MALIGN: A Multiple Sequence Alignment Program

W. C. Wheeler and D. S. Gladstein

MALIGN is a computer program that aligns molecular sequences in a phylogenetic context. Although the program is designed around the analysis of nucleic acid sequence data, proteins can be aligned through conversion to nucleotide ambiguities. The program produces sequence alignments that will yield parsimonious phylogenetic reconstructions. MALIGN provides a number of heuristic and exact procedures (analogous to those used in phylogenetic reconstruction programs) to align the sequences and yields alignments in a variety of formats suitable for direct input into other programs. MALIGN is available for DOS, DOS-386, Macintosh, and SUN Unix Workstations. The program will also operate in DOS windows provided by MS-Windows or OS2.

Multiple sequence alignment is computationally laborious and extremely memory intensive. The basic idea of multiple alignment is a relatively straightforward extension of the Needleman and Wunsch (1970) dynamic programming algorithm for two sequences. An *n*-dimensional hypercube is created with each of the *n* sequences to be aligned as an axis. The least-cost path through this matrix from corner to corner defines the multiple alignment. This procedure is unrealizable for all but the most depauperate of data sets.

In order to achieve a result, it is possible to align multiple sequences by sequential accumulation via pairwise alignments. This has been suggested and implemented by many authors (Feng and Doolittle 1987, 1990; Hein 1989, 1990; Higgins and Sharp 1988, 1989). The problem of order dependency, however, remains. The order in which the pairwise alignments are accomplished affects the final multiple alignment, and the number of possible orders explodes combinatorially as the number of sequences increases. MALIGN aligns sequences in this pairwise manner, but, to overcome the problem of order dependence, manipulates the trees that determine the order of sequence accretion in its search for better alignments. Multiple scenarios are examined, and the one (or several) that implies the fewest evolutionary events (most parsimonious) among the sequences is chosen and output.

The cost of the alignments is usually calculated by determining the most parsimonious (shortest) evolutionary tree derived from an alignment. This phylogenetic cost involves the calculation of minimum length spanning trees (an NP-complete problem), and a variety of options, from the most simple heuristics, to those with several forms of branch swapping, to exact branch-and-bound (Hendy and Penny 1982) solutions, are provided. The length and number of input sequences are limited to 32,767. In practice, the amount of memory and patience available to the user will more severely limit these parameters.

The program can be employed through an interactive interface or through command line instructions. The parameters of the analysis (such as gap costs and other variables) may be entered through the specification of a parameter file or entered directly from the command line. The input data file format is basically a modified GenBank sequence format explained fully in the documentation. IUPAC nucleotide ambiguity codes are accepted. Amino acid sequences can be freely intermixed with nucleic acid sequences, and MALIGN will automatically convert them to nucleotide ambiguity representations. Amino acid designations may be reassigned (for different triplet codes), and the digits 0–9 may also be assigned triplet equivalents (when ambiguities or multiple triplet codes occur within a single data set).

Since the sequences are aligned via trees of relationship (Sankoff and Cedergren 1983), it is possible to force the alignments to occur along a particular path or set of paths. In this way, closely related sequences can be aligned to each other before less related sequences are added (e.g., aligning bird and crocodile sequences to each other before a lizard is added), as has been advocated by Mindell (1991) in systematic analysis. This procedure is accomplished by the "groups" command, which is fully detailed in the documentation.

Complex matrices may be used as cost functions, including transition–transversion bias as well as the relative cost of gaps and base changes. The costs of various sorts of gaps can be specified. Leading and trailing gaps can be set individually and differently from those that occur within sequences. Gaps can be treated as independent insertions (gaps of length two would cost twice as much as a single gap), or with subsequent contiguous gaps as less or more costly. Gaps with lengths in multiples of three that can maintain reading frames may also be assigned their own costs.

The alignments themselves examine many tree structures and here (as with phylogenetic tree reconstruction) require a variety of heuristic and exact approaches. Simple addition of taxa can be accomplished with very rapid (but not particularly effective) algorithms. More aggressive procedures involving branch swapping and even exact solution branch-and-bound searches can be performed. MALIGN will produce multiple equally costly alignments when found.

MALIGN will output sequences in a variety of formats. Simple text strings may be produced as can "dot" representations showing only differences. Of most use and interest, though, are the outputs to the phylogenetic reconstruction programs such as PAUP (Swofford 1989) and Hennig86 (Farris 1988). Files generated by MALIGN can be used directly by these programs.

For a copy of MALIGN and its documentation, contact Ward Wheeler. MALIGN is available for DOS, DOS-386, Macintosh, and SUN Unix Workstations. The program will operate in DOS windows provided by MS-Windows or OS2.

From the Department of Invertebrates, American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192. Address correspondence to Dr. Wheeler at the address above.

The Journal of Heredity 1994:85(5)

References

Farris JS, 1988. Hennig86 version 1.5, Program and documentation. Port Jervis, New York.

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

Feng D and Doolittle RF, 1990. Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzymol 183:375–387.

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 Enzymol 183:626–644.

Hendy MD and Penny D, 1982. Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59:277–290.

Higgins DG and Sharp PM, 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237–244.

Higgins DG and Sharp PM, 1989. Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5:151–153.

Mindell D, 1991. Aligning DNA sequences: homology and phylogenetic weighting. In: Phylogenetic analysis of DNA sequences (Miyamoto MJ and Cracraft J, eds). New York: Oxford University Press; 73–89.

Needleman SB and Wunsch CD, 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 DD and Cedergren RJ, 1983. Simultaneous comparison of three or more sequences related by a tree. In: Time warps, string edits, and macromolecules: the theory and practise of sequence comparison (Sankoff D and Kruskal JB, eds). Reading, Massachusetts: Addison-Wesley; 253–264.

Swofford DL, 1989. PAUP version 3.0q, Program and documentation. Champaign, Illinois.