

Alignment-Ambiguous Nucleotide Sites and the Exclusion of Systematic Data

JOHN GATESY,* ROB DESALLE,† AND WARD WHEELER*

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

Received December 9, 1992; revised June 23, 1993

MATERIALS AND METHODS

Sequences

Sequence data were collected by standard PCR protocols (DeSalle, 1992; Gatesy and Amato, 1992) using 12S mt rDNA universal vertebrate primers (Kocher *et al.*, 1989; Simon *et al.*, 1991) and 16S mt rDNA insect primers (DeSalle, 1992). 12S mt rDNA fragments from crocodylian taxa (*Caiman crocodilus*, *Melanosuchus niger*, *Paleosuchus palpebrosus*, *P. trigonatus*, *Alligator sinensis*, and *A. mississippiensis*) were combined with four published sequences (*Crocodylus rhombifer*, *Tomistoma schlegelii*, *C. latirostris*, and *Gavialis gangeticus*; Gatesy and Amato, 1992). This data set samples all extant species in the clade Alligatoridae and three outgroup taxa. 16S mt rDNA fragments from *Blaberus craniifer* (Blattaria), *Heptagenia* sp. (Ephemera), *Schistocerca americana* (Orthoptera), *Dorocordula lepida* (Odonata), and *Cerastipsocus venosus* (Psocoptera) were combined with the published *Drosophila yakuba* (Diptera; Clary and Wolstenholme, 1985), *Aedes subpictus* (Diptera; Hsu Chen *et al.*, 1984), *Cicindella dorsalis* (Coleoptera; Vogler *et al.*, 1993), and *Apis mellifera* (Hymenoptera; Vlasek *et al.*, 1987) sequences to form the insect data set.

Alignment

An optimal alignment for a group of nucleotide sequences is determined by the assumptions of the alignment algorithm and by the values that are assigned to critical alignment parameters (gap cost, nucleotide substitution cost, alignment order, etc.). Radically different optimal alignments may be favored if these parameters are changed (Fitch and Smith, 1983; Lake, 1991; Mindell, 1991; Wheeler and Gladstein, 1991). Alternatively, a fixed set of parameters may favor several optimal alignments (Wheeler and Gladstein, 1991).

In this study, MALIGN, the parsimony-based alignment procedure of Wheeler and Gladstein (1991), was used to estimate minimum cost multiple taxon alignments for a variety of alignment parameter values. This program allows the specification of alignment costs for different sequence change events such as

Molecular systematists generally rely on computer algorithms to establish the alignment of DNA sequences. However, when alignment regions are characterized by multiple insertions and deletions, these gap-filled stretches of DNA are often excised before phylogenetic reconstruction. This exclusion of systematic data is generally determined by subjective criteria. We explore a replicable methodology in which the comparison of several multiple sequence alignments can be used to eliminate regions of unstable sequence alignment. Using crocodylian and insect mitochondrial (mt) ribosomal (r) DNA as examples, we caution against the removal of sequence data prior to phylogenetic reconstruction. © 1993 Academic Press, Inc.

INTRODUCTION

Molecular systematists commonly delete blocks of multiple sequence alignments that are deemed too divergent to be phylogenetically useful (e.g., Berbee and Taylor, 1992; Bowman *et al.*, 1992; Turbeville *et al.*, 1991). A common justification for this exclusion of data is that "positions exhibiting high variability or length variation could not be reliably aligned" (Turbeville *et al.*, 1992). However, in most cases, subjective criteria are used to distinguish regions of excessive variation from regions of acceptable variation.

What constitutes a reliable alignment is not readily apparent. In principle, any series of DNA sequences, no matter how divergent, can be aligned. However, the nature of the alignment is dictated by the values assigned to different alignment parameters and by the assumptions of the alignment algorithm.

We explore a replicable data exclusion method in which nucleotide positions whose states vary depending on alignment are discarded. Through the comparison of several minimum cost multiple sequence alignments, stable alignment positions can be identified and retained for phylogenetic analysis. Using the mt rDNA of crocodylians and insects as examples, the implications of this method and data exclusion in general are considered.

insertions/deletions and nucleotide substitutions and builds alignments through multiple pairwise operations arranged in a tree topology fashion. By evaluating many possible alignment order topologies, MALIGN chooses those alignments that yield the shortest cladograms.

Each data set was aligned 15 times with gap weight increasing from two-thirds to 300 (2/3, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, and 300) while nucleotide substitution cost was kept constant at one (transition and transversion costs were equal). Each single-base insertion/deletion was treated as an independent evolutionary event. Other MALIGN options were: score 4, build, alignswap, treeswap, keepaligns 20, keeptrees 100, and iter.

Removal of Data

We suspect that molecular systematists choose to remove data because they have visualized many equally viable sequence alignments, with a computer algorithm or by eye, and are unable to recognize one alignment as significantly better than another. When there are several alignments for a data set, nucleotide positions can be classified as either alignment-ambiguous or alignment-invariant. Alignment-invariant nucleotide positions are those that are constant across all alignments for all taxa (Fig. 1). Alignment-ambiguous positions do not remain constant for all taxa when different alignments are compared (Fig. 1).

In this study, alignment-ambiguous positions were

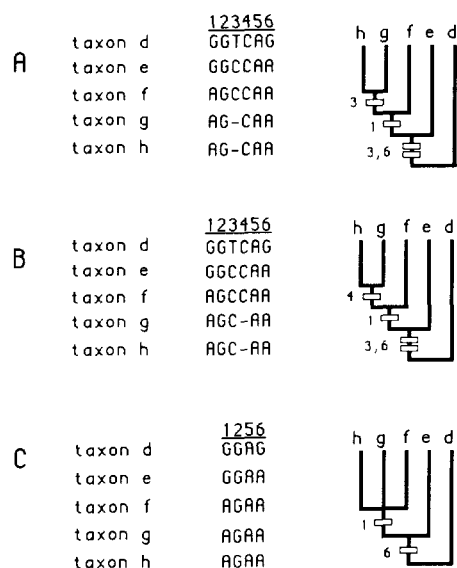


FIG. 1. Two equally costly alignments, A and B, and alignment-invariant sites, C, for hypothetical taxa (d-h). Nucleotide positions are labeled 1-6 in each alignment. Cladograms derived from each alignment are shown to the right with character changes numbered by position. The alignment-ambiguous sites, 3 and 4, support the clade g+h in the initial alignments. Support for g+h is lost by excluding the alignment-ambiguous positions.

determined by the comparison of optimal alignments for gap to substitution cost ratios ranging from 2/3 to 300. The removal of these alignment-ambiguous positions leaves sites that are alignment-invariant across a broad range of gap to substitution cost ratios (e.g., Fig. 2).

The above procedure is analogous to that described by Lake (1991), where ambiguous nucleotide positions were eliminated through the comparison of suboptimal and optimal alignment orders for a single set of alignment parameters. However, we eliminated alignment-ambiguous sites through the comparison of optimal alignment orders for each of several different sets of alignment parameters.

Examination of all gap to substitution cost ratios is an impossible task, but there are minimum and maximum limits to these ratios. At some point all data sets "asymptote." That is, as gap to substitution cost ratio is increased during alignment, a point is reached where a further increase of this ratio does not alter the favored alignment or alignments. For the crocodile 12S rDNA data, alignments at gap to substitution cost ratios of 100:1 and 300:1 are identical, suggesting that alignment stabilizes somewhere between ratios of 50:1 and 100:1. The insect data set apparently asymptotes between ratios of 20:1 and 50:1.

At the other end of the spectrum, the minimum gap cost must be greater than one-half of the substitution cost. This follows from the triangle inequality (Wheeler, 1993). If the gap cost is set less than 1/2, the transformation matrix is nonsensical. For example, if the gap to substitution cost ratio were 1/3:1, a change from an adenine to a gap to a guanine would be less costly than a direct change from an adenine to a guanine.

Even given these limits, many optimal alignments may not be discovered in a reasonable amount of time. Nevertheless, with each new set of parameters that is tested, more ambiguous characters can be eliminated. Ultimately, the assumptions of the alignment program and the range of parameters that are explored will determine the number of nucleotide positions that are considered alignment-ambiguous.

Phylogenetic Analysis

Each initial alignment and each set of alignment-invariant sites was analyzed individually. All nucleotide substitutions were weighted equally, and individual insertions and deletions were weighted proportional to their cost assigned during alignment (from 2/3:1 to 300:1). The branch and bound algorithm in PAUP (Swofford, 1990) was used to find minimum length cladograms. Crocodile cladograms were rooted with non-alligatorid crocodylians (*T. schlegelii*, *G. gangeticus*, and *C. rhombifer*), and insect cladograms were rooted with the mayfly (*Heptagenia* sp).

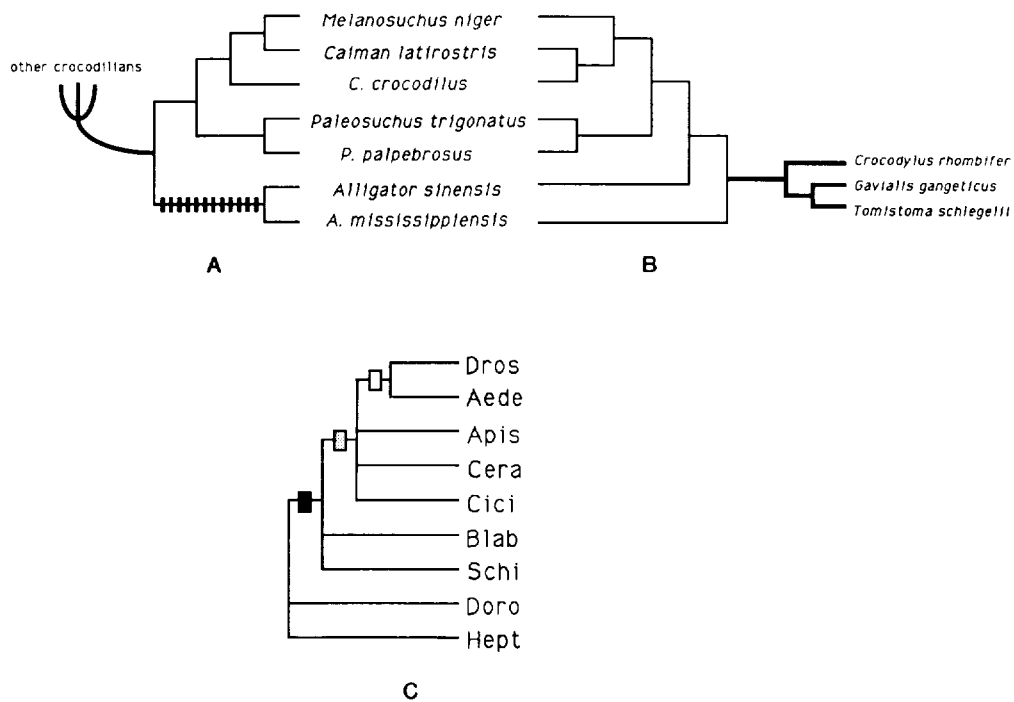


FIG. 3. The independent corroborated phylogeny for alligatorids, A, and the hypothesis favored by the majority of 12S rDNA sequence alignments, B. Twelve unambiguous morphological synapomorphies (Norell, 1988) support the monophyly of *Alligator* in A. (C) Insect clades supported by previous studies: white bar, Diptera; gray bar, Eumetabola; black bar, Neoptera. (Dros, *Drosophila yakuba*; Aede, *Aedes subpictus*; Blab, *Blaberus cranifer*; Schi, *Schistocerca americana*; Cici, *Cicindella dorsalis*; Hept, *Heptagenia* sp., Doro, *Dorocordula lepida*; Cera, *Cerastipsocus venosus*; Apis, *Apis mellifera*).

metabola, and Diptera) are supported by both molecular and morphological characters and will be used to judge our results. All cladograms derived from the initial 15 alignments retain Neoptera, the majority retrieve a monophyletic Diptera, and one of the cladograms favors Eumetabola monophyly (Fig. 4B). Cladogram topology changes with only minor alterations in alignment parameters.

Admittedly, the 16S mt rDNA segment is probably inappropriate for this level of analysis. All alignments were riddled with gaps, but we hoped that reliable phylogenetic information could be extracted from this data set if confounding alignment ambiguity was excluded. Unfortunately, when the data set is trimmed using the criteria established here, only 12 of approximately 250 sites are alignment-invariant, no positions characterized by gaps were alignment-invariant, and only two phylogenetically informative characters remain. The cladogram that is based on the alignment-invariant sites is contrary to prior evidence and not well resolved (Fig. 5B).

Data Exclusion

The exclusion of sequence data before phylogenetic reconstruction can be seen as an extreme form of character weighting. That is, regions of aligned sequences that are thought to be hopelessly scrambled by multiple nucleotide substitutions and insertions/deletions

are given a weight of zero in phylogenetic reconstruction. Such harsh character weights should not be determined by the whim of each individual investigator.

If DNA sequence data must be removed before testing for character congruence, positions should be eliminated according to a repeatable objective protocol. The method presented here is based on the following logic: nucleotide positions that do not align consistently over a variety of alignment parameters are seen as unreliable positional homologies relative to sites that are alignment-invariant. This argumentation takes Patterson's original (1982) concept of similarity, as the initial criterion for homology, to an extreme. Alignment-ambiguous sites are removed from the pool of potential homologies (synapomorphies).

The range of alignment parameters explored in this paper may seem unreasonable (e.g., gap to substitution cost ratios of 2/3:1 or 300:1) especially considering the proportion of alignment-ambiguous positions that were identified. A more limited range of parameters could be used to score fewer sites as alignment-ambiguous. However, the problem then becomes how to justify this range of parameters.

No matter how many alignment parameters are tested, the exclusion of alignment-ambiguous sites does not guarantee the elimination of all bad data and the retention of all good data. This is not surprising, given that a group of alignment-invariant sequences

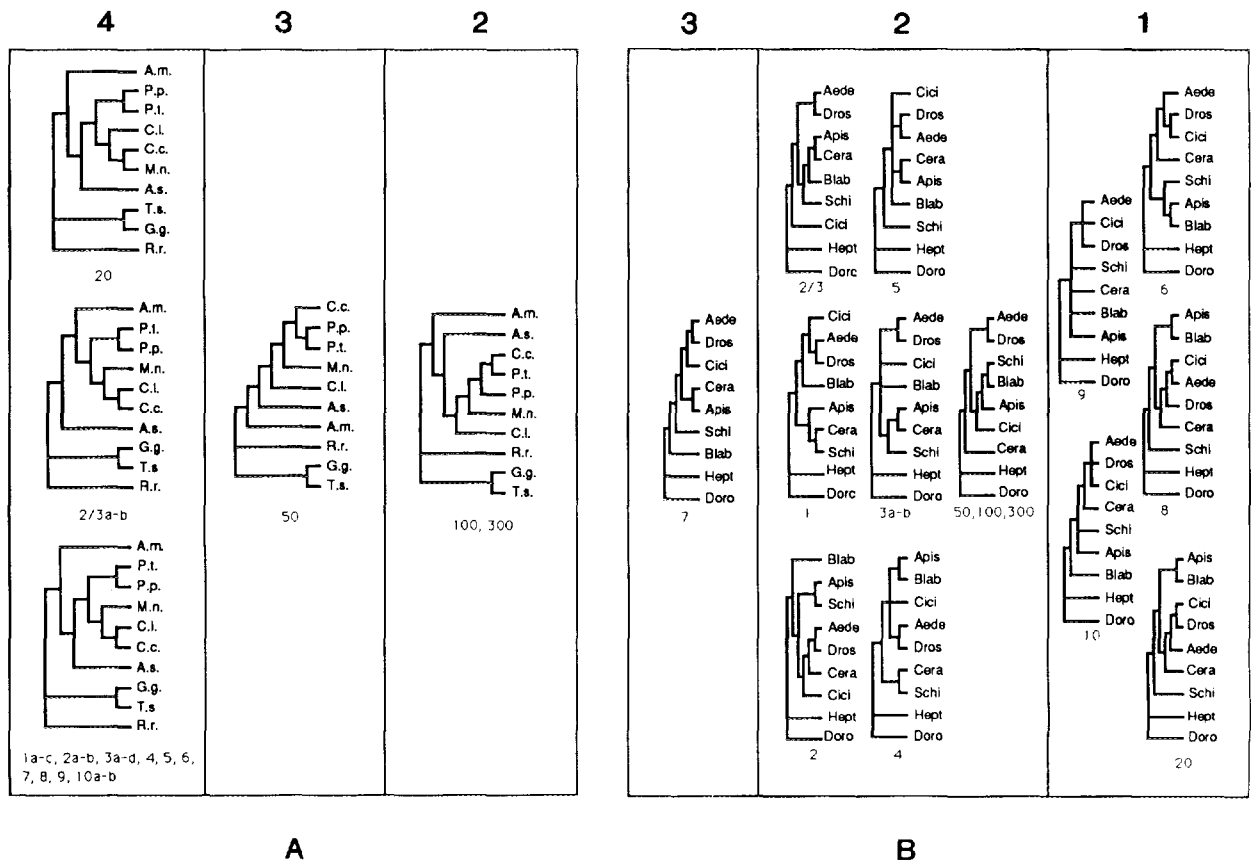


FIG. 4. Cladograms derived from initial alignments of the crocodile 12S mt rDNA data, A, and the insect 16S mt rDNA data, B. Gap to substitution cost ratios of alignments are shown beneath each minimum length cladogram. Letters that follow gap to substitution cost ratios represent multiple equally costly alignments for a given ratio. Numbers above the columns of trees indicate the number of clades that are strictly congruent with the corroborated hypotheses of Fig. 3A and C. (R.r., *Crocodylus rhombifer*; M.n., *Melanosuchus niger*; C.c., *Caiman crocodilus*; C.l., *Caiman latirostris*; P.t., *Paleosuchus trigonatus*; P.p., *Paleosuchus palpebrosus*; A.m., *Alligator mississippiensis*; A.s., *Alligator sinensis*; G.g., *Gavialis gangeticus*; T.s., *Tomistoma schlegelii*). Abbreviations for insect taxa as in Fig. 3.

may be well stricken in multiple substitution events and that insertions/deletions may be unambiguous phylogenetic indicators in spite of alignment ambiguity (Fig. 1).

The removal of problematic alignment regions has had wideranging influence. Much of our knowledge of deep branches in the tree of life is based on rDNA

sequences that have been trimmed prior to phylogenetic analysis (Elwood *et al.*, 1985; Gouy and Li, 1989; Gunderson *et al.*, 1987; Hedges *et al.*, 1990; Lake, 1987; Olsen, 1987; Perasso *et al.*, 1989). These basic biological subdivisions as well as phylogenetic relationships at lower taxonomic levels may be radically rearranged by the inclusion or exclusion of various blocks of data. Therefore, researchers should define and strictly justify their personal protocol for the exclusion of data. The method presented here is objective and replicable, but often entails an undesirable loss of information. In order to assess the effects of this and other data exclusion methods, Kluge's (1989) notion of "total evidence" should be extended to the use of scrambled alignment regions in phylogenetic reconstruction.

ACKNOWLEDGMENTS

We thank C. Hayashi, A. Williams, and D. Yeates for reading this manuscript, B. Dashevge for laboratory help, G. Amato, C. Remington, and H. Keshishian for tissue samples, and D. Mindell for his comments.

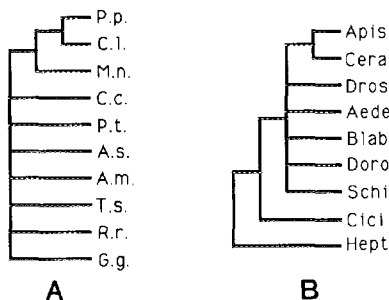


FIG. 5. Trees for the two sets of alignment-invariant sites: (A) Strict consensus tree for the crocodile data derived from eight equally parsimonious cladograms. (B) The most parsimonious cladogram for the insect data.

REFERENCES

- Berbee, M. L., and Taylor, J. W. (1992). Two Ascomycete classes based on fruiting-body characters and ribosomal DNA sequences. *Mol. Biol. Evol.* **9**: 278–284.
- Boudreaux, H. B. (1979). "Arthropod Phylogeny with Special Reference to Insects." Wiley, New York.
- Bowman, B. H., Taylor, J. W., Brownlee, A. G., Lee, J., Lu, S. D., and White, T. J. (1992). Molecular evolution of the fungi: Relationship of the Basidiomycetes, Ascomycetes and Chytridiomycetes. *Mol. Biol. Evol.* **9**: 285–296.
- Clary, D., and Wolstenholme, D. (1985). The mitochondrial DNA molecule of *Drosophila yakuba* Nucleotide sequence, gene organization and genetic code. *J. Mol. Evol.* **22**: 252–271.
- Densmore, L. (1983). Biochemical and immunological systematics of the order Crocodylia. In "Evolutionary Biology" (M. Hecht, B. Wallace, and G. Prance, Eds.) Vol. 16, pp. 397–465, Plenum, New York.
- DeSalle, R. (1992). The phylogenetic relationships of flies in the family Drosophilidae deduced from mtDNA sequences. *Mol. Phylogenet. Evol.* **1**: 1–11.
- Elwood, H., Olsen, G., and Sogin, M. (1985). The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata*. *Mol. Biol. Evol.* **2**(5): 399–410.
- Fitch, W., and Smith, T. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* **80**: 1382–1386.
- Gatesy, J., and Amato, G. (1992). Sequence similarity of 12S ribosomal segment of mitochondrial DNAs of gharial and false gharial. *Copeia* **1**: 241–243.
- Gouy, M., and Li, W.-H. (1989). Phylogenetic analysis based on rRNA sequences supports the archaeobacterial rather than the eocyte tree. *Nature* **339**: 145–147.
- Gunderson, J., Elwood, H., Ingold, A., Kindle, K., and Sogin, M. (1987). Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. *Proc. Natl. Acad. Sci. USA* **84**: 5823–5827.
- Hedges, S., Moberg, K., and Maxson, L. (1990). Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. *Mol. Biol. Evol.* **7**(6): 607–633.
- Hsu Chen, C. C., Koten, R. M., and Dubin, D. T. (1984). Sequences of the coding and flanking regions of the large ribosomal subunit RNA gene of the mosquito mitochondria. *Nucleic Acids Res.* **12**: 7771–7785.
- Kluge, A. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38**: 2–25.
- Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Paabo, S., Villablanca, F. X., and Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. USA* **86**: 6196–6200.
- Kristensen, N. P. (1975). The phylogeny of hexapod 'orders': A critical review of recent accounts. *Z. Zool. Evol. Forsch.* **13**: 1–44.
- Kristensen, N. P. (1981). Phylogeny of insect orders. *Annu. Rev. Entomol.* **26**: 135–157.
- Lake, J. (1987). Prokaryotes and archaeobacteria are not monophyletic: Rate invariant analysis of rRNA genes indicates that eukaryotes and eocytes form a monophyletic taxon. *Cold Spring Harbor Symp. Quant. Biol.* **LII**: 839–846.
- Lake, J. (1991). The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.* **8**(3): 378–385.
- Mindell, D. P. (1991). Aligning DNA sequences: Homology and phylogenetic weighting. In "Phylogenetic Analysis of DNA Sequences" (M. Miyamoto and J. Cracraft, Eds.), pp. 73–89, Oxford University Press, Oxford.
- Norell, M. (1988). "Cladistic Approaches to Evolution and Paleobiology as Applied to the Phylogeny of Alligatorids." PhD. thesis, Yale University, New Haven.
- Olsen, G. (1987). Earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp. Quant. Biol.* **LII**: 825–837.
- Patterson, C. (1982). Morphological characters and homology. In "Prospects in Systematics" (K. Joysey and A. Friday, Eds.), pp. 21–74, Academic Press, London.
- Perasso, R., Baroin, A., Liang, H., Bachelierie, J., and Adoutte, A. (1989). Origin of the algae. *Nature* **339**: 142–144.
- Simon, C., Franke, A., and Martin, A. (1991). The polymerase chain reaction: DNA extraction and amplification. In "Molecular Taxonomy NATO Advanced Studies Institute" (G. M. Hewitt, A. Johnston, J. Young, Eds.), pp. 329–355, Springer Verlag, Berlin.
- Swofford, D. L. (1990). "Phylogenetic Analysis Using Parsimony: Program and Documentation," Ill. Nat. Hist. Surv., Champaign-Urbana.
- Turbeville, J. M., Field, K. G., and Raff, R. A. (1992). Phylogenetic position of the Nemertini, inferred from 18S rRNA sequences: Molecular data as a test of morphological character analysis. *Mol. Biol. Evol.* **9**: 235–249.
- Turbeville, J. M., Pfeifer, D. M., Field, K. G., and Raff, R. A. (1991). The phylogenetic status of arthropods as inferred from 18S rRNA sequences. *Mol. Biol. Evol.* **8**: 669–686.
- Vlasek, I., Bungschaiger, S., and Krell, G. (1987). Honey bee mitochondrial large ribosomal RNA. *Nucleic Acids Res.* **15**: 2388–2388.
- Vogler, A. P., DeSalle, R., Assmann, T., Knisely, C. B., and Schultz, T. D. (1993). Molecular population genetics of the endangered tiger beetle, *Cicindela dorsalis* Say (Coleoptera: Cicindelidae). *Ann. Entomol. Soc. Am.* **86**: 142–152.
- Wheeler, W. (1989). The systematics of insect ribosomal DNA. In "The Hierarchy of Life" (B. Fernholm, K. Bremer, and H. Jornvall, Eds.), pp. 307–321, Elsevier Science Publishers.
- Wheeler, W., and Gladstein, D. (1991). "MALIGN: Program and Documentation-version 1.5," American Museum of Natural History, New York.
- Wheeler, W. (1993). The triangle inequality and character analysis. *Mol. Biol. Evol.* **10**(3): 707–712.