BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 9 2009, pages 1189–1191
doi:10.1093/bioinformatics/btp033

Sequence analysis

Jalview Version 2—a multiple sequence alignment editor and analysis workbench

Andrew M. Waterhouse[1,†,‡], James B. Procter[1,†], David M. A. Martin[1], Michèle Clamp[2] and Geoffrey J. Barton[1,*]

[1]School of Life Sciences Research, College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK and [2]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Received and revised on November 24, 2008; accepted on January 8, 2009

Advance Access publication January 16, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Jalview Version 2 is a system for interactive WYSIWYG editing, analysis and annotation of multiple sequence alignments. Core features include keyboard and mouse-based editing, multiple views and alignment overviews, and linked structure display with Jmol. Jalview 2 is available in two forms: a lightweight Java applet for use in web applications, and a powerful desktop application that employs web services for sequence alignment, secondary structure prediction and the retrieval of alignments, sequences, annotation and structures from public databases and any DAS 1.53 compliant sequence or annotation server.

Availability: The Jalview 2 Desktop application and JalviewLite applet are made freely available under the GPL, and can be downloaded from www.jalview.org

Contact: g.j.barton@dundee.ac.uk

1 INTRODUCTION

Sequences of DNA, RNA and proteins are the fundamental currency of modern biological research that links the different levels of the biological hierarchy, from gene to 3D structure. Multiple sequence alignments (MSAs) permit the identification of common features between species or identify functionally important residues. MSAs provide the foundation for a range of computational methods including the prediction of protein secondary structure and solvent accessibility, functional sites and interaction sites. MSAs are also the essential first step in studying molecular phylogeny and the identification of genomic rearrangements. In journal publications, MSAs provide a convenient framework for displaying common features and complex annotations relating to sequences and their functions. It is therefore important to obtain the best alignment possible. Many multiple alignment techniques exist (Notredame, 2007), but no single method is perfect for all situations (Blackshields *et al*., 2006; Raghava *et al*., 2003). As a consequence, all alignments require inspection and interpretation, and often adjustment by hand, in order to produce an alignment that best represents the biological context of the sequences. Editing tools are essential for this task, not least because they provide visual feedback on an alignment's quality in the light of all known and computationally predicted annotation.

Jalview Version 1.0 (Clamp *et al*., 2004) was an alignment editor first developed in 1996 as an advance over static alignment visualization tools such as ALSCRIPT (Barton, 1993). As well as alignment editing, colouring and generation of figures as postscript or HTML, it included methods for alignment conservation analysis, phylogenetic tree construction and a simple linked view of 3D structure that could colour residues in the same way as the alignment. Many alignment formats were supported, and feature annotation extracted from SwissProt (Boeckmann *et al*., 2003) flat-files could be plotted on the alignment to highlight important regions of a sequence. Jalview V1.0 gained a strong following, and was best known in its lightweight web applet form which was adopted as an alignment viewing/editing tool by many web sites worldwide including major databases such as PFAM (Finn *et al*., 2008) and SRS (Etzold *et al*., 1996). It has also been embedded in stand-alone analysis tools such as ModView (Ilyin *et al*., 2003). However, the original program had many limitations—only one multiple alignment could be edited at a time, and an alignment's colouring and tree-based conservation analysis could only be exported as a figure, not stored and returned to later. Introduction of new functionality to the program was also difficult. Jalview 1's software architecture was developed for optimum performance within the constraints of the Java runtime environment; and the addition of extensions could become complex and lead to unmaintainable code. In summary, Jalview V.1.0's capabilities are now insufficient for the larger, longer and more detailed analysis tasks that a researcher may now routinely perform. Stability, usability and extensibility are now also of prime importance for software used in research, and to this end, we re-engineered the original Jalview code to develop Jalview Version 2 (JV2).

2 IMPLEMENTATION

The new JV2 software architecture and alignment-rendering model provides the foundation for two JV2 program flavours: JalviewLite (JVL) and Jalview Desktop (JVD). JVL is a web optimized, Java 1.1 compliant applet that replaces Jalview V1.0 where it is used on a

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

‡Present address: Genome Exploration Research Group, RIKEN Omics Science Center and the Functional RNA Research Program, Frontier Research System, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho Tsurumi-ku Yokohama, Kanagawa 230-0045, Japan.

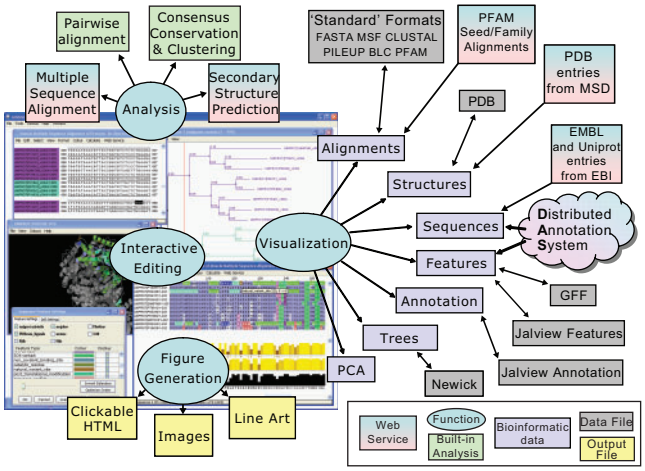

Fig. 1. Capabilities of the Jalview 2 desktop application. Ovals depict major capabilities: visualization, interactive editing, analysis and WYSIWYG figure generation. Arrows connect bioinformatics data handled by JV2 with flat-file or web-service data sources. Analysis includes built-in alignment conservation and tree building algorithms, and web services for MSA and secondary structure prediction methods. The screenshot of the application shows two sets of sequences for proteins in the lactate dehydrogenase family. One contains an alignment of protein sequences retrieved from the Uniprot database, the other their coding sequences retrieved from the EMBL database. Interactive highlighting shows the region corresponding to the amino acid or codon position near the mouse pointer in both the alignment windows and the Jmol structure display of a PDB record associated with one protein.

web page. In contrast, JVD is a fully-fledged desktop application that can be installed easily on the user's machine and launched in batch or interactive mode from the command line, or started *via* Java WebStart (Java 1.4 or later). The capabilities of the JVD that are summarized in Figure 1 and described below include the ability to generate high-quality alignment figures for publication, and to exploit web services for data retrieval and analysis. JVL and JVD both utilize Jmol, an open source molecular graphics viewer, to present linked views of PDB files associated with an aligned sequence.

The core JV2 functionality present in JVL and JVD provides significantly enhanced editing and viewing capabilities when compared to JV1. Interactive editing, colouring and annotation can be performed *via* the mouse or in a keyboard-editing mode. Alignment edits can be undone, and any number of independent views may be created in tabs or as separate windows opened on the same alignment. Navigation in a view is facilitated by an overview window, and each view also has its own layout and display settings. Specific sequences or columns can be hidden from a view, and arbitrary regions may be selected for analysis either by the built-in algorithms or remote web services, cut or copied to another alignment, or defined as named groups and coloured with one of 11 built-in or user-defined alignment colour schemes, or shaded by conservation or quantitative alignment annotation. Annotation rows may be interactively created and displayed below the columns of the alignment. They may contain labels, secondary structure symbols, coloured histograms or line graphs. Sequence features may also be overlaid onto an alignment. Non-positional information, such as sequence database accession numbers and literature references, are viewed as a tooltip displayed when the mouse hovers over a sequence's ID. Positional features such as metal ion binding sites are rendered as transparent or opaque shading over the visible regions of the alignment. Features may also be edited interactively or created from the results of a regular expression search over the alignment. New JV2 file parsers have been developed to generate annotated alignments from Stockholm alignment files, to import features from GFF, and to read and write Newick formatted phylogenetic trees. Three new formats have also been developed. Sequence regions and groups, colouring and alignment annotation are recorded in Jalview annotation files, whereas Jalview feature files are used to exchange sequence feature annotation. The JVD also supports an additional XML document format, the Jalview Project Archive, which enables all alignments, trees, structures, views and DNA/protein/structure mappings to be recorded and returned to at a later date.

2.1 Embedding JV2 in web applications

The JVL applet provides JV2's core MSA visualization, annotation, analysis and editing facilities as a lightweight web application component. Input data and initial display settings are specified using the comprehensive set of start-up parameters. Furthermore, a Javascript API (described on the web site) allows access to user selections, alignment and annotation data, and control of group and feature display settings in a particular alignment view. JVL has been successfully deployed on many servers, including MyHits (Pagni *et al*., 2007), and the structural genomics target optimization pipeline, TarO (Overton *et al*., 2008). Interaction with web application developers has been an important influence on development. For example, the sequence feature settings interface arose to support the needs of MACSIMS (Thompson *et al*., 2006).

2.2 Web services access from the Jalview desktop

Figure 1 provides an overview of the capabilities of the JVD. Its primary role is to support alignment creation, editing and in-depth analysis, and enables the visual integration of local and distributed sources of sequence annotation and structural data. Command-line parameters passed via Java Webstart provides a route for the JVD to be launched from the JVL or directly by a bioinformatics web application, but it can also access public sequence, structure and alignment databases with WSDbFetch (Pillai *et al*., 2005) to retrieve or transfer database accessions and annotation from their records. Menus on the JVD interface enable the researcher to gather sequence and annotation data from external databases, and utilize Jalview's own dedicated SOAP web services for sequence alignment with ClustalW (Thompson *et al*., 1994), Muscle (Edgar, 2004) and MAFFT (Katoh *et al*., 2005) and secondary structure prediction with Jpred3 (Cole *et al*., 2008).

JVD also interacts with Distributed Annotation System (DAS) servers conforming to the DAS 1.53 specification (Dowell *et al*., 2001; Prlic *et al*., 2007). Sequence and annotation servers can be manually added or discovered from the public DAS server registry. Sequences can be retrieved from or matched against any registered sequence source, and features retrieved from annotation sources mapped onto the alignment sequence's local coordinate frame.

3 DISCUSSION

The first version of JV2 appeared in May 2005, and after the release of Jalview 2.4 in September 2008, a search for 'Jalview' in Google returns over 450 000 hits. Furthermore, estimates derived from HTTP logs suggest that the Jalview Desktop is launched between 1500 and 2500 times per week. Naturally, Jalview is not the only program to have editing/analysis and display features, though it is perhaps surprising that relatively few of the 25 or so interactive programs distributed since 1985 appear as widely used. Many, such as HOMED (Stockwell and Petersen, 1987) and MALIGNED (Clark, 1992), seem not to be actively supported or undergoing further development. Programs that are maintained include DCSE RNA alignment editor (De Rijk and De Wachter, 1993), which is now a component of RnaViz (De Rijk *et al*., 2003), and CINEMA (Parry-Smith *et al*., 1998), which is now distributed as part of Utopia (Pettifer *et al*., 2004). SeaView (Galtier *et al*., 1996) is a specialized cross-platform alignment editor developed for molecular phylogeny studies. ClustalX (Thompson *et al*., 1997) provides a graphical interface to the clustal multiple alignment algorithm, but does not allow manual manipulation of the alignment. A more recent introduction is the PFAAT Java alignment editor (Johnson *et al*., 2003), which has novel residue-level annotation tools and uses Jmol for protein structure display. Like Jalview it also provides tree viewing options, and the PFAAT authors kindly acknowledge Jalview as a source of their inspiration.

In this article, we have described the new capabilities and features available in Jalview 2, which enable both the expert bioinformatician and novice alike to perform sequence analysis investigations of ever-increasing size and complexity. Exploiting distributed access to computation and data resources is integral to modern bioinformatics, and to our knowledge, JVD was the first program capable of retrieving and visualizing DAS annotation on MSA. The new JV2 architecture also provides a solid foundation to extend the program further, to provide specialized support for next generation sequencing, improved support for the rendering of quantitative and symbolic annotation, and exchange data with other molecular visualization and analysis applications.

ACKNOWLEDGEMENTS

The authors would like to thank everyone who has contributed to Jalview 2 and previous versions of Jalview, including James Cuff and Steve Searle. Andreas Prlic contributed the DAS client library 'dasobert' as part of BioJava. Benjamin Schuster-Bockler developed the first version of Jalview's Stockholm parser. We also gratefully acknowledge all JV2's users for their encouragement, bug reports, helpful comments and suggestions.

Funding: Biotechnology and Biological Sciences Research Council (grant number BBSB16542).

Conflict of Interest: none declared.

REFERENCES

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

Blackshields,G. *et al*. (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol.*, **6**, 321–339.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Clark,S.P. (1992) MALIGNED: a multiple sequence alignment editor. *Comput. Appl. Biosci.*, **8**, 535–538.

Cole,C. *et al*. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, **36**, W197–W201.

De Rijk,P. and De Wachter,R. (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. *Comput. Appl. Biosci.*, **9**, 735–740.

De Rijk,P. *et al*. (2003) RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics*, **19**, 299–300.

Dowell,R.D. *et al*. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

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

Etzold,T. *et al*. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Galtier,N. *et al*. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.

Ilyin,V.A. *et al*. (2003) ModView, visualization of multiple protein sequences and structures. *Bioinformatics*, **19**, 165–166.

Johnson,J.M. *et al*. (2003) Protein family annotation in a multiple alignment viewer. *Bioinformatics*, **19**, 544–545.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.

Overton,I.M. *et al*. (2008) TarO: a target optimisation system for structural biology. *Nucleic Acids Res.*, **36**, W190–W196.

Pagni,M. *et al*. (2007) MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Res.*, **35**, W433–W437.

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

Pettifer,S.R. *et al*. (2004) UTOPIA-user-friendly tools for operating informatics applications. *Comp. Funct. Genomics*, **5**, 56–60.

Pillai,S. *et al*. (2005) SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.*, **33**, W25–W28.

Prlic,A. *et al*. (2007) Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, **8**, 333.

Raghava,G.P. *et al*. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.

Stockwell,P.A. and Petersen,G.B. (1987) HOMED: a homologous sequence editor. *Comput. Appl. Biosci.*, **3**, 37–43.

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*. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*, **7**, 318.