
Tutorial section

Multiple sequence alignment – the gateway to further analysis

Whether the ultimate aim is a phylogenetic analysis of several orthologues, the identification of a pattern for particular feature or motif, or the basis for structural modelling, multiple sequence alignments allow the researcher to gather more biological information than a single sequence can offer. Possibly the most popular method for comparing three or more sequences is the clustering¹ algorithm used in applications such as the Clustal (ClustalW and ClustalX) series of programs.^{2,3} It is certainly by no means the only method of alignment, but will be used to illustrate this text.

Initial clustering of sequence pairs reduces the computing time required to align multiple sequences and this can be achieved using one of two possible methods. Slow clustering is the more rigorous of the two options, but is noticeably much slower for approximately 20 or more sequences, or fewer, longer regions. It uses the dynamic programming method of Needleman–Wunsch⁴ to align each sequence with another according to a weight matrix^{5–9} and gap penalties. The ultimate aim of the computer program is to achieve the highest score possible, within the constraints the program has been placed under.

Weight matrices have been developed using homologous sequences, and allocate a score to each residue or nucleotide base indicating the probability of it replacing a different residue or nucleotide base as a possible mutation. In the case of protein sequences, this has been done for all 20 amino acid residues, together with the

three ambiguity codes (B = Asp and Asn, Z = Glu and Gln, X = any residue) using several different methods. Nucleotide matrices have also been developed, and in general indicate a positive score for an identical match, and no score, or a negative one for a mismatch. Because of its very nature, and the existence of only four common bases, more information for the alignment can be obtained by using protein sequences, and it often makes sense to translate regions of coding DNA into protein sequence before aligning them.

Once a high score has been achieved for each of the sequence pairs in the alignment, they are clustered together in accordance with their relative scores, using the neighbour-joining method¹⁰ to link the closest pairings together, and less similar sequences more remotely. This information is stored as a series of numerical distances arranged by means of nested brackets in a dendrogram file. This file is in no way representative of evolutionary distances, and should not be presented as such. It merely represents the proximity of each sequence within a cluster, and each cluster to another and is used to form the final alignment. The information retained in the dendrogram file may be kept and used to align other multiple sequence sets.

Larger sequence volumes may be compared using a faster method, in order to reduce computing time. This is based on the algorithm of Wilbur and Lipman¹¹ and is quicker but less accurate than the dynamic programming methods of the slow comparison. It involves definition of

a k -tuple value which will be used to create short, identical fragments of the sequences included in the alignment. The default k -tuple value is often 2 for protein sequences. In this instance, every possible combination of two adjacent fragments throughout all sequences are identified, and the k -tuples compared with each other. Only exact matches are considered, and the sequences with the greatest number of matches are clustered more closely together, to define the dendrogram information that steers the final alignment. For nucleotide alignment, the k -tuple average falls between four and six bases. Larger k -tuples lead to faster but less sensitive alignments. Smaller k -tuples increase sensitivity of the comparison.

Owing to the automated nature of the alignment, it is highly likely that there will be misaligned regions of the sequence and these must be manually altered in an alignment editor. Alterations may be accomplished in several different ways, depending on the final aim of the alignment. Should information be tailored towards building a phylogenetic tree, the alignment should represent most closely areas that match exactly, together with those representing a similarity. It is not always obvious how such areas align, and there may be redundant regions relevant to only a minority of sequences in the alignment. Depending on the results a phylogenetic analysis is expected to produce, sequence regions either side of a relevant alignment may be discarded

before proceeding. This is particularly important as phylogenetic analysis relies on substitutions observed to create an evolutionary tree and mismatching regions would confuse the process.

Further investigation of motifs, domains or features within the sequence such as promotor regions relies on small areas of conservation. As many multiple sequence alignment programs assume global alignment, it is not imperative in these cases to align the entire sequence. However, high regions of conservation must be obvious to show a convincing homology. Such regions of a protein sequence can be used as a seed alignment to create profiles and hidden Markov models in order to search the databases more sensitively for similar regions. The more conserved the alignment, the fewer sequences are required to build up these statistical searches.

Nucleotide regions may also be aligned in this way, allowing the design of polymerase chain reaction (PCR) primers based on a conserved region to identify further, similar regions in other areas of the genome.

Should the goal of a sequence alignment be to underlie definition of protein structure, analysis must be towards aligning regions that fold into relevant structures. Alignment for structural analysis differs from conventional alignments, as convincing columns of conserved regions are not always the best scenario, and regions may be separated from one another in order to best represent each fold or secondary structural element of the protein. It is often best for this kind of analysis to identify an experimentally determined structure, by mining either the literature or the Protein Data Bank (PDB, which holds all experimentally and some theoretically determined protein structures¹²), and align sequence regions according to the two- or three-dimensional structure detailed. Structural information is often more conserved than sequence information and is thus important information for the alignment.

Table 1 ClustalW and multiple sequence alignment programs on the web

ClustalW web servers:

<http://www.ebi.ac.uk/clustalw/>
http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html
<http://transfac.gbf.de/programs/clustalw/clustalw.html>
<http://www.clustalw.genome.ad.jp/>
<http://www.ch.embnet.org/software/ClustalW.html>
<http://www.genebee.msu.su/clustal/basic.html>

Alignment viewers and editors:

http://www.uk.embnet.org/Software/EMBOSS/Jemboss*
<http://prodes.toulouse.inra.fr/ESPript/>
<http://pbil.univ-lyon1.fr/software/seaview.html>
<http://bioinf.man.ac.uk/dbbrowser/CINEMA2.1/>
<http://www.compbio.dundee.ac.uk/Software/Alscript/alscript.html>

*Jalview is currently available in Jemboss

Typically ClustalW and other multiple sequence alignment programs are available on web servers throughout the world (Table 1), as well as embedded into academic and commercial packages. Alignment editors and viewers are also available on the WWW, although many must be downloaded onto the user's computer to function.

Lisa J. Mullan
MRC UK Human Genome Mapping Project
(HGMP) Resource Centre,
Genome Campus,
Hinxton, Cambridge CB10 1SB, UK
Tel: +44 (0) 1223 494500;
Fax: +44 (0) 1223 494512
E-mail: lmullan@hgmp.mrc.ac.uk

References

1. Higgins, D. G. and Sharp, P. M. (1988), 'CLUSTAL: A package for performing multiple sequence alignment on a microcomputer', *Gene*, Vol. 73(1), pp. 237–244.
2. Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996), 'Using CLUSTAL for multiple sequence alignments', *Methods Enzymol.*, Vol. 266, pp. 383–402.
3. 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.*, Vol. 22(22), pp. 4673–4680.
4. 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.*, Vol. 48(3), pp. 443–453.
5. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978), 'A model of evolutionary change in proteins', in Dayhoff, M. O., Ed., 'Atlas of Protein Sequence and Structure', Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, DC, pp. 345–352.
6. Schwartz, R. M. and Dayhoff, M. O. (1978), 'Matrices for detecting distant relationships', in Dayhoff, M. O., Ed., 'Atlas of Protein Sequence and Structure', Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, DC, pp. 353–358.
7. Henikoff, S. and Henikoff, J. G. (1992), 'Amino acid substitution matrices from protein blocks', *Proc. Natl Acad. Sci. USA*, Vol. 89, pp. 10915–10919.
8. Gonnet, G. H., Cohen, M. A. and Benner, S. A. (1992), 'Exhaustive matching of the entire protein sequence database', *Science*, Vol. 256(5062), pp. 1443–1445.
9. Gonnet, G. H., Cohen, M. A. and Benner, S. A. (1994), 'Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix', *Biochem. Biophys. Res. Commun.*, Vol. 199(2), pp. 489–496.
10. Saitou, N. and Nei, M. (1987), 'The neighbor-joining method: A new method for constructing phylogenetic trees', *Mol. Biol. Evol.*, Vol. 4, pp. 406–425.
11. Wilbur, W. J. and Lipman, D. J. (1983), 'Rapid similarity searches of nucleic acid and protein data banks', *Proc. Natl Acad. Sci. USA*, Vol. 80(3), 726–730.
12. URL: <http://www.rcsb.org>