BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 21 2007, pages 2947–2948
doi:10.1093/bioinformatics/btm404

Sequence analysis

Clustal W and Clustal X version 2.0

M.A. Larkin[1], G. Blackshields[1], N.P. Brown[3], R. Chenna[3], P.A. McGettigan[1], H. McWilliam[4], F. Valentin[4], I.M. Wallace[1], A. Wilm[1], R. Lopez[4], J.D. Thompson[2], T.J. Gibson[3] and D.G. Higgins[1,*]

[1]The Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland, [2]Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France, [3]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [4]EMBL Outstation-European Bioinformatics Institute, Wellcome Trust Genome Campus Hinxton, Cambridge, CB10 1SD, UK

Received on June 27, 2007; revised on August 3, 2007; accepted on August 3, 2007

Advance Access publication September 10, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: The Clustal W and Clustal X multiple sequence alignment programs have been completely rewritten in C++. This will facilitate the further development of the alignment algorithms in the future and has allowed proper porting of the programs to the latest versions of Linux, Macintosh and Windows operating systems.

Availability: The programs can be run on-line from the EBI web server: http://www.ebi.ac.uk/tools/clustalw2. The source code and executables for Windows, Linux and Macintosh computers are available from the EBI ftp site ftp://ftp.ebi.ac.uk/pub/software/clustalw2/

Contact: clustalw@ucd.ie

1 INTRODUCTION

Multiple sequence alignments are now one of the most widely used bioinformatics analyses. They are needed routinely as parts of more complicated analyses or analysis pipelines and there are several very widely used packages, e.g. Clustal W (Thompson *et al.*, 1994), Clustal X (Thompson *et al.*, 1997), T-Coffee (Notredame *et al.*, 2000), MAFFT (Katoh *et al.*, 2002) and MUSCLE (Edgar, 2004). Clustal is also the oldest of the currently most widely used programs having been first distributed by post on floppy disks in the late 1980s. It was initially written in Microsoft Fortran for MS-DOS and originally ran on IBM compatible personal computers as four separate executable programs, Clustal1–Clustal4 (Higgins and Sharp, 1988, 1989). These were later rewritten in C and merged into a single program, Clustal V (Higgins *et al.*, 1992), that was distributed for VAX/VMS, Unix, Apple Macintosh and IBM compatible PCs. These programs were distributed from the EMBL File server (Stoehr and Omond, 1989), an e-mail and FTP server, based at the EMBL in Heidelberg, Germany.

The current Clustal programs all derive from Clustal W (Thompson *et al.*, 1994), which incorporated a novel position-specific scoring scheme and a weighting scheme for down weighting over-represented sequence groups. The 'W' stands for 'weights'. These programs have been amended and added to many times since 1994 in order to increase functionality and to increase sensitivity. The user-friendliness has also been greatly enhanced by the addition, in 1997, of a full graphical user interface (Thompson *et al.*, 1997). This has made the code complicated to maintain and develop, as the graphical interface must be constantly modified and recompiled for new operating systems and desktop environments (Windows, Macintosh, VMS, Unix and Linux).

By the late 1990s, Clustal W and Clustal X were the most widely used multiple alignment programs. They were able to align medium-sized data sets very quickly and were easy to use. The alignments were of sufficient quality not to require manual editing or adjustment very often. This situation changed greatly with the appearance of the first custom made benchmark test set for multiple alignment programs, BAliBASE (Thompson *et al.*, 1999). This was followed by the appearance of T-Coffee which was able to make very accurate alignments of very divergent proteins but only for small sets of sequences, given its high computational cost. With the increase in processing speed of desktop computers, and subsequent optimisation of the T-Coffee code, the latter is now practical for routine use on moderately sized alignment problems. More recently, MAFFT and MUSCLE appeared; which were, initially, at least as accurate as Clustal, in terms of alignment accuracy, but which were also extremely fast; and able to align many thousands of sequences. Over the past 4 or 5 years, these programs have also gradually become more and more accurate with difficult alignments. Nonetheless, Clustal W and Clustal X continue to be very widely used, increasingly on websites. The EBI Clustal site, gets literally millions of multiple alignment jobs per year.

It is in this context that we developed Clustal W 2.0 and Clustal X 2.0. These programs were rewritten in C++ with a simple object model in order to make it easier to maintain the code and more importantly, to make it easier to modify or even replace some of the alignment algorithms. We have produced two new programs which are very similar in look and feel to the older version 1.83 programs but which can now be managed more easily. We have also made some minor adjustments to the

*To whom correspondence should be addressed.

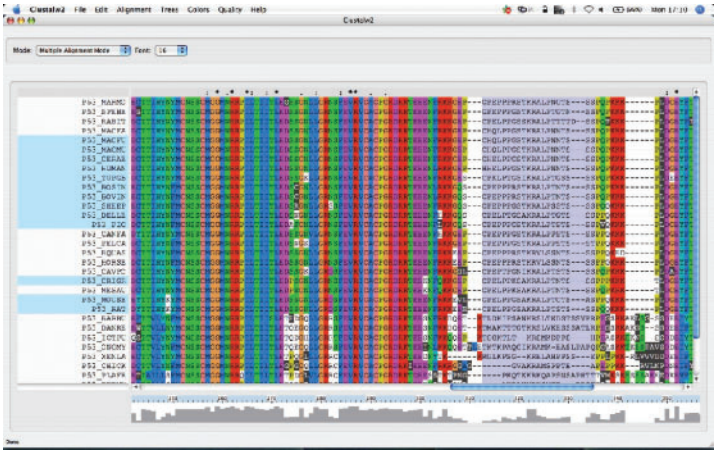

Fig. 1. ClustalX 2.0 Screenshot on Mac OS X.

alignment algorithms. We have included new code for UPGMA guide trees as an alternative to the usual Neighbor-Joining guide trees. This helps speed up the alignment of extremely large data sets of tens of thousands of sequences. We have also included an iterative alignment facility to increase alignment accuracy.

Clustal X 2.0 is the new version of the Clustal X graphical alignment tool. The original Clustal X was developed using NCBI's vibrant toolbox. The vibrant toolbox is no longer supported which led to problems compiling Clustal X on newer versions of operating systems. The graphical interface sections of Clustal X 2.0 have been completely rewritten using the Qt GUI toolbox. Qt is an easy-to-use, multi-platform C++ GUI toolkit. The code need only be compiled once on each of the platforms. The Qt toolbox provides a native look and feel on Windows, Linux and Mac platforms. Clustal X 2.0 has the same functionality as Clustal X.

2 NEW FEATURES

Two new options have been included in Clustal W 2.0, to allow faster alignment of very large data sets and to increase alignment accuracy. The default options of Clustal W And Clustal X 2.0 are the same as Clustal W 1.83, and will give the same alignment results.

The guide trees in Clustal have been calculated using the Neighbor-Joining (NJ) method, for the past 10 years or so. In the earliest versions of the program UPGMA was used. UPGMA is faster than NJ but prone to cluster long branches together when evolutionary rates are very unequal in different lineages. Both algorithms have complexity of $O(N^2)$ but UPGMA is faster for a given data set and the difference becomes pronounced with very large N. On a standard desktop PC, it is possible to cluster 10 000 sequences in less than a minute using UPGMA, while NJ would take over an hour. We have reimplemented a very efficient algorithm for UPGMA which can be called by using the command line option '-clustering=UPGMA'. It is marginally less accurate on the Balibase benchmark, but on large alignments (e.g. 10 000 globin sequences) this is offset by the savings in processing time (2 h versus 12 h).

Iteration is a quick and effective method of refining alignments. A 'remove first' iteration scheme, which optimizes the Weighted Sum of Pairs (WSP) score, has been included in this version of Clustal. During each iteration step, each sequence is removed from the alignment in turn and realigned. If the WSP score is reduced then the resulting alignment is retained. The iteration scheme can be used to either refine the final alignment or at each step in the progressive alignment. Iterating during the progressive alignment tends to be more accurate but also much more time consuming as there are 2N-3 nodes in the guide tree. The command line option '-Iteration=Alignment' refines the final alignment, while the option '-Iteration=Tree' incorporates the scheme into the progressive alignment. The number of iteration cycles is set via the command line option '-numiters' (default is 3).

ACKNOWLEDGEMENT

This work was mainly funded by Science Foundation Ireland.

Conflict of Interest: none declared.

REFERENCES

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Higgins,D.G. and Sharp,P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.*, **5**, 151–153.

Higgins,D.G. *et al.* (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.*, **8**, 189–191.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Notredame,C. *et al.* (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Stoehr,P.J. and Omond,R.A. (1989) The EMBL Network File Server. *Nucleic Acids Res.*, **17**, 6763–6764.

Thompson,J.D. *et al.* (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.

Thompson,J.D. *et al.* (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.

Thompson,J.D. *et al.* (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.