Molecular Systematics and Evolution: Theory and Practice
ed. by R. DeSalle, G. Giribet and W. Wheeler

DNA multiple sequence alignments

Gonzalo Giribet[1], Ward C. Wheeler[2] and Jyrki Muona[3]

[1] *Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge MA 02138, USA*
[2] *Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA*
[3] *Zoological Museum, Division of Entomology, Finnish Museum of Natural History, University of Helsinki, FIN-00014 Helsinki, Finland*

Summary. In this chapter we examine the procedure of multiple sequence alignment. We first examine the heuristic procedures commonly used in multiple sequence alignment. Next we examine sources of ambiguity involved in the alignment procedure. We suggest that several alignment parameters be employed to examine alignment sensitivity. We end by presenting an experiment with humans showing the ambiguity involved in manual alignment.

Introduction

Multiple sequence alignment is a procedure to turn unequal length sequences into equal length character strings via the insertion of gaps. These gaps are mere placeholders which indicate that an insertion or deletion has occurred somewhere after the compared sequences diverged from a common ancestor, resulting in a lack of homologous nucleotides at that position for that taxon.

Despite the existence of new methods for phylogenetic analysis that entirely avoid alignments, the issue of using multiple sequence alignments (fixed alignments) as a source for the primary homology statements for phylogenetic analysis is still important for certain areas of knowledge. An investigator may choose a fixed alignment *versus* a dynamic alignment, and base-to-base correspondences *versus* fragment-to-fragment correspondences for several reasons. For example, below the population level, or to study molecular evolution (of both DNA and proteins), certain methods that are commonly applied require the use of fixed alignments (see Wheeler 2002).

Three issues appear important to us in this respect. First, how multiple sequence alignments are generated algorithmically (and the inherent problematica of alignments). Second, how the available software performs alignments (implementation). Third, how parameter sensitivity enters into exploring phylogenetic hypotheses at the alignment level. This last issue also applies to all other methods of sequence comparison (i.e., optimization of DNA fragments).

Background

The first step in any phylogenetic analysis involves some sort of pairwise comparisons of DNA data (or amino acids; but from here on, we will refer to DNA data analyses). Two families of comparison methods are available: local comparisons, meant to search for homologous domains among sequences, such as the BLAST family of programs (e.g., Altschul et al., 1997), and global comparisons, the type applied to phylogenetic inference, where the entirety of two putatively homologous strings of DNA is compared to assign base-to-base correspondences.

The fundamental method of pairwise sequence alignment was first described by Needleman and Wunsch (1970), and extended to multiple dimensions by Sankoff and Cedergren (1983). The Needleman and Wunsch algorithm calculates the minimum edit distance between two DNA sequences, which is the minimum number of transformations required to go from one sequence to another. In its simplest incarnation, two parameters need to be specified, the gap penalty (or indel cost: the cost assigned to insertion or deletion events), and the change cost (the cost assigned to go from one base to any other). This change cost can be categorized in many different ways, assigning independent costs for every particular type of transformation, or assigning costs to certain categories (i.e., transversions, transitions, etc.). These costs need to be explicit in any algorithmic comparison, and have some lower boundaries delimited by the triangle inequality (Wheeler, 1993). The Needleman and Wunsch algorithm can be expressed as a minimization process, but other optimization procedures for DNA sequence comparisons might be used as well (e.g., maximization of base matches). The specific mechanics of the Needleman and Wunsch algorithm have been reviewed elsewhere (Wheeler, 1994), and we are not going to review the process in detail, but just note certain relevant aspects.

In order to align two sequences of length $(N-1)$ and $(M-1)$, a matrix of $N \times M$ cells is created, and the minimum cost path through this matrix (given specific parameter costs) is calculated. The matrix is traversed in such a fashion that only the adjacent three cells (usually the cells above, to the left, and diagonally up 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 each 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, extensions to phylogenetically interesting numbers of sequences are extremely computationally intensive. 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 Cedergren, 1983).

These complicating factors have made true multiple alignment unachievable for anything but the smallest number of taxa. Real data sets require, at least, heuristic solutions (Wheeler, 2000b). In fact Slowinski (1998) showed that there are 1.05×10^{18} different alignments for five DNA sequences of five nucleotides each, and thus recommends not even attempting to perform multiple sequence alignments, since any optimality criterion is "virtually guaranteed to fail".

The heuristic strategy followed in multiple sequence alignment procedures is quite simple. Since aligning two sequences is easy, the procedure adds sequences via a "guide tree". All programs for multiple sequence alignment in common use today follow this idea, but differ in how they get the binary "guide" tree, and how they add the pairwise results together to generate the complete alignment.

Three implementations of heuristic multiple sequence alignment algorithms that are in some use today are the CLUSTAL family (Higgins and Sharp, 1988, 1989; Higgins et al., 1992, 1996; Higgins, 1994; Thompson et al., 1994, 1997; Jeanmougin et al., 1998), TREEALIGN (Hein, 1989, 1990), and MALIGN (Wheeler and Gladstein, 1994, 1995). These three programs rely on guide trees to accrete pairwise alignment. In the case of CLUSTAL and TREEALIGN, a distance tree is calculated from all the pairwise sequence similarity scores, 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 the other programs 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 partial alignment of 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. The entire procedure involves two levels of heuristics, one to generate the alignment and another to perform tree searches on each one of the alignments. The alignment (or alignments, if multiple solutions are found) which produces the most parsimonious phylogenetic result (i.e., lowest cost) 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 n^3 for n sequences). Not surprisingly, CLUSTAL and TREEALIGN frequently generate results more rapidly than MALIGN.

Furthermore, sequence comparison can well include evolutionary models and be based on statistical approaches. Maximum likelihood methods for alignment of DNA sequences have been proposed (Thorne et al., 1991; Thorne and Churchill, 1995), although these methods have not yet been applied to phylogenetically interesting data sets.

In summary, irrespective of which program or method is used, multiple sequence alignment is a computationally expensive technique, and only heuristic solutions can be achieved.

Sources of ambiguity

That alignments originated from different sources might result in alternative phylogenetic hypotheses is logical, and has also been demonstrated empirically (e.g., Wägele and Stanjek, 1995; Winnepenninckx and Backeljau, 1996). In a recent review on DNA sequence alignments, Wheeler (1994) enumerated three sources of ambiguity in multiple sequence alignments (sources of non-unique alignments). An extra source of difficulty, as mentioned above, is the necessity of heuristics in solving alignment problems. Three sources of ambiguity are:

Parameter variation

That different parameters can result in different alignments, and consequently in alternative phylogenetic hypotheses, is a well-known phenomenon, first described by Fitch and Smith (1983; see also Waterman et al., 1992; Wheeler, 1995; Morrison and Ellis, 1997; Cerchio and Tucker, 1998; Giribet and Wheeler, 1999b). Since there is *a priori* no way to determine directly the appropriate gap or change values, more or less arbitrary decisions must be made when choosing a particular cost regime (Giribet and Wheeler, 1999b). An obvious solution to this problem is to examine a wide space of parameters. For example, using a large regime of gap and change costs would show which areas of the alignment are conserved, and which are more parameter-dependent. What the investigator does with this information is another matter.

Describing the enormous parameter space that can be explored by multiple sequence alignments, Higgins et al. (1996) stated that:

> "We justify this by asking the user to treat CLUSTAL W as a data exploration tool rather than as a definitive analysis method. It is not sensible to automatically derive multiple alignments and to trust particular algorithms as being capable of always getting the correct answer".

Many investigators remove "gappy" areas (whether obtained automatically or manually), appealing to the idea that these areas do not reflect true homolo-

gies, or that the pattern of homology cannot be recognized. This could lead to extremes in which all informative data are removed, especially if hundreds of sequences are examined, even if they were coding genes (never underestimate the power of mutation!). Furthermore, many times this is done because the alignments have been generated manually, or by using bogus algorithms. Other more objective alternatives have been proposed, such as Cull or Elision (Gatesy et al., 1993; DeSalle et al., 1994; Wheeler et al., 1995). However, Cull could also end up with all the information removed from the alignment. Elision is neater in the sense that it acts as a weighting function, downweighting all these positions with ambiguous alignments.

A third solution was proposed by W.C. Wheeler (1994, 1995), which is the use of congruence with other sources of information to decide which alignment best explains evolution of all sources of phylogenetic evidence. Character congruence (Mickevich and Farris, 1981; Farris et al., 1995) or topological congruence (Wheeler, 1999a) are our preferred criteria. In these cases, no information is discarded or downweighted, which accounts for analyses that accommodate a wider scheme of phylogenetic variation. This does not mean that the alternative alignments should not be explored to test for phylogenetic stability to parameter choice. Obviously, more parameters can always be analyzed. Another critique of the use of parameters is that the scheme of parameters is applied uniformly to all the positions in the analyses. But as the phylogenetic data come today, it seems the best way to account for the first source of ambiguity in multiple sequence alignment.

Multiple order-dependent solutions

When multiple alignments are created, whether by exact or by heuristic means, the notion of alignment order comes into play. Heuristic multiple alignment solutions are built typically from a series of pairwise alignments. Initially two sequences are aligned and this result aligned to a third sequence, maintaining the relative alignment between the first two ("once a gap, always a gap"; Feng and Doolittle, 1987, 1990) and so on. This procedure is obviously order-dependent. A different addition order might well yield a different alignment, even when the exact same parameters are chosen. So, not only can different parameter sets result in different alignments, but also the same parameter sets might yield different alignments if a different guide tree is used. This was considered by Wheeler (1994) to be analogous to the existence of multiple optimal trees in a standard parsimony phylogenetic analysis.

Multiple path-dependent solutions

The third source of alignment ambiguity is path variation. Path variation occurs when the alignment algorithm can follow multiple paths through the alignment

space, yielding again multiple solutions (even for the same parameter space and for the same guide tree). Path variation occurs when the alignment can either insert a gap or match the bases with equal cost. Every time that this happens, the number of optimal solutions multiplies, and if this happens repeatedly, the result is a large number of equally costly, but different alignments.

Alignments are just hypotheses of homology

Alignments are not "given static hypotheses of homology", or any phenomenon that we can observe in nature. This is a truth that is hard to accept for many investigators. The same applies to particular base transformations, insertions, deletions, etc. The path from one sequence to another, connected by a common ancestor, may suggest such a phenomenon, but alignments of multiple taxa are missing way too many of these events, too many terminals, and too many ancestors (nodes). This implies that any information that we report in the form of an alignment is the most accurate estimate of these unobserved processes, and thus we should not be afraid to explore the solutions suggested by alternative alignments. Otherwise we would be fooling ourselves by believing that we got "the" alignment.

As an example, there are several possible alignments for the following two sequences (1) AATCGCG and (2) AACCCGG. Four of these possibilities are shown here:

(a)
```
AATCGCG
AACCCGG
```

(b)
```
AATCGCG-
AACC-CGG
```

(c)
```
AATCGCG-
AA-CCCGG
```

(d)
```
AATCGC-G-
AA-C-CCGG
```

Depending on the parameter set adopted, some alignments will be "better" (shorter) than others. For example, if we consider all transformations as equal and assign them a cost of 1 (gap cost = 1; tv cost = 1; ts cost = 1), alignment (a) requires three transformations (a total cost of 3), as it does alignment (b) and (c), while alignment (d) requires four transformations (a total cost of 4). Thus, applying this model, alignments (a), (b) and (c) are equally supported. If we apply a second model with gap costs weighted twice as much as base transformations (gap cost = 2; tv cost = 1; ts cost = 1), then alignment (a) requires no gaps, 2 transversions and 1 transition (3 base transformations; total cost of 3). Alignment (b) requires two indel events and one transition (total cost of 5), alignment (c) requires two indel events and one transversion (total cost of 5), and alignment (d) requires 4 indel events and no base transformations (total cost of 8). Yet other models could be applied, resulting in favored alignments (a) and (b) (gap cost = 2; tv cost = 2; ts cost = 1); (b) (gap cost = 1; tv cost = 2; ts cost = 1), etc. (Tab. 1).

Table 1. Total cost of alignments (a), (b), (c) and (d) at different parameter values

gap	tv	ts	alignment	total
1	1	1	(a)	3
			(b)	3
			(c)	3
			(d)	4
2	1	1	(a)	3
			(b)	5
			(c)	5
			(d)	8
2	2	1	(a)	5
			(b)	5
			(c)	6
			(d)	8
1	2	1	(a)	5
			(b)	3
			(c)	4
			(d)	4

What we are trying to illustrate with this example is the notion that the decision on which is the "best" alignment is not trivial, and certainly decisions made "by eye" would probably choose alignment (a) *versus* the alternative ones, although (b) and (c) might be as good or even better under a wide range of parameters. This, we guess, stresses the necessity of being explicit and repeatable, two conditions only mutually satisfied by automatic alignments. The lack of sufficiently good algorithms to perform multiple alignments should not be taken as a critique of a philosophically superior method.

Due to the existence of sources of ambiguity in multiple sequence alignments, different alignments based on different parameter sets should be explored. With these multiple hypotheses of positional homology, phylogenetic analyses increase in complexity, but decrease in the degree of arbitrariness.

Experimenting with humans

In order to evaluate "manual alignments," eight student investigators were given a set of sequences, the "original data", and were asked to align them. The original data consisted of ten sequences, most starting with the motif "AAGAAGAAT", and all of them ending with the motif "TTTATTTTGA". The students knew what homology meant and were supposed to align the sequences so that they would be equally long and have the "highest possible" base-to-base concordance, any way they could.

The same ten sequences were submitted to ClustalW and Malign for multiple sequence alignments with parameters set at gap cost = 10; change cost = 1; gap extension penalty options off (each gap was given the same cost value, as if they were independent). The sequences were also optimized in POY

Table 2. Tree length of the alignment of 10 sequences

One	388
Two	362 (some bases deleted)
Three	374 (some bases deleted)
Four	367 (some bases deleted)
Five	388
Six	367 (some bases deleted)
Seven	362 (some bases deleted)
Eight	357 (some bases deleted)
CLUSTAL	395
MALIGN	386
POY	379

One to Eight indicate manual alignments generated by 8 students. MALIGN. CLUSTAL and POY indicate the alignments obtained with the respective programs.

(Gladstein and Wheeler, 1997) and the "implied alignment" corresponding to the topology optimized was compared to the other alignments. Manual alignments and computer-generated alignments were evaluated using the parsimony program NONA v 2.0 (Goloboff, 1994), counting gaps as a character state (gap cost arbitrarily set at 1) using tbr branch swapping (h1000;h/10;mult*100;).

For the manual alignments, in general a few gaps were added, but in most cases, a few bases were removed as well, making tree-length comparisons impossible. Of course removing bases was incorrect, but these were just examples. Tree lengths for all the alignments are given in Table 2. What we can observe from this simple experiment is that manual alignments are unpredictable. In addition, they incorporate a high degree of subjectivity and are error-prone.

Computer-generated alignments dependent on guide trees are more parsimonious if several guide trees are examined (MALIGN *versus* CLUSTAL), although shorter alignments might exist (implied alignment from POY).
Multiple sequencing alignment is a complicated process that requires considerably large amounts of computation. Methods using a single guide tree can be improved by doing multiple runs with different starting points, giving several randomly generated guide trees. However, this is tedious and this is why programs such as MALIGN, which examines multiple guide trees (using multiple random addition), are superior. Even when multiple guide trees are examined, there is no guarantee, as in any other heuristic procedure, that the optimal (shortest) alignment will be found. Shorter alignments can be found much faster by outputting the implied alignment with a tree generated via DNA direct optimization (Wheeler, 1996).

Acknowledgements
We want to thank Rob DeSalle and John Gatesy for comments that improved this chapter.