Cladistics 15, 51–64 (1999)
Article ID clad.1999.0084, available online at http://www.idealibrary.com on IDEAL®

Noise

John W. Wenzel* and Mark E. Siddall†

**Department of Entomology, The Ohio State University, Columbus, Ohio 43210; and*
†Laboratory of Phylohirudinology, Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109

Accepted October 8, 1998

The proliferation of DNA sequence data has generated a concern about the effects of "noise" on phylogeny reconstruction. This concern has led to various recommendations for weighting schemes and for separating data types prior to analysis. A new technique is explored to examine directly how noise influences the stability of parsimony reconstruction. By appending purely random characters onto a matrix of pure signal, or by replacing characters in a matrix of signal by random states, one can measure the degree to which a matrix is robust against noise. Reconstructions were sensitive to tree topology and clade size when noise was added, but were less so when character states were replaced with noise. When a signal matrix is complemented with a noise matrix of equal size, parsimony will trace the original signal about half the time when there is only one synapomorphy per node, and about 90% of the time when there are three synapomorphies per node. Similar results obtain when 20% of a matrix is replaced by noise. Successive weighting does not improve performance. Adding noise to only some taxa is more damaging, but replacing characters in only some taxa is less so. The bootstrap and g1 (tree skewness) statistics are shown to be uninterpretable measures of noise or departures from randomness. Empirical data sets illustrate that commonly recommended schemes of differential weighting (e.g., downweighting third positions) are not well supported from the point of view of reducing the influence of noise nor are more noisy data sets likely to degrade signal found in less noisy data sets.

INTRODUCTION

For the purpose of this paper, "noise" is random data. Although this noise may form a pattern by chance, such a pattern is not due to phylogeny or to systematic error. Cracraft and Helm-Bychowski (1991:208) equated large amounts of noise with homoplasy, but the relationship is not so simple. Homoplasy can emerge as reversal or convergence on a cladogram and is analogous to noise in the present sense if there is no particular bias in how such homoplasy occurs. In contrast, as a bias grows, say, by convergent changes in limb, ear, and body shape accompanying increasingly aquatic habit, homoplasy will be less noisy in the present sense and become an adaptive signal in its own right. In a molecular framework, bias caused by base composition, codon usage, or transition–transversion ratios also confers structure on the data and is not random. Thus, there exists a spectrum between homoplasy that is noisy and homoplasy that is structured. Whether such structure is problematic (e.g., Collins *et al.*, 1994; Lockhart *et al.*, 1994) or should be seen as additional and legitimate

0748-3007/99 $30.00

historical information (see also Eyre-Walker, 1991) is outside the scope of this investigation. The present paper deals only with the end of the spectrum at which a pattern disagrees with historical ancestry by chance alone. This focus does not diminish the force of our findings because noise is frequently cited as a threat (below), but it seems to be very little studied in comparison to the issue of conflicting signals.

There is a general concern that random base changes (noise) can be a problem. One reasonable origin for this view would be that silent substitutions at the third position of a codon may not be part of any historical signal. For example, Avise (1994:34) reports "many molecular data, even those in the form of qualitative character states such as protein electromorphs or DNA sequences, are not particularly well suited for strict Hennigian cladistic analysis, in part because of the high risk of homoplasy at the level of individual electromorph or nucleotide character states." In a similar vein, Hillis (1991:278) states "In order to use these data . . . it is critical to distinguish sequences that are saturated by change from those that are phylogenetically informative." This importance of noise seems to be a given (see also Felsentstein, 1988b; Simon *et al.*, 1994), such that weighting schemes are often validated as being "an essential step in extraction of the phylogenetic signal from a background of random noise in DNA sequence change" (Knight and Mindell, 1993:18). Down-weighting or eliminating third positions is common as a way to account for saturation, that is, noise (e.g., Mindell *et al.*, 1995; Naylor and Brown, 1997). Sidow and Wilson (1992:53) discarded all transitions so that "historically relevant information can be indentified and distinguished from noise." Similarly, Naylor and Brown (1997) reinforce this view by offering "a way to select objectively for data with maximum 'signal-to-noise' potential." Yet, even as the problem grows in status, there is little to demonstrate whether it is only one of many small thorns or a crippling flaw that requires drastic surgery. It is rare to find helpful statements as to when noise is or is not a problem for phylogeny reconstruction, and the few that exist are closely tied to specific empirical examples. Friedlander *et al.* (1994) stated that a sequence divergence threshold of 20% marks the limit of reliable reconstruction; Eernisse and Kluge (1993) found that 0.2% error in recording the primary data can change results 12% of the time. Yet, for all of its apparent rarity, such work regarding the various ways in which noise can be introduced to an analysis and its effects would seem to be a necessary foundation for the perception that noise generally is bad.

One common view is that larger data sets may overwhelm smaller ones even if the smaller of the two contains more phylogenetic information. This concern is raised particularly when large molecular data sets are to be combined with smaller morphological ones (Miyamoto, 1985; Hillis, 1987; Hedges and Maxson, 1996). The root of this opinion is that the greater volume of the molecular data can include sufficient random noise to contravene phylogenetic signal in the morphological data. Yet, there is an important distinction between great qualitative morphological change and great quantitative molecular change insofar as it relates to the methodology of phylogeny reconstruction. Morphologists rightly code qualitatively different parts as different characters. Both a bird and a rhinocerous have three functional digits, but these are coded differently because one animal has a wing and the other a foot. Morphological data tend toward more characters and fewer states—in the extreme, binary data for all characters. In contrast, DNA sequence data generally do not expand the number of possible states as the taxa increase. Consequently, there are more possibilities to introduce noise into molecular data. However, largely unrecognized is the fact that because there are so many ways to arrange four states among taxa, the chance of assembling DNA data by random to controvert legitimate signal also is lower. Thus, although the chance for noise may be higher in molecular data, the chance of noisy data forming a pattern is lower. The degree to which these opposite tendencies balance has never been addressed.

This paper offers a direct examination of the effects of noise on hierarchical signal. Noise can be introduced either as a result of more (unstructured) characters added to an initial data set or by replacing data with unstructured changes as is argued to occur with long branches. We examine the effects on contrived and empirical data sets. Our direct measures of stability against noise are compared to several other measures that have been promoted to guard against poorly supported phylogenetic hypotheses and are found to have

general implications regarding combinability and the necessity of sanitizing data prior to analysis.

MATERIALS AND METHODS

We considered two fundamental ways in which noise could be investigated: addition of noise or replacement with noise. With the addition of noise, the original data are left undisturbed and a percentage of the original number of characters are added to the matrix as supplemental random characters; if, for example, there were 100 characters in the original matrix and 100% noise is "added," the new matrix would have 200 characters, half of which are random noise. The original data are left undisturbed. In this case, for each additional character added, each taxon is assigned one of four possible states chosen at random. This, then, emulates the addition of nucleotide characters that are meaningless in their entirety, not unlike the supposition that, say, third positions in codons might be wholly meaningless (e.g., Knight and Mindell, 1993; Mindell *et al.*, 1995, 1996; Mindell and Thacker, 1996; but see Cummings *et al.*, 1995). The replacement method, in contrast, confounds otherwise meaningful information in the original matrix by replacing a portion of the character states in that matrix with random states. Here, the original matrix and the perturbed matrix each have the same number of characters, but some of the original information content is replaced. If, for example, there are 20 taxa and 100 characters in the original matrix, and 10% replacement is specified, then each character state for each taxon in the original matrix is visited with the possibility of replacement with a random character state. For each cell in the matrix, an integer between 0 and 99 is chosen at random, if this number is less than the specified percentage (e.g., 10), then that character state is replaced with one of four character states chosen at random. Thus, no individual character necessarily is completely informative nor necessarily completely uninformative.

It has been suggested that trees of differing topologies might be differentially stable (Rholf *et al.*, 1990; Guyer and Slowinski, 1991; Mooers *et al.*, 1995) due to differences in information content. To examine this directly, we assessed the effects of noise on a perfectly imbalanced (pectinate) topology of eight ingroup taxa, (a b (C (D (E (F (G (H (I J)))))))), and on a perfectly balanced topology of eight ingroup taxa, (a b (((C D) (E F))((G H) (I J)))). Also, to assess the effect that noise has in relation to support, for both topologies we began with matrices having identical support for each clade and ranging from one uncontroverted synapomorphy per clade through five uncontroverted synapomorphies per clade.

For each of these topologies and each of these levels of preexisting support, we assessed the frequency of recovery of each clade defined by the unperturbed data in 1000 perturbations of each of 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000% addition of noise, and 1000 perturbations of each of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100% replacement with noise.

All perturbations were accomplished with MooToo (Siddall and Wenzel, 1997) coded in C and compiled for the MS-DOS operating system environment. The mojo option generates the perturbed data sets and spawns Hennig86 (Farris, 1988) as a daughter process—once to determine the clades in the most parsimonious tree(s) from the unperturbed data, and then to determine the clades found from each of the perturbed data sets. The frequency with which a clade is determined to be recovered is like that described by Felsenstein (1985; see also Siddall, 1995). Because some have argued in favor of weighting strategies as a suitable approach to reduce the confounding effects of noise, we repeated each of the foregoing analyses but with additional successive approximations commands included (three rounds of "xs w" and branch breaking).

In addition to the foregoing contrived data sets, we applied the same procedure to empirical data, including Carmean and Crespi's (1995) 18S rDNA data for holometabolous insect orders, Wheeler *et al.*'s (1993) arthropod data, and Hayasaka *et al.*'s (1988) mtDNA data for primates. For each data set, we calculated the frequency of recovery of clades found in the most parsimonious tree(s) in 100 replicates of 20% noise replacement. Also, g1 (skew) statistics were calculated (sensu Hillis, 1991) for each of these data sets with 100 replicates of 10, 20, 30, 40, and 50% replacement with random noise. For comparative purposes, bootstrap support (BS) values were calculated for each clade in each analysis.

RESULTS

Figures 1 and 2 illustrate, by way of shaded spline contour plots, the mojo values found for each combination of parameters investigated. The ordinate is expressed as a percentage of the original data matrix size and the abscissa is expressed in terms of the number of taxa in each clade. Darker shading indicates poorer recovery of clades defined by the unperturbed data.

Preliminary perturbations in which noise was supplemented to the original data (i.e., "added") at less than 100% of the original matrix size were so stable that we did not fully investigate these levels. Instead, we restricted analyses to increments of the size of the full matrix (i.e., 100, 200, ... 1000%). These revealed both a noticeable effect of tree shape and an expectedly strong relationship to the number of synapomorphies per clade (Fig. 1). The pectinate topology was more stable to the addition of noise than was the balanced tree, with the former requiring approximately twice the percentage of added noise to obtain a similar suppression of signal. There also was a noticeable effect of clade size, with intermediate sizes being more susceptible to loss than either the two-taxon or the eight-taxon clade. Irrespective of clade size or tree shape, one uncontroverted synapomorphy was sufficient to ensure a better than even (>50%) chance of recovering clades when only half of the data matrix was signal (i.e., additional 100% noise). The presence of three uncontroverted synapomorphies was sufficient to achieve 90% at this level of noise. Doubling the number of synapomorphies for a clade was accompanied by an approximate doubling of the amount of noise required to suppress the signal.

The ability to recover a clade when a portion of the existing signal is *replaced* with noise was less susceptible to clade size or shape (Fig. 2). All clades, regardless of number of included taxa, were recovered with approximately equal frequency in the presence of the same amount of noise replacement wherein the amount replaced was less than 60% of the matrix size. At more severe levels of replacement, a clade-size effect was noticed. Also, clades in the pectinate tree were only marginally more stable than those in the balanced tree. Generally, replacement of between 20 and 40% of the character states in these matrices was well tolerated depending on the number of synapomorphies per clade. With even only a single synapomorphy, at least one-fifth of the character states must be noise for there to be a less-than-even chance of recovery. With three synapomorphies, 20% overwrite with noise yields approximately 90% recovery, and roughly one-third of the matrix had to be replaced with noise to reduce recovery rates to less than 50%. Additional synapomorphies (>3) did not confer proportionally increased stability against noise replacement.

The application of successive approximations did not appreciably counter nor enhance the deleterious effects of having added noise to matrices or of replacing portions of the matrices with noise. A slight benefit was conferred on matrices to which random data had been added, but values were essentially identical irrespective of weighting in matrices that had a portion of their states overwritten with noise.

The 18S rDNA data set for holometabolous insects resolved five clades in a strict consensus of 27 equally parsimonious trees when (and in accordance with Carmean and Crespi's analysis) sites having at least one missing or ambiguous state were excluded (Fig. 3a). The inclusion of these sites resolved one most parsimonious tree (Fig. 3b). With or without missing and ambiguous sites, the clade delimiting all ingroup taxa was equally stable to 20% overwrite (i.e., replacement) with noise (mojo = 100) as was the dipteran clade (mojo = 100) and the clade consisting of flea and scorpionfly (mojo = 61). The neuropteran clade was more stable with missing and ambiguous data excluded (mojo = 73) than it was when these data were included (mojo = 53), whereas the halterian clade [= Strepsiptera + Diptera (see Whiting *et al.*, 1997)] was more stable with all sites included (mojo = 97) than without (mojo = 91).

The arthropod data set resolved a single optimal tree (Fig. 4) when all data were included from morphological characters, 18S rDNA sites, and ubiquitin characters. Not all of the clades found in the total evidence solution were found by the individual data set partitions. In any case, the frequency of recovery of clades in 100 replicates of 20% noise replacement (Table 1) indicates that the combined analysis generally was more stable to these perturbations than could be accounted for by any one of the individual data sets. Twenty of 23 clades were more stable in the combined analysis, and 12 of these were more stable than the sum of mojo values found in the parts. Three clades

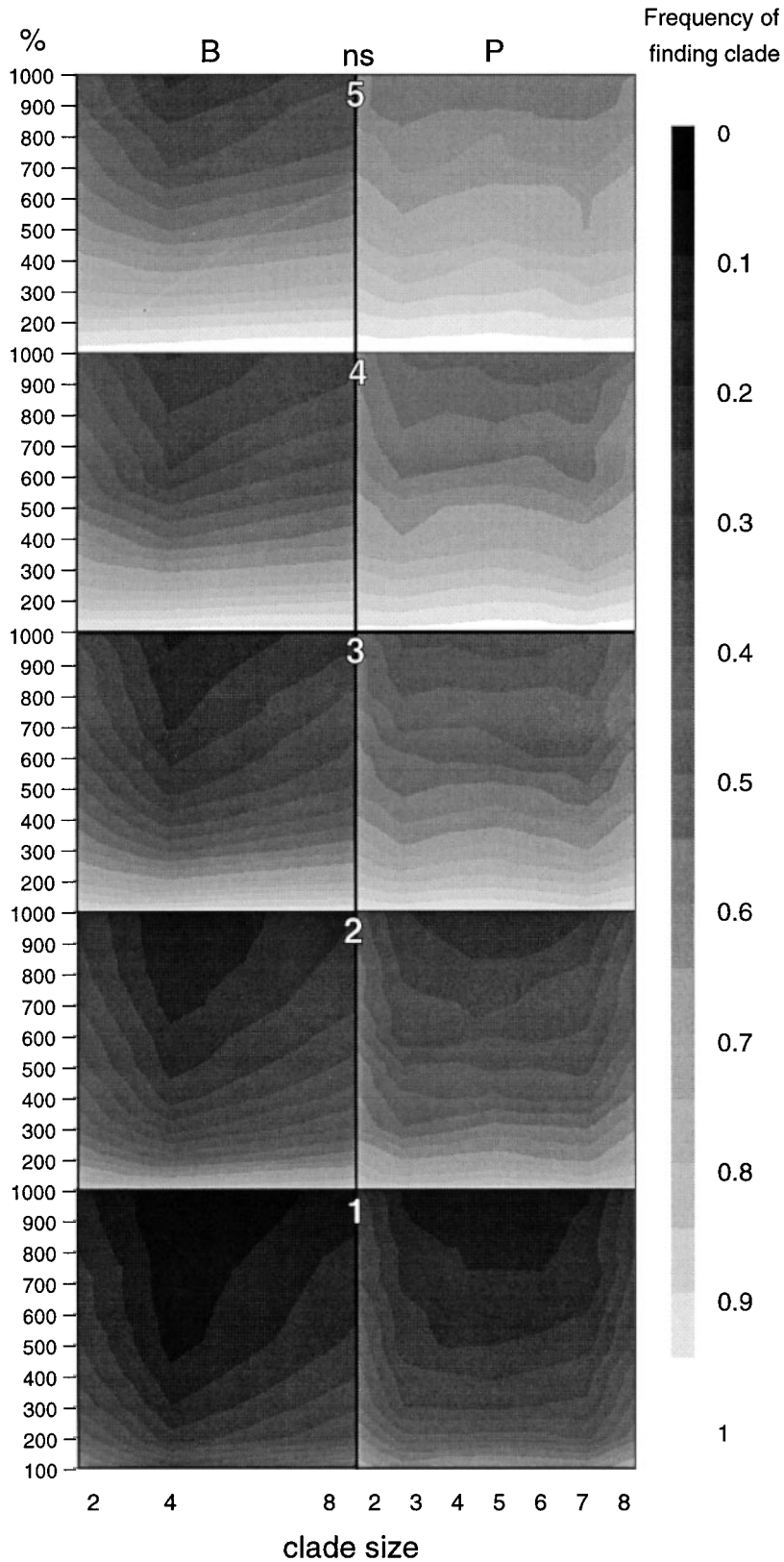

FIG. 1. Effect of adding supplemental noise to contrived matrices in 1000 perturbations for each of 10 different percentages added (%) for the fully balanced (B) and the fully pectinate (P) eight-taxon topology, and for different numbers of uncontroverted synapomorphies per clade (ns).

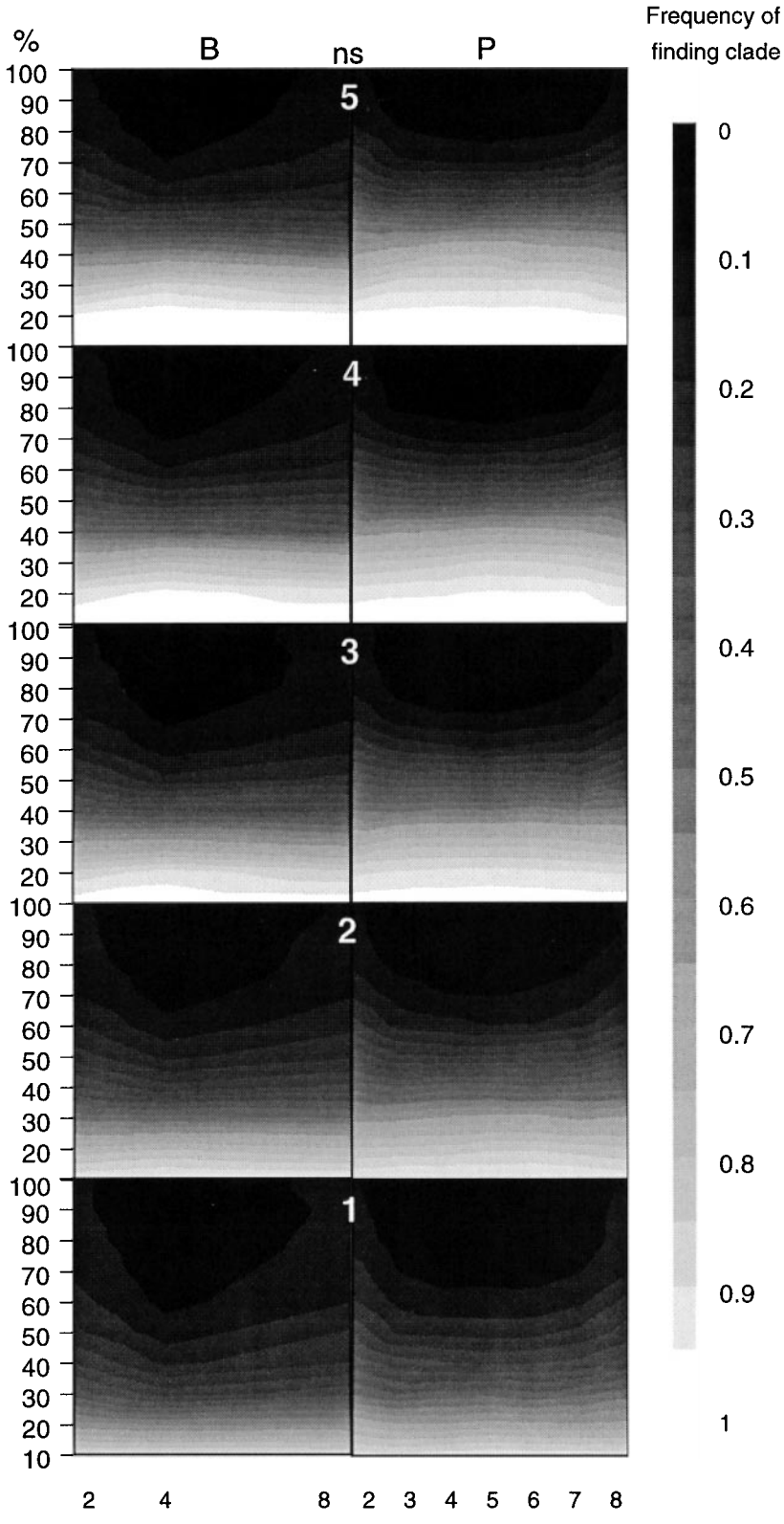

FIG. 2. Effect of replacing character states in contrived matrices with noise in 1000 perturbations for each of 10 different percentages replaced (%) for the fully balanced (B) and the fully pectinate (P) eight-taxon topology, and for different numbers of uncontroverted synapomorphies per clade (ns).

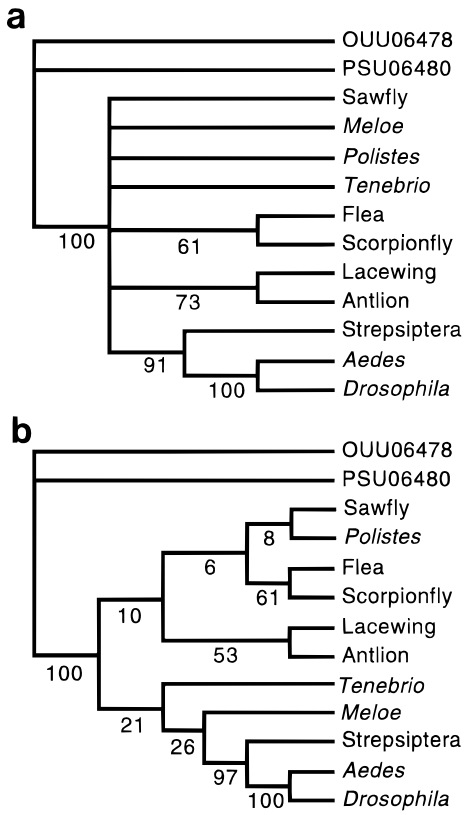

FIG. 3. Strict consensus of 27 equally parsimonious trees (a) resulting from parsimony analysis of Carmean and Crespi's (1995; see also Huelsenbeck, 1997) molecular data for holometabolous insects in which sites with missing and ambiguous states were excluded, and single most parsimonious tree (b) resulting when these sites are included. Values at nodes are mojo values obtained from 100 replicates of replacing 20% of the matrix with noise.

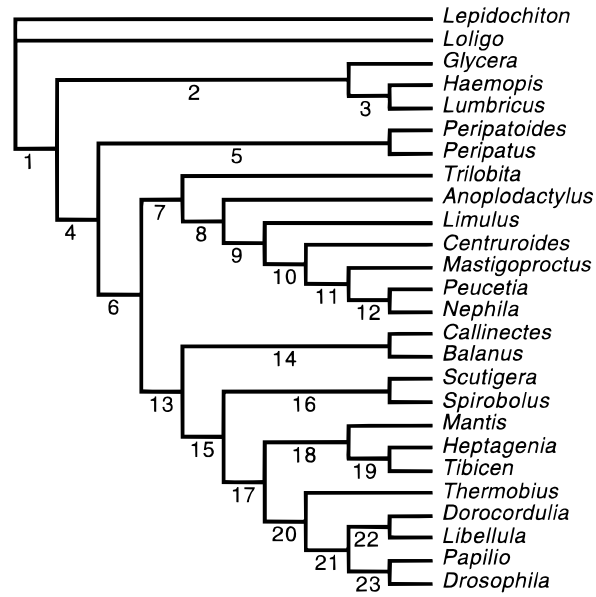

FIG. 4. Single optimal tree resulting from parsimony analysis of the combined morphological, 18S rDNA, and ubiquitin characters in the arthropod data set (see Wheeler *et al.*, 1993). Numbers at nodes correspond to clade numbers in Table 1.

in particular are notable. The arachnid clade of *Mastigoproctus*, *Peucetia*, and *Nephila* was not recovered in any of the individual data set analyses and yet had a better-than-even chance of recovery in the combined data set (mojo = 53). Similarly, two of these taxa (*Peucetia* and *Nephila*), although unstable to noise in the 18S rDNA data analysis (mojo = 28), were considerably more stable to noise in the combined analysis (mojo = 70), even though the shortest trees for the other two data sets did *not* support this clade. Similar results were found for the grouping of *Drosophila* and *Papilio*.

Figure 5a illustrates a plot of pairwise taxonomic determinations of transitions and transversions versus total divergence for the third position of codons in the primate mtDNA data set. As divergence increases, numbers of transversions increase monotonically, but

TABLE 1

Frequency of Recovery of Clades in the Arthropod Data Set in 100 Replicates of Replacing 20% of the Original Matrix with Noise

Clade	Morph	18S	Ubiquitin	All data
1	75	—	—	84
2	72	—	—	86
3	78	65	—	95
4	74	27	—	85
5	95	—	12	100
6	74	—	—	85
7	36	—	—	25
8	64	30	—	68
9	61	28	—	79
10	74	11	—	90
11	—	—	—	53
12	—	28	—	70
13	31	—	—	24
14	65	18	—	78
15	48	—	—	66
16	30	30	71	85
17	71	25	—	81
18	—	—	—	36
19	—	13	—	39
20	—	—	—	6
21	—	—	—	27
22	—	94	—	92
23	—	43	—	87

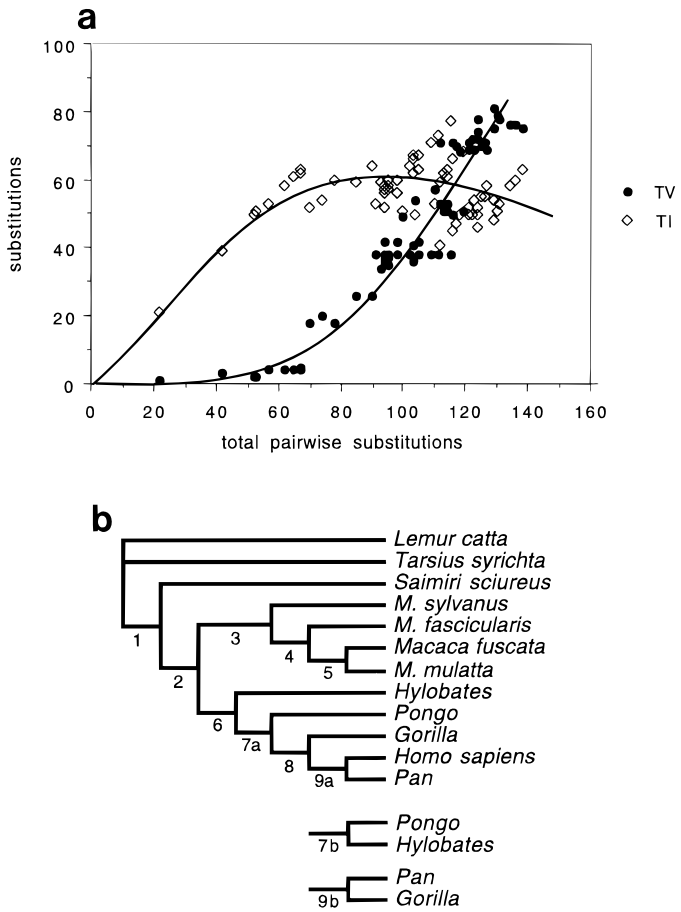

FIG. 5. The primate data set of Hayasaka *et al.* (1988) yielded a typical saturation curve for third-position transitions relative to transversions versus total divergence (a). Numbers at nodes (b) correspond to clades in Table 2.

numbers of transitions quickly reach a maximum and even begin to fall slightly at the highest divergence levels. Table 2 details the frequency of recovery of clades found in 100 replicates of 20% overwrite with noise for various partitions of the codon serial homology in the primate mtDNA data set. When first positions are considered alone, one clade was more stable to noise than it was when either second positions or third positions were considered alone. Second positions considered alone revealed five clades that were more stable than in the other two positions. Three clades were more stable to noise considering only third positions than in the foregoing two partitions. Considering only third positions, two of these clades also were more stable to noise than when first and second positions were combined. All clades were maximally

TABLE 2

Frequency of Recovery of Clades in the Primate Data Set in 100 Replicates of Replacing Portions of the Original Matrix with Noise

	Codon positions included[a]					All positions included		
Clade	1	2	3	1 and 2	1 + 2 + 3TV	20%[b]	40%[b]	50%[b]
1	48	98	87	99	100	99	84	73
2	63	76	72	85	97	93	89	67
3	98	100	99	100	100	100	100	93
4	45	65	96	81	71	99	70	53
5	23	31	94	80	76	97	56	53
6	61	78	61	96	100	97	64	48
7	26	80	56	68	78	84	59	49
7b	31	—	—	—	—	—	—	—
8	100	61	67	99	100	100	87	51
9a	60	31	78	—	—	53	50	28
9b	35	39	—	54	39	40	27	31

[a] All with 20% replacement with noise.
[b] Amount of replacement with noise for complete dataset.

stable to the addition of noise when all sites and all transformation types were included, and six clades were more stable than when third positions were excluded.

Skewness statistics (g1) obtained from tree length distributions of random trees (Fig. 6) showed a general positive relationship with increasing proportions of data matrices replaced with random noise. However, this increase was not proportional to the increase in noise. That is, for the arthropod and primate data sets, the increase appeared to be sigmoidal, with the greatest increase being between 30 and 40% replacement. Although the holometabolan data set had the most nodes unstable to addition of noise (see above), followed by the arthropod and primate data sets, respectively, g1 statistics for these data sets did not follow a similar arrangement.

DISCUSSION

Our findings suggest that, in most cladistic applications, noise may not be a general problem. The perturbations in which we "added" noisy characters as a proportion of the original informative matrix size were

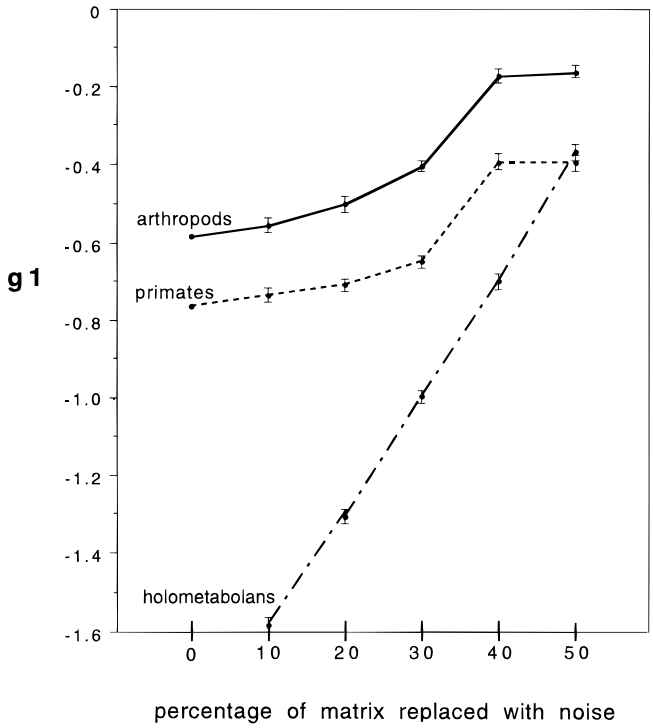

FIG. 6. Values of g1 obtained for the three empirical data sets with increasing amounts of noise overwriting the respective matrices.

intended to emulate expectations of a class of characters being wholly unstructured (e.g., third positions or gapped sites), and yet we have shown that fully one-half of the characters in an analysis must be devoid of phylogenetic signal prior to having a greater than even chance of obliterating a single uncontroverted synapomorphy (Fig. 1). We suspect that most systematists would shy away from choosing a gene with this magnitude of random nucleotide sites and would have great difficulty even coding morphological data of this sort. In particular, the concern that one-third of a matrix (third positions) might be composed only of noise can no longer serve as justification for avoiding these genes, nor necessarily for eliminating these sites because this is far less than what is required for noisy characters to contravene even one synapomorphy. The simulations that replaced (or overwrote) the character states in the original contrived matrices were intended to simulate the problem of multiple substitutions that might prevent the recovery of ancestral states. Again, parsimony analyses are seen to be remarkably stable. If the historical information in even 20% of the original matrix has been obliterated by random point mutations, the presence of one uncontroverted synapomorphy still yields a better-than-even chance of the signal being recovered. If there are three such synapomorphies (i.e., Bremer support (Bremer, 1988, 1994) or $b \geq 3$), the stability that the data have against the confounding effects of noise is impressive.

Neither of these forms of investigating noise are new ideas. Farris (1969) added wholly random characters to matrices to determine whether successive approximations character weighting would apply lower weights to characters with random phylogenetic meaning (i.e., noise). They did, and our results further corroborate this notion in that we found that slightly higher levels of added noise were required to confound successively weighted data sets. However, if matrices containing signal are converted into matrices containing characters composed partly of signal and partly of noise, as in the character state replacement simulations, there seems to be no effect of weighting. This is perhaps not surprising insofar as any weighting strategy must be applied to a character across all taxa equally, and yet the confounding effects of noise may well be restricted to only some taxa or some clades in an analysis. The matrix replacement method of overwriting character states also has been investigated previously (Eernisse and Kluge, 1993) in relation to the possible deleterious effects that reverse transcriptase sequencing errors can hold for phylogeny reconstruction. In light of the marked stability we have found here in the presence of 20% of the matrix being overwritten, the problems found by Eernisse and Kluge (1993) in the presence of only 0.2% replacement render the Haematothermia conclusions of Hedges *et al.* (1990) even more questionable than was previously thought.

The Mismeasure of Noise

Because clade size and tree balance biases were apparent in the simulations in which we added noisy characters, and not when we replaced (overwrote) signal with noise, we considered the latter procedure to be more appropriate as an index of the stability of clades to the problems that noise might pose for real empirical data sets. The mojo index of 20% noise overwrite was chosen because one uncontroverted synapomorphy preserved the clade in more than half of the pseudoreplicates, and three or more synapomorphies

yielded values exceeding 90% recovery (numbers which are perceived, by some, to be desirable).

Mooers *et al.* (1995) indicated that many published trees had less balance than would be expected from a Markovian null distribution and concluded that this was an indication that many hypotheses had been confounded by the effects of noise. Our findings support their view that pectinate trees are more stable to wholly noisy characters than are balanced topologies (Fig. 1); however, we doubt that half of the characters in very many data sets are composed only of noise. Mooers *et al.* (1995) did not acknowledge that *even if* there are many topologies with more balance than the one obtained, the one obtained still is a legitimate solution in every case they examined. Given a series of four tosses of a coin in which "heads" lands face up each time and the knowledge that there are more ways (i.e., 15) a series of four tosses will yield at least one result of "tails" face up does not deny that a series of four "heads" face up is possible and, in fact, was observed (Stoppard, 1967).

There have been attempts to discover specifics of signal and noise content in data sets. Cracraft and Helm-Bychowski (1991:209; see also Swofford, 1991) stated that constructing a majority rule consensus of the top percentile of shortest trees "permits a detailed description of the phylogenetic signal contained in the data." However, a clade not found in a tree only one step longer than the most parsimonious may yet appear in much longer trees, and there is no biological (or even mathematical) rationale for choosing the first percentile as opposed to the tenth, for example. Cracraft and Helm-Bychowski (1991, see also Helm-Bychowski and Cracraft, 1993) also advocated use of BS values as a measure of signal to noise ratios, although not in a statistical framework (contra Felsenstein, 1985). However, for bootstrapping to be relevant to noise, that noise must be distributed identically across all taxa and characters, otherwise there is no justification for the sampling routine (Felsenstein, 1985; Noreen, 1989; Carpenter, 1992). In contrast, because mojo values are obtained from applying noise across a matrix evenly, this will enhance the effects of noise (or overwhelm inadequate signal) in certain clades over others where this preexists differentially. By way of example, a 70% BS value is considered to be sufficient to represent "a true clade" (Lafay *et al.*, 1995) or to be "reliable" and "significantly" well supported (Hillis and Bull, 1993; Felsenstein and Kishino, 1993), but in the arthropod data set, two clades with approximately equal BS values (72 and 73, respectively) have markedly different susceptibilities to noise (mojo = 53 and 78, respectively).

Unlike the nodal values determined by mojo (or even BS and majority rule consensus), Hillis (1991) proposed the use of g1 (skew) statistics derived from the distributions of lengths of all possible, or randomly generated, tree topologies (for like treatments published elsewhere, see Huelsenbeck, 1991, as well as Hillis and Huelsenbeck, 1992) as a whole-tree measure of noise, and others (e.g., Lafay *et al.*, 1995; Swofford, 1991) have used this as a reliable indicator of signal strength. This idea (e.g., Fitch, 1979; Goodman *et al.*, 1979) is driven by the premise that a "good" phylogeny will have many steps between the shortest tree and the next shortest trees. The results from the empirical data sets (Fig. 6) demonstrate that although g1 is related to amount of noise (that is, as more noise is introduced to a particular data set, the g1 values are of lower magnitude for *that* data set), the change in g1 is not proportional to change in amount of noise, and it is not comparative across data sets. For two of the data sets, g1 values did not change after replacement of more than 40% of the matrix with noise. If g1 was a reasonable measure of the amount of noise in these data sets, it would suggest that this level (40%) should be sufficient to completely randomize the signal. However, it is clear (Table 2) that this is not the case. If 40% replacement with noise is sufficient to maximally reduce the magnitude of g1 (Fig. 6), it is nonetheless insufficient to prevent the recovery of most clades in the primate data set. All clades that were recovered at a frequency of greater than 50% with a 20% noise overwrite still had a better-than-even chance of recovery with the replacement of 40% of the matrix with noise, as did all but three with 50% replacement. More dramatically, even with 50% replacement with noise, the monkey clade still was recovered remarkably frequently (mojo = 93). Therefore, g1 cannot be considered a reliable indicator of the amount of noise present in a matrix in general, nor is it in any way indicative of the relative effects of noise on the various clades that comprise the most parsimonious tree(s).

Surely if one is interested in how much noise exists in a data set, and the effect that it might have on a

topology, the appropriate course of action is to use noise directly as opposed to indirect measures like BS, PTP, g1, or a majority rule consensus tree of an arbitrary portion of near-optimal trees. Although we do not advocate interpretation of mojo values with an arbitrarily chosen cutoff level as an indicator of "truth," it is clear that they can give a reasonable and comparative indication of the magnitude of sensitivity to noise in different clades. A marked sensitivity to noise, however, is just that. It is not a rationale for rejection of clades any more than highly variable bivariate data are reason to reject the least-squares line as the best estimate of the parametric regression line (see also Farris, 1983). It can, however, provide an indication of which clades might be expected to be most unstable to the addition of more data or which clades may not justify bold biological statements (cf. Huelsenbeck, 1997).

Combining Data Sets

The results reported here have strong and simple implications for the question of whether it is better to combine two data sets for phylogeny reconstruction. For noise to overwhelm signal, some character states must conspire by chance to contravene those that mark phylogeny. Our results show that signal is additive across different matrices, but that noise is averaged. Consider two data sets of equal size with *s* support and *n*% noise. Their combination will produce 2*s* support, but still only *n*% noise. This can be expected to decrease the problem of noise by half (Fig. 1), even though the analysis now includes twice as many noisy characters. This is the basis for the results obtained by Barrett *et al.* (1991), who showed that a signal too weak to determine the tree in either of two data sets can emerge to dominate in a combined analysis (see also Chippindale and Weins, 1994; Bremer, 1996; Sullivan, 1996; DeSalle and Brower, 1997). Because the probability of recovering signal increases with raw number of synapomorphies, combining a few synapomorphies from a bad data set with the few from a good data set will go a very long way to obviating concerns about noise from the bad data set. Therefore, our recommendation is to combine data sets always (see also Wenzel, 1997).

Despite all we present here, some readers still may be concerned that a small volume of good morphological data could be corrupted by a larger volume of molecular data. The arthropod data (Table 1) are instructive. With a total of twice as many informative characters contributed from the molecular data sets (200) as from the morphological data set (97). Mojo values show that, for 46 clade-by-clade comparisons between the morphological data set and either of the molecular data sets, the morphological data set is less stable to noise in only five cases (11%). This suggests that the molecular data form a substantially less coherent pattern. Yet, when all three data sets are combined, only 2 of 23 clades (clades 7 and 13) have lower mojo values than they do according to morphology alone. Thus, even though the molecular data were noisier, or have conflicting signals, hierarchical grouping was improved in 21 of 23 cases by including them (and both of the remaining 2 clades have rather low mojo values from morphology alone to begin with).

Some have argued that morphology is not worth including because molecules are more reliable (Hedges and Maxson, 1996; Givnish and Sytsma, 1997; but see Albert *et al.*, 1994; and Eyre-Walker, 1991, for discussion of molecular functional constraints). However, large and reliable data sets will not be compromised by a smaller set on the basis of noise alone (Fig. 1).

Saturation and Disinformation

Mindell and Honeycutt (1990) argued that the elimination of characters or taxa wherein multiple substitutions have occurred "will reduce the amount of phylogenetically uninformative, or misleading, nucleotide character change." Graybeal (1994) advocated the avoidance of genes showing more than 20% divergence between some taxa. Hillis and Dixon (1991) chose 30% divergence as an appropriate cutoff, whereas Friedlander *et al.* (1994) equivocated between these two values. Hillis and Dixon (1991), citing Swofford and Olsen (1990), concluded that "it is best to delete from analyses any regions where the alignment is questionable." Huelsenbeck (1997) followed this rubric in his attempt to refute the halterian clade by eliminating more than two-thirds of the data set, as did Carmean and Crespi (1995). Others have taken the approach of investigating the effects of alignment parameters more rigorously (e.g., Wheeler, 1995; Whiting *et al.*, 1997) rather than run the risk of arbitrarily defining "questionable." Often, concern is restricted to the behavior of third positions,

which "tend to become saturated with change" (Mindell *et al.*, 1996). Nanney *et al.* (1989) argued that only those sites that have changed once (i.e., ditypic) should be included. Sidow and Wilson (1992), in their modification of Lake's (1987) method of invariants, suggested that by confining their "analysis to transversional differences only, the historically relevant information can be distinguished from noise." Knight and Mindell (1993) stated that "a level of TIs near 50% indicates that this class of change is saturated with multiple changes and is therefore not an indicator of phylogeny but is largely 'noise' " (see also Mindell and Honeycutt, 1990). In some cases (e.g., Mindell *et al.*, 1995) all gapped regions in alignments *as well as* third position transitions are eliminated, even though using all of the data can refute the principle findings of such a study (Siddall, 1997).

Novice phylogeneticists (and even perhaps the well versed) surely must be frustrated with such a bewildering array of imperatives for ignoring their hard-won data points. Martin (1995) was explicit in his preference for a transversion-only solution of 10 trees over a total-evidence single tree because of "the disturbing features of [a] sister taxa relationship between *Prionace* and *Galdeocerdo*." Graur (1985) too disbelieved his parsimony analyses because they gave solutions that conflicted with his preconceived notions of relationships. If including all of the data results in a tree that coincides with conventional wisdom, would proponents of data triage still advocate the downweighting or elimination of whole portions of data, even if doing so results in a radically *unconventional* hypothesis? The argument that third positions always are uninformative, or that transitions are, has been so soundly refuted (e.g., Eyre-Walker, 1991; Albert *et al.*, 1994; Cummings *et al.*, 1995; Sullivan, 1996; Arias and Sheppard, 1996) that we cannot understand the persistence of these ideas (e.g., Mindell and Thacker, 1996).

In the very worst case, truly saturated data will not necessarily be misinformative. They might be misinformative, uninformative, or even informative. The primate data set reveals that the average sequence divergence between *Lemur* and the other taxa is 30.5%, which exceeds all of the cutoff values suggested above (Hillis and Dixon, 1991; Friedlander *et al.*, 1994; Graybeal, 1994) and this data set should be thought to be confounded by saturating noise across all sites. With respect to third positions alone, the average divergence across all taxa was 31.5%, which would appear to necessitate at least the elimination of this portion of the data. The data plotted in Fig. 5a show a classic saturation of transitions in the third position of codons in the primate data set. On the whole then, we should expect that analyses of these data should be very unstable and readily confounded by the imposition of even *more* noise. They are not. All three positions independently find all of the clades found in the combined data set. In many cases (Table 2) third positions alone find clades with greater stability than do either first or second positions alone. Moreover, in most cases, the elimination of only transitions in the third position actually reduced the stability of clades to the confounding effects of noise relative to that found when third position transitions are included. Thus, saturation curves or percentage divergences cannot be anything but misleading indications of whether one should include or exclude data. The reason for this is obvious. The absolute divergence or relative number of transitions between *Lemur* and *Homo* indicates little of substance in an analysis in which *Homo* groups with *Pan* and not with *Lemur*. It is precisely because transitions or third positions are going to be *more* informative than transversions in *recent* divergences that they should be left *in* an analysis.

Huelsenbeck's (1997; see also Mindell *et al.*, 1995) assertion that elimination of all sites with gaps somehow lessens the effects of alignment ambiguity also is readily refuted in Fig. 3. His premise for avoiding these sites was to reduce the effects of contributing noise due to alignment. That premise is seen to be empty because including those sites actually improves stability to noise.

Weighting or eliminating portions of data is done to reduce the putative effects of noise. It cannot be intended to reduce the contribution a conflicting signal might hold. With multiple signals one cannot presume to know which is the "correct" one. Phylogeneticists, even compartmentalists like Miyamoto and Fitch (1995), are interested in understanding or uncovering conflicting signals, not in eliminating or ameliorating them (contra Bull *et al.*, 1993 and Chippindale and Wiens, 1994). Nor can weighting be intended to achieve a more palatable hypothesis lest systematics be rendered unempirical [or worse, self-fulfilling (Naylor and

Brown, 1997)]. If one knows in advance what the relationships *should* be, there is not much point in looking for them.

ACKNOWLEDGMENTS

We are grateful to the students of the Cladistic Methods seminar at OSU for discussions that motivated us to write this paper, in particular Brian Mark and Brady Porter. Special thanks for generous hospitality go to Donna, proprietor of the Outgroup Inn, Dublin, Ohio. James Carpenter provided timely and reasonably-priced electronic delivery of the arthropod data set. We found much that was relevant to the publication of this study on the World Wide Web, specifically Kluge's phylogenetic literature database.

REFERENCES

Albert, V. A., Buckland, A., and Bremer, K. (1994). DNA characters and cladistics: The optimization of functional history. *In* "Models in Phylogeny Reconstruction" (R. W. Scotland, D. L. Siebert and D. M. Williams, Eds.), pp. 249–272. Clarendon Press, Oxford.

Arias, M. C., and Sheppard, W. S. (1996). Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L.) inferred from mitochondrial DNA sequence. *Mol. Phylog. Evol.* **5,** 557–566.

Avise, J. C. (1994). "Molecular Markers, Natural History and Evolution." Chapman & Hall, New York.

Barrett, M., Donoghue, M. J., and Sober, E. (1991). Against consensus. *Syst. Zool.* **40,** 486–493.

Bremer, B. (1996). Combined and separate analyses of morphological and molecular data in the plant family Rubiaceae. *Cladistics* **12,** 21–40.

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.

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

Carmean, D., and Crespi, B. J. (1995). Do long branches attract flies? *Nature* **373,** 666.

Carpenter, J. M. (1992). Random cladistics. *Cladistics* **10,** 215–220.

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

Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* **43,** 482–496.

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

Cummings, M. P., Otto, S. P., and Wakeley, J. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12,** 814–822.

DeSalle, R., and 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.

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

Eyre-Walker, A. C. (1991). An analysis of codon usage in mammals: Selection or mutation bias? *J. Mol. Evol.* **33,** 442–453.

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

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

Farris, J. S. (1988). "Hennig86." Software and documentation. Port Jefferson Station, NY.

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

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39,** 783–791.

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

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

Friedlander, T. P., Regier, J. C., and Mitter, C. (1994). Phylogenetic information—content of 5 nuclear gene-sequences in animals—initial assessment of character sets from concordance and divergence studies. *Syst. Biol.* **43,** 511–525.

Givnish, T. J., and Sytsma, K. J. (1997). Consistency, characters, and the likelihood of correct phylogenetic inference. *Mol. Phylog. Evol.* **7,** 320–330.

Goodman, M., Czelusniak, J., and Moore, G. W. (1979). Further remarks on the parameter of gene duplication and expression events in parsimony reconstructions. *Syst. Zool.* **28,** 379–385.

Graur, D. (1985). Pattern of nucleotide substitution and the extent of purifying selection in retroviruses. *J. Mol. Evol.* **24,** 221–231.

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

Guyer, C., and Slowinski, J. B. (1991). Comparison of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution* **45,** 340–350.

Hayasaka, K., Gojobori, T., and Horai, S. (1988). Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.* **5,** 626–644.

Hedges, S. B., and Maxson, L. R. (1996). Re: Molecules and morphology in amniote phylogeny. *Mol. Phylog. Evol.* **6,** 312–314.

Hedges, S. B., Moberg, K. D., and Maxson, L. R. (1990). Tetrapod phylogeny inferred from 18S and 28S riboosomal RNA sequences and a review of the evidence for amniote relationships. *Mol. Biol. Evol.* **7,** 607–633.

Helm-Bychowski, K., and Cracraft, J. (1993). Recovering phylogenetic signal from DNA sequences: Relationships within the corvine assemblage (class Aves) as inferred from complete sequences of the mitochondrial DNA cytochrome-b gene. *Mol. Biol. Evol.* **10,** 1196–1214.

Hillis, D. M. (1987). Molecular versus morphological approaches to systematics. *Annu. Rev. Ecol. Syst.* **18,** 23–42.

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

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42,** 182–192.

Hillis, D. M., and Dixon, M. T. (1991). Ribosomal DNA: Molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66,** 411–453.

Hillis, D. M., and 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.

Knight, A., and Mindell, D. P. (1993). Substitution bias, weighting of DNA sequence evolution, and the phylogenetic position of Fea's viper. *Syst. Biol.* **42,** 18–31.

Lafay, B., Smith, A. B., and Christen, R. (1995). A combined morphological and molecular approach to the phylogeny of asteroids (Asteroidea: Echinodermata). *Syst. Biol.* **44,** 190–208.

Lake, J. A. (1987). Determining evolutionary distances from highly diverged nucleic acid sequences. *J. Mol. Evol.* **26,** 59–73.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11,** 605–612.

Martin, A. P. (1995). Mitochondrial DNA sequence evolution in sharks: Rates, patterns, and phylogenetic inferences. *Mol. Biol. Evol.* **12,** 1114–1123.

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

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

Mindell, D. P., Shultz, J. W., and Ewald, P. W. (1995). The AIDS pandemic is new, but is HIV new? *Syst. Biol.* **44,** 77–92.

Mindell, D. P., Knight, A., Baer, C., and Huddleston, C. J. (1996). Slow rates of molecular evolution in birds and the metabolic rate and body temperature hypothesis. *Mol. Biol. Evol.* **13,** 422–426.

Miyamoto, M. M. (1985). Consensus cladograms and general classifications. *Cladistics* **1,** 186–189.

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

Mooers, A. Ø., Page, R. D. M., Purvis, A., and Harvey, P. H. (1995). Phylogenetic noise leads to unbalanced cladistic tree reconstructions. *Syst. Biol.* **44,** 332–342.

Nanney, D. L., Preparata, R. M., Preparata, F. P., Meyer, E. B., and Simon, E. M. (1989). Shifting ditypic site analysis: Heuristics for expanding the phylogenetic range of nucleotide sequences in Sankoff analyses. *J. Mol. Evol.* **28,** 451–459.

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

Noreen, E. W. (1989). "Computer-Intensive Methods for Testing Hypotheses: An Introduction." Wiley, New York.

Rholf, F. J., Chang, W. S., Sokal, R. R., and Kim, J. (1990). Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. *Evolution* **44,** 1671–1684.

Sidow, A., and Wilson, A. C. (1990). Compositional statistics—an improvement of evolutionary parsimony and its application to deep branches in the tree of life. *J. Mol. Evol.* **31,** 51–68.

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

Sidall, M. E. (1997). The AIDS pandemic is new, but is HIV not new? *Cladistics* **12,** 267–273.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., and Flook, P. (1994). Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annu. Entomol. Soc. Am.* **87,** 651–701.

Stoppard, T. (1967). "Rosencrantz and Guildenstern are Dead." Grove Press, New York.

Sullivan, J. (1996). Combining data with different distributions of among-stie variation. *Syst. Biol.* **45,** 375–380.

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

Swofford, D. L., and Olsen, G. J. (1990). Phylogeny reconstruction. *In* "Molecular Systematics" (D. M. Hillis and C. Mortiz, Eds.), pp. 411–501. Sinauer Associates, Inc., Sunderland, MA.

Wenzel, J. W. (1997). When is a phylogenetic test good enough? *Mem. Mus. Nat. D'Histoire Naturelle (Paris)* **173,** 31–45.

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

Wheeler, W. C., Cartwright, P., and Hayashi, C. Y. (1993). Arthropod phylogeny: A combined approach. *Cladistics* **9,**1–39.

Whiting, M. F., Carpenter, J. C., Wheeler, Q. D., and Wheeler, W. C. (1997). The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Syst. Biol.* **46,** 1–68.