Molecular Phylogenetics and Evolution
Vol. 13, No. 1, October, pp. 132–143, 1999
Article ID mpev.1999.0643, available online at http://www.idealibrary.com on IDEAL®

On Gaps

Gonzalo Giribet and Ward C. Wheeler

Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024

Received September 24, 1998; revised February 5, 1999

Gaps result from the alignment of sequences of unequal length during primary homology assessment. Viewed as character states originating from particular biological events (mutation), gaps contain historical information suitable for phylogenetic analysis. The effect of gaps as a source of phylogenetic data is explored via sensitivity analysis and character congruence among different data partitions. Example data sets are provided to show that gaps contain important phylogenetic information not recovered by those methods that omit gaps in their calculations. However, gap cost schemes are arbitrary (although they must be explicit) and thus data exploration is a necessity of molecular analyses, while character congruence is necessary as an external criterion for hypothesis decision. © 1999 Academic Press

Key Words: **gaps; sequence alignments; phylogenetic analysis; character congruence; parsimony; maximum likelihood.**

INTRODUCTION: WHAT IS A GAP?

The first step in any systematic or evolutionary study is to establish provisional homology statements ("primary homology" *sensu* de Pinna, 1991; "topographic identity" *sensu* Brower and Schawaroch, 1996). In the case of DNA data, sequence alignment represents the primary homology hypothesis. The kernel of sequence alignment is the dynamic programming algorithm of Needleman and Wunsch (1970). In this procedure, transformation and insertion/deletion (indel) costs are established, and sequences are aligned via the insertion of gaps.

From a molecular point of view, Li (1997: 28) described gaps as follows:

> Deletions and insertions are collectively referred to as gaps (or indels), because when a sequence involving either an insertion or a deletion is compared with the original sequence, a gap will appear in one of the two sequences.

Li (1997: 28) also differentiates two sorts of gaps based on their length and origin:

> The length of the gaps essentially exhibit a bimodal type of frequency distribution, with short gaps (up to 20–30 nucleotides) being mostly caused by errors in the process of DNA replication, such as the slipped-strand mispairing [. . .], and with long insertions and deletions occurring mainly because of unequal crossing-over or DNA transposition.

From this definition, it can be implied that gaps originate from particular biological events such as mutation (insertion or deletion) and thus they may contain the same historical information as observed in nucleotide changes.

The insertion of gaps to accommodate homologous DNA sequences of unequal length is a necessity in the first steps of the phylogenetic analyses:

> . . . a gap assumes that a deletion or an insertion has occurred at this position in one of the two sequences. Thus, an alignment represents a specific hypothesis about the evolution of the two sequences. (Li, 1997: 91)

These alignment gaps allow the nucleotide base correspondences to be interpreted as putative homologies and phylogenetic analysis to proceed.

The preceding is the case of the "classical" approach to analyze DNA sequence data: the pairwise/multiple sequence alignment approach. In this case a general pattern of observation (sequence data) is followed by the alignment (insertion of gaps) to phylogeny reconstruction (parsimony or other method). During this process, sequence 'gaps' are created and treated as a fifth character state, although they are not observations but rather placeholders signifying a specific type of transformation event (Wheeler, 1996). Stated simply, nucleotide bases are observable while gaps are not.

The novel method (direct optimization of DNA sequences) proposed by Wheeler (1996), contrary to static alignments, avoids the problem of alignment by generalizing phylogenetic character analysis to include insertion/deletion events (indels). By doing this, the analysis proceeds directly from the sequence data to phylogenetic reconstruction, obviating the necessity to create gap characters. Indels appear not as states but as transformations linking ancestral and descendent nucleotide sequences.

Although in the direct optimization approach gaps are not treated as characters but as transformations, it is necessary in both approaches to define explicitly the gap:change cost to be used. It is well known that different gap:change costs result in distinct phylogenetic hypotheses (e.g., Fitch and Smith, 1983).

1055-7903/99 $30.00

Alignment-derived gap costs are not directly measurable in the absence of a predetermined phylogeny, although it may be possible to estimate their values through appeal to an external optimality criterion. This idea led Wheeler (1995) to choose congruence (for taxonomic congruence, see Nelson, 1979; for character-based congruence, see Mickevich and Farris, 1981) as the optimality criterion for phylogenetic analysis. More recently, character-based congruence has been embraced to be a more appropriate criterion to measure precision objectively (the agreement among data) and choose among competing hypotheses (Wheeler, 1995, 1998; Wheeler and Hayashi, 1998).

HOW ARE GAPS TREATED?

Despite the fact that gaps are a necessity (and should be explicitly weighted) for sequence alignment (as well as direct optimization), a surprising number of articles have disregarded them as a source of phylogenetic information. Most of the parsimony analyses of molecular data published in the literature treat gaps as missing data, or if they treat gaps as a character state, do not assign them the same gap cost used to generate the alignment. Very few use gap costs in the analyses as defined in the alignment. Moreover, none of the other phylogenetic reconstruction methods account for gap information at all. In fact, it is a prerequisite of some programs and methods to eliminate gaps from the alignment:

> The sequences must be aligned, and sites involving gaps will be removed from all sequences before analysis, with appropriate adjustment to the sequence length. (Yang, 1997)

> . . . insertions and deletions are ignored, and it is assumed that the sequences are aligned with gaps removed. (Yang and Rannala, 1997)

In particular, neighbor-joining methods are unable to accommodate indel information in their calculations. It has been said that maximum likelihood could incorporate gap information into a model of evolution, although to date only one model that accommodates single indel events has been developed (Thorne *et al.,* 1991).

In what has been claimed to be the most cited article for phylogenetic inference methods, very little is said about gaps and their importance in phylogenetic analyses:

> Although the character state "gap" is sometimes treated as a fifth base [. . .], the processes responsible for base substitution and for insertion and deletion are evolutionarily and mechanistically distinct. Because a proper treatment is not obvious, sequence positions with gaps are usually omitted from analyses in one of two ways (e.g. Kumar *et al.,* 1993; Swofford, 199[8]). (Swofford *et al.,* 1996: 453)

Although it is true that the processes responsible for base substitution and for insertion and deletion are evolutionarily and mechanistically distinct from base transformations, to our criterion the "proper treatment" of gaps should be to explore their influence on both alignments and phylogenetic inference, not to disregard them.

Continuing,

> Once again we emphasize that regions of the sequence alignment that contain substantial numbers of alignment gaps should be omitted from the analysis; positional homology is too uncertain for reliable estimates to be made from these regions. (Swofford *et al.,* 1996: 453)

But data removal is a problematic issue, and an objective criterion is required if data have to be excluded. In that sense, Gatesy *et al.* (1993) and Wheeler *et al.* (1995) proposed an objective criterion for the accommodation of ambiguous sites in the case of multiple alignments (see a review in DeSalle *et al.,* 1994).

Some authors have studied the phylogenetic effect of gaps when coded as missing data or as a fifth character state (but only in the context of parsimony). However, the relationship between the gap cost assigned in alignments and the gap weight used in phylogenetic reconstruction is often disconnected in such studies. The logical approach would be to use the same gap cost in both steps (alignment and phylogenetic inference) of the process. Nevertheless, very few authors employ the same cost matrices in alignment and phylogenetic inference or conduct a sensitivity analysis. In the case of direct optimization, there is no way to disregard gap information because there is no intermediate step (alignment). Gap costs must always be explicitly defined and used as historical information.

Due to the lack of discussion in the literature and the apparent confusion about how to treat gaps, our intention is to discuss the use of gap information in phylogenetic analyses. Gaps are also crucial in hypothesis testing from a theoretical point of view. We provide evidence of the importance of gap information in the phylogenetic analysis of sequence data and we demonstrate how methods that dismiss gap information *a priori* will fail to recover historical information concordant to that of methods that consider gaps.

HOW SHOULD GAPS BE TREATED?

The phylogenetic inference process via multiple sequence alignments is divided into two stages: (1) alignment of DNA sequences and (2) tree inference. Gap costs are required for stage 1; otherwise, sequences of different lengths would not be suitable for analysis. It is therefore illogical to omit gaps from the second part of the process.

Whether alignments are generated manually (by eye) or automatically (by a computer program), gaps should be assigned a specific cost. In manual alignments, primary homology is inferred through intuition, through inference from molecular structures, or through combination of both factors (Titus and Frost, 1996). In this case, an implicit weight is assigned to gaps, being

the gap value always ≥½ change cost according to the triangle inequality (Wheeler, 1993). Manual alignments have been criticized for the lack of objectivity, lack of repeatability, and lack of a criterion for the exclusion of data (Gatesy *et al.,* 1993; DeSalle *et al.,* 1994). Gap cost is explicit in automatic alignments. However, it is arbitrarily determined (generally higher than change costs) and in some cases special cost values can be defined to favor or to penalize gap extension.

Which gap:change cost ratio should be chosen for a particular data set? There is no general answer to this question *a priori.* DeSalle *et al.* (1994) commented on the possibility of using a range of gap:change cost ratios and assessing the situation based on this range of alignments (see also Waterman *et al.,* 1992; Gatesy *et al.,* 1993). We stress this necessity for data exploration and the requirement of an optimality criterion to choose among alternative topologies, as presented by Wheeler (1995). The "parameter sensitivity" approach has its own drawbacks (i.e., the number of parameters to be explored is unbounded) but in this case they are operational and not epistemological. New values (in this case more gap:change cost ratios) could always be added. Particularly, fractional costs could be incorporated to points surrounding the parameter that maximized congruence in the first round, with subsequent search of new points (adaptive sampling).

Many alignment programs offer the option of having lower cost gaps ("extension" cost) after the initial or "opening" gap. These lower extension costs are employed to favor longer, more contiguous gaps as opposed to a myriad of small individual gaps. The extension costs can also be modulated to reflect coding regions: gaps whose lengths are divisible by three with lower costs than those which might disrupt a reading frame. The central assumption of this sort of approach is that insertion–deletion events should be treated wherever possible as single events. That all seven gaps in a row were actually created by a single deletion of the whole series of bases might well be true but any analysis which creates such a tight dependency of costs among aligned positions will run afoul of the postulates of the phylogenetic analysis of characters—that they are at least logically independent.

All cost regimes that do not treat gaps as independent events, but which treat character columns as independent, are logically inconsistent. This flows from the simple fact that when diagnosing a cladogram, the determination of whether a particular nucleotide base adds length requires only that we know the cladogram topology and the character states at the terminals. With the interpretation of gap cost based on how many gaps precede or follow that gap, the character positions are no longer independent. To determine whether that gap adds length to a cladogram requires knowledge of all the other positions in the matrix. Furthermore, if gaps are found in two taxa, but one of them is an "initial" gap and one "extended," do we assign a cost of transformation between these two sorts of gaps? How would we do this and how could we justify a transformation cost between two identical character states?

It is important to remember the distinction between what actually happened historically and how we analyze data. It may be that gap positions are highly interdependent. However, we are still required to analyze them separately because the logic of our analysis demands it. No matter what technique we employ to reconstruct cladograms, we always assume epistemological independence in our characters. Without this notion, phylogenetic analysis would be impossible.

CONGRUENCE AS AN OBJECTIVE WAY TO DECIDE GAP:CHANGE COST RATIOS

Once different gap costs have been explored for a particular data set, a decision should be made to favor one of the competing hypotheses. In general this decision is made by the investigator based on an *a priori* hypothesis of relation-

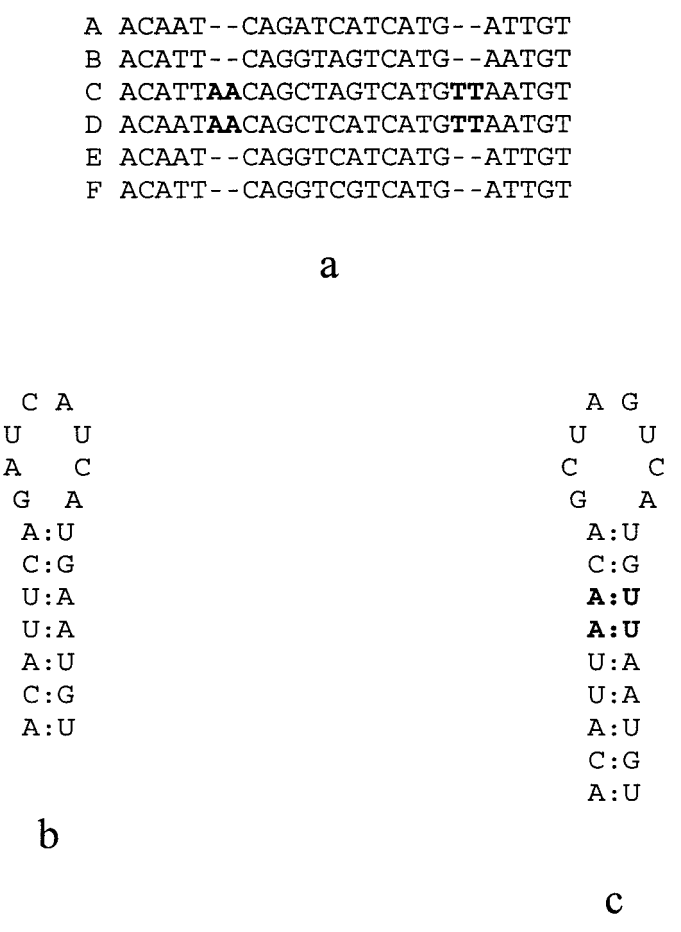

FIG. 1. (a) Data set representing a hypothetical stem of a ribosomal-like gene. (b) Secondary structure prediction of sequence B. (c) Secondary structure prediction of sequence C.

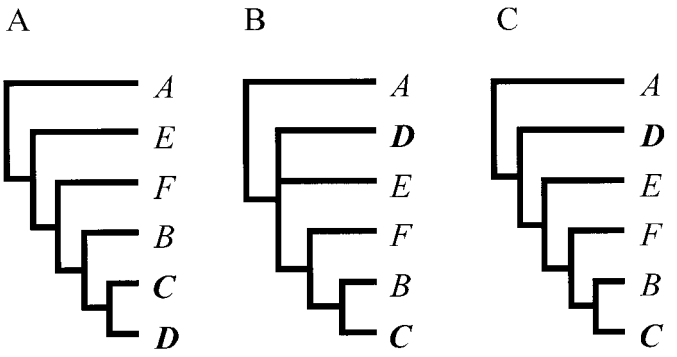

FIG. 2. Topologies obtained for the data set represented in Fig. 1 under (A) parsimony when gaps are treated as a character state (all gap weights ranging from 1 to ∞); (B) parsimony when gaps are treated as missing data (strict consensus of three equally parsimonious trees); and (C) maximum likelihood analysis (F81).

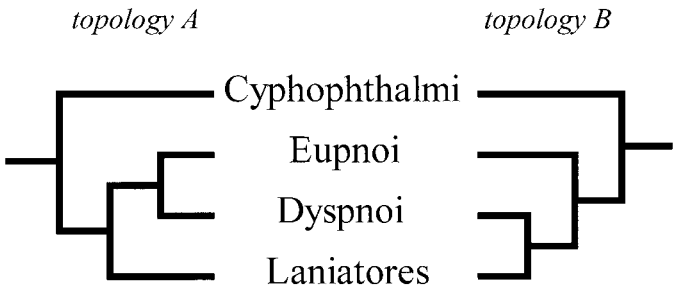

FIG. 3. The two competing hypotheses. *Topology A* implies monophyly of Palpatores (Eupnoi + Dyspnoi). *Topology B* implies paraphyly of Palpatores, being the Dyspnoi sister group to Laniatores.

TABLE 1

Taxa Used in this Study with the Supraspecific Categories

Class Chelicerata	
Subclass Merostomata	
Order Xiphosura	*Limulus polyphemus*
	Carcinoscorpius rotundicaudatus
Subclass Arachnida	
Order Solifugae	*Eusimonia wunderlichi*
Order Ricinulei	*Pseudocellus pearsei*
Order Scorpionida	*Androctonus australis*
Order Opiliones	
Cyphophthalmi	
Family Sironidae	*Siro rubens*
	Parasiro coiffaiti
Family Stylocellidae	*Stylocellus* sp.
'Palpatores'—Eupnoi	
Superfamily Phalangioidea	
Family Phalangiidae	*Odiellus troguloides*
Family Leiobunidae	*Nelima sylvatica*
Superfamily Caddoidea	
Family Caddidae	*Caddo agilis*
'Palpatores'—Dyspnoi	
Superfamily Ischyropsalidoidea	
Family Ischyropsalidae	*Ischyropsalis luteipes*
Superfamily Troguloidea	
Family Dicranolasmatidae	*Dicranolasma soerenseni*
Family Nemastomatidae	*Centetostoma dubium*
Laniatores	
Superfamily Travunioidea	
Family Triaenonychidae	*Equitius doriae*
Superfamily Oncopodoidea	
Family Oncopodidae	*Oncopus* cf. *alticeps*
Superfamily Gonyleptoidea	
Family Phalangodidae	*Maiorerus randoi*
	Scotolemon lespesi
Family Cosmetidae	*Gnidia holnbergii*
Family Gonyleptidae	*Pachyloides thorellii*

Note. Molecular and morphological data extracted from Giribet *et al.* (1999).

ships, background knowledge, or some other subjective criteria. This is why Wheeler (1995) introduced the sensitivity analysis concept to phylogenetic systematics, in which external criteria such as taxonomic or character congruence were used to make a decision on the hypothesis to be chosen, without relying on subjectivity.

Taxonomic congruence as used by Wheeler (1995) is probably not a good criterion to choose among competing hypotheses because the groups have to be defined arbitrarily (perhaps based on current taxonomy). However, taxonomic congruence can be very informative in comparing some taxonomic groups to hypotheses obtained. Character congruence is thus the only logical, objective, and operational criterion to be used in determining gap costs in sequence alignments and phylogenetic analyses. If measures of character congruence are not available (e.g., a single partition or extremely large data sets), a decision should not be made and results of a full set of gap costs should be shown, although recognizing that extreme gap costs may be unlikely to yield a congruent hypothesis of relationships.

TABLE 2

Tree Length for Each Gap Cost for the Different Data Sets

	Gap = 1	Gap = 2	Gap = 4	Gap = 8
18S rDNA	750	792	876	1055
18S rDNA (no gap)	648	704	699	717
Morphology 1	70	70	70	70
Morphology 2	70	140	280	560
Combined 1	823	864	947	1126
Combined 2	823	934	1159	1619
Combined 3	721	776	772	790
MF1	0.0036	0.0023	0.0011	0.0009
MF2	0.0036	0.0021	0.0026	0.0025
MF3	0.0042	0.0026	0.0039	0.0038

Note. 18S rDNA; 18S rDNA (No Gap; Gap = Missing Data); Morphology 1 (Morphology Always Weighted 1); Morphology 2 (Morphology = Gap Cost); Combined 1 (18S rDNA + Morphology 1); Combined 2 (18S rDNA + Morphology 2); Combined 3 (18S rDNA (No Gap) + Morphology 1). MF1 (Mikevich-Farris incongruence length difference for combined 1); MF2 (idem for combined 2); MF3 (idem for combined 3).

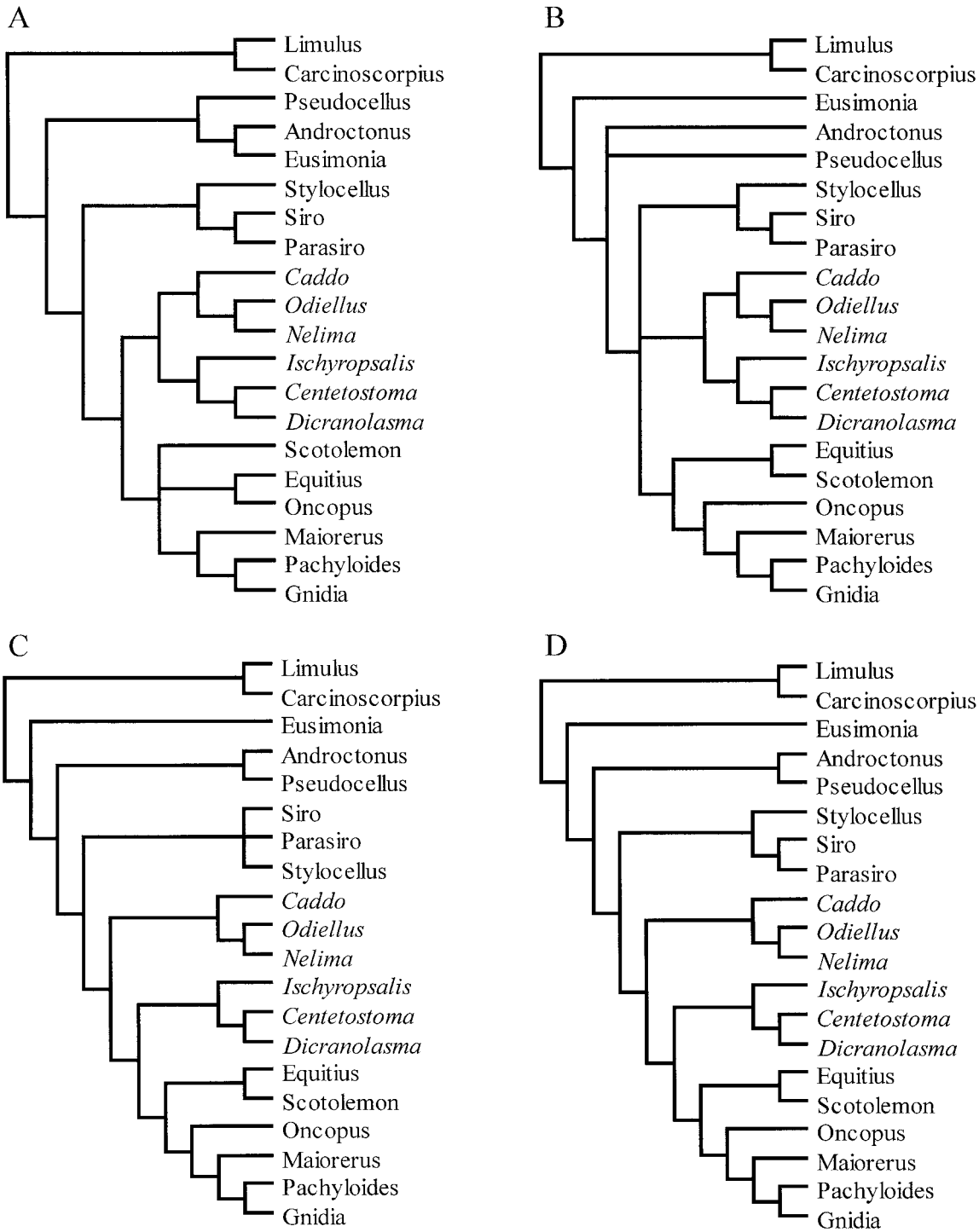

FIG. 4. Consensus trees (when applicable) of the MPT's for the parsimony analyses of the 18S rDNA data set, using gap costs of 1, 2, 4, and 8 (A, B, C, and D, respectively).

WHY THE METHODS THAT DO NOT CONSIDER GAP INFORMATION ARE 'POSITIVELY MISLEADING'?

If gaps originate from particular biological events (mutation) (e.g., Li, 1997) and not as simple placeholders in an alignment, thus reflecting historical information, ignoring them in the phylogenetic analyses may yield misleading topologies. In the following examples we stress this aspect of gap information.

Example I

We present the sequence data of a hypothetical stem of a ribosomal-like gene (Fig. 1). Sequence A is consid-

ered to be the outgroup. The data set (Fig. 1a) is trivial in terms of alignment (gap costs ranging from 1 to ∞) and secondary structure prediction (Figs. 1b and 1c). In this case we observe that gaps provide essential information for reconstructing the phylogeny of these sequences. Sequences C and D share two insertions of two nucleotides each, which are in turn corroborated by the subsequent compensatory mutation.

Phylogenetic reconstruction of these data using parsimony differ depending on whether gaps are considered as a character state or as missing data (Fig. 2A and 2B, respectively). A sister-taxon relationship between sequences C and D is obtained when gap information is considered in the analysis. Other methods that do not take into account gap information, such as maximum likelihood (Fig. 2C), yield topologies like those obtained when gaps are not considered in the parsimony analysis.

Example II

In a data set of 18S rDNA sequences of the arachnid order Opiliones (Giribet, 1997; Giribet *et al.,* 1999; see Table 1 for the taxa employed) gaps reveal essential information to support one of the clades (Dyspnoi + Laniatores) and to discern between two competing hypotheses (summarized in Fig. 3). We have reanalyzed the 18S rDNA data set using the multiple sequence alignment program MALIGN parallel version 1.5 (Wheeler and Gladstein, 1994, 1995). Gap costs were explored at the alignment level for values of 1, 2, 4, and 8. Phylogenetic analyses of the alignments obtained with MALIGN were analyzed with PAUP* 4.0b1 (Swofford, 1998) under parsimony and maximum-likelihood criteria. For the parsimony analyses, we analyzed the 18S rDNA data set with gap values as in the alignment step. The combined analyses of 18S rDNA + morphology were analyzed for each gap cost. The relative weight between morphology and molecules was explored (morphology = 1 vs morphology = gap cost). Congruence among partitions was measured by the ILD metrics (Mickevich and Farris, 1981) (see Table 2). This value is calculated by dividing the difference between the overall tree length and the sum of its data components:

$$\text{ILD} = \frac{(\text{Length}_{\text{Combined}} - \text{Sum Length}_{\text{Individual Sets}})}{\text{Length}_{\text{Combined}}}$$

Maximum-likelihood analyses were conducted for the same four alignments, considering gaps as missing data (the only option implemented in PAUP*). Different models and assumptions were tested, from the simplest substitution models, adding more parameters such as base composition, numbers of substitutional classes, and incorporating among-site rate variation with gamma distribution (see Cunningham *et al.,* 1998).

Finally, we analyzed the four alignments obtained at different gap costs under parsimony criterion but coding gaps as missing data at the phylogenetic tree reconstruction step.

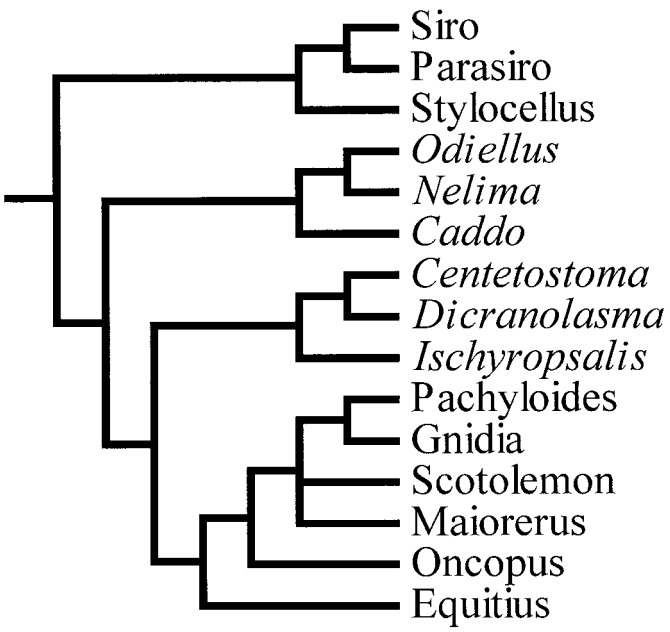

FIG. 5. Strict consensus tree (rooted after the combined analysis with molecular data) based on the morphological data set of Giribet (1997) and Giribet *et al.* (1999).

We define *topology A* to be that of monophyletic Palpatores as follows: (Cyphophthalmi (Palpatores + Laniatores)). *Topology B* refers to a paraphyletic Palpatores (Cyphophthalmi (Eupnoi (Dyspnoi + Laniatores))) (see Fig. 3). In the case of the 18S rDNA data set alone, *topology A* is obtained for gap costs of 1 and 2, while *topology B* is obtained for gap costs of 2, 4, and 8 (Fig. 4). *Topology B* is also obtained by analyzing the morphological data matrix (Fig. 5) (for details see Giribet, 1997; Giribet *et al.,* 1999). For the combined analysis of the 18S and morphological data sets, *topology A* is obtained for a gap cost equal to 1, while *topology B* is obtained for gap costs equal to 1, 2, 4, and 8 (Fig. 6). The maximum-likelihood analyses (with all models tested; F81 model represented in Fig. 7) and the parsimony analyses with gaps coded as missing data (Fig. 8) always yielded *topology A* (only in one case *topology A* was shared with a third topology; see Fig. 8D). Tree lengths for all the analyses (18S under parsimony; morphology; 18S + morphology; are shown in Table 2.

We have examined character congruence among partitions (18S rDNA + morphology). Morphology has been weighted to the unity (MF1) or equal to the gap cost (MF2). We have also measured congruence for the four alignments (generated at gap costs 1, 2, 4, and 8) when gaps were coded as missing data (MF3). The best hypothesis in the three cases corresponds to *topology B*. When morphology is weighted equal to the gap cost, the best supported tree is that with a gap cost of 2. However, when morphology weight is set at unity, we

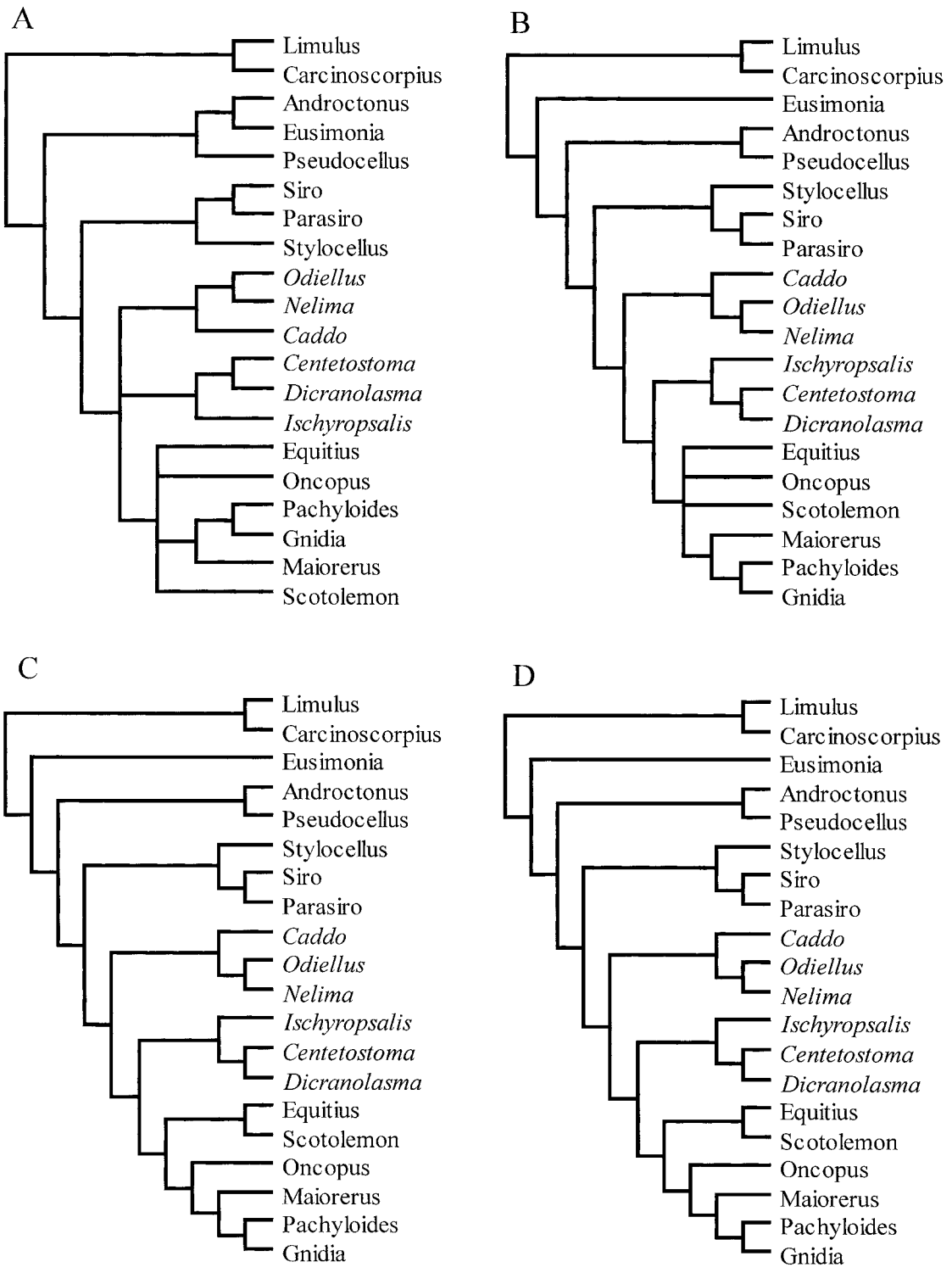

FIG. 6. Consensus trees (when applicable) of the MPT's for the parsimony analyses of the combined analyses of the 18S rDNA and morphological data sets (morphology weighted to the unity), using gap costs of 1, 2, 4, and 8 (A, B, C, and D, respectively).

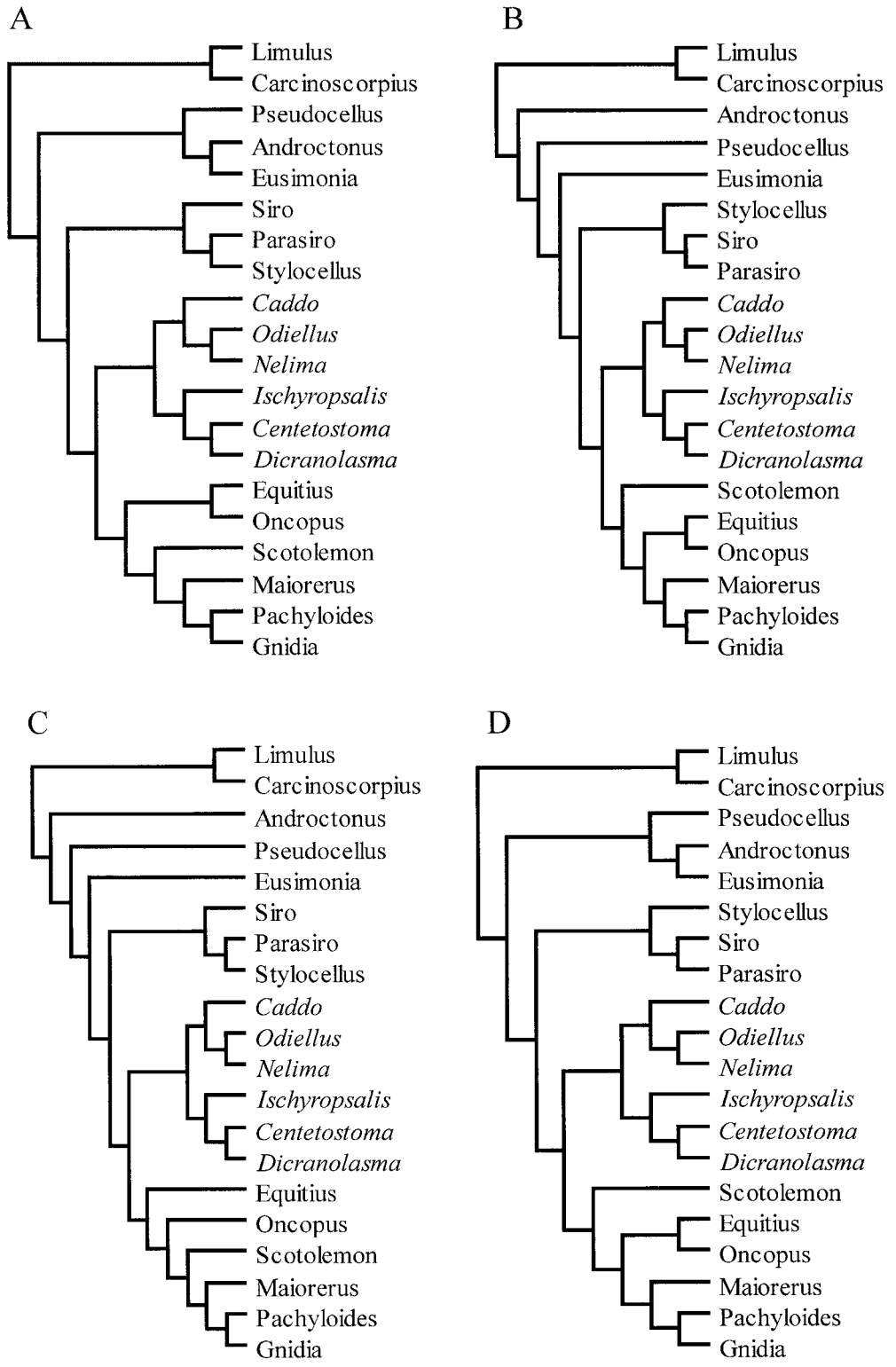

FIG. 7. Maximum likelihood (F81 model) trees of the data sets obtained after MALIGN for gap costs of 1, 2, 4, and 8 (A, B, C and D, respectively). Gaps are treated as missing data in the phylogenetic analysis.

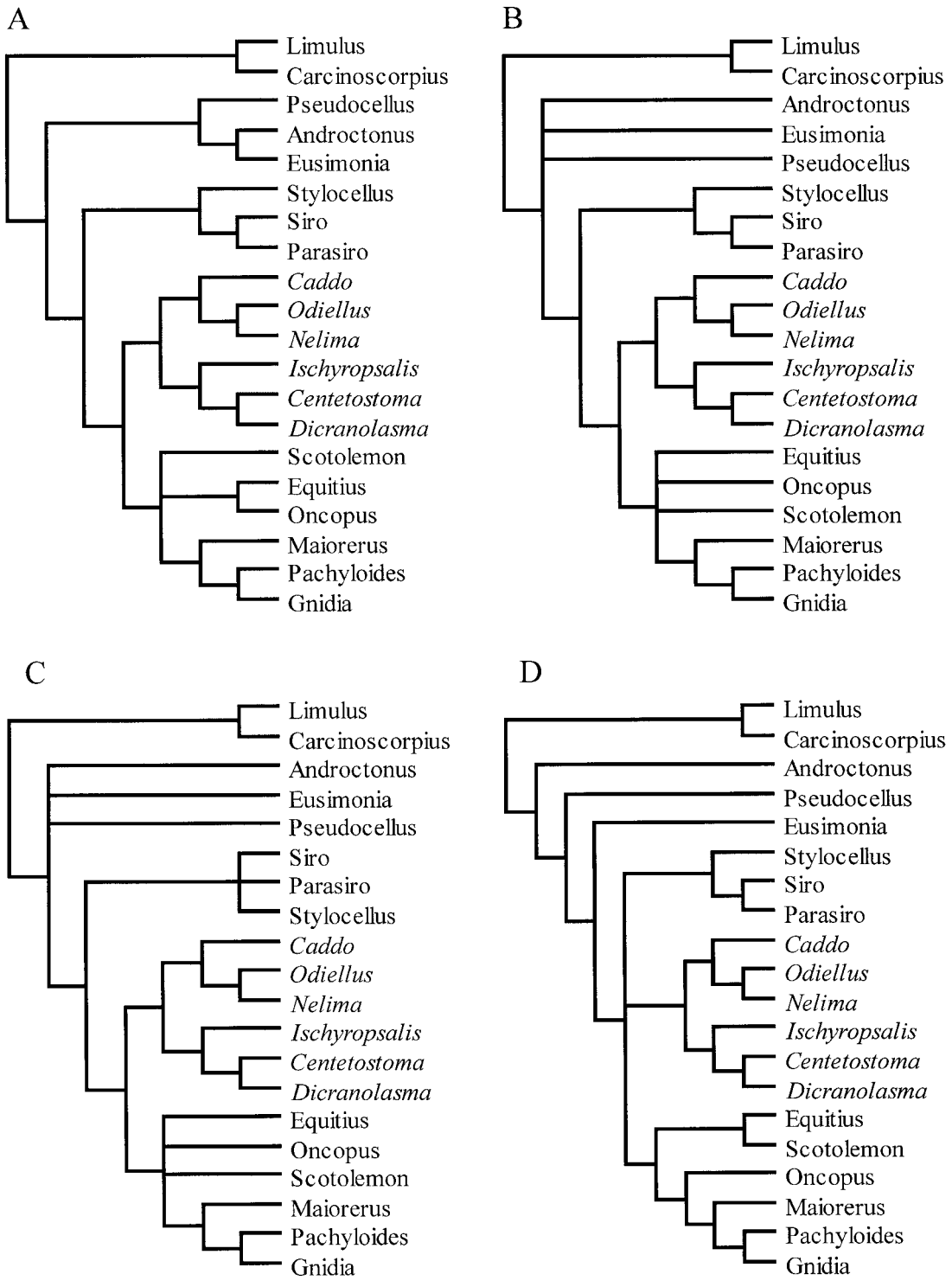

FIG. 8. Consensus trees (when applicable) of the MPT's for the parsimony analyses of the 18S rDNA data set when gaps were considered as missing data, using alignments for gap costs at 1, 2, 4, and 8 (A, B, C, and D, respectively).

observe an unusual behavior in the MF metrics. It descends (congruence augments) as the gap cost increases. This is, to our knowledge, the first described case in which congruence augments as gap cost increases. For the four alignments, the least congruent results are those in which gaps are coded as missing data. The most overall congruent result (according to ILD) of the 12 combined analyses performed is for gap cost of 8 when morphology is weighted to unity. The behavior of the ILD metrics is shown in Fig. 9.

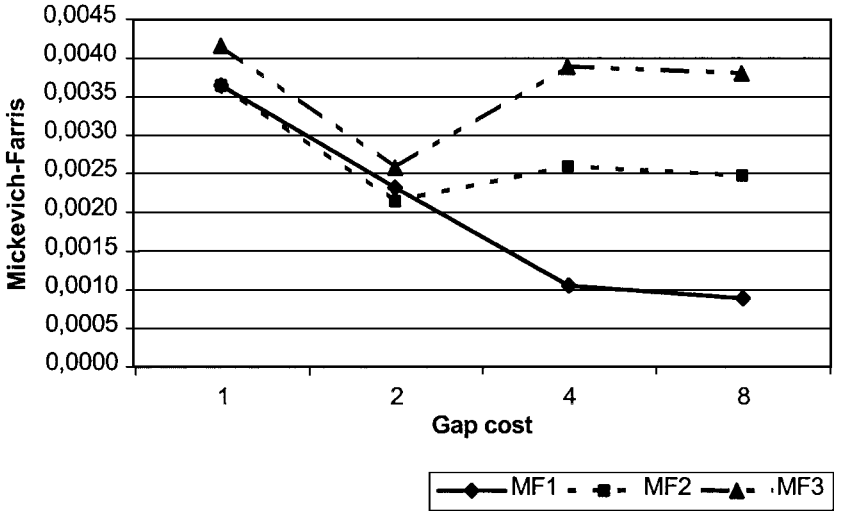

FIG. 9. Plot of the ILD metrics (Mickevich-Farris) for the three combined analyses (MF1, MF2, and MF3; see Table 2) for the alignments generated at gap costs 1, 2, 4, and 8.

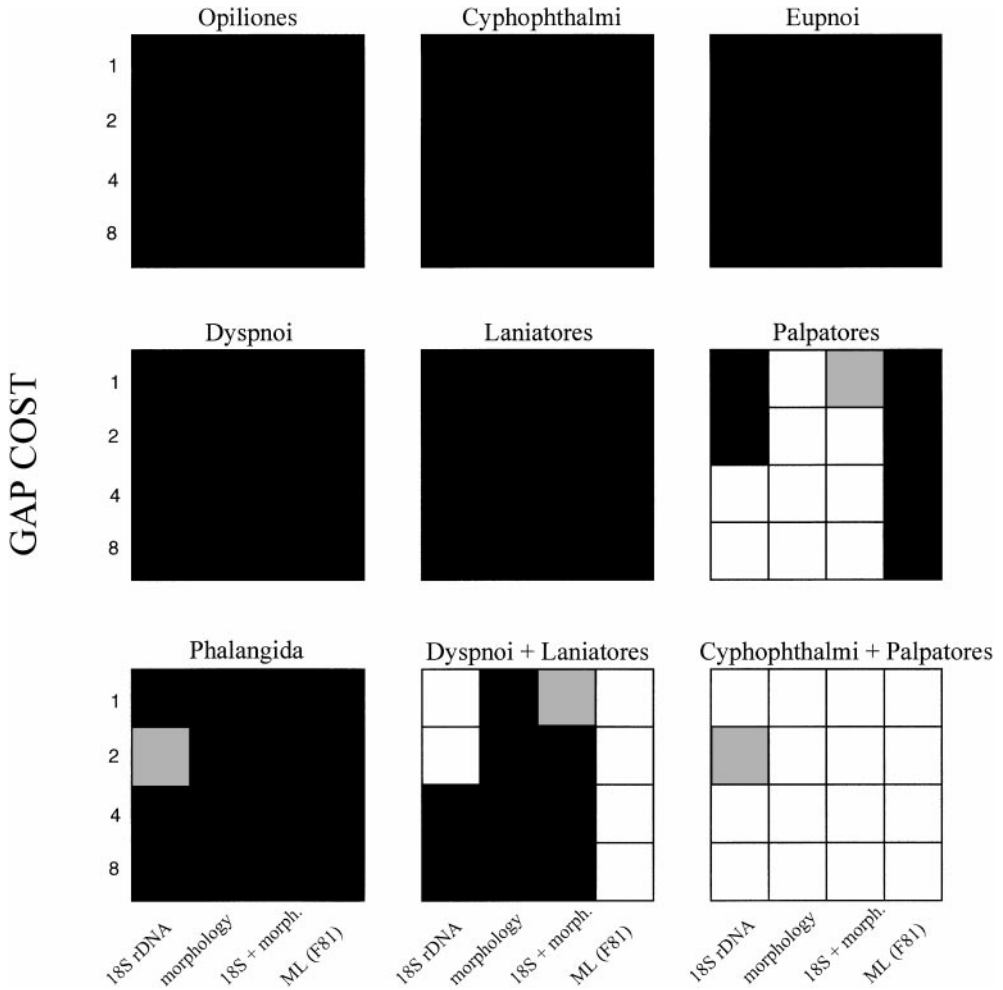

FIG. 10. Topological congruence plots for the clades indicated above the squares at gap costs of 1, 2, 4, and 8 for the following partitions: 18S rDNA, morphology, 18S + morphology (morphology weighted to the unity), and maximum likelihood (F81 model). ■, monophyletic; ▨, some of the MPT's consistent with monophyly; □, nonmonophyletic.

In this particular case (see Fig. 10), the most congruent hypothesis (combined analyses with a gap cost of 8) is compatible with the morphological analysis, as well as with the 18S rDNA analyses for gap costs of 4 and 8, corresponding to *topology B*. *Topology A* is recovered under certain parameters for the parsimony analysis of the 18S data set alone (gap cost of 1 and 2), for the combined data set (gap cost of 1), and for the parsimony analyses in which gaps are coded as missing data. The maximum-likelihood analyses match *topology A* (the least corroborated), apparently because they cannot take into account gap information.

From the analyses of these two example data sets, we not only demonstrate the importance of using gap information (in terms of congruence) but also show that the effect of gap costs are unpredictable for any data set unless a sensitivity analysis of some sort is conducted.

CONCLUSIONS AND RECOMMENDATIONS

1. Gaps constitute a valuable source of phylogenetic information and thus should be considered in phylogenetic analyses. Methods that do not consider gap information may be less explanatory since they dismiss some of the historical information.

2. Gap:change cost ratios should be defined explicitly and consistently. As a consequence, manual alignments should be avoided.

3. The same regime of gap costs used in the alignment should be used in the phylogenetic reconstruction step.

4. Gap costs are arbitrarily defined. In consequence, a range of different gap costs should be explored. Data exploration is a necessity of the phylogenetic reconstruction process to avoid hypotheses supported by unique combinations of parameter values.

5. The choice of a hypothesis based on a particular parameter scheme should be the result of a sensitivity analysis. A criterion by means of which we can decide which topology is preferred among competing hypotheses should be specified prior to the analysis. When measures of character congruence are not available (i.e., in the case of a single partition or when analyzing extremely complicated data sets), the results of the different gap schemes should be shown. An arbitrary choice of a 'preferred' tree is not defensible epistemologically.

ACKNOWLEDGMENTS

We thank Rob DeSalle, Lorenzo Prendini, Paul Goldstein, Dan Janies, Maureen O'Leary, John Wenzel, and one of the two anonymous reviewers for discussion and comments. This work was supported by a Lerner-Gray Research Fellowship to G.G.

REFERENCES

Brower, A. V. Z., and Schawaroch, V. (1996). Three steps of homology assessment. *Cladistics* **12:** 265–272.

Cunningham, C. W., Zhu, H., and Hillis, D. M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: Best empirical tests with known phylogenies. *Evolution* **54:** 978–987.

De Pinna, M. C. C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7:** 367–394.

DeSalle, R., Wray, C., and Absher, R. (1994). Computational problems in molecular systematics. *In* "Molecular Ecology and Evolution: Approaches and Applications" (B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, Eds.), pp. 353–370. Birkhäuser, Basel.

Fitch, W. M., and Smith, T. F. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* **80:** 1382–1386.

Gatesy, J., DeSalle, R., and Wheeler, W. C. (1993). Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.* **2:** 152–157.

Giribet, G. (1997). "Filogenia Molecular de Artrópodos Basada en la Secuencia de Genes Ribosomales," PhD dissertation, Universitat de Barcelona, Barcelona.

Giribet, G., Rambla, M., Carranza, S., Baguñà, J., Riutort, M., and Ribera, C. (1999). Phylogeny of the arachnid order Opiliones (Arthropoda) inferred from a combined approach of complete 18S and partial 28S ribosomal DNA sequences and morphology. *Mol. Phylogenet. Evol.* **11:** 296–307.

Kumar, S., Tamura, K., and Nei, M. (1993). "MEGA: Molecular Evolutionary Genetics Analysis. Version 1.0", Pennsylvania State University, University Park, PA.

Li, W. H. (1997). "Molecular Evolution," Sinauer, Sunderland, MA.

Mickevich, M. F., and Farris, J. S. (1981). The implications of congruence in *Menidia*. *Syst. Zool.* **27:** 143–158.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

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

Swofford, D. L. (1998). PAUP* 4.0: Phylogenetic Analysis Using Parsimony (* and Other Methods) ver. 4.0, Sinauer, Sunderland, MA.

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

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114–124.

Titus, T. A., and Frost, D. R. (1996). Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). *Mol. Phylogenet. Evol.* **6:** 49–62.

Waterman, M. S., Eggert, M., and Lander, E. (1992). Parametric sequence comparisons. *Proc. Natl. Acad. Sci. USA* **89:** 6090–6093.

Wheeler, W. C. (1993). The triangle inequality and character analysis. *Mol. Biol. Evol.* **10:** 707–712.

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. (1998). Molecular systematics and arthropods. *In* "Arthropod Fossils and Phylogeny" (G. D. Edgecombe, Ed.), pp. 9–32. Cambridge Univ. Press, New York, NY.

Wheeler, W. C., Gatesy, J., and DeSalle, R. (1995). Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phylogenet. Evol.* **4:** 1–9.

Wheeler, W. C., and Gladstein, D. S. (1994). MALIGN: a multiple sequence alignment program. *J. Hered.* **85:** 417–418.

Wheeler, W. C., and Gladstein, D. (1995). MALIGN ver. 2.7, American Museum of Natural History, New York, NY.

Wheeler, W. C., and Hayashi, C. Y. (1998). The phylogeny of extant chelicerate orders. *Cladistics* **14:** 173–192.

Yang, Z. (1997). Phylogenetic Analysis by Maximum Likelihood (PAML) ver. 1.3, University of California at Berkeley, Berkeley, CA.

Yang, Z., and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14:** 717–724.