Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3497–3500***
DOI: 10.1093/nar/gkg500

Multiple sequence alignment with the Clustal series of programs

Ramu Chenna, Hideaki Sugawara[1], Tadashi Koike[1], Rodrigo Lopez[2], Toby J. Gibson, Desmond G. Higgins[3] and Julie D. Thompson[4,*]

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany, [1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka-Ken 411-8540, Japan, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [3]Department of Biochemistry, University College Cork, Cork, Ireland and [4]Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP), BP 10142, 67404 Illkirch Cedex, France

Received January 29, 2003; Revised and Accepted March 4, 2003

ABSTRACT

The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequences and for preparing phylogenetic trees. The popularity of the programs depends on a number of factors, including not only the accuracy of the results, but also the robustness, portability and user-friendliness of the programs. New features include NEXUS and FASTA format output, printing range numbers and faster tree calculation. Although, Clustal was originally developed to run on a local computer, numerous Web servers have been set up, notably at the EBI (European Bioinformatics Institute) (http://www.ebi.ac.uk/clustalw/).

INTRODUCTION

One of the cornerstones of modern bioinformatics is the comparison or alignment of protein sequences. With the aid of multiple sequence alignments, biologists are able to study the sequence patterns conserved through evolution and the ancestral relationships between different organisms. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). The most widely used programs for global multiple sequence alignment are from the Clustal series of programs. The first Clustal program was written by Des Higgins in 1988 (1) and was designed specifically to work efficiently on personal computers, which at that time, had feeble computing power by today's standards. It combined a memory-efficient dynamic programming algorithm (2) with the progressive alignment strategy developed by Feng and Doolittle (3) and Willie Taylor (4). The multiple alignment is built up progressively by a series of pairwise alignments, following the branching order in a guide tree. The initial pre-comparison used a rapid word-based alignment algorithm (5) and the guide tree was constructed using the UPGMA method (6). In 1992, a new release was made, called ClustalV (7,8), which incorporated profile alignments (alignments of existing alignments) and the facility to generate trees from the multiple alignment using the Neighbour-Joining (NJ) method (9). The third generation of the series, ClustalW (10), released in 1994, incorporated a number of improvements to the alignment algorithm, including sequence weighting, position-specific gap penalties and the automatic choice of a suitable residue comparison matrix at each stage in the multiple alignment. In addition, the approximate word search used for the pre-comparison step was replaced by a more sensitive dynamic programming algorithm, and the dendogram construction by UPGMA was replaced by NJ. The ClustalW program looked very similar to ClustalV, with simple text menus for interactive use and the possibility of running the program in batch mode by specifying the input file and the parameter options on the command line.

The rationale behind the development of the Clustal series has been to provide robust, portable programs that are capable of providing good, biologically accurate alignments within a reasonable time limit. A close collaboration between biologists and computer scientists is probably one of the main reasons for the success and continued widespread use of the Clustal programs. ClustalW has given rise to a number of developments, including the latest member of the family, ClustalX (11). Although the alignments produced are the same as those produced by the current release of ClustalW, the user can better evaluate alignments in ClustalX. The program displays the multiple alignment in a scrollable window and all parameters are available using pull-down menus. Within alignments, conserved columns are highlighted using a customizable colour scheme and quality analysis tools are available to highlight potentially misaligned regions. ClustalX is easy to

*To whom correspondence should be addressed. Tel: +33 388653200; Fax: +33 388653276; Email: julie@igbmc.u-strasbg.fr


```
1fdr/100-247    -------DEVPHCETLWMLATGTAIGPYLSILR-LGKDLDR----FKNLVLVHAARYAAD
1ndh/129-270    RPDKKSSPVIKTVKSVGMIAGGTGITPMLQVIRAIMKDPDD----HTVCHLLFANQTEKD
1fnc/132-287    ---------KDPNATIIMLGTGTGIAPFRSFLWKMFFEKHDDYKFNGLAWLFLGVPTSSS
2pia/106-223    ---------DKRAKSFILVAGGIGITPMLSMARQLRAEGLR------SFRLYYLTRDPEG
                              :. ::. * .* *  ..   :  :           *       .

1fdr/100-247    LSYLPLMQELEKRYEGKLRIQTVVSRETAAGSLTGRIPALIESGELESTIGLPMNK--ET
1ndh/129-270    ILLRPELEELRNEHSARFKLWYTVDRAPEAWDYS---QGFVNEEMIRDHLPPPE----EE
1fnc/132-287    LLYKEEFEKMKEKAPDNFRLDFAVSREQTNEKGE---KMYIQTRMAQYAVELWEMLKKDN
2pia/106-223    TAFFDELTS--DEWRSDVKIHHDHGDPTKAFDFW---SVFEKSK--------------PA
                      : .  ..    .::    .     .         :

1fdr/100-247    SHVMLCGNPQMVR-----DTQQLLKETRQMTKHLR--RRPGHMTAEHYW
1ndh/129-270    PLVLMCGPPPMIQ-----YACLPNLER---VGHPK--ERCFAF------
1fnc/132-287    TYVYMCGLKGMEKGIDDIMVSLAAAEGIDWIEYKRQLKKAEQWNVEVY-
2pia/106-223    QHVYCCGPQALMD-------TVRDMTG----HWPSGTVHFESF------
                  *  **   :                       :
```

Figure 1. A multiple alignment of four oxidoreductase NAD binding domain protein sequences. Residues are coloured according to the following criteria: AVFPMILW are shown in red, DE are blue, RHK are magenta, STYHCNGQ are green and all other residues are grey. The residue range for each sequence is shown after the sequence name.

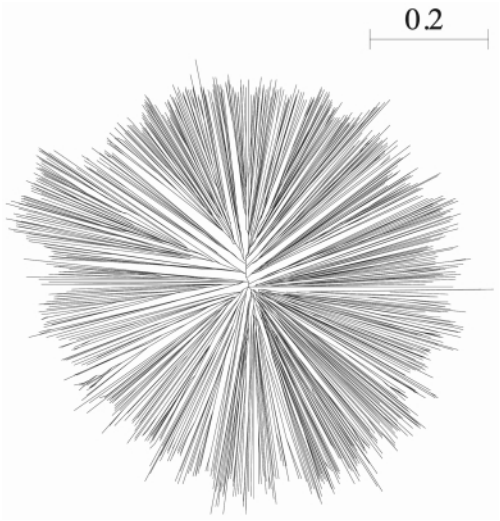

Figure 2. A tree calculated from an alignment of more than 1100 ring finger domains, using ClustalW 1.83. The full tree calculation, including the distance matrix calculation, took 22 s on a 1 GHz Pentium III. The output tree was displayed with Unrooted (18).

install, is user-friendly and maintains the portability of the previous generations through the NCBI Vibrant toolkit (ftp://ncbi.nlm.nih.gov/toolbox/ncbitools/). Numerous options are provided, such as the realignment of selected sequences or selected blocks of the alignment and the possibility of building up difficult alignments piecemeal, making ClustalX an ideal tool for working interactively on alignments.

Parallel versions of ClustalW and ClustalX have been developed by SGI (http://www.sgi.com/industries/sciences/chembio/resources/clustalw/parallel_clustalw.html), which show increased speeds of up to 10× when running ClustalW/X on 16 CPUs and significantly reduce the time required for data analysis. A number of other significant developments have been based on the ClustalW program. For example, ClustalNet (12) is a Clustal alignment CORBA server and DbClustal (13) is a program for aligning sequences detected by database searches, which uses local alignment information to anchor the global multiple alignment. DbClustal is available on the Web at http://www-igbmc.u-strasbg.fr/BioInfo/DbClustal and forms part of the WU-Blast2 (Washington University BLAST version 2.0) server at the EBI (http://www.ebi.ac.uk/blast2/).

ClustalWWW WEB SERVER

Numerous Web servers have exploited the command line interface of ClustalW, notably the EBI's ClustalWWW Web server, which currently runs between 2000–10 000 jobs/day and the SRS server at the same site (http://srs.ebi.ac.uk/), which has ClustalW built in. The EBI ClustalWWW interface provides extensive help, ranging from an introduction to multiple alignments for new users to detailed descriptions of each alignment option. An important factor in obtaining a high-quality alignment is the ability to change the numerous alignment parameters available in ClustalW. While the default values of the parameters have been optimised to work in the majority of cases, they are not necessarily optimal for any given alignment problem. In the ClustalWWW interface, all the options are easily accessible on the top page.

Sequences can be entered by either pasting them or by uploading a file from the user's local computer. In both cases, the sequences should be in one of seven different formats (GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or SWISS-PROT). Although users are encouraged to submit large numbers of sequences, there is no guarantee that the alignment will be completed within the job run limits. Therefore, users who experience problems when attempting to make very large alignments are advised to download the software and run it locally. In addition to the input format, the user can also specify the preferred output format for the multiple sequence alignment. The options are currently ALN, GCG, PHYLIP, PIR and GDE. It is also possible to configure the browser to automatically load the results files from ClustalW into a suitable external application. A list of some example URLs for obtaining such applications for MS-Windows, Macintosh and UNIX systems is provided in Table 1. Many commercial packages, e.g. the GCG package (Wisconsin Package, Genetics Computer Group, Madison, WI) and its X Window graphical user interface, SeqLab, can also accept ClustalW alignments.

The resulting multiple alignments can be displayed as either black and white or colour coded text. An example of the colour coded display is shown in Figure 1. The alignment consists of four oxidoreductase NAD binding domains. The colouring of residues takes place according to physicochemical criteria highlighting conserved positions in the sequences. A consensus line is also displayed below the alignment with the following symbols denoting the degree of conservation observed in each column: '*' (identical residues in all sequences), ':' (highly conserved column), '.' (weakly conserved column).

A recent enhancement to the ClustalWWW interface has been the addition of an option that allows the user to upload the results of ClustalW into an alignment editor, using a Java

Table 1. Example URLs of some external applications compatible with the output from ClustalW and ClustalX

Program name	Description	Operating systems	URL
BelVu	Multiple alignment viewer	UNIX	ftp://www.cgr.ki.se/cgr/groups/sonnhammer/Belvu.html
CINEMA	Multiple alignment editor	UNIX, Macintosh, MS-Windows	http://www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/
Se-Al	Multiple alignment editor	Macintosh	http://evolve.zoo.ox.ac.uk/software/Se-Al/main.html
GeneDoc	GCG MSF file viewer	MS-Windows	http://www.psc.edu/biomed/genedoc/
ClustalX	Graphical interface version of ClustalW	UNIX, Macintosh, MS-Windows	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/ ftp://ftp.ebi.ac.uk/pub/software/dos/clustalx/ ftp://ftp.ebi.ac.uk/pub/software/mac/clustalx/ ftp://ftp.ebi.ac.uk/pub/software/unix/clustalx/
AMAS	Multiple alignment analysis	UNIX	http://www.compbio.dundee.ac.uk/amas/
EMMA	EMBOSS open software interface	UNIX	http://www.emboss.org/
SeaView	Multiple alignment editor	UNIX, Macintosh, MS-Windows	http://pbil.univ-lyon1.fr/software/seaview.html
Phylip	Phylogeny	UNIX, Macintosh, MS-Windows	http://evolution.genetics.washington.edu/phylip.html
njplot	Tree viewer	UNIX, Macintosh, MS-Windows	http://pbil.univ-lyon1.fr/software/njplot.html
TreeView	Tree viewer	UNIX, Macintosh, MS-Windows	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Table 2. A comparison of execution times

Number of sequences	Original NJ		New NJ	
	NJ algorithm only	Complete multiple alignment	NJ algorithm only	Complete multiple alignment
200	0′ 6″	0′ 11″	0.1″	0′ 5″
500	6′ 55″	7′ 27″	1.1″	0′ 33″
1000	XXX	XXX	16″	2′ 18″

A comparison of two different implementations of the NJ algorithm (not including the time taken for the calculation of the distance matrix) for different sizes of alignments. The time required for the NJ algorithm depends only on the number of sequences, while the complete multiple alignment depends also on the lengths of the sequences. The timings reported here were all performed for sequences of ~40 residues. XXX, the algorithm did not complete. The timings were performed on a Compaq Alpha EV67 running True64 UNIX.

Applet called JalView (http://www.compbio.dundee.ac.uk/). JalView is a fully featured multiple sequence alignment editor which allows the user to perform further alignment analysis. Special features include the definition of sequence sub-groups, links to the SRS server at the EBI and an option to output the alignment as a colour postscript file for printing purposes.

As well as constructing multiple alignments, ClustalWWW can also calculate trees from a multiple alignment using the NJ method, a widely used and relatively fast algorithm that clusters sequences by minimising the sum of branch lengths. The resulting evolutionary relationships can be viewed either as cladograms or phylograms, with the option to display branch lengths (or 'tree graph distances').

NEW FEATURES

Both ClustalW and ClustalX are being actively maintained and updated. Recent enhancements have included the possibility of saving both alignments and phylogenetic trees in the NEXUS format (14) for compatibility with a number of phylogeny programs. Some work has also been done to optimise the alignment parameters, for example the Gonnet series of residue comparison matrices (15) is now used by default for protein sequence alignments. The latest version of the programs (version 1.83), which was released early this year, contained four main enhancements. The first modification is the facility to save the multiple alignment result as a FASTA format file, for compatibility with a number of other software packages. Another is to provide a percent identity matrix, which some users have asked for. A third new option is the possibility of saving the residue range in the output file when saving a user-specified range of the alignment. This is particularly useful when extracting a single domain from the alignment of multi-domain proteins. For example, in Figure 1 the NAD binding domain was extracted from a multiple alignment of the full-length oxidoreductase protein sequences and the residue range was automatically appended to the sequence names. Perhaps the most important enhancement in the latest version, however, is the incorporation of a faster implementation of the NJ algorithm used to construct guide trees during the multiple alignment process and also to construct phylogenetic trees based on the final alignment. Table 2 contains examples of the time required by the NJ algorithm for the construction of a phylogenetic tree from alignments containing different numbers of sequences. The increased speeds obtained mean that it is now possible to construct phylogenetic trees for very large sets of sequences, which were previously only feasible on very large computer systems. As an example, Figure 2 shows a phylogenetic tree constructed from an alignment of more than 1100 ring finger domain sequences taken from the PFAM database (16) entry PF00097. The new NJ implementation was written by T. Koike. An independent acceleration of the NJ algorithm has been published and is freely available as the QuickTree program (17). Though coding details differ, both

implementations addressed the major slow points of the original code and so will not produce combinatorial improvement.

ACKNOWLEDGEMENTS

We thank the many Clustal users who have provided feedback, bug reports and feature requests. T.K. would like to express his thanks to the Life Science Systems Division at Fujitsu Ltd for allowing him to participate in the bioinformatics research and development at the National Institute of Genetics. J.D.T. was supported by institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg and the Fond National de la Science (GENOPOLE).

REFERENCES

1. Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
2. Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Applic. Biosci.*, **4**, 11–17.
3. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
4. Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, **28**, 161–169.
5. Wilbur,W.J. and Lipman,D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl Acad. Sci. USA*, **80**, 726–730.
6. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy.* WH Freeman, San Francisco, CA, pp. 230–234.
7. Higgins,D.G. (1994) CLUSTAL V: multiple alignment of DNA and protein sequences. *Methods Mol. Biol.*, **25**, 307–318.
8. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.*, **8**, 189–191.
9. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
10. 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.*, **22**, 4673–4680.
11. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
12. Compagne,F. (2000) Clustalnet: the joining of Clustal and CORBA. *Bioinformatics*, **16**, 606–612.
13. Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
14. Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) NEXUS: an extensible file format for systematic information, *Syst. Biol.*, **46**, 590–621.
15. Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.*, **7**, 1323–1332.
16. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **1**, 276–280.
17. Howe,K., Bateman,A. and Durbin,R. (2002) Quick Tree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, **11**, 1546–1547.
18. Perrière,G. and Gouy,M. (1996) WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364–369.