13

HOMOLOGY AND DNA SEQUENCE DATA

WARD WHEELER

Department of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024

INTRODUCTION

The phylogenetic analysis of DNA sequences, like that of all other comparative data, is based on schemes of putative homology which are then tested via congruence to determine synapomorphy schemes and cladistic relationships. Unlike some other data types, however, the matrix of putative homologies or "characters" is not directly observable. When sequences are unequal in length, the correspondences among sequence positions are not preestablished and some sort of procedure is required to determine which positions are "homologous." This is the traditional province of multiple sequence alignment (= alignment here). Alignment generates a collection of column vectors through the insertion of gaps, which form the character set. Whether accomplished manually, or via some computational algorithm, these characters are then submitted to phylogenetic analysis in the same manner as other forms of data. This scheme of correspondences or putative homologies has two salient features. First, alignment precedes the phylogenetic analysis (i.e., cladogram search) and is

The Character Concept in Evolutionary Biology

never revised in light of systematic hypotheses. Second, alignment-based homology schemes rest on a notion of base-to-base homology where individual nucleotide bases transform among five states (A, C, G, T/U, and gap) within a single character. Two methods have recently been proposed ("Optimization-Alignment," Wheeler, 1996 and "Fixed-State Optimization," Wheeler, 1999) which avoid multiple alignment altogether and question these two tenets of sequence analysis. Although these approaches are parsimony methods, and rely on testing homology through synapomorphy, they differ in the entities they propose for testing and this has implications for the interpretation of DNA sequence homology.

In discussing these concepts, a shorthand will be used. To describe those correspondences among states frequently referred to as putative homologies, the lowercase "homology" will be used. To describe those correspondences that have been tested through congruence on a cladogram (i.e., synapomorphy), the uppercase "Homology" will be used. The discussion here is mainly concerned with methods of deriving homology statements, but all of these would then be tested with other data to determine which homologies are Homologies.

STATIC VERSUS DYNAMIC HOMOLOGY

The standard precursor to the phylogenetic analysis of DNA sequences is alignment. This procedure takes the unequal length strings of nucleotide bases and inserts place-holding gaps ("-") to make the corresponding (homologous) bases line up into intelligible columns. These columns (characters) comprise the data used to reconstruct cladograms. However this alignment is created, once phylogenetic analysis has begun, it will not be revised. That is, the homologies explicitly defined in the alignment will not be reexamined during the cladogram-search process.

Consider four sequences: I GGGG, II GGG, III GAAG, and IV GAA. An alignment can be generated to be supplied to standard phylogenetic analysis. In this case, insertion-deletion events are given a cost of two and base substitutions one. The most parsimonious (minimum cost) cladogram relating these four taxa would be that which holds I and III to be sister taxa with an overall length of six (1 indel and 4 base changes—Fig. 1.) Given this alignment, the two other phylogenetic scenarios are less favored (7 and 8 steps). There is another alignment, however, which generates the same minimum length for topology (((I III) II) IV) yet yields the same length (6 steps) for one of the other two possible topologies (Fig. 2). Using this alignment, two of the topologies are equally parsimonious.

		Topology		
Alignment		((I II) III) IV)	(((I III) II) IV)	(((I IV) II) III)
I	GGGG	7	6	8
II	–GGG			
III	GAAG			
IV	--GAA			

Insertion-Deletion events cost 2
Base changes cost 1

FIGURE 1 Possible alignment for four simple sequences and the cladogram cost (length) for the possible topologies for these taxa.

The point of this example is that the alignment process yields static homology schemes which is not optimized for any particular topology. Once the alignment is determined, all testing of the alignment itself stops. Although homologies are tested on each cladogram, there may be no single homology matrix which optimizes Homology (yields the most parsimonious result) for each cladogram. In order to give each topology its shortest length, homologies need to be generated which are optimal for that particular topology. It is this need which motivates the method of "Optimization-alignment" (Wheeler, 1996).

		Topology		
Alignment		((I II) III) IV)	(((I III) II) IV)	(((I IV) II) III)
I	GGGG	7	6	8
II	–GGG			
III	GAAG			
IV	--GAA			
I	GGGG	6	6	8
II	GGG--			
III	GAAG			
IV	GAA–			

Insertion-Deletion events cost 2
Base changes cost 1

FIGURE 2 Comparison of the implications of two different alignments on the cladogram costs for the sequences in Fig. 1.

	Seq. 1	Seq. 2	Seq. 3
Seq. 1	0	C_{21}	C_{31}
Seq. 2	C_{12}	0	C_{32}
Seq. 3	C_{13}	C_{23}	0

$C_{ij} = C_{ji}$

FIGURE 5 Matrix of minimum transformation cost between sequence pairs.

When employing blocks of contiguous sequence as characters, with observed sequences as states, dynamic programming methods must be used to optimize cladograms and determine their length. The procedure is identical to the optimization of Sankoff-style characters ("step-matrix" characters), just modified for large numbers of states ($n_{states} \leq n_{taxa}$). This approach relies on the postulate that only observed sequences may be optimized to hypothetical ancestors. This restricts the possible world of reconstructed sequences, but also requires that these sequences exist. Each sequence becomes a state in an extremely complex character. The first step in the optimization procedure is the determination of the transformation cost matrix among all the states. This is defined as the minimum transformation cost (including all forms of base substitutions and insertion-deletion costs) between each pair of states (Fig. 5). Once these transformation costs are known, standard dynamic programming implicitly examines the assignment of each of these states to each internal node and determines the optimal set of states and cladogram length (Fig. 6).

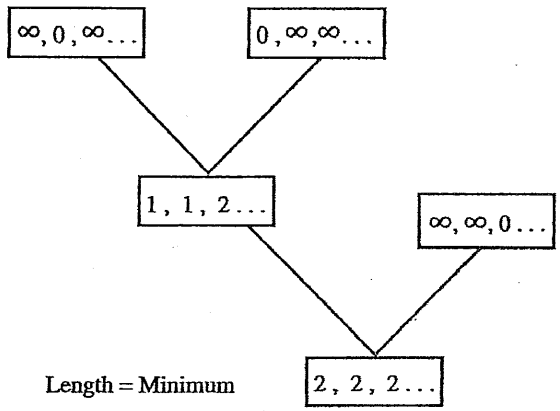

FIGURE 6 An example of down-pass cladogram optimization via the fixed-state approach.

Given that the cladogram length is based on nucleotide sequence, it might seem strange to say that the bases themselves are not homologous. This effect is derived from the pairwise nature of the character transformation matrix. Consider three sequences I AAATTT, II TTT, and III AAA. When transformation costs are determined, the first "T" in sequence II (position 1) corresponds with the first "T" of sequence I (position 4). This same "T" in sequence I also corresponds with the first "A" of sequence III (position 1). If our logic were transitive, this would imply that position 4 of sequence I would correspond to position 1 of sequence III. It does not. Position 1 of sequence III ("A") corresponds to position 4 of sequence I (also "A"). No circle of correspondence can be drawn among these nucleotides describing state transformations. They are not homologous (Fig. 7).

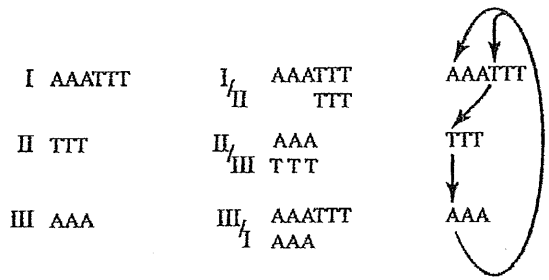

FIGURE 7 Scheme of base correspondences implied by the fixed-state approach.

Two other aspects of this approach affect ideas of homology. One of the salient features of base-to-base methods, whether built upon static or dynamic homologies, is difficulty in tracing complex homologies through the cladogram (or alignment)—in other words, messy data. When there is extensive sequence length variation coupled with base changes, tremendous uncertainty in homology can occur in both multiple-alignment and optimization alignment. The requirement that such variation be accommodated over the entire cladogram can make local uncertainties propagate throughout the analysis. Since the sequence level homology approach transforms the complex states with their variations in length and nucleotide base composition into simple numbered states with pairwise costs, this problem does not occur. Such seemingly confusing variation patterns will certainly lead to longer cladograms, but the homologies (at the fragment level) will remain clear.

A second feature of fragment level homology is the requirement that the character homologies be defined *a priori*. Whether entire loci, structurally or functionally defined regions are employed as homologies, they are determined by the investigator. This is akin to the delimitation of variation in complex morphological features. Are complex structures such as complete development in the endopterygote insects single or multiple characters? As with all such seemingly arbitrary decisions, what matters most is the effect of changing these character delimitations on phylogenetic results.

The notion of synapomorphy as a shared derived feature might also seem to be altered by the homology concept implicit in sequence fragment comparisons. Since each taxon may well express a unique character state, it might appear that synapomorphy (as a shared state) would be impossible. This criticism would only apply if the characters were completely unordered. State transformation costs are not equal among states, hence are more akin to synapomorphy in the context of ordered characters. Two taxa might present states 1 and 2 of an ordered series $0 \rightarrow 1 \rightarrow 2$. These taxa are united by the transformation implied by the ordering with 1 and 2 sharing special derived similarity not found in 0 (Platnick, 1979). The concept of synapomorphy (or Homology) is unaffected by the fixed-state approach.

COMPARISONS

For these distinctions (static versus dynamic; base-to-base versus fragment) to be anything more than nomenclature, some means of comparing these methods and judging superiority must be offered. Congruence could be that measure. When analyzing single data sets, via whatever method, the best solution is that which minimized discord among data (i.e., characters). This may be measured by simplicity (parsimony) or with respect to complex statistical models (likelihood). The three methods of viewing homology here define characters in somewhat different ways, hence simple counting of change for single data sets (i.e.,

cladogram length) cannot be used. The things that are counted are just not the same. This notion of character congruence, however, can be extended to the broader concept of congruence among data sets. Character congruence has been used to discriminate among analysis parameters (Wheeler, 1995; Whiting *et al.*, 1997; Wheeler and Hayashi, 1998) and could reasonably be used to compare the behavior of methods (although numerous other means could also be employed).

Two types of congruence measures can be used: character based and topological. The relative merits and demerits of these approaches have been explored in the literature (Mickevich and Farris, 1981; Wheeler, 1995) and character congruence will be used here due to its link with parsimony and combined data analysis. Phylogenetic methods are judged to be superior if they accommodate variation in multiple data sets efficiently as measured by the Mickevich-Farris incongruence length metric (Mickevich and Farris, 1981).

EXAMPLE—ARTHROPODS

In order to compare these three homology-determination methods, the arthropod data of Wheeler *et al.* (1993) are used. These data consist of 100 morphological characters, ~650 18S rDNA nucleotides, and 228 Ubiquitin nucleotides. To these data ~350 28S rDNA nucleotides were added. The 18S and Ubiquitin data were determined for 25 extant taxa and the morphological data scored for these taxa and "Trilobita," an extinct clade. The 28S rDNA data were determined for 15 of the extant taxa (Table I).

TABLE I Taxon List

Mollusca		
	Cephalopoda	*Loligo pealei*
	Polyplacophora	*Lepidochiton cavernae*
Annelida		
	Polycheata	*Glycera sp.*
	Oligocheata	*Lumbricus terrestris*
	Hirudinea	*Haemopis marmorata*
Onychophora		
	Peripatoidae	*Peripatus trinitatis*
	Peripatopsidae	*Peripatoides novozealandia*
Trilobita		groundplan of Ramsköld and Edgecombe, 1991. (morphological analysis only)
Chelicerata		
	Pycnogonida	*Anoplodactylus portus*
	Xiphosura	*Limulus polyphemus*
	Scorpiones	*Centruroides hentzii*
	Uropygi	*Mastogoproctus giganteus*
	Araneae	*Nephila clavipes*
	Araneae	*Peucetia viridans*
Crustacea		

	Cirrepedia	*Balanus sp.*
	Malacostraca	*Callinectes sp.*
Myriapoda		
	Chilopoda	*Scutigera coleoptrata*
	Diplopoda	*Spirobolus sp.*
Hexapoda		
	Zygentoma	*Thermobius sp.*
	Ephemerida	*Heptagenia sp.*
	Odonata	*Libellula pulchella*
	Odonata	*Dorocordulia lepida*
	Dictyoptera	*Mantis religiosa*
	Auchenorrhyncha	*Tibicen sp.*
	Lepidoptera	*Papilio sp.*
	Diptera	*Drosophila melanogaster*

Three analyses were performed. In each case, the insertion-deletion cost was set at two and all base substitutions set at one. When morphological characters were used, character transformations were set at two. In the first analysis, the data were aligned (via MALIGN; Wheeler and Gladstein, 1994) and phylogenetic analysis was performed using PHAST (Goloboff, 1996). The second analysis employed optimization-alignment as implemented in POY (Gladstein and Wheeler, 1996). The third used the fixed-state optimization technique also as implemented in POY. Gaps/indels were included and given the same weight (2) in all length calculations. All searches employed TBR branch swapping and 10 random addition sequences. The results of the individual data partitions, combined results, and congruence calculations are summarized in Table II and Figs. 8-10.

TABLE II Comparison of Methodologies

Data	Alignment	Method Optimization-alignment	Fixed-state
18S rDNA	503	501	584
Ubiquitin	387[1]	392	484
28S rDNA	919	848	943
Morphology	252[2]	252	252
Combined	2123	2007	2271
Incongruence[3]	0.0292	0.00698	0.00352

[1] This length of 387 steps is shorter than that of the optimization-alignment purely due to the treatment of ambiguities. When all ambiguities are treated as missing data, both alignment (MALIGN-PHAST) and optimization-alignment (POY) yield the same length of 387 steps.
[2] This length is 2 times the length of 126 steps.
[3] Calculated as (Combined – 18S rDNA – Ubiquitin – 28S rDNA – Morphology)/Combined.

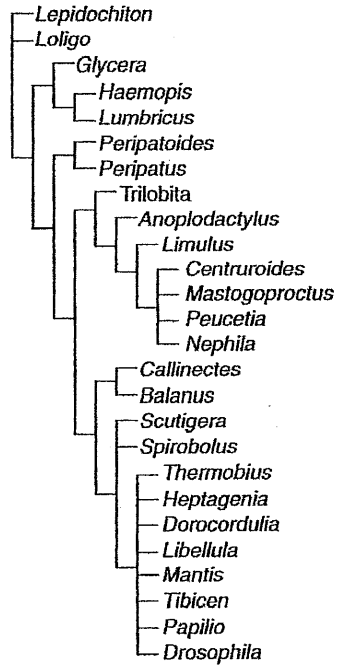

FIGURE 8 Morphologically based cladogram of arthropod relationships (Wheeler *et al.*, 1993).

The dynamic homology approach of optimization-alignment resulted in more parsimonious cladograms in all the cases where sequences were unequal in length. This is due, no doubt, to the simultaneous optimization of synapomorphy and homology uniquely for each topology. The cladograms derived from the fixed-state approach were the longest. The restriction on the possible range of internal node (HTU) sequences is responsible for this. Since internal node sequences are chosen from the range of observed terminal sequences, longer cladograms frequently arise (Wheeler, 1999). Overall character incongruence was lowest (0.00352 vs. 0.00698 and 0.0292) for the fixed-state analysis.

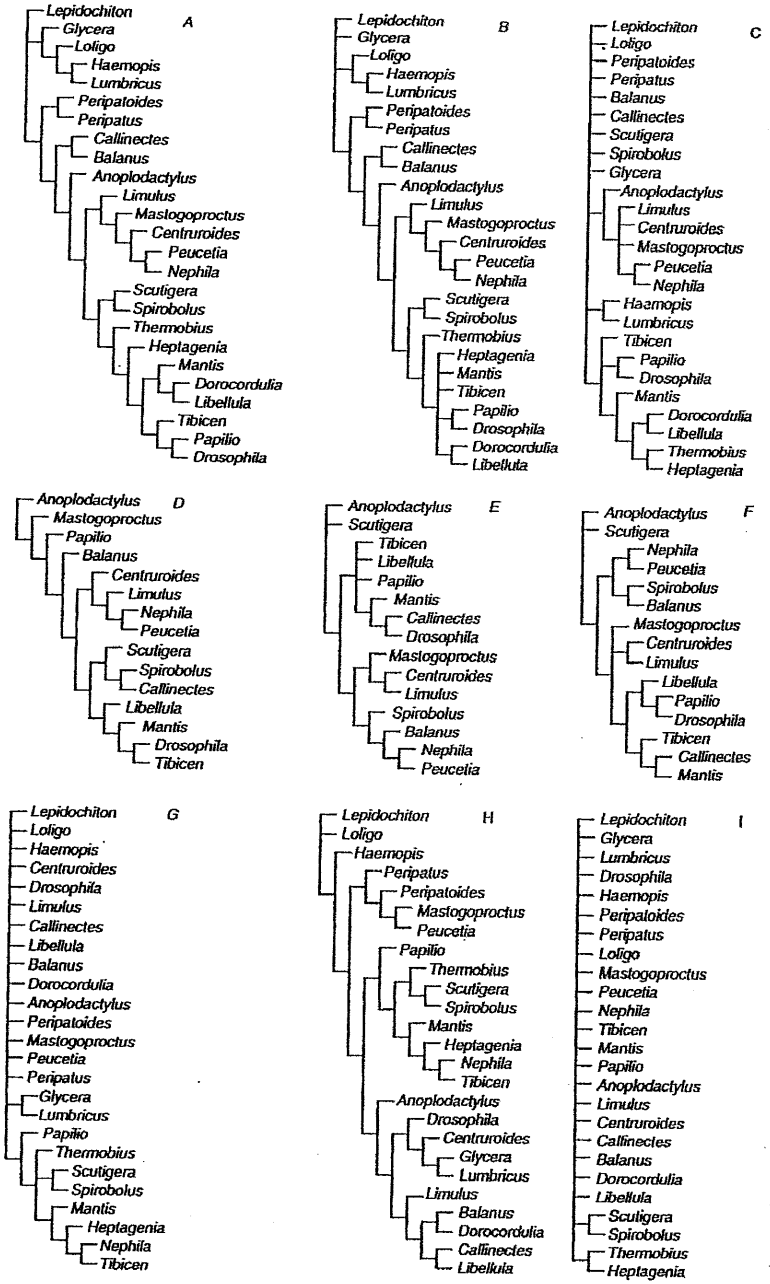

FIGURE 9 Cladograms of individual data partitions when subjected to different analytical techniques. A. 18S rDNA and multiple sequence alignment. B. 18S rDNA and optimization-alignment. C. 18S rDNA and fixed-state optimization. D. 28S rDNA and multiple sequence alignment. E. 28S rDNA and optimization-alignment. F. 28S rDNA and fixed-state optimization. G. Ubiquitint and multiple sequence alignment. H. Ubiquitin and optimization-alignment. I. Ubiquitin and fixed-state optimization.

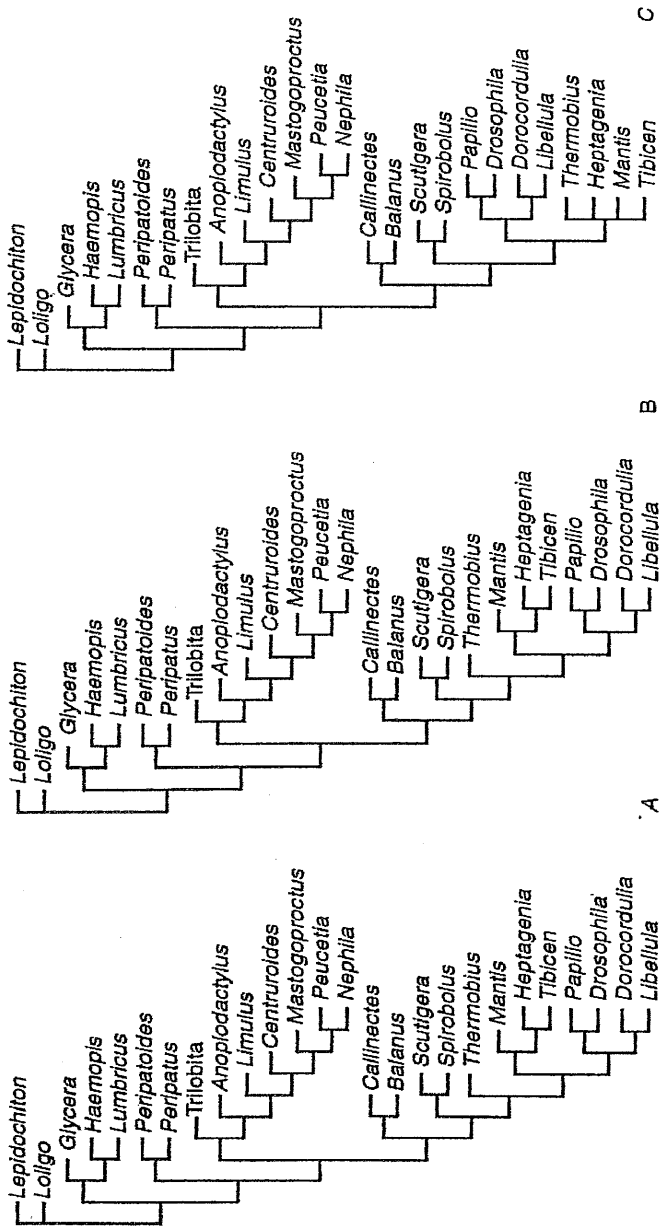

FIGURE 10 Cladograms of combined data (18S rDNA, Ubiquitin, 28S rDNA, morphology) for arthropod taxa when subjected to different analytical techniques. A. Multiple sequence alignment. B. Optimization-alignment. C. Fixed-state optimization.

DISCUSSION

Clearly, the way we view sequence homology has tremendous implications for the elucidation of phylogenetic pattern. The three modes discussed here (static, dynamic, and fragment level) imply different patterns of relationship in the small test case used here. Furthermore, since the reconstructions of hypothetical ancestral sequences vary with the method, the types of evolutionary events reconstructed on these patterns differ as well.

An additional feature of the base-to-base methods, the nonindependence of the nucleotide characters, remains largely unexplored. In both alignment and optimization-alignment, the homology scheme for each nucleotide is determined in concert with all the other bases that surround it. The relative position and number of indels and nucleotide substitutions in adjacent sequence positions fundamentally affect positional homology. Clearly, such character statements are not independent. However, when cladograms are constructed, the cost of changes and indels are summed linearly over the data—an assumption of rigid character independence. Since the individual bases play no role in homology with the fixed-state approach, this character dependence problem vanishes. The indels and base substitutions only determine the cost of transformation between states, there is no requirement that these changes be independent. This inconsistency of base-to-base homology is avoided.

With character incongruence levels at half or lower levels than the other techniques and character nonindependence removed, the fixed-state approach to sequence homology is clearly worth consideration.

ACKNOWLEDGMENTS

I would like to acknowledge the contributions of Daniel Janies, Gonzalo Giribet, Norman Platnick, Lorenzo Prendini, Randall Schuh, Susanne Schulmeister, and Mark Williams to this work through discussion and abuse. I would also like to thank Portia Rollins for expert art work.

LITERATURE CITED

Gladstein, D. S., and Wheeler W. C. (1997). "POY: The Optimization of Alignment Characters. Program and Documentation. New York, NY. Available at "ftp.amnh.org" /pub/molecular.

Goloboff, P. (1996). PHAST. Program and Documentation. Version 1.5.

Mickevich, M. F., and Farris, S. J. (1981). The implications of congruence in *Menidia*. *Syst. Zool.* **30**:351-370.

Platnick, N. I. (1979). Philosophy and the transformation of cladistics. *Syst. Zool.* **28**:537-546.

Ramsköld, L., and Edgecombe, G. D. (1991). Trilobite monophyly revisited. *Hist. Biol.* **4**:267-283.

Wheeler, W. C. (2000). Heuristic reconstruction of hypothetical-ancestral DNA sequeces: sequence alignment versus direct optimization. *In* "Homology and Systematics: Coding Characters for Pylogenetic Analysis" (R. W. Scotland, ed.), pp. 106-113. Taylor and Francis, London.

Wheeler, W. C. (1999). Fixed character states and the optimization of molecular sequence data. *Cladistics* **15**:379-385.

Wheeler, W. C. (1996) Optimization alignment: the end of multiple sequence alignment in phylogentics? *Cladistics* **12**:1-10.

Wheeler, W. C. (1995). Sequence alignment, parameter sensitivity, and the phylogentic analysis of molecular data. *Syst. Biol.* **44**:321-332.

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

Wheeler, W. C., and Gladstein, D. M. (1992-1996). Malign: A Multiple Sequence Alignment Program. Program and Documentation. New York, NY. available ftp.amnh.org /pub/molecular/malign

Wheeler, W. C., and Hayashi, C. Y. (1998). The phylogeny of the chelicerate orders. *Cladistics* **24**:173-192.

Wheeler, W. C., Cartwright, P., and Hayashi, C. (1993). Arthropod phylogenetics: a total evidence approach. *Cladistics* **9**:1-39.

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