Syst. Biol. 49(4):628–651, 2000

Integrating Ambiguously Aligned Regions of DNA Sequences in Phylogenetic Analyses Without Violating Positional Homology

FRANÇOIS LUTZONI,[1] PETER WAGNER,[2] VALÉRIE REEB[1], AND STEFAN ZOLLER

Departments of [1]Botany and [2]Geology, The Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago, Illinois 60605-2496, USA; E-mail: flutzoni@fmnh.org

Abstract.—Phylogenetic analyses of non-protein-coding nucleotide sequences such as ribosomal RNA genes, internal transcribed spacers, and introns are often impeded by regions of the alignments that are ambiguously aligned. These regions are characterized by the presence of gaps and their uncertain positions, no matter which optimization criteria are used. This problem is particularly acute in large-scale phylogenetic studies and when aligning highly diverged sequences. Accommodating these regions, where positional homology is likely to be violated, in phylogenetic analyses has been dealt with very differently by molecular systematists and evolutionists, ranging from the total exclusion of these regions to the inclusion of every position regardless of ambiguity in the alignment. We present a new method that allows the inclusion of ambiguously aligned regions without violating homology. In this three-step procedure, first homologous regions of the alignment containing ambiguously aligned sequences are delimited. Second, each ambiguously aligned region is unequivocally coded as a new character, replacing its respective ambiguous region. Third, each of the coded characters is subjected to a specific step matrix to account for the differential number of changes (summing substitutions and indels) needed to transform one sequence to another. The optimal number of steps included in the step matrix is the one derived from the pairwise alignment with the greatest similarity and the least number of steps. In addition to potentially enhancing phylogenetic resolution and support, by integrating previously nonaccessible characters without violating positional homology, this new approach can improve branch length estimations when using parsimony. [Character coding; character-state weighting; crocodile 12S mtrDNA; indel-rich region; insect 16S mtrDNA; intron; large-scale phylogenies; molecular phylogenies; multiple sequence alignment; non-protein-coding DNA sequences; ribosomal RNA genes.]

Among the most fundamental practices in systematic studies are the characterization of intrinsic features of organisms, character coding, and the assessment of homology (Wiley, 1981; Patterson, 1982, 1988; Roth, 1988; Stevens, 1991). Because of the nature of nucleotide sequences, the two first aspects are not so much a concern for molecular systematists and evolutionists. Homology assessment, however, is a major concern (Mindell, 1991; Hillis, 1994; Doyle and Davis, 1998). In this paper we assume that the nucleotide sequences compared are orthologous. The new method described here tackles the problem of positional homology when the position of a specific nucleotide is equivocal because of the potential presence of nearby gaps. This problem is particularly acute when aligning fast-evolving or highly diverged non-protein-coding portions of genomes, a situation frequently encountered in large-scale phylogenetic studies. As sequences from more distant taxa are added to an alignment covering a broad spectrum of organisms, these ambiguously aligned regions have a tendency to become larger and more numerous. If these regions are included in phylogenetic analyses, fundamental assumptions of homology are likely to be violated and phylogenetic accuracy might be lowered considerably. If excluded, however, resolving power and branch length estimates may be jeopardized.

Ambiguous alignments, or multiple equally optimal alignments, are most easily detected when sequences vary in length. This length variation requires inserting gaps in the alignment to maintain positional homology. The number of gaps needed and their exact position can be uncertain. Most alignment procedures necessitate the assignment of a cost for a nucleotide substitution versus the insertion of a gap (DeSalle et al., 1994; Wheeler, 1994). Different parameters often result in the inclusion of various numbers of gaps that can also vary in their position. The inability to justify one set of parameters over another (Vingron and Waterman, 1994; Kjer, 1995; Doyle and Davis, 1998) leads to alternative sequence alignments for the same data set. Problems occur when different topologies are revealed by phylogenetic analyses of these different alignments (Wheeler, 1995; Wheeler et al., 1995; Soltis et al., 1996). Thorne et al. (1991, 1992) developed an

objective technique to estimate alignment parameters, based on an evolutionary model using a maximum likelihood approach. In their method, the weights of evolutionary events such as nucleotide substitutions and indels are a function of evolutionary rates and divergence times, estimated from the two sequences to be aligned. However, this method has yet to be extended to the simultaneous alignment of more than two sequences.

Another source of ambiguity is the presence of multiple equally optimal alignments for a specific set of alignment parameters (Wheeler et al., 1995). This is best visualized as multiple equally optimal paths (Kruskal, 1983; Weir, 1990; Wheeler, 1994) when aligning two sequences by using a method derived from the dynamic programming algorithm of Needleman and Wunsch (1970). Finally, the order in which sequences are integrated in a multiple alignment process can affect the resulting alignments (Lake, 1991; Mindell, 1991). In this context, the question becomes which order should be chosen. This situation leads to the multiple alignment problem described by Gatesy et al. (1993) and Wheeler (1995). The consequences of these multiple potential alignments for a given data set can be severe because different alignments can support different topologies regardless of the degree of sophistication of the phylogenetic method used afterward. This can lead systematists to conflicting conclusions (Cerchio and Tucker, 1998). Despite the primary importance of positional homology in phylogenetic studies, alignment issues have received far less attention and discussion by systematists than theoretical aspects of phylogenetic reconstruction (Doyle and Davis, 1998).

There are several options available to phylogeneticists dealing with ambiguous regions of an alignment. These different methods can lead to different phylogenetic results (e.g., Vogler and DeSalle, 1994). A common, and often considered the most conservative, approach is the exclusion of these regions from phylogenetic searches (e.g., Bruns et al., 1992; Berbee and Taylor, 1993; Hibbett and Vilgalys, 1993; Spatafora and Blackwell, 1993; Lutzoni, 1995; Spatafora, 1995; Berbee, 1996). However, the subjectivity associated with this process can lead to different phylogenies, depending on which combination of sites is excluded. The other main caveat of this method is the loss of resolution (Lutzoni, 1995; Wheeler et al., 1995) and the removal of data essential for a more accurate estimation of branch lengths. This is because ambiguously aligned regions can contain a large fraction (sometimes >50%) of all potentially parsimony-informative characters present in a given alignment. For example, Fernandez et al. (1999) used a 1-kb fragment at the 5′ end of the large subunit nuclear ribosomal RNA (nrRNA) gene to unveil relationships of a group of pyrenomycetous ascomycetes at the ordinal and family level. In all, 16 ambiguously aligned regions covering 157 sites and corresponding to 17% of the entire alignment were delimited. They estimated that these ambiguous regions would provide 120 parsimony-informative characters, corresponding to ~40% of all parsimony-informative characters that the entire data set would provide. The small subunit nrRNA gene, which is on average more conserved than the large subunit nrRNA gene, is not devoid of this problem. For example, Lutzoni et al. (in prep.), in their phylogenetic study of lichenized and nonlichenized ascomycetes, found that, at the ordinal level, 20% of all potential parsimony-informative characters resided in ambiguously aligned regions of a specific alignment of the 1-kb fragment they sequenced. This percentage goes up as more distantly related taxa (such as basidiomycetes) are included in the alignment.

The exclusion or inclusion of ambiguously aligned regions can have a determinant effect on the results of phylogenetic analyses (Giribet and Wheeler, 1999). This was the case for the controversial Mysticeti/Physeteroidea clade, for which Cerchio and Tucker (1998) showed that the phylogenetic signal supporting this hypothesis was contained predominantly in the ambiguously aligned regions of the 12S and 16S mitochondrial ribosomal DNA (mtrDNA). The different ways systematists working on these data sets treated these regions resulted in incongruences.

Another problem associated with the exclusion of ambiguously aligned regions is how to delimit them accurately. The use of a range of gap-to-substitution cost ratios (e.g., ranging from 2:3 to 300:1) has been proposed to circumscribe alignment-ambiguous sites (Waterman et al., 1992; Gatesy et al., 1993). In such a procedure, sites that are not constant among all the alignments resulting from the different cost ratios are

considered ambiguously aligned and subsequently removed from phylogenetic analyses. The use of extreme cost ratios, however, such as 2:3 and 300:1, causes unambiguous regions to be unstable among alignments, such that parsimony-informative sites are removed that are clearly not violating positional homology. Even if this method were more accurate in delimiting ambiguously aligned regions, this does not alleviate the loss in resolution and branch length accuracy associated with the exclusion of data. The other extreme, and by far the worst-case scenario, is the inclusion of all sites in the phylogenetic analysis based on one of many equally most optimal alignments and simultaneously treating gaps as a fifth character state. One problem with this strategy is the overweighting of adjacent gaps by treating them as independent indels when they are very likely part of a single genetic change.

To minimize the detrimental effect of inserting gaps in an alignment, gaps have been treated as missing data. Although it might be tempting to believe that replacing gaps by question marks in ambiguously aligned regions is a safe alternative, even if gaps are treated as missing data, their positional homology remains highly questionable and very likely wrong. When such ambiguously aligned sites are parsimony-informative, they will often have a negative effect on phylogenetic accuracy. This practice has the unfortunate potential of generating highly resolved trees, which is erroneously interpreted as a sign of phylogenetic accuracy (Hillis et al., 1994) and can give a sense of overconfidence in the resulting topologies. As was demonstrated by Hillis and Huelsenbeck (1992), the analysis of random molecular data sets can yield a single most-parsimonious (highly resolved) tree that may also be considerably shorter than the second-best alternative. Bremer (1988) support values generated for this single most-parsimonious, but meaningless, tree might even increase. In our opinion, the inclusion of all sites of an alignment with ambiguously aligned regions in a phylogenetic analysis should be done only if all competing alignments are shown to generate the same topology. At least, the sensitivity of the phylogenetic analysis to the inclusion of various alignments for a given ambiguous region should be explored (see Baum et al., 1994).

Hibbett et al. (1995) pointed out two other caveats associated with the treatment of gaps as missing data: (1) the exclusion of potentially parsimony-informative indels as characters, and (2) the potential for assigning impossible states to ancestors (Platnick et al., 1991; Maddison, 1993). Barriel (1994) developed a method that combined the use of question marks and the implementation of a coding scheme for regions with multiple adjacent gaps. The method proposed by Barriel (1994) takes care of the first problem mentioned by Hibbett et al. (1995) but not the second one. This is because treating gaps as missing data, especially when the placement of gaps is not ambiguous and when the gaps are not treated as a fifth character state, can result in the ancestors of taxa with gaps being assigned a nucleotide unequivocally when, more likely, the ancestor had a gap at that position. Finally, several hybrid approaches consisting of excluding some indel-rich regions and including others as coded without treating gaps as missing data, or where gaps are considered as missing information, have been implemented to maximize the integration of phylogenetic signal provided by indels without violating criteria for positional homology (e.g., Baldwin et al., 1995; Hibbett et al., 1995; Kjer, 1995; Kretzer et al., 1996; Manos, 1997). All of the above methods assume that gap positions are correct, that is, unambiguous, which is often not the case. Therefore, ambiguously aligned sequences with indels of variable lengths and equivocal positions are still excluded from phylogenetic analyses.

Wheeler et al. (1995) suggested a solution to the problem of equally optimal alignments and multiple alignments resulting from different cost parameters that can lead to different phylogenetic trees. Their method, termed elision, consists of joining end to end all optimal alignments obtained from all cost parameters into a single grand alignment. Sites that are identical among all combined alignments will have, by default, the most weight, corresponding to the number of optimal alignments that were fused. Positions that vary among alignments are automatically downweighted proportionally to the degree of interalignment variability. We have identified four problems with the elision approach. First, all equally optimal alignments for a high number of cost parameters and for different orders of sequence entry

in the alignment process should be part of the grand alignment; otherwise, the resulting topology or topologies could be biased. Finding all these optimal alignments could be problematic when dealing with large data sets, that is, with a high number of OTUs, or with highly diverged sequences. Moreover, the number of gap:substitution cost ratios used has a direct impact on the result. The greater the number of cost ratios used, the greater the potential to find a high number of different alignments and the more downweighted the variable sites will be. Second, although the proportional downweighting of these ambiguously aligned sites compensates to some degree for the introduction of phylogenetic noise, this weighting scheme is too drastic when numerous equally optimal alignments are involved. For most interalignment variable sites this procedure becomes equivalent to the exclusion of these sites, when they might be essential for resolving specific portions of the phylogenetic tree. Third, for data sets with highly divergent sequences and relatively poor resolving power, this could involve the fusion of a prohibitively high number of equally optimal alignments. Fourth, as pointed out by Wheeler et al. (1995:5), this method has "the disturbing property of assigning multiple putative homologies to the same datum"; therefore, it inevitably introduces many sites for which positional homology is violated.

On the basis of the initial work of Sankoff and Cedergren (1983), Feng and Doolittle (1987, 1990), Hein (1990), and others to align sequences and reconstruct phylogenies simultaneously, Wheeler (1996) proposed a new approach for the analysis of ambiguously aligned sequences (POY). His method proceeds directly from the original nucleotide sequences to phylogeny reconstruction; that is, it does not first insert gaps in a multiple sequence alignment. This direct optimization of DNA sequences offers a potential solution to the problem of integrating ambiguously aligned regions in phylogenetic analyses without violating positional homology; however, its implementation might be too time consuming, especially when dealing with large-scale phylogenies. Moreover, with a direct optimization approach, according to Giribet and Wheeler (1999), "there is no way to disregard gap information because there is no intermediate step (alignment)". Therefore, this practice of eliminating the use of alignments prevents the potential detection of (1) site-to-site variation in terms of parameters such as transition:transversion ratios, rates of nucleotide substitution, and base frequency biases; (2) regions saturated by changes that could mislead the search toward incorrect topologies because of the loss of most, if not all, of the phylogenetic signal; (3) the contribution of ambiguously aligned sites to the resulting phylogenetic tree; and (4) sequences so divergent (e.g., outgroup compared with ingroup sequences) that an alignment step would quickly reveal these sequences could not be incorporated into a given phylogenetic analysis.

Because gaps and ambiguously aligned regions are a class of molecular characters that can be exceptionally reliable for phylogenetic analyses (Lloyd and Calder, 1991; Giribet and Wheeler, 1999), it is crucial to find a viable way to integrate this type of information in phylogenetic analyses. In this paper we present a new method that accommodates every type of ambiguously aligned region except the ones where saturation caused by multiple changes has most likely resulted in the complete loss of phylogenetic signal. Such regions still should be excluded from phylogenetic analyses (Swofford et al., 1996). The method presented here also provides a criterion to detect the ambiguously aligned regions most likely to be saturated by multiple changes. We have restricted this paper to the case of nucleotide sequences, but this method can be extended to accommodate ambiguous alignments of amino acid sequences. The first part of this paper describes the three steps of this new procedure: (1) delimiting of homologous regions that contain ambiguously aligned sequences, (2) unequivocal coding of homologous ambiguously aligned regions, and (3) optimal weighting of changes among character states (step matrices) of unequivocally coded regions. Ambiguously aligned regions are delimited by sliding gaps laterally until there is no justification to push the gap further, based on nucleotide composition of neighboring sites. Each ambiguously aligned region is recoded as a single character without involving any gaps and without violating positional homology. Gaps are used only to determine the optimal number of steps necessary to go from one coded character state to another by reducing the alignment

task from comparing multiple sequences to only pairwise comparisons. The second part of this paper will use this new approach to reassess the insect 16S mtrDNA and crocodile 12S mtrDNA data sets that were analyzed previously by the elision method (Wheeler et al., 1995).

Materials and Methods

Data Sets and Alignments

We used the same nucleotide sequences (insect 16S and crocodile 12S mtrDNA) used by Wheeler et al. (1995). The 16S data as presented in that publication has two serious problems. First, the reverse complement form is shown; second, some of the sequences differ from those reported in GenBank. To facilitate the comparison between the two papers, and to ensure that differences in our respective analyses are strictly the result of using different methods, these two data sets were used here exactly as presented in Wheeler et al. 1999. The sequences were aligned by using the automated assembly option of Sequencher 3.0 (Gene Codes) and subsequently optimized by eye. Using the same program, the resulting alignment was exported as a Nexus file for phylogenetic analyses.

Phylogenetic Analyses

All phylogenetic analyses were implemented by using the maximum parsimony optimization criterion in PAUP* 4.0d64 (Swofford, 1998). Constant sites were removed from all analyses. For the unambiguously aligned sites, a step matrix was constructed from the negative natural logarithms of the relative frequencies of each possible transformation (Felsenstein, 1981; Wheeler, 1990; Maddison and Maddison, 1992). The frequency of each possible genetic change was estimated with the PAUP* option that provides the current status of each character. This command provides a list of all character states for each site of interest, from which the frequency of all potential changes can be compiled. These unambiguous sites were subjected to this step matrix in all phylogenetic analyses.

All ambiguously aligned regions were excluded from phylogenetic analyses. However, the coded versions of these regions were added to the end of the data matrix and each coded character was subjected simultaneously to its own step matrix. The coding and the building of these step matrices are described in detail in steps 2 and 3 below. The branch-and-bound search algorithm was implemented, branches were collapsed if maximum branch length was 0, and MULPARS was in effect. The support for the internodes of the most-parsimonious trees was estimated by 1,000 bootstrap replicates (Felsenstein, 1985) using the branch-and-bound algorithm and the same options described above. The Wilcoxon signed-ranks test (Templeton, 1983), as implemented in PAUP* 4.0d64, was used to determine whether some topologies were significantly worse than the best topology.

Integration of Ambiguously Aligned Nucleotide Sequences in Maximum Parsimony Analyses

Step 1: Delimitation of Homologous Ambiguously Aligned Regions

The first step of the method presented here is to delimit homologous regions of sequences for which the number of gaps inserted or their placement (or both) is ambiguous. The advantage of delimiting ambiguous regions for their exclusion and coding is that the alignment within these regions does not need to be optimized. These are the regions most sensitive to the order in which the sequences are aligned (Lake, 1991; Mindell, 1991) or to the gap-to-substitution cost ratios used in the alignment process (Gatesy et al., 1993). Theoretically, the alignment procedure has five possible outcomes:

1. regions that lack indels no matter which parameters or entry order of sequences was used (Fig. 1a);
2. regions that consistently include the same number of indels with identical placement across all optimal alignments (Figs. 1b, c);
3. regions in which the presence/absence of indels varies when alignments derived from different gap-to-substitution cost ratios are compared, but their positions, when present, does not change among all optimal alignments (Fig. 1d);
4. regions that consistently include a specific number of indels, but their placement varies among optimal alignments (Fig. 1e); and
5. regions that combine variation in the number and position of indels when compared across all putative alignments (Fig. 1f).

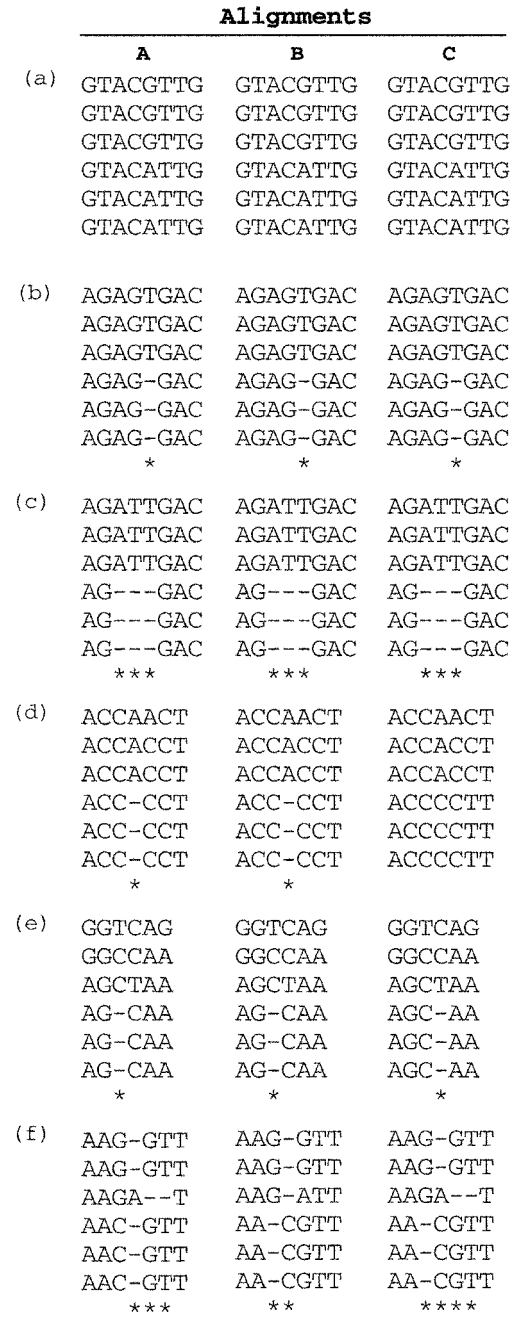


FIGURE 1. Six different regions (a–f) with three hypothetical alignments (A–C) for each. An asterisk indicates the position of an indel, and a dash represents a gap. See text (Step 1) for descriptions of a–f.

The first three cases (Figs. 1a–c) described above are unambiguous even if cases b and c include indels. The first case is not problematic and can be included "as is" in phylogenetic analyses. The second and third cases can be problematic even if the placement of gaps is not ambiguous and their number is identical for all sequences with one or more gaps in that region. The problem resides in the weighting scheme. In the case presented in Figure 1b, where we have a 1×3 block of gaps, the transformation T⟷– can be attributed a cost of 1, or >1 if based on the negative natural logarithm of the probability of that change (Felsenstein, 1981; Wheeler, 1990; Maddison and Maddison, 1992). The problem is accentuated when several gaps are adjacent to one another (e.g., 3×3, Fig. 1c). If each adjacent gap is considered part of an independent indel region, this region can quickly become overweighted. Indels that are likely to be part of one genetic event, that is, more than one gap long and of equal length across all taxa at a specific site, can be coded as a new character that replaces that region (e.g., Manos, 1997). However, the estimation of the cost for the transformation of one sequence to another, for sequences that are part of a specific indel region, compared with the cost of this insertion or deletion, remains problematic. The weighting scheme presented here for character-state changes within a coded character partially solves this problem.

The fourth example (Fig. 1d) represents a special case, in which the placement of the indel is not ambiguous but its presence or absence is. This case is likely when an excessively low gap-to-substitution cost ratio is included as part of the range of cost ratios used to reveal different potential alignments. Such low cost ratios should be used only when dealing with highly divergent sequences of highly variable lengths, for which a high frequency of gaps are expected. However, the sequences in such cases are expected to be so different that alignment is not possible, and their orthology or paralogy is questionable. This is an additional reason not to use a range of gap:substitution cost ratios to delimit ambiguous regions. If the presence of gaps is extremely frequent for a given alignment, a totally different approach should be used (see region of data set of internal transcribed spacer region of Lutzoni, 1995). A multiple alignment method derived from the maximum likelihood approach proposed by Thorne et al. (1991, 1992), where a cost ratio would be estimated for each pair of sequences compared, might solve this problem. The new method presented here does

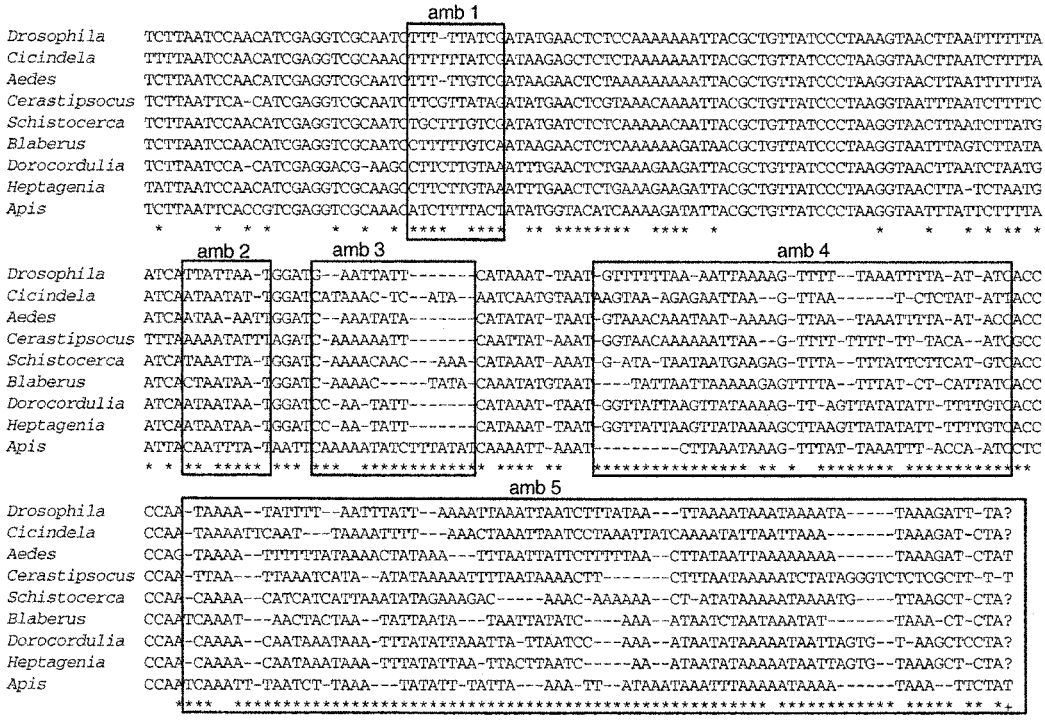


FIGURE 2. Alignment generated in this study from the insect 16S mtrDNA data set as presented by Wheeler et al. (1995). Note that this data set was inadvertently written in its reverse complement form by Wheeler et al. (1995), and that their sequences for some taxa are different from the sequences found in GenBank. To facilitate the comparison with their paper and to ensure that differences in our respective results were strictly due to different methods, however, we used the same data set as presented in their 1995 publication. Boxes delimit homologous ambiguously aligned sequences. −, gap; ?, missing data; ★ variable site; +, constant site including missing data. All sites without the symbols ★or + on the line below the alignment are constant and without missing data.

not address this specific case of presence or absence of indels in a given region of the alignment.

The fifth and sixth cases (Figs. 1e, f) represent typical situations where many systematists will choose to exclude ambiguous regions, especially when the different placement of gaps, even when treated as missing data, suggests conflicting phylogenetic relationships. The method we present was specifically designed to deal with the cases illustrated in Figures 1c, 1e, and 1f, but can also be applied to case 2 (Fig. 1b).

In our proposed approach, alignment is seen as the procedure for determining where the variation in length among sequences is located. This can usually be achieved by implementing commonly used alignment programs such as CLUSTAL (Higgins et al., 1996), Sequencher, and others, followed by a thorough inspection by eye to correct obvious alignment errors and determine where the gaps are located.

Figures 2 and 3 provide examples of delimited ambiguous regions from the insect 16S mtrDNA and crocodile 12S mtrDNA data sets respectively, of Wheeler et al. (1995). The approach used here was as follows:

1. Inspect each region with at least one gap.
2. Slide the gap(s) laterally, in an outward direction from where they are located, to determine whether the nucleotide compositions at adjacent sites, and the secondary structure, can provide any justification for alternative position(s) for the gap(s).
3. Continue this outward sliding of gaps, in both directions, until the sliding of gaps, by one more position cannot be justified, thus marking the boundaries for that region.
4. If the nucleotide compositions at adjacent sites or if the secondary structure justifies an alternative position for the gap(s)

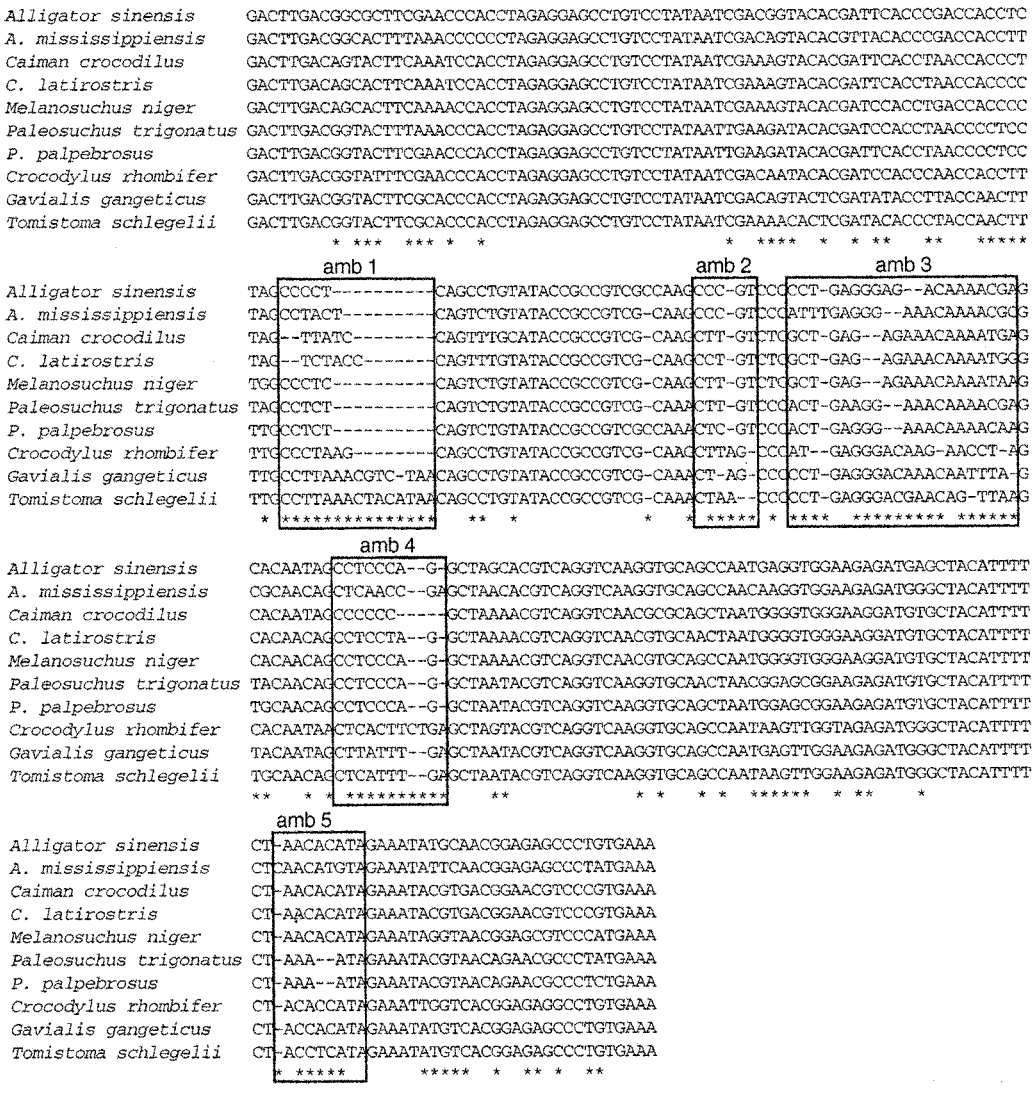


FIGURE 3. Alignment generated in this study for the crocodile 12S mtrDNA data set of Wheeler et al. (1995). Boxes delimit homologous ambiguously aligned sequences. −, gap; ★ variable site. All sites without the ★symbol on the line below the alignment are constant.

within the delimited region, assess the potential effect on phylogenetic relationships for these different gap positions.

5. If these different alignments do not support different phylogenetic relationships, retain this region, which does not need to be excluded from the parsimony analysis.
6. If the alignments do support different phylogenetic hypotheses, replace this region by a coded character.
7. A first approximation of the limits of these regions can be made by using invariant flanking regions as a guide (sites without asterisks in Figs. 2 and 3).

Step 2: Unequivocal Coding of Ambiguously Aligned Regions

Once all ambiguous regions of an alignment have been delimited, each region is treated separately (Fig. 4). Because the ambiguity resides in the presence/absence of gaps and their placement, the first operation is to recover the original sequences for this region. This is done by removing all gaps from that region (Figs. 4a, b). Next, the sequences are inspected for the presence of missing data, uncertain base calling (e.g., presence of IUPAC-IUB codes), and polymorphisms. All sequences with missing data at one or more

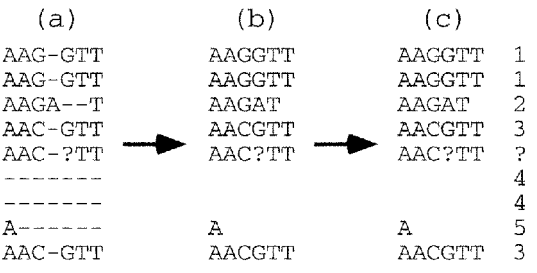


FIGURE 4. Unequivocal coding procedure for one hypothetical delimited ambiguously aligned region. −, gap; ?, missing data. (a) one delimited homologous region containing ambiguously aligned sequences. (b) The first operation is to remove all gaps. (c) Pairwise comparisons of sequences for that region are then implemented. Identical sequences within that homologous region (including the complete absence of nucleotides) are given the same code. Any sequence with at least one missing datum is coded as "?". At the end of the second step of our approach, the ambiguously aligned region is reduced to a single coded character.

sites are coded as missing data (Fig. 4c) because it is impossible to determine whether they are identical to any other sequence in that ambiguous region. Identical sequences (both in length and base composition) are attributed the same character state (Fig. 4c). This principle also applies to OTUs without nucleotides for a given homologous but ambiguously aligned region. For example, in Figure 4c, the two OTUs that do not have nucleotides for that region of the alignment are both coded as character state 4. This last case illustrates why these delimited regions need to be flanked by very conserved regions and, therefore, can be considered homologous. For sequences that include uncertain base calling or polymorphic sites, all possible coded character states can be attributed to these sequences. This multistate coded version of these sequences can then be treated as polymorphic or uncertain when using PAUP*. This practice, however, can lead to the introduction of a large number of character states for a given coded character, which can be problematic (see Discussion). In such instances, coding sequences with uncertain base calling or polymorphisms simply as missing data might be preferable. Finally, although the ambiguous region is excluded from phylogenetic analyses, it is replaced by its respective coded character, which is added to the end of the alignment. This procedure is repeated for each ambiguous region that is delimited for a given alignment and judged worthy of coding.

Two major factors affecting the result of this coding procedure are the width of the ambiguous region and the rates at which these sequences are evolving. For example, identical sequences are less likely to be found for an ambiguous region of 40 sites than for a region of 4 sites; the same is true for a fast-evolving (hypervariable) region versus a slowly evolving one. The implementation of this coding method on a data set with a large number of OTUs could also be problematic because not enough symbols might be available for all the different sequences within a given ambiguous region. The NEXUS format (Maddison et al., 1997) allows the use of all possible ASCII characters for this purpose; however, PAUP*4.0b2 allows a maximum of 32 character states per character. If the number of different sequences for a given ambiguous region exceeds the maximum number of character states per character allowed by a given program, all sequences that are found more than once in that region could be coded with ASCII characters and all unique sequences could be coded as "?". As pointed out earlier, the use of missing data in phylogenetic studies can be misleading; therefore, completely excluding this region from phylogenetic analyses, that is, without replacement by a coded character, might be preferable. This problem could be viewed as a warning that saturation, and consequently the loss of phylogenetic signal in that region, might be an issue, which would give additional support for complete exclusion of the region from the phylogenetic study. This is more likely to be the case when almost all sequences within a delimited region are different. Accordingly, it is preferable to delimit regions as narrowly as possible. However, the risk of delimiting nonhomologous regions increases with the circumscription of narrower regions. Because the validity of this method depends entirely on the delimitation of ambiguous regions, which doubtless contain some homologous sequences, including a few unambiguously aligned sites from the flanking regions to delimit slightly broader regions might be desirable. The latter approach could be preferable even if it led to the complete exclusion of this region from phylogenetic analyses. The use of loop and stem information provided by the secondary structure can be often useful in delimiting narrow ambiguously aligned regions without risking delimiting nonhomologous regions.

If an ambiguous region is at the end of the alignment, such as ambiguous region 5 (amb 5) in Figure 2, there may be no highly conserved region between the end of the alignment and the ambiguous region to be delimited. Such an uncertain delimitation at the end of an alignment could be used to justify the complete exclusion of that ambiguous region from future analyses, because the portion of the sequences included in that delimited region might not be homologous.

Step 3: Optimal Weighting of Ambiguously Aligned Regions

The weighting scheme proposed here reflects the most likely nature of the changes from one sequence to another within an ambiguous region and includes a combination of two factors: similarity (S_{xy}) and number of changes (C_{xy}) for all pairwise comparisons of sequences (x) and all possible alignments for each pair of sequences (y). S_{xy} is equal to the number of sites that are identical (N_i) divided by the total number of sites (N_t):

$$S_{xy} = N_i / N_t$$

C_{xy} is equal to the total number of changes (number of nucleotide susbstitutions + number of separate indels) for a given pair of sequences. A series of consecutive gaps or a single gap is counted as one indel (= one change). The optimal number of steps is obtained by maximizing S_{xy} and minimizing C_{xy}. The highest optimal value (O_{xy}) is used to select the C_{xy} value that will be entered in the symmetric step matrix to represent the optimal number of steps for the transformation of one sequence into the other:

$$O_{xy} = S_{xy}/C_{xy}$$

First, all alignments with the fewest changes and the greatest similarity for a given pair of sequences must be found. In this process, a gap or a series of gaps will be introduced only if O_{xy} stays the same or increases. For example, in Figure 5a the optimal number of steps that would be included in the step matrix for these two sequences would be 2, because the highest optimal value is 28.57. This example, has two equally optimal alignments, O_{11} and O_{12}. Both involve the same number of changes (C_{xy} = 2). In the case where two equally optimal values are obtained but each indicates a different number of fewest changes (i.e., different C_{xy} values), the fewest changes (lowest C_{xy} value) would be integrated in the step matrix. By doing this, we assert a preference for slightly underweighting than overweighting a change from one sequence to the other. This also minimizes the possibility of violating triangle inequality for any particular step matrix (Maddison and Maddison, 1992). Figure 5b shows how sliding the sequence by one site and increasing N_t by one can still increase O_{xy}.

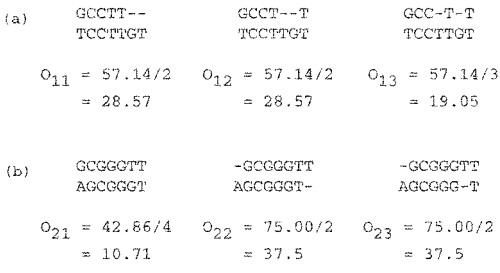


FIGURE 5. Determination of the optimal C_{xy} value, to be included in a step matrix and applied to a specific coded character representing an ambiguously aligned region, by maximizing O_{xy} (see text for further explanation). (a) In this case the optimal number of changes (C_{xy} = 2) would be chosen as the optimal number of steps and included in the step matrix to account for the transformation between the two characters states representing these two sequences. (b) Example showing a case where increasing N_t can increase O_{xy}.

The output of this method consists of a series of coded characters, one for each designated ambiguous region of the alignment, each with its own step matrix (Fig. 6). The coded characters can be added at the end of the data matrix. The number of characters specified in the NEXUS file needs to be increased accordingly, and the ambiguous regions need to be excluded because they are now replaced by their coded counterparts. Finally, the step matrices need to be added in the assumptions block of the NEXUS file. The user needs to ensure that each step matrix is applied to its specific coded character only. PAUP* tests each step matrix for triangle inequality automatically and makes any necessary modifications.

AN EXAMPLE USING THE INSECT 16S AND CROCODILE 12S MTRDNA DATA SETS

Insect 16S mtrDNA

Wheeler et al. (1995) used 10 different ratios of gap-to-substitution costs to align 9 sequences from the mitochondrial genome of

(a) INSECT DATA SET

amb 1

		b	d	h	k	m	n	r	v
TTTTTATCG	b		b	d	h	k	m	n	r	v
TTTTTTATCG	d	[b]	.	1	1	3	3	3	5	4
TTTTTGTCG	h	[d]	1	.	2	3	3	3	5	4
TTCGTTATAG	k	[h]	1	2	.	4	2	2	4	3
TGCTTTGTCG	m	[k]	3	3	4	.	4	5	4	5
CTTTTTGTCA	n	[m]	3	3	2	4	.	3	4	5
CTTCTTGTAA	r	[n]	3	3	2	5	3	.	2	3
CTTCTTGTAA	r	[r]	5	5	4	4	4	2	.	3
ATCTTTTACT	v	[v]	4	4	3	5	5	3	3	.

amb 2

		b	d	h	k	m	n	r	v
TTATTAAT	b		b	d	h	k	m	n	r	v
ATAATATT	d	[b]	.	3	4	3	3	2	2	3
ATAAAATT	h	[d]	3	.	1	2	3	2	1	4
AAAATATTT	k	[h]	4	1	.	2	3	3	2	3
TAAATTAT	m	[k]	3	2	2	.	3	3	3	3
CTAATAAT	n	[m]	3	3	3	3	.	3	3	2
ATAATAAT	r	[n]	2	2	3	3	3	.	1	3
ATAATAAT	r	[r]	2	1	2	3	3	1	.	4
CAATTTAT	v	[v]	3	4	3	3	2	3	4	.

amb 3

		b	d	h	k	m	n	r	v
GAATTATT	b		b	d	h	k	m	n	r	v
CATAAACTCATA	d	[b]	.	5	3	3	3	4	3	3
CAAATATA	h	[d]	5	.	3	3	4	2	3	5
CAAAAAATT	k	[h]	3	3	.	3	3	1	2	2
CAAAACAACAAA	m	[k]	3	3	3	.	2	3	3	2
CAAAACTATA	n	[m]	3	4	3	2	.	3	3	4
CCAATATT	r	[n]	4	2	1	3	3	.	3	3
CCAATATT	r	[r]	3	3	2	3	3	3	.	3
CAAAAATATCTTTATAT	v	[v]	3	5	2	2	4	3	3	.

amb 4

		b		d	h	k	m	n	r	v	w
GTTTTTTAAAATTAAAAGTTTTTAAATTTTAATATC	b		b	d	h	k	m	n	r	v	w
AAGTAAAGAGAATTAAGTTAATCTCTATATT	d	[b]	.	7	6	7	10	8	11	10	7
GTAAACAAATAATAAAAGTTAATAAATTTTAATACC	h	[d]	7	.	10	6	11	9	6	6	7
GGTAACAAAAAATTAAGTTTTTTTTTTTACAATC	k	[h]	6	10	.	7	12	10	8	7	8
GATATAATAATGAAGAGTTTATTTATTCTTCATGTC	m	[k]	7	6	7	.	12	9	9	8	8
TATTAATTAAAAAGAGTTTTATTTATCTCATTATC	n	[m]	10	11	12	12	.	9	10	10	8
GGTTATTAAGTTATAAAAGTTAGTTATATATTTTTTGTC	r	[n]	8	9	10	9	9	.	9	9	9
GGTTATTAAGTTATAAAAGCTTAAGTTATATATTTTTTGTC	v	[r]	11	6	8	9	10	9	.	2	10
CTTAAATAAAGTTTATTAAATTTACCAATC	w	[v]	10	6	7	8	10	9	2	.	10
		[w]	7	7	8	8	8	9	10	10	.

(b) CROCODILE DATA SET

amb 1

			1	2	0	K	5	6	7	8	9
CCCCT	1		1	2	0	K	5	6	7	8	9
CCTACT	2	[1]	.	2	2	3	2	1	2	3	3
TTATC	0	[2]	2	.	4	2	2	1	2	3	2
TCTACC	K	[0]	2	4	.	2	3	3	4	3	4
CCCTC	5	[K]	3	2	2	.	3	2	4	4	4
CCTCT	6	[5]	2	2	3	3	.	2	2	3	3
CCTCT	6	[6]	1	1	3	2	2	.	3	2	2
CCCTAAG	7	[7]	2	2	4	4	2	3	.	2	3
CCTTAAACGTCTAA	8	[8]	3	3	3	4	3	2	2	.	2
CCTTAAACTACATAA	9	[9]	3	2	4	4	3	2	3	2	.

amb 2

			1	2	0	K	5	6	7
CCCGT	1		1	2	0	K	5	6	7
CCCGT	1	[1]	.	2	1	1	3	3	2
CTTGT	2	[2]	2	.	1	1	2	2	2
CCTGT	0	[0]	1	1	.	2	3	3	2
CTTGT	2	[K]	1	1	2	.	2	2	2
CTTGT	2	[5]	3	2	3	2	.	1	2
CTCGT	K	[6]	3	2	3	2	1	.	1
CTTAG	5	[7]	2	2	2	2	2	1	.
CTAG	6								
CTAA	7								

amb 3

			1	2	3	K	5	6	7	8	9	0
CCTGAGGGAGACAAAACGA	1		1	2	3	K	5	6	7	8	9	0
ATTTGAGGGAAACAAAACGC	2	[1]	.	4	4	5	5	3	3	6	3	4
GCTGAGAGAAACAAAATGA	3	[2]	4	.	5	5	5	5	4	5	5	5
GCTGAGAGAAACAAAATGC	K	[3]	4	5	.	1	1	4	4	6	6	5
GCTGAGAGAAACAAAATAA	5	[K]	5	5	1	.	2	5	4	7	5	5
ACTGAAGGAAACAAAACGA	6	[5]	5	5	1	2	.	5	3	7	5	5
ACTGAGGGAAACAAAACAA	7	[6]	3	5	4	5	5	.	2	6	5	5
ATGAGGGACAAGAACCTA	8	[7]	3	4	4	4	3	2	.	5	4	4
CCTGAGGGACAAACAATTTA	9	[8]	6	5	6	7	7	6	5	.	5	5
CCTGAGGGACGAACAGTTAA	0	[9]	3	5	6	5	5	5	4	5	.	3
		[0]	4	5	5	5	5	5	4	5	3	.

amb 4

			1	2	0	K	5	6	7
CCTCCCAG	1		1	2	0	K	5	6	7
CTCAACCGA	2	[1]	.	4	2	1	5	4	4
CCCCCC	0	[2]	4	.	4	4	3	3	2
CCTCCTAG	K	[0]	2	4	.	3	4	2	3
CCTCCCAG	1	[K]	1	4	3	.	4	4	4
CCTCCCAG	1	[5]	5	3	4	4	.	3	2
CCTCCCAG	1	[6]	4	3	2	4	3	.	1
CTCACTTCTGA	5	[7]	4	2	3	4	2	1	.
CTTATTTGA	6								
CTCATTTGA	7								

amb 5

			1	2	0	K	5	6
AACACATA	1		1	2	0	K	5	6
CAACATGTA	2	[1]	.	3	2	2	1	2
AACACATA	1	[2]	3	.	3	3	3	4
AACACATA	1	[0]	2	3	.	2	2	2
AACACATA	1	[K]	2	3	2	.	2	2
AAAATA	0	[5]	1	3	2	2	.	1
AAAATA	0	[6]	2	4	2	2	1	.
ACACCATA	K							
ACCACATA	5							
ACCTCATA	6							

FIGURE 6. Unequivocal coding and optimal character-state weighting (step matrices), according to the method proposed here, for each homologous ambiguously aligned region. (a) Insect 16S mtrDNA. (b) Crocodile 12S mtrDNA.

insects, which yielded 12 alignments. When they analyzed these alignments individually, 11 unique topologies were generated. When they combined these 12 alignments into one concatenated alignment and removed (culled) all sites that differed among these 12 alignments, their parsimony analysis generated 2 equally most-parsimonious trees. They reported that the strict consensus of these two trees was entirely unresolved except for a single questionable *Apis–Cerastipsocus* clade (Fig. 7a). When they downweighted, rather then culled, the variable sites among these 12 alignments proportionally to the amount of interalignment variability, that is, after their elision method,

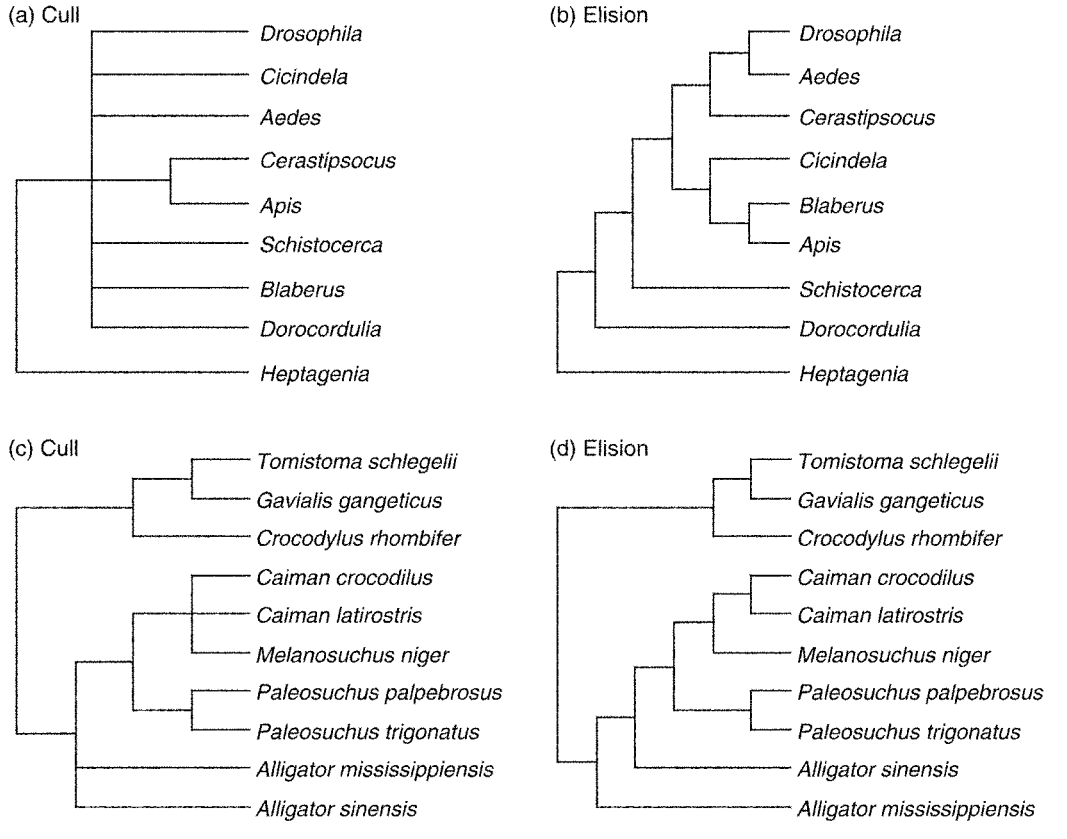


FIGURE 7. Results obtained by Wheeler et al. (1995). (a, b) Phylogenetic trees resulting from the cull (a) and elision (b) procedures applied to the insect 16S mtrDNA data set. (c, d) Phylogenetic trees resulting from the cull (c) and elision (d) procedures applied to the crocodile 12S mtrDNA data set.

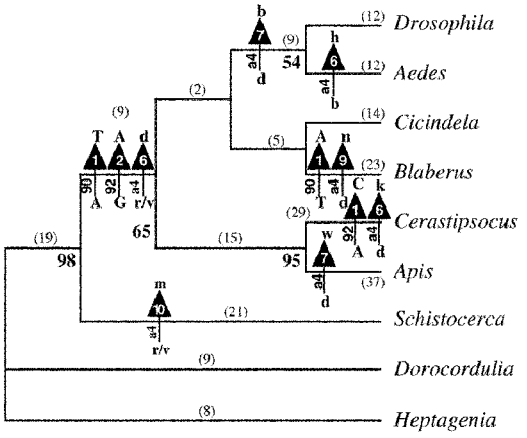
they obtained a single most-parsimonious tree (Fig. 7b).

We found five regions in this data set that could not be aligned unequivocally (amb 1 to amb 5; Fig. 2). Amb 1 to amb 4 were easily delimited. They are flanked by highly conserved (often invariable) sites and shifting the gaps into these flanking regions could not be justified under any optimization criterion to improve the alignment. Amb 5, however, lacks a conserved region at its 3′ end, which makes it impossible to determine if the sequences in amb 5 are homologous, as shown by the "?" inserted at the end of the alignment (Fig. 2). With this fundamental assumption in doubt, we did not code amb 5; accordingly, we completely removed it from subsequent phylogenetic analyses.

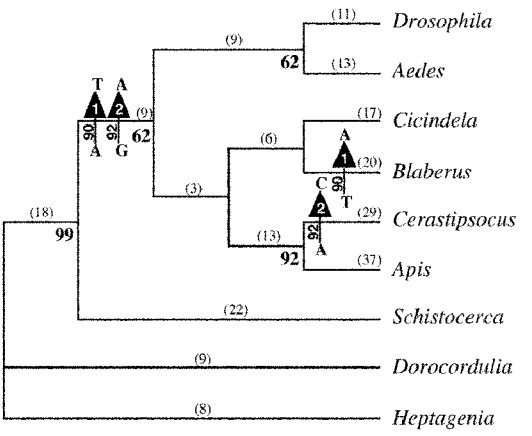
The unambiguously aligned portions of this data set, as established with our method, provided 21 parsimony-informative characters. Of the four ambiguously aligned regions coded with our method, none was parsimony-informative (Fig. 6a) if implemented under a model in which all changes among coded character-states had equal costs (i.e., in a step matrix in which all changes = 1 step). However, under the model of unequal weighting of changes among character-states implemented with our method, phylogenetic signal was recovered from these ambiguous regions (Table 1).

When the phylogenetic search was restricted to unambiguous sites, analogous to the "cull" analysis of Wheeler et al. (1995), three equally most-parsimonious trees were recovered (Figs. 8a, b, e; Table 1). Wheeler et al. (1995) reported that their cull procedure yielded two topologies with an entirely unresolved strict consensus except for a single questionable *Apis–Cerastipsocus* clade (Fig. 7a). The strict consensus of the three trees recovered with our approach was much more resolved and is identical to the topology

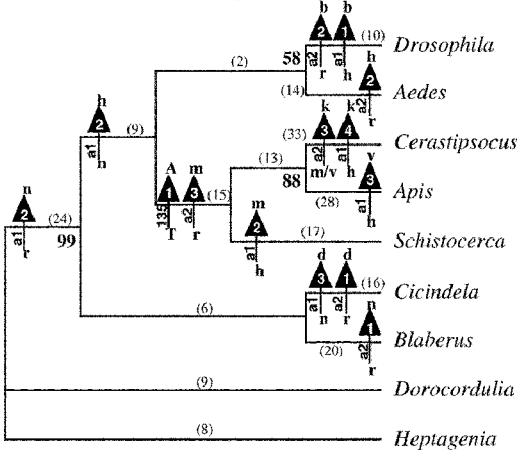
(a) Tree length = 123 steps



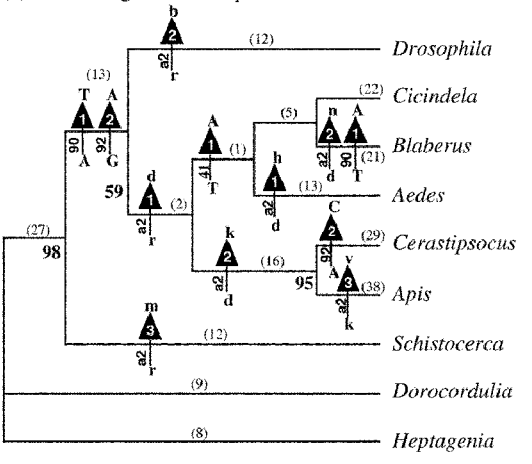
(b) Tree length = 123 steps



(c) Tree length = 125 steps



(d) Tree length = 124 steps



(e) Tree length = 123 steps

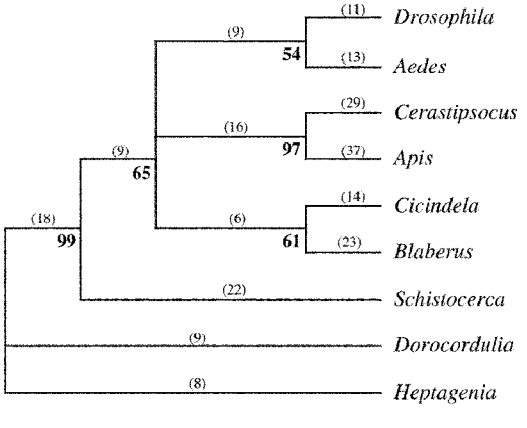


TABLE 1. Summary of phylogenetic analyses for the insect 16S mtrDNA data set. Coded amb 1–4 correspond to ambiguous regions shown in Figure 2.

Characters included	Tree length	No. most-parsimonious trees	Topology[a]	Consistency index	Rescaled consistency index
Unambiguous sites only	123	3	a, b, e	0.846	0.542
Unamb. + coded amb 1	141	3	a, b, e	0.837	0.521
Unamb. + coded amb 2	138	4	a, b, d, e	0.841	0.516
Unamb. + coded amb 3	138	3	a, b, e	0.855	0.560
Unamb. + coded amb 4	176	3	a, b, e	0.875	0.579
Unamb. + coded amb 1–4	224	4	a, b, c, e	0.866	0.549

[a]See Figure 8.

shown in Figure 8e. Only one internode is missing for this topology to be completely resolved.

When added individually, the coded versions of amb 1, amb 3, and amb 4 (Figs. 2, 6) generated the same three topologies as when the phylogenetic search was restricted to the unambiguous portions of the alignment (Figs. 8a, b, e; Table 1). The addition of amb 2 alone (Figs. 2, 6) resulted in one additional equally most-parsimonious tree (Fig. 8d; Table 1). This additional tree corresponds more or less to the collapse of the internode supporting the *Drosophila–Aedes* clade, which was associated with the lowest bootstrap values (Fig. 8). The new relationships revealed for these two taxa by the addition of amb 2 are all very weakly supported (Fig. 8d).

When all four coded ambiguous regions were included simultaneously in the phylogenetic search with the unambiguously aligned regions, four equally most-parsimonious trees of 224 steps were revealed (Table 1). These were the same three topologies derived from the analysis restricted to the unambiguously aligned portions of the data set (Figs. 8a, b, e) plus one additional tree (Fig. 8c). The latter topology was 2 steps longer (tree length = 125 steps) when only the unambiguously aligned sites were used to calculate the tree length (Fig. 8). The additional tree (Fig. 8d) found when adding amb 2 alone was only one step longer (tree length = 124) than the three most-parsimonious trees (Figs. 8a, b, e). These longer trees were not statistically significantly worse than the most-parsimonious trees when subjected to the Templeton test. Because the four coded ambiguous regions are not parsimony-informative under an equally weighted (step matrix) scheme, and because we believe that the inclusion of such characters with our method can cause phylogenetic artifacts (see Discussion), we concluded that restricting the analysis to the unambiguously aligned sites was the only justifiable solution.

The elision procedure of Wheeler et al. (1995) generated one most-parsimonious tree (Fig. 7b). The internode supporting the *Drosophila–Aedes* clade, the one separating *Dorocordulia* and *Heptagenia* from the rest of the taxa, and the bipartition separating

FIGURE 8. The five most-parsimonious trees generated from analysis of the insect 16S mtrDNA data set, with and without the inclusion of ambiguously aligned regions, as obtained by the new method proposed here (see also Table 1). Only characters that provided unequivocal support, in terms of gains and reversals, to different phylogenetic relationships revealed by Wheeler et al. (1995) or in this study, were mapped onto the topologies. Arrows describe the nature of the changes within these characters by showing the character states involved in each change. The symbol on the left of the tail of each arrow refers to a specific character. When this symbol is a number, it refers to a specific site in the alignment (Fig. 2). If the number is preceded by "a", it refers to a coded ambiguously aligned region. For example, "a4" corresponds to ambiguously aligned region 4 (Figs. 2, 6a). Numbers in each arrowhead indicate the number of steps (from the step matrix) associated with each specific change. Bold numbers below internodes are bootstrap percentages, and numbers above internodes in parentheses are branch lengths (unequivocal changes). Tree lengths shown on this figure were calculated when only unambiguously aligned sites were analyzed. (a, b, e) Three equally most-parsimonious trees consistently generated, whether ambiguous regions were included or not in the six phylogenetic analyses listed in Table 1. (c) Single most-parsimonious tree generated only when all four coded ambiguous regions were included with unambiguously aligned sites (Table 1). (d) Single most-parsimonious tree obtained when only the coded amb 2 region was included with the unambiguously aligned sites (see Table 1, Fig. 2).

Dorocordulia, *Heptagenia*, and *Schistocerca* from the rest are congruent with the strict consensus of our three most-parsimonious trees. Both methods revealed a close relationship between *Cicindela* and *Blaberus*. There was one marked difference between the elision result and ours: The latter consistently showed *Apis* as sister to *Cerastipsocus*. This odd relationship (according to Wheeler et al., 1995) could well be an artifact resulting from long branch attraction. These two taxa have the two longest branches (Fig. 8) nested within a portion of the tree with short internodes.

Two other factors could have contributed to the incongruent results between the elision approach and our method. First, Wheeler et al. (1995) assumed that all types of nucleotide substitutions had the same cost. Our model took into consideration that some nucleotide substitutions were more frequent than others and, therefore, were attributed different costs, which we implemented through a step matrix on the unambiguous portion of the alignment. Second, our method did not allow us to include region amb 5, which is by far the largest ambiguously aligned region, covering almost one third of the alignment (Fig. 2). This region is highly variable and would provide a very high fraction (probably close to 50%) of all the parsimony-informative sites from the complete data matrix. Even if most sites in this region are downweighted by the elision method, the signal it provides will probably have a considerable impact on the resulting topology. However, this region is extremely AT-rich (A = 52% and T = 37%) and requires the inclusion of many gaps. Therefore, we feel justified to at least question the phylogenetic quality of the signal it provides. If this region were saturated by changes that led to an accumulation of As and Ts followed by multiple changes between As and Ts, relatively little (if any) phylogenetic signal would actually still be present in this region. This raises a fundamental question. Should such a region be included in a phylogenetic analysis even if we were able to include it without violating positional homology?

Crocodile 12S mtrDNA

The 10 different gap-to-substitution cost ratios used by Wheeler et al. (1995) generated 23 different alignments. When analyzed separately, each of their 23 phylogenetic analyses yielded a single most-parsimonious tree. When compared, however, many of these trees were found to be identical, yielding a total of four different topologies. By joining these 23 alignments end to end to form one concatenated alignment and then removing (culling) all interalignment variable sites, the phylogenetic analysis of this concatenated alignment generated four equally most–parsimonious trees. The strict consensus of these four topologies (Fig. 7c) shows a lack of resolution within the *Caiman–Melanosuchus* clade and also for the relationship between the two *Alligator* species (i.e., paraphyly vs. monophyly). Maximum parsimony analysis by Wheeler et al. (1995) of the concatenated alignment resulting from the elision of all 23 alignments (i.e., by downweighting the interalignment variable sites rather than excluding them from the analysis) revealed a single most-parsimonious tree with *Melanosuchus* sister to the *Caiman* clade and *Alligator* as paraphyletic (Fig. 7d).

Our alignment of the same data resulted in the recognition of five ambiguously aligned regions, all of which were clearly flanked by highly conserved nucleotide sequences (Fig. 3). All five ambiguous regions were subjected to our method, that is, were coded and assigned step matrices (Fig. 6b). Two of the five coded ambiguous regions (amb 2 and amb 5) are potentially parsimony-informative if implemented under a model in which every type of change has an equal cost. Under a model of unequal weighting among coded character-states, however, only amb 2 and amb 4 affected the phylogenetic searches, compared with a search restricted to unambiguous sites only (Table 2).

When our maximum parsimony analysis was restricted to the unambiguously aligned portions of the data, we obtained 52 parsimony-informative characters that resulted in one most-parsimonious tree (Fig. 9b). As with the insect 16S data set, our cull analysis (Fig. 9b) provided much more resolution than did the cull analysis (Fig. 7c) implemented by Wheeler et al. (1995). This topology is also identical to the most-parsimonious tree derived from the alignment obtained by using a gap-to-substitution cost ratio of 1:2 (Wheeler et al. 1995).

When only coded region amb 2 was added to our phylogenetic analysis of unambiguous sites, two equally most-parsimonious trees were recovered (Table 2). One of these

TABLE 2. Summary of phylogenetic analyses for the crocodile 12S mtrDNA data set. Coded amb 1–5 correspond to ambiguous regions shown in Figure 3.

Characters included	Tree length	No. most-parsimonious trees	Topology[a]	Consistency index	Rescaled consistency index
Unambiguous sites only	145	1	b	0.697	0.496
Unamb. + coded amb 1	159	1	b	0.723	0.518
Unamb. + coded amb 2	154	2	a, b	0.701	0.497
Unamb. + coded amb 3	170	1	b	0.741	0.539
Unamb. + coded amb 4	158	1	c	0.702	0.500
Unamb. + coded amb 5	154	1	b	0.714	0.513
Unamb. + coded amb 1–5	215	1	c	0.772	0.563
Unamb. + coded amb 2, 5	163	2	a, b	0.718	0.513

[a] See Figure 9.

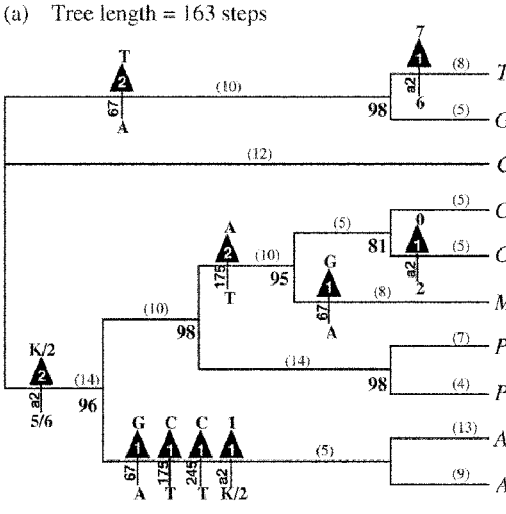
topologies is identical to the one derived from the analysis of the unambiguous sites only (Fig. 9b). The second topology shows the two *Alligator* species as sister species (Fig. 9a). The strict consensus of these two topologies would show a completely resolved tree except for the unresolved relationship between the two *Alligator* species. This unresolved portion of the tree certainly represents the main weakness of the data set to resolve relationships between this pair of species (see differences among trees in terms of topologies and bootstrap values in Fig. 9).

When only coded region amb 4 was added to our phylogenetic analysis of unambiguous sites, a single most-parsimonious tree was obtained (Fig. 9c). This topology is identical to the single most-parsimonious tree that resulted from our all-inclusive analysis of the unambiguous sites plus all five coded ambiguous regions (Table 2). This is also the same tree recovered by Wheeler et al. (1995) when they analyzed individually each alignment derived from gap-to-substitution cost ratios of 1:1, 2:1, 4:1, 8:1, and 16:1. They also recovered this topology with a cull analysis on alignments derived from gap-to-substitution cost ratios of 1:1, 8:1, and 16:1, and when their elision analysis was restricted to alignments derived from gap-to-substitution cost ratios of 2:1, 8:1, and 16:1. Finally, applying their elision analysis to all alignments combined also generated this topology (Figs. 7d, 9c). For the crocodile 12S mtrDNA data set, the resulting trees from the elision method (Fig. 7d) and from our new approach if applied to all ambiguous regions simultaneously (Fig. 9c) were identical. One advantage to our approach is that it also revealed that the phylogenetic signal contained in ambiguous region 4 (amb 4) is what led to this topology (Table 2). Sequences for *Alligator sinensis, Melanosuchus niger, Paleosuchus trigonatus*, and *P. palpebrosus* are identical in this region (Figs. 3, 9c). The differences between these sequences and the ones for the two *Caiman* species involve only one and two steps, whereas for this same region (amb 4), on the internode between the two *Alligator* species, four steps are involved (Fig. 9c). We emphasize that the coded version of amb 4 is parsimony-uninformative when subjected to an equally weighted step matrix.

This result raises an important question. Should a coded, ambiguously aligned region (such as amb 4) that is parsimony-uninformative when changes are equally weighted, but parsimony-informative when subjected to an unequally weighted step matrix, be included in phylogenetic analyses? Based on empirical evidence (results not shown), the inclusion of such characters favors an asymmetric topology (paraphyletic relationships). To avoid this problem, we would be justified to exclude all coded regions that are parsimony-uniformative when subjected to an equally weighted step matrix. In the case of the 12S data set, the use of this criterion would justify the exclusion of coded regions amb 1, amb 3, and amb 4 from phylogenetic analyses. When the parsimony analysis was implemented by using the unambiguously aligned portions plus amb 2 and amb 5 only, two equally most-parsimonious trees resulted (as shown in Figs. 9a, b), one of which shows the two alligators as sister species.

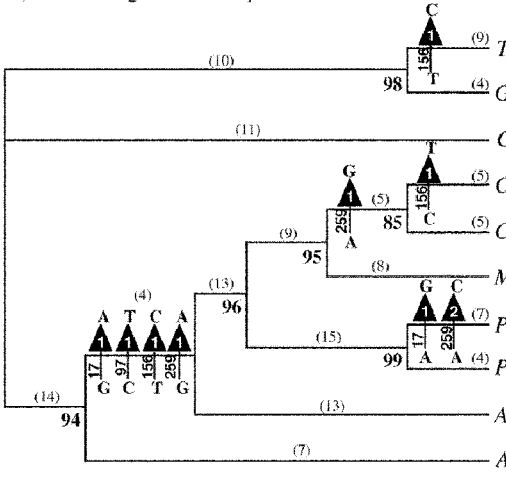
Phylogenetic analyses by Brochu (1997) of morphological as well as combined morphological and molecular data sets strongly support the monophyly of *Alligator sinensis* and *A. mississippiensis*. The suboptimal topology shown in Figure 9c, identical to the tree recovered by Wheeler et al. (1995), was only one step longer and was

(a) Tree length = 163 steps



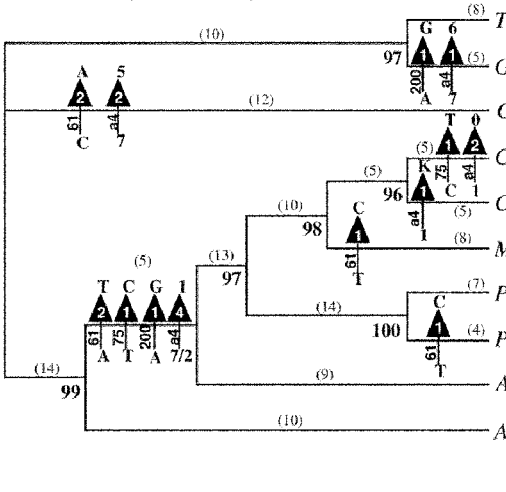
	amb 2 = a2 67	175	245	(coded char.)
Tomistoma schlegelii	T	T	T	CTAA (7)
Gavialis gangeticus	T	T	T	CTAG (6)
Crocodylus rhombifer	A	T	T	CTTAG (5)
Caiman crocodilus	A	A	T	CTTGT (2)
Caiman latirostris	A	A	T	CCTGT (0)
Melanosuchus niger	G	A	T	CTTGT (2)
Paleosuchus palpebrosus	A	T	T	CTCGT (K)
Paleosuchus trigonatus	A	T	T	CTTGT (2)
Alligator mississippiensis	G	C	C	CCCGT (1)
Alligator sinensis	G	C	C	CCCGT (1)

(b) Tree length = 163 steps



	17	97	156	259
Tomistoma schlegelii	G	C	C	G
Gavialis gangeticus	G	C	T	G
Crocodylus rhombifer	G	C	T	G
Caiman crocodilus	A	T	T	G
Caiman latirostris	A	T	C	G
Melanosuchus niger	A	T	C	A
Paleosuchus palpebrosus	G	T	C	C
Paleosuchus trigonatus	A	T	C	A
Alligator mississippiensis	A	T	C	A
Alligator sinensis	G	C	T	G

(c) Tree length = 164 steps



	61	75	200	amb 4 = a4 (coded char.)
Tomistoma schlegelii	A	T	A	CTCATTTGA (7)
Gavialis gangeticus	A	T	G	CTTATTTGA (6)
Crocodylus rhombifer	C	T	A	CTCACTTCTGA (5)
Caiman crocodilus	T	T	G	CCCCCC (0)
Caiman latirostris	T	C	G	CCTCCTAG (K)
Melanosuchus niger	C	C	G	CCTCCCAG (1)
Paleosuchus palpebrosus	T	C	G	CCTCCCAG (1)
Paleosuchus trigonatus	C	C	G	CCTCCCAG (1)
Alligator sinensis	T	C	G	CCTCCCAG (1)
Alligator mississippiensis	A	T	A	CTCAACCGA (2)

not statistically significantly worse than the most-parsimonious trees (Figs. 9a, b) according to the Templeton test. Clearly, this 12S data set (Gatesy et al., 1993; Wheeler et al., 1995) alone cannot resolve with any certainty the relationships between these two *Alligator* species.

DISCUSSION

Criteria for Delimiting Homologous Ambiguously Aligned Sequences

The suggestion by Wheeler et al. (1995) to use a broad range of gap-to-nucleotide substitution cost ratios to detect ambiguously aligned sites, with ratios as low as 1:2 and as high as 256:1, is more likely to identify sites as being ambiguously aligned when they are not. At the low end of this spectrum of cost ratios (e.g., 1:2, 1:1, 2:1), many gaps are included among the aligned sequences resulting in a high level of similarity. At the other end of this spectrum (e.g., 256:1, 300:1), resulting alignments will have fewer gaps but the average similarity among aligned sequences will be much lower. The alignments resulting from the latter cost ratios are extremely poor. For the insect 16S data set, if we exclude the five regions that we identified as being ambiguously aligned (amb 1–5), only 9% of the remaining alignment resulting from a 256:1 cost ratio, as published by Wheeler et al. (1995), would be correctly aligned according to our alignment. This means that only the first 10 sites of this data set (Fig. 10) would not differ among all alignments and, therefore, would not be downweighted or culled using the elision approach. Our approach suggests that 100 more sites of this alignment are unambiguously aligned. On this basis, we predict that more resolution and higher support values will result when phylogenetic searches are restricted to unambiguously aligned regions delimited by using our approach. The exclusion of extreme cost ratios in the elision process would improve the situation. However, this would require a clear definition of the threshold beyond which a cost ratio would be considered extreme. For example, the alignment of Wheeler et al. (1995) using an 8:1 cost ratio for the same insect data set also included several obvious misalignments. The selection of the range of gap-to-nucleotide substitution costs is very subjective, which is what Wheeler et al. (1995) were trying to avoid with the elision method. In an attempt to solve this problem, Giribet and Wheeler (1999) proposed using character congruence among data sets, obtained with ILD metrics (Mickevich and Farris, 1981), as the criterion to evaluate gap costs in sequence alignment and phylogenetic analyses.

When maximum parsimony analysis of the insect 16S mtrDNA data set was restricted to unambiguously aligned sites identified with our method versus those identified using the elision approach, our strict consensus tree was much more resolved (Fig. 8e vs. Fig. 7a). Our approach also obtained higher resolution for the crocodile 12S mtrDNA data set under the same circumstances. Four equally most-parsimonious trees resulted from the analysis after culling ambiguously aligned sites identified by the elision method (Fig. 7c), whereas a single most-parsimonious solution (Fig. 9b)

FIGURE 9. The three most-parsimonious trees generated from the crocodile 12S mtrDNA data set, with and without the inclusion of ambiguously aligned regions, as obtained by the new method proposed here (see also Table 2). Only characters that provided unequivocal support for relationships between the two *Alligator* species (mono- vs. paraphyletic) were mapped onto the topologies. Arrows describe the nature of the changes within these characters by showing the character states involved in each change. The symbol on the left of the tail of each arrow refers to a specific character. When this symbol is a number, it refers to a specific site of the alignment (Fig. 3). If the number is preceded by "a", it refers to a coded ambiguously aligned region. For example, "a4" corresponds to ambiguously aligned region 4 (Figs. 3, 9c). Numbers in each arrowhead indicate the number of steps (from the step matrix) associated with each specific change. For example, on topology (c) the change in amb 4 from character states 7 or 2 to 1 involved four steps. Bold numbers below internodes are bootstrap percentages and numbers above internodes in parentheses are branch lengths (unequivocal changes). Tree lengths shown on this figure were calculated when unambiguous sites and coded ambiguous regions amb 2 and amb 5 were analyzed simultaneously. (a) Tree obtained when coded amb 2, with or without amb 5, was included with the unambiguous sites (Table 2, Fig. 3). (b) Tree obtained when only unambiguous sites were included, and when coded amb 1, amb 2, amb 3, amb 5, or amb 2 and amb 5, were added to unambiguous sites. (c) Single most-parsimonious tree generated when the entire data set was included, that is, when all ambiguously aligned regions (amb 1–5) were added to unambiguous sites. The same topology was also generated when only amb 4 was added to the unambiguous sites (Table 2; amb 4, Fig. 3).

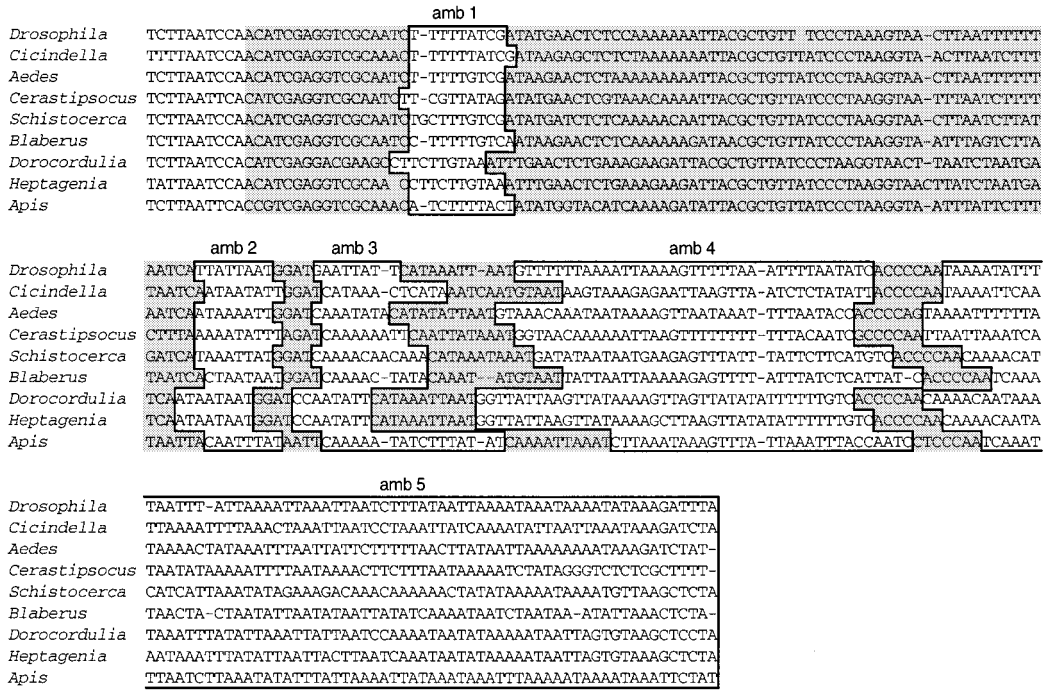


FIGURE 10. Alignment of the insect 16S mtrDNA data set presented by Wheeler et al. (1995), using a gap-to-nucleotide substitution cost ratio of 256:1. Ambiguously aligned regions (amb 1–5) identified with our method (see Fig. 2) are delimited by clear boxes. Shaded areas delimit regions that we identified as unambiguously aligned but that are aligned differently because the 256:1 cost ratio is used (compare with our alignment shown in Figure 2). These shaded regions would be detected as ambiguously aligned and, therefore, would be downweighted with the elision approach. Adapted from the appendix published by Wheeler et al. (1995).

resulted when the analysis was restricted to unambiguously aligned sites selected by using our approach.

Assuming that conservation of secondary structure exceeds that of its nucleotides, Kjer (1995) demonstrated the high potential of the secondary structure as a guide for assigning homologous positions and for improving alignments and phylogenetic accuracy. Our experience using secondary structure to enhance alignments of nrDNA sequences confirms Kjer's conclusions. However, that practice does not solve all alignment problems. Ambiguously aligned sites are highly concentrated in loops, where the secondary structure is of no use in refining the alignment; or the sites can be located where the secondary structure is itself unresolved. Despite this limitation, secondary structure can be often useful in delimiting narrower ambiguously aligned regions without violating homology. For this reason, the use of secondary structure information (when available) is an integral part of the first step of our procedure. Unfortunately, some degree of subjectivity is still involved in the delimitation of ambiguous regions regardless of the method used. This specific aspect needs further investigation.

In this article we did not use secondary structure to help delimit the ambiguously aligned regions because we wanted to demonstrate that the method can be implemented even if the secondary structure is unknown. Moreover, the 16S sequences published by Wheeler et al. (1995), which differ from those in GenBank, favor stem disruptions that would lead to an incorrect secondary structure for some taxa. For a fair comparison of elision with our method, however, we used the data set as it was published by Wheeler et al. in 1995.

Contribution of Ambiguously Aligned Sequences to Phylogenetic Resolution

Ambiguously aligned regions generally include the fastest evolving sites of a given

molecule. Therefore, these sites are most likely to provide phylogenetic signal that can resolve rapid radiations, characterized by consecutive short internodes that are often unresolved in phylogenetic studies and are almost always associated with very low bootstrap values. The integration of these rapidly evolving, ambiguously aligned regions by the method proposed here is most likely to provide synapomorphies along these problematic short internodes. We therefore expect that greater resolution will be gained when implementing our method, especially in comparison with the results obtained from the common and conservative practice of completely excluding these regions. In a study by Fernandez et al. (1999), based on a data set of 898 sites from the 5′ end of the nrRNA large subunit for 22 fungal taxa, 16 ambiguously aligned regions were delimited. These 16 regions included 157 of the 898 sites. The integration of these regions into the phylogenetic analysis, according to our method, resulted in a single most-parsimonious tree, compared with six equally most-parsimonious trees when these regions were completely excluded from the analysis (see Fernandez et al., 1999).

The phylogenetic analysis of the crocodile 12S mtrDNA provides an example of another possible outcome resulting from the integration of ambiguously aligned regions as proposed here. The exclusion of the five ambiguously aligned regions resulted in a single most-parsimonious tree, whereas the inclusion of amb 2 and 5 resulted in one additional tree, which showed the two alligators as sister species. In this case the inclusion of ambiguous regions by our method resulted in the loss of resolution, but this larger pool of equally most-parsimonious trees includes what is thought to be the correct tree, according to a combined morphological and molecular study (Brochu, 1997). If the phylogenetic signal contained in ambiguously aligned regions is essential for phylogenetic searches to converge on the correct tree, and when maximum parsimony is consistent (sensu Felsenstein, 1978), our method should increase phylogenetic accuracy, because of the additional characters it provides, that would otherwise be excluded or downweighted.

Should All Ambiguously Aligned Regions Be Included in Phylogenetic Analyses?

Despite the high potential for ambiguously aligned regions to contribute greatly to phylogenetic accuracy, in some cases these regions will provide no gain or will even be detrimental to phylogenetic accuracy. For example, if these regions are evolving so fast that all sequences within a region are different, the coded version of these regions will not be parsimony-informative. Yet, when subjected to an unequally weighted step matrix, these coded regions can become informative and provide a large portion of the signal (e.g., the contribution of amb 4 in Table 2 and Fig. 9c). However, the phylogenetic signal from these hypervariable regions might have been totally lost in the multitude of genetic changes over time. Moreover, preliminary empirical work we have done (results not shown) indicates that such characters are biased toward recovering asymmetric topologies. The uninformativeness of coded ambiguously aligned regions, and perhaps the saturation of genetic changes, seem to best describe what is taking place with the insect 16S mtrDNA (Figs. 6a, 8; Table 1) and ambiguously aligned regions 1, 3, and 4 of the crocodile 12S mtrDNA in this study (Figs. 6b, 9; Table 2). For these reasons we strongly recommend that users of the method proposed here exclude from phylogenetic analyses those coded ambiguously aligned regions that are uninformative when subjected to an equally weighted step matrix. If the homology of a given delimited ambiguously aligned region is doubtful, this region cannot be dealt with in our approach and therefore should be excluded from all analyses (e.g., amb 5 in Fig. 2).

Another case for which users of the method proposed here may consider excluding a coded ambiguously aligned region from phylogenetic analyses is when a prohibitively high number of character states are required to code a given region. PAUP* 4.0b2a allows a maximum of 32 character states per character. If the coding of an ambiguous region would require >32 character states, that region should be excluded from phylogenetic analyses; however, this is an arbitrary cutoff point.

This leads us to the question: Should a coded ambiguous region with a high number

of character states be included in phylogenetic analyses, even if it can be accommodated? Our experience is that it becomes very difficult to handle ambiguously aligned regions that require >15 character states. Such situations are usually associated with large (many sites) ambiguous regions, and large regions are more likely to include sequences with drastic differences in length, which can cause inconsistencies in their pairwise alignment and in the determination of optimal number of changes. In these situations triangle inequalities are likely to be so large that the adjustments needed would modify the step matrix in a way that does not truly represent the relative number of steps required to transform one sequence into another for most of the cells of that step matrix. Finally, coded regions with a large number of character states, even if they are parsimony-informative when subjected to an equally weighted step matrix, can generate the same phylogenetic artifact (asymmetric phylogenies) as the uninformative coded regions that become informative when subjected to an unequally weighted step matrix. Simulation studies are needed to determine when such an artifact is more likely to happen. Based on the implementation of our proposed method on several different data sets, our rule of thumb has been to exclude any ambiguously aligned regions requiring >15 character states.

Therefore, in addition to integrating ambiguously aligned sequences in phylogenetic analyses without violating positional homology, this new method provides an objective criterion to exclude regions that are likely to jeopardize phylogenetic accuracy. The exclusion of sites involves a certain degree of subjectivity and, therefore, is open to criticism (Gatesy et al., 1993) and improvement. However, the choice of gap:nucleotide substitution cost ratios needed for the elision or direct optimization (e.g., POY) of DNA sequences also involves subjectivity (see Giribet and Wheeler, 1999). Most likely, the complete removal of subjectivity in the delimitation of ambiguously aligned regions is utopian (Swofford et al., 1996). However, with the method proposed here, compared with direct optimization (POY) of DNA sequences, the investigator still has access to the full range of analytical flexibility and complexity available in PAUP*.

Contribution of Ambiguously Aligned Sequences to Estimations of Branch Length

Estimation of branch lengths is expected to be improved with our method, even when compared with estimates corrected by an evolutionary model under the maximum likelihood optimization. Branch lengths are likely to be underestimated in an uneven way if they are derived exclusively from the unambiguous and often slowest-evolving regions of a molecule. Unfortunately, there is no satisfactory way, to our knowledge, to integrate ambiguously aligned regions into a phylogenetic search by using maximum likelihood as the optimization criterion.

Contribution of Ambiguously Aligned Sequences to Large-Scale Phylogenies

One of the major impediments in large-scale phylogenetic studies is the low ratio for number of informative sites to number of sequences. A large number of very short or zero-length internodes, resulting from few or no synapomorphies, will greatly reduce the efficiency of a search because of the overwhelming abundance of equally parsimonious suboptimal and optimal trees. Another problem associated with large-scale phylogenetic studies involving a broad taxonomic sampling is the ever-increasing expansion of ambiguously aligned regions and the creation of additional ambiguous regions associated with the inclusion of more distant taxa in the alignment. This translates into a net loss of unambiguously aligned informative sites. Our method provides accessibility to the phylogenetic signal residing in ambiguously aligned characters, which otherwise would be excluded or downweighted (elision). Because these characters are likely to be rapidly evolving, they can provide additional synapomorphies in regions of the tree that previously had very short or zero-length internodes. The resulting reduction in the number of optimal and suboptimal equally parsimonious solutions should reduce the time needed for search algorithms to converge on the most-parsimonious tree or trees.

Data sets with large numbers of sequences and high intersequence divergences are likely to generate a prohibitively high number of equally optimal alignments in the elision approach. Because the degree to

which an ambiguously aligned site is downweighted is proportional to its interalignment variability with elision, many sites could be downweighted to the extent that they would be effectively excluded from the phylogenetic analysis. Therefore, the contribution of the elision approach implemented within large-scale phylogenetic studies is likely to be minimal and could even be detrimental if too broad a range of gap-to-substitution cost ratios is used (Fig. 10).

Our method offers several advantages over elision (Wheeler et al., 1995) and direct optimization of DNA sequences (POY; Wheeler, 1996), especially for dealing with large data sets. The delimitation of ambiguously aligned regions imposed by our method requires a thorough and careful inspection of alignments. Whereas, this familiarization with the data set allows one to refine the alignment, it can also lead to the identification of specific evolutionary attributes (e.g., rate heterogeneity across sites, base composition bias, genetic change bias) that must be accounted for in a rigorous phylogenetic analysis of large-scale data sets. Another advantage of our approach is that regions that are not ambiguously aligned but include gaps at two or more contiguous sites (e.g., Fig. 1c) can be systematically recoded. This allows the user to use gaps as a fifth character state on the remaining data set without worrying about the effect of overweighting stretches that contain gaps at more than one contiguous site.

Future Developments

We have implemented our method by using the program INAASE (INtegration of Ambiguously Aligned SEquences). In future developments of our method we expect to use a better pairwise alignment procedure than the one implemented in the current version of our program INAASE 0.2c1, provide a weighting scheme for the coded regions that would integrate probabilities of every possible type of change in those regions, and extend this approach to phylogenetic searches by using maximum likelihood as the optimization criterion. The current version of INAASE 0.2c1 uses a dynamic programming alignment approach (Needleman and Wunsch, 1970) to find all the best alignments of two sequences.

Contrary to the step matrices applied to the unambiguous portions of the alignment, the step matrices applied to the coded ambiguous regions do not provide differential weights for different types of changes among the coded character states. For example, a change from A to C or C to A could have a weight of two steps, based on the frequency of this type of change derived from the unambiguously aligned portion of the alignment. Currently, however, every type of change within a delimited ambiguously aligned region is counted as one step, in INAASE. This discrepancy between the two weighting schemes—for the unambiguously aligned versus ambiguously aligned regions—arises from the difficulty in estimating the frequency of changes in a region of the alignment that is ambiguously aligned and likely to be subject to a very different evolutionary model. However, we do not believe that this problem is insurmountable.

Currently, only phylogenetic searches using parsimony can benefit from our approach. This means that ambiguously aligned regions still need to be excluded from maximum likelihood searches. To extend this method to phylogenetic searches that use maximum likelihood as the optimization criterion is by far our greatest future challenge.

PROGRAM AVAILABILITY

The program INAASE (INtegration of Ambiguously Aligned SEquences) has been developed to implement the unequivocal coding and optimal weighting of ambiguously aligned sequences described in this paper. INAASE can be obtained by contacting F.L. The program was written by P.W. in C for the power PC (Apple) platform and modified by S.Z.

ACKNOWLEDGMENTS

We are particularly thankful to Karl Kjer for his thorough reading of the manuscript. His comments and suggestions greatly improved this paper. We thank Paul Lewis for his useful comments on how to improve the alignment algorithm implemented in our program INAASE, Rob DeSalle and John Gatesy for providing data matrices, and Kathleen Pryer, Chris Brochu, Jutta Buschbom, Andrew Miller, and Rebecca Rincker for a critical reading of the manuscript. This research was supported by a National Science Foundation DEB-9615542 grant awarded to F.L.

NOTE ADDED IN PROOF

W. Wheeler (1999; Cladistics 15:379–385) published a similar approach while our manuscript was in press.

REFERENCES

BALDWIN, B. G., M. J. SANDERSON, J. M. PORTER, M. F. WOJCIECHOWSKI, C. S. CAMPBELL, AND M. J. DONOGHUE. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. Ann. Mo. Bot. Gard. 82:257–277.

BARRIEL, V. 1994. Phylogénies moléculaires et insertions-délétions de nucléotides. C. R. Acad. Sci. Paris 317:693–701.

BAUM, D. A., K. J. SYTSMA, AND P. C. HOCH. 1994. A phylogenetic analysis of *Epilobium* (Onagraceae) based on nuclear ribosomal DNA sequences. Syst. Bot. 19:363–388.

BERBEE, M. L. 1996. Loculoascomycete origins and evolution of filamentous ascomycetes morphology based on 18S rRNA gene sequence data. Mol. Biol. Evol. 13:462–470.

BERBEE, M. L., AND J. W. TAYLOR. 1993. Dating the evolutionary radiations of the true fungi. Can. J. Bot. 71:1114–1127.

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

BROCHU, C. A. 1997. Morphology, fossils, divergence timing, and the phylogenetic relationships of *Gavialis*. Syst. Biol. 46:479–522.

BRUNS, T. D., R. VILGALYS, S. M. BARNS, D. GONZALEZ, D. S. HIBBETT, D. J. LANE, L. SIMON, S. STICKEL, T. M. SZARO, W. G. WEISBURG, AND M. L. SOGIN. 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. Mol. Phylogenet. Evol. 1:231–241.

CERCHIO, S., AND P. TUCKER. 1998. Influence of alignment on the mtDNA phylogeny of Cetacea: questionable support for a Mysticeti/Physeteroidea clade. Syst. Biol. 47:336–344.

DESALLE, R., C. WRAY, AND R. ABSHER. 1994. Computational problems in molecular systematics. Pages 353–370 *in* Molecular ecology and evolution: approaches and applications (B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, eds.). Birkhäuser Verlag, Boston.

DOYLE, J. J., AND J. I. DAVIS. 1998. Homology in molecular phylogenetics: a parsimony perspective. Pages 101–131 *in* Molecular systematics of plants, 2nd edition (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

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

FELSENSTEIN, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linn. Soc. 16:183–196.

FELSENSTEIN, J. 1985. Confidence intervals on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

FENG, D., AND R. F. DOOLITTLE. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360.

FENG, D., AND R. F. DOOLITTLE. 1990. Progressive alignment and phylogenetic tree reconstruction of protein sequences. *In* Molecular evolution: computer analysis of protein and nucleic acid sequences (R. F. Doolittle, ed.). Methods Enzymol. 183: 375–387 .

FERNANDEZ, F., F. LUTZONI, AND S. M. HUHNDORF. 1999. Teleomorph–anamorph connections: the new pyrenomycetous genus *Carpoligna* and its *Pleurothecium* anamorph. Mycologia 91:251–262.

GATESY, J., R. DESALLE, AND W. WHEELER. 1993. Alignment–ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2:152–157.

GIRIBET, G., AND W. C. WHEELER. 1999. On gaps. Mol. Phylogenet. Evol. 13:132–143.

HEIN, J. 1990. Unified approach to alignment and phylogenies. *In* Molecular evolution: computer analysis of protein and nucleic acid sequences (R. F. Doolittle, ed.). Methods Enzymol. 183:626–644.

HIBBETT, D. S., AND R. VILGALYS. 1993. Phylogenetic relationships of *Lentinus* (Basidiomycotina) inferred from molecular and morphological characters. Syst. Bot. 18:409–433.

HIBBETT, D. S., Y. FUKUMASA-NAKAI, A. TSUNEDA, AND M. J. DONOGHUE. 1995. Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. Mycologia 87:618–638.

HIGGINS, D. G., J. D. THOMPSON, AND T. J. GIBSON. 1996. Using CLUSTAL for multiple sequence alignments. *In* Computer methods for macromolecular sequence analysis (R. F. Doolittle, ed.). Methods Enzymol. 266:383–402.

HILLIS, D. M. 1994. Homology in molecular biology. Pages 339–368 *in* Homology: the hierarchical basis of comparative biology (B. K. Hall, ed.). Academic Press, San Diego.

HILLIS, D. M., AND J. P. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83:189–195.

HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

KJER, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4:314–330.

KRETZER, A., Y. LI, T. SZARO, AND T. D. BRUNS. 1996. Internal transcribed spacer sequences from 38 recognized species of *Suillus* sensu lato: phylogenetic and taxonomic implications. Mycologia 88:776–785.

KRUSKAL, J. B. 1983. An overview of sequence comparison. Pages 1–45 *in* Time warps, string edits, and macromolecules (D. Sankoff and J. B. Kruskal, eds.). Addison-Wesley, Reading, Massachusetts.

LAKE, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8:378–385.

LLOYD, D. G., AND V. L. CALDER. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. 4:9–21.

LUTZONI, F. 1995. Phylogeny of lichen- and non-lichen-forming omphalinoid mushrooms and the utility of testing for combinability among multiple data sets. Syst. Biol. 46:373–406.

MADDISON, D. R., D. L. SWOFFORD, AND W. P. MADDISON. 1997. NEXUS: An extensible file format for systematic information. Syst. Biol. 46:590–621.

MADDISON, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42:576–581.

MADDISON, W. P., AND D. R. MADDISON. 1992. MacClade: analysis of phylogeny and character evolution, vers. 3.0. Sinauer, Sunderland, Massachusetts.

MANOS, P. S. 1997. Systematics of *Nothofagus* (Nothofagaceae) based on rDNA spacer sequences (ITS): taxonomic congruence with morphology and plastid sequences. Am. J. Bot. 84:1137–1155.

MICKEVICH, M. F., AND J. S. FARRIS. 1981. The implications of congruence in *Menidia*. Syst. Zool. 27:143–158.

MINDELL, D. P. 1991. Aligning DNA sequences: homology and phylogenetic weighting. Pages 73–89 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

NEEDLEMAN, S. B., AND C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

PATTERSON, C. 1982. Morphological characters and homology. Pages 21–74 *in* Prospects in systematics (K. Joysey and A. Friday, eds.). Academic Press, London.

PATTERSON, C. 1988. Homology in classical and molecular biology. Mol. Biol. Evol. 5:603–625.

PLATNICK, N. I., C. E. GRISWOLD, AND J. A. CODDINGTON. 1991. On missing entries in cladistic analysis. Cladistics 7:337–343.

ROTH, V. L. 1988. The biological basis of homology. Pages 1–26 *in* Ontogeny and systematics (C. J. Humphries, ed.). Columbia Univ. Press, New York.

SANKOFF, D. D., AND R. J. CEDERGREN. 1983. Simultaneous comparison of three or more sequences related by a tree. Pages 253–264 *in* Time warps, string edits, and macromolecules: the theory and practice of sequence comparison (D. Sankoff and F. B. Kruskal, eds.). Addison-Wesley, Reading, Massachusetts.

SOLTIS, D. E., L. A. JOHNSON, AND C. LOONEY. 1996. Discordance between ITS and chloroplast topologies in the *Boykinia* group (Saxifragaceae). Syst. Bot. 21:169–185.

SPATAFORA, J. W. 1995. Ascomal evolution of filamentous ascomycetes: Evidence from molecular data. Can. J. Bot. 73(Suppl. 1):S811–S815.

SPATAFORA, J. W., AND M. BLACKWELL. 1993. Molecular systematics of unitunicate perithecial ascomycetes: The Clavicipitales–Hypocreales connection. Mycologia 85:912–922.

STEVENS, P. F. 1991. Character states, morphological variation, and phylogenetic analysis: A review. Syst. Bot. 16:553–583.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods), vers. 4.0. Sinauer, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

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

THORNE, J. L., H. KISHINO, AND J. FELSENSTEIN. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.

THORNE, J. L., H. KISHINO, AND J. FELSENSTEIN. 1992. Inching toward reality: An improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.

VINGRON, M., AND M. S. WATERMAN. 1994. Sequence alignment and penalty choice: Review of concepts, case studies and implications. J. Mol. Biol. 235:1–12.

VOGLER, A. P., AND R. DESALLE. 1994. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. Mol. Biol. Evol. 11:393–405.

WATERMAN, M. S., M. EGGERT, AND F. LANDER. 1992. Parametric sequence comparisons. Proc. Natl. Acad. Sci. USA 89:6090–6093.

WEIR, B. 1990. Genetic data analysis. Sinauer, Sunderland, Massachusetts.

WHEELER, W. C. 1990. Combinatorial weights in phylogenetic analyis: a statistical parsimony procedure. Cladistics 6:269–275.

WHEELER, W. C. 1994. Sources of ambiguity in nucleic acid sequence alignment. Pages 323–352 *in* Molecular ecology and evolution: approaches and applications (B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, eds.). Birkhäuser Verlag, Boston.

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

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

WHEELER, W. C., J. GATESY, AND R. DESALLE. 1995. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4:1–9.

WILEY, E. O. 1981. Phylogenetics: the theory and practice of phylogenetic systematics. Wiley, New York.

Received 11 January 1999; accepted 15 April 1999
Associate Editor: R. Olmstead