BIOINFORMATICS APPLICATIONS NOTE

Vol. 20 no. 3 2004, pages 426–427
DOI: 10.1093/bioinformatics/btg430

# The Jalview Java alignment editor

Michele Clamp[1,2,4,*], James Cuff[1,2], Stephen M. Searle[1,2] and Geoffrey J. Barton[2,3,4]

[1] The Wellcome Trust Sanger Institute and [2] The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK, [3] School of Life Sciences, University of Dundee, Dow St, Dundee, DD1 5EH, UK and [4] The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

Received on April 16, 2003; revised on August 4, 2003; accepted on August 6, 2003
Advance Access Publication January 22, 2004

**ABSTRACT**

**Summary:** Multiple sequence alignment remains a crucial method for understanding the function of groups of related nucleic acid and protein sequences. However, it is known that automatic multiple sequence alignments can often be improved by manual editing. Therefore, tools are needed to view and edit multiple sequence alignments. Due to growth in the sequence databases, multiple sequence alignments can often be large and difficult to view efficiently. The Jalview Java alignment editor is presented here, which enables fast viewing and editing of large multiple sequence alignments.

**Availability:** The Jar file and source code for `Jalview` is freely available via the World Wide Web at http://www.jalview.org. A Jalview mailing list is also available by e-mailing majordomo@sanger.ac.uk with *subscribe Jalview* in the body of the mail.

**Contact:** michele@sanger.ac.uk

## INTRODUCTION

The alignment of biological sequences has a long history and the development of automatic techniques has eased the difficulty of generating alignments from unaligned sequences. However, even the best multiple sequence alignment methods only achieve <50% accuracy per position in the alignment of sequences with <20% identity (Thompson *et al*., 1999). Biologists can often use other information about the sequence and structure of a family of proteins to improve a multiple sequence alignment. Therefore, biologists striving for the best possible alignment will often need to edit manually an automatically generated alignment.

There exist a large number of software packages that allow the viewing of multiple sequence alignments. These include Belvu, Alscript (Barton, 1993), ClustalX (Thompson *et al*., 1997) and Chroma (Goodstadt and Ponting, 2001). These packages do not allow editing of multiple sequence alignments. Although alignments can be edited in word processing software, such as Microsoft Word or emacs, it is often difficult to see conserved patterns without a specific colouring of the alignment that these programs do not provide. In addition, specialized multiple sequence alignment editors can provide extra features for the user including grouping and analysis of the conservation patterns in the alignment. A small number of software packages exist that allow editing of multiple sequence alignments, such as GeneDoc (Nicholas and Nicholas, 1997, http://www.cris.com/~Ketchup/genedoc.shtml), BioEdit, Seaview (Galtier *et al*., 1996), MPSA (Blanchet *et al*., 2000), ANTHEPROT (Deleage *et al*., 2001) and CINEMA (Parry-Smith *et al*., 1998) amongst others. Of these, CINEMA has most similarities with Jalview as it is written in Java. However, Jalview provides extra functionality with the ability to calculate trees, conservation within subfamilies and on the fly pairwise alignments.

The Jalview program was written with the following design goals in mind. First, it should be platform independent; second, it should be fast and capable of editing of large multiple sequence alignments without significant degradation of performance; and third, it should allow multiple integrated views of the alignment and other data. These goals were addressed by coding the software in the platform independent Java version 1.1 language.

## FEATURES OF JALVIEW

Jalview has a rich functionality based on its core alignment viewing and editing options. These features are described in outline in the following section. Jalview can input and output multiple sequence alignments in a variety of common formats including MSF, aligned Fasta and Clustal format. Once loaded into Jalview the alignments are coloured by default according to the ClustalX colouring scheme (Thompson *et al*., 1997). A number of other colouring options are available via the edit menu including a user configurable scheme. If the user does not have a sequence alignment, a set of unaligned sequences can be aligned using ClustalW either locally or via the web at the EBI ClustalW server (Brooksbank *et al*., 2003).

[*] To whom correspondence should be addressed.

Editing multiple sequence alignments in Jalview simply requires the user to drag residues to the left to remove gaps and to the right to insert gaps at the cursor position. Editing can be carried out on multiple sequences by applying group selection, found in the edit menu. Grouping sequences can speed up editing of large numbers of similar sequences. Jalview allows users to calculate UPGMA or neighbour-joining trees (Saitou and Nei, 1987). Upon selecting this option, a new window is opened to display the tree. These trees can be used to reorder the sequences in a multiple alignment as well as to select groups of sequences for group editing.

Sequence features on a multiple sequence alignment can be viewed in Jalview. If the sequence identifiers in the alignment are Swiss-Prot/TrEMBL identifiers Jalview can access the EBI website via SRS to download feature table elements and display them on the alignment (Brooksbank *et al*., 2003). By right-clicking on Swiss-Prot/TrEMBL sequence identifiers in the alignment window, the entry is retrieved from an SRS server and displayed in Jalview's lightweight web-browser. If a structure is known for one of the sequences in the alignment, this can also be downloaded from an SRS server and displayed in the Jalview structure viewer. The colour scheme from the alignment is projected on to the structure to highlight regions of conservation.

Principal component analysis (PCA) can help in understanding the relationship between sequences of an alignment. The method of clustering sequences implemented in Jalview is based on the method applied in SequenceSpace (Casari *et al*., 1995). When PCA is selected from the calculate menu a PCA viewer window is created that shows the sequences projected on to the first three eigenvectors. Clicking on points in the PCA window selects the corresponding sequence in the alignment window and in the tree window if it is visible.

Multiple sequence alignments often contain sub-families of sequences and applying a colour scheme across the whole alignment can make it difficult to identify these families. Jalview allows the user to define sequence groups easily by using the tree panel. Clicking on the tree defines a maximum distance apart any two sequences can be in a group and the alignment is split into groups accordingly. Conservation across each group can then be calculated by considering the different amino acid properties across each column in the group (Zvelebil *et al*., 1987). Columns that are most conserved have the most intense colour schemes fading to no colouring at all for unconserved columns.

The Jalview software was originally written in 1997 and is now widely used with over 20 000 downloads. It has been used to produce publication quality alignment figures as well as to provide a platform independent method to view multiple sequence alignments by databases, such as Pfam (Bateman *et al*., 2002). Jalview is run as an applet via the Pfam web pages, for an example see http://www.sanger.ac.uk/Software/Pfam/cgi-bin/getacc.pl?PF00045. Jalview is also used by the EBI ClustalW server (Brooksbank *et al*., 2003) as well as in the Apollo genome annotation editor (Lewis *et al*., 2002). The supplementary information available at http://www.jalview.org/bioinf/supp.html includes a figure showing a screenshot of the main Jalview windows.

## REFERENCES

Barton,G.J. (1993) ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.*, **6**, 37–40.

Bateman,A., Birney,E. *et al*. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Blanchet,M. *et al*. (2000) MPSA: integrated system for multiple protein sequence analysis with client/server capabilities. *Bioinformatics*, **16**, 286–287.

Brooksbank,C., Camon,E. *et al*. (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **31**, 43–50.

Casari,G., Sander,C. *et al*. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Galtier,N., Gouy,M. *et al*. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–538.

Goodstadt,L. and Ponting,C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.

Deleage,G. *et al*. (2001) ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. *Comput. Biol. Med.*, **31**, 259–267.

Lewis,S.E., Searle,S.M. *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.

Nicholas,K.B. and Nicholas,H.B.Jr (1997) GeneDoc: Analysis and Visualization of Genetic Variation.

Parry-Smith,D.J., Payne,A.W. *et al*. (1998). CINEMA—a novel colour interactive editor for multiple alignments. *Gene*, **221**, GC57–GC63.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Thompson,J.D., Gibson,T.J. *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., Plewniak,F. *et al*. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Zvelebil *et al*. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957–961.