Cladistics 17, S3–S11 (2001)
doi:10.1006/clad.2000.0154, available online at http://www.idealibrary.com on IDEAL®

Homology and the Optimization of DNA Sequence Data

Ward Wheeler

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

Accepted December 12, 2000

Three methods of nucleotide character analysis are discussed. Their implications for molecular sequence homology and phylogenetic analysis are compared. The criterion of inter-data set congruence, both character based and topological, are applied to two data sets to elucidate and potentially discriminate among these parsimony-based ideas.

INTRODUCTION

Sequence data do not come in neat packages. Unlike most non-sequence character data sets, the series of appropriate comparisons to be made among observed variants (characters) is not clear. In short, when comparable sequences vary in length, it is not obvious which nucleotides to compare. The same A's, C's, G's, and T's are found in long strings in all the taxa in all positions. Obviously, the choice of which nucleotides to compare is fundamental to systematic analysis and will have serious impact on the historical scenarios we examine. The process of transforming nucleotide observations into homology statements and cladograms involves choices of both the means of optimizing homology and the entities that are to be homologized in the first place. Currently, there are three means of parsimoniously diagnosing sequence characters: multiple sequence alignment followed by standard analysis (Higgins *et al.*, 1996; Wheeler and Gladstein, 1994), optimization alignment (Wheeler, 1996), and fixed-state optimization (Wheeler, 1999b). Each of these methods creates homology statements and parsimoniously diagnoses them on cladograms. Multiple sequence alignment precedes cladogram analysis, while the other two methods do not "preprocess" the data, proceeding directly to cladogram optimization.

HOMOLOGY ASPECTS

The homology notions embedded in these three analysis methods can be distinguished in two ways. Multiple alignment and fixed-state analysis rely on putative homologies that are invariant—static—and each candidate cladogram is diagnosed based on the same set of putative homologies. The optimization alignment method, however, creates potentially unique homologies for each historical hypothesis. Another means of distinguishing among the methods is by treating the entities as comparable. In the case of multiple alignment and optimization alignment, the individual nucleotides are potentially homologous, as opposed to fixed-state analysis, which treats entire sequence fragments as characters.

0748-3007/01 $35.00

MULTIPLE SEQUENCE ALIGNMENT

Multiple sequce alignment (MSA) is a procedure to turn unequal length sequences into equal length character strings via the insertion of gaps. These gaps ("–") are place holders that signify that an insertion or deletion has occurred somewhere, resulting in a lack of homologous nucleotides at that position in that taxon. Most alignment procedures attempt to minimize some cost (e.g., evolutionary length) or maximize some benefit (e.g., overall similarity). While the procedure to determine a minimum cost alignment of two sequences is well understood and easily accomplished (Needleman and Wunsch, 1970), the extension to multiple sequences becomes difficult rapidly. Briefly, in order to align two sequences, a matrix of $(N + 1) \times (M + 1)$ cells is created and the minimum cost path through this matrix given specific cost to base transformations and insertion–deletion events is obtained. The matrix is traversed in such a fashion that only the adjacent three cells (usually the cells above, to the left and diagonally up, and to the left) are examined to determine the cost of each cell and the most efficient path to it (Needleman and Wunsch, 1970). This means that for each of the $N \times M$ cells, three cells are involved in the calculation of every other internal cell. While this is manageable for two sequences (the cost of computation being roughly proportional to the product of the sequence lengths) and significant shortcuts are known, extension to phylogenetically interesting numbers of sequences is extremely time-consuming. The alignment matrix for "n" sequences would have "n" axes, and each cell would require knowledge of $2^n - 1$ other cells. Furthermore, while the cost of spanning two sequences is simply the summed difference, when four or more sequences are involved, some tree search or prior knowledge is required to determine the alignment and its overall cost (Sankoff and Cedegren, 1983). These complicating factors have made true multiple alignment unachievable for anything but the smallest numbers of taxa. Real data sets require heuristic MSA solutions.

The heuristic thrust for multiple alignment is quite simple. Since aligning two sequences is easy, build the multiple alignment out of a series of pairwise alignments guided by a binary tree. All the multiple alignment procedures in common use today follow this idea, but differ in how they obtain the binary "guide" tree and how they add the pairwise results together to generate the complete alignment. Three implementations of heuristic alignment algorithms that are in some use today are CLUSTAL (Higgins and Sharp, 1988; Higgins *et al.*, 1992, 1996; Higgins, 1994; Jeanmougin *et al.*, 1998; Thompson *et al.*, 1994, 1997), TREEALIGN (Hein, 1989, 1990), and MALIGN (Wheeler and Gladstein, 1994, 1991–1998). Each of these methods relies on guide trees to accrete pairwise alignment. In the cases of CLUSTAL and TREEALIGN, a distance tree is calculated from the pairwise alignment costs and this distance tree becomes the guide tree. In the case of CLUSTAL, this is a Fitch–Margoliash tree; TREEALIGN uses a method developed by Hein (1989, 1990). At the nodes (vertices) of the guide trees, consensus (CLUSTAL) or quasi-optimized (TREEALIGN) single sequences are created from the aligned pair, which is then submitted to another pairwise alignment further down the tree. When the root of the guide tree is reached, the various gaps inserted on the way down are placed into the sequences at the tips, creating sequences of equal length—the multiple alignment.

MALIGN also uses guide trees, but differs from CLUSTAL and TREEALIGN in that it examines multiple guide trees. These guide trees are generated through standard tree search procedures of tree building and branch swapping. Furthermore, no individual sequences are created at the internal vertices, but the sequences descending from that node are carried along and aligned in a modified pairwise manner. During the search procedure, a complete multiple alignment is generated for each candidate guide tree, and a heuristic phylogenetic search is performed on the multiple alignment. That alignment (or alignments if multiple solutions are found) that produces the most parsimonious (i.e., lowest cost) phylogenetic result is chosen as the "best" multiple alignment. As a result of this search procedure, MALIGN will frequently examine many thousands or millions of candidate alignments (usually about n^3 for n sequences). Not surprisingly, CLUSTAL and TREEALIGN frequently generate results more rapidly than MALIGN.

Each of these methods, as an implementation of multiple alignment, yields the same type of result—a series of column vectors that are then submitted to phylogenetic analysis. When phylogenetic analysis takes place, the putative homologies are not altered or reexamined

in any way. The alignment-generated homologies are created a priori and are never revised. In this sense, they are static. Since these homologies exist at the level of the nucleotide position, they are "base-to-base" indicating the independence of transformation at each position.

PROBLEMS WITH MULTIPLE ALIGNMENT

Since multiple alignment creates putative homology schemes without reference to the diagnosis of any specific phylogenetic topology, it may be that some alignments are more favorably disposed to certain topologies than others. Consider four sequences, "GGGG," "GAAG," "GGG," and "GAA," and a cost regime of 2 for indels and 1 for base substitutions of all flavors. If we were to use the alignment

I G G G G
II - G G G
III G A A G
IV - G A A

the topology ((I III)(II IV) is optimal, with a length of six steps. The topology linking I and II [(I II) (III IV)] would require seven steps and [(II III)(I IV)] eight steps. However, suppose we had based our analysis on the following alignment:

I G G G G
II G - G G
III G A A G
IV G A A -.

In this case, topology [(I II)(III IV)] is optimal at six steps and the other two topologies would require seven steps. Furthermore, a third alignment,

I G G G G
II G G G -
III G A A G
IV G A A -,

would require six steps for both [(I II)(III IV)] and [(I III)(II IV)] with eight steps for [(I IV)(II III)].

Each of these three alignments appears reasonable and can achieve a most parsimonious solution. If we had used either alignments 1 or 2, however, we would not have missed a solution. These alignments would have been specified before the analysis, and no provision could have been made to allow individual topologies to be diagnosed with their best-case alignment. In short, cladograms were determining final homology (i.e., synapomorphy), but had no role whatsoever in putative homology.

OPTIMIZATION ALIGNMENT

A response to the lack of interaction between topology and putative homology is optimization alignment (OA; Wheeler, 1996). This method of directly optimizing sequence variation creates a unique set of putative homologies (base correspondences) for each topology. Not only are nucleotide changes and insertion–deletion events minimized, but the base correspondences themselves are chosen to minimize tree length. The goal of the process is to yield parsimonious cladograms in terms of the weighted sum of nucleotide transformations and insertion–deletion events. Like alignment, homologies are created at the level of nucleotide position or base to base. Unlike alignment, however, these homologies are dynamically determined and uniquely tailored to each topology.

The procedure is based on the heuristic determination of hypothetical ancestral sequences. The algorithm begins with an internal node (or vertex) with two terminal observed sequences as descendants (Fig. 1). The most parsimonious hypothetical ancestral sequence is

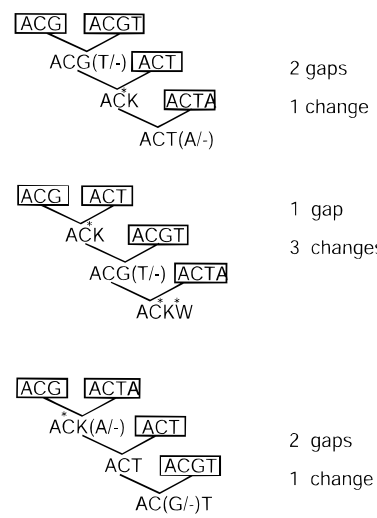

FIG. 1. Optimization alignment process for four sequences and a given topology.

created for this node via a dynamic programming procedure whereby all ancestral sequences are implicitly examined. This is accomplished in a manner very close to a pairwise alignment, except that the cost to be minimized is the weighted union/intersection cost usually used to diagnose phylogenetic hypotheses and "gaps" are optimized out since real sequences do not contain unambiguous gaps. This is explained in more detail elsewhere (Fig. 2; Wheeler, 1996, 2000, 2001). A second difference is that the product of this operation is a single sequence and the cost of producing that sequence from its descendants. This sequence is then used as one of the descendants of its parent node, and the operation is repeated. The weighted sum of all the events required to form the nodal sequences is the cost of the cladogram as with other Sankoff-style character optimizations. After this "down-pass," an "up-pass" can be performed to determine the most parsimonious set of character states at each position for each hypothetical ancestor.

In the case of the three sequences mentioned above, OA yields both of the length-6 cladograms directly and the alternate patterns of homology are presented as alternate synapomorphy schemes.

FIXED-STATE OPTIMIZATION

Even with the dynamic homology schemes created via OA, considerable confusion may remain with "gappy" sequences of greatly unequal length. Situations frequently occur in which the members of one group (or individual sequence) have hundreds of nucleotides while others seem to have none or just a few. In these situations, huge numbers of insertion–deletion events need to be postulated (they must go somewhere) and the placement of these gaps can seem haphazard. Furthermore, these indels must be accounted for over several ancestral nodes of the cladogram (OA) or in all taxa (multiple alignment), causing considerable damage downstream.

The motivating force behind this behavior is the treatment of the individual nucleotides as potential homologies. If the entire fragment were to be treated as a unified structure, as with a large complex morphological feature, the extreme apomorphy of such a sequence would not spill over into other areas of comparison. This is the idea behind fixed-state character optimization (FSO; Wheeler, 1999b). FSO treats contiguous strings of nucleotides as character states in a complex character defined by the extent of homologous regions of nucleotide sequence. In this way, the entire 18S rDNA might be a character, with the actual sequences exhibited by individual taxa as the character states. Each unique sequence defines a character state. Potentially, there are as many character states as taxa. This would seem to be uninformative, and it would be if all the transformations among all the states were taken to be equally costly. This is not the case, however. The transformation costs among the states are determined by the weighted sum of the various events required to transform one sequence into another, i.e., the minimum edit cost in indels, transitions, and transversion (or any other specified transformation types). These edit costs define a matrix of pair-wise transformation costs among the states (Fig. 3). The optimization step then proceeds via a standard dynamic optimization (Sankoff-style) procedure for a multistate character with predetermined transformation costs (Sankoff and Rousseau, 1975).

Since the states of the character are defined by the

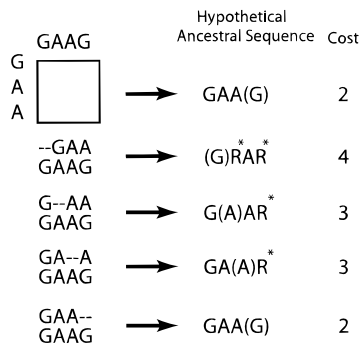

FIG. 2. Determination of hypothetical ancestral sequences during the "down-pass" of optimization alignment.

	I	II	III	IV
I	-	2	2	4
II	2	-	3	2
III	2	3	-	2
IV	4	2	2	-

FIG. 3. Transformation cost matrix for four sequences with an indel cost of 2 and a nucleotide substitution cost of 1.

actual sequence variation observed, the range of hypothetical ancestral states (=sequences) is limited. Arbitrary combinations of nucleotides are not permitted; hence the cladogram length will in most cases be longer. This phenomenon will become less pronounced with increases in the numbers of taxa, which can increase the number of possible nodal states. Furthermore, "impossible" secondary structure, stop codons, etc. cannot occur, since only observed sequences are reconstructed at the internal nodes.

CONGRUENCE AND COMPARISON

Although each of these three methods of optimizing sequence characters is based on parsimony, the entities that are minimized differ. MSA seeks to minimize global, static nucleotide changes; while optimization seeks to do this also, it allows for topology-specific putative homology schemes. These two methods, however, are minimizing the same combinations of base changes and indels, but in different ways. FSO seeks to minimize the overall change (in nucleotide substitutions and indels) as well, but in the form of transformations among a limited set of extremely complex character states. If the actual lengths of cladograms presented by these methods are not directly comparable, how can meaningful choices be made among these techniques?

Congruence among data is the fulcrum of parsimonious phylogenetic analysis. The cladogram indicating the greatest congruence among characters (most parsimonious) is the least contradicted and the most favored. This logic can be extended to among-data set comparisons. Methods that result in greater congruence among data sets are more desirable than others. This congruence may be measured in two ways: character based and topology based. Character congruence seeks to assay the level of homoplasy incurred through the combination of data, usually measured by the ILD (Mickevich and Farris, 1981) and expressed as a percentage of combined cladogram length due to among-data set homoplasy. Topological measures of congruence attempt to describe the degree of agreement among the results of phylogenetic analyses. The similarity of cladograms generated from several data sets

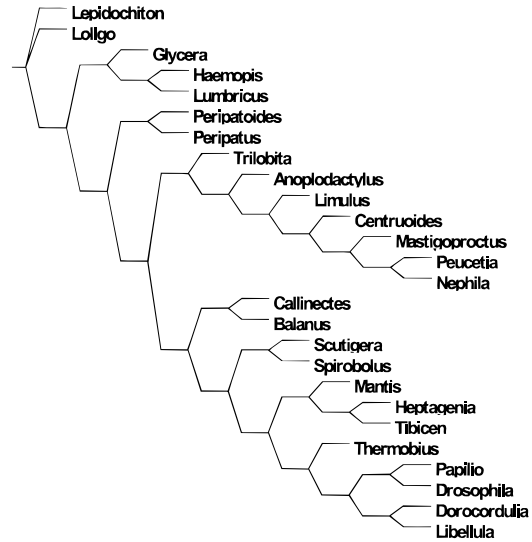

FIG. 4. Cladogram of arthropod relationships of Wheeler (2001) with indel cost of 2 and transitions and transversions 1.

TABLE 1
Taxon List for the Analysis of Arthropod Relationships

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)
Chelicerata	
Pycnogonida	*Anoplodactylus portus*
Xiphosura	*Limulus polyphemus*
Scorpiones	*Centruroides hentzii*
Uropygi	*Mastigoproctus giganteus*
Araneae	*Nephila clavipes*
Araneae	*Peucetia viridens*
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*

may be compared via some metric (e.g., Wheeler, 1999a).

Here we will apply both character and topological congruence measures to analysis of arthropod (Wheeler *et al.*, 1993; Wheeler, 1999a,b) and chelicerate (Wheeler and Hayashi, 1998) data sets to examine the relative performance of MSA, OA, and MSO analyses.

Each of the two test data sets (arthropod and chelicerate) consists of morphological and multiple molecular data sets. The character incongruence (as measured by ILD) and taxonomic incongruence (as measured by TILD) will be used to distinguish among these three approaches.

Arthropods

The arthropod data set is an enlarged version of the data set of Wheeler *et al.* (1993) used in Wheeler (1999a,b). This data set consists of 26 taxa, one of which, Trilobita, is extinct and "missing" for all molecular data. For the extant taxa, three sources of molecular information are used: 18S rDNA (an approximately 1000-bp fragment), 28S rDNA (an approximately 350-bp fragment), and Ubiquitin (228 bp). A morphological data set of 100 discrete characters for all of the taxa was also analyzed (Table 1 and Fig. 4).

Chelicerates

The chelicerate data set is that of Wheeler and Hayashi (1998). This data set consists of 34 taxa (one with morphology only), 93 morphological characters, 18S rDNA, and 28SrDNA fragments (Table 2 and Fig. 5).

MSA, OA, and FSO were each performed on these data sets. In the case of the arthropods, morphological character transformations and indels were assigned a cost of 2 and transversions and transitions 1. For the chelicerate case, all transformations (morphological, indels, transversions, and transitions) were assigned equal weight (1). Multiple alignment was performed using the program MALIGN (Wheeler and Gladstein, 1991–1998, 1994). This is not necessarily the best or certainly the only way to perform multiple alignment, but MALIGN attempts to create parsimonious alignments, those that generate parsimonious cladograms, and hence most appropriate to parsimony-based phylogenetic reconstruction. OA and FSO were accomplished through the use of POY (Gladstein and Wheeler, 1997).

TABLE 2
Taxon List for the Analysis of Chelicerate Relationships

Onychophora	
Peripatopsidae	*Peripatopsis capensis*
Chelicerata	
Pycnogonida	*Anoplodactylus portus*
	Anoplodactylus lentus
	Colossendeis sp.
Xiphosura	*Limulus polyphemus*
Scorpiones	*Centruroides hentzii*
	Androctonus australis
	Hadrurus arizonensis
	Paruroctonus meaensis
Araneae	*Hypochilus pococki*
	Gea heptagon
	Eurypelma californica
	Thelechoris striatipes
	Heptathela kimurai
	Liphistius bristowei
Palpigradi	sp.
Pseudoscorpiones	*Americhenernes* sp.
Solifugae	*Chanbria regalis*
Opiliones	*Vonones ornata*
	Leiobunum sp.
Acari	*Amblyomma americanum*
	Rhiphicephalus sanguineus
	Tetranychus urticae
Ricinulei	Ricinoididae sp. (juvenile)
Amblypygi	Amblypygi sp.
Thelyphonida	*Mastigoproctus giganteus*
Schizomida	*Trithyreus pentapeltis*
Crustacea	
Reptantia	*Callinectes* sp.
Anostraca	*Artemia salina*
Thoracica	*Balanus* sp.
Myriapoda	
Chilopoda	*Scutigera coleoptrata*
Diplopoda	*Spirobolus* sp.
Hexapoda	
Odonata	*Agrion maculatum*
Hymenoptera	*Monobia* sp.

In both cases, character congruence and topological congruence are maximized by FSO and minimized by MSA. The results of these analyses are summarized in Tables 3 and 4.

DISCUSSION

The central question raised by these modes of analysis is what are the units to be homologized? Since OA and MSA purport to optimize numerically identical

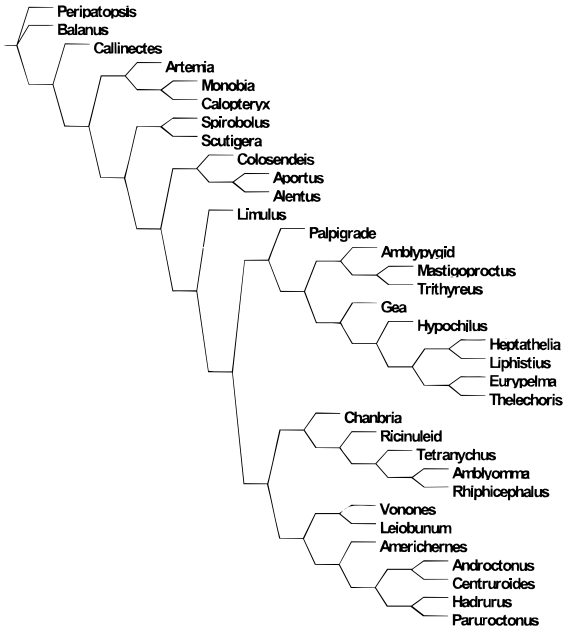

FIG. 5. Cladogram of chelicerate relationships from Wheeler and Hayashi (1998) with indel cost of 1 and transitions and transversions 1.

criteria, they must be viewed as different means of optimizing (Fig. 6) the same thing: the minimum weighted number of nucleotide transformations and indels. In these implementations, OA is superior in both cladogram length and topological congruence to MSA. If we accept for a moment the superiority of OA over alignment, the comparison to be made is between OA and FSO.

The cladogram lengths of these two methods are not directly comparable (as stated above). In the comparisons here, both character and topological congruence measures were higher for FSO. If congruence were the undisputed determinate of methodological superiority, the case would rely on the accumulation of such examples until a clear pattern presented itself. If the conclusions of the two data sets examined here held generally, FSO would be the way to go and sequence level homology would be preferred over nucleotide-level comparisons. This is not necessarily the case, however (see Edgecombe *et al.*, 1999 for a counter example).

The choice of comparable or homologous entities is also one of epistemology. How do we perceive variation, information, and transformation among variable sequences? The choice can be (and perhaps should be) made on grounds completely unrelated to the behavior of any data. The choice can be made on an ontological or even an aesthetic basis. On the other hand, an operational or optimality-based perspective is also possible. Which method exhibits the best numerically defined behavior (such as congruence, however measured)?

I have no answer to this question. I believe strong arguments can be made from both views. We can say now, however, that these different notions of homology in sequence data affect historical hypotheses and are fundamental to our choice among competing scenarios.

ACKNOWLEDGMENTS

I acknowledge the contributions of Steve Farris, Arnold Kluge, Daniel Janies, Gonzalo Giribet, James Carpenter, Norman Platnick, and Randall Schuh for discussing these ideas.

TABLE 3
Anthropod Congruence

	Combined data	Morphology	18S rDNA	28S rDNA	Ubiquitin	ILD or TILD
			Character congruence			
MSA	2123	252	503	919	387	0.0292[a]
OA	2007	252	501	848	392	0.00698[a]
FSO	2271	252	584	943	484	0.00352[a]
			Topological congruence			
MSA	85	14	19	12	8	0.18[b]
OA	75	14	19	10	8	0.14[b]
FSO	48	14	19	12	1	0.096[b]

[a] ILD.
[b] TILD.

TABLE 4
Chelicerate Congruence

	Combined data	Morphology	18S rDNA	28S rDNA	ILD or TILD
			Character congruence		
MSA	3027	201	1406	1217	0.067[a]
OA	2322	201	1082	966	0.031[a]
FSO	2470	201	1210	996	0.026[a]
			Topological congruence		
MSA	130	13	29	25	0.485[b]
OA	126	13	32	25	0.444[b]
FSO	93	13	23	21	0.387[b]

[a] ILD.
[b] TILD.

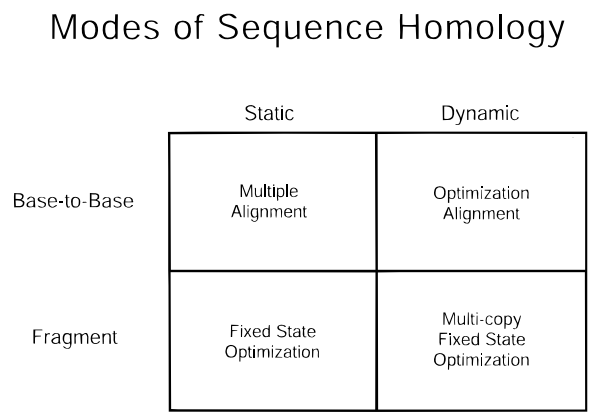

FIG. 6. Cartoon of the relationship among different homology notions and methods.

REFERENCES

Edgecombe, G. D., Giribet G., and Wheeler, W. (1999). Phylogeny of Chilopoda: combining 18S and 28S rDNA sequences and morphology. *In* "Evolución y filogenia de Arthropoda" (A. Melic, Ed.), pp. 293–331. Sociedad Entomológica Aragonesa, Zaragoza.

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

Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when a phylogeny is given. *Mol. Biol. Evol.* **6,** 649–668.

Hein, J. (1990). Unified approach to alignment and phylogenies. *In* "Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences" (R. F. Doolittle, Ed.), Methods in Enzymology, Vol. 183, pp. 626–644. Academic Press, San Diego.

Higgins, D. G. (1994). CLUSTAL V: Multiple alignment of DNA and protein sequences. *Methods Mol. Biol.* **25,** 307–318.

Higgins, D. G., and Sharp, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73**(1), 237–244.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992). CLUSTAL V: Improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**(2), 189–191.

Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266,** 383–402.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**(10), 403–405.

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

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.

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

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

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

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

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**(24), 4876–48821.

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

Wheeler, W. (1999a). Measuring topological congruence by extending character techniques. *Cladistics* **15,** 131–135.

Wheeler, W. C. (1999b). Fixed character states and the optimization of molecular sequence data. *Cladistics* **15,** 379–386.

Wheeler, W. C. (2000). Heuristic reconstruction of hypothetical-ancestral DNA sequences: Sequence alignment versus direct optimization. *In* "Homology and Systematics" (R. Scotland and R. T. Pennington, Eds.), pp. 106–113. Systematics Society, London.

Wheeler, W. C. (2001). Homology and DNA sequence data. *In* "The Character Concept in Evolutionary Biology" (G. P. Wagner, Ed.), pp. 303–318. Academic Press, New York.

Wheeler, W. C., and Gladstein, D. S. (1991–1998). Program and documentation. Documentation by Daniel Janies and Ward Wheeler.

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

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

Wheeler, W. C., Cartwright, P., and Hayashi, C. Y. (1993). Arthropod phylogeny: A combined approach. *Cladistics* **9,** 1–39.