Chapter 5

Heuristic reconstruction of hypothetical–ancestral DNA sequences: sequence alignment vs direct optimization

Ward Wheeler

Introduction

The problem of historical reconstruction of nucleic acid sequences can be reduced to one of the determination of ancestral (i.e. nodal) sequences. The composition of these hypothetical sequences determines the length and shape of any cladogram. The efficient or parsimonious placement of these nodes is the fundamental systematic act.

Nucleic acid (and protein) sequence data present problems not normally seen in other types of comparative data. The main distinction is that the terminal taxa do not present the same number of features. This, coupled with the limited number of character states, causes the correspondences among features to be undetermined. In other words, the putative homologies among the sequence bases are unknown. Most anatomical data do not have this problem. A feature of a vertebrate forelimb will never be compared to the retinular structure of the eye or even a superficially similar hind limb. The positional information implicit in the character definition and the effectively infinite number of states leaves no room to confuse character states among characters (the case described by Whiting and Wheeler (1994) is presumed to be rare).

The normal course of events for the systematic analysis of sequence information begins with some form of multiple sequence alignment. Although some workers still argue for alignments 'by eye', this technique clearly is not based on any quantitative optimality criterion, is highly subject to preconceived notions of relationship, and is non-reproducible. Algorithmically (i.e. computer-) generated multiple alignments may not be without fault, but at least they can be reproduced by other investigators. These alignments embody the putative synapomorphies on which standard phylogenetics relies. After alignment, character reconstructions may be completely isomorphic (unordered or non-additive) or specified in great detail with step matrices, and may even include asymmetrical character transformation costs.

A method has been proposed recently which seeks to optimize the sequence characters directly, without the intervening alignment step (Wheeler, 1996). The thrust of this idea is the generalization of character optimization techniques to allow for terminals with unequal numbers of characters. Existing optimization techniques such as those of Farris (1970), Fitch (1971) and Sankoff and Rousseau (1975) all require equal numbers of characters among all taxa. Once optimization has been generalized to accommodate sequence length variation, multiple sequence alignment is unnecessary.

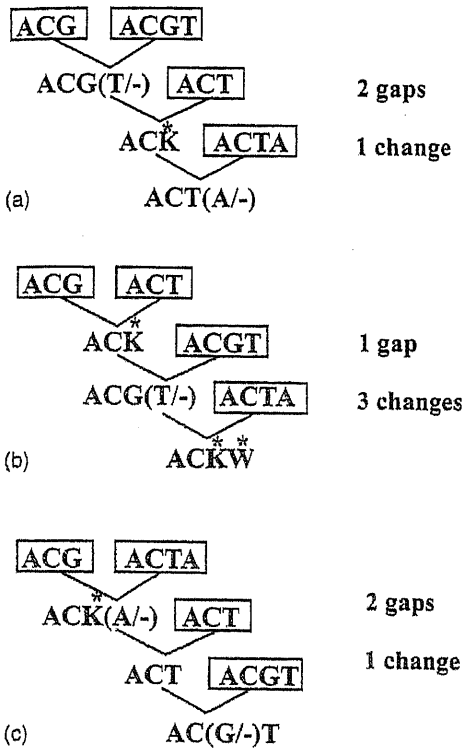

Figure 5.1 Examples of direct optimization (optimization–alignment).

In order to compare these two methods (sequence alignment and direct optimization), some measure must be agreed upon to assess their relative merits. Given the unknowability of 'truth' and the problems of simulation, congruence is the most appropriate means of comparing methods. Certainly consistency of results over multiple sources of data is a desirable property of any historical method. There may be others, but for the purpose of this discussion, consideration is limited to the congruence of systematic results in the face of diverse phylogenetic information.

Here, the methodologies of multiple sequence alignment and direct optimization are discussed and compared empirically with respect to character congruence as measured by the Mickevich and Farris (1981) (MF) incongruence length difference in three real data sets.

Direct optimization/optimization alignment

A simple restatement of the procedure described earlier (Wheeler, 1996) is shown in Fig. 5.1. In this example, four simple sequences are diagnosed on three cladograms. As I have pointed out elsewhere (Wheeler, 1998), this algorithmic procedure is a heuristic determination of minimum tree length. The exact case is computationally prohibitive (within current thinking). Since the heuristic method is greedy

and short-sighted, not examining all the global possibilities, the tree length returned by this algorithm is an upper bound on the minimal length.

The essence of the method is the determination of efficient hypothetical ancestral sequences at each internal node. In short, this down-pass sequence is calculated as the sequence, which requires the minimum change to its two descendants. In practice, the sequence is calculated via dynamic programming (as with the Needleman-Wunsch (1970) alignment algorithm), with the exception that the procedure minimizes the sum cost of transformation from the nodal sequence to its descendants. The sum cost will depend on the relative gap cost and any transition:transversion bias or more complex character model. Where there are several minimal length base (or indel) assignments possible, ambiguities are noted in the nodal sequence. The length of the cladogram is determined by summing the cost of the creation of each of these nodes until the cladogram is complete. The node sequences created during the down-pass are not the final hypothetical ancestral sequences. As with standard analysis, an up-pass is required to get the final character states and branch lengths. The final nodal sequences are created to minimize distance among the three nodal sequences (final ancestor and two preliminary descendants) connected to it.

In the case of Fig. 5.1, not only optimization but also the choice of optimal topology would be affected by differential gap and change cost. If indels were relatively expensive compared to base substitutions, topology 'b' would be minimal, whereas 'a' would be favoured with relatively expensive base changes.

Data

The three data sets discussed here are derived from literature sources. They are: (1) Chelicerata – 34 taxa with 93 morphological characters, approximately 1000 bp 18S rDNA, and approximately 350 bp 28S rDNA (Wheeler and Hayashi, 1998); (2) Carnivora – 39 taxa with 265 bp mt cytochrome b, approximately 210 bp mt 12S rDNA, and approximately 310 bp mt 16S rDNA (based on Vrana *et al.*, 1994); and (3) Hemiptera – 21 taxa with approximately 780 bp 18S rDNA, approximately 350 bp 28S rDNA, approximately 570 bp mt cytochrome oxidase I, and approximately 400 bp mt 16S rDNA (based on Wheeler *et al.*, 1993).

Methods

Multiple sequence alignment

The program MALIGN (Wheeler and Gladstein, 1992, 1994) was used to align the sequences. In each case, the indel (gap) cost (penalty) was 2 and the base substitution cost was 1 (for both transitions and transversions). Each gap inserted in a sequence was treated independently, i.e. a gap of length three cost three times the individual gap cost. Ten random addition sequences were performed; each was then subjected to TBR branch swapping on the alignment tree. This means that many (millions in fact) multiple alignment orders were tried and for each a multiple alignment created, and a cladogram search performed. The cost of each alignment was determined by the length of the shortest tree derived from the multiple alignment.

Table 5.1 Character incongruence among data sets by method

Taxon	*Data set*	*Direct*	*Malign*	*MF direct*	*MF Malign*
Hemiptera	All	3263	3764	0.0362	0.0404
	16	792	850		
	18	632	781		
	28	772	1013		
	COI	949	968		
Carnivora	All	3802	4365	0.0326	0.0575
	12	853	987		
	16	1337	1666		
	Cyb	1488	1461		
Chelicerata	All	4691	5686	0.0284	0.0990
	Morph	804	804		
	18	1845	2205		
	28	1909	2114		
	Mol	3849	4539	0.0247	0.0485

Direct = direct optimization (optimization–alignment); Malign = multiple alignment produced by Malign; MF = Michevich-Farris incongruence value; COI = cytochrome oxidase I; Cyb = cytochrome B; all = combined data; 12 = 12S rDNA; 16 = 16S rDNA; 18 = 18S rDNA; 28 = 28S rDNA; mol = combined molecular data; morph = morphological data.

When these tree searches were performed SPR branch swapping was performed on multiple trees. The reason the less stringent searches were used for the choice of multiple alignment was that tree searches were performed for each one of the millions of multiple alignments performed. After the best (lowest cost) alignment was found, the program PHAST (Goloboff, 1996) was used to search for the shortest cladogram based on the best alignment using TBR branch-swapping and 10 random addition sequences.

Direct optimization

The program POY (Gladstein and Wheeler, 1997) was used to perform the direct analysis of the sequence data. As with alignment, the indel (gap) cost was 2 and the base substitution cost was 1 (for both transitions and transversions) and gap costs were linear with respect to length. Ten random addition sequences were performed; each was then subjected to TBR branch swapping. Since the results of the direct optimization of the sequences are parsimonious topologies, no further operations were required.

In both cases, the programs (POY and MALIGN) were executed on a parallel cluster consisting of 23 Unix workstations of various types united by Parallel Virtual Machine (PVM v. 3.4).

Results

The results of the combined and separate analyses are given in Table 5.1. Of the 13 analyses performed by both direct optimization and multiple alignment, 12 cases resulted in a shorter cladogram for the direct analysis than for that based on multiple alignment. This has been shown before (Wheeler, 1996) and appears to be general at

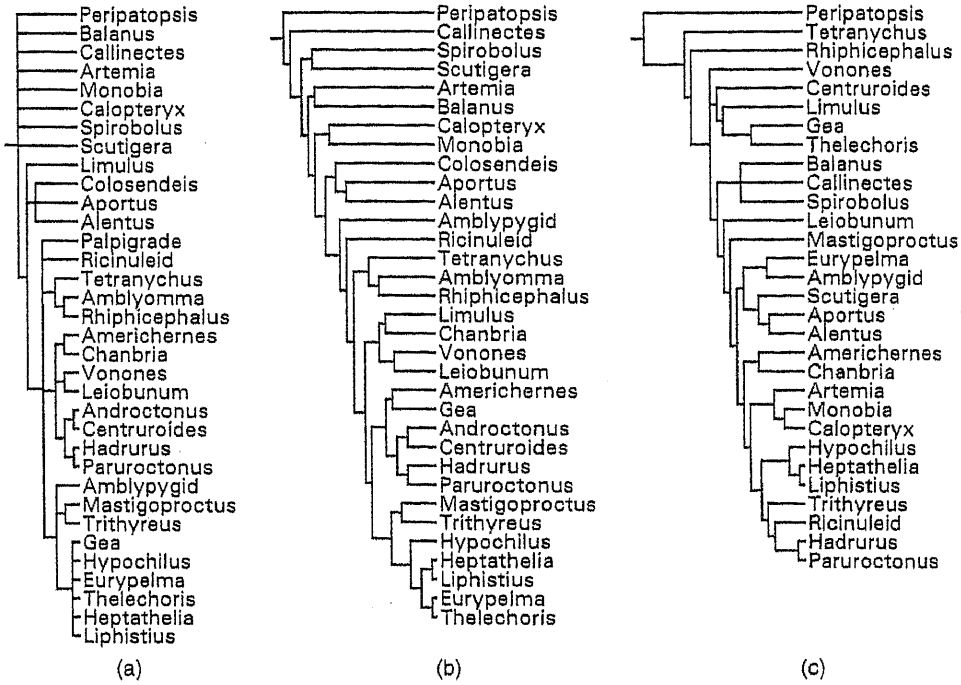

Figure 5.2 Chelicerate cladograms generated by direct optimization of separate analyses: (a) 93 morphological characters; (b) 18S rDNA; (c) 28S rDNA. The character incongruence level of 2.6% (MF) is much lower than any topological measure of congruence would suggest, since they share only four of 32 non-trivial groups.

least as far as these sequences are concerned. The enhancement of parsimony is most likely due to the fact that the putative homologies of bases in direct optimization are determined for each topology separately. In multiple alignment, these correspondences are predetermined and invariant with respect to the phylogenetic topology. The single case of alignment yielding more parsimonious results, carnivore cytochrome b, contained a great deal of missing data and may be more complex than the lengths suggest. The term 'shorter' in this case signifies that fewer insertions, deletions, and base substitutions were required by the direct analyses than those based on multiple alignment. Although I feel that these numbers are directly comparable, others have suggested that since the sequences are treated in such different ways, the simple lengths of the cladograms are not comparable. Also, MALIGN may well not be the best way to generate parsimonious multiple alignments. Assuming these factors, the most important criterion for comparison is the MF incongruence measures.

In each of the three cases examined – Hemiptera, Carnivora, and Chelicerata – the analyses based on direct optimization had lower MF incongruence levels than those derived from multiple alignment. The topological implications of these analyses are shown for the chelicerate topologies generated by direct optimization (Fig. 5.2). As can be seen in this example, although character congruence is very high, there is little common resolution among these topologies.

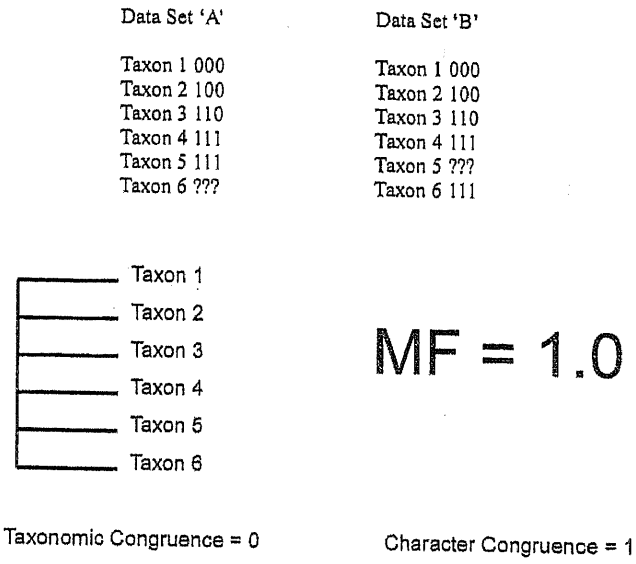

Figure 5.3 Topological vs character incongruence. One missing taxon in sets A and B causes zero taxonomic congruence (as measured by strict consensus) but complete congruence at the character level.

Discussion

Although we can never measure whether real data sets are accurate or not, we can assess their consistency. By definition, accurate methods will be convergent. Most current discussions of congruence are based on one of two notions of agreement: taxonomic and character. Taxonomic congruence (*sensu* Mickevich and Farris, 1981) measures the similarity in the topologies of systematic hypotheses. This has the strength of addressing the similarity of historical conclusions among data sets, but is very sensitive to small shifts in a single taxon. As has been shown many times (e.g. Wheeler, 1995), a single unstable taxon can reduce taxonomic congruence to such low levels that meaningful comparisons are impossible. Additionally, data sets must contain precisely equal taxon samples to be comparable (pruned comparisons aside). If some taxa are not available for analysis by certain data sets (e.g. extinct taxa and molecular information), taxonomic incongruence is not easily measured. Character-based incongruence, on the other hand, avoids these problems (Fig. 5.3). In character congruence, the extra homoplasy required by combining data is used to assay the agreement among characters. Taxa without data from a particular source are no problem, they are merely missing entries. This metric (Mickevich and Farris, 1981) (MF) has been used frequently to measure congruence (Kluge, 1989; Wheeler *et al.*, 1993; Wheeler, 1995; Whiting *et al.*, 1997). However, character-based incongruence is not without criticism. When comparing phylogenetic methods (or even assumptions), the MF value measures extra homoplasy as a fraction of total character change. This is accomplished by summing up the lengths of the constituent data sets when analysed independently and subtracting that number from the length of the cladogram calculated from the combined data.

This number of extra steps is then normalized via division by this same length of the cladogram derived from the combined data. Hence, MF is the fraction of the total length of this combined data cladogram that is due to combining the data. The greater the difference between the individual analyses and the combined, the greater is the incongruence. Suppose, however, that a poor method is used. In this case, so much homoplasy is already present in the separate analyses that combining them changes little. The number of extra steps and the MF value are low. An example of this would be to use cursory searches for individual data (say without branch swapping) and diligent, time-consuming (or even exact) searches for the combined data. The MF values would be artificially decreased since the individual tree lengths would be artificially long. At present, it is unclear how to deal with this problem except to suspect MF values whose base cladogram lengths are vastly different or perhaps to adjust the MF numbers to reflect the maximum possible incongruence. The rationale behind this adjustment would be that bad methods would express a greater fraction of the possible incongruence in the values used to calculate MF.

Conclusions

Desirable phylogenetic methods are consistent. Although we may not ever know whether an answer is correct ('true') or not, we can assay methods by the constancy of pattern they derive for the same taxa based on different data. Although the measurement of congruence is far from zipless, it does afford us a metric to gauge the behaviour of different reconstruction procedures.

The cases examined here compared nuclear, mitochondrial sequence characters and morphology – a total of six different sources. These data were presented in three different systematic areas and yet demonstrate a consistent pattern of greater congruence for direct optimization than for multiple sequence alignment. If consistency is our guide, direct optimization is superior.

There are several caveats, however, the most important being the use and measure of congruence. The argument has been made (Goloboff, personal communication) that congruence (or at least MF character congruence) cannot be used because it will favour methods that amplify homoplasy. As discussed above, if homoplasy is put in the analysis early, there is little left over to create incongruence when data are combined. This scenario would require that the raw cladogram lengths from which MF is calculated be very different. The 'bad' method would have to have much longer constituent cladograms and similar combined length results (hence lower MF). The direct optimization has *shorter* constituent cladograms. Additionally, the differences in length are in the 10–20% range, not hugely variant (see Table 5.1).

Secondly, no general conclusion about a method as complex as multiple sequence alignment could ever be resolved based on a single implementation. These results will have to be verified with other alignment methods and other data. Even with these reservations, however, the pattern remains. Direct optimization results are more congruent than those derived from sequence alignment.

Acknowledgements

I would like to thank and acknowledge the contributions of Robert Scotland, Mario de Pinna, Andrew Brower, Rob DeSalle, Gonzalo Giribet, Daniel Janies, Amy Litt, and Pablo Goloboff, without whose assistance this work would not have been possible.

References

Farris, J.S. (1970) A method for computing Wagner trees, *Systematic Zoology*, 34, 21–34.

Fitch, W.M. (1971) Toward defining the course of evolution: minimum changes for a specific tree topology, *Systematic Zoology* 20, 406–416.

Gladstein, D.S. and Wheeler, W.C. (1997) *POY: the Optimization of Alignment Characters* program and documentation, New York. Available from ftp.amnh.org/pub/molecular.

Goloboff, P. (1996) *PHAST version 1.0* program and documentation. Available from the author.

Kluge, A. (1989) A concern for evidence and a phylogenetic hypothesis for relationships among *Epicrates* (Boidae, Serpentes), *Systematic Zoology* 38, 1–25.

Mickevich, M.F., and Farris, J.S. (1981) The implications of congruence in *Menidia*, *Systematic Zoology*, 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, *Journal of Molecular Biology* 48, 443–453.

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

Vrana, P.B., Milinkovich, M.C., Powell, J.R. and Wheeler, W.C. (1994) Higher relationships of the arctoid Carnivora based on sequence data and 'total evidence', *Molecular Phylogenetic Evolution*, 3, 47–58.

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

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

Wheeler, W.C. (1998) Alignment characters, dynamic programming, and heuristic solutions, in Schierwater, B., Streit, B., Wagner, G.P. and DeSalle, R. (eds) *Molecular Approaches to Ecology and Evolution*, 2nd edn. Basel: Birkhäuser Verlag, pp. 243–251.

Wheeler, W.C., Bang, R. and Schuh R.T. (1993) Cladistic relationships among higher groups of Heteroptera: congruence between morphological and molecular data sets, *Ent. Scand.*, 24, 121–138.

Wheeler, W.C. and Gladstein, D.G. (1992) *Malign: a multiple sequence alignment program*, program and documentation, New York. Avaliable from ftp://ftp.amnh.org/pub/molecular.

Wheeler, W.C. and Gladstein, D.G. (1994) Malign: a multiple nucleic acid sequence alignment program, *Journal of Heredity*, 85, 417.

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

Whiting, M. and Wheeler, W.C. (1994) Phylogenetic position of the Strepsiptera: evidence for a homeotic reciprocal thoracic transformation *Nature*, 368, 696.

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