

---

# Software review

## Protein family alignment annotation

### Abstract

For bioscientists studying protein structure and function, the Protein Family Alignment Annotation Tool (Pfaat) is a useful and simple program for annotating collections of proteins. This open-source software includes methods for viewing and aligning protein families, and for annotating sequence structure and residues with known functions. It offers new options to aid the study of proteins, and an extensible annotation tool for bioinformatics developers.

**Keywords:** protein families, alignment, annotation, Java, open source software, sequence analysis

### INTRODUCTION

Properties of particular proteins and families of proteins, such as conservation of sequence and structure, and the role of particular residues in function and structure, are of interest to a range of biologists, biochemists and medical researchers. A molecular biologist can experimentally modify the residues of a protein to determine their functions. What is the consequence of mutating an amino acid, or of blocking a residue chemically? How does a residue interact with its substrate? Experimental answers to these questions need to be correlated with a large body of related work to fully understand protein structure and function, or to design new pharmaceuticals.

A protein family may have thousands of members known from a wide range of organisms. With proper access to aligned families of proteins and their experimental literature, a scientist can use research on equivalent residues in homologous proteins to answer such questions. Research covering decades is available on roles of particular residues in function and structure of some protein families. The Pfam<sup>1</sup> database includes over 5,000 protein families (February 2003), while GenomeWeb<sup>2</sup> and BioNetBook<sup>3</sup> collections list scores of web sites for protein family databases.

Software tools are needed to make such

information more accessible, visually in alignments and protein structure displays, and annotation tools to allow a scientist to relate research literature to alignments and residues. *Pfaat*<sup>4</sup> (Protein Family Alignment Annotation Tool) is such a tool designed to facilitate annotation and analysis of large protein families. It runs on most computers with Java 1.3 or later software and is freely available from Sourceforge.net under a GNU licence for open-source software.

Features of Pfaat include the ability to visually align collections of sequences, grouping proteins into families using similarity and phylogenetic evidence, analyse them based on similarity criteria, and annotate sequences and residue positions with descriptive text. Pfaat was designed and written by members of Neogenesis Drug Discovery and Pfizer Discovery Technology Center (DTC) groups. It builds on related sequence alignment software, notably Jalview,<sup>5</sup> and incorporates additional open-source software tools for several of its functions.

### Details of operation

Documentation provided with Pfaat is brief but useful, outlining all parts of the program, and its operations. These include standard application functions for opening and saving data files, printing and help. The primary view and focus of this

program is a display of aligned protein sequences, as shown in Figure 1, similar to other multiple sequence alignment software. Sequence alignment functions include ability to select blocks of sequences, columns of residues, and to shift residues in a sequence to align with others. An accessory editing window allows one to fully edit a sequence. Methods to group sequences into families or subfamilies include options to name and arrange groups.

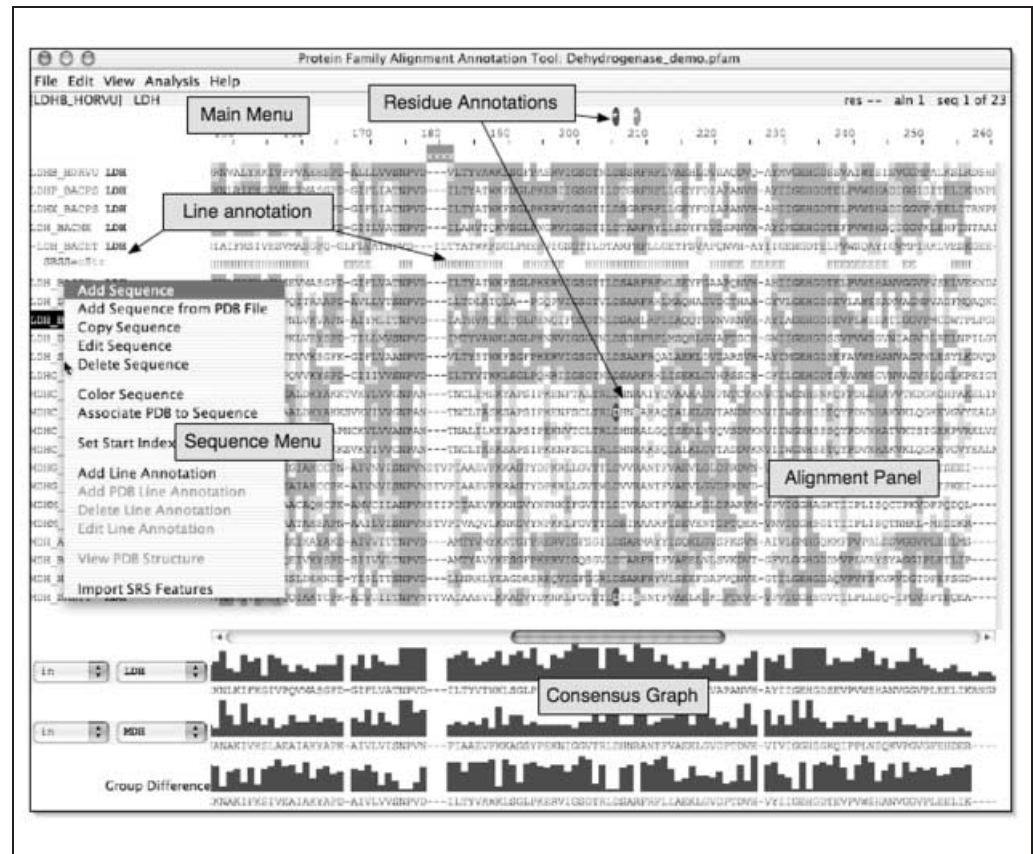
The Pfaat distribution website employs the easy to use InstallAnywhere program for installing Java programs on different computer systems: MS Windows, Mac OS X, Linux and other Unix systems. Normal operation is as an application running on the scientist's workstation, but with proper configuration, it may be run through a web browser as an applet. This can be useful for an organisation that provides web-based bioinformatics services; however, applets in general have

various limitations, and Pfaat applet is limited to viewing without editing functions.

I tested this application on a current Mac OS X computer. The authors support MS WindowsNT/2000 and Linux systems, but state it works for Mac OS X. I was able to run it on Mac OS X with caveats and extra effort. The installation creates an application that is easy to use, but lacks information when the program failed to work properly. Using a Java command line invocation identified the problem as a missing Java3D vecmath library. The installation package lacked this, but the source distribution<sup>6</sup> contains this. Adding that Java library resulted in proper basic operation. There is no Mac OS X release for Java3D at this time, preventing use of 3D display of protein structure.

The default alignment annotation file format is Stockholm,<sup>1,7</sup> usable with PFAM, HMMer and other software. Pfaat

**Figure 1:** Primary display of Pfaat, showing the main panel of 23 aligned dehydrogenase proteins, with LDH and MDH groups. The bottom panel shows three consensus graphs – one for LDH, one for MDH and a third graph of the difference in these groups. Residue annotations are highlighted above the alignment and for two sequences in the alignment. For a selected sequence, a pop-up menu of possible operations is shown, and above it the annotation line for another sequence



can read/write other formats (FastA, Clustal, MSF) but will not save annotation information in those. One bug found in this review: the program failed to read Stockholm format alignments downloaded from a Pfam server, owing to line wrapping of the aligned sequences. Using the source code provided with the program, I was able to locate and correct this problem in one hour, and rebuild the program. As the authors use Sourceforge, which provides methods for interested software developers to contribute to the maintenance of programs, I was able to post this patch to the Pfaat project website for others to use.

A consensus of the aligned sequences can be calculated using several criteria. The consensus display includes a bar graph of similarity for residue columns. Where your alignment includes multiple groups, you can have the consensus graph show differences between groups (see Figure 1). Several colour schemes can easily be applied to the alignment display, based on alignment similarity and amino acid function groupings. These colourings are an aid to building alignments and identifying residues of interest. Fonts and other display preferences can be set for this program.

If your protein family data includes Protein Databank (PDB) entries, you can use the built in function to associate this structure to a sequence alignment. It is then possible to use the PDB secondary structure and view its 3D structure as an aid to annotating your protein.

Highlighted residues in the alignment and 3D structure are linked in the display.

The Neighbor Joining Tree operation provides a phylogenetic tree view that is very helpful for viewing, selecting and grouping entries based on similarity relations. It is provided through incorporation of the open-source ATV<sup>8</sup> phylogenetic tree display tool. This tree view includes several options to select, reorder and re-root the tree, move branches using mouse operations, along with options to save as a file, print the tree, and use it to order the alignment

groups. One can also import a phylogenetic data file for grouping the protein family.

Searching and extracting sequences using patterns, or regular expressions, are supported with Perl regular expression syntax. Though such are now widely used in bioinformatics software,<sup>9</sup> additional documentation with examples of pattern matching and extraction would help this software. Comparison matrix and pairwise alignment analyses of the sequence selection are available, produced as a separate text window. Though no menu options are included in these, one can use standard software control keys (Control-C) to copy then paste into a text editor for saving.

Web access to SRS (Sequence Retrieval System) servers is included as a handy way for adding and updating protein sequences. This option is limited to retrieval by ID of SWISSPROT and TrEMBL data, though SRS web pages allow retrieval on a number of other criteria. A useful extension would be retrieval from PFAM databases, however one can use a web browser for this, then import from the saved file.

Plug-in or accessory programs can be added to align sequences including bioinformatics standards ClustalW and HMMer. These plug-in functions require computer-specific compiled programs, and proper configuration with Pfaat. These accessories are useful, and are examples of how a bioinformatician can use external analyses with this program (other alignment tools, phylogenetic or protein structure/function analyses). I was unable get Pfaat to use ClustalW properly on my computer, but reading the source code suggests this is easily correctable for a programmer.

## DISCUSSION

Pfaat provides good value as a basic, residue-level annotation tool for aligned protein families, including its use of the standard Pfam file format, and use of similarity tree views. The source code is available to others and is well written,

offering ways to plug into other database, analysis and data access methods. Pfaat retains much of the flavour and several functions of one of its influences, Jalview, but extends and improves upon Jalview in a number of ways for better ease of use and functionality, to make Pfaat a logical successor.

It is very encouraging for bioinformatics and bioscience communities to see commercially developed software such as Pfaat provided in an open-source way for public use and further development. The Pfizer Discovery Technology Center and authors are congratulated for this, especially as this tool provides a good basis for others to build advanced protein family annotation systems and use with project or company databases. The ease with which this reviewer fixed a bug in reading Pfam data is one of the compelling reasons that open-source software benefits all in the bioinformatics community.

Areas where additions and improvements to Pfaat software could make it more useful have been noted above. The program could use more user-interface methods and features to offer easier hand-alignment and annotation. Sequence editing ability is limited, though in practice this less important than its annotation and grouping methods. The print function, which provides only postscript output, could be improved using standard Java print methods.

How large a set of protein sequences is this program capable of handling? I tested this using Pfam alignment sets from <http://pfam.wustl.edu/>. For the COX1 family of 10,359 aligned sequences (PF00115; average length of 224 aa), the program ran out of memory with an allocation of 200 MB of memory. Likewise for 1,000 sequences, and 350, it showed memory limitations. The program proved able at working with 120 aligned sequences. There are programming efficiencies that would improve Pfaat's handling of large data sets, such as eliminating use of separate Java

objects for each amino acid (>2 million in the full COX1 family).

If you use Pfam or other protein family data in your work, Pfaat will be a useful addition to your software toolkit. Despite limitations and suggestions for improvements, this tool is quite usable in this initial release. Many of the problems this reviewer ran into are similar to those found in early releases of any free software, ones that attention of developers can quickly correct. Its use of a bioinformatics standard file format will ensure anyone using it can reuse data in other tools. It offers a new, needed entry to annotating and viewing information on aligned protein families. It also is promising as a basic framework that other bioinformaticians and software engineers can extend for particular needs of a protein family research project. Relational database access is one such possible addition, to provide for project or company database access to annotations.

#### Acknowledgments

This work was supported in part by NSF grant 0090782 and NIH grant 1R01HG002733-01 to D. Gilbert.

*Don Gilbert,  
Biology Department and Center for  
Genomics and Bioinformatics,  
Indiana University,  
Bloomington, Indiana 47405, USA  
Tel: +1 812 855 0587  
E-mail: gilbertd@bio.indiana.edu*

#### References

1. Bateman, A., Birney, E., Durbin, R. *et al.* (2000), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 28, pp. 263-266.
2. GenomeWeb (URL: <http://www.hgmp.mrc.ac.uk/GenomeWeb/>).
3. BioNetBook (URL: <http://www.pasteur.fr/recherche/BNB/bnb-en.html>).
4. Johnson, J. M., Mason, K., Moallemi, C. *et al.* (2003), 'Protein family annotation in a multiple alignment viewer', *Bioinformatics*, Vol. 19(4), pp. 544-545 (URL: <http://pfaat.sourceforge.net/>).
5. Clamp, M. (1998), 'Jalview, a Java multiple alignment editor', European Bioinformatics Inst. (URL: <http://www.ebi.ac.uk/~michele/jalview/>).

6. URL: <http://sourceforge.net/projects/pfaat/>
7. Sonnhammer, E., 'Stockholm multiple alignment markup format' (URL: <http://www.cgb.ki.se/cgr/groups/sonnhammer/Stockholm.html>).
8. Zmasek, C. M. and Eddy, S. R. (2001), 'ATV: Display and manipulation of annotated phylogenetic trees', *Bioinformatics*, Vol. 17, pp. 383–384 (URL: <http://www.genetics.wustl.edu/eddy/atv/>).
9. Tisdall, J. (2001), 'Beginning Perl for Bioinformatics', O'Reilly & Assoc., Sebastopol, CA.