

Multiple sequence alignment with Clustal X

The Clustal series of programs are widely used for multiple alignment and for preparing phylogenetic trees. The programs have undergone several incarnations, and 1997 saw the release of the Clustal W 1.7 upgrade and of Clustal X, which has a windows interface. Although we like to think that people use Clustal programs because they produce good alignments, undoubtedly one of the reasons for the programs' wide usage has been their portability to all computers. Portability can have a downside, and for many years the Clustal interface has had to be kept simple. Clustal X (Ref. 1) now provides a much nicer, graphical user interface (see Fig. 1) for X-, Mac and PC windows, while maintaining portability. It presents alignments in which residue conservation is shown in colour, and has a very useful new tool for marking poor regions of the alignment. In addition, the user can select such regions for realignment. Thus, Clustal X adds further flexibility to the available strategies for preparing multiple alignments.

A history of the Clustal programs

Clustal programs have undergone continual development for over ten years, so the versions available do not all give the same results. This can be confusing to new users; we therefore felt that a short history of Clustal development would help to clarify matters.

The first Clustal program^{2,3}, written by Des Higgins in 1988, was designed to perform efficient alignment on PCs, which then had feeble computing power by today's standards. It harnessed a memory-efficient, recursive alignment algorithm⁴ with the progressive alignment strategy introduced by Feng and Doolittle⁵ and Willie Taylor⁶. The essence of progressive alignment is to align the most-closely related sequences first and the difficult divergent ones last. The pre-comparison used a rapid FASTA-like word search, and the dendrogram was constructed using the UPGMA method^{7,8}. Simple text menus made Clustal easy to use. Although conceived as a 'poor man's' alignment program, available to anyone who could afford a microcomputer, Clustal was actually one of the most up-to-date programs of its type.

Alan Bleasby and Rainer Fuchs helped Higgins⁹ to revamp Clustal extensively for a brand new release, Clustal V, in 1992. They incorporated profile alignments (alignments of old alignments) and the facility to generate trees from the alignment by using the fast Neighbour-Joining method¹⁰. The user could also test the tree for robustness by using a simple bootstrap test of tree topology¹¹.

Julie Thompson and Toby Gibson collaborated on the third generation, Clustal W (W for Weighting, which was applied to sequences and gap penalties)¹², which was released in 1994. Clustal W looks very similar to Clustal V, but there are many internal differences. We incorporated position-specific gap penalties so that gap penalties can be lowered at hydrophilic residues and wherever gaps are already introduced into the alignment¹³. The sequence pre-comparison in Clustal W uses more-sensitive dynamic programming, which yields a much better dendrogram. The dendrogram itself is now calculated by Neighbour-Joining, which improves tree topology and provides a method for

weighting sequences on the basis of their divergences. In later releases of Clustal W, gap-penalty masks can guide the alignment (e.g. in cases where secondary-structure information is available; see Fig. 1). The program can also merge alignments together or add a list of new sequences to an old alignment, thereby becoming a very flexible aligning tool. Although we developed Clustal W to run on a local computer, numerous Web servers have been set up – for example, at the EBI (<http://www2.ebi.ac.uk/clustalw/>).

François Jeanmougin and Frédéric Plewniak joined us in the development of Clustal X. We focused on providing a modern windows-type interface, maintaining portability through the NCBI Vibrant Toolkit (which can be found at ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools/). Although the alignments produced are the same as those produced by the current release of Clustal W, the user can better evaluate alignments in Clustal X (Ref. 1). Within alignments, conserved columns are highlighted (using a colour scheme that the user can customize). Beneath the sequence alignment, Clustal X provides a plot of residue conservation. Quality-analysis tools that highlight misaligned regions are also available. The user can then target problems by realigning either selected sequences or selected blocks of the alignment and build up difficult alignments



Figure 1

Screenshot of a session with Clustal X in split-window mode for profile alignment. Archaeal TFIIIB sequences (lower window) are aligned with prealigned eukaryotic TFIIIBs (upper window). A structure mask from the solved structure of TF2B_Human has been applied. Structural information and the quality curves for each alignment are displayed. Horizontal scrolls are locked. a/A, α -helix.

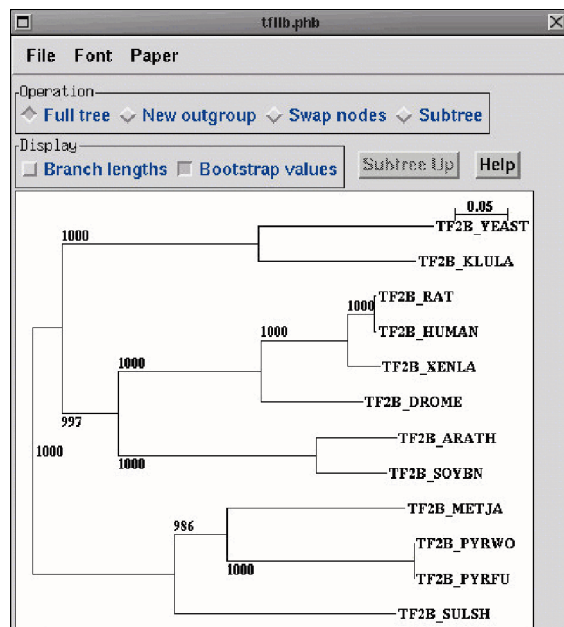


Figure 2

Screenshot of NJplot displaying a tree calculated by Clustal X from the TF1IB alignment generated in Fig. 1. The root is placed on the branch that links eukaryotes and archaea. Branch lengths are proportional to sequence divergence and can be measured relative to the bar shown (top right). Branch labels record the stability of the branches over 1000 bootstrap replicates.

piecemeal, so that problem areas can be dealt with one at a time. Clustal X is therefore a tool for working on multiple alignments, rather than simply an alignment program.

Getting started with Clustal X

The Clustal W and Clustal X programs have self-explanatory layouts, and on-line help is available, so that using the programs should not be difficult. For inexperienced users, the chief hurdle seems to be getting the program to read their sequences. Sequences must be collected into a single file in a format that Clustal can read. The simplest format is FASTA format, but the EMBL and SWISS-PROT database formats can be read directly. Usually, the set of sequences will be exported from some other sequence-analysis package (most of which support FASTA format). Web users can extract sets of database sequences via SRS servers (e.g. from <http://srs.ebi.ac.uk/>). Finally, if need be, the user can assemble and edit sequences in a word-processing package, provided that the sequences are saved as text with linebreaks. Once loaded in Clustal X, the sequences can be aligned immediately using default parameter settings. Note that the default parameters – in particular the gap penalty settings – do not always give the best alignment. New users should therefore do

some trial runs, using higher and lower gap penalties to see how the program performs. In difficult alignment cases, it will usually pay to test different parameters. The alignment-quality tools in Clustal X can help greatly in evaluating alignment results.

When and how to use Clustal X – and when not to!

The wide usage of Clustal W and X might seem to imply that they always align sequences well. In fact, this is not always so. The alignment algorithm has been optimized to align sets of sequences that are entirely collinear – that is, the sequences have the same protein domains, and these domains are in the same order. If this condition is not met (and it often is not), Clustal X can produce serious misalignments. The user must give a little thought to the nature of the sequence set.

When to use Clustal X

The Clustal X program can be used to align any group of protein or nucleic acid sequences that are related to each other over their entire lengths. However, some problems can still be encountered.

Divergent sequences. Clustal tries to align the most-closely related sequences first, in order to build a representative profile of the family. Divergent sequences are delayed by default until this profile is available. If you only have divergent members of a family, this can result in most (if not all) of the sequences being delayed, and the progressivity of the alignment is lost. In this case, you can change the delay parameter or use the Profile Alignment Mode to drive the alignment order yourself. In extreme cases, the sequences can simply be too divergent to be correctly aligned.

Composition bias. Clustal uses position-specific gap penalties to help introduce gaps in hydrophilic regions of the alignment. By default, the residues G, P, S, N, D, Q, E, K and R are considered to be hydrophilic. If your sequences show a bias in one (or more) of these residues, you should remove that residue from the list in the Protein Gap Parameters menu.

Few sequences. An alignment of a small number (<10) of very distantly related proteins could be unreliable. You should check such an alignment carefully.

When not to use Clustal X (or when to use it with great caution)

The sequences do not share common ancestry. This is attempted surprisingly often – mostly, but not always, by accident. (This most often happens to us when we extract a set of sequences by using a keyword search.)

The sequences have large, variable, N- and C-terminal overhangs (e.g. kinesins). The unconserved termini must be removed, or the 'Use Negative Matrix' option must be invoked – otherwise a completely false alignment can result.

The sequences are partially related. Multidomain proteins that have complex evolutionary histories often share some, but not all, of the domain set. The alignments produced in these cases can be unpredictable.

The sequences including short non-overlapping fragments. Sometimes, people attempt to align a set of short fragments to a single reference sequence. This will not work in Clustal X. If the sequences are non-overlapping, they are of course completely unrelated, and the guide tree and sequence weighting generated are nonsense! Find another way to do this.

The alignment quality can be checked using the analysis tools provided by Clustal X, as well as the very powerful residue-colouring scheme. The alignment process can be traced by saving the progress messages in an optional log file. From here, you can see which sequences have been delayed in the multiple-alignment order until the core profile has been built. You can also examine the guide tree (*.dnd file) using NJplot, but remember that this is not a reliable phylogenetic tree.

Displaying trees with NJplot

Clustal X can calculate trees by using the Neighbour-Joining method¹⁰, a widely used and relatively fast algorithm that clusters sequences by minimizing the sum of branch lengths. However, Clustal X does not display trees. A simple tree-display program, NJplot (Ref. 14), is included in the Clustal-X distribution package. Like Clustal X, NJplot is available for all computer platforms. NJplot reads the phylip-format tree output of Clustal X and displays trees as dendrograms. Basic manipulations of the text labels, branch flips and

rerooting of the tree can be performed, but the underlying tree topology cannot be changed. Note that Neighbour-Joining trees are unrooted, so the user must decide if there is a biologically valid root or not. Figure 2 shows a tree generated by Clustal X and displayed by NJplot. Other useful tree-display packages include TreeTool, available for Sun UNIX only¹⁵, and TreeView, which runs on Macs/PCs¹⁶. These programs can display trees radially as well as in dendrograms. Clustal alignments can also be used as input for more-comprehensive tree packages, such as PHYLIP¹⁷, PhyloWin¹⁸ or PAUP (which was developed by D. L. Swofford at the Smithsonian Institute).

Conclusion

In this article, we provide some guidance that we hope will prove useful to Clustal users. In the not-too-distant future, progressive alignment – the dominant strategy for the last ten years – will probably be rendered obsolete. Iterative alignment strategies, such as PRRP¹⁹ and SAGA²⁰, are reported to perform as well as, or better than, Clustal X for small numbers of sequences but are currently still too slow to handle large datasets. More-efficient iterative strategies, harnessed to increasingly powerful desktop computers, might soon be able to provide high-quality alignments for everyone who needs them.

Acknowledgements

Clustal development has been supported by Trinity College, Dublin, the EMBL and ICGEB, Strasbourg. Current Clustal development is supported by funds from INSERM, CNRS and Ministère de l'Éducation nationale de la Recherche et de la Technologie and the EMBL. We thank former developers for their input and the Clustal users for their informative feedback.

References

- 1 Thompson, J. D. et al. (1997) *Nucleic Acids Res.* 25, 4876–4882
- 2 Higgins, D. G. and Sharp, P. M. (1988) *Gene* 73, 237–244
- 3 Higgins, D. G. and Sharp, P. M. (1989) *Comput. Appl. Biosci.* 5, 151–153
- 4 Myers, E. W. and Miller, W. (1988) *Comput. Appl. Biosci.* 4, 11–17
- 5 Feng, D. F. and Doolittle, R. F. (1987) *J. Mol. Evol.* 25, 351–360
- 6 Taylor, W. R. (1988) *J. Mol. Evol.* 28, 161–169
- 7 Wilbur, W. J. and Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U. S. A.* 80, 726–730
- 8 Sneath, P. H. A. and Sokal, R. R. (1973) in *Numerical Taxonomy* (pp. 230–234), W. H. Freeman
- 9 Higgins, D. G., Bleasby, A. J. and Fuchs, R. (1992) *Comput. Appl. Biosci.* 8, 189–191

- 10 Saitou, N. and Nei, M. (1987) *Mol. Biol. Evol.* 4, 406–425
- 11 Felsenstein, J. (1985) *Evolution* 39, 583
- 12 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucleic Acids Res.* 22, 4673–4680
- 13 Thompson, J. D. (1995) *Comput. Appl. Biosci.* 11, 181–186
- 14 Perriere, G. and Gouy, M. (1996) *Biochimie* 78, 364–369
- 15 Maidak, B. L. et al. (1997) *Nucleic Acids Res.* 25, 109–111
- 16 Page, R. D. (1996) *Comput. Appl. Biosci.* 12, 357–358
- 17 Felsenstein, J. (1989) *Cladistics* 5, 164–166
- 18 Galtier, N., Gouy, M. and Gautier, C. (1996) *Comput. Appl. Biosci.* 12, 543–548
- 19 Gotoh, O. (1996) *J. Mol. Biol.* 264, 823–838
- 20 Notredame, C. and Higgins, D. G. (1996) *Nucleic Acids Res.* 24, 1515–1524

FRANÇOIS JEANMOUGIN AND JULIE D. THOMPSON

Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP163 67404, Illkirch CEDEX, France.
Email: jeanmougin@igbmc.u-strasbg.fr

MANOLO GOUY

Laboratoire de Biometrie, Génétique et Biologie des Populations, Université Lyon I, UMR CNRS 5558, 69622 Villeurbanne, France.


DESMOND G. HIGGINS

Dept of Biochemistry, University College, Cork, Ireland.

TOBY J. GIBSON

European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany.

The Post-docs'

EPISTEMIC DICTIONARY of Rap  No 54

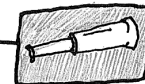
I IS FOR INNER SPACE...

THE CRADLE'S CELLING FIRST DEFINES THE LIMITS OF THE WORLD...
A UNIVERSAL DISTANCE DELICATELY UNFURLED,
LEARNING THAT THIS DISTANCE IS NO DISTANCE WHATSOEVER...
THE CHILD'S FIRST MODEL IS LEFT IN SEARCH OF BETTER



VENTURING AT NIGHT IN A DARK + CLOUDLESS PLACE
A NEWER COSMIC ORDER SLIPS QUIETLY INTO "SPACE"

A SYSTEM OF SOLAR RINGS OR ORBITS EXTENDING FURTHER
THE ADOLESCENT LEARNS THAT OUTER SPACE KNOWS NO CONCRETE BORDER...



CATAPULTED THROUGH THE ENIGMA OF INFINITE SPACE...
THE ADULT GAZE RETURNS TO ITS STARTING PLACE...
CONCENTRATION UPON A SPOT – THIS INNER ISLE OF TIME
COSMIC ORDER TRACED, A CIRCUMSCRIBING LINE...



SEEING FOR THE FIRST TIME THAT WHICH WAS INVISIBLE...
UNITING FOR THE FIRST TIME THAT WHICH WAS DIVISIBLE...
BREATHING TRULY FOR THE FIRST TIME THAT WHICH IS OUR MEDIUM...
OUTER SPACE TO INNER SPACE IN PERFECT EQUILIBRIUM.

NEXT MONTH "J"

Pete Jeffs is a freelancer working in Paris, France.