TESTING SPECIES PHYLOGENIES AND PHYLOGENETIC METHODS WITH CONGRUENCE

MICHAEL M. MIYAMOTO1,3 AND WALTER M. FITCH2

1Department of Zoology, University of Florida, Gainesville, Florida 32611, USA
2Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, California 92717-2515, USA

Abstract.—We assessed the utility of congruence and multiple data sets to test species relationships and the accuracy of phylogenetic methods. The ongoing controversy about whether to combine data sets for phylogenetic analysis was evaluated against the naturalness of different types of data (as commonly recognized by systematists) and character independence. We defend the recommendation that independent data sets (defined in terms of process partitions; sensu Bull et al., 1993, Syst. Biol. 42:384–397) should rarely be combined but should be kept separate for phylogenetic analysis because their independence increases the significance of corroboration. Trees of natural taxa, well supported by many independent lines of evidence, should be used in the same way as the known phylogenies of simulations and of certain laboratory and domesticated groups, i.e., as standards for evaluating the accuracy of different phylogenetic methods. Although compromised by their imperfect reliabilities, such tests using well-supported trees of wild taxa provide important reality checks on the conclusions of the other two approaches by encompassing more of the complexity and diversity of natural systems and their evolutionary processes. In this way, a combination of testing with the well-supported trees of natural groups, with simulations, and with those laboratory and domesticated taxa with known phylogenies is most likely to prove effective in establishing the strengths, weaknesses, and assumptions of different phylogenetic methods. [Accuracy; taxonomic congruence; character congruence; process partitions; character independence; well-supported trees; phylogenetic methods.]

A principal objective of most phylogenetic studies is to contribute to the discovery of the true species phylogeny of life underlying biological diversity (Penny and Hendy, 1986; Lanyon, 1993; Hillis et al., 1994). Heritable characters of taxa are the products of this history and are therefore of potential value in its elucidation. Thus, different data sets, when subjected to reliable methods of phylogenetic analysis, are expected to converge onto the true species phylogeny for their group (Miyamoto and Cracraft, 1991; Sheldon and Bledsoe, 1993; Miyamoto et al., 1994). This expectation forms the basis in systematics for testing the accuracy or reliability of phylogenetic hypotheses with new independent data. Those hypotheses that are supported by different lines of evidence are preferred over those that are not. Phrased differently, congruent trees obtained from analyses of independent data sets provide the best estimates of the true phylogeny for a group (Penny and Hendy, 1986; Cracraft and Helm-Bychowski, 1991; Lanyon, 1993). This approach is part of the age-old tradition in science of hypothesis testing with new independent information. It remains the most familiar way for practicing systematists to test the accuracy of their phylogenetic conclusions for particular groups.

Less obvious to systematists is that the accuracy of their phylogenetic methods (as well as of their data; see Kumazawa and Nishida, 1993) can also be tested with congruence (Penny and Hendy, 1986; Bledsoe and Raikow, 1990; Miyamoto and Cracraft, 1991). This approach relies on the same premise as before, that the congruent topology is the most accurate estimate of the true phylogeny. Thus, phylogenetic methods that recover congruent topologies are preferred over those that rarely do (Allard and Miyamoto, 1992; Miyamoto et al., 1994). Two of the earliest studies to test alternative tree-building procedures with congruence were by Mickevich and Johnson (1976) and Mickevich (1978), who compared the stabil-
CONGRUENCE, SPECIES PHYLOGENIES, AND PHYLOGENETIC METHODS

ity or agreement (=consensus, sensu Kluge and Wolf, 1993) of different phenetic versus cladistic procedures. In their studies, different pairs of data sets for the same groups (e.g., morphological versus allozyme characters for *Menidia* [silversides]) were individually analyzed by both phenetic and cladistic methods. The resultant trees of each pair were then compared for their agreement, leading these investigators to favor cladistic algorithms because they resulted in more stable topologies. Their conclusions about the greater phylogenetic agreement from cladistic procedures have been controversial, but the focus of those arguments has largely been on the actual protocols of the tests and not on the overall approach (e.g., Colless, 1980; Mickevich and Farris, 1981). All parties involved in the controversy recognized the potential for testing the relative merits of different tree-building procedures with congruence.

The above studies of different data sets are examples of taxonomic congruence (Mickevich, 1978; Kluge, 1989). Taxonomic congruence is the agreement or consensus among the supported topologies of different data sets. Consensus objects (trees or statements) are normally employed to summarize the common groupings and/or nestings among the supported topologies of some or all of the separate data matrices. This approach constitutes one view on the best way to synthesize the information of multiple data sets. A more recent viewpoint, referred to as character congruence, combines the individual data sets to maximize the descriptive efficiency ( informativeness) and explanatory power of the total information (Kluge, 1989; Eernisse and Kluge, 1993; Kluge and Wolf, 1993). Optimal descriptive and explanatory power (which is achieved by the most-parsimonious topology of the total information) is critical because only then, according to this view, can characters be properly judged as evidence of species relationships. The opposing viewpoints of taxonomic congruence and character congruence have become one of the important controversies in modern systematics.

This contribution to the special issue of *Systematic Biology* on assessing phylogenetic accuracy focuses first on the ongoing debate over whether or not data sets should be combined for phylogenetic analysis. In this examination of the debate, we evaluate whether taxonomic congruence offers "something more" than the optimization of descriptive and explanatory power. Our focus then shifts to the use of congruent phylogenies for natural groups to test the reliability of different tree-building procedures. We conclude (1) that taxonomic congruence can be preferred by relating its "something more" to the greater potential for independence of characters from different data sets, and (2) that trees of natural groups, well supported by multiple lines of evidence, provide another important way to compare the accuracies of phylogenetic methods.

**PARTITIONING AND COMBINING DATA SETS**

In this section on congruence, the different algorithms for calculating consensus trees and indices and measures of character fit and incongruence will not be discussed because an excellent review of these procedures is available elsewhere (Swofford, 1991). Rather, we focus on the ongoing controversy as to whether data from different sources (e.g., mitochondrial DNA [mtDNA] versus nuclear DNA sequences [Lanyon, 1993]) should be combined into a single character matrix prior to phylogenetic analysis. The extreme viewpoints on this topic are those of taxonomic congruence (for the phylogenetic analysis of separate data sets: Nelson, 1979; Nelson and Platnick, 1981; Shaffer et al., 1991; Marshall, 1992; Lanyon, 1993) and character congruence (for the combined treatment of multiple data sets: Miyamoto, 1985; Kluge, 1989; Barrett et al., 1991; Donoghue and Sanderson, 1992; Eernisse and Kluge, 1993; Kluge and Wolf, 1993; Chippindale and Wiens, 1994). Bull et al. (1993) and de Queiroz (1993) took the intermediate position that data sets should only be combined when they do not strongly support conflicting trees. Reasons for combining or not combining data sets...
are summarized in Table 1, as synthesized from the preceding contributions to this controversy.

**Different Types of Data and Process Partitions**

An understanding of the controversy about whether or not to combine data sets begins with an examination of reason I.3 in Table 1. This argument in support of combining data sets raises the question of whether natural classes of characters are discoverable in the first place (Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993). The contention is that data sets, as commonly recognized by systematists, do not conform to natural classes of characters but are merely artifacts of tradition and technology. Thus, nothing of significance can be gained by dividing characters into artificial categories and no special importance can be attached to the agreement among the topologies of different data sets. Taxonomic congruence depends on the naturalness of its character partitions and is thereby compromised by problems in their recognition. In contrast, character congruence is not so hampered because it relies on combined data matrices rather than on partitions of the available information (Kluge and Wolf, 1993). At issue then is whether taxonomic congruence can be discounted from the start on the grounds that natural classes of characters are unrecognizable.

The naturalness of the different data sets of systematists can be defended by relating their existence to the concept of process partitions, sensu Bull et al. (1993). These authors defined process partitions as subsets of characters that are evolving under demonstrably different sets of rules. Significant differences in these rules are demonstrated when data sets strongly support conflicting trees. These heterogeneities are not likely the result of stochastic errors alone but are more likely due to separate histories for the different sets of characters and/or to systematic errors and model failures in the phylogeny reconstructions for at least some of the data sets (Swofford and Olsen, 1990; Shaffer et al., 1991; Doyle, 1992; de Queiroz, 1993). A lower limit on the sizes of process partitions is effectively set by the need for sufficient numbers of characters to test statistically for significant heterogeneities among these partitions.

The concept of process partitions, as developed by Bull et al. (1993), is important because of its emphasis on evolutionary processes, but its reliance on strongly supported but conflicting trees for their recognition is regarded as too restrictive. Instead, process partitions may be useful, even when they do not strongly support conflicting phylogenies, by focusing on the biological and evolutionary factors that are the basis of their different sets of rules and thus their greater potential for character independence. Although such recognitions must be made with caution, character sets may be considered as independent process partitions when (1) their genes are not genetically linked, (2) the products of their genes do not interact with each other, (3) the genes do not specify the same function, (4) the gene products are not components of a common pathway (e.g., electron transport system), and (5) the gene products do not regulate the expression of loci in other partitions. This list is intended to be directive rather than prescriptive and is expected to change with experience. Our broadened definition of process partitions does not preclude their recognition in those instances defined by Bull et al. (1993), i.e., where separate data sets support significantly different trees; their cases remain unaffected.

For example, the gene sequences for oxyg-globin and 12S ribosomal RNA (rRNA) of the nuclear and mitochondrial genomes, respectively, can be considered as separate process partitions whether or not they strongly support conflicting phylogenies. This recognition is based on the many differences in such characteristics as transmission genetics (biparental vs. maternal), copy number (duplicate vs. effectively single), biochemistry (e.g., well-developed DNA repair mechanisms vs. none readily apparent), and expression (transcribed and translated vs. transcribed only) that distinguish the two genes and their genomes (Dickerson and Geis, 1983; Brown, 1985). The two genes
constitute separate genetic units encoding distinct products with different functions (polypeptide vs. structural RNA and gas transportation vs. protein translation, respectively). These biological differences underlie the different evolutionary properties of the two genes, as represented by their differing substitution patterns, evolutionary rates, and frequencies of recombination, gene conversion, and duplication (Dickerson and Geis, 1983; Mindell and Honeycutt, 1990; Miyamoto and Goodman, 1990; Avise, 1991). In turn, these differences in their evolutionary characteristics permit the recognition of the \( \beta \)-globin and 12S rRNA gene sequences as independent process partitions. Thus, if both sets of data support the same tree, the correlations among their character state changes are more likely to be the consequence of the species history, whose recovery is the purpose of the investigation. For reasons such as these, most systematists agree that their data sets conform to natural divisions among characters, thereby leaving taxonomic congruence as a viable alternative to character congruence (Swofford, 1991; Lanyon, 1993).

**Taxonomic Congruence and Character Independence**

The primary argument for not combining data sets is directly related to the concept of character independence as it applies to the recovery of the species phylogeny (reason II.1, Table 1) (Kluge, 1989; Shaffer et al., 1991; de Queiroz, 1993). The emphasis on species phylogeny is important because it acknowledges that characters can be interdependent with regard to the reconstruction of their species phylogeny even when they are independent with respect to their character phylogeny (e.g., gene tree) (Doyle, 1992). The relationship between separate data sets and independence assumes that nonindependence is more likely for characters from the same data sets than for characters from different ones (Shaffer et al., 1991; Lanyon, 1993). Thus, characters of one data set are more likely to support the same wrong species phylogeny than are those from different data sets (de Queiroz, 1993).

Combining data sets for phylogenetic analysis assumes (1) that the same history is true for each character set and (2) that the chosen methods of tree reconstruction are equipped to handle the differences in their evolutionary rules (e.g., by the use of unequal weighting for characters and/or changes if heterogeneous rates or transformational probabilities are a problem [Bull et al., 1993; Chippindale and Wiens, 1994; Hillis, 1995]). The existence of strongly supported but conflicting trees is prima facie evidence that either or both of these assumptions for combining data sets has been violated. It follows then that different data sets should not be combined but should be kept separate when they strongly support conflicting trees (Shaffer et al., 1991; Bull et al., 1993; de Queiroz, 1993). But what about when the data sets are not significantly heterogeneous? We support the opinion that data sets still should be kept separate under these conditions for the same reasons as above (Nelson, 1979; Nelson and Platnick, 1981; Swofford, 1991; Lanyon, 1993). As process partitions, the biological and evolutionary differences among data sets make it more likely that the agreement among their topologies is the result of the true species phylogeny rather than the result of the same nonphylogenetic factors responsible for the separate histories of character sets and/or the systematic errors and model failures in phylogeny reconstructions.

This reason in support of keeping data sets separate can be illustrated by referring back to the example for the \( \beta \)-globin and 12S rRNA genes. Suppose that an investigator has shown that species A and B are sister taxa according to an analysis of sequences for the 5' end of the \( \beta \)-globin gene. Imagine that this investigator has a choice of sequencing either an adjacent segment of the \( \beta \)-globin gene or the 5' end of the 12S rRNA locus. All else being equal, the mtDNA sequences for the 5' end of the 12S rRNA locus offer the more powerful test of the hypothesis that A and B are sister species because of the greater likelihood that these data are independent of the nuclear DNA evidence for the 5' end of the \( \beta \)-globin gene (Doyle,
TABLE 1. Reasons to combine or not to combine data sets prior to phylogenetic analysis and our general responses to them.¹

<table>
<thead>
<tr>
<th>Reasons</th>
<th>General response</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. To combine data sets</td>
<td></td>
</tr>
<tr>
<td>1. The goal of character congruence is to maximize the descriptive efficiency (informativeness) and explanatory power of the character data by the direct resolution of character conflicts according to the total information itself (Kluge, 1989; Barrett et al., 1991; Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993). By virtue of its optimal descriptive and explanatory power, the most-parsimonious tree provides the proper context to evaluate characters as evidence of species relationships. In contrast, such power is not guaranteed when final phylogenies are obtained by taxonomic congruence and not by direct reference to the data themselves. Thus, this approach often supports phylogenies with reduced descriptive and explanatory power, thereby rendering these solutions inappropriate for evaluating character homology.</td>
<td></td>
</tr>
<tr>
<td>2. Combining data sets follows from first principles (i.e., the principle of total evidence that hypotheses must be selected on the basis of all available information; Kluge, 1989; Barrett et al., 1991).</td>
<td>1. There are other reasons besides descriptive and explanatory power that must be considered when phylogenetic hypotheses are selected from multiple data sets. Of these, the most important one is the greater potential for independence for characters from different data sets than for characters from the same sets (see reason II.1 below). Thus, taxonomic congruence is preferred because it provides the most direct way to emphasize the special importance of agreement among the topologies of different data sets. If its solution is accepted as the best estimate of the true phylogeny for a group, then this estimate becomes the historical foundation to infer the homologies and nonhomologies of characters. Reason I.1 subsumes two other arguments that are often presented in support of combining data sets (Bull et al., 1993; de Queiroz, 1993). The first is that a true but weak phylogenetic signal, masked by homoplasy, can become amplified over the noise when data sets are combined (Barrett et al., 1991). The second is that a more fully resolved tree can be obtained from the combined data when the resolving power of the individual character matrices varies with hierarchical level (Hillis, 1987). Both possibilities are included under this reason, because they correspond to specific cases of improved resolution by direct reference to the total information itself.</td>
</tr>
<tr>
<td>3. Different types of data, as commonly recognized by systematists, are not natural but are merely artifacts of tradition and technology (Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993). Taxonomic congruence is burdened by this problem, because it relies on the proper identification of natural divisions among data to justify the assignment of special importance to the agreement among their topologies (reason II.1 below). In contrast, character congruence is not so burdened because its goal is to maximize descriptive and explanatory power (reason I.1 above). This goal is achieved by the use of combined data sets rather than the use of separate partitions of the total information.</td>
<td>2. True, but the same correspondence to the principle of total evidence can be claimed as well when all of the available information is analyzed on a data set by data set basis (de Queiroz, 1993; Purvis, 1995). In both cases, hypotheses are rejected and retained “in light of all of the data” (Barrett et al., 1991:487, emphasis in original).</td>
</tr>
<tr>
<td>4. The consensus object (tree or statement), summarizing the topologies of different data sets, is not necessarily conservative because it can positively contradict the phylogeny supported by the combined character matrix (Barrett et al., 1991; Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993).</td>
<td>3. If the different data sets of systematists are artificial, then no special importance can be claimed for the agreement among their topologies. The principle reason not to combine data sets (reason II.1 below) is thereby negated and no primary argument remains for keeping character sets separate. Thus, those who use taxonomic congruence must first document that their data partitions are natural before assigning special significance to the agreement among their phylogenies. Such justifications can be accomplished by relating the data sets of systematists to the concept of process partitions (sensu Bull et al., 1993).</td>
</tr>
<tr>
<td>5. Combining data sets follows from the principle of total evidence that hypotheses must be selected on the basis of all available information (Kluge, 1989; Barrett et al., 1991).</td>
<td>4. True, the two approaches can give contradictory results and thus the use of a consensus procedure cannot be justified on the grounds that it is safe or conservative. However, this reason should not be taken to mean that the topology derived from taxonomic congruence is more likely to be incorrect than is that from character</td>
</tr>
</tbody>
</table>
Table 1. Continued.

<table>
<thead>
<tr>
<th>Reasons</th>
<th>General response</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Many different procedures for constructing consensus objects are available and the final results for each often disagree (Kluge, 1989; Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993).b Unfortunately, there are no clear reasons, other than technical ones, to prefer the results of one method over those of the others. This argument includes the problem of how to compare two or more suites of equally parsimonious trees for different data sets (criticism 6 of Kluge and Wolf, 1993).</td>
<td>congruence (or vice versa; Chippindale and Wiens, 1994) when the two disagree.</td>
</tr>
</tbody>
</table>

II. Not to combine data sets

1. Independence is more likely for characters from different data sets than for characters from the same sets (Swofford, 1991; Lanyon, 1993). As a result, characters from different data sets are less likely to support the same, and perhaps wrong, species phylogeny than are those from the same data sets (de Queiroz, 1993). This reason is the basis for attaching special importance to the agreement among the topologies of different data sets (Nelson, 1979; Nelson and Platnick, 1981; Swofford, 1991; Lanyon, 1993). It also underlies the conclusions of Bull et al. (1993) and de Queiroz (1993) that data sets should not be combined when they strongly support conflicting trees. Thus, taxonomic congruence is called for, because this approach is designed to summarize the agreement among the topologies of independent data sets (Swofford, 1991).

2. Some data sets are noncombinable because they are not collected as character information.b For example, the distance matrix of a DNA–DNA hybridization study cannot be combined in principle with the character data from a morphological analysis (Barrett et al., 1991; Eernisse and Kluge, 1993; Lanyon, 1993; Sheldon and Bledsoe, 1993). | 1. This reason assumes that the different data sets of systematists conform to natural classes of characters (reason 1.3 above). We contend that the best way for systematists to ensure the naturalness of their data sets is to identify them as process partitions (following our modification of the definition of Bull et al., 1993). As process partitions, the different evolutionary and biological properties of data sets make it more likely that the agreement among their topologies is the result of the true species phylogeny and not of the same nonphylogenetic factors responsible for the separate histories of character sets and/or the systematic errors and model failures in phylogeny reconstructions (Miyamoto and Cracraft, 1991; Lanyon, 1993; Sheldon and Bledsoe, 1993). This reason is the basis of our recommendation that data sets should rarely be combined for phylogenetic analysis even when they do not strongly support conflicting trees. |

2. In these cases, the only available option is to test for agreement among the topologies of the incompatible matrices or to deny the informativeness of some categories of evidence (e.g., distance data; Kluge, 1989; Eernisse and Kluge, 1993; Kluge and Wolf, 1993). | 2. In these cases, the only available option is to test for agreement among the topologies of the incompatible matrices or to deny the informativeness of some categories of evidence (e.g., distance data; Kluge, 1989; Eernisse and Kluge, 1993; Kluge and Wolf, 1993). |

---

a An interesting connection between congruence analyses of species relationships and meta-analysis (the use of formal statistical methods to integrate the results of separate disparate studies [Mann, 1990]) has been made by Hillis (1995), who recognized that much of the debate over meta-analysis is being repeated by systematists in their “to combine or not to combine” controversy. Thus, the potential exists that much can be learned about the phylogenetic synthesis of multiple data sets from this ongoing debate over meta-analysis.

b Reasons 1.5 and II.2 concern specific methodological limitations of the two approaches, not their different underlying principles that are the fundamental reasons why one should combine or should not combine multiple data sets.
This conclusion is based on the many biological and evolutionary differences that distinguish the unlinked genes and genomes of these two data sets. Similar arguments do not apply to the two data sets of the $\gamma^1$-globin locus because of their many interrelationships as samples of the same gene (Farris, 1971). Indeed, the possibility of gene conversion constitutes one obvious alternative to the true species phylogeny as an explanation for any agreement between the two samples (e.g., as in regards to the monophyly of species A and B) (Miyamoto and Goodman, 1990). The greater potential for independence of characters from different process partitions is the "something more" that systematists seek when they analyze their data sets separately.

Preference

If the data sets of systematists are artificial and/or independence is no more likely for characters from different data sets than for those from the same data sets, then the total information should be combined into a single matrix prior to phylogenetic analysis to maximize descriptive and explanatory power and to minimize stochastic errors. However, our arguments in support of the naturalness of different types of data (as commonly recognized by systematists) and of the greater potential for independence of characters from different process partitions suggest instead that data sets should rarely be combined for phylogenetic analysis even when they do not strongly support conflicting trees (Nelson, 1979; Nelson and Platnick, 1981; Swofford, 1991; Lanyon, 1993). Given these arguments, taxonomic congruence is preferred because this approach maintains the distinctions among independent data sets and acknowledges the special importance of agreement among their phylogenies. Combining old data with the new destroys the independence of the new information to test the hypotheses derived from the old evidence (Nelson, 1979; Nelson and Platnick, 1981; Lanyon, 1993). It is for these reasons that taxonomic congruence is included among "the traditional virtues of systematics" whereby phylogenetic hypotheses are tested by their replication among "small [independent] samples" (Nelson, 1979:13).

Our preference for taxonomic congruence depends on specific assumptions about the evolutionary process considered important for the estimation of the true species phylogeny. This viewpoint stands in sharp contrast to the perspective of character congruence, whereby descriptive and explanatory power are emphasized and explicit assumptions are avoided (reasons I.1, I.3, and II.1; Table 1) (Eernisse and Kluge, 1993; Jones et al., 1993; Kluge and Wolf, 1993). These philosophical differences are characteristic of those that are fundamental to the general debate in phylogenetic systematics over the use of statistical and nonstatistical approaches in the field (see Sanderson [1995] for a recent discussion). These issues are expected to become even more important as statistical methods continue to gain acceptance for use in phylogenetic inference.

Other Considerations

Traditionally, taxonomic congruence has relied on comparisons of only the best (e.g., most parsimonious) topologies for different data sets (Miczekich, 1978; Kluge and Wolf, 1993). However, this reliance on only optimal trees ignores the fact that all phylogenetic estimates are made with some degree of error and uncertainty (Penny and Hendy, 1986; Lanyon, 1993). Thus, studies of taxonomic congruence must be based on comparisons of the optimal and near-optimal topologies for different data sets rather than only on their most-parsimonious solutions (Swofford, 1991; Rodrigo et al., 1993; Miyamoto et al., 1994; Hillis, 1995). In this regard, the procedures of Swofford (1991: 326), Rodrigo et al. (1993), and Miyamoto et al. (1994) deserve further attention because they are designed to search among the optimal and near-optimal topologies of different data sets for the fully resolved tree that is most strongly supported by each individual matrix. Topologies identified in this way can be retained as the best estimates of
their species phylogeny, given their greater support from the independent data sets.

New statistical tests are being developed for the detection of strong support for conflicting trees in different data sets (Bull et al., 1993; Rodrigo et al., 1993; Larson, 1994). These tests, which can be thought of as the first attempts to delimit "confidence sets" of optimal and near-optimal trees for each data matrix (Sanderson, 1989; Hillis, 1995), complement the procedures of Swofford (1991) and Miyamoto et al. (1994) by establishing with statistical rigor which data sets are incompatible and which are not. Individual data sets that are incompatible with others can be identified in this way and targeted for future studies concerned with the sources of their significant heterogeneities. If these sources are systematic errors and model failures, then additional efforts can be made to accommodate better the assumptions of their tree-building procedures to the biological and evolutionary characteristics of the incompatible data sets (Shaffer et al., 1991; Bull et al., 1993; Lanyon, 1993). Such follow-up studies are necessary whenever different data sets strongly support conflicting trees. This requirement, along with that for the consideration of both optimal and near-optimal topologies for each data set, constitute stipulations of our preference for taxonomic congruence.

Discussion about the use of all available information has also centered on whether fossils should be included in phylogenetic studies (Gauthier et al., 1988; Donoghue et al., 1989; Eernisse and Kluge, 1993). Although related to this discussion by a concern for what is meant by total information, this issue does not bear on our preference for taxonomic congruence, which is based on the importance of character independence. The conclusions of Gauthier et al. (1988), Donoghue et al. (1989), and Eernisse and Kluge (1993) are convincing, and it is clear that fossils should be included in phylogenetic analyses despite their apparent shortcomings (e.g., their limited number of characters). Furthermore, the inclusion of fossils in a study of taxonomic congruence does not pose a problem for this approach even if the extinct taxa are not represented in all of the individual data sets. Although additional research is needed in this area, alternative procedures already exist to compare the topologies for data sets with different memberships (Gordon, 1986; Steel, 1992; Lanyon, 1993; Purvis, 1995).

Systematists routinely place greater confidence in phylogenetic hypotheses that are confirmed by multiple lines of evidence rather than by only one or two (e.g., Swofford, 1991; Helm-Bychowski and Cracraft, 1993; Lanyon, 1993; Sheldon and Bledsoe, 1993; Hillis, 1995). As the accepted best estimates of the true species phylogeny, such topologies provide the historical framework needed to assess characters as homologies and nonhomologies and to trace character evolution. When well supported by many independent lines of evidence, these topologies can also serve as the foundation for testing the accuracy of different phylogenetic methods.

**TESTING THE ACCURACY OF PHYLOGENETIC METHODS**

**Well-Supported Phylogenies of Natural Groups**

The true phylogenetic relationships for the vast majority of real taxa will never be known with certainty (Hillis and Bull, 1991; Miyamoto and Cracraft, 1991; Hillis, 1995). This fact has impeded the testing and development of phylogenetic methods, because accurate estimates of success and failure in reconstructing phylogenies require direct knowledge about the true relationships of a group (Nei, 1991; Hillis et al., 1994). To circumvent this limitation, investigators have relied (1) on simulation analyses whereby the simulated taxa or their sequences diverge according to a given phylogeny and model of evolution (Hulsenbeck, 1995) or (2) on those very few groups of laboratory and domesticated organisms where their actual history is known (e.g., experimental phylogeny for T7 bacteriophages [Hillis et al., 1992] and inbred strains of mice [Atchley and Fitch, 1991]). The availability of known phylogenies in these cases allows for precise esti-
mates of the frequencies of success and failure of different phylogenetic methods.

In the absence of such certainty, systematists have tested the successes and failures of their phylogenetic procedures with the well-supported phylogenies of natural groups (e.g., Allard and Miyamoto, 1992; Miyamoto et al., 1994; see also below). Such phylogenies, obtained by many independent lines of evidence, can be used in tests of methods in the same way as the known phylogenies of simulations and of certain laboratory and domesticated groups, i.e., as the standard references for measuring reliability. For example, the accuracies of three tree-building procedures were evaluated by Gouy and Li (1989), using the topology in Figure 1a for human, fruit fly (Drosophila), rice, and slime mold (Physarum). This topology in support of a human/fruit fly grouping is widely accepted by biologists because of the large body of congruent evidence in its favor from morphology, molecular biology, biochemistry, physiology, and other sources (Margulis and Schwartz, 1982; Fernholm et al., 1989). This support justifies its acceptance as an accurate representation of the true phylogeny for these taxa. In turn, this acceptance of its accuracy permitted Gouy and Li (1989) to test the strengths, weaknesses, and assumptions of the three tree-building procedures despite the lack of direct knowledge of these species' relationships.

The accepted grouping of the two animals (human and fruit fly; Fig. 1a) was supported by the maximum-parsimony and neighbor-joining analyses of the small subunit rRNA (SSU rRNA) sequences for the four representative species (Gouy and Li, 1989). In contrast, a different and presumably wrong arrangement was obtained when the same sequences were analyzed by the evolutionary parsimony procedure (Fig. 1b). This apparent failure of the evolutionary parsimony procedure was related to its sensitivity to species differences in G + C content and to unequal transversion frequencies for the same base (e.g., the rate for A → C does not equal the rate for A → T). The greater robustness of the other two methods to these conditions was illustrated in the process. These insights into the behaviors and relative merits of the three tree-building procedures complement the results of simulation analyses with an example for natural taxa (Gouy and Li, 1989; Jin and Nei, 1990; Nei, 1991).

Although the example in Figure 1a is for a set of distantly related taxa (see also Kumazawa and Nishida, 1993), this fact does not necessarily imply that tests of phylogenetic accuracy with the well-supported phylogenies of natural groups are largely or wholly restricted to topologies representing older divergences. Instead, we predict that a significant number of comparable examples also exist for more closely related taxa, and we recommend that specialists in particular

**Figure 1.** (a) Accepted arrangement for human, fruit fly (Drosophila), rice, and slime mold (Physarum). (b, c) The two possible dichotomous alternatives for the four representative taxa. The accepted arrangement is supported by maximum-parsimony and neighbor-joining analyses of the SSU rRNA sequences for these species (Gouy and Li, 1989). In contrast, evolutionary parsimony analysis of the same sequences results in statistically significant support for alternative b ($P < 0.05$) but not for the accepted topology (a) or the third possibility (c). The results for the evolutionary parsimony analysis are summarized by the $\chi^2$ (df = 1) values for the three alternatives.
groups and their publications should be consulted for the identification of such trees. Our familiarity with primates allows us to propose that the phylogenies of hominoids and of major anthropoid lineages are sufficiently known to qualify as well supported (Miyamoto and Goodman, 1990; Patterson et al., 1993). Thus, well-supported trees for different natural groups, spanning a broad range of hierarchical and taxonomic levels and evolutionary times, do exist for testing the accuracy of different phylogenetic methods.

Strengths, Weaknesses, and Assumptions of the Approach

One obvious weakness with using well-supported trees of natural groups to test phylogenetic accuracy is that some degree of uncertainty, however small, always remains with them despite their extensive support. Thus, although they constitute best-case situations for natural taxa, they can never be equated with truth in contrast to the known phylogenies of simulations and of certain laboratory and domesticated groups. Well-supported phylogenies of natural groups must be recognized for what they are in tests of phylogenetic methods: the congruent topologies are assumptions of their methodological investigations. Because all scientific studies depend on assumptions, tests of phylogenetic methods based on well-supported trees of natural taxa are not invalid simply because they rely on phylogenies with some degree of uncertainty as their standards. Under the premise that extensive congruence is evidence of the true phylogeny for a group, the validity of these tests depends instead on the degree to which their accepted standards can be defended with available information. Thus, congruent topologies, such as Figure 1a, allow for stronger inferences about the accuracy of phylogenetic methods than do phylogenies that are more poorly supported. These assumptions are the same ones that behavioralists, ecologists, and other comparative biologists make when they accept the phylogenetic hypotheses of systematists to trace character evolution within their groups (Lanyon, 1993).

The known phylogenies of simulation analyses and of certain laboratory and domesticated groups are free of the uncertainties about accuracy that limit the use of well-supported trees for natural taxa in comparisons of phylogenetic methods (Hillis et al., 1994). However, these approaches are not without their own limitations (Hillis, 1995). Simulation analyses necessarily rely on simplifications of the real world and are therefore always open to questions about their relevance to reality and about investigator bias (Hillis and Bull, 1991; Miyamoto and Cracraft, 1991). Such concerns are reduced, but not eliminated, in studies of laboratory and domesticated taxa with known phylogenies because evolution in these instances is determined to a greater extent by the properties of the organisms themselves rather than by the simulator. However, even this approach does not entirely eliminate the problem of simplification because such studies are restricted to laboratory or agricultural situations, very short periods of evolutionary time, and/or the simplest biological systems (e.g., bacteriophages) (Fitch and Atchley, 1987; Hillis et al., 1992; Hillis, 1995). These limitations make it difficult to generalize the conclusions of both approaches to the more complex and diverse natural systems that characterize the tree of life (Hillis et al., 1993b; Sober, 1993).

Thus, one obvious strength of using well-supported trees of natural taxa to evaluate phylogenetic accuracy is that such tests are not limited in terms of biological complexity, as are the other two approaches using known phylogenies or simulations. By definition, the taxa in these tests are the constituents of the natural world. They are the end products of evolution over the dimensions of space and time that embody the tree of life. Thus, one obvious way to confirm the generalities of studies based on simulation analyses and on certain laboratory and domesticated groups with known phylogenies is to compare their conclusions with those of well-supported trees for different natural taxa. Known phylogenies for laboratory and domesticated groups provide important reality checks on the conclusions of simulation
studies (Hillis et al., 1994). In the same way, the general relevance of both of these approaches can be verified by tests of phylogenetic accuracy with the well-supported trees of different natural groups.

**Epilogue**

Some investigators may find it unusual or even amusing that one of us (M.M.M.) has reversed his position on the question of whether or not to combine multiple data sets (cf. Miyamoto, 1985). However, the field of phylogenetic systematics has grown considerably over the last decade, and our viewpoints on congruence have correspondingly changed with it. Agreement among the topologies of different data sets is powerful support for species relationships and the reliability of phylogenetic methods. As better procedures become available to compare the optimal and near-optimal trees of different data sets, we anticipate that tests of taxonomic congruence will come to fill a primary role in the development of phylogenetic methods as they already have in the recovery of the tree of life.

**Acknowledgments**

We thank A. de Queiroz, D. L. Swofford, M. R. Tannant, and the reviewers of our paper (J. J. Bull, D. M. Hillis, J. P. Huelsenbeck, and one anonymous referee) for their comments about the research and H. Van and G. R. Kiltie for assistance with manuscript preparation. We also thank A. G. Kluge for his comments on our manuscript, which helped us to better understand his position on the “to combine or not to combine” controversy. However, he still strongly disagrees with the conclusions of our paper and is preparing a reply that argues for character congruence and against taxonomic congruence. This study was supported by grants from the National Science Foundation (BSR-8857264 to M.M.M. and BSR-9096052 to W.M.F.) and from the National Institutes of Health (GM42397 to W.M.F.).

**References**


Rodrigo, A. G., M. Kelly-Borges, P. R. Bergquist, and P. L. Bergquist. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates


Received 1 September 1994; accepted 21 September 1994