further tests*" (Popper, 1979, p. 13, italics in original). What matters scientifically is that the evidence supports a hypothesis, not the degree of support, which is why the strict consensus of most-parsimonious cladograms is especially beneficial as a summary of universally corroborated groups. Weakly supported groups are still supported by the evidence, and we see no *epistemological* reason to exclude them (but see Farris, 1998; see also below).

Our concept of support is heuristic in that it identifies cases in which refutation of competing hypotheses is weak, because weakly corroborated hypotheses may be more easily refuted than those that are more strongly corroborated. Rather than underscoring strongly supported clades by indicating them with asterisks and arrows, providing detailed discussions, and formalizing them with special taxonomic ranks, we believe that science would be better served by focusing on weakly supported clades and the potential means of more severely testing them (Kluge, 1997). Furthermore, epistemologically, all clades in the least refuted cladogram(s) provide an equally valid basis for testing evolutionary scenarios (e.g., biogeographic hypotheses) and designing future phylogenetic studies (e.g., selecting outgroups), and although it may be tempting to base such studies exclusively on the more strongly supported parts of a given cladogram, that procedure is a slippery slope to a verificationist interpretation of better-supported clades as more reliable or certain or closer to truth.

Interpretation and justification of methods

We judge a method to be scientific if and only if it involves an empirical test. For example, the unweighted parsimony method employed in phylogenetic systematics is scientific because propositions of relative recency of common ancestry are tested according to the congruence/incongruence of the available character evidence. Moreover, in a progressive research program such as phylogenetic systematics, the results of tests always point to new problems and other testable hypotheses (Lakatos, 1978; Kluge, 1997, 1998, 1999). In this regard, phylogenetic systematics can be considered ampliative.

Those methods that do not provide a valid test and therefore are not scientific may nonetheless be heuristic, but if and only if they are guided by the cognitive considerations of evidential ambiguity and decisiveness of tests, as outlined above. Obviously relevant to the concept of heurism, the method must point to strategies for further testing or to testable hypotheses. In other words, scientific objectivity must be evident for a method to be heuristic. From this it follows that a heuristic method cannot protect a hypothesis from being refuted. For example, the auxiliary assumptions—the major premise(s) in causal explanation, such as background knowledge—must not diminish testability. Indeed, those kinds of included assumptions are admissible only if they *increase* testability. Likewise, methods that are not protected from ad hoc hypotheses cannot be relevant, nor can those leading to tautology (e.g., Faith and Trueman, 2001).

Part II: Detailed evaluations

Data exploration as a test

Many of the methods of data exploration have been applied as a kind of optimality criterion in choosing among competing auxiliary assumptions and hypotheses. For this to be valid, a clear relationship between results of data exploration and hypothesis testing must exist. Wheeler's (1995) sensitivity analysis and Mindell and Thacker's (1996) relative rate comparisons are evaluated for this relationship between exploration and testing.

Wheeler's sensitivity analysis as a test. Numerous authors have employed Wheeler's sensitivity analysis as a test of relationships or auxiliary assumptions (Wheeler, 1995, 1999; Allard and Carpenter, 1996; Wheeler and Hayashi, 1998; Giribet and Wheeler, 1999; O'Leary, 1999; Phillips et al., 2000; Frost et al., 2001a; Janies, 2001; McGuire and Bong Heang, 2001; Wheeler et al., 2001; Giribet et al., 2000, 2001, 2002).

As a test of relationships, insensitivity of groups to variation in the relative weights assigned to transitions, transversions, and insertion–deletion events (indels) is treated as an optimality criterion to decide whether to reject or accept a hypothesis of monophyly (Wheeler, 1995). Groups that are more sensitive to variation in cost ratios are rejected, whereas those that are more robust to different ratios are accepted. Group robustness may be represented graphically as a sensitivity plot (binary Cartesian graph) or a consensus of clades recovered under an arbitrary number of parameter sets (e.g., Wheeler, 1995; Giribet and Wheeler, 1999; Wheeler et al., 2001).

To justify this procedure, Wheeler (1995, p. 328; italics added) argued:

> If a high fraction of the *total analysis space* supports a group, the group is generally *supported by the data* because most combinations of analytical parameters will yield that clade...

However, leaving aside for the moment (see below) the unjustifiable practice of specifying a priori classes of phylogenetic evidence (Kluge and Wolf, 1993, p. 190; see also Allard et al., 1999), robustness to variation in weights of classes of data does not quantify evidential support. For example, an uncontradicted clade corroborated by a large number of transitions is certainly strongly supported by the data, given that any other hypothesis of relationships would entail extensive incongruence, yet that group would disappear in transversion parsimony (sensu Swofford et al., 1996, p. 422), giving the impression that it lacks support. Likewise, a group corroborated by many transversions and contradicted by only a few indels and transitions is also strongly supported by the data, but that group would disappear in any weighting scheme that were to upweight transitions and/or indels. Stability under a range of cost ratios indicates only that evidence for a clade does not derive from a single synapomorphy class; it provides no indication of the actual amount of evidence that supports a group. Consequently, an uncontradicted group corroborated by 15 transitions, 15 indels, and

15 transversions would seem no more supported than one corroborated by 1 transition, 1 indel, and 1 transversion. Clearly, this kind of sensitivity analysis does not evaluate support by the data, so that justification fails to provide a basis for using this approach to decide whether to accept or reject a hypothesis of monophyly.

Alternatively, using this approach to sensitivity analysis as a test of relationships has been justified as measuring support by assumptions. Wheeler et al. (2001, p. 139) argued that "[s]upport for all groups is dependent on the analytical assumptions we make," and they went on to interpret stability to auxiliary assumptions as a measure of clade support distinct from evidential (character) support (which they measured with Bremer support). Giribet et al. (2000, p. 547; see also Giribet et al., 2002, p. 16) similarly explained that they "considered this an effective way to explore the data and discern between well-supported relationships (those supported through a wide range of parameters) and poorly supported relationships (those that appear only with very particular parameter sets)," and Giribet and Wheeler (2002, p. 288) explicitly considered groups stable to different assumptions to be "well corroborated." However, this argument misplaces the formal role of auxiliary assumptions in hypothesis testing: a hypothesis is corroborated *by* empirical evidence *in light of* auxiliary assumptions, the critical issue being the *validity* of auxiliary assumptions, not their differential effects on the outcome of an analysis. That is, auxiliary assumptions provide the background knowledge necessary to perform a valid test and in turn assess degree of corroboration, but they do not themselves corroborate or refute a hypothesis.

The futility of this approach is further underscored by the fact that it views all weighting schemes that do not violate the triangle inequality as equally plausible a priori (Wheeler, 1995; W.C. Wheeler, pers. comm.) and provides no justification for weighting only the character classes of transitions, transversions, and indels or for choosing the few cost ratios evaluated. This argument is relevant because for all but the simplest data sets every possible cladogram may be supported under some set of relative weights (Kluge, 1998), so unless some nonarbitrary boundary is placed on permissible character classes and cost ratios, a rigorous application of this approach would find that all cladograms are unsupported; finding that a clade is supported would simply mean that the sampling of weighting schemes was not sufficiently exhaustive.

Instead of inferring support for relationships, Wheeler (1995) also argued that sensitivity analysis provides a test of auxiliary assumptions. Still believing transversion–transition and indel–substitution cost ratios to be otherwise arbitrary, he treated congruence among data partitions or sets (as judged by Mickevich and Farris's (1981) original measure of taxonomic incongruence, the $\mathbf{i}_{\mathbf{MF}}$ of Kluge (1989, Table 3), or its rescaled form (Wheeler and Hayashi, 1998)) as an optimality criterion, where "the set of values for the transversion–transition ratio and gap–change ratio that maximize congruence would be chosen" (Wheeler, 1995, p. 323).

Wheeler (1995, p. 321; see also Phillips et al., 2000, p. 327) defended this position on the grounds that it increases precision, claiming that

> Without any way of objectively measuring the accuracy of reconstruction, only precision (the agreement among data) can be used to arbitrate among competing hypotheses.

The logical basis of this assertion is indisputable. Precision, thus defined, is clearly related to explanatory power and testability, and it logically translates into the test of congruence/incongruence in phylogenetic inference. Accordingly, Giribet et al. (2002, p. 17; see also Giribet et al., 2000, p. 548) stated that

> Character congruence is thus used as the criterion to choose the best (most corroborated) tree, the tree that minimizes overall character conflict among the data.

However, the belief that these arguments justify minimization of incongruence among partitions is incorrect and stems from employing the $\mathbf{i}_{\mathbf{MF}}$—a character-based measure of *taxonomic* congruence[2]—as a measure of *character* congruence (e.g., Wheeler, 1995, p. 321, 323; Giribet et al., 2002, p. 17). Even though all data are included, this approach employs a kind of taxonomic congruence, not total evidence (Kluge, 1997), because (1) it weights data differentially with regard to partitions (transitions, transversions, and indels) and (2) it evaluates fit with regard to congruence among partitions (e.g., morphology and DNA sequences). Precision is actually maximized by minimizing incongruence among independent data (characters), not sets of data (partitions). Consequently, precision provides an argument for equal weighting of all data because any weighting scheme that favors a different topology necessarily increases the number of events required to explain the data (or the number of bits required to describe them), making

[2] Since its inception, *taxonomic congruence* has referred to congruence among partitions (sets) of characters (e.g., Sokal and Sneath, 1963, pp. 85–86; Farris, 1971; Sneath and Sokal, 1973, p. 97; Rohlf, 1974; Mickevich and Johnson, 1976; Mickevich, 1978; Rohlf and Sokal, 1980; Mickevich and Farris, 1981), and may be evaluated by topology-based measures (e.g., Farris, 1967, 1969, 1973a; Mickevich, 1978; Nelson, 1979; Colless, 1980; Mickevich and Farris, 1981; Rohlf, 1982; Wheeler, 1995, 1999; Levasseur and Lapointe, 2001) and character-based measures (e.g., Rohlf, 1963; Throckmorton, 1968; Farris et al., 1970; Sneath and Sokal, 1973; Mickevich and Farris, 1981; Miyamoto et al., 1994). Terminology obscures this fact and has led some authors to view taxonomic congruence and topological congruence as coextensive (e.g., Chavarría and Carpenter, 1994, p. 243; Allard and Carpenter, 1996; Dolphin et al., 2000; see also Barker and Lutzoni, 2002).

the preferred hypothesis less precise, less parsimonious, and less corroborated (Kluge, 1998, p. 356; Allard et al., 1999). That is, equally weighted parsimony maximizes congruence over *all* data by minimizing the *total* number of hypothesized transformations, whereas differential weighting preferentially maximizes congruence among higher-weighted characters at the expense of congruence among lower-weighted characters. Although the effect is obscured by the fact that differently weighted cladogram costs are not directly comparable, differential weighting leads ultimately to a less efficient overall distribution of states and a greater number of hypothesized transformations (i.e., increased incongruence). Therefore, even though differential weighting may improve the $\mathbf{i}_{\mathbf{MF}}$ score, the apparent increase in character congruence is fallacious, and in fact character congruence has decreased. Given that "no degree of abundance of homoplasy is by itself sufficient to defend choice of a less parsimonious genealogy over a more parsimonious one" (Farris, 1983, p. 14), there can be no justification for differential weighting.

Giribet et al. (2000, p. 548; see also Giribet et al., 2002, p. 17) also argued that minimizing incongruence among partitions is

> ...understood as an extension of parsimony (or any other minimizing criteria); in the same sense that parsimony tries to minimize the number of overall steps in a tree, the "character [sic, = taxonomic] congruence analysis" tries to find the model that minimizes incongruence for all the data sources.

However, the epistemological justification for parsimony as minimizing hypotheses of transformation and maximizing explanatory power (Farris, 1983) does not extend logically to congruence among data partitions (sources) because minimization of incongruence among partitions through differential weighting may actually increase character incongruence, thereby increasing the number transformations required to explain the observations and reducing explanatory power.

The arguments for Wheeler's (1995) sensitivity analysis are predicated entirely on the belief that choice of transversion–transition and indel–substitution cost ratios is externally arbitrary and that even "[s]imple homogeneous weighting does not avoid the issue of arbitrary, yet crucial assumptions" (Wheeler, 1995, p. 321; see also Mindell and Thacker, 1996). That sentiment was echoed more recently by Geiger (2002, p. 192), who asserted that "all DNA sequence alignment is inherently subjective." In addition to the problem that differential weighting entails nonindependence of tests (Siddall and Kluge, 1997; Kluge, 1998; Siddall, 2002b), claims of arbitrariness or subjectivity overlook the fact that phylogenetic inference is historical and that historicity places a nonarbitrary, objective constraint on phylogenetic discovery operations (Siddall and Kluge, 1997; Kluge, 1998, 2002; Grant, 2002). Differential weighting of classes of transformations—including indel transformations (e.g., Wheeler, 1996)—relies ultimately on frequentist probability arguments (Kluge, 1998, 2002). Yet, these arguments are relevant only in nomothetic sciences, where discovery operations must contend with the objective indeterminism of the future. Because phylogenetic inference is strictly historical, and history is objectively determinate[3] (i.e., it has already happened, making it fixed), there can be no objective, frequency-based probability relating to the *necessarily* unique phylogenetic events of the past (Popper, 1990; Siddall and Kluge, 1997; Grant, 2002; Kluge, 2002). That is, even if the overall cost ratios were known for the different classes of molecular transformations, they would be uninformative of whether a particular transformation occurred. Moreover, the fact that history has already happened means that historical scientists can search for evidence of past events. Observed evidence may be false for a number of reasons, including observer error, multiple events having left the same kind of marks interpreted as evidence, or some sort of information-destroying process (Sober, 1988), but, unlike nomothetic scientists, ideographic scientists may base inferences on tests of each piece of evidence against all other evidence simultaneously, thereby maximizing severity of test and detection of errors. As noted above, equal weighting of all evidence in a total-evidence analysis maximizes precision and provides the severest test (see also Kluge, 1997, 1998; Allard et al., 1999; Frost et al., 2001a,b). Transformation cost ratios in phylogenetic inference are therefore *nonarbitrary*, equal weighting of all transformations being the only objectively defensible parameter set.[4]

Relative rate comparison (saturation analysis) as a test. Separately evolved, homoplasious character states do not, by definition, identify the same historical entity (e.g., Hennig, 1966, p. 89). Further, there is the potential for saturation of gene sequence data with such changes because of the small number of possible character states and the potential for multiple substitutions at any nucleotide site in the sequence. More specifically, given the four possible nucleotide states and only a modest rate of change occurring at random among sites, identical independently evolved states are expected at any site.

[3] That history is predetermined with respect to the present should not be confused with historical determinism (historicism), the inevitability of a certain general sequence of events or the existence of laws of history. To the contrary, it is the frequentist/probabilistic approach to historical inference that we oppose that entails historical determinism (Siddall, 2002b; Kluge, 2002).

[4] Note that this does not preclude the incorporation of character state additivity into cladistic analysis. Although additivity is operationalized as differential transformation costs (e.g., given the transformation series $0 \leftrightarrow 1 \leftrightarrow 2$, a "transformation" from 0 to 2 implies a cost of 2 steps), the extra cost is defensible on the grounds that it corresponds to inferred transformations, each of cost 1.

Saturation is assumed to have occurred for a given part of the genome when differences among taxa are less than expected. Molecular biologists (e.g., Brown et al., 1982) have long claimed saturation in those observations that show transitions to occur more frequently than transversions, a bias that arguably must be corrected to provide "*reliable estimation* of sequence distance and phylogeny reconstruction" (Yang and Yoder, 1999, p. 274; italics added).

In general practice, some form of pairwise taxon comparison is used to assess the level of saturation. The differences between pairs of taxa are plotted according to the relative age of origin of the pairs, often measured with regard to taxonomic rank. The comparisons are usually referred to as a "test" when the saturation curves obtained for different kinds of nucleotide substitutions are compared—thus, the term "relative rate test." It is assumed that the comparisons in the saturated, tapering-off portion of a given curve provide inaccurate phylogenetic estimates due to multiple substitutions and randomization of observed states, whereas those compared from the positively increasing portion of the curve do not. The saturated, randomized portion is assumed to reflect excessive homoplasy, which is taken to be evidence for down-weighting such a class of data to improve the reliability of the phylogenetic estimate.

Relative rate tests are employed commonly in assessing codon and transversion–transition bias (e.g., Mindell and Thacker, 1996; Wakeley, 1996), not withstanding the unjustifiable practice of specifying a priori classes of evidence (Kluge and Wolf, 1993, p. 190; see also Allard et al., 1999). Generalizing, the method consists of the following steps: the amount of evolutionary change (discrete or continuous character change) that terminal taxa (A,B) exhibit relative to that of their most recent common ancestor (Y) or to that of an outgroup lineage (C) is determined (Mindell and Thacker, 1996, p. 281). Thus, character change in A:Y can be contrasted to that in B:Y, assuming additivity of character state change, or character change in A:C can be contrasted to that in B:C. In either case, an equal rate of evolution is assumed when those two sets of numbers are the same. Such relative rate comparisons have also been used to test the evolutionary clock hypothesis (Mindell and Thacker, 1996).

The advantage of the relative rate comparisons described above is supposed to be due to the equal amounts of time that have passed (by definition) between sister species and their common ancestor. However, in that advantage is the very undoing of such comparisons and the very concept of relative rate "test." Relative rate comparisons involving patristic or path-length distances, such as A:C, B:C, cannot avoid the criticism of nonindependence, i.e., the amount of evolution exhibited by the nonterminal intervals, between C and the common ancestor of A and B, is redundant. For example, to use traditional statistics, such as the binomial distribution (Mindell and Honeycutt, 1990; Mindell and Thacker, 1996), to test the null hypothesis (i.e., departure from an expected 50% of all distance change between any two sister species) is invalid due to failure to meet the assumption of independence. Relative rate comparisons involving steps, such as A:Y, B:Y, can avoid the issue of nonindependence by recording only those pairs of sister species that are exclusive of one another. However, with such exclusivity comes both a loss of statistical power and the need for a prior phylogenetic hypothesis, a pattern that can then be used to justify what is, and what is not, an exclusive comparison.

Additionally, recent maximum likelihood studies (e.g., Yang and Yoder, 1999) indicated that relative rate comparisons are biased by taxonomic sampling and, by extension, the density of the taxonomic samples relative to parts of phylogeny. They also indicated, as have most other studies of relative rate, that transversion–transition rate ratios vary among different parts of phylogeny. Thus, relative rate comparisons, as tests of codon bias or the molecular clock, are unjustified as tests and cannot be rationally claimed as bases for a priori character weighting.

Data exploration as a heuristic

In a progressive research program, the results of past tests inform future problem choice and test design by indicating heuristically which areas of knowledge are especially worthy of further inquiry (Lakatos, 1978). That is, although all empirical problems remain open to further investigation, prior experience can indicate which problems are most likely to be scientifically fruitful. An example of this progressiveness in phylogenetic testing is Hennig's (1966) reciprocal illumination or clarification, whereby finding that some synapomorphies are incongruent with the most-parsimonious hypothesis heuristically suggests that the initial hypothesis of their homology was incorrect. This, in turn, indicates the need for additional, independent testing (e.g., character reanalysis) and, ultimately, the possibility of eliminating error in the identification of synapomorphy (Kluge, 1998; contra Mindell, 1991). All synapomorphies are worthy of further investigation, but incongruent ones are especially interesting scientifically because the balance of the evidence refutes them, and they are critical in phylogenetic systematics because the optimal, most-parsimonious hypothesis relies on the independence of instances of incongruence (Farris, 1983, 1995; Carpenter, 1992; Kluge, 1997). As this example illustrates, genuine empirical tests may be both scientific and heuristic (for elaboration on this point, see Gattei, 2002).

Likewise, operations that do not perform empirical tests themselves may still be useful tools if they point to highly testable hypotheses, and we suggest that data exploration methods that cannot be defended as tests may still serve this function. The sensitivity analysis of Wheeler (1995) and the relative rate comparisons of Mindell and Thacker (1996) are technically impressive but do not constitute valid scientific tests. Still, there remains the possibility that their applications may be heuristic, and we now turn to judging them in that sense.

Wheeler's sensitivity analysis as a heuristic. For Wheeler's (1995) sensitivity analysis to be heuristically useful, it would have to point out ambiguously corroborated hypotheses, and sensitivity to parameter variation has been interpreted as an indicator of evidential ambiguity. For example, Giribet et al. (2000, p. 557) summarized their sensitivity analyses as "discerning among well-corroborated versus unstable hypotheses of relationships," concluding that for highly parameter-sensitive clades "inferences based on the currently available data are, at least, poorly supported." Likewise, Janies (2001, p. 1247) interpreted sensitivity analysis as having "pinpointed areas of weakness in our understanding of echinoderm relationships" by identifying groups for which "the available character evidence is equivocal." That is, rather than rejecting a hypothesis of monophyly based on parameter sensitivity, this interpretation points to relatively unstable groups as more ambiguously corroborated and poorly supported than groups that are more robust to parameter variation. However, as discussed above, this approach to sensitivity analysis does not evaluate the amount of evidential support and is therefore unable to identify those hypotheses that are weakly or ambiguously corroborated. As such, it is not heuristic.

Relative rate comparison as a heuristic. That relative rate comparisons may be heuristic is also without foundation because that methodology assumes that homoplasy, as measured by incongruence, *necessarily* misinforms the inference of phylogeny. If nucleotide evolution were to occur at random then it might be misinformative, uninformative, or even informative. As Wenzel and Siddall (1999) showed in simulations, half the characters in an analysis must be random for there to be a greater than even chance of overwhelming even a single unique and unreversed synapomorphy. This should come as no surprise, considering that, given even a moderate number of randomized sites, there are so many ways to arrange the four possible nucleotide states among taxa that the chance of forming a pattern such that historically relevant data are contravened is extremely low (Wenzel and Siddall, 1999).

Empirical findings also argue against relative rate comparisons being heuristic. That third codon positions confound phylogenetic inference because they are relatively more homoplasious than first or second codon positions has become conventional wisdom, in large part due to examining saturation curves, and accordingly those transformations are often down-weighted a priori. However, Källersjö et al. (1999) showed, for a 1428-base plastome gene recorded on more than 2500 green plant species, that third codon positions, while relatively more homoplasious than first or second position states, were nonetheless phylogenetically more informative (they had a higher mean character retention index, **ri**) than first or second position states. Such a finding demonstrates that frequency weighting cannot be presumed generally, the problems of nonindependence in relative rate comparisons and saturation plots being beside the point (*contra* Mindell and Thacker, 1996). Additionally, the variable constraints on mutation and fixation rates summarized by Mindell and Thacker (1996, Table 1) would appear to have had no effect on the results reported by Källersjö et al. (1999; see, however, Farris, 2002b, Table 3).

The relevance of these simulation and empirical studies lies in their ability to illustrate the extent to which relative rate comparison tests can misinform phylogenetic inference. Evidence is the primary concern. Any independently evolved synapomorphy is evidentially significant, and it is only by including all available evidence in a simultaneous test that severity of test is maximized (Kluge, 1997, 1998). If a priori weights are based on distributional values (e.g., base compositions, transversion–transition ratios) across all characters or across a character class (e.g., third positions) then the independence of these potential homologues is in fact lost (Kluge, 1998, p. 357; Siddall and Kluge, 1997). The end result of such data purification procedures is a violation of independence and a negative impact on severity of test. There can be no heurism in this.

Part III: Broad survey of methods

Our evaluations above of one kind of sensitivity analysis and one kind of quality analysis exemplify the critical arguments that we believe can be used to evaluate any method of data exploration for its scientific and heuristic merits. We now turn our attention to somewhat briefer assessments of additional data exploration methods. Each of these assessments includes a short description and answers to the following questions: Does the method entail an empirical test? If not, is it heuristic?

Other kinds of sensitivity analysis

Decisiveness/ambiguity. One of the aims of sensitivity analysis is to assess evidential decisiveness (or its converse, ambiguity), defined as the degree to which an optimal solution is preferred over alternatives. As noted by Källersjö et al. (1992, p. 283), "[a]mbiguity is usually detected by finding multiple most parsimonious trees, the unambiguous part of the structure (that common to the several trees) being recovered as a consensus tree.'' Although that is a direct and objective assessment of evidential ambiguity, as Farris et al. (2001) recently discussed in the context of branch lengths, it is not uncommon for a great deal of evidence to favor a particular hypothesis, but for that preference to be extremely weak. Therefore, additional approaches have been sought to assess relative evidential decisiveness/ambiguity in more detail. A low degree of decisiveness cannot alter preference for the most-parsimonious hypothesis (Farris, 1983), but it does give reason to suspect that the preferred hypothesis may be easily disconfirmed with additional testing.

Bremer support (branch support; Bremer, 1988, 1994) assesses the decisiveness of corroboration of a given clade by comparing the length of the most-parsimonious cladogram(s) with those of suboptimal solutions to determine how much worse (i.e., longer) a topology must be for that clade to be absent. Gatesy et al. (1999) attributed the basic index of character support to Templeton (1983; see also Prager and Wilson, 1988). For a particular data set, a particular clade, and a particular character, character support is just the minimum number of steps for that character on the shortest cladogram(s) that does not contain that clade, minus the minimum number of steps for that character on the shortest cladogram(s) that does contain that clade. Källersjö et al. (1992) calculated ''total support'' as the decisiveness of a data matrix for a cladogram by summing the Bremer support values for all nodes. Bremer (1994) rescaled total support by dividing by the greatest possible sum of Bremer supports, the sum of branch lengths.

Wilkinson et al. (2000, p. 757) argued that, because Bremer support focuses only on the clades common to all most-parsimonious cladogram(s), it

> . . .is unable to distinguish cases in which instability in trees is associated with a few terminals and most relationships are otherwise well supported, from cases in which instability and lack of support for relationships are more ubiquitous.

The solution that they proposed, called double decay, is to evaluate the Bremer support for all groups (*n*-taxon statements of Wilkinson, 1994) present in *any* of the most-parsimonious cladograms, not just those groups present in the strict consensus. As a result, all groups with a positive Bremer value are reported, including mutually incompatible groups, allowing highly ambiguously placed terminals to be identified.

None of these methods of data exploration tests phylogenetic hypotheses, but they are explicit indicators of ambiguity of evidential support, making them heuristically useful in deciding which problems to pursue next. Bremer support indicates directly the nodes for which evidence is ambiguous and therefore which hypotheses could be most easily refuted. Double decay analysis takes this a step further by identifying the terminals that are most responsible for ambiguity, which allows them to be targeted specifically for further study in future rounds of testing. By focusing on ambiguously corroborated groups and the synapomorphies that delimit and contradict them, character reanalysis is facilitated, as is the discovery of new synapomorphies relevant to testing their placement. Because total support does not identify the more weakly corroborated portions of the overall hypothesis, it does not point to particular tests and is therefore not heuristic.

Clade stability index. Given a small data set with little incongruence, determining how many and which synapomorphies are crucial to clade delimitation is trivial. However, as matrix size and character conflict increase, it becomes more difficult to assess the relationship between characters and clades. The complexity of character interactions is underscored by considering that clade resolution may be crucially dependent on both synapomorphies of that clade and synapomorphies of other clades (Davis et al., 1993). To identify the characters and character combinations crucial for the recovery of a clade, Davis (1993; see also Davis et al., 1993) proposed to sequentially remove characters and sets of characters and record the presence or absence of each clade. Davis (1993, p. 201) defined the clade stability index (**CSI**) as ''the minimum number of characters that, when removed, cause resolution of the clade to be lost,'' where a clade is considered lost if it is absent from the strict consensus. **CSI** is measured as the ratio of the minimum number of informative characters removed to the total number of informative characters, giving it a range from 0 (when a clade is absent prior to character removal) to 1 (when a clade is lost only when all characters are removed). Gatesy et al. (1999) referred to the unscaled version of **CSI** as the character removal index, **CRI**.

Implementation of **CSI** is hindered by the computational difficulty of the problem. The number of character combinations to be tested in an exhaustive analysis is given by $c = n!/r!(n-r)!$ for n informative characters and r characters to be removed in a given round of analysis. Given that a full parsimony search must be performed for each combination, evaluation of all character combinations is impractical for even moderate-sized matrices (e.g., for a matrix of only 20 informative characters, evaluation of exhaustive character combinations would require 1,048,575 cladistic analyses). Consequently, for removal of more than 2 characters, Davis (1993) estimated **CSI** by examining 500 random character combinations. Like other methods of data exploration that rely on heuristic search strategies (e.g., Bremer support), the reported (observed) **CSI** is equal to or greater than the true **CSI**.

CSI does not test phylogenetic hypotheses, and the stability that it identifies is not justifiable epistemologically. However, by revealing the proportion of characters that is crucial to clade resolution, **CSI** is heuristic. Additional testing may more easily refute clades that are dependent on fewer characters than clades that are supported by more characters. The heurism of this approach is augmented by determining precisely which character or character combination(s) is critical (see Davis et al., 1993), as this allows those characters to be targeted specifically in future rounds of character analysis/reanalysis. A significant drawback is that **CSI** does not distinguish between congruent and incongruent synapomorphies. In this regard simple examination of incongruent characters on the most-parsimonious cladogram is heuristically more effective. The union of the two procedures would increase heurism by pointing to the incongruent synapomorphies that are most crucial to clade resolution.

Transformation series additivity. Transformation series additivity refers to the hypothesized path of evolution of a multistate character. During rounds of separate testing (character analysis/reanalysis), independent evidence is brought to bear in an attempt to choose among the competing hypotheses of character evolution. Successful refutation may result in a defensible preference for one or several (i.e., evidence may be partially ambiguous) of the possible hypotheses of additivity. In either case, the least refuted hypothesis(es) of transformation series additivity is then employed as an auxiliary assumption in the simultaneous test of character congruence. As with binary characters, incongruence with the weight of the evidence disconfirms the original hypothesis of transformation, and each independent instance of incongruence requires an additional hypothesis. If the initial (separate) attempted refutation is not successful, none of the competing hypotheses of transformation series additivity is supported (i.e., available evidence is completely ambiguous), and all of the competitors must be considered in the simultaneous test of congruence. Referring to characters as "nonadditive" or "unordered" in cladistic analysis means that all possible hypotheses of transformation series additivity are submitted in that test.

Incorporation of auxiliary assumptions is unavoidable in science, but their inclusion is justified only to the extent that they *increase* testability (Popper, 1959, pp. 82–83). Consequently, it is preferable to propose a single auxiliary assumption of transformation additivity because this increases testability by (1) minimizing the number of auxiliary assumptions (where each possible character state tree is an auxiliary assumption in cladistic analysis) and (2) maximizing the empirical content (i.e., by prohibiting more). If, however, the preference for a hypothesis of additivity is not the result of an objective test, then that preference is ad hoc and decreases the severity of test and explanatory power of the resulting cladistic hypothesis. As such, it is essential that the evidential basis for the choice of additivity be explicated clearly.

Many studies have evaluated the sensitivity of results to auxiliary assumptions of additivity by rerunning analyses with all transformations nonadditive (e.g., Kluge, 1991; Wilkinson, 1992; O'Leary and Geisler, 1999; Asher, 1999; Prendini, 2000). Clades that disappear when transformation series are considered nonadditive are often considered less supported than those that do not, but this interpretation is unfounded, given that this procedure does not evaluate the objective support of data for the hypothesis or the validity of the assumed additivity.

A more technical extension of this procedure is transformation series analysis (TSA), an iterative method of character analysis/reanalysis intended to remove incongruence in multistate characters and thereby make them consistent with the historical pattern implied by the rest of the data (Mickevich, 1982; Lipscomb, 1990, 1992; Mickevich and Weller, 1990; Mickevich and Lipscomb, 1991; Pogue and Mickevich, 1990). The method begins with a postulated additive transformation series. A tree is constructed from the data matrix in which that initial additive transformation series hypothesis is included. That character's topology is optimized on the most-parsimonious cladogram, a nearest neighbor matrix is constructed, and a new character state tree is formed by joining those states that have the greatest frequency of being nearest neighbor on the preferred cladogram. That character state tree then becomes the new additive transformation series for another round of tree searching and so on, until the transformation series does not change between iterations.

Analysis of nonadditive characters can only result in equal or fewer steps than analysis of additive characters, so discovery that nonadditive analysis returns a shorter topology alone does not disconfirm the hypothesized transformation series. That is, this outcome is a logical necessity, so if the shortest length is demanded, irrespective of independent evidence of additivity, then nonadditive analysis should be preferred a priori. Although it may be considered a minimal methodological requirement of TSA, any iterative procedure that "assured maximum congruence of all the data and always converged to the same solution(s)" (Buckup and Dyer, 1991, p. 502) would simply converge on the results (or a subset of the results) of unordered analysis and would therefore be redundant and less efficient. Furthermore, for preference of a particular auxiliary assumption of additivity to be valid, its determination must be external to the results of cladistic analysis. By recoding additivity based only on cladistic results, TSA leads to nonindependence of evidence, i.e., circularity instead of reciprocal illumination. Just as "[i]nitial hypotheses of character-state transformations should not be allowed to bias the construction of the very cladogram that is used to generate the final cladogram characters unless there is strong justification by a transmodal theory for those particular hypotheses" (Buckup and Dyer, 1991, p. 502), TSA alone should not be allowed to freely overturn independent evidence of additivity. That is, if the initial hypothesis of additivity was based on evidence, then overturning that evidence must be counted as an additional ad hoc-ism (which TSA does not do). If there was no basis for preferring one hypothesis of additivity over another, then the analysis should not have been constrained in the first place. TSA also fails in the case of partially ambiguous evidence, given that it could prefer (1) an a priori excluded transformation series because it is shorter than any of the a priori permissible transformation series or (2) a suboptimal transformation series (due to initial input sensitivity; Buckup and Dyer, 1991). A superior approach in this case is to test competing hypotheses of additivity directly and explicitly by running analyses with each of the alternatives and preferring the least refuted one(s).

Rerunning analyses with unordered characters indicates how dependent the optimal solution is on hypotheses of additivity and points to those hypotheses of additivity that are contradicted by external data (other characters). It therefore draws attention to potentially problematic transformation series and indicates a need for further study (Wilkinson, 1992), which can lead to detection of error. Although the iterative aspect of TSA is problematic, the construction of nearest neighbor matrices can facilitate reciprocal illumination. For example, multiple occurrences of a state as the nearest neighbor "suggest errors or failures to define character states properly" (Pogue and Mickevich, 1990, p. 330).

Methodological concordance. Methodological concordance applies multiple methods of phylogenetic analysis—such as maximum likelihood, neighbor-joining, parsimony, and UPGMA—to the same data set, accepts groups that are insensitive to choice of operation as well-supported and reliable, and rejects sensitive groups as weakly supported and unreliable. Although methodological concordance is, without doubt, one of the most popular methods of data exploration, few authors have given explicit justifications for its use. For example, Wägele and Misof (2001, p. 167, italics added) claimed that "[f]or independent support different genes and—if possible—*different methods of data analysis* are needed," yet they did not explicate the epistemological relevance or evidential significance of their assertion. Similarly, Barkman et al. (2000, pp. 13170–13171; see also Carranza et al., 2002, p. 247) concluded that congruence among methods is necessary to attain confidence, but neither they nor the papers that they cited (Miyamoto and Cracraft, 1991; Miyamoto and Fitch, 1995) provided arguments in defense of that position. As a consequence of the lack of explicit justification, there is little consistency in the way that methodological concordance is performed or interpreted. Here we address the most explicit justifications of which we are aware.

Flores-Villela et al. (2000, p. 714) defended methodological concordance on the grounds that

> Presumably, analysis and comparison of empirical data sets will influence opinions on methods...

As a prediction of social change, this may be true, but it does not provide a cognitive justification for this influence. Indeed, although they considered concordance among methods to be "an important conclusion," Flores-Villela et al. (2000, p. 732) immediately clarified:

> We do not claim that congruence among these alternative methods is necessarily evidence for strongly supported nodes but only that the stability of these clades is not sensitive to very different assumptions of character evolution.

Aside from proving the tautology that stable clades are not sensitive, if methodological concordance does not measure evidential support, then what good is it? In fact, despite the considerable importance that these authors ascribe to methodological concordance, their only reason for using it is that "no general consensus has been reached about the 'best' approach to phylogeny reconstruction" (Flores-Villela et al., 2000, p. 713). The same reason was given by Nei and Kumar (2000, p. 292), who advised

> Of course, if there is controversy over the method to be used, it is advisable to try several other methods and derive the most reasonable conclusion.

However, in the absence of an explicit definition or criterion of reasonableness (which is precisely what phylogenetic discovery operations aim to provide), methodological concordance is arbitrary and unscientific. Such unqualified pluralism defeats the purpose of performing quantitative analysis in the first place, and it returns phylogenetic inference to the days of subjective story telling.

Kim (1993) provided the most explicit justification of methodological congruence in his consideration of neighbor-joining, UPGMA, and parsimony. He used simulated data to assess the correlation between the average of Rohlf's (1982) strict consensus index ($\mathbf{CI_c}$) from pairwise comparisons of methods (method concordance index, **MCI**) and the average accuracy of the methods. He went on to propose differential weighting of empirical data sets to increase the value of **MCI** and, he argued, accuracy.

As has been pointed out repeatedly (Siddall and Kluge, 1997; Siddall, 1998; Pol and Siddall, 2001; Grant, 2002), simulation studies of this sort may provide important insight into the behavior of different methods, but they offer no indication of the accuracy or reliability of methods in empirical studies. Even under ideal, simulated conditions Kim's (1993) findings were not universal, and Kim (1993) reported that increasing **MCI** through reweighting actually decreased accuracy in some cases. This suggests that the correlation between accuracy and **MCI** may be highly dependent on particular simulation conditions. As has been demonstrated in related contexts (e.g., Tuffley and Steel, 1997; Steel and Penny, 2000), under certain conditions different methods will prefer the same phylogenetic hypothesis regardless of its veracity, and it is trivial to imagine cases in which any method or set of methods may give the same erroneous results (Farris, 1986, 1999). Kim (1993, p. 339) was aware of the potential limitations of his simulation-based findings, and for empirical data he suggested that:

> The weighted data set [chosen to maximize **MCI**] may be examined against available corroborative evidence to see whether the excluded characters were truly unreliable.

Exactly what this corroborative evidence could be and why it should not be included in the quantitative phylogenetic analysis in the first place was not stated, which, once again, renders methodological concordance arbitrary and unscientific.

For methodological concordance to be employed defensibly in empirical phylogenetic inference, a more general justification than mere correlation in simulations is required. To that end, Kim (1993, p. 333; see also Wheeler, 2000, p. 111) argued

> A high correlation is expected when the methods are accurate because all three trees must approach the same true tree.

However, although it is logically true that accurate methods necessarily converge, this does not entail that convergent methods are necessarily accurate. Such an invalid inference is an example of affirming the consequent (if p, then q; q, therefore p) and is an elementary error of logic.

Nevertheless, Kim (1993, p. 333) also found that

> . . .when the trees estimated by each method differ from the true tree, they also differ from each other. . .

which would imply that, even though agreement alone cannot be taken to imply accuracy, lack of accuracy could be inferred from disagreement among methods—a logically valid synthetic inference, provided that it is linked to a causal explanation. The explanation that Kim (1993, pp. 337–338) offered was that

> Presumably, the homoplastic characters cause the different methods to estimate erroneously the true tree. The results indicate that different methods are affected in different ways by the same set of homoplastic characters (i.e., in their estimations of different erroneous trees).

The implication is that disagreement among methods must be due to different erroneous interpretations of homoplastic characters, while agreement must be due to a lack of influence of homoplastic characters (leaving only the true signal). However, different methods are also affected in different ways by the same set of *non*homoplastic characters, this being a function of underlying assumptions applied to all data (not just homoplasies), and, even in this case of simulated data, which class of characters is responsible for absence of methodological concordance is unknown. Consequently, neither accuracy nor falsity can be inferred from methodological concordance.

That methodological concordance has been difficult to defend rationally should come as no surprise. Rohlf and Sokal (1965, p. 25; see also Sokal and Sneath, 1963) were unable to decide whether it was better to use a distance or a correlation coefficient for clustering when data include a mix of size-dependent and size-independent characters, and they suggested "that both coefficients should be computed and comparisons made, since both are valid measures of similarity." This pluralism set the tone for the development of numerical taxonomy and led to the proliferation of equally good phenetic clustering statistics, which ultimately resulted in the demise of phenetics. Later, Rohlf and Sokal (1980) proposed methodological concordance as a measure of stability, but it was summarily dismissed (Mickevich, 1980; Schuh and Farris, 1981; Farris, 1982). Although they did not retract it explicitly, Rohlf and Sokal (1981) no longer included methodological concordance as a measure of stability, and no attempt to justify it was made until their student resurrected it formally over a decade later (Kim, 1993). More recently, Kim (2000) incorporated methodological concordance into his common geometric framework, but this is simply an arbitrary mathematical construct designed to help intuition, not to make phylogenetic inferences. Advocates of methodological concordance have yet to offer reasons for excluding the many methods that they ignore, and the only reasons that they have given for performing methodological concordance at all are (1) lack of consensus of the best method of phylogenetic inference and (2) increased accuracy. Neither of these concerns is scientific. Furthermore, as discussed above, variation in assumptions is unable to assess the objective support of data for a hypothesis. Given that this procedure is unable to judge the validity of those competing assumptions, it has no relevance in hypothesis testing and is nonscientific.

As a heuristic method of data exploration, methodological concordance is an inefficient, if not indecipherable, approach. The various optimality criteria differ in their underlying assumptions to such an extent that only the simplest of disagreements can point to potential sources of error. The heuristic futility of this approach is further underscored by the fact that "[t]here are an infinite number of possible methods that could collectively yield any possible topology" (Brower, 2000, p. 148).

Sensitivity to prior probabilities (Bayesian phylogenetic inference). Two main approaches to Bayesian phylogenetic inference have been proposed in about the last decade. The first, by Wheeler (1991), is a more traditional Bayesian application that provides an explicit basis for assigning prior probabilities, calculates posterior probabilities analytically, and proposes an explicit method of assessing and interpreting sensitivity of results to the priors. The more recent trend in Bayesian phylogenetic inference (reviewed by Huelsenbeck et al., 2002) has only just begun to address the fundamental issue of selection of prior probabilities, approximates the posterior probability by sampling tree space using the Markov Chain Monte Carlo (MCMC) procedure (similar to Goloboff's (1999) familiar tree searching method of tree drifting), and has yet to propose a means of assessing and interpreting sensitivity to priors, suggesting merely that "the influence of the priors on the posterior distribution can be examined by rerunning the analysis with different priors" (Huelsenbeck et al., 2002, p. 681). Wheeler's method is therefore a more completely developed Bayesian application, and we focus on it primarily. However, the incompleteness of the currently popular approach is by no means a virtue, and our general criticisms of Wheeler's method are equally applicable to it.

Wheeler's (1991) Bayesian approach relies heavily on arguments for taxonomic congruence, assuming that confidence increases in proportion to agreement among cladograms derived from independently determined data sets. Thus, the approach is intended to capture both the strengths of the data and the confidence that obtains from taxonomic congruence. Wheeler (1991) developed the method in response to the fact that many consensus methods used to measure agreement among cladograms do not take account of weight of evidence for groups. A further strength of the Bayesian approach to taxonomic congruence is that it relies not on problematic arguments relating to the objective independence of classes of data (see Data partition methods, below) but merely on the temporal independence of analyses.

In all Bayesian approaches, the conclusion, i.e., posterior (Bayesian) probability, $p(\mathbf{h},\mathbf{e})$ is simply the product of the likelihood, $p(\mathbf{e},\mathbf{h})$, and the prior probability of the hypothesis alone, $p(\mathbf{h})$, normalized by the sum of those products for all $\mathbf{h}_{\mathrm{n}}$. In the present circumstance, **e** is a set of molecular data, and $\mathbf{h}_{\mathrm{n}}$ is a set of cladograms. This approach is inductive because the premises of the prior probabilities and the model employed in calculating the likelihood establish the conclusion as more or less probably true.[5]

Interpreting a Bayesian decision as minimizing risk and assuming a simple loss function for all possible values of **h**,[6] coupled with prior probability, $p(\mathbf{h})$, and data probability (likelihood), $p(\mathbf{e},\mathbf{h})$, values, Wheeler calculated the risk (cost) of any decision as the sum of the cost of all the decisions, where decision risk is the compliment of the final probability. Risk is interpreted as a measure of cladogram support, where minimizing risk is a function of maximizing the probability. "A single cladogram, or several, may be accepted until risk is sufficiently minimized" (Wheeler, 1991, p. 341).

Implementing Wheeler's (1991, pp. 339–440) approach requires calculating prior probabilities and likelihoods. In general, the prior probability of a cladogram analyzed with one kind of data (e.g., molecular) is its ability to explain another kind of data (e.g., morphological), a probability that is calculated for all possible cladograms. The prior probability of each cladogram is calculated by using parsimony to optimize a previously studied morphological data set onto each topology. The resulting lengths (numbers of evolutionary steps) are converted into probabilities by considering each step to be equivalent to a decrease in probability of a factor of e (the base of natural logarithms). The likelihoods are calculated for each topology by determining either the minimum length (unweighted or weighted, converted to a probability) or the maximum likelihood (where the rate of evolution is a modeled assumption, as in standard maximum likelihood applications) of a newly obtained molecular data set. In either the unweighted or the weighted parsimony approach, "[t]he most probable (least risk) cladogram will be the most-parsimonious." In any case, having calculated the prior probabilities and likelihoods, the posterior probability of each cladogram is simply the product of those probabilities divided by the sum of likelihoods of all cladograms. As noted above, the criterion for choosing among those cladograms is one of minimizing risk.

Wheeler (1991, pp. 337–338) appealed to an explicit form of sensitivity analysis based on simplex space and decision theory as a way of measuring the effect of prior probabilities, and in doing so he attempted to blunt the likelihoodists' claim that prior probabilities cannot be determined except in the most trivial cases. The posterior probabilities for a range of priors are plotted in simplex space, and their proximity to the decision lines is observed. "If this point is very close to one or several of these lines, small variations in the priors can affect the estimate of [the parameter], undermining our faith in the results" (Wheeler, 1991, p. 337).

[5] Wheeler's (1991, p. 336) "logical" probability is actually a conditional probabilistic measure of uncertainty, one based on sampling a sequence of events. These kinds of probabilities are typical of inductive inference (*sensu* Popper, 1959; Kluge, 2001, 2002).

[6] Simple in the sense that all incorrect hypotheses are assumed to be equally undesirable, and all correct hypotheses are assumed to be equally desirable, which allows a cost matrix to be constructed in the simplest of terms, 0 and 1, respectively.

Problems with this approach to data exploration include the following: (1) Sensitivity of posterior probabilities to choice of prior probabilities has no bearing on the objective validity of the priors and is therefore unrelated to hypothesis optimality and has no logical bearing on hypothesis preference. (2) As is the case for inference from taxonomic congruence generally, this Bayesian method is inferior to the simultaneous analysis of all critical evidence. The strength of Bayesian statistics is its ability to count prior statistical data indirectly by incorporating the results of prior tests, where it is assumed that prior and new data cannot be combined directly. However, this does not apply to phylogenetic analysis, where direct combination of all relevant evidence—including morphological (prior) and molecular (new) data—is nonproblematic. Likewise, we agree with Huelsenbeck et al. (2002, p. 680) that "[i]t is not only sensible to base conclusions on all the available information (total evidence); in some cases it is critical," but logic demands that all such "information" be objectively relevant to the inference of phylogeny, and, given that all relevant information can be analyzed simultaneously, there is no reason to isolate one set of observations as informing the prior probability and another as informing the likelihood. There is no epistemological justification for deriving confidence from taxonomic congruence and disregarding the increased explanatory power derived from a simultaneous analysis of the total evidence. (3) More generally, degree of belief (or risk) is unrelated to scientific discovery and evidential support (Hacking, 1965; Lakatos, 1998). This is illustrated by Huelsenbeck et al.'s (2002) approach, which employed the arbitrary opinions of "both systematists and amateurs" as prior probabilities relevant to the inference of phylogeny. Why should such undefended beliefs be attributed evidential significance in evaluating empirical knowledge claims? Should the beliefs of Creationists also be factored into phylogenetic inference? Or those of small children? Or the mentally ill? Should dreams and visions provide a basis for priors? If not, then some criterion of relevance must be formulated to objectively validate their exclusion. Most scientists would of course dismiss these questions as ridiculous, but they are central to the rational implementation of Bayesian statistics generally, and their absence from the recent systematics literature suggests that the current trend may be driven primarily by a fascination with the MCMC, which is exalted as "elegant and computationally efficient" (Huelsenbeck et al., 2002, p. 674), rather than a concern for cognitive advance. Although we disagree with his arguments, one of the reasons that we consider Wheeler's implementation to be more complete is that he at least attempted to address directly the fundamental and enormously problematic issue of assigning priors. (4) Induction provides no legitimate inference of truth (or probable truth). (5) Bayesian methods generally lack heurism because the effect of the prior probability in the calculus of the posterior probability is to reduce the effect of the likelihood, i.e., it reduces the ability of new observations to lead to new conclusions that contradict prior beliefs; yet, it is only by discovering inconsistencies with prior beliefs that new problems are pointed out. On the other hand, the objective interpretation "gives the more detailed account of the inferences within its domain, and hence has the virtue of being more readily open to refutation and subsequent improvement" (Hacking, 1965, p. ix).

Skewness test. The skewness of the distribution of cladogram lengths has long been advocated as a measure of phylogenetic structure or decisiveness in a data matrix (Fitch, 1979, p. 376; see also Le Quesne, 1989), most recently by Hillis and Huelsenbeck (Hillis, 1991; Huelsenbeck, 1991; Hillis and Huelsenbeck, 1992). The test consists of calculating the length of each bifurcating cladogram, or a "random" sample of those cladograms, for a given data matrix, and the $\mathbf{g}_1$ statistic of Sokal and Rohlf (1981) is then used to quantify the skewness of the resulting distribution of lengths. Typically, it is negative when the distribution is left-skewed, i.e., when the median exceeds the mean. The test has been applied to two or more sets of data, such as molecular and morphological partitions (Larson and Dimmick, 1993). According to Hillis (1991), the $\mathbf{g}_1$ statistic measures the strength of the phylogenetic signal.

The skewness test has several deficiencies. As Källersjö et al. (1992) demonstrated, the test can give indefensible conclusions, particularly when the frequency of the state within characters outweighs the congruence among characters. Also, Källersjö et al. demonstrated that the criterion is insensitive to the number of characters. Thus, skewness in the distribution of tree lengths does not accurately measure the degree to which a cladogram is supported. Especially condemning is the fact that the skewness test does not assess phylogenetic signal in proportion to being strongly left-skewed, as Hillis and Huelsenbeck contended. That test is determined mostly by the central mass of the distribution in tree lengths, whether the left tail of the distribution is strongly attenuate or not. Finally, when confronted with multifurcating cladograms, arbitrary resolutions may be counted as distinct, but that results in "exactly the wrong assessment of ambiguity in these matrices" (Källersjö et al., 1992, p. 286).

Hillis's (1991) skewness-based significance test concludes significant structure when $\mathbf{g}_1$ for the distribution of tree lengths for a data matrix is below (for example) the fifth percentile of distribution of tree lengths for data matrices produced under his null model, those being generated randomly and independently, with all states having the same expected frequency. Because skewness is influenced by both congruence and state frequency, Hillis's skewness-based significance test confounds the two effects. Huelsenbeck's (1991) advocacy of skewness is based on results from simulations, in which he assumed the same probability of character change on all branches of the cladogram. His correlation between most-parsimonious cladogram for simulated data and accuracy, as matches the simulated cladogram, when the distribution of tree length is strongly left-skewed, is then simply a function of his assumption that all branches have the same change probability (Källersjö et al., 1992, p. 279).

Skewness does not test phylogenetic hypotheses, and in light of these several criticisms there is no basis for recommending skewness as a heuristic. Nonetheless, the Hillis–Huelsenbeck skewness approach, and $\mathbf{g}_1$ significance test, continues to be used in phylogenetic studies (e.g., Crandall and Fitzpatrick, 1996; Jackman et al., 1999; Burbrink et al., 2000; Wiens, 2001; Reeder and Montanucci, 2001; Floyd, 2002; Gahn and Kammer, 2002; Lamb and Bauer, 2002; Salducci et al., 2002). It would appear that some investigators are compelled more by the appearances of statistical elegance than by scientific evaluation criteria (Grant, 2002). Perhaps the uncritical use of $\mathbf{g}_1$ is even better explained by the ready access given to it by popular computer software packages (e.g., PAUP*).

Computer-intensive sampling. We recognize two kinds of computer-intensive data exploration: Monte Carlo sampling and approximate randomization tests. Simulation lies outside our concept of computer-intensive data exploration because to simulate something is to subject a model to imaginary changes. The design of the model is modeling, and the modeling–simulation pair is a thought experiment. For example, we do not include the parametric bootstrap method (Goldman, 1993; Huelsenbeck et al., 1996c) in our analysis of data exploration methods because it uses a stochastic model of change for simulated data. We also exclude MoJo (Wenzel and Siddall, 1999) from our review because that method focuses on simulating the effect of noise (randomly generated character states), according to an equally probable model, while manipulating data from the original matrix.

Noreen (1989, p. 6) clarified the application of Monte Carlo sampling and approximate randomization methods:

> Monte Carlo sampling can be used when the hypothesis concerns a parameter of the population from which a random sample has been drawn. A randomization test can be used when the null hypothesis is that one variable is unrelated to another—whether or not the observations constitute a random sample.

The Monte Carlo approach applies to problems with and without inherent probabilistic structure. The bootstrap and jackknife applications that have become so popular in phylogenetic inference are examples of the former. In phylogenetic inference, approximate randomization tests are widely known as permutation methods, the most familiar being the permutation tail probability method (PTP).

Felsenstein (1985b; see also Felsenstein, 1988) proposed the bootstrap to estimate the reliability of phylogenetic inferences by estimating the uncertainty in the original matrix of data and placing confidence intervals on monophyletic groups. This application is model dependent, but it does not require that the model be specified explicitly because the model is inferred from the data by resampling characters at random from that matrix. Assuming that the number of replicates is large enough and that the original character matrix is representative of the population of all characters, then the resampling is expected to correspond to that which would be obtained by sampling repeatedly from the "real" population of all characters. As such, and as emphasized below, this application is conditional on the statistical sampling assumptions of independence and identical distribution (Felsenstein, 1985b). The level of confidence in a group is equated to the proportion of the times that the group is found among the bootstrap replicates, with the expected frequency depending on both the number of uncontradicted characters and the total number of characters in the matrix, which must be very large.

The jackknife, the other class of resampling methods commonly used to assess reliability in phylogenetic inference, involves the deletion of elements in the original data matrix, either taxa (Lanyon, 1985; Siddall, 1995) or characters (e.g., Farris et al., 1996). The original data set is sampled, usually without replacement, in forming the pseudoreplicate matrix, which is then analyzed for taxonomic relationships. Lanyon (1985) suggested a single taxon deletion approach as a way of assessing the stability of phylogenetic hypotheses, where a majority rule consensus of the results of the analyses of the jackknifed pseudoreplicates was advocated, each replicate lacking one taxon. The jackknife monophyly index of Siddall (1995) is also a measure of clade stability, but it excludes alternative suboptimal clades from the consensus.

Jackknifing on characters attempts to find or eliminate clades that are weakly supported by the data, which can be a function of either character incongruence or zero-length branches. The parsimony jackknife method (character jackknife) of Farris et al. (1996) was originally proposed as an efficient means of analyzing large data sets. It involves deleting sets of characters randomly and independently from the original matrix, each character having the same chance e^{-1} (approximately 0.3679) of being omitted from a given pseudoreplicate matrix (Farris, 1998). A new terminal taxon order is generated randomly in the formation of each pseudoreplicate, thereby diminishing the order sensitivity of terminal taxa. A most-parsimonious cladogram is calculated from each pseudoreplicate. A large number of pseudoreplicate matrices/most-parsimonious cladograms are calculated (e.g., 1000), and it is on the basis of that set of results that a resampling frequency for each clade is determined.

Unlike the bootstrap, the parsimony jackknife does not require very large matrices to validate the expected frequency of a group G set off by k uncontroverted synapomorphies as $1 - e^{-k}$ (Farris et al., 1996; Farris, 1998). There is also an important, if subtle, conceptual distinction between the bootstrap and the jackknife when applied to phylogenetic character data. While both methods manipulate the data at hand, the results of a bootstrap analysis are affected by the differential reweighting of the characters, whereas the results of a jackknife analysis are affected by changing sample size.

While still widely used, numerous authors have indirectly or directly criticized the bootstrap as a measure of uncertainty in phylogenetic inference (e.g., Mickevich, 1980; Farris, 1983; Felsenstein, 1985b, p. 785; Penny et al., 1990, p. 26; West and Faith, 1990, p. 18; Faith and Cranston, 1991, p. 121; Carpenter, 1992, p. 150, 1996; Felsenstein and Kishino, 1993; Jones et al., 1993, p. 97; Kluge and Wolf, 1993; Harshman, 1994; Sanderson, 1995; Siddall and Kluge, 1997; Farris, 1998; Sanderson and Wojciechowski, 2000; Siddall, 2002a), with many of those criticisms also applying to the jackknife. For example, a criticism that can be leveled at both approaches is that there is no universe of character-state transformations from which a probabilistic sampling distribution can be specified, because each such evolutionary event is necessarily unique. Additional criticisms include the following: (1) Even if there were such a universe, the characters and character-state entries in the original data matrix do not represent a random sample. Furthermore, (2) the relevant unknown parameter of phylogenetic inference, the tree, does not have frequentist probabilities associated with its nodes because each is necessarily unique. (3) Characters are not necessarily independent (the individual probabilities cannot then be multiplied) nor are they identically distributed (each character is not representative of a single common stochastic process). (4) The absence (and duplication in bootstrapping) of some synapomorphies in pseudoreplicates represents an unjustified form of differential character weighting, with the accompanying biases being unpredictable. (5) The bootstrap for a large clade is known to decline with increased taxon sampling, which has been interpreted as a statistical bias in bootstrap proportion. (6) In bootstrapping, the claim that monophyletic groups should be rejected when they appear in less than 95% of the pseudoreplicates has yet to be justified with regard to sampling theory—thus, any claim that the bootstrap has a bearing on accuracy, exclusive of precision, is without foundation (see also Siddall, 2002a, pp. 82–83).

A number of criticisms are also specific to the jackknife. Generally, the elimination of data, which comes with either taxon or character deletion, cannot be considered a virtue in science. Certainly, a problem with interpreting the jackknife statistically is that diminished power always obtains because those estimates are based on fewer observations than provided by the original data set. Lanyon's appeal to majority rule consensus as an optimality criterion for phylogenetic hypothesis choice is an obvious example of enumerative induction without a rational justification. In addition, Felsenstein (1988) severely criticized Lanyon's approach for technical reasons. Unfortunately, Lanyon's taxon deletion approach, like the bootstrap, continues to be used without justification (e.g., Hutchinson and Donnellan, 1992; Cicero and Johnson, 2001, 2002; Duffels and Turner, 2002).

The efficiency of parsimony jackknifing cannot be denied when applied to large matrices, such as the *rbc*L data set of Chase et al. (1993; Farris, 1998). However, Rice et al. (1997, p. 559) claimed epistemological deficiencies in the method. The major issues in question are whether parsimony jackknifing abandons the phylogenetic parsimony criterion and whether it is consistent with a refutationist philosophy. Farris (1998, p. 304) responded to these concerns by pointing out that the most-parsimonious cladogram(s) is determined for each replicate and further argued that the "purpose of resampling is not to discard the optimality criterion [of parsimony], but simply to allow ambiguous conclusions (poorly supported groups) to be identified efficiently." However, the accompanying "support" values cannot be interpreted as assessing the relative objective support provided by those data because (1) a simultaneous test including all critical evidence is never performed and (2) *resampling* frequencies are logically unrelated to degree of corroboration, although degree of corroboration can be increased by *accumulating* statistical evidence (i.e., increasing sample size; Popper, 1959, p. 411). Likewise, Siddall's (2002a, p. 88) defense of parsimony jackknifing as "only resolv[ing] clades that would also appear in all of the most parsimonious trees if one could actually find those trees…[and] not resolv[ing] clades that are not in those trees" is simply unfounded and was never claimed by Farris et al. (1996).

The approximate randomization class of computer-intensive tests involves permuting any of the different linear arrangements that can be made of a given set of objects. For example, in phylogenetic inference a data matrix serves as the basis for the randomizations of included characters. This involves permuting at random the entries (character states) within each column (character) of that matrix. A separate permutation can also be chosen at random for each character, so that, for example, congruence among characters in a randomization is just that produced by chance associations (Källersjö et al., 1992).

The permutation tail probability method (PTP) is supposed to assess the degree of phylogenetic structure in a data matrix. The method involves (1) reassigning characters to the original data matrix, under an equally probable random model with the number and frequency of states maintained, (2) finding the most-parsimonious cladogram for the data matrix that results from step 1, (3) repeating steps 1 and 2 many times, and (4) inferring phylogenetic structure from the frequency of the cladograms having a length at least as short as that for the original data—when the lengths of the cladograms from the original and permuted data matrices do not differ significantly it is assumed that the original matrix does not support the most-parsimonious cladogram calculated from it.

Rohlf's (1965) randomization test of the nonspecificity hypothesis (Sokal and Sneath, 1963; Farris, 1971) anticipated the PTP method. The PTP method was originally proposed as a measure of the statistical significance of phylogenetic conclusions (Le Quesne, 1969; Archie, 1989a) and subsequently interpreted erroneously by Faith (1992) and Salisbury (1999) as a measure of Popperian degree of corroboration. In actuality, PTP reflects only departure from a model of randomness, not corroboration (Carpenter, 1992), and PTP seems to have no phylogenetic interpretation at all (Källersjö et al., 1992; Carpenter et al., 1998). As Farris et al. (1995, p. 571; see also Swofford et al., 1996, p. 507) pointed out:

> ...the procedure models complete independence of characters [and] [a]ny kind of structure in the data—not just incongruence between matrices in particular—might cause significant departure from that model. If that method were used, interpreting significance as indicative of incongruence could thus easily be misleading.

Therefore, PTP does not necessarily measure phylogenetic structure in a matrix.

Topology-dependent cladistic permutation tail probability (conditional PTP, or T-PTP) is a modification of the PTP test that is supposed to measure the significance of the support by constraining a monophyletic group (Faith, 1991). A T-PTP test can be performed a posteriori (on monophyletic groups, as determined on a most-parsimonious hypothesis; e.g., Ballard et al., 1992) or a priori (on the data before they are analyzed for a most-parsimonious hypothesis; e.g., Faith, 1991). However, the method is flawed because it assigns "significant" support to both of two contradictory conclusions and when support is zero or even negative (Farris et al., 1994a; Swofford et al., 1996; Carpenter et al., 1998; Farris, 1998). Ancestor replacement is a serious problem with the a posteriori test, while randomized data are not the correct basis for an a priori statistical test because it cannot be ruled out that randomized data will have structure.

The reciprocal topology-dependent permutation tail probability (RT-PTP) test has been used in the analysis of data heterogeneity, where each minimum length cladogram or consensus cladogram is used as a constraint for the other data set (e.g., a morphological cladogram is constrained in the analysis of the molecular data set and vice versa; Thiele, 1993). When the difference in cladogram length, with and without the constraint, is equal to or greater than the differences obtained in some proportion of the randomized matrices (say, 50/1000) then the null hypothesis is rejected at the 5% level of significance, and the data sets are claimed to be uncombinable (see also the homoplasy excess ratio, HER, below). In not being able to reject the null hypothesis, the data sets are argued to be combinable, assuming that they mark the same underlying phylogeny. However, RT-PTP cannot be recommended because the distribution of randomized lengths is for a constraint (fixed) tree, and, like the a priori T-PTP test, it cannot be ruled out that randomized data will have structure.

The HER is a congruence index that permutes characters in a data matrix, assigning characters randomly to terminal taxa, thereby rendering characters independent of each other and of phylogeny (Archie, 1989a,b). In this method, congruence is assessed simply from the length of the most-parsimonious tree(s) for a data matrix. Most-parsimonious trees are calculated for the observed data and for each of a sample comprising a number **W** of randomizations. The lengths of the most-parsimonious trees for some number **E** of those randomizations exceed those for the observed data. If the lower tail probability (error rate, α'), $\alpha' = 1 - \mathbf{E}/(\mathbf{W}+1)$, is small enough (no greater than, say, 5%), the data differ significantly from random (Källersjö et al., 1992, p. 277). Unlike the distribution of tree length skewness (see above), α' is sensitive to the number of characters involved; however, α' is not sensitive to character state frequencies, as is the skewness index, because permutation does not change those frequencies. The single greatest weakness of the homoplasy excess ratio is that even though a data matrix can exhibit shorter-length trees than most of the randomizations from that matrix, that does not necessarily mean that the original data exhibited unambiguous hierarchic structure, i.e., strength of support is not measured.

Given the aforementioned criticisms of computer-intensive sampling, it is clear that they do not represent scientific tests. Likewise, these methods fail to measure objective support and therefore lack heurism. For example, the parsimony jackknife relies on sampling frequencies derived from partitioned analyses and never evaluates the congruence of all critical evidence in a simultaneous test; therefore it cannot be said to measure objective support. Nevertheless, the parsimony jackknife remains a useful part of an efficient strategy to analyze large data sets, as Farris et al. (1996) originally intended, without compromising severity of test. For example, Nixon's (1999) extremely efficient parsimony ratchet includes a jackknife procedure to escape local optima, and the command -jackstart in POY (Wheeler et al., 1996–2002) uses the parsimony jackknife to generate starting trees that can be fused (Goloboff, 1999) and submitted to swapping.

Long-branch attraction. Two methods of data exploration to test for long-branch attraction are currently available (see Siddall and Whiting (1999) for refutation of others). The simplest and most common approach is to apply parsimony and maximum likelihood to the same data set to examine the sensitivity of long branches to choice of method (e.g., Huelsenbeck, 1997). The validity of this procedure relies on the assumption that maximum likelihood is immune to long-branch attraction, but this was recently shown to be incorrect in simulation studies of 10 taxa (Pol and Siddall, 2001). Further, finding that maximum likelihood separates long branches does nothing to rule out the possibility of long-branch repulsion (Siddall, 1998), and maximum likelihood's reliance on counterfactual model assumptions and frequentist probability to address ideographic (historical) problems renders its results generally suspect (Siddall and Kluge, 1997). This procedure is therefore neither scientific nor heuristic.

Noting that long branches cannot attract each other when they are not simultaneously part of the same analysis, Siddall and Whiting (1999) proposed a parsimony-based method of pruning one and then the other long branch ("long-branch extraction") to determine whether the remaining branch is placed elsewhere in the tree. Insensitivity to long-branch extraction demonstrates that the placement of the long branches is not due to interactions of the two long branches, but this method falls short of an empirical test of the relationships of the long branch taxa (and therefore of long-branch attraction) because neither sensitivity nor robustness brings empirical evidence to bear on whether placement of long branches is real or artifactual. That is, this operation does not test the competing hypotheses, and preference for the most-parsimonious tree is justified rationally on the basis of increased explanatory power and testability, regardless of the placement of long branches relative to each other. Nonetheless, this method is strongly heuristic in that it may guide researchers in taxon sampling (e.g., by targeting taxa that may subdivide long branches) and character sampling (e.g., by targeting morphological characters that are less susceptible to long-branch attraction).

Likelihood ratio test (LR, Λ) for model selection. It is generally recognized that no single common mechanism (model) of molecular evolution is valid for all taxa and that the probabilistic model employed in a maximum likelihood analysis of phylogenetic relationships is deterministic of the results (Siddall and Kluge, 1997; Sullivan and Swofford, 1997; Cunningham et al., 1998; Kelsey et al., 1999; Wilgenbusch and de Queiroz, 2000; Posada and Crandall, 2001b). In an attempt to overcome this set of problems, the likelihood ratio is used as a test to choose from among a set of a priori plausible candidate models the most appropriate model for a particular group of taxa.

The statistical legitimacy of maximum likelihood is usually discussed with regard to the likelihood ratio test because the sum of those likelihoods has no particular meaning, each being a point, not a cumulative, probability (Hacking, 1965). The test is simply $\Lambda = L(\mathbf{h}_\mathrm{n}, \mathbf{e})/L(\mathbf{h}_\mathrm{a}, \mathbf{e})$,[7] where the numerator is the maximum likelihood of the null hypothesis, and the denominator is the maximum likelihood of the alternate hypothesis. According to Felsenstein (1983, p. 317; see also Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Pagel, 1999; Posada and Crandall, 2001a, b), the ratios of maximum likelihoods in phylogenetic inference "test whether a less general hypothesis can be rejected as compared to a more general one that includes it."

In model testing, the test becomes $\Lambda = L_0(\mathbf{h}, \mathbf{e})[= p(\mathbf{e}|\mathbf{M}_\mathrm{n}, \mathbf{h})]/L_1(\mathbf{h}, \mathbf{e})[= p(\mathbf{e}|\mathbf{M_a}, \mathbf{h})]$, where the numerator $L_0(\mathbf{h}, \mathbf{e})$ is the maximum likelihood of the function including the null model ($\mathbf{M}_\mathrm{n}$), and the denominator $L_1(\mathbf{h}, \mathbf{e})$ is the maximum likelihood of the function including an alternate model ($\mathbf{M}_\mathrm{a}$). The ratio is the degree to which one model maximizes the likelihood relative to that of another model. Assuming parameter comparability, more complex (parameter-rich) models always produce a higher maximum likelihood, but preference for simpler (less parameter-rich) models has been defended on the basis that (1) complex models require that a large number of parameters be estimated, which makes analyses computationally difficult and slow, and (2) greater complexity increases the error with which each parameter is estimated (e.g., Huelsenbeck and Rannala, 1997; Posada and Crandall, 2001a,b). It may also be argued that the likelihood ratio test refutes particular assumptions by comparing the maximum likelihoods of models that differ in a single parameter (Posada and Crandall, 2001b).

If the two models are special cases of one another (i.e., they involve nested sets of parameters), then the likelihood ratio is assumed to approximate a χ^2 statistic, with degrees of freedom equal to the difference in the number of free parameters estimated under the two models (e.g., Pagel, 1999). Goldman (1993) pointed out a number of problems with that assumption (see below) and suggested using simulations to generate the null distribution of the likelihood ratio test statistic. However, this is rarely done, either because of computation/time constraints or because models are not included in available simulation software (e.g., Buckley et al., 2001), and the χ^2 distribution is almost always assumed.

[7] Or, more simply yet, $LR = -2\log_e[H_1/H_2]$, where H_1 is the likelihood of the hypothesis that fits the data less well (Pagel, 1999).

Otherwise, by convention, $\Lambda > 4.0$ is taken to be evidence that one of the two hypotheses explains the evidence significantly better than the other (Edwards, 1972).

As currently applied in phylogenetic analysis, the likelihood ratio test must be dismissed as entirely ad hoc. Different phylogenies (topologies) may entail different best-fit models (e.g., Sullivan and Swofford, 1997; Kelsey et al., 1999; Buckley et al., 2001; Sanderson and Kim, 2000). For the test to be statistically valid in the proposed methods of model selection, the phylogeny (**h**) must be known a priori, i.e., it must be derived independently of the subsequent analysis (Goldman, 1993). However, the purpose of phylogenetic analysis is to infer relationships among taxa, meaning that the assumptions of the model are contingencies that require independent testing outside the model itself (Edwards, 1972; Thompson, 1975, p. 11; Farris, 1986; Sober, 1988; Goldman, 1990; see also Popper, 1979, pp. 191–193). Instead, in phylogenetics, the phylogeny (the unknown variable of interest) is first treated as known to estimate the model, and then that estimated model is treated as known to estimate the phylogeny (i.e., the probability of the model is conditional on the tree, and the probability of the tree is conditional on the model). This violation of empirical independence renders the approach statistically invalid.

Several authors have attempted to dismiss or mitigate this problem. Yang et al. (1995, p. 391) admitted that tree topology is a theoretical difficulty, but they dismissed its practical relevance because "the likelihoods of several reasonable trees, including the ML tree and (presumably) the true tree, are very similar," meaning that competing models affect the maximum likelihood score more than do those competing tree topologies. However, this begs the question as to what a "reasonable" tree is and why only such trees should be considered (see also Sanderson and Kim, 2000). In simulations, Posada and Crandall (2001b) found that initial neighbor-joining trees led to selection of the true model, but that random trees did not. This finding cannot be generalized to empirical data because the counterfactual premises that it relies on render it evidentially inert (Grant, 2002). Sullivan and Swofford (1997) proposed an iterative approach—beginning with a tree, selecting the best-fit model, searching under that model, selecting a new best-fit model, searching, etc., until stability is reached—but this does not mitigate the ad hoc-ness of the approach, nor does it necessarily avoid an infinite loop (e.g., where the best-fit model of tree A gives tree B, and the best-fit model of tree B gives tree A). Minimally, for phylogenetic model selection using the likelihood ratio test to be defensible, it must be established that selection among candidate models is *completely* insensitive to choice of initial topology (which is theoretically possible—albeit "a computationally chilling prospect" (Sanderson and Kim, 2000, p. 821)—given that the competing hypotheses form a closed set).

Additional criticisms of the likelihood ratio test of models as usually applied in phylogenetic inference include the following: (1) A difficulty in using the commonly assumed χ^2 distribution is that the number of parameters (and, therefore, the degrees of freedom) represented by a phylogenetic hypothesis is unclear. Goldman (1993) discussed problems with tree parameterization, but parameterization of nucleotide sequence evolution may also present difficulties. For example, transversion–transition ratio and base composition bias (base frequencies) relate to different model parameters, but transversion–transition ratios are not independent of base composition biases. A constraint to be A-T-rich seems certain to result in more transversions than transitions. (2) Likewise, parameter nonindependence (e.g., rate heterogeneity, Γ, and differential transversion–transition rates are clearly nonindependent) conflates the effects of parameter addition and invalidates inferences of evidential support for particular model assumptions. (3) Also, the asymptotic validity of the χ^2 distribution may not hold (Goldman, 1993), particularly when one or more parameters is fixed on the boundary of the set of permissible values (Whelan and Goldman, 1999; Ota et al., 2000; Goldman and Whelan, 2000). (4) Of special concern in phylogenetic applications of the likelihood ratio test is the inclusion of counterfactual models among the set of candidate models, and the exclusion of other, more realistic models. As Burnham and Anderson (1998, p. 8, italics in original; see also Goldman, 1993) cautioned: "*If a particular model (parameterization) does not make biological sense, it should not be included in the set of candidate models.*" "Biological sense" was foremost among the considerations that led Farris (1973b) to develop his model, but was dismissed by Felsenstein (1973, 1978) and many subsequent workers (e.g., Swofford et al., 1996) in favor of simpler calculations and statistical consistency. However, assurances of statistical consistency are irrelevant if the model is contradicted by reality because the resulting inferences are conditional on counterfactual premises (such as a common mechanism of evolution for nonhomologous transformations). There is no *statistical* justification for preferring consistency over other considerations, such as robustness or efficiency, and there is no *epistemological* justification for purposefully disregarding biological knowledge merely to simplify calculations (Farris, 1999). Of even more fundamental concern is the validity of the statistical (probabilistic) approach to phylogenetic inference (Siddall and Kluge, 1997; Grant, 2002; Kluge, 2002). Use of the likelihood ratio test presupposes the objective indeterminism of the system under study. However, as discussed above, that assumption is not valid in phylogenetic inference. The inferred phylogenetic events are historical, and history is objectively determinate, a fact that denies the objective validity of phylogenetic inferences derived from model-based methods such as maximum likelihood (and therefore the likelihood ratio test). (5) There is the presumption that a model "can be made indefinitely more complicated and realistic by adding parameters" (Felsenstein, 1983, p. 319; see also Thompson, 1975; Burnham and Anderson, 1998). Leaving aside whether counterfactual assumptions are *useful* in discriminating between alternative hypotheses (an instrumentalist interpretation), there is no stopping rule for adding parameters, which would seem to be a failure to find a scientifically relevant likelihood ratio test. That is, finding that the most parameter-rich (and therefore most realistic) model overfits the data (i.e., the model is overparameterized or overspecified) only suggests that the next-simpler model was not sufficiently realistic. (6) A related problem is that, as a test of competing candidate models, the likelihood ratio test does not actually test for goodness-of-fit; rather, it is a test for the significance of how much better the fit is among alternative models (i.e., relative goodness of fit, relative adequacy). Therefore, it is possible for one model to provide a significantly better fit than another and yet for that better-fitting hypothesis to not provide a significantly good fit. A procedure for evaluating model goodness-of-fit (adequacy) is given by Goldman (1993), but it has yet to be applied generally. Along those lines, and especially in consideration of (5), above, there would appear to be no statistical reason to exclude the highly parameter-rich (and presumably more realistic) "no common mechanism" model of Tuffley and Steel (1997) in tests of model adequacy and relative fit. That inclusion is crucial, given that it has been shown that maximum likelihood using the "no common mechanism" model selects the same tree(s) as parsimony under Fitch optimization (Tuffley and Steel, 1997; Steel and Penny, 2000). (7) The likelihood function, even for simple models, is not necessarily optimized at a unique point for a tree (Steel, 1994; Tuffley and Steel, 1997). Thus, it must be demonstrated that multiple optima do not exist when employing the likelihood ratio test.

In light of these several problems, it seems ironic that the likelihood ratio test continues to be cited as the basis for the credibility of maximum likelihood. In phylogenetics, at least, maximum likelihood would appear to have nothing to say about causal hypotheses that is not confounded by assuming what is at issue in the argument (petitio principii), the appearance of pursuing causality with the likelihood ratio test simply being more apparent than real. One still might make the argument for this test on the basis of its heuristic value, but that too would require justification of the counterfactual conditionals of the assumed models.

Amount of evidence. Numerous methods explore the sensitivity of results to variation in amounts of evidence. Although they are usually cast in terms of evaluating the effects of "missing" evidence, those procedures do not actually assess the effects of including unknown states ("?" entries), but rather they assess the effects of including *known* states. The effects of missing character-state entries are known a priori—they have no effect on cladogram length and may only decrease the ability to choose among competing hypotheses. What is at issue in these procedures is the decisiveness of available evidence, i.e., the decisiveness of the limited known character states in choosing among competing hypotheses.

Likewise, many methods are claimed to explore the effects of adding or removing taxa or characters, but instead explore the effects of adding or removing evidence. To better understand these methods, it is useful to divide the matrix into its components and to consider each independently. A matrix is composed of taxa, characters, and character-state entries/evidence. The effects of varying number of taxa and characters are determined logically, without recourse to data exploration. The sole effect of decreasing taxa is to reduce the empirical content of the competing hypotheses. The empirical content of a hypothesis is defined by its logical improbability (Popper, 1959; Kluge, 2001). In the special case of phylogenetic systematics, all possible hypotheses comprise a closed set, where the number of competing hypotheses is defined solely by the number of taxa; the number of competing hypotheses and, correspondingly, the logical improbability of any one hypothesis increases exponentially as a function of the number of taxa. Similarly, the sole effect of removing characters is to decrease the severity of the simultaneous, total-evidence test. For the verificationist employing frequentist probability, missing taxa and characters are relevant in that they necessarily alter observed frequencies (Siddall and Kluge, 1997; Siddall, 2001).

The most common procedure to explore the effects of adding evidence is to run analyses with and without classes of character-state entries for which some portion of entries is missing. Character-state entries may be classified according to either the kind of taxa from which they were coded, such as fossil and extant taxa, or the characters of which they are part, such as molecular and morphological characters. This procedure was used in studies of amniote phylogeny to demonstrate the importance of including evidence from fossils in phylogenetic analysis. Gardiner (1982; see also Patterson, 1981) stated that fossil evidence could not overturn his hypothesis of amniote phylogeny based on evidence from extant taxa alone. Gauthier et al. (1988) refuted that conjecture with a set of empirical "experiments," where evidence from amniote fossils was excluded and included in the reanalysis of relationships. Gauthier et al. also discussed the basis for their finding with regard to the patterns of evidence potential in fossils, and

Donoghue et al. (1989) extended those arguments of evidence to taxa in general. More recently, Eernisse and Kluge (1993) were unable, with the addition of gene sequence data, to refute Gauthier et al.'s finding that the pattern of the evidence attributed to fossils is important in the analysis of amniote relationships. Minimally, fossils must be included when they are available because (1) in principle they increase the testability of phylogenetic hypotheses and (2) in practice it cannot be known a priori when they will not make a difference to the results of any particular study.

Although there is now widespread recognition of the importance of taxa with missing evidence, such as fossils, there is also much concern that such taxa may obscure otherwise well-corroborated relationships by being placed almost anywhere in a cladogram without increasing length, which leads to a proliferation of most-parsimonious solutions that collapse into a polytomy in consensus (Nixon and Wheeler, 1992; Novacek, 1992a,b; Wilkinson, 1995; Gao and Norell, 1998; Kearney, 1998, 2002; Anderson, 2001). Mere a priori exclusion of taxa on the basis of degree of evidential completeness is an inadequate procedure because degree of completeness is not necessarily correlated with this wildcard behavior (Novacek, 1992a,b; Kearney, 1998, 2002; see also Gauthier et al., 1988; Gao and Norell, 1998). To discover and eliminate wildcard taxa, Wilkinson (1995) provided rules of "safe taxonomic reduction" that allow taxa with missing character-state entries to be removed if and only if their removal does not affect the placement of other taxa. Alternatively, Anderson (2001) proposed the "phylogenetic trunk" method to eliminate the most ambiguously placed taxa. In Anderson's (2001) method, a total-evidence analysis is run and the most problematic taxa are identified by either Adams consensus (Adams, 1972) or systematic deletion of taxa. The most variably placed taxon is then excluded from the analysis, and "[f]urther iterations are performed by using this procedure until the desired level of resolution is achieved" (Anderson, 2001, p. 174). Arnedo et al. (2002) also removed taxa with missing character-state entries, but they evaluated the effect of taxon removal by measuring incongruence among data partitions with the RILD (Wheeler and Hayashi, 1998). An improved RILD score was used as the basis for permanent taxon removal.

These methods of data exploration are not tests. Minimization of the number of most-parsimonious cladograms and maximization of resolution and taxonomic congruence are not scientifically defensible optimality criteria (Grant, 2002); any attempt to decrease ambiguity or incongruence through elimination of evidence results in a lack of independence and renders conclusions nonempirical. We agree with Kearney's (2002, p. 380) conclusion that "[a]mbiguity of results calls for reexamination of data and addition of new data, rather than use of methods that may imply more resolution than the data support," which we take as an endorsement of the strict consensus, given that it collapses clades that are not unambiguously supported by the data. The methods of Anderson and Arnedo et al. must be rejected because they fail to distinguish between ambiguity due to lack of evidence and ambiguity due to conflict of evidence. Indeed, Anderson's own finding that a taxon with 76.9% missing evidence is stable, while another with 67.6% missing evidence is not, suggests that character conflict may be more important. Wilkinson's approach is clearly superior in that any ambiguity attributable exclusively to taxonomic equivalents cannot be due to coded character states.

Identification of the taxa for which available evidence is indecisive is heuristic in that it allows investigators to give priority to those taxa and characters when gathering additional evidence (cf. Kearney, 2002; quoted above). For example, the discovery that a single taxon with many missing character-state entries is primarily responsible for ambiguity would allow researchers to expend a disproportionate amount of their limited resources on obtaining the missing data (e.g., special field work to collect more specimens or special protocols for DNA extraction). A clear advantage of Wilkinson's (1995) method is that it ensures that taxon elimination during heuristic data exploration does not alter the fundamental topology of the remaining taxa (although other considerations may be affected, such as character state optimizations or measures of nodal support). Curiously, Anderson (2001, p. 174, italics added) claimed that one of the strengths of his method is that "by reanalyzing the matrix after each pruning cycle, the phylogenetic trunk method permits the discovery of a topology *different from those within the component trees*." Similarly, Arnedo et al. (2002, p. 317) reasoned that

> If the presence of a certain taxon without any information for some of the data partitions was responsible for obtaining spurious results, then the congruence between data partitions should increase with the removal of the incomplete taxon. The rationale is that character transformations of the combined [= complete] analysis selected only because of the presence of missing data are very likely to be in strong disagreement with character transformations supported in the partial analyses, which do not have missing data.

However, missing character-state entries cannot affect the placement of other taxa or increase cladogram length, so the differences in topology and partition incongruence must be due entirely to the elimination of the *coded* character-state entries, not the missing entries! Wilkinson (1995) had already recognized this, and his method was explicitly designed to *avoid* the problem.

Poe (1998, p. 18) proposed a method of measuring the effect of including evidence from different numbers of taxa "by mapping characters from a matrix of culled taxa onto optimal trees for that reduced matrix and onto the pruned optimal tree for the entire matrix, then comparing the length of the reduced tree to the length of the pruned complete tree." In other words, a difference in tree length is considered the degree to which adding or removing taxa changes the evidentiary *estimate* of phylogeny. In refining this estimator, Poe analyzed 29 different data sets, from which he calculated a second-order regression equation describing the relationship of the fraction of taxa sampled to the sensitivity to sampling. This equation was then transformed to a linear relationship that described the total number of taxa sampled. Two significant problems are evident in just these preliminaries: (1) Foremost, the application of frequentist statistics in estimating optimal taxonomic sampling cannot be rationally justified because the history of species is necessarily unique. Logic denies the application of a statistical test in this context. (2) There is no compelling justification for tree length as the preferred measure of sensitivity to taxon sampling, as opposed to the distortion index, which is a *relative* measure of number of extra steps (Farris, 1989a,b). One must not lose sight of the fact that what is predicted by the independent variables in a multiple regression analysis depends on the dependent variable, i.e., on one's notion of sensitivity to taxonomic sampling and how it is defined.

The limitations of the chosen statistical model, multiple regression analysis, must also be considered: (1) Correlation may be indicated, but the set of causal mechanisms is not. (2) Including as many predictors as possible increases one's chance of finding a significant correlation; however, the number of observations per independent variable must be large enough to ensure that the estimate of the regression line is stable. That scientific knowledge springs from correlation is denied, and while only few variables are analyzed and the sample sizes seem reasonable, as Poe admitted (p. 25), "experiments with the data from this paper suggest that both retention index and number of characters may eventually turn out to be significant."

Finally, consider the assumptions of regression analysis and whether they are violated in the case of the particular independent (predictor) variables that Poe chose, i.e., number of taxa, number of informative characters, degree of homoplasy (retention index), total (Bremer) support, and index (I) of symmetry. The standard assumptions of regression analysis are (1) interval or near-interval data, (2) data whose range is not truncated, (3) linear relationships among variables, (4) homoscedasticity (same range of relationship) throughout the range of the independent variable, (5) normal distribution of residuals (predicted minus observed values), and (6) absence of multicollinearity (redundancy of statistical indicators) and matrix ill-conditioning. As Poe acknowledged, the range of the data is severely truncated in the case of the number of taxa and number of informative characters and cannot be judged predictive of actual phylogenetic research, where both numbers are much larger. Also, as acknowledged by Poe, at least some of the independent variables exhibited multicollinearity, and this casts further doubt on the meaningfulness of the author's interpretation that number of taxa is the most, and only, significant predictor of sensitivity of taxon sampling. Such flawed statistical approaches cannot be judged heuristic.

Polymorphism. Several methods for dealing with the ambiguity of variable terminal taxa have been proposed: (1) ambiguity coding, (2) excluding variable characters, (3) frequency coding, such as majority or modal coding, (4) splitting taxa into monotypic terminals, and (5) inferring ancestral states (for another classification of methods see Kornet and Turner, 1999). Various arguments have been advanced for and against each of these methods in phylogenetic inference, and sensitivity analysis has been used to evaluate the alternatives. For example, Wiens (1995; see also Wiens, 2000a, pp. 133–138) and Smith and Gutberlet (2001) performed sensitivity analyses using criteria such as number of most-parsimonious trees, number of informative characters, skewness ($\mathbf{g}_1$ statistic), and bootstrap support to determine which method of treating polymorphic characters is optimal. However, as Grant (2002, p. 105) pointed out, "these evaluation criteria are not sufficient to defensibly select one discovery operation over another because they are unrelated to the scientific principles of explanatory power and severity of test." Thus, the Wiens and the Smith and Gutberlet studies represent misapplications of sensitivity analysis in science.

Even more general conclusions follow from the fact that the polymorphism ascribed to terminal taxa, including higher taxa (Nixon and Davis, 1991; Donoghue, 1994; Simmons, 2001), is only investigator error: (1) To employ a polymorphic terminal taxon that is not the smallest historical individual is a potential failure because the common ancestral species of a group "is *identical* with *all* the species that have arisen from it" (Hennig, 1966, p. 71; italics added). (2) Also, to use two or more different semaphoronts in the description and codification of a character can give the appearance of polymorphism where none actually exists among comparable individuals. (3) Although some polymorphism may be irreducible, as may occur within an organism (e.g., heterozygosity) or semaphoront of a smallest historical individual (species), that kind of polymorphism is due only to the inability of the investigator to discriminate between character history and organism or taxon history. That is, taxon phylogeny is inferred from hypothesized transformations from one character state to another (Hennig, 1966). Polymorphism is observed when those transformations do not unambiguously demarcate the cladistic events that gave rise to the taxa in question. As such, the problem of polymorphism in phylogenetic inference is due not to characterizing taxa with regard to observations made on organisms (Campbell and Frost, 1993, p. 62) but to not matching the *individuality* of taxon and character histories.

This distinction affects the epistemological validity of the analytical solutions that have been proposed to eliminate the ambiguity of polymorphism. For example, conversion of polymorphism into a separate, "polymorphic" state or frequency eliminates ambiguity, but that is a case of overreductionism (Frost and Kluge, 1994, p. 266) because it mistakenly equates character-state transformations, such as A → B, with changes in the distribution of those states among organisms, A → AB → B. The "state" AB logically could not occur in the evolution of the character and is therefore evidentially irrelevant. Approaches that convert polymorphism into frequency are further dismissed because frequency is an abstraction; it is neither heritable nor a thing (Murphy, 1993; Wiens, 2000a, p. 130; contra de Queiroz, 1987), rendering it evidentially irrelevant in phylogenetics. Frequencies are defensibly interpreted as objective probabilities in nomothetic, predictive sciences, such as population genetics (e.g., Kimura, 1955), but such justifications are irrelevant to the ideographic science of phylogenetics, which derives evidence from concrete, spatiotemporally restricted events (character-state transformations), not abstractions (contra Wiens, 1998, 1999, 2000a, 2001; see also Swofford and Berlocher, 1987; Berlocher and Swofford, 1997). In light of these considerations, we must disagree with Smith and Gutberlet's (2001, p. 166) conclusion that "frequency coding is philosophically sound and consistent with the tenets of phylogenetic systematics," a conclusion that they reached merely on the basis that intraspecific variation may be observed empirically.

Consequently, beyond simply coding polymorphism as ambiguous information, the only course of action for the phylogeneticist is to discover the basis for the error and eliminate it from the data matrix. Any attempt to model the error of polymorphism in phylogenetic inference, or to apply a methodology that attempts to deal with the error by codification (e.g., Wiens, 1998, 2000a), cannot be heuristic because that kind of error has no ontological standing in science. Indeed, the exclusion of frequency data is not "contrary to the maxim of total evidence," as asserted by Wiens (1999, p. 343; see also Wiens, 2000a, p. 130), because that rule of scientific conduct covers only data that are *relevant* to the inference, which, in the case of species and the natural groups of which they are a part, are necessarily unique character-state transformations. That Wiens's (2000a, p. 138) simulation studies indicate that the majority method for coding variable higher taxa is to be preferred, and from which he claimed "the common-equals-primitive assumption may have some predictive value… because it uses some information on the distribution of states within the variable higher taxon," only underscores additional erroneous reasoning. As Hennig (1966, Fig. 21) clearly demonstrated, only the apomorphic state can be informative of species relationships.

Clade concordance index (CC). Nixon and Carpenter (1996a, p. 314) defined a measure of ambiguity, or "inter-cladogram character conflict for all characters among a set of cladograms," as clade concordance, $CC = 1 - (((\sum GLn) - PL)/(CL - PL))$, where GL is the greatest length of each character n observed among the cladograms, PL is the length of the most-parsimonious cladogram(s), and CL is the length of the strict consensus of the set of most-parsimonious cladograms. This index measures the conflict over all characters that occurs *between* equally most-parsimonious cladograms by making use of the length of the strict consensus topology. As such, the index may be an efficient way to detect an overall wildcard effect; however, to actually remove putatively wildcard taxa has the effect of reducing the empirical content of the competing hypotheses. Furthermore, the resulting increase in resolution creates the impression of increased empirical knowledge, where in fact empirical knowledge has been decreased through the exclusion of evidence. While the clade concordance index may be useful in computer programming (Nixon and Carpenter, 1996a), it does not provide a "test" of ambiguity, neither of kind nor precisely where it occurs, and it is therefore not heuristic.

A more efficient assessment of ambiguity may be achieved through simple inspection of the most-parsimonious hypothesis(es) of a total-evidence analysis. In that context, an unambiguously defined group is one that appears in all members of the set of equally most-parsimonious cladograms, and an unambiguously optimized synapomorphy is one that diagnoses just that group; otherwise, the group is ambiguously defined and objectively unsupported. An ambiguous character varies in the number of steps that it exhibits in the neighborhood of the taxa that defines the ambiguous group. To remove ambiguity scientifically, one performs more decisive tests, either through character reanalysis or inclusion of additional critical evidence in a simultaneous, multiple test, not by eliminating evidence (Kearney, 2002, p. 380).

Other kinds of quality analysis

Character compatibility. This test was first proposed by Wilson (1965) and formalized by Le Quesne (1969). Two characters are said to be compatible when their state transformations can be mapped on the same branching pattern as unique and unreversed; otherwise, those characters are incompatible. Compatible characters are both congruent with the same hypothesis of relationships and consistent with the same explanation of inheritance, that of synapomorphies interpreted as homologues. Generalizing, the character compatibility test is the following: (1) Given any pair of characters, i and j, each with two states, x and y, (e.g., 0,1), find the counts $N(0,0)$, $N(0,1)$, $N(1,0)$, and $N(1,1)$, where $N(x,y)$ is the number of terminal taxa in the data matrix exhibiting state x in the ith character and state y in the jth character. The counts of these possible combinations can be efficiently summarized in the form of a 2×2 table ($i_{x,y}$, $j_{x,y}$). (2) Making no assumptions about which state is plesiomorphic, if one or more of these four Ns is zero then i and j are said to be compatible; otherwise, they are incompatible. Assuming that one state is plesiomorphic (e.g., $0 \rightarrow 1$) and finding N to be zero for one or more of the three derived combinations, $N(0,1)$, $N(1,0)$, and $N(1,1)$, is evidence that characters i and j are compatible; otherwise, they are incompatible. It can be easily proven by explicit enumeration that if all four Ns are nonzero, then any phylogenetic hypothesis that requires no independent evolution in character i must require at least one case of independent evolution in character j and conversely. However, all that can be deduced from a compatibility analysis is that two incompatible characters cannot both be explained as homologous, indicating that at least one of the homology statements is false. Such pairwise comparisons are tests and may therefore play a valid part in the cycle of cladistics research (Kluge, 1997, p. 90; see, however, Kluge, 1998, p. 351), although they are not as severe as the simultaneous test of character congruence provided by parsimony when it is applied to a matrix of three or more characters (Kluge, 1997), and they say nothing about character reliability (*contra* Penny and Hendy, 1985a, 1986). Compatibility tests are certainly heuristic in that they indicate the need for additional, independent testing (i.e., character reanalysis) because incompatible characters cannot both be homologous (i.e., one must be erroneous; Farris, 1983, p. 9).

Spectral analysis. Spectral analysis (Hendy and Penny, 1993; Hendy and Charleston, 1993; Penny et al., 1993; Steel et al., 1993a; Hendy et al., 1994) may be viewed as related to character compatibility in that it also evaluates pairwise conflict among hypotheses of synapomorphy without assessing conflict at all levels. However, the stated goal of spectral analysis is to provide accurate and reliable estimates of phylogeny, where statistical consistency is given primacy, and it aims to achieve this by improving the quality of the data prior to evaluation of competing phylogenetic hypotheses. Swofford et al. (1996, p. 472) highlighted spectral analysis as a method of data exploration, suggesting that, "[a]part from their use in estimating trees, spectral analysis methods are useful as aids in understanding the peculiarities of particular data sets."

Spectral analysis begins by calculating the relative frequency of bipartitions (splits) implied by each character in isolation (observed sequence spectrum, **s**). Next, under a chosen probabilistic model (mechanism) of sequence evolution (e.g., 3ST of Kimura, 1981), the Hadamard transform is applied (giving the conjugate spectrum, γ) to provide a global "correction" for all unobserved substitutions prior to selection of the preferred hypothesis of relationships—an essential aspect of phylogenetic analysis in this paradigm (e.g., Penny et al., 1993, 1996; Steel et al., 1993a; Lento et al., 1995). An optimality criterion (e.g., parsimony) can then be applied to the transformed data to select an optimal cladogram. However, a full spectral analysis uses the Hadamard conjugation to interconvert between a given tree (including branch lengths) and the expected sequence spectrum (or tree spectrum, **q**), which enables the closest tree criterion (Hendy, 1989; Hendy and Charleston, 1993; Hendy and Penny, 1993) to employ a least squares procedure to select the tree (with branch lengths) for which the distance between **q** and γ is minimal.

Of primary concern in spectral analysis is the validity of the global corrections applied to observed sequences. The authors see data correction as an essential step in phylogenetic analysis because statistical consistency is model specific (Farris, 1983, p. 17), meaning that, irrespective of the method of analysis, statistical consistency can only be guaranteed if data do not deviate from the assumed model (Penny et al., 1993, 1996; Steel et al., 1993a). The Hadamard conjugation provides a means of transforming data to conform to the assumed model, thereby "correcting" the data for the multiple, unobserved changes that must have occurred, given the truth of the model. However, no empirical *evidence* is actually brought forth to allow the unobserved changes to be inferred, and the claim remains entirely untested.[8] We see no increase in knowledge to be claimed from the antiempirical practice of forcing data to conform to a preconceived model, especially considering that the models in question are demonstrably counterfactual and are deterministic to the outcome of analysis (Siddall and Kluge, 1997). Furthermore, statistical consistency can never be guaranteed in practice because the truth of the model can never be guaranteed (Farris, 1999), so we see no reason to prefer spectral analysis over a method that guarantees to maximize explanatory power and testability (viz., phylogenetic parsimony; Kluge, 1997, 1999). More simply, we consider *logical* consistency to take precedence over *statistical* consistency.

As a method of data exploration, a full spectral analysis involving the Hadamard conjugation and data

[8] Although it is often claimed in this paradigm that the data falsify (reject) the model, "model" in this case refers to a given tree with specified branch lengths (e.g., Penny et al., 1993; Steel et al., 1993a). The probabilistic model (or mechanism) of change is not tested in this procedure.

"correction" can serve no heuristic purpose. Nonetheless, plotting the observed sequence spectrum is minimally heuristic in that it provides some indication of the extent of conflict in the data and may point to alternative (suboptimal) hypotheses worthy of special consideration. However, because it considers only bipartitions implied by characters in isolation and does not evaluate the results of the interactions among all the characters, the observed sequence spectrum provides a weaker test of congruence than that given by a parsimony analysis (Kluge, 1997) and a weaker indication of strength of preference or signal than other methods of data exploration. Furthermore, it does not identify the individual characters responsible for conflict, nor does it identify instances of conflict at all levels, all of which denies its heurism.

Relative apparent synapomorphy analysis (RASA). RASA has been described as "a tree-independent conceptual framework of phylogenetic data exploration" (Lyons-Weiler and Hoelzer, 1997, p. 375). The method begins by counting, for each pair of taxa, *i* and *j*, the number of times that a three-taxon statement in which grouping *i* and *j* is supported by that character (or, more simply, for each character in which *i* and *j* share the same character state, the number of taxa that have a different state), summed over all characters. This sum is referred to as the relative apparent synapomorphy, *RAS* (although this is a misnomer; see below). Next, the number of characters for which *i* and *j* share the same state and at least one other taxon has a different state is counted. This value is referred to as phenetic similarity, *E*. A least squares linear regression of *RAS* on *E* is performed, and the resulting slope, *b*, is compared to a null slope, β, originally (Lyons-Weiler et al., 1996) obtained from $(\sum_{ij} RAS_{ij})/(\sum_{ij} E_{ij})$ and later (Lyons-Weiler and Hoelzer, 1999) obtained from a permutation approach in which a large number of randomized matrices are generated by permuting entries within each character and the slope of *RAS* against *E* for each randomized matrix, averaged over the number of randomizations (Archie, 1989b). Student's *t* test is carried out by calculating the test statistic, $t_{RASA} = (b - \beta)/s_b$ (where s_b is the standard error of *b*), and degrees of freedom, $\gamma = m - N - 3$, for *m* taxon pairs and *N* taxa (see Farris (2002a) for a detailed statistical discussion).

RAS analysis (RASA) was proposed as a means of assessing the quality of data with regard to phylogenetic signal (Lyons-Weiler et al., 1996). In the interest of detecting and eliminating "problematic" evidence, an increased t_{RASA} score has been invoked as the basis for discarding data to avoid long-branch attraction (Lyons-Weiler and Hoelzer, 1997), use optimal outgroups (Lyons-Weiler et al., 1998), detect lineage sorting (Lyons-Weiler and Milinkovitch, 1997), and eliminate "noise" (Barkman et al., 2000). The method has been employed by numerous authors working with diverse taxa (e.g., Hall et al., 1998; Milinkovitch and Lyons-Weiler, 1998; Teeling et al., 2000; Chek et al., 2001; Austin et al., 2002).

However, Simmons et al. (2002), Faivovich (2002), and Farris (2002a) have pointed out a large number of flaws in so-called "RASA theory." Summarizing those authors' findings, those flaws include the following: (1) As a count of three-taxon statements, *RAS* is a measure of phenetic similarity, not synapomorphy (Kluge and Farris, 1999). As such, regressing *RAS* (a phenetic measure) on *E* (another phenetic measure) has no relation to phylogenetic signal, and RASA would better be considered "relative apparent similarity analysis" (Farris, 2002a, p. 336). (2) RASA attributes significant hierarchic structure when there is none. (3) Remarkably, RASA also fails to detect hierarchic structure in highly structured data sets! (4) More generally, the RASA regression does not meet the minimum requirements of a statistically valid regression analysis, making it only a regression analogy. Not the least of these requirements is that the dependent variable, *Y*, be sampled randomly and independently and that *Y* be a linear function of the independent variable, *X*. In RASA, both *RAS* and *E* are calculated deterministically from the data matrix, meaning that, if anything, RASA measures "the inaccuracy of the premise that [*RAS*] is a linear function of *E* for the particular character matrix in question" (Farris, 2002a, p. 343). (5) The Student's *t* distribution is inappropriate in this case, meaning that the *t* test statistic is also invalid in this case. (6) Using the RASA slope as a test for hierarchic structure is counterproductive because "the rejection region is cluttered with matrices that have high RASA slope but are poorly structured, while matrices that are in fact highly structured but have lower slopes are forced out of the rejection region" (Farris, 2002a, p. 347). (7) t_{RASA} is highly sensitive to character state frequency. (8) RASA's ability to detect hierarchic structure is highly sensitive to departure from a clock. (9) RASA fails as a detector of long-branch attraction. RASA may indicate long-branch attraction when none is present and may also fail to detect long-branch attraction when it is present. Moreover, the specific long-branch attraction avoidance strategy of Lyons-Weiler and Hoelzer (1997) may actually cause long-branch attraction to occur. Simmons et al. (2002) found that RASA identified a zero-length branch as a problematic long branch in a matrix that did not involve long-branch attraction and that removal of that problematic terminal actually caused convergent terminals to attract. (10) The recommended procedure of discarding evidence to increase t_{RASA} scores is contrary to the basic principles of science. It is true that "[t]ests of normality are widely used before parameter estimation and hypothesis testing" (Barkman et al., 2000, p. 13166), but the purpose of those tests is to evaluate the appropriateness of the chosen test statistic, not the accuracy of the data. Finding that a sample does not fit a normal distribution disallows standard parametric tests that assume normality; it does not justify forcing observations into conformity through data purification. In standard statistics, outliers may be excluded only if the source of observer error can be determined or if appropriate tests demonstrate that the outliers were drawn from a different population. Likewise, under certain, well-defined conditions a suitable transformation, such as a logarithmic transformation when the factor effects are multiplicative, not additive, may be applied to the data. However, such transformations do not discard data. Rather than improve phylogenetic inference, RASA's data removal protocols violate independence, decrease severity of test and explanatory power, and actually obscure the hierarchic structure present in the data. There can be no heurism in such a procedure.

Data partition methods (taxonomic congruence). The approaches and justifications for analyzing data partitions separately have changed considerably since the idea first grew out of the pheneticists' nonspecificity hypothesis some 40 years ago (e.g., Sneath and Sokal, 1962; Rohlf, 1963; Sokal and Sneath, 1963; Rohlf, 1965; see Kluge (1989) for a brief history up to that time). In recent years, the focus has shifted yet again. Contrary to workers who advocate deriving phylogenetic inferences from *either* total evidence (e.g., Kluge, 1997, 1998) *or* taxonomic congruence (e.g., Cracraft and Helm-Bychowski, 1991; Miyamoto and Cracraft, 1991; Swofford, 1991; Miyamoto and Fitch, 1995), the majority of contemporary workers advocate combined analysis but also explore the effects of separate analyses of data partitions. Indeed, that kind of data exploration seems to surpass methodological concordance in its current popularity. Underlying this data exploration approach is a concern for the quality of data and the strength and validity of the phylogenetic inferences based on them.

Two lines of argument in defense of partition methods of data exploration, which we refer to as the *strong* and *weak* interpretations, have emerged. The justification of the strong interpretation, where the concern is for the homogeneity of the data with respect to a chosen evolutionary model, is explicitly statistical, because, for example, "[i]n a phylogenetic context, data homogeneity can be defined as the sharing of a single history...and uniform probabilities of change among character states" (Barker and Lutzoni, 2002, p. 625). Accordingly, Bull et al. (1993, p. 385; italics added) argued that the simultaneous analysis of many different characters "increases the chance that support for *true* phylogenetic groupings coming from *reliable* characters may be diluted by random or systematic errors from unreliable characters." As such, they suggested "that a combined analysis of potentially diverse data is inappropriate unless it is shown that the different data sets are not significantly heterogeneous with respect to the reconstruction model." de Queiroz et al. (1995, p. 659) were somewhat more ambivalent, but they also advocated the use of separate analyses "as a means of exploring possible disagreements among data sets," with the ultimate goal of identifying the source of the statistically significant conflict and correcting model assumptions prior to deriving phylogenetic inferences. We refer to this as the strong interpretation because incongruence among partitions is used to alter phylogenetic inferences directly. The statistical combinability of classes of evidence is usually "tested" with partition methods (e.g., Bull et al., 1993; Huelsenbeck et al., 1996a; Yoder et al., 2001; Barker and Lutzoni, 2002).

Numerous authors have endorsed the second argument for data exploration using partition methods, many of whom have also argued strongly for the superiority of simultaneous, total-evidence analysis. For example, although Nixon and Carpenter (1996b, p. 221) concluded unequivocally that simultaneous, total-evidence analysis is superior to the partition methods of taxonomic congruence, they nonetheless asserted that "[s]eparate analyses are useful and of interest to understanding the differences among data sets." Similarly, Remsen and DeSalle (1998, p. 233; see also DeSalle and Brower, 1997; Baker and DeSalle, 1997) contended that "a test for congruence between and among data partitions should always be performed, even if one intends to combine the data partitions from the start," because "without knowledge of the signal emanating from the various partitions, it will not be possible to diagnose particularly striking interactions among them." That is, although total evidence provides the severest test of competing phylogenetic hypotheses and maximizes explanatory power, analysis of congruence among partitions or between the partitions and the total evidence is supposed to lead to increased empirical knowledge. We refer to this as the weak interpretation because the results of partitioned analyses are not used to alter phylogenetic inferences directly.

One of the most common ways to explore data partitions is simply to inspect the results of separate and variously combined analyses. However, a wide variety of explicit methods are also used to assess precisely the degree of incongruence of data partitions; here, we review briefly several of the more popular methods. As indicated in footnote 2, above, partition incongruence has been judged with both topology-based measures, which assess differences in the branching patterns of trees obtained from separate analyses, and character-based measures, which quantify differences in the fit of data in combined and/or separate analyses. The topological incongruence test (Rodrigo et al., 1993) and the global congruence approach (Levasseur and Lapointe, 2001) are examples of topology-based methods, while all other partition methods that we deal with in any detail are examples of character-based methods.

The topological incongruence index of Rodrigo et al. (1993) is intended to assess the significance of the difference among data partitions. This index begins by calculating the symmetric distance (**SD**) between most-parsimonious trees from each data partition (see also Penny and Hendy, 1985b). The distribution of **SD** is then determined for a partition by calculating the mean **SD** between most-parsimonious trees obtained from bootstrap pseudoreplicates derived from that partition. There will be a wide distribution of **SD**s when there is a large degree of variation between the trees supported by each bootstrap. Like Templeton's test (see below), the topological incongruence index requires additional conventions when equally most-parsimonious trees exist.

The global congruence approach of Levasseur and Lapointe (2001) uses a method of calculating consensus branch lengths (the average consensus procedure of Lapointe and Cucumel, 1997) of character and taxonomic congruence results. Character sets are converted to distance matrices and phylogenetic trees are computed by the least squares method based on those distances.

The familiar contingency χ^2 test has been used as a goodness-of-fit statistical test for relative amounts of congruence among different classes of characters (e.g., Larson and Dimmick, 1993). Typically, a contingency table constitutes the basis for evaluating the congruence and incongruence among molecular and morphological characters, given a particular tree. As in a typical χ^2 test, marginal totals are determined from the observed values on the tree, from which expected frequencies are estimated.

Templeton's test assesses whether a most-parsimonious tree obtained from one data matrix of discrete characters is significantly less parsimonious than another (suboptimal) hypothesis of relationships (Templeton, 1983; see also Kishino and Hasegawa, 1989; Larson, 1994). This test has also been used to test whether trees obtained from discrete character data can significantly discriminate distance hypotheses (see also Rzhetsky and Nei, 1992). Templeton's test is a modification of the nonparametric Wilcoxon paired-sample (signed rank) test. The phylogenetic hypotheses in question are compared character-by-character to determine whether the number of steps required of each character differ on the competing historical propositions. The differences in the numbers of steps between the competing hypotheses are ranked, and the significance of these rankings, from random error, is evaluated with regard to the binomial distribution. A one-tailed test would seem to be appropriate, given that the relative optimality of the competing hypotheses is usually known a priori; however, the more conservative two-tailed test has been recommended (Felsenstein, 1985a).

The Mickevich–Farris incongruence index ($\mathrm{i_{MF}}$)[9] measures the scaled proportion of the total number of transformations in a data matrix, $\mathbf{x_T}$, that are due to the incongruence between partitions of that data matrix, $\mathbf{x_B}$ (Mickevich and Farris, 1981). Thus, $\mathbf{i_{MF}} = \mathbf{x_B}/\mathbf{x_T}$, where $\mathbf{x_B} = \mathbf{x_T} - \mathbf{x_W}$, and $\mathbf{x_W}$ is defined as the sum of the transformations calculated separately for each of the partitions. The number of transformations is determined only on most-parsimonious trees. Wheeler and Hayashi (1998; see also Wheeler et al., 2001, p. 128) provided a rescaled incongruence length difference index (RILD) that "does not exhibit the trivial minimum (0) as data set weights become increasingly disproportionate."

The Miyamoto incongruence index ($\mathbf{i_M}$) is the proportion of the total transformations in a data matrix, $\mathbf{x_T}$, that is due to the incongruence that occurs between partitions of that data matrix, $\mathbf{x_B}$ (M.M. Miyamoto, pers. comm., as reported in Kluge, 1989). This incongruence index is calculated in the same way as the Mickevich–Farris index (see above), except that $\mathbf{x_T}$ is the sum of the transformations required to explain the characters of one partition on the tree derived from the other partition and vice versa. As with the Mickevich–Farris index, the number of steps is determined only on most-parsimonious trees. The Miyamoto index does not always yield a proportion that seems reasonable, and it is quite sensitive to the distribution of congruent characters among the partitions (Swofford, 1991, p. 317). Moreover, an additional problem arises when two or more equally most-parsimonious trees result from the analysis of one, or both, of the partitions: more or fewer transformations may be attributed to a partition depending on which of the equally most-parsimonious trees is chosen. A further deficiency compared to the $\mathbf{i_{MF}}$ is that the $\mathbf{i_M}$ evaluates incongruence in reference to partitioned hypotheses and not to the total-evidence hypothesis (Kluge, 1989).

The incongruence length difference test (ILD, discordance test, Farris incongruence test, partition homogeneity test, Swofford's test) is currently the most widely used partition measure. The incongruence length difference is just the numerator of the $\mathbf{i_{MF}}$ ($\mathbf{x_B}$ in the above equation), which measures the number of transformations in a data matrix that are attributable to incongruence among partitions of that matrix, or the difference (**D**) between the length of the total-evidence hypothesis and the sum of lengths for each partitioned hypothesis (Farris et al., 1994b). The statistical test of that measure is accomplished by resampling (Farris et al., 1995). For example, two data matrices (**X** and **Y**) are incongruent when the sum of their most-parsimonious tree lengths (**L**) is shorter than that obtained from

[9] This is the metric defined by Kluge (1989); Dowton and Austin (2002, p. 20) referred to it as the "WILD" test and mistakenly attributed it to "Wheeler [and Hayashi, 1998]."

equivalent matrices, **P** and **Q**, obtained by randomly resampling the pooled original data sets. Thus, $(\mathbf{L_X} + \mathbf{L_Y}) < (\mathbf{L_P} + \mathbf{L_Q})$. The null distribution for the ILD test is obtained by regarding the observed matrices of **X** and **Y** characters as having been sampled at random from a single statistical population. On that assumption, any partition of the total $\mathbf{X} + \mathbf{Y}$ characters into two matrices of the same two sizes should be equally likely. The null distribution would then be determined by averaging over all possible partitions of the $\mathbf{X} + \mathbf{Y}$ characters into sets of sizes **X** and **Y**. In practice, to obtain a significance test it is necessary to compute **D** for only a small number of partitions, these being chosen at random from among those possible. Thus, the value of **D** is found for the partitions of the observed data matrix and for a number **W** of randomly selected partitions of the same sizes as the original partitions. If a number **S** of the **D** values from randomly selected partitions is smaller than the observed **D**, then the type I error rate (tail probability) of rejecting the null hypothesis is $1 - \mathbf{S}/(\mathbf{W} + 1)$. A 5% level of significance is indicated, for example, when $\mathbf{W} = 99$ and $\mathbf{S} = 95$.

According to Cunningham (1997), the ILD test can be recommended over the other tests of incongruence because of its (1) ease of calculation, (2) application to multiple data partitions simultaneously, (3) successive reapplications (e.g., to a data set that was unweighted and then weighted; but see Allard et al., 1999), and (4) effectiveness at discriminating significant data partition incongruence. However, recent studies involving simulation (Dolphin et al., 2000; Dowton and Austin, 2002; Barker and Lutzoni, 2002) and comparison to a "known" phylogeny (Yoder et al., 2001; but see Grant, 2002) have cast doubt on its effectiveness.

Partitioned Bremer support (partitioned branch support) aims to evaluate the distribution of evidential support for a particular clade (node) from different classes of data on the total-evidence hypothesis. Partitioned Bremer support is defined as the length (or mean length, if multiple most-parsimonious trees obtain) of a given data partition on the most-parsimonious tree(s) not containing a given clade minus the length (or mean length) of that partition on the total-evidence tree(s) (Baker and DeSalle, 1997). Gatesy et al. (1999; see also O'Grady et al., 2002, Table 1) defined several other related indexes, including hidden branch support, partitioned hidden branch support, hidden character support, hidden synapomorphy support, data set removal index, nodal data set influence, hidden nodal data set influence, and data set influence. All of these measures involve assessing clade stability in reference to data partitions. Reed and Sperling (1999) defined branch support as a function of partition weight ratios.

The presumed relevance of all of the above methods of data exploration is that significant incongruence suggests that some phenomenon caused the partitions to evolve differently. Under the strong interpretation (e.g., Bull et al., 1993; de Queiroz et al., 1995; Huelsenbeck et al., 1996a), significant incongruence provides clear evidence either of a violation of model assumptions or of different histories for the different partitions; in either case, equivalent treatment of all data in a simultaneous analysis is prohibited. Explicit in this interpretation (see quote above from Bull et al., 1993) is that incongruent partitions contain unreliable characters. Conversely, Nixon and Carpenter (1996b, p. 233) were careful to clarify that they "would advocate combination and simultaneous analysis even if the amount of incongruence is deemed 'significant'"; instead, they interpreted a large amount of incongruence heuristically as pointing to particular hypotheses that the investigator may wish to investigate. In a similar vein, a lack of significant incongruence among partitions has been taken to indicate increased support and increased confidence in the total-evidence hypothesis (e.g., Hillis, 1995). However, several problems compromise both interpretations.

The most obvious problem faced by both interpretations is the arbitrariness of the chosen partitions (Kluge and Wolf, 1993). As illustrated by Siddall (1997), patterns of congruence/incongruence depend crucially on the choice of partitions, yet there is a multitude of ways in which a data set can be partitioned, leading to contradictory conclusions and requiring arbitrary resolutions. Some have argued (e.g., Miyamoto and Fitch, 1995) that functional classes of data exist in nature (i.e., have discoverable boundaries) and that the different subsets of evidence, $\mathbf{e}_1|\mathbf{e}_2|\mathbf{e}_3|\cdots|\mathbf{e}_\mathrm{n}$, should not be analyzed simultaneously because they do not represent the same kind of evolutionary process. However, Siddall's examples included conflicting functional class partitions, demonstrating that some arbitrary decision as to which functional class should have precedence is required. Moreover, it has yet to be made clear why being of the same functional class or process partition is relevant to the inference of phylogeny, when the nature of the evidence in phylogenetic inference is the evolutionary event or transformation, for which there is the common currency of a unit of change from one state to another (Hennig, 1966, Fig. 21; Kluge and Wolf, 1993). As such, the decision not to combine or to introduce novel model assumptions or weighting schemes is also arbitrary, as is any inference(s) derived from that decision. Other authors (e.g., Nixon and Carpenter, 1996b, p. 225) have merely taken an unscientific, pragmatic position as to the reality of classes of data, arguing that the boundary between data sets exists as long as "we choose to recognize it"!

A further problem with the strong interpretation is that the inference of different histories (e.g., paralogy, lineage sorting, introgression, horizontal gene transfer, ancestral polymorphism, "gene trees *versus* species trees"), different evolutionary processes (e.g., evolutionary rates, selective constraints), or character unreliability (i.e., that they are misleading) is entirely ad hoc. No independent evidence of confounding processes is presented (DeSalle and Brower, 1997). Instead, this approach eliminates or down-weights contradictory evidence *solely* on the basis that it is contradictory. Moreover, those contradictory data are dealt with *en masse* and are not permitted to count fully against the resulting phylogenetic hypothesis. This interpretation relies on the assumption that data within a partition unanimously support the same phylogeny (i.e., share the same, contradictory signal), with discrepancy attributable to sampling error or random homoplasy. However, discovery that one partition is incongruent with another (or with the total-evidence tree) or that a partition has a negative partitioned Bremer support value at a given node does not deny that some of the characters in that partition (i.e., a subpartition) may strongly support that node nor does it indicate which characters in the partition are responsible for the conflict. As observed by DeSalle and Brower (1997, p. 759), tests of incongruence "cannot serve as criteria for determining if (and which) evidence should be deleted or downweighted."

Many workers hold the belief that hypotheses recovered in separate analysis of data partitions are better supported. For example, one of the foremost requirements for a phylogeny to be designated as "known" is that it be supported by multiple partitions (Miyamoto et al., 1994; Miyamoto and Fitch, 1995; Wiens, 1998, 2000a,b; Smith and Gutberlet, 2001; Buckley and Cunningham, 2002; see below). Likewise, Hillis (1995, p. 3; italics added) underscored the independence of data assigned to separate partitions, claiming that "congruence studies of multiple data sets can be used to assess the degree to which *independent* results agree and thus the minimum proportion of the findings that can be attributed to an underlying phylogeny" and further (p. 11) that

> Although a combined analysis of several data sets (assuming that they are appropriate for combining) may give the best estimate of phylogeny...the conclusion would be greatly strengthened if it were compatible with that of each of the individual data sets as well...

However, that argument for partitioning evidence loses sight of the fact that the important assumption of independence still obtains *within* each recognized partition (Kluge and Wolf, 1993). That is, as long as the assumption of independence applies to the characters within each partition, there is nothing more to be gained by also claiming independence among partitions.

Superficially, it would seem that the weak interpretation may have extensive heurism, a consideration that could establish the utility of partition methods of data exploration. Although partition incongruence is insufficient to claim discovery of different histories or evolutionary processes (see above), it is consistent with those phenomena and may therefore suggest interesting hypotheses worthy of independent testing. Indeed, Huelsenbeck et al. (1996b) suggested that discovery of different evolutionary processes and histories can be achieved only through the use of partitioned analyses. However, the heurism of partitioned analysis is illusory because the indication of particular hypotheses judged especially worthy of investigation derives from interaction of independent characters in a simultaneous analysis and not from a procedure that explicitly prohibits such interactions. It is epistemologically inconsistent to claim (e.g., Remsen and DeSalle, 1998; Gatesy et al., 1999; O'Grady et al., 2002) that the total-evidence analysis maximizes explanatory power—in part because characters interact synergistically to produce novel results—while also claiming that the less explanatory hypotheses of separate analyses are essential to detect interactions among the characters of different partitions. As discussed above, epistemology indicates that the least refuted, most highly corroborated hypothesis has the greatest heurism.

We do not mean to suggest that consideration of character partitions cannot be heuristic. Rather, our contention is that for such considerations to be truly heuristic, they must be based on the results of the total-evidence analysis. Indeed, we believe that there is great potential for the development of heuristic methods of a posteriori analysis of sets of characters. For example, the heurism of the character **ci** (Kluge and Farris, 1969) and **ri** (Farris, 1989a) on the total-evidence cladogram can be extended to partitions by averaging those values over each partition, as done by Källersjö et al. (1999). In addition to the epistemological strength of such a posteriori analysis of the total-evidence hypothesis, the practical advantage of this approach over separate analyses is that within-partition patterns can also be detected. For example, a large range of **ci** or **ri** values may indicate that the chosen partition did not contain a single, strong signal and that alternative partitions may have more heuristic value. Such a posteriori analysis could be refined further by plotting the distribution of those values against any variable of interest (e.g., alignment position, codon position, secondary structure, functional regions of translated proteins) and comparing within and among possible partitions. Moreover, provided that the necessary assumptions of the chosen test can be met, the statistical precision sought by many workers could be attained by performing multivariate analyses of variance or by simply employing such statistical tests as the χ^2 test of homogeneity. Similarly, past studies can indicate heuristically the expected "utility" of different character classes for resolving different phylogenetic questions (a concern when designing any phylogenetic study) by examining the transformations associated with different levels of divergence, and such expectations could even provide the basis for prior probabilities in a Bayesian framework.

The simultaneous parsimony analysis of the total, equally weighted evidence provides the severest possible test of competing phylogenetic hypotheses and identifies the hypothesis(es) of greatest explanatory power (Kluge, 1997, 1998). As such, we see no scientific or heuristic reason to favor separate analysis of arbitrary partitions over procedures that derive independently testable hypotheses from the patterns of character state transformations implied by the phylogenetic hypothesis of greatest explanatory power.

Congruence with an empirically "known" phylogeny. In this method of data exploration, a phylogeny that is "well-supported" (defined as a phylogenetic hypothesis supported by two or more data partitions; Miyamoto et al., 1994; Miyamoto and Fitch, 1995; Wiens, 1998, 2000a,b; Smith and Gutberlet, 2001), "well-corroborated" or "strong" (supported by quantitative analysis; Allard and Miyamoto, 1992; Marshall, 1992; Friedlander et al., 1996; Mindell and Thacker, 1996; Zardoya and Meyer, 1996; Cunningham, 1997; Hillis, 1999; Buckley and Cunningham, 2002), or "firmly established," "noncontroversial," "widely accepted," or "conservative" (no explicit operation or criteria employed to select the initial hypothesis; Friedlander et al., 1994; Graybeal, 1994; Russo et al., 1996; Zardoya and Meyer, 1996; Cunningham, 1997; Naylor and Brown, 1997, 1998; Ballard et al., 1998; Miya and Nishida, 2000; Posada and Crandall, 2001c) is designated as "known," "correct," or "expected." Data (or methods) that are congruent with that phylogeny are deemed to be of high quality and greater reliability, whereas those that are incongruent are of lower quality and are accordingly down-weighted or excluded.

Grant (2002) rejected this procedure as an empirical test of discovery operations, and it fails as a test of data quality for largely the same reasons. Most importantly, none of the proposed criteria is sufficient to justify conclusive acceptance of a hypothesis of relationships as "known," "correct," or "expected," so there is no reason to demand that new observations conform to previously supported hypotheses. Moreover, those new data are in fact potential falsifiers of the previous hypotheses, so judging data quality by how well they conform to those hypotheses results in a complete loss of independence. Likewise, this operation is not heuristic because it serves only to protect a preferred phylogenetic hypothesis from refutation by forcing new data into conformity.

Summary and conclusions

The current paradigm in phylogenetic systematics is clearly dominated by data exploration. In the above review, we identified over 20 approaches commonly used to explore data. However, it is equally clear that much more attention has been paid to the development and application of data exploration methods than to the critical evaluation of the scientific merits of those methods. As a result, many authors carry out elaborate, superficially impressive data exploration for no apparent reason. For example, McGuire and Bong Heang (2001) provided detailed descriptions of the procedures and results of extensive, technically sophisticated data exploration, but they made no attempt to explicate the significance of their procedures. Moreover, McGuire and Bong Heang actually dismissed the results of those analyses altogether on the grounds that previous studies compelled them to choose the $GTR + \Gamma + I$ maximum likelihood result as optimal for their data set, leaving the reader to wonder why such extensive data exploration was important enough to merit publication, but was also irrelevant to the inference of phylogeny.

In this paper we reviewed a wide variety of methods of data exploration in an attempt to understand their relevance to the science of phylogenetic systematics. We recognize three kinds of methods or operations:

- *Methods that perform empirical tests (i.e., discovery operations).*

Such methods are scientific, and, insofar as the results of scientific tests point to new or highly testable problems and hypotheses, they are also heuristic. Of the methods of data exploration that we examined, only character compatibility can be construed as an empirical test, although character congruence of phylogenetic parsimony is superior because it maximizes severity of test and explanatory power through simultaneous analysis of all critical evidence and the minimization of transformations.

- *Methods that are nonscientific but point to new or highly testable problems and hypotheses.*

Such methods are heuristic; i.e., they do not perform tests themselves, but they point to the weaker areas in our system of knowledge, thereby providing an indicator of the relative strength of evidential support and the expected fruitfulness of additional inquiry. Clear examples of heuristic approaches to data exploration include Bremer support (Bremer, 1988), long-branch extraction (Siddall and Whiting, 1999), and safe taxonomic reduction (Wilkinson, 1995), and we see considerable potential for the development of methods of a posteriori analysis of patterns of character transformations on the total-evidence phylogeny. Misunderstanding the data exploration method may hinder interpretation of heuristic results. For example, although most workers have employed safe taxonomic reduction and similar approaches to evaluate the effects of "missing data" and have provided empirical justification for excluding taxa, these methods actually evaluate the effects of including additional evidence and provide a heuristic justification for targeting certain taxa and characters in future rounds of testing.

- *Methods that are neither scientific nor heuristic.*

Such methods amount to mere sophistry and are irrelevant to phylogenetic inference. A disappointing number of the data exploration methods examined exemplify this category. For example, most methods of quality analysis function as data purification routines, whereby evidence is discarded or manipulated to make it conform with some notion of goodness. Such methods serve no purpose in the scientific enterprise, and their continued use seems to be more a function of systematists' fascination with the cult of impressive technicalities than any genuinely scientific concern (see epigraph, above).

The concept of support is central to data exploration. We propose an explicit concept of support, defined as the degree to which critical evidence refutes competing hypotheses. Our concept contrasts sharply with the verificationist interpretation of support as a measure of confidence, probability, or reliability. Instead, our concept of support is concerned with the relative degree of corroboration of competing hypotheses, and as Popper (1979, p. 18, italics in original) clarified,

> Being a report on past performance only, [degree of corroboration] has to do with a situation which may lead us to prefer some theories to others. *But it says nothing whatever about future performance, or about the 'reliability' of a theory.*

Furthermore, our concept of support is objective in that it focuses on the support of data for a hypothesis and therefore opposes the subjective concept of support commonly applied in sensitivity analysis (e.g., Wheeler, 1991, 1995; Flores-Villela et al., 2000; Nei and Kumar, 2000; Wheeler et al., 2001), whereby support is inferred from the effects of different assumptions about or interpretations of the data. Given that auxiliary assumptions are logically incapable of providing empirical support for hypotheses, the claimed support cannot be justified, except as a measure of subjective (and therefore relativistic) belief. However, justification by subjective belief is irrelevant to science. As Lakatos (1998, p. 21) observed,

> The cognitive value of a theory has nothing to do with its psychological influence on people's minds. Belief, commitment, understanding are states of the human mind. But the objective, scientific value of a theory is independent of the human mind which creates it or understands it. Its scientific value depends only on what objective support these conjectures have in facts.

According to our analysis, the most common interpretations of the results of data exploration in phylogenetic systematics are mistaken. Results of data exploration are primarily used to highlight strongly supported hypotheses as more accurate, reliable, or probably true and in effect protect those hypotheses from refutation by indicating that they are beyond additional testing. This misapplication of support is exemplified by common taxonomic practice, wherein strongly supported groups are recognized formally, while weakly supported groups remain nameless and are thus hidden, often allowing paraphyletic groups to be retained. Such formal recognition effectively protects those so-called reliable groups from future refutation by fiat, i.e., by imposing legally the principle of stability, while the groups that are especially interesting scientifically are simply ignored. This practice is generally defended in the interest of "conservatism," but we fail to see how this justifies overturning empirical evidence. Moreover, scientifically, the most conservative taxonomy is the one that strays least from available evidence (D.R. Frost, pers. comm.). Instead of drawing attention to strongly supported clades, we suggest that methods of data exploration be used to further the goals of science by highlighting *weakly* supported hypotheses by indicating cases in which choice among competing hypotheses is ambiguous or hypotheses have been less severely tested (tests have been less decisive), and, therefore, scientific inquiry aimed at them is likely to be more fruitful.

Also of concern is the emphasis placed on data exploration in empirical phylogenetic studies. Given that the only legitimate role of most methods of data exploration is heuristic, not scientific, the current emphasis on these methods is unwarranted. Empirical tests are the only source of scientific knowledge, and the science of phylogenetic systematics would be better served by emphasizing that fact in publications and when considering proposals for funding. Of course, pointing out the weaker areas in our system of knowledge is important and should not be abandoned altogether; but, in our judgment, the resources and interest currently devoted to data exploration are grossly disproportionate to their cognitive worth. Consequently, despite the current popularity of data exploration techniques and the accompanying social pressures to include them in published studies, we urge phylogeneticists not to accept these methods uncritically, but to consider their cognitive merits with regard to the logic of scientific discovery and to use them accordingly.

Acknowledgments

This paper grew out of a tag-team presentation given at the 2001 S'gebengu Symposium at the Cladistics Institute, Harbor Springs, Michigan, and was refined during the follow-up Plataforma Workshop held in New York. We are grateful to our colleagues for their willingness to discuss disagreements and their patience in critically reading and improving drafts of the manuscript; in particular, we thank Paulo Buckup, Julián Faivovich, Darrel Frost, Gonzalo Giribet, Maureen Kearney, Jyrki Muona, Diego Pol, Leo Smith, Mark Siddall, Ward Wheeler, and two anonymous reviewers. All errors that remain are ours alone. T.G. was supported by an AMNH Graduate Student Fellowship, a Columbia University/CERC Faculty Fellowship, and NSF DEB 0309226.

References

Adams III, E.N., 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. 21, 390–397.

Allard, M.W., Carpenter, J.M., 1996. On weighting and congruence. Cladistics 12, 183–198.

Allard, M.W., Farris, J.S., Carpenter, J.M., 1999. Congruence among mammalian mitochondrial genes. Cladistics 15, 75–84.

Allard, M.W., Miyamoto, M.M., 1992. Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. Mol. Biol. Evol. 9, 778–786.

Anderson, J.S., 2001. The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the Lepospondyli (Vertebrata, Tetrapoda). Syst. Biol. 50, 170–193.

Archie, J.W., 1989a. Phylogenies of plant families: a demonstration of phylogenetic randomness in DNA sequence data derived from proteins. Evolution 43, 1796–1800.

Archie, J.W., 1989b. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38, 239–252.

Arnedo, M.A., Oromí, P., Ribera, C., 2002. Radiation of the spider genus *Dysdera* (Aranae, Dysderidae) in the Canary Islands: cladistic assessment based on multiple data sets. Cladistics 17, 313–353.

Asher, R.J., 1999. A morphological basis for assessing the phylogeny of the "Tenrecoidea" (Mammalia, Lipotyphla). Cladistics 15, 231–252.

Austin, J.D., Lougheed, S.C., Tanner, K., Chek, A.A., Bogart, J.P., Boag, P.T., 2002. A molecular perspective on the evolutionary affinities of an enigmatic neotropical frog, *Allophryne ruthveni*. Zool. J. Linnean Soc. 134, 335–346.

Baird, B.F., 1989. Managerial Decisions under Uncertainty: An Introduction to the Analysis of Decision Making. Wiley, New York.

Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46, 654–673.

Ballard, J.W.O., Olsen, G.J., Faith, D.P., Odgers, W.A., Rowell, D.M., Atkinson, P.W., 1992. Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. Science 258, 1345–1348.

Ballard, J.W.O., Thayer, M.K., Newton Jr., A.F., Grismer, E.R., 1998. Data sets, partitions, and characters: philosophies and procedures for analyzing multiple data sets. Syst. Biol. 47, 367–396.

Barker, F.K., Lanyon, S.M., 2000. The impact of parsimony weighting schemes on inferred relationships among toucans and Neotropical barbets (Aves: Piciformes). Mol. Phylogenet. Evol. 15, 215–234.

Barker, F.K., Lutzoni, F.M., 2002. The utility of the incongruence length difference test. Syst. Biol. 51, 625–637.

Barkman, T.J., Chenery, G., McNeal, J.R., Lyons-Weiler, J., Ellisens, W.J., Moore, G., Wolfe, A.D., dePamphilis, C.W., 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl. Acad. Sci. USA 97, 13166–13171.

Berlocher, S.H., Swofford, D.L., 1997. Searching for phylogenetic trees under the frequency parsimony criterion: an approximation using generalized parsimony. Syst. Biol. 46, 211–215.

Braun, D., 1998. The role of funding agencies in the cognitive development of science. Res. Policy 27, 807–821.

Bremer, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42, 795–803.

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

Brower, A.V.Z., 2000. Evolution is not a necessary assumption of cladistics. Cladistics 16, 143–154.

Brown, J.K.M., 1994. Bootstrap hypothesis tests for evolutionary trees and other dendrograms. Proc. Natl. Acad. Sci. USA 91, 12293–12297.

Brown, W.M., Prager, E.M., Wang, A., 1982. Mitochondrial DNA sequences of primates, tempo and mode of evolution. J. Mol. Evol. 18, 225–239.

Buckley, T.R., Cunningham, C.W., 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. Evol. 19, 394–405.

Buckley, T.R., Simon, C., Chambers, G.K., 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50, 67–86.

Buckup, P.A., Dyer, B.S., 1991. Transformation series analysis (TSA) is dependent on initial order of character states. Syst. Zool. 40, 500–502.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384–397.

Bunge, M., 1998. Philosophy of Science: From Explanation to Justification. Transaction Publishers, New Brunswick, NJ.

Burbrink, F.T., Lawson, R., Slowinski, J.B., 2000. Mitochondrial DNA phylogeography of the polytypic North american rat snake (*Elaphe obsoleta*): a critique of the subspecies concept. Evolution 54, 2107–2118.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York.

Campbell, J.A., Frost, D.R., 1993. Anguid lizards of the genus *Abronia*: revisionary notes, descriptions of new species, a phylogenetic analysis, and key. Bull. Am. Mus. Nat. Hist. 216, 1–121.

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

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

Carpenter, J.M., 1994. Successive weighting, reliability, and evidence. Cladistics 10, 177–181.

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

Carranza, S., Arnold, E.N., Mateo, J.A., Geniez, P., 2002. Relationships and evolution of the North African geckos, *Geckonia* and *Tarentola* (Reptilia: Gekkonidae), based on mitochondrial and nuclear DNA sequences. Mol. Phylogenet. Evol. 23, 244–256.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y., Kron, K.A., Rettig, J.H., Conti, E., Palmer, J.D., Manhart, J.R., Sytsma, K.J., Michaels, H.J., Kress, W.J., Karol, K.G., Clark, W.D., Hedren, M., Gaut, B.S., Jansen, R.K., Kim, K., Wimpee, C.F., Smith, J.F., Furnier, G.R., Strauss, S.H., Xiang, Q., Plunkett, G.M., Soltis, P.S., Swensen, S.M., Williams, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn, G.H.J., Graham, S.W., Barrett, S.C.H., Dayanandan, S., Albert, V.A., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann. Missouri Bot. Gard. 80, 528–580.

Chavarría, G., Carpenter, J.M., 1994. "Total evidence" and the evolution of highly social bees. Cladistics 10, 229–258.

Chek, A.A., Lougheed, S.C., Bogart, J.P., Boag, P.T., 2001. Perception and history: molecular phylogeny of a diverse group of frogs, the 30-chromosome *Hyla* (Anura: Hylidae). Mol. Phylogenet. Evol. 18, 370–385.

Chippindale, P.T., Wiens, J.J., 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. Syst. Biol. 43, 278–287.

Cicero, C., Johnson, N.K., 2001. Higher-level phylogeny of new world vireos (Aves: Vireonidae) based on sequences of multiple mitochondrial DNA genes. Mol. Phylogenet. Evol. 20, 27–40.

Cicero, C., Johnson, N.K., 2002. Phylogeny and character evolution in the *Empidonax* group of tyrant flycatchers (Aves: Tyrannidae): a test of W.E. Lanyon's hypothesis using mtDNA sequences. Mol. Phylogenet. Evol. 22, 289–302.

Colless, D.H., 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. Syst. Zool. 29, 288–299.

Cracraft, J., Helm-Bychowski, K., 1991. Parsimony and phylogenetic inference using DNA sequences: some methodological strategies. In: Miyamoto, M.M. and Cracraft, J. (Eds.), Phylogenetic analysis of DNA Sequences. Oxford University Press, New York, pp. 184–220.

Crandall, K.A., Fitzpatrick Jr., J.F., 1996. Crayfish molecular systematics: using a combination of procedures to estimate phylogeny. Syst. Biol. 45, 1–26.

Cunningham, C., 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14, 733–740.

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

Davis, J.I., 1993. Character removal as a means for assessing stability of clades. Cladistics 9, 201–210.

Davis, J.I., Frohlich, M.W., Soreng, R.J., 1993. Cladistic characters and cladogram stability. Syst. Bot. 18, 188–196.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26, 657–681.

de Queiroz, K., 1987. Phylogenetic systematics of iguanine lizards: a comparative osteological study. Univ. Calif. Publ. Zool., 118.

DeSalle, R., Brower, A.V.Z., 1997. Process partitions, congruence, and the independence of characters: inferring relationships among closely related Hawaiian *Drosophila* from multiple gene regions. Syst. Biol. 46, 751–764.

Dolphin, K., Belshaw, R., Orme, C.D.L., Quicke, D.L.J., 2000. Noise and incongruence: interpreting results of the incongruence length difference test. Mol. Phylogenet. Evol. 17, 401–406.

Donoghue, M.J., 1994. Progress and prospects in reconstructing plant phylogeny. Ann. Missouri Bot. Gard. 81, 405–418.

Donoghue, M.J., Ackerly, D.D., 1996. Phylogenetic uncertainties and sensitivity analyses in comparative biology. Phil. Trans. R. Soc. Lond. B 351, 1241–1249.

Donoghue, M.J., Doyle, J.A., Gauthier, J., Kluge, A.G., Rowe, T., 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20, 431–460.

Dowton, M., Austin, A.D., 2002. Increased incongruence does not necessarily indicate increased phylogenetic accuracy—the behavior of the incongruence length difference test in mixed-model analysis. Syst. Biol. 51, 19–31.

Duffels, J.P., Turner, H., 2002. Cladistic analysis and biogeography of the cicadas of the Indo-Pacific subtribe Cosmopsaltriina (Hemiptera: Cicadoidea: Cicadidae). Syst. Entomol. 27, 235–261.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge.

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

Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93, 7085–7090.

Faith, D.P., 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40, 366–375.

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

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

Faith, D.P., Trueman, J.W.H., 2001. Towards an inclusive philosophy for phylogenetic inference. Syst. Biol. 50, 331–350.

Faivovich, J., 2002. On RASA. Cladistics 18, 324–333.

Farris, J.S., 1967. The meaning of relationship and taxonomic procedure. Syst. Zool. 16, 44–51.

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

Farris, J.S., 1971. The hypothesis of nonspecificity and taxonomic congruence. Annu. Rev. Ecol. Syst. 2, 227–302.

Farris, J.S., 1973a. On comparing the shapes of taxonomic trees. Syst. Zool. 22, 50–54.

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

Farris, J.S., 1982. Simplicity and informativeness in systematics and phylogeny. Syst. Zool. 31, 413–444.

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

Farris, J.S., 1986. On the boundaries of phylogenetic systematics. Cladistics 2, 14–27.

Farris, J.S., 1989a. The retention index and homoplasy excess. Syst. Zool. 38, 406–407.

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

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

Farris, J.S., 1998. The future of phylogeny reconstruction. Zool. Scripta 26, 303–311.

Farris, J.S., 1999. Likelihood and inconsistency. Cladistics 15, 199–204.

Farris, J.S., 2002a. RASA attributes highly significant structure to randomized data. Cladistics 18, 334–353.

Farris, J.S., 2002b. Support weighting. Cladistics 17, 389–394.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Farris, J.S., Källersjö, M., De Laet, J., 2001. Branch lengths do not indicate support—even in maximum likelihood. Cladistics 17, 298–299.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994a. Permutations. Cladistics 10, 65–76.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994b. Testing significance of incongruence. Cladistics 10, 315–319.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995. Constructing a significance test for incongruence. Syst. Biol. 44, 570–572.

Farris, J.S., Kluge, A.G., Eckardt, M.J., 1970. On predictivity and efficiency. Syst. Zool. 19, 363–372.

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

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

Felsenstein, J., 1983. Methods for inferring phylogenies: A statistical view. In: Felsenstein, J. (Ed.), Numerical Taxonomy. Springer-Verlag, New York, pp. 315–334.

Felsenstein, J., 1985a. Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34, 152–161.

Felsenstein, J., 1985b. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22, 521–565.

Felsenstein, J., 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.R., Rohlf, F.J., 1969. Robustness of numerical taxonomic methods and errors in homology. Syst. Zool. 18, 33–36.

Fitch, W.M., 1979. Cautionary remarks on using gene expression events in parsimony procedures. Syst. Zool. 28, 375–379.

Flores-Villela, O., Kjer, K.M., Benabib, M., Sites Jr., J.W., 2000. Multiple data sets, congruence, and hypothesis testing for the phylogeny of basal groups of the lizard genus *Sceloporus* (Squamata, Phrynosomatidae). Syst. Biol. 49, 713–739.

Floyd, J.W., 2002. Phylogenetic and biogeographic patterns in *Gaylussacia* (Ericaceae) based on morphological, nuclear DNA, and chloroplast DNA variation. Syst. Bot. 27, 99–115.

Friedlander, T.P., Regier, J.C., Mitter, C., 1994. Phylogenetic information content of five nuclear gene sequences in animals: Initial assessment of character sets from concordance and divergence studies. Syst. Biol. 43, 511–525.

Friedlander, T.P., Regier, J.C., Mitter, C., Wagner, D., 1996. A nuclear gene for higher level phylogenetics: phosphoenolpyruvate carboxykinase tracks Mesozoic-age divergences within Lepidoptera (Insecta). Mol. Biol. Evol. 13, 594–604.

Frost, D.R., Etheridge, R., Janies, D., Titus, T.A., 2001a. Total evidence, sequence alignment, evolution of polychrotid lizards, and a reclassification of the Iguania (Squamata: Iguania). Am. Mus. Novit. 3343, 1–38.

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

Frost, D.R., Rodrigues, M.T., Grant, T., Titus, T.A., 2001b. Phylogenetics of the lizard genus *Tropidurus* (Squamata: Tropiduridae: Tropidurinae): direct optimization, descriptive efficiency, and sensitivity analysis of congruence between molecular data and morphology. Mol. Phylogenet. Evol. 21, 352–371.

Fuller, S., 1993. Philosophy of Science and its Discontents. The Guilford Press, New York.

Gahn, F.J., Kammer, T.W., 2002. The cladid crinoid *Barycrinus* from the Burlington Limestone (early Osagean) and the phylogenetics of Mississippian botryocrinids. J. Paleont. 76, 123–133.

Gao, K.-Q., Norell, M.A., 1998. Taxonomic revision of *Carusia* (Reptilia: Squamata) from the Late Cretaceous of the Gobi Desert and phylogenetic relationships of anguimorphan lizards. Am. Mus. Novit. 3230, 1–51.

Gardiner, B.G., 1982. Tetrapod classification. Zool. J. Linnean Soc. 74, 207–232.

Gatesy, J., O'Grady, P., Baker, R.H., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15, 271–313.

Gattei, S., 2002. The positive power of negative thinking. Cladistics 18, 446–452.

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

Geiger, D.L., 2002. Stretch coding and block coding: two new strategies to represent questionably aligned DNA sequences. J. Mol. Evol. 54, 191–199.

Giribet, G., Distel, D.L., Polz, M., Sterrer, W., Wheeler, W.C., 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. Syst. Biol. 49, 539–562.

Giribet, G., Edgecombe, G.D., Wheeler, W.C., 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature 413, 157–161.

Giribet, G., Edgecombe, G.D., Wheeler, W.C., Babbit, C., 2002. Phylogeny and systematic position of Opiliones: a combined analysis of chelicerate relationships using morphological and molecular data. Cladistics 18, 5–70.

Giribet, G., Wheeler, W.C., 1999. On gaps. Mol. Phylogenet. Evol. 13, 132–143.

Giribet, G., Wheeler, W.C., 2002. On bivalve phylogeny: a high-level analysis of the Bivalvia (Mullusca) based on combined morphology and DNA sequences. Invertebr. Biol. 121, 271–324.

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

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

Goldman, N., Whelan, A.S., 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. 17, 975–978.

Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Grant, T., 2002. Testing methods: the evaluation of discovery operations in evolutionary biology. Cladistics 18, 94–111.

Graybeal, A., 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. 43, 174–193.

Hacking, I., 1965. The Logic of Statistical Inference. Cambridge University Press, Cambridge.

Hall, J.S., Adams, B., Parsons, T.J., French, R., Lane, L.C., Jensen, S.G., 1998. Molecular cloning, sequencing, and phylogenetic relationships of a new potyvirus: sugarcane streak mosaic virus, and a reevaluation of the classification of the Potyviridae. Mol. Phylogenet. Evol. 10, 323–332.

Harshman, J., 1994. The effect of irrelevant characters on bootstrap values. Syst. Biol. 43, 419–424.

Hendy, M.D., 1989. The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. 38, 310–321.

Hendy, M.D., Charleston, M.A., 1993. Hadamard conjugation: a versatile tool for modelling nucleotide sequence evolution. New Zealand J. Bot. 31, 231–237.

Hendy, M.D., Penny, D., 1993. Spectral analysis of phylogenetic data. J. Classif. 10, 5–24.

Hendy, M.D., Penny, D., Steel, M.A., 1994. A discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA 91, 3339–3343.

Hennig, W., 1966. Phylogenetic Systematics. University of Chicago, Chicago.

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

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

Hillis, D.M., 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. USA 96, 9979–9981.

Hillis, D.M., Huelsenbeck, J.P., 1992. Signal, noise, and reliability in molecular phylogenetic analysis. J. Hered. 83, 189–195.

Huelsenbeck, J.P., 1991. Tree-length distribution skewness: an indicator of phylogenetic information. Syst. Zool. 40, 257–270.

Huelsenbeck, J.P., 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46, 69–74.

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

Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996b. Combining data in phylogenetic analysis: reply from J.P. Huelsenbeck, J.J. Bull, and C.W. Cunningham. Trends Ecol. Evol. 11, 335.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28, 437–466.

Huelsenbeck, J.P., Hillis, D.M., Jones, R., 1996c. Parametric bootstrapping in molecular phylogenetics: applications and performance. In: Ferraris, J.D., Palumbi, S.R. (Eds.), Molecular Zoology: Advances, Strategies, and Protocols. Wiley-Liss Inc, New York, pp. 19–45.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.

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

Hutchinson, M.N., Donnellan, S.C., 1992. Taxonomy and genetic variation in the Australian lizards of the genus *Pseudemoia* (Scincidae: Lygosominae). J. Nat. Hist 26, 215–264.

Jackman, T.R., Larson, A., de Queiroz, K., Losos, J.B., 1999. Phylogenetic relationships and tempo of early diversification in *Anolis* lizards. Syst. Biol. 48, 254–285.

Janies, D., 2001. Phylogenetic relationships of extant echinoderm classes. Can. J. Zool. 79, 1232–1250.

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.

Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy *increases* phylogenetic structure. Cladistics 15, 91–93.

Källersjö, M., Farris, J.S., 1998. Recent advances in large-scale plant phylogenetic studies. In: Nordenstam, B., El-Ghazaly, G. and Kassas, M. (Eds.), Plant Systematics for the 21st Century. Portland Press, London, pp. 59–63.

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

Kearney, M., 1998. Systematics of the amphisbaenian family Rhineuridae: missing data and resolution. J. Vert. Paleo. 18, 55A.

Kearney, M., 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51, 369–381.

Kelsey, C.R., Crandall, K.A., Voevodin, A.F., 1999. Different models, different trees: the geographic origin of PTLV-I. Mol. Phylogenet. Evol. 13, 336–347.

Kim, J., 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. 42, 331–340.

Kim, J., 2000. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. Mol. Phylogenet. Evol. 17, 58–75.

Kimura, M., 1955. Random genetic drift in multi-allelic locus. Evolution 9, 419–435.

Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78, 454–458.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of Hominoidea. J. Mol. Evol. 29, 170–179.

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., 1991. Boine snake phylogeny and research cycles. Misc. Publ. Mus. Zool. Univ. Michigan 178, 1–58.

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

Kluge, A.G., 1998. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic analysis. Zool. Scripta 26, 349–360.

Kluge, A.G., 1999. The science of phylogenetic systematics: explanation, prediction, and test. Cladistics 14, 151–158.

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

Kluge, A.G., 2002. Distinguishing "or" from "and" and the case for historical identification. Cladistics 18, 585–593.

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

Kluge, A.G., Farris, J.S., 1999. Taxic homology = overall similarity. Cladistics 15, 205–212.

Kluge, A.G., Wolf, A.J., 1993. Cladistics: what's in a word? Cladistics 9, 183–199.

Kornet, D.J., Turner, H., 1999. Coding polymorphism for phylogeny reconstruction. Syst. Biol. 48, 365–379.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Kuhn, T.S., 1977. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press, Chicago.

Lakatos, I., 1978. The Methodology of Scientific Research Programmes. Cambridge University Press, Cambridge, UK.

Lakatos, I., 1998. Science and pseudoscience. In: Curd, M. and Cover, J.A. (Eds.), Philosophy of Science: The Central Issues. W.W. Norton & Company, New York, pp. 20–26.

Lamb, T., Bauer, A.M., 2002. Phylogenetic relationships of the large-bodied members of the African lizard genus *Pachydactylus* (Reptilia: Gekkonidae). Copeia 2002, 586–596.

Lanyon, S.M., 1985. Detecting internal inconsistencies in distance data. Syst. Zool. 34, 397–403.

Lapointe, F.J., Cucumel, G., 1997. The average weighted consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46, 306–312.

Larson, A., 1994. The comparison of morphological and molecular data in phylogenetic systematics. In: Schierwater, B., Streit, B., Wagner, G.P. and DeSalle, R. (Eds.), Molecular Ecology and Evolution: Approaches and Applications. Birkhauser Verlag, Basel, pp. 371–390.

Larson, A., Dimmick, W.W., 1993. Phylogenetic relationships of the salamander families: an analysis of congruence among morphological and molecular characters. Herpetol. Monogr. 7, 77–93.

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

Le Quesne, W., 1989. The normal deviate test of phylogenetic value of a data matrix. Syst. Zool. 38, 51–54.

Lento, G.M., Hickson, R.E., Chambers, G.K., Penny, D., 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. 12, 28–52.

Levasseur, C., Lapointe, F.-J., 2001. War and peace in phylogenetics: a rejoinder on total evidence and consensus. Syst. Biol. 50, 881–891.

Lipscomb, D., 1990. Two methods for calculating cladogram characters: transformation series analysis and the iterative FIG/FOG method. Syst. Zool. 39, 277–288.

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

Lyons-Weiler, J., Hoelzer, G., 1999. Null model selection, compositional bias, character state bias, and the limits of phylogenetic information. Mol. Biol. Evol. 16, 1400–1405.

Lyons-Weiler, J., Hoelzer, G.A., 1997. Escaping from the Felsenstein Zone by detecting long branches in phylogenetic data. Mol. Phylogenet. Evol. 8, 375–384.

Lyons-Weiler, J., Hoelzer, G.A., Tausch, R.J., 1996. Relative apparent synapomorphy analysis (RASA) I: the statistical measurement of phylogenetic signal. Mol. Biol. Evol. 13, 749–757.

Lyons-Weiler, J., Hoelzer, G.A., Tausch, R.J. Optimal outgroup analysis. Biol. J. Linn. Soc. 64, 493–511.

Lyons-Weiler, J., Milinkovitch, M., 1997. A phylogenetic approach to the problem of differential lineage sorting. Mol. Biol. Evol. 14, 968–975.

Marshall, C., 1992. Substitution bias, weighted parsimony, and amniote phylogeny as inferred from 18S rRNA sequences. Mol. Biol. Evol. 9, 370–377.

McGuire, J.A., Bong Heang, K., 2001. Phylogenetic systematics of Southeast Asian flying lizards (Iguania: Agamidae: *Draco*) as inferred from mitochondrial DNA sequence data. Biol. J. Linn. Soc. 72, 203–229.

Mickevich, M.F., 1978. Taxonomic congruence. Syst. Zool. 27, 143–158.

Mickevich, M.F., 1980. Taxonomic congruence: Rohlf and Sokal's misunderstanding. Syst. Zool. 29, 162–176.

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

Mickevich, M.F., Farris, J.S., 1981. The implications of congruence in *Menidia*. Syst. Zool. 30, 351–370.

Mickevich, M.F., Johnson, M.S., 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. Syst. Zool. 25, 260–270.

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

Mickevich, M.F., Weller, S.J., 1990. Phylogenetic character analysis: tracing character evolution on a cladogram. Cladistics 6, 137–170.

Milinkovitch, M.C., Lyons-Weiler, J., 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogenet. Evol. 9, 348–357.

Mindell, D., 1991. Similarity and congruence as criteria for molecular homology. Mol. Biol. Evol. 8, 897–900.

Mindell, D.P., Honeycutt, R.L., 1990. Ribosomal RNA in vertebrates: evolution and phylogenetic applications. Annu. Rev. Ecol. Syst. 21, 541–566.

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

Miya, M., Nishida, M., 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. 17, 437–455.

Miyamoto, M.M., Allard, M.W., Adkins, R.M., Janacek, L.L., Honeycutt, R.L., 1994. A congruence test of reliability using linked mitochondrial DNA sequences. Syst. Biol. 43, 236–249.

Miyamoto, M.M., Cracraft, J., 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. In: Miyamoto, M.M. and Cracraft, J. (Eds.), Phylogenetic Analysis of DNA Sequences. Oxford University Press, New York, pp. 3–17.

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

Mueller, L.D., Ayala, F.J., 1982. Estimation and interpretation of genetic distance in empirical studies. Genet. Res. 40, 127–137.

Murphy, R.W., 1993. The phylogenetic analysis of allozyme data: invalidity of coding alleles by presence/absence and recommended procedures. Biochem. Syst. Ecol. 21, 25–38.

Naylor, G.J.P., Brown, W.M., 1997. Structural biology and phylogenetic estimation. Nature 388, 527–528.

Naylor, G.J.P., Brown, W.M., 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. 47, 61–76.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Nelson, G.J., 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's *Familles des Plantes* (1763–1764). Syst. Zool. 28, 1–21.

Nickles, T., 2000. Discovery. In: Newton-Smith, W.H. (Ed.), A Companion to the Philosophy of Science. Blackwell Publishers Ltd, Oxford, pp. 85–96.

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., Carpenter, J.M., 1996a. On consensus, collapsibility, and clade concordance. Cladistics 12, 305–321.

Nixon, K.C., Carpenter, J.M., 1996b. On simultaneous analysis. Cladistics 12, 221–241.

Nixon, K.C., Davis, J.I., 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics 7, 233–241.

Nixon, K.C., Wheeler, Q.D., 1992. Extinction and the origin of species. In: Novacek, M.J. and Wheeler, Q.D. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 119–143.

Noreen, E.W., 1989. Computer-Intensive Methods for Testing Hypotheses: An Introduction. John Wiley and Sons, New York.

Novacek, M.J., 1992a. Fossils as critical data for phylogeny. In: Novacek, M.J. and Wheeler, Q.D. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 46–88.

Novacek, M.J., 1992b. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Syst. Biol. 41, 58–73.

O'Grady, R.T., Remsen, J., Gatesy, J., 2002. Partitioning of multiple data sets in phylogenetic analysis. In: DeSalle, R., Giribet, G., Wheeler, W.C. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser Verlag, Basel, Switzerland, pp. 102–119.

O'Leary, M.A., 1999. Parsimony analysis of total evidence from extinct and extant taxa and the cetacean-artiodactyl question (Mammalia, Ungulata). Cladistics 15, 315–330.

O'Leary, M.A., Geisler, J.H., 1999. The position of Cetacea within Mammalia: phylogenetic analysis of morphological data from extinct and extant taxa. Syst. Biol. 48, 455–490.

Ota, R., Waddell, P.J., Hasegawa, M., Shimodaira, H., Kishino, H., 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. 17, 798–803.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. Nature 401, 877–884.

Pannell, D.J., 1997. Introduction to Practical Linear Programming. John Wiley and Sons, New York.

Patterson, C., 1981. Significance of fossils in determining evolutionary relationships. Annu. Rev. Ecol. Syst. 12, 195–223.

Penny, D., Hendy, M.D., 1985a. Testing methods of evolutionary tree construction. Cladistics 1, 266–278.

Penny, D., Hendy, M.D., 1985b. The use of tree comparison metrics. Syst. Zool. 34, 75–82.

Penny, D., Hendy, M.D., 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3, 403–417.

Penny, D., Hendy, M.D., Lockhart, P.J., Steel, M.A., 1996. Corrected parsimony, minimum evolution, and Hadamard conjugations. Syst. Biol. 45, 596–606.

Penny, D., Hendy, M.D., Zimmer, E.A., Hamby, R.K., 1990. Trees from sequences: Panacea or Pandora's box? Aust. Syst. Bot. 3, 21–38.

Penny, D., Watson, E.E., Hickson, R.E., Lockhart, P.J., 1993. Some recent progress with methods for evolutionary trees. New Zealand J. Bot. 31, 275–288.

Phillips, A., Janies, D., Wheeler, W.C., 2000. Multiple sequence alignment in phylogenetic analysis. Mol. Phylogenet. Evol. 16, 317–330.

Poe, S., 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47, 18–31.

Pogue, M.G., Mickevich, M.F., 1990. Character definitions and character state delineation: the bête noire of phylogenetic inference. Cladistics 6, 319–361.

Pol, D., Siddall, M., 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17, 266–281.

Popper, K.R., 1959. The Logic of Scientific Discovery [1992 reprint of 1968 edition]. Routledge, London.

Popper, K.R., 1979. Objective Knowledge: An Evolutionary Approach. Oxford University Press, New York.

Popper, K.R., 1983. Realism and the Aim of Science. Routledge, London.

Popper, K.R., 1990. A World of Propensities. Thoemmes, Bristol.

Posada, D., Crandall, K.A., 2001a. Selecting models of nucleotide substitution: an application to Human Immunodeficiency Virus 1 (HIV-1). Mol. Biol. Evol. 18, 897–906.

Posada, D., Crandall, K.A., 2001b. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50, 580–601.

Posada, D., Crandall, K.A., 2001c. Simple (wrong) models for complex trees: a case from Retroviridae. Mol. Biol. Evol. 18, 271–275.

Prager, E., Wilson, A., 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J. Mol. Evol. 27, 326–335.

Prendini, L., 2000. Phylogeny and classification of the superfamily Scorpionoidea Latrelle 1802 (Chelicerata, Scorpiones): an exemplar approach. Cladistics 16, 1–78.

Reed, R.D., Sperling, F.A., 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 16, 286–297.

Reeder, T.W., Montanucci, R.R., 2001. Phylogenetic analysis of the horned lizards (Phrynosomatidae: *Phrynosoma*): evidence from mitochondrial DNA and morphology. Copeia 2001, 309–323.

Remsen, J., DeSalle, R., 1998. Character congruence of multiple data partitions and the origin of the Hawaiian Drosophilidae. Mol. Phylogenet. Evol. 9, 225–235.

Resnik, D.B., 2001. Financial interests and research bias. Perspect. Sci. 8, 255–285.

Rice, K.A., Donoghue, M.J., Olmstead, R.G., 1997. Analyzing large data sets: *rbc*L 500 revisited. Syst. Biol. 46, 554–563.

Rodrigo, A.G., Kelly-Borges, M., Bergquist, P.R., Bergquist, P.L., 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric tree. New Zealand J. Bot. 31, 257–268.

Rohlf, F.J., 1963. Congruence of larval and adult classifications in *Aedes* (Diptera: Culicidae). Syst. Zool. 12, 97–117.

Rohlf, F.J., 1965. A randomization test of the nonspecificity hypothesis in numerical taxonomy. Taxon 14, 262–267.

Rohlf, F.J., 1974. Methods of comparing classifications. Annu. Rev. Ecol. Syst. 5, 101–113.

Rohlf, F.J., 1982. Consensus indices for comparing classifications. Math. Biosci. 59, 131–144.

Rohlf, F.J., Sokal, R.R., 1965. Coefficients of correlation and distance in numerical taxonomy. Univ. Kansas Sci. Bull. 45, 3–27.

Rohlf, F.J., Sokal, R.R., 1980. Comments on taxonomic congruence. Syst. Zool. 29, 97–101.

Rohlf, F.J., Sokal, R.R., 1981. Comparing numerical taxonomic studies. Syst. Zool. 30, 459–490.

Russo, C.A., Takezaki, N., Nei, M., 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13, 525–536.

Rzhetsky, A., Nei, M., 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum evolution methods of phylogenetic inference. J. Mol. Evol. 35, 367–375.

Salducci, M.-D., Marty, C., Chappaz, R., Gilles, A., 2002. Molecular phylogeny of French Guiana Hylinae: implications for the systematic and biodiversity of the Neotropical frogs. C.R. Biol. 325, 141–153.

Salisbury, B.A., 1999. Strongest evidence: maximum apparent phylogenetic signal as a new cladistic optimality criterion. Cladistics 15, 137–149.

Salmon, W.C., 1966. The Foundations of Scientific Inference. University of Pittsburgh Press, Pittsburgh, PA.

Sanderson, M.J., 1995. Objections to bootstrapping: a critique. Syst. Biol. 44, 299–320.

Sanderson, M.J., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817–829.

Sanderson, M.J., Wojciechowski, M.F., 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). Syst. Biol. 49, 671–685.

Schuh, R.T., Farris, J.S., 1981. Methods for investigating taxonomic congruence and their application to the Leptopodomorpha. Syst. Zool. 30, 331–351.

Siddall, M.E., 1995. Another monophyly index: revisiting the jackknife. Cladistics 11, 33–56.

Siddall, M.E., 1997. Prior agreement: arbitration or arbitrary? Syst. Biol. 46, 765–769.

Siddall, M.E., 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. Cladistics 14, 209–220.

Siddall, M.E., 2001. Computer-intensive randomization in systematics. Cladistics 17, S35–S52.

Siddall, M.E., 2002a. Measures of support. In: DeSalle, R., Giribet, G. and Wheeler, W.C. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser Verlag, Basel, Switzerland, pp. 80–101.

Siddall, M.E., 2002b. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. Cladistics 17, 395–399.

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

Siddall, M.E., Whiting, M.F., 1999. Long-branch abstractions. Cladistics 15, 9–24.

Simmons, M.P., Randle, C.P., Freudenstein, J.V., Wenzel, J.W., 2002. Limitations of Relative Apparent Synapomorphy Analysis (RASA) for measuring phylogenetic signal. Mol. Biol. Evol. 19, 14–23.

Simmons, N.B., 2001. Misleading results from the use of ambiguity coding to score polymorphisms in higher-level taxa. Syst. Biol. 50, 613–620.

Smith, E.N., Gutberlet Jr., R.L., 2001. Generalized frequency coding: a method of preparing polymorphic multistate characters for phylogenetic analysis. Syst. Biol. 50, 156–169.

Sneath, P.H.A., Sokal, R.R., 1962. Numerical taxonomy. Nature 193, 855–860.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy. W.H. Freeman, San Francisco.

Sober, E., 1988. Reconstructing the Past. The MIT Press, Cambridge, MA.

Sokal, R.R., Rohlf, F.J., 1981. Biometry. W.H. Freeman, San Fransisco.

Sokal, R.R., Sneath, P.H.A., 1963. Principles of Numerical Taxonomy. Freeman, San Francisco.

Steel, M.A., 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43, 560–564.

Steel, M.A., Hendy, M.D., Penny, D., 1993a. Parsimony can be consistent!. Syst. Biol. 42, 581–587.

Steel, M.A., Lockhart, P.J., Penny, D., 1993b. Confidence in evolutionary trees from biological sequence data. Nature 364, 440–442.

Steel, M.A., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.

Sullivan, J., Swofford, D., 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 4, 77–86.

Swofford, D., 1991. When are phylogeny estimates from molecular and morphological data incongruent? In: Miyamoto, M.M. and Cracraft, J. (Eds.), Phylogenetic Analysis of DNA Sequences. Oxford University Press, New York, pp. 295–333.

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

Swofford, D.L., Berlocher, S.H., 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst. Zool. 36, 293–325.

Systematic Biology: Instructions for authors. 2002. Systematic Biology: Instructions for authors. Available from http://systbiol.org/info/instrauth.html.

Teeling, E.C., Scally, M., Kao, D.J., Romagnoli, M.L., Springer, M.S., Stanhope, M.J., 2000. Molecular evidence regarding the origin of echolocation and flight in bats. Nature 403, 188–192.

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

Thiele, K., 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics 9, 275–304.

Thompson, E.A., 1975. Human Evolutionary Trees. Cambridge University Press, Cambridge.

Throckmorton, L.H., 1968. Concordance and discordance of taxonomic characters in *Drosophila* classification. Syst. Zool. 17, 355–387.

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

von Wright, G.H., 1984. Philosophical Logic. Cornell University Press, Ithaca.

Wägele, J.-W., Misof, B., 2001. On quality of evidence in phylogeny reconstruction: a reply to Zrzavý's defence of the 'Ecdysozoa' hypothesis. J. Zool. Syst. Evol. Res. 39, 165–176.

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

Watkins, J., 1997. Popperian ideas on progress and rationality in science. Crit. Rationalist 2, 2–11.

Wenzel, J.W., 1997. When is a phylogenetic test good enough?. In: Grandcolas, P. (Ed.), The Origin of Biodiversity in Insects: Phylogenetic Tests Evolutionary Scenarios. Mém. Mus. Natn. Hist. Nat. Paris, pp. 31–45.

Wenzel, J.W., Carpenter, J.M., 1994. Comparing methods: adaptive traits and tests of adaptation. In: Eggleton, P., Vane-Wright, R.I. (Eds.), Comparing Methods: Adaptive Traits and Tests of Adaptation. Academic Press, London, pp. 51–64.

Wenzel, J.W., Siddall, M., 1999. Noise. Cladistics 15, 51–64.

West, J.G., Faith, D.P., 1990. Data, methods and assumptions in phylogenetic inference. Aust. Syst. Bot. 3, 9–20.

Wheeler, W.C., 1991. Congruence among data sets: A Bayesian approach. In: Miyamoto, M.M. and Cracraft, J. (Eds.), Phylogenetic Analysis of DNA Sequences. Oxford University Press, New York, pp. 334–346.

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

Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.

Wheeler, W.C., 1999. Measuring topological congruence by extending character techniques. Cladistics 15, 131–135.

Wheeler, W.C., 2000. Heuristic reconstruction of hypothetical-ancestral DNA sequences: Sequence alignment vs optimization. In: Scotland, R.W. and Pennington, R.T. (Eds.), Homology and Systematics. Taylor and Francis, New York, pp. 106–113.

Wheeler, W.C., Gladstein, D., De Laet, J., 1996–2002. POY: Phylogeny Reconstruction via Optimization of DNA Data. Ver. 3.0. Available from ftp://ftp.amnh.org/pub/molecular/poy.

Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 14, 173–192.

Wheeler, W.C., Whiting, M.F., Wheeler, Q.D., Carpenter, J.M., 2001. The phylogeny of extant hexapod orders. Cladistics 17, 113–169.

Whelan, S., Goldman, N., 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol. 16, 1292–1299.

Wiens, J.J., 1995. Polymorphic characters in phylogenetic systematics. Syst. Biol. 44, 482–500.

Wiens, J.J., 1998. Testing phylogenetic methods with tree congruence: phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Syst. Biol. 47, 427–444.

Wiens, J.J., 1999. Polymorphism in systematics and comparative biology. Annu. Rev. Ecol. Syst. 30, 327–362.

Wiens, J.J., 2000a. Coding morphological variation within species and higher taxa for phylogenetic analysis. In: Wiens, J.J. (Ed.), Phylogenetic Analysis of Morphological Data. Smithsonian Institution Press, Washington, pp. 115–145.

Wiens, J.J., 2000b. Reconstructing phylogenies from allozyme data: comparing method performance with congruence. Biol. J. Linn. Soc. 70, 613–632.

Wiens, J.J., 2001. Character analysis in morphological phylogenetics: problems and solutions. Syst. Biol. 50, 689–699.

Wilgenbusch, J., de Queiroz, K., 2000. Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrian DNA sequences generated by heterogeneous evolutionary processes. Syst. Biol. 49, 592–612.

Wilkinson, M., 1992. Ordered versus unordered characters. Cladistics 8, 375–385.

Wilkinson, M., 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43, 343–368.

Wilkinson, M., 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44, 501–514.

Wilkinson, M., Thorley, J.L., Upchurch, P., 2000. A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. Syst. Biol. 49, 754–776.

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

Yang, Z., Goldman, N., Friday, A., 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. 44, 384–399.

Yang, Z., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Yoder, A.D., Irwin, J.A., Payseur, B.A., 2001. Failure of the ILD to determine data combinability for slow loris phylogeny. Syst. Biol. 50, 408–424.

Zardoya, R., Meyer, A., 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13, 933–942.