Cladistics (1996) 12:1–9

OPTIMIZATION ALIGNMENT: THE END OF MULTIPLE SEQUENCE ALIGNMENT IN PHYLOGENETICS?

Ward Wheeler

Department of Invertebrates, American Museum of Natural History, Central Park West at 79th St., New York, NY 1024–5192, U.S.A.

Received for publication 10 December 1994; accepted 20 June 1995

Abstract — A method is described to assess directly the number of DNA sequence transformations, evolutionary events, required by a phylogenetic topology without the use of multiple sequence alignment. This is accomplished through a generalization of existing character optimization procedures to include insertion and deletion events (indels) in addition to base substitutions. The crux of the model is the treatment of indels as processes as opposed to the patterns implied by multiple sequence alignment. The results of this procedure are directly compatible with parsimony-based tree lengths. In addition to the simplicity of the method, it appears to generate more efficient (simpler) explanations of sequence variation than does multiple alignment.

Introduction

Phylogenetic analysis of nucleotide sequences presents problems not frequently found in other forms of character data. Although each base position presents one of four identical states (A, C, G or T), the number of these positions is likely to vary, that is homologous nucleotide sequences may differ in length. These differences may be small to non-existent in protein coding regions or extensive in non-translated sequences such as ribosomal DNAs and introns. This sequence length variation has lead to the development of procedures (multiple sequence alignment) to line-up these bases via the insertion of gaps. These alignment gaps allow the nucleotide base correspondences to be interpreted as putative homologies and phylogenetic analysis to proceed.

Many methods have been proposed to accomplish multiple sequence alignment either simultaneously or sequentially, but the process is computationally intensive and there is not yet consensus on what defines a “good” or “best” multiple alignment. Methods based on the dynamic programming algorithm of Needleman and Wunsch (1970) have generated alignments based on similarity (Feng and Doolittle, 1987, 1990; Higgins and Sharp, 1988, 1989), inferred phylogenetic order (Hein, 1989, 1990), external knowledge of phylogeny (Mindell, 1991), and parsimony (Sankoff and Cedergren, 1983; Wheeler and Gladstein, 1992, 1994). In each case, alignments are created with the expressed purpose of constructing historical schemes of evolutionary relationship.

Methods which use multiple alignment follow a general pattern of observation (sequence data) to alignment (insertion of gaps) to phylogeny reconstruction (parsimony or other method). During this process, sequence “gaps” are created and treated as fifth nucleotide base although they are not observations but rather

0748-3007/96/010001+09/$18.00/0

place-holders signifying a specific type of transformation event. Nucleotide bases are observable, gaps are not. Hence, a certain amount of logical inconsistency is introduced into the analysis since a process (insertion or deletion of bases) could be treated as a pattern (synapomorphy). The method proposed here avoids this problem by generalizing phylogenetic character analysis to include insertion/deletion events (indels). By doing this, analysis proceeds directly from the sequence data to phylogeny reconstruction, obviating the need to create gap characters. Indels appear not as states but as transformations linking ancestral and descendent nucleotide sequences.

The Method

The optimization procedure is a straight-forward generalization of non-additive or unordered optimization (Farris, 1970; Fitch, 1971). In strict non-additive optimization, all transformations receive equal weight and this is the means followed here (however, the generalization shown here can be extended to encompass transition–transversion bias or other modifications as with standard optimization techniques). In Fig. 1, the down-pass of the Fitch algorithm is illustrated for four simple sequences arranged on a phylogenetic tree. The process begins at the top and proceeds down through the tree creating each node in turn. In each case, node sequences (hypothetical ancestor state sets) are determined by finding first the intersection of the corresponding bases of its two descendants. If the bases are identical (A and A) or overlapping (A and R=A or G), the ancestral base is taken (initially) to be that intersection. If the intersection is empty (C and T), the union is assigned to the ancestral node. Each union operation requires a base transformation hence lengthens the tree. The scheme of Fig. 1 requires five base transformations.

Generalized optimization is depicted in Fig. 2. In this case, there are five sequences but they have unequal lengths. Without prior knowledge of the aligned base correspondences, it is impossible to construct a hypothetical ancestor or determine how costly that operation is (in terms of transformations). Hence, correspondences, or putative homologies, must be constructed as we go down the tree for the comparisons made at each node. In essence, all possible schemes of comparison must be examined for each node and that scheme which minimizes the number of union events (weighted by the cost of a base transformation) and indels (weighted by the gap cost) is assigned to the node. In this way, the most efficient (i.e. lowest cost) hypothetical ancestor is constructed. The implicit examination of the complete set of ancestors can be accomplished easily via dynamic programming.

As with the non-additive analysis, the procedure begins at the top of the tree (or more specifically at an ancestral node of two terminal taxa) with the sequences "ACG" and ACGT. The construction of the hypothetical ancestor can be broken down into two operations. The first can be thought of as an alignment step. The sequences are aligned to minimize the weighted cost of indels (gaps) and base transformations as determined by union/intersection counts. This is accomplished with the proviso that if a gap is inserted in one sequence to correspond to a gap in the other, this is done at no cost (the sequences would have a non-empty

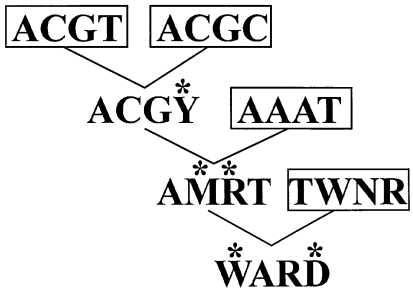


Fig. 1. Schematic of non-additive or unordered character optimization (Farris, 1970; Fitch, 1971). This topology of the aligned sequences requires five nucleotide transformations. Ambiguous hypothetical ancestor bases are represented by IUPAC codes (Y=T and C etc.). The asterisks (*) denote nucleotide changes. The boxed sequences are terminal taxa.

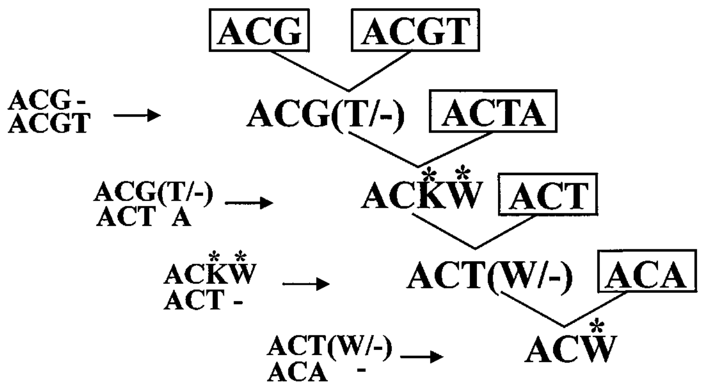


Fig. 2. Schematic of the optimization method presented here. The topology of the sequences requires three nucleotide changes and two indels. Symbols as in Fig. 1 with parentheses denoting indels.

intersection). Each possible alignment is considered (via dynamic programming) as in the Needleman and Wunsch (1970) procedure. The best alignment contains three matched bases and one gap. In the second operation, the hypothetical ancestor is constructed from this alignment by taking the union/intersection position by position along the sequence yielding "ACGT(T/-)". The ambiguity of the fourth position is derived from the indel required to reconstruct the ancestor and signifies that, as far as the down-pass is concerned, there may have been three or four bases in the ancestral sequence.

Proceeding to the next node, "ACG(T or -)" is compared to "ACTA". Here, the alignment step requires two nucleotide transformations yielding an ancestral sequence with two ambiguities denoting sequence differences. The first three bases are reconstructed as before as "ACK". The reconstruction of the fourth, however, is more complicated. Standard, non-additive optimization would allow each of the three possibilities "A, T or gap" in the fourth position. If the cost of an indel is equal to that of a base transformation, this method would do the same. If, however, indels are given a higher cost than nucleotide transformations, the possibilities are limited to the two bases "A or T". This is due to the existence of two classes

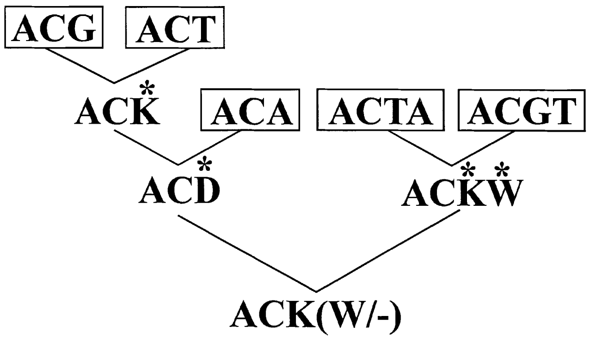


Fig. 3. Diagnosis of alternate topology. The topology of the sequences requires four nucleotide changes and a single indel. Symbols as in Fig. 2.

of events in this cost regime, indels and base transformations. In essence, each descendant presents a nucleotide base as a possible state for the ancestor whereas only one shows a gap as a possibility. The information common to the two descendants is that the ancestral sequence is likely to contain a base. When indel costs are not greater than base transformations, there are no such classes, hence no exclusion of the "gap" possibility. For the sake of this example, it is assumed that indels have a greater cost than base transformations and the ancestor is set to "ACKW".

The third node is reconstructed similarly to the first, yielding "ACT(W/-)" at a cost of one indel. The reconstruction of the next and final node shows another peculiarity of this analysis procedure. When "ACA" and "ACT(W/-)" are compared and combined to form the ancestor, it would appear that this sequence should be "AC(T/-)A". However, the correct ancestral reconstruction is "ACW", for two reasons. The first is the peculiarity of the alignment cost function. If gaps are inserted opposite positions which are ambiguously defined and contain gaps as a possibility, the insertion comes at no cost. This is because, at the later stage in the ancestral determination, these gaps would have a "-" intersection and would require no explanation (cost). The second source of this effect comes from the fact that sequences do not contain gaps. As stated before, gaps are hypotheses of transformation, not observations. Hence, the gap in the fourth position of "ACW-" is removed to yield "ACW". The ancestral sequence length is most parsimoniously reconstructed as three. This is logical when thinking of comparing the lengths alone. One descendent sequence has a length of three and the other of three or four. Hence, length three is common to both and simplest. An insertion or deletion event can still occur on the branch leading to a group. The feature defining the group, however, would be its length in that case, not a "gap".

The entire topology has been diagnosed at a cost of two indels and three base transformations. An alternate topology (Fig. 3) which grouped similar length sequences first, would be diagnosed with four base changes and a single indel. Although both topologies require five events the types of events required differ and regimes of indel and base change costs would effect their relative costs.

An "up-pass" could be performed on these topologies to determine the actual state assignments (sequences) of the nodes as with standard analysis favoring early,

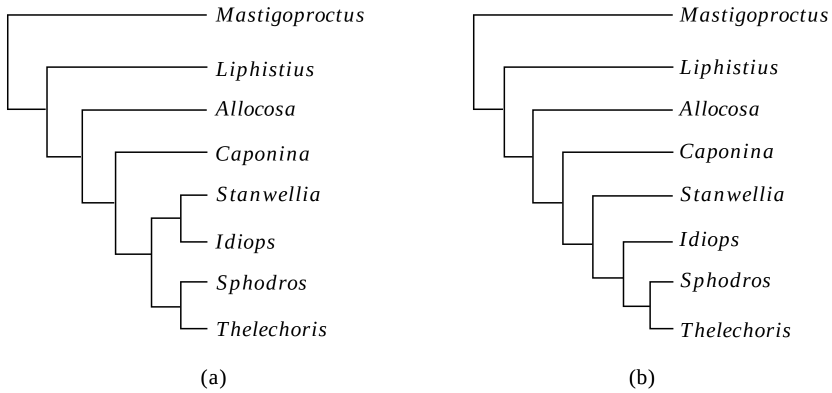


Fig. 4. Phylogenetic topologies of selected spider taxa based on 16s rDNA. (a) The arrangement derived from multiple-alignment required 540 nucleotide changes and 114 indels for a weighted length of 2418; (b) the arrangement derived from the method proposed here required 533 nucleotide changes and 109 indels for a weighted cost of 2362.

late, or any other scheme of transformations. The correspondences among the bases between ancestors and descendants would be those found during the down pass.

Concerns could be raised about ignoring or "overwriting" indels, this method not only allows but *requires* indel information to decide among competing topologies. The manner in which it does this, however, differs from standard methods. The indels are events which occur between nodes on the diagram, the lengths of the sequences are the synapomorphies which characterize the groups. No one would say that a transversion characterized a group. Only an observation (like a nucleotide "A" or "C") can characterize anything. The indels are treated like transitions or transversions, they link observed or inferred character states.

An Example

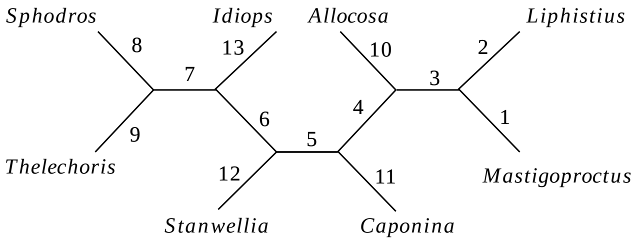
In order to demonstrate this method, 16S mt rDNA sequence data (approximately 400–450 bases) from seven spiders and an outgroup were analyzed. These taxa (*Mastigoproctus giganteus, Liphistius bristowei, Sphrodos abboti, Thelechoris striatipes*, *Idiops* sp., *Stanwellia* sp., *Caponina chilensis*, and *Allocosa* sp.) represent a small sampling of spider diversity. The programs MALIGN (Wheeler and Gladstein, 1992, 1994) and NONA (Goloboff, 1994) were used to perform the analyses.

Two situations were analyzed. In the first, the sequences were aligned and phylogeny reconstructed from the multiple alignment. The gap cost was arbitrarily set to seven and the nucleotide change cost to three. Multiple alignment was performed in a sequential pairwise manner determined by a branching diagram. At each node on the diagram, a pairwise alignment was performed. The basalmost node yielded the multiple alignment of all eight sequences. Since this process is order (topology) dependent all 135 135 alignment orders (rooting matters) for these sequences were performed. In each case, a branch-and-bound search was

Table 1
Branch Lengths (this=one Optimization of (potentially) many)

Branch		Insertions/deletions						Changes			Cost		
From	To	Leading/trailing			Internal			This	Minimum	Maximum	This	Minimum	Maximum
		This	Minimum	Maximum	This	Minimum	Maximum						
10	9	0	0	0	9	4	9	21	17	22	126	79	129
11	10	0	0	0	3	3	8	31	23	34	114	105	143
12	11	0	0	0	5	3	5	35	34	35	140	123	140
13	12	0	0	0	7	7	9	20	16	22	109	103	123
14	13	0	0	0	9	9	9	40	36	43	183	171	192
root	14	0	0	0	48	0	49	28	0	66	420	0	538
root	*Mastigo-pro*	0	0	0	1	0	49	37	0	113	118	0	538
14	*Liphistius*	0	0	0	2	2	2	63	60	67	203	194	215
9	*Sphodros*	0	0	0	3	3	3	56	53	58	189	180	195
9	*Thelech-oris*	0	0	0	5	5	5	40	38	43	155	149	164
10	*Idiops*	0	0	0	2	2	7	34	33	43	116	113	163
11	*Stanwellia*	0	0	0	4	4	4	37	34	40	139	130	148
12	*Caponina*	0	0	0	5	5	7	53	53	56	194	194	211
13	*Allocosa*	0	0	0	6	6	6	38	36	40	156	150	162

total length 2362 with 0 leading/trailing gaps, 109 internal gaps and 533 changes.



Root	Cost	Gaps	Changes
1	2362	109	533
2	2391	108	545
3	2384	107	545
4	2393	107	548
5	2398	109	545
6	2404	106	554
7	2393	104	555
8	2399	107	550
9	2408	107	553
10	2406	108	550
11	2412	108	552
12	2394	108	546
13	2413	106	557

Fig. 5. Variation in phylogenetic topology cost with the placement of the root. Unlike standard phylogenetic analysis, the costs of cladograms calculated by the method proposed here depend on the root position. The cost and number of indels and nucleotide changes required by each of the thirteen rootings of the eight taxon network are shown.

performed to determine the length of the most parsimonious cladogram for each of the multiple alignments. This “best” alignment yielded a tree with a weighted length of 2418 (Fig. 4a).

In the second analysis, the optimization regime proposed here was used to determine the optimal tree. Here as well, all rooted topologies for the eight taxa were analyzed (implicitly) and the best topology determined. This topology had a weighted cost of 2362 (Fig. 4b). Since there is no multiple alignment to examine, none is presented. Branch lengths and hypothetical ancestors are calculable and are shown in Table 1.

The derived topologies are similar, differing only in the status of *Idiops+Stanwellia*. The lengths of the topologies, however, differ by 2.4%. The multiple-alignment based cladogram required 114 indels and 540 nucleotide changes while the method proposed here yielded a cladogram requiring only 109 indels and 533 changes. The diagnosis of the best multiple-alignment derived tree by this method required 2389 steps (1.2% shorter).

Discussion and Conclusions

Although this method of optimization is an extension of currently used methods, the extradimensionality of this process yields areas which behave differently from standard methods. Two of these areas are rooting and the non-uniqueness of ancestors.

In standard analysis, the root placement has no effect on cladogram length or cost. Here, however, gaps can preclude base changes and vice versa, hence the position of the root can effect its length (cost). This is illustrated in Figure 5. Note that all rootings of the most parsimonious network have lower cost than that derived from multiple alignment (I do not know if this is a general trend or not, but if it is, large search efficiencies will be realized—a factor of 2n–3). I believe that this rooting "problem" is artifactual—related to the non-uniqueness of ancestors discussed below.

Another area of difference is in the non-uniqueness of ancestors. There may be several equally costly ancestral sequences for any pair of descendants. In the analyses performed here, alignments were performed to maximize contiguous gaps—MALIGN option "contig". There are potentially many locally equally optimal ancestors which may not be globally equal in cost.

A third difference comes in the area of assaying levels of homoplasy in the analysis. Since there are no base correspondence (characters) in the absence of a topology, simple CI or RI calculations cannot be performed. Analogues of these could be calculated by following the lines of correspondence through the ancestor-descendent relationships of the topologies. These lines would correspond to characters in standard analysis (completely if there were no gaps), but would begin and end with indel synapomorphies, respectively. While they endure, these lines would trace nucleotide changes in homologous positions.

As mentioned above, this method of optimizing features is a generalization of standard parsimony-based character analysis. In fact, if there are few or no sequence gaps, this method will yield numerically identical results. The extension of standard analysis to include indels allows a more efficient inference of phylogenetic pattern from sequence data than can be accomplished with multiple sequence alignment. In the case shown here, this method yielded cladograms which were more parsimonious than those generated by multiple sequence alignment coupled with parsimony analysis. I have not demonstrated this to be always true, however, I have never seen a situation where this was not the case.

Although the method behaves in novel ways in some situations, specifically rooting and non-uniqueness of ancestors, it appears to be more efficient at ferreting out parsimonious schemes of sequence transformation than other methods, and it does this entirely without multiple-alignment.

Acknowledgements

I would like to thank Rob DeSalle, Paul Vrana, Michael Whiting, Cheryl Hayashi, Daryl Frost, Mark Norell, Michael Novacek, Amy Litt, James Carpenter, Norman Platnick, Alfried Vogler, Andrew Brower, Joel Cracraft, Pablo Goloboff and Steven Farris for discussion and "encouragement."

REFERENCES

Feng, D. and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25: 351–360.

FENG, D. AND R. F. DOOLITTLE. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzomol. 183: 375–387.

FARRIS, J. S. 1970. A method for computing Wagner trees. Syst. Zool. 34: 21–34.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20: 406–416.

GOLOBOFF, P. A. 1994. NONA/Pee-Wee. Ver. 1.1. The American Museum of Natural History, New York.

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. Methods Enzomol. 183: 626–644.

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

HIGGINS, D. G. AND P. M. SHARP. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. Comput. Appl. Biosci. 5: 151–153.

MINDELL, D. 1991. Aligning DNA sequences: homology and phylogenetic weighting. *In*: M. J. Miyamoto and J. Cracraft (eds.). Phylogenetic Analysis of DNA Sequences. Oxford University Press, New York, pp. 73–89.

NEEDLEMAN, S. B. AND C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443–453.

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

WHEELER, W. C. AND D. G. GLADSTEIN. 1992. MALIGN: A Multiple Sequence Alignment Program. Program and documentation. Vers. 2.0. The American Museum of Natural History, New York.

WHEELER, W. C. AND D. G. GLADSTEIN. 1994. MALIGN: a multiple nucleic and sequence alignment program. J. Hered. 85: 417.