BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 4 2003, pages 544–545
DOI: 10.1093/bioinformatics/btg021

Protein family annotation in a multiple alignment viewer

Jason M. Johnson [1,*,†], Keith Mason [2], Ciamac Moallemi [2], Hualin Xi [1], Shyamal Somaroo [1] and Enoch S. Huang [1,*]

[1]Pfizer Discovery Technology Center, 620 Memorial Drive, Cambridge, MA 02139, USA and [2]Neogenesis Drug Discovery, 840 Memorial Drive, Cambridge, MA 02139, USA

Received on July 11, 2002; revised on October 14, 2002; accepted on October 25, 2002

ABSTRACT

Summary: The Pfaat protein family alignment annotation tool is a Java-based multiple sequence alignment editor and viewer designed for protein family analysis. The application merges display features such as dendrograms, secondary and tertiary protein structure with SRS retrieval, subgroup comparison, and extensive user-annotation capabilities.

Availability: The program and source code are freely available from the authors under the GNU General Public License at http://www.pfizerdtc.com.

Contact: jason_johnson@merck.com; enoch_huang@cambridge.pfizer.com

INTRODUCTION

Studying sequence–structure–function relationships within a protein family requires integrating the results of many different bioinformatics analyses in the context of a multiple sequence alignment. To position each residue accurately in an alignment and discern how it relates to structure and function often requires a substantial amount of thought and manual intervention in addition to careful use of automated methods. Several excellent sequence alignment editors and viewers are currently available to assist in this process, including Jalview (Clamp, 1998), Seaview (Galtier *et al.*, 1996), Cinema (Parry-Smith *et al.*, 1998), GeneDoc (Nicholas *et al.*, 1997), FarOut (Madsen, 2000), Belvu (Sonnhammer, 1999), and tools more oriented to protein structure and comparative modeling, such as Cn3D (Wang *et al.*, 2000), SWISS-PDB-VIEWER (Guex *et al.*, 1999), DINAMO (Bentz *et al.*, 1999), and STRAP (Gille and Frommel, 2001), although none combined the analysis tools and user-annotation features we needed for in-depth study and citation within the alignment itself. The protein family alignment annotation tool (Pfaat) is a Java-based, multi-platform sequence alignment editor with many features inspired by Jalview and Belvu, such as connectivity with Sequence Retrieval Service (SRS; Etzold and Argos, 1993) and interactive phylograms, but with additional analysis, display, and annotation features as described below.

*To whom correspondence should be addressed.
† Present address: Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, WA 98034, USA.

FEATURES

Main window: annotations, groups, alignment, mutual information, editing, and viewing. Pfaat provides for six different types of user annotation (residue, line, sequence, column, group, and color) which are stored with the alignment in a modified version of Stockholm format. Residue annotations are pop-up enabled text notes attached to individual amino acid residues within the alignment, shown by a colored oval around the residue and above the alignment column. These annotations follow their associated residues when the alignment is modified. Line annotations (which may be created and modified individually or in a master control window) allow a row of symbols below a sequence to track secondary structure, solvent accessibility, or other properties across a sequence. A user may also load the SwissProt annotations for a protein via the sequence retrieval system (SRS) and store them as line annotations. Sequence annotations allow notes to be saved for each row. Column annotations do not track with individual residues or sequences and are useful for annotation of protein family properties such as the positions of transmembrane helices for display or printing. Predefined and user-defined color schemes can be used to indicate sequence properties such as secondary structure propensity or show alignment conservation by percentage identity or similarity.

Pfaat also provides for grouping of sequences, group editing and for simultaneous display of consensus sequences and conservation plots for different groups. Also useful is a group analysis feature, which identifies columns that are similar within sequence subsets but dissimilar across the subsets. This feature uses absolute distance to measure similarity between the residue distributions of two groups. Specifically, for a given alignment column, if $p_1(i)$ is the fraction of sequence group 1 with

amino acid i and $p_2(i)$ is the fraction of sequence group 2 with amino acid i, then the distance is $\sum |p_1(i) - p_2(i)|$ over all amino acids.

Several automatic alignment options are interoperable with Pfaat, including ClustalW (Thompson *et al.*, 1994) and the hmmbuild and hmmalign functions of HMMer (Eddy, 2000), which may be used to add a new sequence into an existing alignment. Editing features include right or left justification of a sequence block to clean up gaps, deleting or adding residues at the cursor position, deleting columns, extracting columns of interest, and removing redundant sequences. Pfaat recognizes MSF, Stockholm, FASTA, and tab-delimited formats, and can write the alignment and display components to color postscript for printing to plotters and printers. Another feature unique to Pfaat is calculation of mutual information between a protein classification scheme and each column of residues. For instance, for a protein family in which more than one biochemical function is present, such as the lactate and malate dehydrogenases, these functional labels may be uploaded and the mutual information can be plotted below the alignment to identify residue columns that have a similar classification pattern. Users may also use Pfaat to search for a sequence pattern or regular expression, or to generate a regular expression from a selected set of alignment columns and rows.

Dendrogram, multidimensional analysis, and 3D-structure windows

Pfaat uses the ATV program window from Forester 1.92 (http://www.genetics.wustl.edu/eddy/forester/; Zmasek and Eddy, 2001) to display and manipulate similarity dendrograms or phylogenetic trees. This tool allows users to sort, print, or zoom in and out of large dendrograms, to import New Hampshire format trees (http://evolution.genetics.washington.edu/phylip.html), and is connected to the main window, such that sequences selected in the dendrogram are highlighted in the main window and vice versa. Columns can be weighted by their degree of sequence conservation (computed as the Shannon entropy), and removed from consideration if the percentage of gaps is above a user-defined threshold. Optionally, the integrity of tree nodes can be tested using the bootstrap procedure. Finally, a 3D-structure tool allows loading of a PDB sequence into the alignment; viewing, selecting residues, and manipulating the structure in a separate window; and extraction of PDB secondary structure data stored in the header as line annotation in the alignment window.

IMPLEMENTATION

A Java VM and Java3D are required to run Pfaat. For all platforms except SGI Irix, Java 2 Standard Edition v 1.3 and Java 3D v 1.1.3 (http://www.javasoft.com) must be installed, as distributed by Sun Microsystems. SGI Irix users should use Java 2 Standard Edition v 1.2.2 and Java 3D v 1.1.3 (http://www.sgi.com/developers/devtools/languages/java.html) as distributed by SGI. Pfaat is supported for Linux and WindowsNT/2000, and has been successfully installed on MacOS X and SGI-IRIX.

A version of Pfaat can also be launched from a Java applet, which facilitates the sharing of alignments on the web. The applet only provides the viewing and analysis functionality of Pfaat while disabling editing capabilities.

ACKNOWLEDGEMENTS

We would like to thank Qing Cao, Daniel Caffrey, Giles Day, Jeff Carlson, Patricia Soulard, Paul Dana, Colin Groom, Andrew Hopkins, and James Mills for suggestions of features and improvements or other project assistance. This work was funded in its entirety by Pfizer Inc.

REFERENCES

Bentz,J., Baucom,A., Hansen,M. and Gregoret,L. (1999) Dinamo: interactive protein alignment and model building. *Bioinformatics*, **15**, 309–316.

Clamp,M. (1998) *Jalview—a Java Multiple Alignment Editor*. European Bioinformatics Institute, Cambridge, (http://www.ebi.ac.uk/∼michele/jalview/).

Eddy,S.R. (2000) *HMMER: Profile Hidden Markov Models for Biological Sequence Analysis*. Washington University School of Medicine, St. Louis, MO, (http://hmmer.wustl.edu/).

Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.

Galtier,N., Gouy,M. and Gautier,C. (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.

Gille,C. and Frommel,C. (2001) STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics*, **17**, 377–378.

Guex,N., Diemand,A. and Peitsch,M.C. (1999) Protein modelling for all. *Trends Biochem. Sci.*, **24**, 364–367.

Madsen,D. (2000) *FarOut*. Uppsala University, Uppsala, Sweden, (http://xray.bmc.uu.se/dennis/manual/).

Nicholas,K.B., Nicholas,Jr,H.B. and Deerfield,II,D.W. (1997) GeneDoc: Analysis and Visualization of Genetic Variation (http://www.psc.edu/biomed/genedoc)..

Parry-Smith,D.J., Payne,A.W., Michie,A.D. and Attwood,T.K. (1998) Cinema—a novel colour interactive editor for multiple alignments. *Gene*, **221**, C57–C63.

Sonnhammer,E. (1999) *Belvu*. Karolinska Institutet, Stockholm, Sweden, (http://www.cgr.ki.se/cgr/groups/sonnhammer/Belvu.html).

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: 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.

Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.