Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cladistics 19 (2003) 261–268
Cladistics
www.elsevier.com/locate/yclad

Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search

Ward C. Wheeler*

Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192, USA

Accepted 7 April 2003

Abstract

A method to align sequence data based on parsimonious synapomorphy schemes generated by direct optimization (DO; earlier termed optimization alignment) is proposed. DO directly diagnoses sequence data on cladograms without an intervening multiple-alignment step, thereby creating topology-specific, dynamic homology statements. Hence, no multiple-alignment is required to generate cladograms. Unlike general and globally optimal multiple-alignment procedures, the method described here, implied alignment (IA), takes these dynamic homologies and traces them back through a single cladogram, linking the unaligned sequence positions in the terminal taxa via DO transformation series. These "lines of correspondence" link ancestor–descendent states and, when displayed as linearly arrayed columns without hypothetical ancestors, are largely indistinguishable from standard multiple alignment. Since this method is based on synapomorphy, the treatment of certain classes of insertion–deletion (indel) events may be different from that of other alignment procedures. As with all alignment methods, results are dependent on parameter assumptions such as indel cost and transversion:transition ratios. Such an IA could be used as a basis for phylogenetic search, but this would be questionable since the homologies derived from the implied alignment depend on its natal cladogram and any variance, between DO and IA + Search, due to heuristic approach. The utility of this procedure in heuristic cladogram searches using DO and the improvement of heuristic cladogram cost calculations are discussed.

Multiple-alignment procedures begin with sequence data (usually, but not necessarily of unequal length) and transform them into identical-length character sets via the insertion of gaps ("-"), which signify the insertion or deletion of nucleotides (or amino acids). This is done, in a phylogenetic context, to allow the application of standard analysis tools that require predefined putative homologies (i.e., column vector data). This is not a logical requisite, since character-based methods that directly analyze sequence data as a superset of standard optimization protocols have been proposed (Wheeler, 1996, 1999). Alignments have other phylogenetic uses, and there are many nonphylogenetic applications, such as motif searching, that employ these structures.

Here, a synapomorphy-based alignment procedure—implied alignment (IA)—that is an outgrowth of direct optimization (DO; earlier termed optimization alignment; Wheeler, 1996) is described. Implied alignment takes the topology-specific homologies of DO and extracts and represents them as a multiple alignment. This alignment can then be submitted to standard phylogenetic reconstruction procedures (e.g., PHAST (Goloboff, 1996), PAUP (Swofford, 2001), POY (Wheeler et al., 2002)) and cladograms generated from this. When DO is used to search for parsimonious cladograms and an IA is generated from the most parsimonious cladogram, diagnosis with PHAST or PAUP should return the same cladogram cost as that of the IA-based cladogram (given appropriate character transformation weights, indel cost, etc.). The use of IA coupled with phylogenetic analysis is entirely redundant (heuristics aside), but can be comforting nonetheless. Any differences between the DO output and the PAUP or PHAST analysis of IA would be due to the heuristics of different cladogram search procedures and would not represent any difference of approach.

* Fax: +212-769-5233.
E-mail address: wheeler@amnh.org.

0748-3007/03/$ - see front matter © 2003 The Willi Hennig Society. Published by Elsevier Science (USA). All rights reserved.
doi:10.1016/S0748-3007(03)00049-5

This and related methods have their origin in the Sankoff (1975) and Sankoff and Cedergren (1983) exact solution of the NP-complete (Wang and Jiang, 1994) tree alignment problem. The method proposed here is much simpler than that of Sankoff, since he was concerned with deriving the optimal solution over all cladogram topologies. Here, the cladogram search would be responsible for the quality of the solution, with the IA restricted to representing the homologies derived from that cladogram. Schwikowski and Vingron (1997) present heuristic solutions to the Sankoff exact solution that have approaches similar to that of IA but are also concerned with global multiple-alignment solutions.

Method

Consider a simple data set of five sequences (AA, AA, AGGG, ATT, and ATTG) related by a pectinate cladogram (Fig. 1). If we set the cost of indels to two and all types of nucleotide base transformation to one, this cladogram has a cost of nine weighted steps and the hypothetical ancestral sequences (H0–H3) specified in the figure. Following the DO downpass as described in Wheeler (1996), the preliminary states are established (often with many ambiguities of length and nucleotide). An uppass is then performed (Wheeler, 1996, 2002) to determine the hypothetical final ancestral sequences. The downpass established not only the preliminary states but also the correspondences among the pairs of descendant sequences and preliminary sequences at each node. These node-centered correspondences can be traced, like a rope, up from the root node to each descendent until the terminal taxa are reached. If positions correspond throughout the sequence set, these paths will span the entire cladogram. If, however, a deletion is encountered, that path will end. On the other hand, an insertion will create a new path, which will then move out and up from the node until it is deleted later or arrives at a terminal taxon. These paths can then be examined individually as a collection of individual nucleotides linked by putative homology. For those paths, which did not span the entire cladogram (i.e., involved indels), gap characters are placed in all the taxa without corresponding bases since the paths can link only actual (or reconstructed) nucleotides. These completed paths and nucleotides can then be arrayed in a 5′ to 3′ order and the hypothetical ancestral states removed. This is the implied alignment.

Two points worth noting come from the fact that these implied alignments are entirely dependent on the hypothetical ancestral sequence reconstructions. First, given that these reconstructions are unlikely to be minimal (given the complexity of the problem), the IA is unlikely in turn to be the minimal cost alignment for that cladogram and those terminals. Second, there may be multiple, equally costly ancestral reconstructions in heuristic optimal solutions, and the resulting ambiguity can effect the implied alignment. The sources of ambiguity and shortcomings of DO have been discussed (Wheeler, 1996, 2002) and these will apply equally to IA.

A synthetic example

Consider the sequences of Fig. 1. The positional correspondences, preliminary or downpass states, and final or up-pass sequences are determined as in Wheeler (1996, 2002). The lines of correspondence are shown in Fig. 2, with each path traced from its origin to its termination. Only two paths in this example span the entire cladogram and both insertions and deletions are inferred. The figure shows the same cladogram and paths with those paths extended to all nodes and with the vertices filled in with gaps. The paths are extracted, the internal nodal sequences are removed, and the IA of Fig. 3 is generated. The cladogram cost for this IA is nine weighted steps.

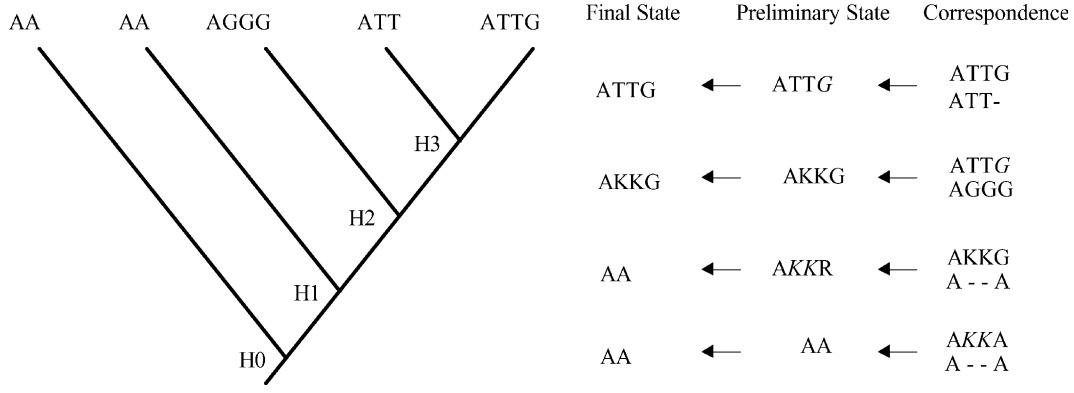

Fig. 1. Direct optimization after Wheeler (1996). H0–H3 represent hypothetical ancestral nodes. Correspondence refers to the correlated nucleotides (and inferred indels) derived from comparing two descendent taxa. The preliminary state is that derived from the downpass and the final state that from the uppass. IUPAC codes are used to represent ambiguous optimization, with italics denoting ambiguity with respect to indel (e.g., A = A or "-").

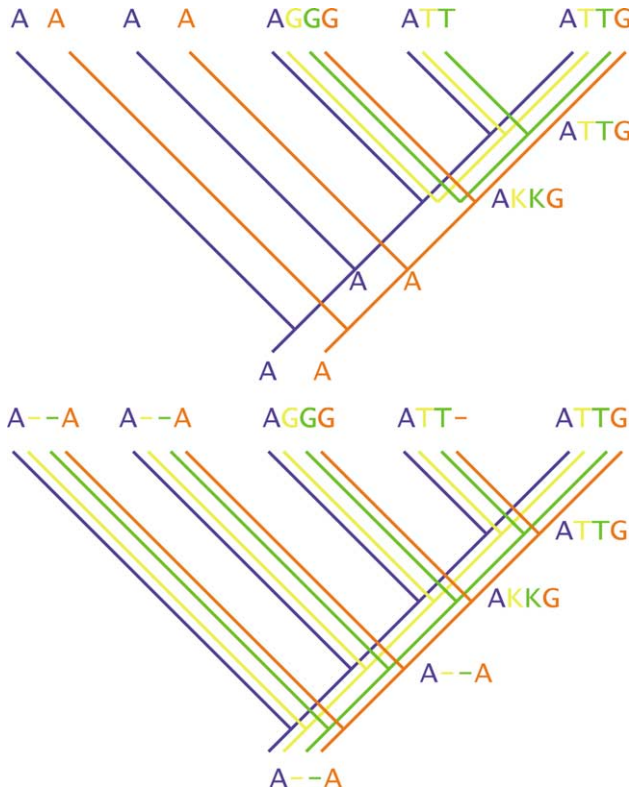

Fig. 2. Lines of correspondence. Each color line (blue, orange, yellow, green) traces the corresponding (homologous) nucleotides from the root of the cladogram through HTUs to terminal taxa.

TAX_A	A--A
TAX_B	A--A
TAX_C	AGGG
TAX_D	ATT-
TAX_E	ATTG

Fig. 3. Implied alignments based on Fig. 2.

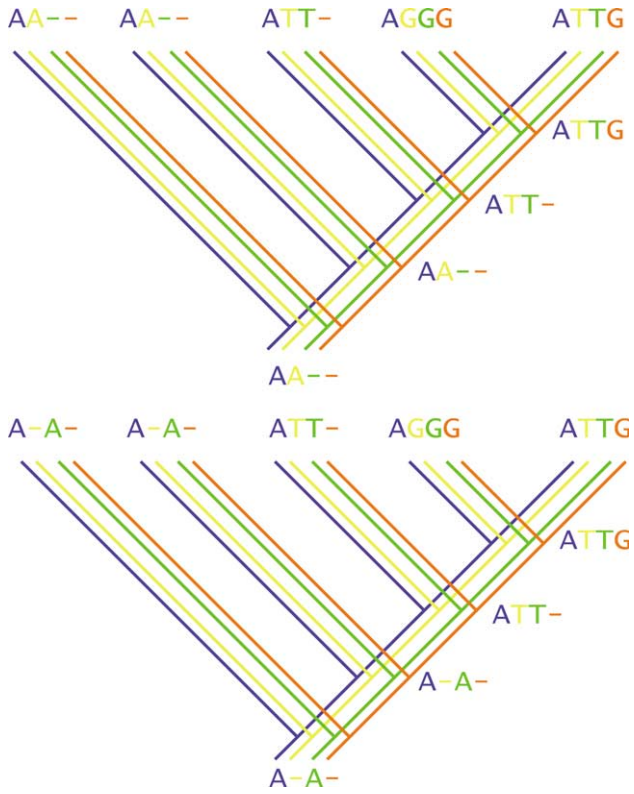

Fig. 4. Lines of correspondence based on an alternate cladogram showing ambiguous alignment.

TAX_A	AA--	A-A-
TAX_B	AA--	A-A-
TAX_C	AGGG	AGGG
TAX_D	ATT-	ATT-
TAX_E	ATTG	ATTG

Fig. 5. Implied alignments based on Fig. 4.

As mentioned above, the IA is not necessarily unique, however, and the example here demonstrates this property. If we do not force this scheme of relationships but instead search for the most parsimonious relationships, there is a solution at the cost of seven weighted steps. The relative placement of the paths which are filled to gaps in the base of the cladogram can vary. Here there are two relative placements of the paths, hence homology schemes, at equal cost (Fig. 4). The IA of each is shown in Fig. 5. Each of these homology schemes yields the cladogram (A (B (D (C E)))) at a cost of seven steps (which was the cladogram from which the IA was generated).

Comparisons with CLUSTAL and MALIGN

A short fragment of the small ribosomal subunit RNA (18S rDNA) from 17 chelicerates was used to illustrate the differences between the alignments produced by CLUSTALW (vers. 1.60; Higgins and Sharp, 1988, 1989; Thompson et al., 1994, 1997), MALIGN (vers. 2.8; Wheeler and Gladstein, 1994, 1991–1998), and IA (implemented in POY vers. 3.0.4; Wheeler et al., 2002). In each case, gaps were set to 2, with no differential extension penalty and unbiased costs for transitions and transversions. The alignments produced are shown in Figs. 6–8. Each alignment was then subjected to phylogenetic analysis using PHAST (Goloboff, 1996) with character state transformation costs set as in the alignment, 2 for indels, and 1 for base transformations. The cladograms generated from these three alignment procedures are quite different in both cost and topology. The CLUSTAL-based cladogram had a cost of 462 steps, the MALIGN 404 steps, and the IA 386 steps (Table 1; Fig. 9). Notably, the IA was the longest (greatest number of aligned positions) but still yielded the lowest-cost cladograms.

```
Americhenernes    TCGAGCCTCCAATGATACGTTGA-AAGGCGTTTA-TCGTTGGGGCCGAC---AGCGCGTCGT-GG--G-C-TC-------G-GT-TGGCC
Chanbria          TCTAGAC-TGGT-GGTCCGCC-T-CTGGTGGTTACTACCTGGCCTAAACA--ATTTGCCGGT-TT--T-C-CC---T---T-GG-TGCTC
Gea               TCCGGCC-GGACGGGTCCGCCTA-CCGGTGGTTACTGTTCGCTGCCGAG---CTTCAGGGGG-CC--G-C-TG---T---C-GA-TGATC
Hypochilus        TCCAGAC-GGGC-GGTCCGCCTA-ACGGTGGTTACTGCCTGGCCTGAAC---AACCAGCCGG-TT--T-C-CC---T---A-GA-TGATC
Thelechoris       TCCAGAC-GGGC-GGTCCGCCTA-ACGGTGGTTACTGCCTGGCCTGAAC---AGCCAGCCGG-TT--T-C-CC---T---A-GA-TGATC
Amblypygid        TCCAGAC-TGGC-GGTCCGCCTA-GCGGCGAGTACTGTCAGGCCTGAAC---ATGGCGCCGG-TT--T-T-CC---T---T-GG-TTCTC
Mastigoproctus    TCCAGAC-GGGT-GGTCCACCGC-CCGGTGGCTACTGCCCGGCCTGAAC-A-ATCTCGCCGG-TT--T-T-CC---T---T-GA-TTCTC
Rhipicephalus     TCCAGAC-GAGT-AGTGCATCTA-CCCGATGCTACGGCTCGGACTGAAC---ATCATGCCGGTTC--T-T-TC---T---T-GG-TGCAC
Vonones           TCGAGGC-TGGC-GGTCCGCCTA-CAGGCGGTCACTGCCAGTACTCAAC---ATCCTGCCGGTTT--T-C-CC---T---T-GG-TGCTC
Centruroides      TCCAGAC-AAGC-GGTCCACC-C-GCGGTGGTTACTGTTTGGACTGGAC---GTTTGGCCGG-AT--T-C-CT---T---T-GA-TGCTC
Hadrurus          TCCGGAC-TGTC-GGTCCGC--C--GCAAGCTTACTGGCAGGACCGGAC---GTCTAGCCGG-AC--T-C-TC---T---C-GTATCCTC
Paruroctonus      TCCAGAC-TGTC-GGTCCGCA-C--CGGAGGTTACTGGCAGGACCGGAC---GTCTAGCCGG-CC--T-C-CC---T---C-GT-TGCTC
Limulus           TCTAGAC-TGGC-GGTCCGCT-T-CCGGCGGTTACTGCCTGGCCTAAAC---ATCTGCCGGT-TT--T-C-CC---T---C-GG-TGCCC
Aportus           TCCAGAC-TGAC-GGTCCACC-G-CTCG-GGCGACTGTCAGGCCTGAAC---ATTCCGTCGGTTTCGACAGATTTTTCCCGCGG-TGCTC
Alentus           TCCAGAC-TGAC-GGTCCACC-GAATCG-GGCGACTGTCAGGCCTGAAC---ATCCGGTCGGTTTC-TCTCCCTTTTCCCGCGG-TGCTC
Artemia           CTCGGTC-GGGTGGTGCCGCCTC-ACGGTGGTCACTGCCTCGATCGGAC--AATTCATTGGA-TC--G-T-TC---G---G-GG-TGCTC
Thermobia         CTCGGAC-GATC-GGTTCGCC-G-CCCGTGTTAACTGATCGTTCCGGAC---GTCCTGCCGGTTT--T-T-CC-TTT--CTCGG-TGCTC

Americhenernes    TTAAAAAGCTGATCGG-GTTCTCCGGCAATTTTACTTTGAAAAAATTAGGGTGCTCAAAGCA-G-----G--C-CTGGTGCC
Chanbria          TTCACCGAGTGTCTTG-GGGGACTGGTACGTTTACTTTGAAGAAACTAGAGTGCTCAAAGCA--G----G--C-GTAACGCC
Gea               TTCATCGGTTATCTTC-CGTAACCCTCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--C-GCGACGCC
Hypochilus        TTCATTGATTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--C-GTGACGCC
Thelechoris       TTTACCGGTTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--G-CTGACGCC
Amblypygid        TTTACTGAGTGTCTTG-GGCGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA---G---G--C--TGTCGCC
Mastigoproctus    TTCACCGAGTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTTAAAGCA-------G--C-GTAACGCC
Rhipicephalus     TTCATTGTGTGCCTCGAGATGGCCGGTGCTTTTACTTTGAAAAAATTAGAGTGCTCAACGCA---G---G--C-GAGTCGCC
Vonones           TTCGCTGAGTGTCTCG-GGTGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA---G---G--C-GCGTAGCC
Centruroides      TTTGCCGAGTGTCTTG-GGTGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--C-GTACCGCC
Hadrurus          TTCACCGGGTGTCTTG-GGTGTCCGGCAATTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA----GC-G--C-GATCCGCC
Paruroctonus      TTCACCGGGTGTCTTG-GGTGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--C-GACCCGCC
Limulus           TTGATTGAGTGTCTTG-GGTGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA-------G--C-GCAACGCC
Aportus           TTCGGTGAGTGTCGC--GGAGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA---G---G--C-G-GTCACC
Alentus           TTCGTTGAGTGTCGCG-GGAGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCA---G---G--C-G-GTCACC
Artemia           TTAACCGAGTGTCCTG-GGTGGCCGATACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGG----TG--C-ACCGCGCC
Thermobia         TTCATTGAGTGTCTTG-ATTGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTCAAAGCA---G---GTCCAGTGTCGCC
```

Fig. 6. Implied alignment of chelicerate fragment from Wheeler and Hayashi (1998) using POY (vers. 3.0.4) with TBR branch swapping after simple addition.

```
Americhenernes    TCGAGCCTCCAATGAT-ACGTTG-A-AAGGCGTTTATCGTTGGG-GCCGAC-AGCGCGTCGT-----G-GG--CTC-G--G-T-TGGCCT
Chanbria          TCTAGACTGGT-GG-T-CCGCCT-C-TGGTGGTTACTACCTGGC-CTAAAC-AATTTGCCGG-----T-TT-TCCCTT--G-G-TGCTCT
Gea               TCCGGCCGGACGGG-T-CCGCCTAC-CGGTGGTTACTGTTCGCT-GCCGAG-CTTCAGGGGG-----C-CG--CTGTC--G-A-TGATCT
Hypochilus        TCCAGACGGGC-GG-T-CCGCCTAA-CGGTGGTTACTGCCTGGC-CTGAAC-AACCAGCCGG-----T-TT--CCCTA--G-A-TGATCT
Thelechoris       TCCAGACGGGC-GG-T-CCGCCTAA-CGGTGGTTACTGCCTGGC-CTGAAC-AGCCAGCCGG-----T-TT--CCCTA--G-A-TGATCT
Amblypygid        TCCAGACTGGC-GG-T-CCGCCTAG-CGGCGAGTACTGTCAGGC-CTGAAC-ATGGCGCCGG-----T-TT--TCCTT--G-G-TTCTCT
Mastigoproctus    TCCAGACGGGT-GG-T-CCACCGCC-CGGTGGCTACTGCCCGGC-CTGAACAATCTCGCCGG-----T-TT--TCCTT--G-A-TTCTCT
Rhipicephalus     TCCAGACGAGT-AG-TGCATCTA-C-CCGATGCTACGGCTCGGA-CTGAAC-ATCATGCCGG-----T-TC-TTTCTT--G-G-TGCACT
Vonones           TCGAGGCTGGC-GG-T-CCGCCTAC-AGGCGGTCACTGCCAGTA-CTCAAC-ATCCTGCCGG-----T-TT-TCCCTT--G-G-TGCTCT
Centruroides      TCCAGACAAGC-GG-T-CCACCC-G-CGGTGGTTACTGTTTGGA-CTGGAC-GTTTGGCCGG-----A-TT--CCTTT--G-A-TGCTCT
Hadrurus          TCCGGACTGTC-GG-T-CCGCCG-C-AAG-CTTA-CTGGCAGGA-CCGGAC-GTCTAGCCGG-----A-CT--CTCTC--GTA-TCCTCT
Paruroctonus      TCCAGACTGTC-GG-T-CCGC-A-C-CGGAGGTTACTGGCAGGA-CCGGAC-GTCTAGCCGG-----C-CT--CCCTC--G-T-TGCTCT
Limulus           TCTAGACTGGC-GG-T-CCGCTT-C-CGGCGGTTACTGCCTGGC-CTAAAC-ATC-TGCCGG-----T-TT-TCCCTC--G-G-TGCCCT
Aportus           TCCAGACTGAC-GG-T-CCACCG-C-TCG-GGCGACTGTCAGGC-CTGAAC-ATTCCGTCGGTTTCGACAGATTTTTCCCGCGGTGCTCT
Alentus           TCCAGACTGAC-GG-T-CCACCG-AATCG-GGCGACTGTCAGGC-CTGAAC-ATCCGGTCGGTTTC-TCTCCCTTTTCCCGCGGTGCTCT
Artemia           CTCGGTCGGGT-GG-TGCCGCCT-CACGGTGGTCACTGCCTCGATCGGACA-ATT-CATTGG-----A-TC-GTTCGG--G-G-TGCTCT
Thermobia         CTCGGACGATC-GG-T-TCGCCG-C-CCGTGTTAACTGATCGTT-CCGGAC-GTCCTGCCGGTTT--T-TCCTTTCTC--G-G-TGCTCT

Americhenernes    TAAAAAGCTGATCGG-GTTCTCCGGCAATTTTACTTTGAAAAAATTAGGGTGCTCAAAGCAGGC--C-T-GGTGC--C
Chanbria          TCACCGAGTGTCTTG-GGGGACTGGTACGTTTACTTTGAAGAAACTAGAGTGCTCAAAGCAGGC--G-T-AACGC--C
Gea               TCATCGGTTATCTTC-CGTAACCCTCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-C--G-C-GACGC--C
Hypochilus        TCATTGATTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-C--G-T-GACGC--C
Thelechoris       TTACCGGTTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-G--C-T-GACGC--C
Amblypygid        TTACTGAGTGTCTTG-GGCGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-G--C-T-GTCGC--C
Mastigoproctus    TCACCGAGTGTCTTG-GGTGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTTAAAGCAG-C--G-T-AACGC--C
Rhipicephalus     TCATTGTGTGCCTCGAGATGGCCGGTGCTTTTACTTTGAAAAAATTAGAGTGCTCAACGCAG-GC-G-A-GTCGC--C
Vonones           TCGCTGAGTGTCTCG-GGTGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-G--C-GCGTAGC--C
Centruroides      TTGCCGAGTGTCTTG-GGTGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-C--G-T-ACCGC--C
Hadrurus          TCACCGGGTGTCTTG-GGTGTCCGGCAATTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGCGC-G-A-TCCGC--C
Paruroctonus      TCACCGGGTGTCTTG-GGTGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-C--G-A-CCCGC--C
Limulus           TGATTGAGTGTCTTG-GGTGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-C--G-C-AACGC--C
Aportus           TCGGTGAGTGTCGCG-G-AGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-G--C-G-GTCAC--C
Alentus           TCGTTGAGTGTCGCG-GGAGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-G--C-G-GTCAC--C
Artemia           TAACCGAGTGTCCTG-GGTGGCCGATACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAG-GT-G-C-ACCGCGCC
Thermobia         TCATTGAGTGTCTTG-ATTGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTCAAAGCAG-GTCCAGTGTCGC--C
```

Fig. 7. Multiple alignment of chelicerate fragment from Wheeler and Hayashi (1998) using MALIGN (vers. 2.8) with TBR and root swapping after simple addition.

```
Americhernes      -TCGAGCCTCCAATGATACGTTGAAAGGCGTTTATCGTTGGGGCCGACAGCGCGTCGTGGGCT---------------CGGTTGGCCTT
Chanbria          TCTAGACTGGTGGT---CCGCCTC-TGGTGGTTACTACCTGGCCTAAACAATTT-GCCGGTTT----------T-CCCTTGG-TGCTCTT
Gea               TCCGGCCGGACGGGT--CCGCCTACCGGTGGTTACTGT-TCGCTGCCGAGCTTCAGGGGGCCG-------------CTGTCGATGATCTT
Hypochilus        TCCAGACGGGCGGT---CCGCCTAACGGTGGTTACTGCCTGGCCTGAACAACCA-GCCGGTTT----------C--CCTAGA-TGATCTT
Thelechoris       TCCAGACGGGCGGT---CCGCCTAACGGTGGTTACTGCCTGGCCTGAACAGCCA-GCCGGTTT----------C--CCTAGA-TGATCTT
Amblypygid        TCCAGACTGGCGGT---CCGCCTAGCGGCGAGTACTGTCAGGCCTGAACATGGC-GCCGGTTT----------T--CCTTGG-TTCTCTT
Mastigoproctus    TCCAGACGGGTGGT---CCACCGCCCGGTGGCTACTGCCCGGCCTGAACAATCTCGCCGGTTT----------T--CCTTGA-TTCTCTT
Rhipicephalus     TCCAGACGAGTAGT---GCATCTACCCGATGCTACGGCTCGGACTGAACATCAT-GCCGGTTC----------T--TTCTTGGTGCACTT
Vonones           TCGAGGCTGGCGGT---CCGCCTACAGGCGGTCACTGCCAGTACTCAACATCCT-GCCGGTTT----------T-CCCTTGG-TGCTCTT
Centruroides      TCCAGACAAGCGGT---CCACCCG-CGGTGGTTACTGTTTGGACTGGACGTTTG-GCCGGATT----------C--CTTTGA-TGCTCTT
Hadrurus          TCCGGACTGTCGGT---CCGCC-G-CA-AGCTTACTGGCAGGACCGGACGTCTA-GCCGGACT----------C--TCTCGTATCCTCTT
Paruroctonus      TCCAGACTGTCGGT---CCGCA-C-CGGAGGTTACTGGCAGGACCGGACGTCTA-GCCGGCCT----------C--CCTCGT-TGCTCTT
Limulus           TCTAGACTGGCGGT---CCGCTTC-CGGCGGTTACTGCCTGGCCTAAACA-TCT-GCCGGTTT----------T-CCCTCGG-TGCCCTT
Aportus           TCCAGACTGACGGT---CCACCGC-TCG-GGCGACTGTCAGGCCTGAACATTCC-GTCGGTTTCGACAGATTTTTCCCGCGG-TGCTCTT
Alentus           TCCAGACTGACGGT---CCACCGAATCG-GGCGACTGTCAGGCCTGAACATCCG-GTCGGTTTCT-CTCCCTTTTCCCGCGG-TGCTCTT
Artemia           CTCGGTCGGGTGGTG--CCGCCTCACGGTGGTCACTGCCTCGATCGGACAATTCATTGGATCG-------------TTCGGGGTGCTCTT
Thermobia         CTCGGACGATCGGT---TCGCCGC-CCGTGTTAACTGATCGTTCCGGACGTCCT-GCCGGTTT--------TTCCTTTCTCGGTGCTCTT

Americhernes      AAAAAGCTGATCGGGT-TCTCCGGCAATTTTACTTTGAAAAAATTAGGGTGCTCAAAGCAGGCCTGGTGCC----
Chanbria          CACCGAGTGTCTTGGG-GGACTGGTACGTTTACTTTGAAGAAACTAGAGTGCTCAAAGCAGGCGTAACGC-C---
Gea               CATCGGTTATCTTCCG-TAACCCTCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-CGCGACGC---C-
Hypochilus        CATTGATTGTCTTGGG-TGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGCGTGACGC-C----
Thelechoris       TACCGGTTGTCTTGGG-TGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGGCTGACGCC-----
Amblypygid        TACTGAGTGTCTTGGG-CGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGGCTGTCGCC-----
Mastigoproctus    CACCGAGTGTCTTGGG-TGACCGGCACGTTTACTTTGAAAAAATTAGAGTGCTTAAAGCAGCGTAACGCC-----
Rhipicephalus     CATTGTGTGCCTCGAGATGGCCGGTGCTTTTACTTTGAAAAAATTAGAGTGCTCAACGCAGGCGAGTCGCC----
Vonones           CGCTGAGTGTCTCGGG-TGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGGCGCGTAGCC----
Centruroides      TGCCGAGTGTCTTGGG-TGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGCG--TACCGCC---
Hadrurus          CACCGGGTGTCTTGGG-TGTCCGGCAATTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGCGCGATCCGC-C--
Paruroctonus      CACCGGGTGTCTTGGG-TGTCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGCG--ACCCGCC---
Limulus           GATTGAGTGTCTTGGG-TGGCCGGCACGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAG-CGCAACGCC----
Aportus           CGGTGAGTGTCGCGG--AGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGGCGGTCAC-----C
Alentus           CGTTGAGTGTCGCGGG-AGGCCGACAAGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAGGCGGTCACC-----
Artemia           AACCGAGTGTCCTGGG-TGGCCGATACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGTGCACCGCGCC--
Thermobia         CATTGAGTGTCTTGAT-TGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTCAAAGCAGGTCCAGTGTCGCC-
```

Fig. 8. Multiple alignment of chelicerate fragment from Wheeler and Hayashi (1998) using CLUSTALW (vers. 1.6) with default guide tree calculation.

Table 1
Implied alignment, CLUSTAL, and MALIGN analyses

Alignment method	Aligned positions	Execution time (s)	Cladogram cost
Implied alignment	172	2	386
CLUSTAL	165	3	462
MALIGN	168	100	404

Static approximation

Cladogram cost calculations of unequal length sequences via DO can be time consuming—in general $O(n^2)$ for each node with sequences of length "n" (although use of the algorithm in Ukkonen (1985) can result in near-linear performance for well-behaved data). Such calculations made from aligned sequences are linear with sequence length. It seems logical, therefore, to use the IA to provide a heuristic estimate of the cladogram cost of unaligned sequences and to perform the cladogram optimizations (during branch swapping, etc.) much more rapidly. This would approximate the "dynamic" topology-specific homologies of DO with the "static" global statements of multiple alignment (Wheeler, 2001).

The method is quite simple. For each new cladogram of minimum cost found, an IA is performed and cladogram cost calculations are performed on the aligned data until a new shortest cladogram is found, this cost is verified by a complete DO downpass, and a new IA is generated. This process is repeated until no new cladograms are found as with any search (POY option "-staticapprox").

In a simple test data set (20 complete 18S rRNA chelicerate sequences of approximately 1750 bp; G. Giribet, pers. com.), a round of TBR branch swap (using POY) to completion resulted in a final cladogram cost of 1497 steps (indels = 2, base changes = 1) examining 2574 candidate cladograms. On a 800-MHz PIII running Windows 2000, this took 257 s, hence evaluating 10.02 cladograms per second. Using the option "-staticapprox" in POY resulted in an identical TBR search in 59 s, hence 43.63 cladograms per second—a speedup of over 4×. Since the cladogram length calculation involves another level of heuristic, results may not be identical to those of standard searches. In my experience, speedups of a factor of 4 or 5 are not uncommon.

"Exact" cost calculations

In addition to creating shortcuts in search strategies, implied alignments can improve the estimates of cladogram cost. As discussed in Wheeler (2002), the DO cladogram costs are upper bounds of the NP-complete

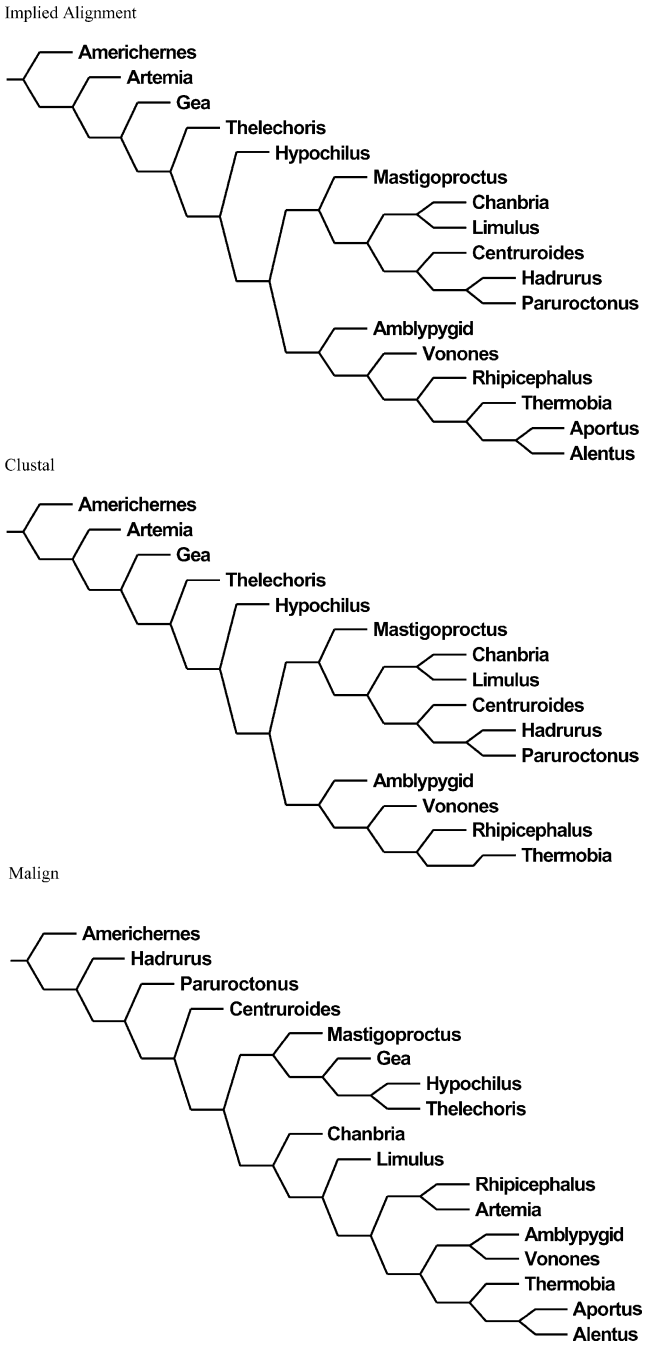

Fig. 9. Cladograms generated from the alignments of Figs. 6–8 using PHAST (vers. 1.9) with all base transformations set to equality and indels set to two.

problem of tree. There are two levels of heuristics in the cladogram cost calculation, nucleotide homology and weighted cladogram cost given that homology. The DO algorithm relies on greedy approximations for each of these problems (Wheeler, 2002). IA can make the cost calculation (but not the homology statements) exact by applying dynamic programming (Sankoff and Rousseau, 1975) to the aligned sequences (POY option "-exact"). During cladogram search, an IA is generated for each candidate topology and dynamic programming is used to determine cladogram cost. This can be time consuming. When the chelicerate 18S data set used above undergoes a simple Wagner + TBR search in which indels cost 4, transversions 2, and transitions 1, one cladogram of cost 2508 was found. With the "-exact" option specified in POY, two cladograms (both slightly different from the "nonexact" cladogram) of cost 2506 were found. The costs of these cladograms were verified by PHAST. For more homogeneous parameter regimes (e.g., indels = transversions = transitions) on these chelicerate data, this difference is not found. I have found that the improvements by "-exact" are more pronounced when more extreme weighting schemes are employed (although they can be found when all events are equally weighted as well).

Discussion

In my experience, IA-based multiple alignments are more parsimonious than those generated by MALIGN or CLUSTAL in all but trivial cases. This is not surprising since the search-based IA will examine many more guide scenarios than CLUSTAL and is closer to cladogram diagnosis than MALIGN. As such, IA is a superior multiple-alignment procedure for analyses where parsimonious results are desired. Furthermore, IA has great utility in improving both execution time and cladogram cost calculation for some data.

When comparing IA with the procedures of MALIGN and CLUSTAL, the phylogenetic topology generated by DO serves as a type of guide tree. The main difference between the methods is in the determination of the homologies at the internal nodes and how these are traced through the cladogram. As such, IA may look very different from standard multiple alignments when dealing with insertions (see Figs. 6–8). Since independent insertions are regarded by IA as nonhomologies, they will not "line up" in the alignment. Approximately 50 bases in from the 5′ end of the IA in Fig. 6, are three aligned positions, each of which has an "A" in a single taxon and gaps elsewhere (Fig. 10a). This might appear obviously "wrong" at first since this would not allow for the situation where they were a homologous insertion. This is not the case, however. Since the IA was based on a cladogram where the three "A" taxa were scattered, these insertions must have had independent origins and should not appear putatively homologous. If one forces these three taxa into a clade, the resultant IA compresses these three positions into one; hence, they are putatively homologous (Fig. 10b). This illustrates one of the central distinctions between IA and traditional multiple alignment. IA is based on synapomorphy, hence, it cannot escape its dependence on the cladistic relationships of the taxa. This is its greatest strength.

a

```
                                                             ***
Americhenernes   TCGAGCCTCCAATGATACGTTGA-AAGGCGTTTA-TCGTTGGGGCCGAC---AGCGCGTCGT-GG--G-C-TC-------G-GT-TGGCC
Chanbria         TCTAGAC-TGGT-GGTCCGCC-T-CTGGTGGTTACTACCTGGCCTAAACA--ATTTGCCGGT-TT--T-C-CC---T---T-GG-TGCTC
Gea              TCCGGCC-GGACGGGTCCGCCTA-CCGGTGGTTACTGTTCGCTGCCGAG---CTTCAGGGGG-CC--G-C-TG---T---C-GA-TGATC
Hypochilus       TCCAGAC-GGGC-GGTCCGCCTA-ACGGTGGTTACTGCCTGGCCTGAAC---AACCAGCCGG-TT--T-C-CC---T---A-GA-TGATC
Thelechoris      TCCAGAC-GGGC-GGTCCGCCTA-ACGGTGGTTACTGCCTGGCCTGAAC---AGCCAGCCGG-TT--T-C-CC---T---A-GA-TGATC
Amblypygid       TCCAGAC-TGGC-GGTCCGCCTA-GCGGCGAGTACTGTCAGGCCTGAAC---ATGGCGCCGG-TT--T-T-CC---T---T-GG-TTCTC
Mastigoproctus   TCCAGAC-GGGT-GGTCCACCGC-CCGGTGGCTACTGCCCGGCCTGAAC-A-ATCTCGCCGG-TT--T-T-CC---T---T-GA-TTCTC
Rhipicephalus    TCCAGAC-GAGT-AGTGCATCTA-CCCGATGCTACGGCTCGGACTGAAC---ATCATGCCGGTTC--T-T-TC---T---T-GG-TGCAC
Vonones          TCGAGGC-TGGC-GGTCCGCCTA-CAGGCGGTCACTGCCAGTACTCAAC---ATCCTGCCGGTTT--T-C-CC---T---T-GG-TGCTC
Centruroides     TCCAGAC-AAGC-GGTCCACC-C-GCGGTGGTTACTGTTTGGACTGGAC---GTTTGGCCGG-AT--T-C-CT---T---T-GA-TGCTC
Hadrurus         TCCGGAC-TGTC-GGTCCGC--C--GCAAGCTTACTGGCAGGACCGGAC---GTCTAGCCGG-AC--T-C-TC---T---C-GTATCCTC
Paruroctonus     TCCAGAC-TGTC-GGTCCGCA-C--CGGAGGTTACTGGCAGGACCGGAC---GTCTAGCCGG-CC--T-C-CC---T---C-GT-TGCTC
Limulus          TCTAGAC-TGGC-GGTCCGCT-T-CCGGCGGTTACTGCCTGGCCTAAAC---ATCTGCCGGT-TT--T-C-CC---T---C-GG-TGCCC
Aportus          TCCAGAC-TGAC-GGTCCACC-G-CTCG-GGCGACTGTCAGGCCTGAAC---ATTCCGTCGGTTTCGACAGATTTTTCCCGCGG-TGCTC
Alentus          TCCAGAC-TGAC-GGTCCACC-GAATCG-GGCGACTGTCAGGCCTGAAC---ATCCGGTCGGTTTC-TCTCCCTTTTCCCGCGG-TGCTC
Artemia          CTCGGTC-GGGTGGTGCCGCCTC-ACGGTGGTCACTGCCTCGATCGGAC--AATTCATTGGA-TC--G-T-TC---G---G-GG-TGCTC
Thermobia        CTCGGAC-GATC-GGTTCGCC-G-CCCGTGTTAACTGATCGTTCCGGAC---GTCCTGCCGGTTT--T-T-CC-TTT--CTCGG-TGCTC
```

b

```
Americhenernes   TCGAGCCTCCAATGAT-ACGTTGA-AAGGCGTTTA-TCGTTGGGGCCGACAGC-GCGTCGT--GG--G-C--T------C-GGT-TGGCC
Chanbria         TCTAGAC-TGGT-GGT-CCGCCTC--TGGTGGTTACTACCTGGCCTAAACAAT-TTGCCGGT-TT--T-C--C------CTTGG-TGCTC
Gea              TCCGGCC-GGACGGGT-CCGCCTA-CCGGTGGTTACTGTTCGCTGCCGAGCTT-CAGGGGG--CC--G-C--T------GTCGA-TGATC
Hypochilus       TCCAGAC-GGGC-GGT-CCGCCTA-ACGGTGGTTACTGCCTGGCCTGAACAAC-CAGCCGG--TT--T-C--C------CTAGA-TGATC
Thelechoris      TCCAGAC-GGGC-GGT-CCGCCTA-ACGGTGGTTACTGCCTGGCCTGAACAGC-CAGCCGG--TT--T-C--C------CTAGA-TGATC
Amblypygid       TCCAGAC-TGGC-GGT-CCGCCTA-GCGGCGAGTACTGTCAGGCCTGAACATG-GCGCCGG--TT--T-T--C------CTTGG-TTCTC
Mastigoproctus   TCCAGAC-GGGT-GGT-CCACCGC-CCGGTGGCTACTGCCCGGCCTGAACAATCTCGCCGG--TT--T-T--C------CTTGA-TTCTC
Rhipicephalus    TCCAGAC-GAGT-AGT-GCATCTA-CCCGATGCTACGGCTCGGACTGAACATC-ATGCCGG--TT--C-T-TT------CTTGG-TGCAC
Vonones          TCGAGGC-TGGC-GGT-CCGCCTA-CAGGCGGTCACTGCCAGTACTCAACATC-CTGCCGG--TT--T-T-CC------CTTGG-TGCTC
Centruroides     TCCAGAC-AAGC-GGT-CCACC-C-GCGGTGGTTACTGTTTGGACTGGACGTT-TGGCCGG--AT--T-C--C------TTTGA-TGCTC
Hadrurus         TCCGGAC-TGTC-GGT-CCGC--C--GCAAGCTTACTGGCAGGACCGGACGTC-TAGCCGG--AC--T-C--T------CTCGTATCCTC
Paruroctonus     TCCAGAC-TGTC-GGT-CCGCA-C--CGGAGGTTACTGGCAGGACCGGACGTC-TAGCCGG--CC--T-C--C------CTCGT-TGCTC
Limulus          TCTAGAC-TGGC-GGT-CCGCTT--CCGGCGGTTACTGCCTGGCCTAAACAT--CTGCCGG--TT--T-T-CC------CTCGG-TGCCC
Aportus          TCCAGAC-TGAC-GGT-CCACCG--CTCG-GGCGACTGTCAGGCCTGAACATT-CCGTCGG-TTTCGACAGATTTTTCCCGCGG-TGCTC
Alentus          TCCAGAC-TGAC-GGT-CCACCG-AATCG-GGCGACTGTCAGGCCTGAACATC-CGGTCGG-TTTC-TCTCCCTTTTCCCGCGG-TGCTC
Artemia          CTCGGTC-GGGT-GGTGCCGCCTC-ACGGTGGTCACTGCCTCGATCGGACAAT-TCATTGGA-TC--G-T--T------CGGGG-TGCTC
Thermobia        CTCGGAC-GATC-GGT-TCGCCG--CCCGTGTTAACTGATCGTTCCGGACGTC-CTGCCGG-TTT--T-T-CC-TTT--CTCGG-TGCTC
```

Fig. 10. Implied alignments of chelicerate fragment from Wheeler and Hayashi (1998), highlighting the areas of difference when *Chanbria*, *Mastigoproctus*, and *Artemia* sequences are treated as a clade (a) and as polyphyletic (b). The sequences correspond to the upper portion of Fig. 6.

Acknowledgments

I acknowledge the help and prodding of Cyrille D'Haese, Jan De Laet, Gonzalo Giribet, Daniel Janies, Jyrki Muona, Julian Favovitch, Taran Grant, and an anonymous reviewer for many helpful critisms. Grant support has been generously supported by NSF Systematic Biology and NASA Fundamental Space Biology Program.

References

Goloboff, P., 1996. PHAST. Program and Documentation. Version 1.9.

Higgins, D.G., Sharp, P.M., 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.

Higgins, D.G., Sharp, P.M., 1989. Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5, 151–153.

Sankoff, D.D., 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 78, 35–42.

Sankoff, D.D., Cedergren, R.J., 1983. Simultaneous comparison of three or more sequences related by a tree. In: Sankoff, D., Kruskal, J.B. (Eds.), Time Warps, String Edits, and Macromolecules: the Theory and Practise of Sequence Comparison. Addison-Wesley, Reading, MA, pp. 253–264.

Sankoff, D.D., Rousseau, P., 1975. Locating the vertices of a Steiner tree in arbitrary space. Math. Prog. 9, 240–246.

Schwikowski, B., Vingron, M., 1997. The deferred path heuristic for the generalized tree alignment problem. J. Comp. Biol., 415–431.

Swofford, D., 2001. PAUP. Version 4.0. Smithsonian Institution.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–48821.

Ukkonen, E., 1985. Finding approximate patterns in strings. J. Algorith. 6, 132–137.

Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348.

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

Wheeler, W.C., 1999. Fixed character states and the optimization of molecular sequence data. Cladistics 15, 379–385.

Wheeler, W.C., 2001. Homology and the optimization of DNA sequence data. Cladistics 17, S3–S11.

Wheeler, W.C., 2002. Optimization alignment: down, up, error, and improvements. In: Desalle, R., Giribet, G., Wheeler, W. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser Verlag, Basel Switzerland, pp. 55–69.

Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 24, 173–192.

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

Wheeler, W.C., Gladstein, D.S., 1991–1998. Malign. Program and documentation. New York, NY. Documentation by Daniel Janies and Ward Wheeler.

Wheeler, W.C., De Laet, J., Gladstein, D.S., 2002. POY: The Optimization of Alignment Characters. Version 3.0.4. Program and Documentation. New York, NY. Available at ftp.amnh.org/pub/molecular. Documentation by D. Janies and W.C. Wheeler.