Cladistics **15**, 379–385 (1999)
Article ID clad.1999.0115, available online at http://www.idealibrary.com on IDEAL®

Fixed Character States and the Optimization of Molecular Sequence Data

Ward Wheeler[1]

Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192

Accepted October 2, 1999

A method is proposed to optimize molecular sequence data that does not employ multiple sequence alignment. This method treats entire homologous contiguous stretches of sequence data as individual characters. This sequence is treated as the homologous unit employed in phylogeny reconstruction. The sets of specific sequences exhibited by the terminal taxa constitute the character states. The number of states is then less than or equal to the number of unique sequences (or homologous fragments) exhibited by the data. A matrix of transformation costs is created to relate the states to one another. The cells of this matrix are defined as the minimum transformation cost between each pair of states based on insertion–deletion and base substitution costs. The diagnosis of a topology then follows existing dynamic programming techniques, with the number of states greatly expanded. Since the possible sequences reconstructed at nodes are limited to those exhibited by the terminals, cladograms constructed in this way may be longer than those of other methods in that they require a greater number of weighted evolutionary events. Example data, the effects of missing data, restricted ancestors, and putative long-branch attraction are discussed.

INTRODUCTION

Molecular sequence data present problems in systematic analysis which are often the inverse of those presented by other forms of character data. While sequences offer potentially huge numbers of characters, the states of these characters are limited in number (four or five) and all the characters share the same set of states. When the length of these sequences varies (as it frequently does), this identity of states creates confusion. Which of the positions in one sequence correspond (putatively homologous) to those in another? Other forms of character data are able to avoid this problem through the comparison of the huge variety of possible states complex characters can exhibit. Features of the head of a creature are not easily confused with those of a limb. In both molecular and morphological analyses, position plays a role, but even when both cranial and limb features are scored "condition one and condition two," the descriptions of "condition one" in the two morphological characters make clear their incomparability. DNA characters might exhibit A, C, G, T, or "gap"—at all positions. An A in position one is indistinguishable from another A in position 500. Where do the potential homologies lie?

This is the problem and the thinking that motivated

[1] Fax: 1-212-769-5783. E-mail: wheeler@amnh.org.

0748-3007/99 $30.00

the efforts of multiple sequence alignment. Many procedures have been developed to convert the contiguous strings of nucleotide bases into a series of column vector characters via the insertion of "gaps" to allow the same sorts of analysis used for other types of character data. This desire for topology-specific base-to-base homology also motivated multiple alignment free analysis such as direct optimization or "optimization alignment" (Wheeler, 1996). In both multiple alignment and optimization alignment, the fundamental homology statements are made at the level of the individual nucleotide bases. If this idea is abandoned and the sequences themselves are treated as the fundamental units of homology, a simpler notion of character diagnosis and interpretation can be developed. It is this notion of contiguous sequence homology as opposed to individual nucleotide homology that distinguishes this parsimony method from others.

This method changes what is used as a character in molecular sequence data. In standard analysis, individual nucleotide position variation is the stuff of cladogram diagnosis. Here, it is change in the entire sequence that we use. The homologous unit becomes the contiguous sequence, however defined. This may be an entire locus (such as the 18S rDNA) or smaller pieces of that locus (primer or secondary structure defined regions), but individual nucleotide homologies are never created.

THE METHOD

The procedure is simple. The first step is to convert each terminal taxon sequence to character states (denoted 0 through number of taxa − 1; hence, $n_{\text{taxa}} = n_{\text{states}}$) and create the cost matrix relating these states. The cells of this Sankoff or "step matrix" are defined as the minimum cost of transformation between each pair of states. This transformation cost is calculated as the weighted sum of each type of base transformation as well as insertion and deletion events (as in calculation of an alignment score). The next operation uses this transformation matrix to diagnose a specific phylogenetic topology via the well-known techniques of dynamic programming (Sankoff and Rousseau, 1975). The only possible states at internal nodes are those exhibited by the terminal taxa, and the operation is repeated over multiple topologies and the most parsimonious phylogenetic topology is determined. A complete down-pass optimization by this method of "n" taxa will involve the comparison of a maximum of "n" states at each node ($2n^2$ operations) over $n - 1$ nodes for a total cost of computation proportional to the cube of the number of terminal taxa ($\sim 2n^3$). No doubt, shortcuts can be found to expedite this process (Goloboff, 1994, 1998).

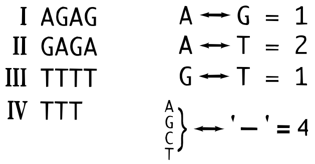

FIG. 1. Four hypothetical DNA sequences and base substitution and insertion–deletion event transformation matrix.

Consider four sequences: I = AGAG, II = GAGA, III = TTTT, and IV = TTT. These sequences differ in both base composition and number. Furthermore, specify a substitution matrix where transformations between A and G cost 1 as do transformations between G and T. Transformations between A and T, however, cost 2 and all flavors of insertions and deletions cost 4 (Fig. 1). As mentioned above, the next step determines the matrix relating the four states to each other. The minimum cost of transformation between I and II would be 4, that between I and III 6, I and IV 8, II and III 6, II and IV 8, and III and IV 4 (Fig. 2). Now each of these states is tested at each internal node via dynamic programming to determine the most parsimonious optimization of this topology in this case for a total length of 14.

	I	II	III	IV
I	-	4	6	8
II	4	-	6	8
III	6	6	-	4
IV	8	8	4	-

FIG. 2. Fixed-character-state transformation matrix. The numbers represent the transformation costs between all sequence pairs in terms of weighted bases, changes, and insertion–deletion events.

AN EXAMPLE

To more fully demonstrate the behavior of this approach in systematic analysis, the data of Wheeler and Hayashi (1998) were subjected to fixed-character-state optimization. These data were drawn from 13 chelicerate clades and several outgroup taxa. A total of 93 morphological characters were culled from the literature and added to sequence data from the nuclear 18S and 28S ribosomal DNA. Based on the presence of primer sequences and secondary structure motifs, the molecular sequences were broken into 16 corresponding pieces (9 for the 18S and 7 for the 28S). The data were accumulated for a total of 34 taxa.

The morphological characters were all treated as unordered characters and each of the 16 corresponding pieces of the molecular data treated as a single character with 34 (or fewer if sequences are repeated or taxa have missing data) states. Using the program POY (Gladstein and Wheeler, 1997), phylogenetic reconstruction was performed with transitions and transversions having equal cost (set at 1) for the molecular data alone and transversions twice transitions for the combined data (set at 2). The insertion–deletion events and morphological transformations were weighted twice the cost of transversions (set at 2 and 4, respectively). These parameter schemes were those used by Wheeler and Hayashi (1998) to minimize character incongruence.

The topology of the total evidence cladogram (Fig. 3) does not differ greatly from that published by Wheeler and Hayashi (1998), but the lengths of the combined data cladogram (4691 published versus 4510 with fixed-state optimization) and the molecular partitions cladogram (2432 published versus 2331 with fixed-state optimization) do differ. In each case, the cladogram lengths are shorter than those of the published direct optimization alignment results (Fig. 4). This shorter length may seem strange in light of comments above and the conditions of Fig. 6, but is most likely due to two effects. The first is related to the data in that the 34 taxa sampled by Wheeler and Hayashi (1998) may have sufficient sequence diversity to leave ancestral reconstructions relatively unconstrained. The second is due to the method itself and is derived from the absence of global base-to-base homology schemes. Since the sequences are only related in a pairwise fashion, the need to accommodate sequence length variation throughout the cladogram is gone. This may remove some of the "mess" created by the expansion and contraction of sequences and the difficulty some optimization schemes have in accommodating these events coherently.

MISSING DATA

Missing data may have unforeseen effects (Platnick *et al.*, 1991; Nixon and Davis, 1991) in any systematic analysis given that they may behave as many states simultaneously. Any cladograms favored may contain multiple sets of results, each with a different combination of effective values for the missing cells. No real observation could behave this way. Furthermore, since no sensible treatment of missing data can add length to a cladogram, data sets with missing data will yield cladograms which cannot be longer and may well be shorter than any actual observation could generate. It is this property which affects the fixed-character-state approach.

The problem does not come from entirely missing data—in the sense that an entire contiguous piece of DNA is unavailable, which can be treated in the usual fashion as missing data. When a fraction of a contiguous piece is missing, however, the problem occurs. If a state, i.e., a particular sequence, contains half unknown bases (string of Ns or Xs—N or gap), the state transformation matrix will be non-metric. In other words, a sequence with a large fraction of missing elements will appear to be most similar to all the states. This will create a violation of the triangle inequality (non-metricity) among the sequence states and make the results unintelligible. The most likely result will be a meaningless optimization with this half-missing sequence character optimized to each node as a hypothetical ancestor (Fig. 5). This problem has been noted with respect to the treatment of sequence indels as missing data earlier (Wheeler, 1993). This version of the problem is an expression of the same analytical shortcoming. The solution to this context would be to break up the fragment into those pieces (i.e., more characters) that are observed and those that are truly missing. This will not allow the missing states to be trivially optimized to internal nodes at zero cost.

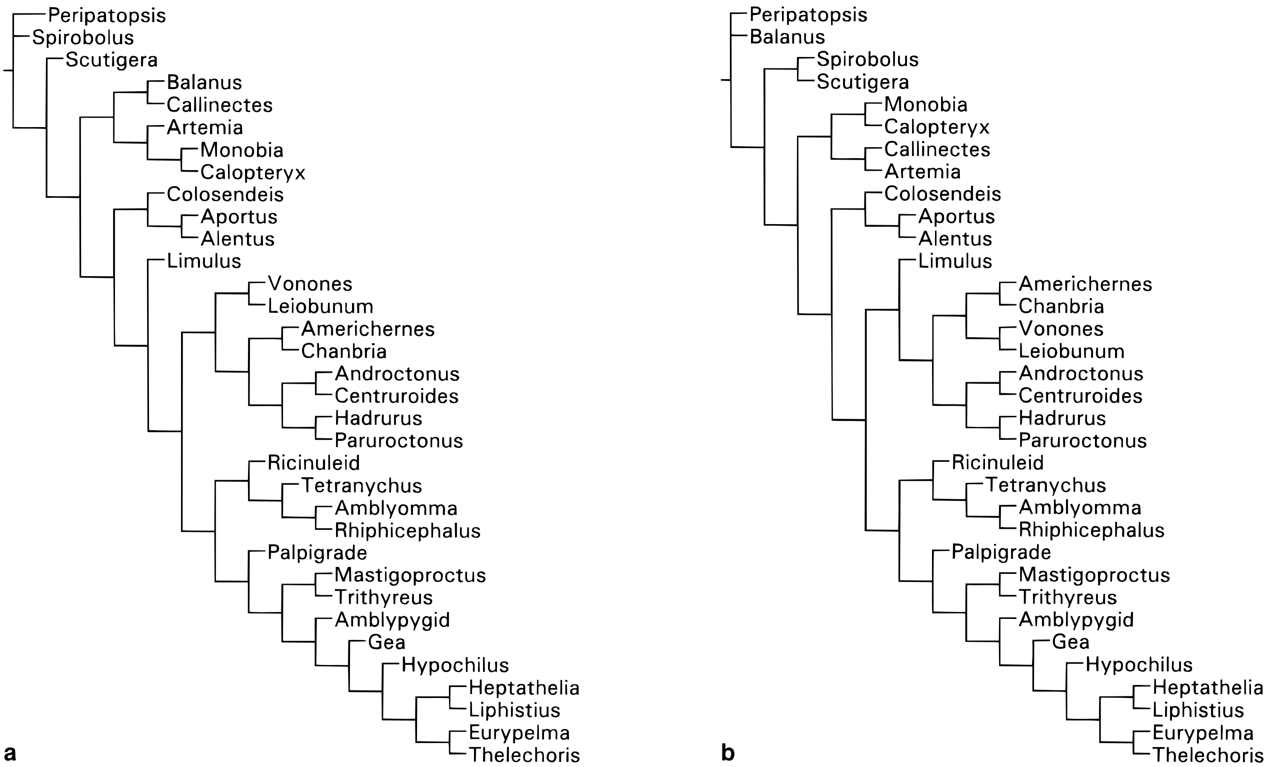

FIG. 3. Total evidence cladograms of chelicerates based on (a) optimization-alignment (Wheeler, 1996) and (b) fixed-character state optimization. In both of these schemes morphological transformations were given a cost of 4, indels 4, transversions 2, and transitions 1. The data are those of Wheeler and Hayashi (1998).

DIFFERENCES FROM OTHER TECHNIQUES

There are several unique features of this mode of analysis. Two of these are directly derived from the restricted set of possible HTU character state reconstructions. The first of these is that additional sequence information can actually decrease cladogram length in some situations. This can occur because additional terminal taxa increase the range of possible HTU states. New states can suggest a more parsimonious optimization than could have occurred before. Consider sequences I, II, and III of Figs. 1 and 2. By themselves these three sequences (given the transformation matrix specified in Fig. 2) require a cost of 10 weighted steps. A standard analysis would require only 8 weighted steps since it is freer to put bases at internal nodes. The addition of a single sequence—TTTT—allows new possibilities at the internal nodes resulting in a more parsimonious reconstruction of length 8 (Fig. 6). This is not so odd when one considers that more observations of variation imply more possible variation.

A second feature of this method is that "impossible" ancestors cannot be posited at internal nodes. Since the terminal sequences (i.e., those that are observed) define the states for the HTUs, stop codons, impossible secondary structures, and other sequence oddities cannot be generated (unless they occur in the terminals—in which case they are hardly impossible).

APOMORPHY AND SYNAPOMORPHY SCHEMES

Since the observed sequences become character states, all the usual sorts of operations can be performed to determine synapomorphy schemes, apomorphy lists, branch lengths, and reconstructed ancestral states (see Table 1). Support measures such as

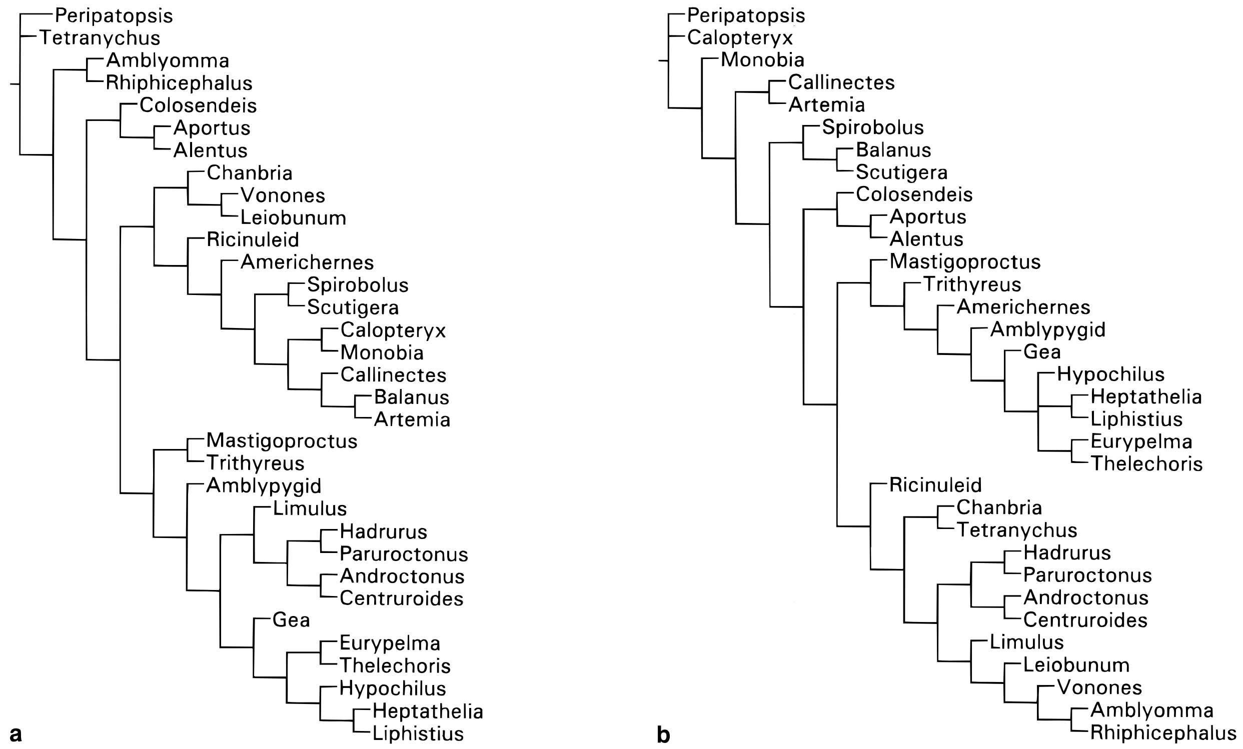

FIG. 4. Molecular evidence (18S + 28S rDNA) cladograms of chelicerates based on (a) optimization alignment (Wheeler, 1996) and (b) fixed-character-state optimization. In both of these schemes indels were given a cost of 2, transversions 1, and transitions 1. The data are those of Wheeler and Hayashi (1998).

Bremer support can be calculated, jackknife and bootstrap frequencies determined, and implied weight applied. These new sequence characters behave as any complex character does and allow any sort of common analysis.

CONCLUSIONS

The treatment of actual sequences as character states has many implications for the definition and interpretation of characters in general. Foremost among these

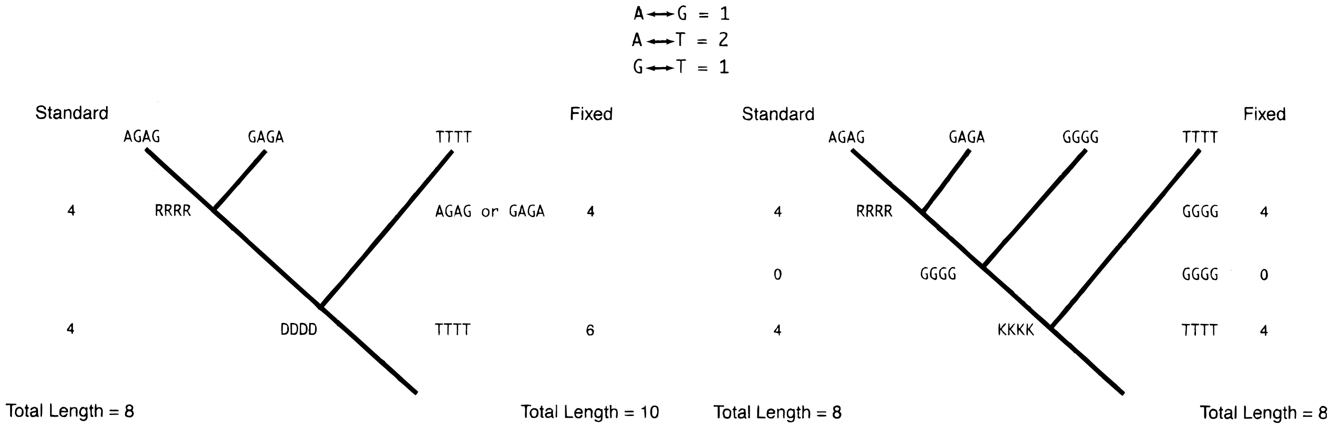

FIG. 5. The effect of optimizing states with a high fraction of missing values to internal nodes, spuriously decreasing cladogram length from 5 to the impossible 4.

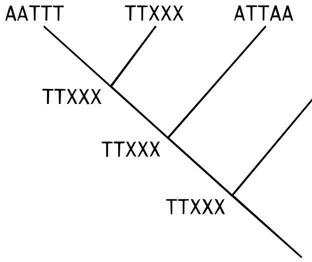

FIG. 6. The effect of adding taxa to an analysis which increases the ancestral character state set. In this case, cladograms may decrease in length with additional taxa by reducing state restrictions on ancestral sequence reconstruction.

effects are the delimitation of sequence characters and the notion of homology this demands.

Since contiguous strings of nucleotide bases or amino acids must be defined externally by the investigator, an element of subjectivity is injected into the analysis (unless one locus equals one character). The overriding principle in delimiting these characters (as with all characters) is the notion of independently varying units, i.e., those fragments which will not be logically determined by other units. An example from arthropod morphology would be the compound eye. At the level of arthropods, this structure comes and goes as a unit. The feature itself is very complex and could be broken down into many constituent characters. These related characters would, however, vary rigidly in step (this might not be the case at lower levels). Although an investigator might describe many features that comprise the compound eye, these multiple features would be logically linked and not vary independently. However, the observer defines the characters. Fixed-state sequence characters can be treated in the same way. For example, ribosomal sequences might be broken into areas of secondary structure and complex loci broken into intron and exon regions, or characters might be delineated by primers. In all of these cases, units are defined by the observer as varying independently (at least potentially) and are treated as independent evidence of history.

Since base-to-base homology schemes are never constructed in fixed-state optimization, only transformation costs between sequence pairs, the fundamental homology statement is at the level of the sequence itself. The contiguous sequences are the homologous units that transform at prescribed costs among various states. The notion of base-to-base homology is rejected at the level of the cladogram and is rechanneled into the determination of sequence level character similarity. The bases determine the relative adjacency (Wheeler, 1990) of the states; the sequences as a whole determine the cladogram.

TABLE 1

Topology of Cladogram to Diagnose at Cost 14: (I (II (III IV))) Hypothetical Ancestral Nodes

Node	Branch length[a]	Description
I	[0–4]	Terminal
II	[0–4]	Terminal
III	[0–0]	Terminal
IV	[4–4]	Terminal
HTU0	[Root Node] =	((II (IV III)) I)
HTU1	[6–6] =	(IV III)
HTU2	[0–4] =	(II (IV III))

Character change list

Nodes			States		
Anc	Desc	Character	AncS	DescS	Definite[b]
HTU0	HTU2	[0]	{0 1}	{0 1}	
HTU2	II	[0]	{0 1}	1	
HTU2	HTU1	[0]	{0 1}	2	*
HTU1	IV	[0]	2	3	*
HTU1	III				
HTU0	I	[0]	{0 1}	0	

Hypothetical node	Ancestral character states	
	[Character number]	State
HTU0	[0]	{0 1}
HTU2	[0]	{0 1}
II	[0]	1
HTU1	[0]	2
IV	[0]	3
III	[0]	2
I	[0]	0

[a] Minimum and maximum branch lengths.
[b] Optimization-independent change.

This notion of analysis cannot be tested by reference to historical “accuracy” in any way. This notion of homology determines the rules of analysis and could be compared to other parsimony methods (multiple sequence alignment, optimization alignment) by appeal to congruence whether at the level of characters or topology (Wheeler, 1995). A reasonable method should yield consistent, empirically supported results.

This type of comparison should illuminate the strengths and weaknesses of this approach.

Given the diversity of molecular character states implied by this fixed-state approach, the chances of random character state matching would seem to be extremely small. In fact, with a simple inverse relationship between the number of states and the probability of "random attraction," the chance of any occurrence of random matches would decrease with the number of taxa (i.e., states). As we gather more data, the probability of random similarity should decrease to marginal levels.

Along with the transfer of sequence data homology from the level of the nucleotide base to that of the sequence string itself, this method offers a certain simplicity which is appealing. Instead of long strings of bases, aligned or directly optimized with sequence gaps and a range of computation complexities and impossibilities, we are presented with much smaller numbers of characters with very informative state relationships. As sequencing technology improves, multiple DNA sequences are routinely determined for phylogenetic studies and these sequences are more and more frequently analyzed in combination with morphological data. This fixed-character-state method more firmly allies the analysis of morphological and molecular data, assuaging the most pernicious fear of sequence data—the monotonous repetition of four states.

Here, the loci become the characters.

ACKNOWLEDGMENTS

I thank James Carpenter, Darryl Frost, Gonzalo Giribet, Daniel Janies, Norman Platnick, Lorenzo Prendini, and Mark Siddall for comments on the manuscript and Steven Farris and Pablo Goloboff for discussion and abuse. I also thank Portia Rollins for original illustrations.

REFERENCES

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

Goloboff, P. A. (1994). Character optimization and calculation of tree lengths. *Cladistics* **9,** 433–436.

Goloboff, P. A. (1998). Tree searches under Sankoff parsimony. *Cladistics* **14,** 229–238.

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.

Nixon, K. N., and Davis, J. I. (1991). Polymorphic taxa, missing values, and cladistic analysis. *Cladistics* **7,** 233–241.

Platnick, N., Griswold, C. E., and Coddington, J. A. (1991). On missing entries in cladistic analysis. *Cladistics* **7,** 337–343.

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

Wheeler, Q. D. (1990). Ontogeny and character phylogeny. *Cladistics* **6,** 225–268.

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

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

Wheeler, W. C., and Hayashi, C. Y. (1998). The phylogeny of the chelicerate orders. *Cladistics* **24,** 1–20.