BMC Bioinformatics

PFAAT version 2.0: A tool for editing, annotating, and analyzing multiple sequence alignments

Daniel R Caffrey*[1], Paul H Dana[1], Vidhya Mathur[1], Marco Ocano[1], Eun-Jong Hong[2], Yaoyu E Wang[3], Shyamal Somaroo[1], Brian E Caffrey[4], Shobha Potluri[1] and Enoch S Huang[1,3]

Address: [1]Pfizer Global Research and Development, 620 Memorial Drive, Cambridge, MA 02139, USA, [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, MA 02139, USA, [3]Bioinformatics Program and Biomedical Engineering Department, Boston University, MA 02215, USA and [4]School of Theoretical Physics, Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland

Email: Daniel R Caffrey* - daniel.caffrey@gmail.com; Paul H Dana - paul.dana@pfizer.com; Vidhya Mathur - vidhya_mathur@yahoo.com; Marco Ocano - myocana@yahoo.com; Eun-Jong Hong - eunjong@csail.mit.edu; Yaoyu E Wang - yew@bu.edu; Shyamal Somaroo - ssomaroo@yahoo.com; Brian E Caffrey - caffreb@tcd.ie; Shobha Potluri - shobha.potluri@pfizer.com; Enoch S Huang - enoch.huang@pfizer.com

* Corresponding author

Abstract

Background: By virtue of their shared ancestry, homologous sequences are similar in their structure and function. Consequently, multiple sequence alignments are routinely used to identify trends that relate to function. This type of analysis is particularly productive when it is combined with structural and phylogenetic analysis.

Results: Here we describe the release of PFAAT version 2.0, a tool for editing, analyzing, and annotating multiple sequence alignments. Support for multiple annotations is a key component of this release as it provides a framework for most of the new functionalities. The sequence annotations are accessible from the alignment and tree, where they are typically used to label sequences or hyperlink them to related databases. Sequence annotations can be created manually or extracted automatically from UniProt entries. Once a multiple sequence alignment is populated with sequence annotations, sequences can be easily selected and sorted through a sophisticated search dialog. The selected sequences can be further analyzed using statistical methods that explicitly model relationships between the sequence annotations and residue properties. Residue annotations are accessible from the alignment viewer and are typically used to designate binding sites or properties for a particular residue.

Residue annotations are also searchable, and allow one to quickly select alignment columns for further sequence analysis, e.g. computing percent identities. Other features include: novel algorithms to compute sequence conservation, mapping conservation scores to a 3D structure in Jmol, displaying secondary structure elements, and sorting sequences by residue composition.

Conclusion: PFAAT provides a framework whereby end-users can specify knowledge for a protein family in the form of annotation. The annotations can be combined with sophisticated analysis to test hypothesis that relate to sequence, structure and function.

Background

Building a multiple sequence alignment (MSA) is a critical step towards understanding the function and evolution of a protein family. Subsequent analysis typically includes phylogenetics, homology modeling, structure prediction, and binding site prediction. There are several excellent software packages that align multiple sequences [1]. Alignment accuracy is usually dependent on the percent amino acid identity between sequences [2] and manual editing is often a necessary step. Alignment editing tools are available in PFAAT as well as several other applications [3-10]. Additionally, MSA viewers provide various tools for sequence and structural analysis [3,5-7,11-16]. More recently, it has been recognized that MSAs can be used to validate and propagate annotations to other sequences [17]. PFAAT specializes in the annotation and analysis of a MSA, and since the release of version 1.0 [18], we have continued to develop and add novel features to PFAAT. We describe some of the main features below.

Implementation

PFAAT is written in Java and runs on several operating systems (Linux, Mac OS X, Solaris, and Windows). Users ini-

tially download and install the program from the home page using Java Web Start technology. Updated versions of the application are automatically downloaded on subsequent launches if the user is connected to the internet; otherwise the cached executable is used. Although PFAAT was not explicitly implemented for viewing nucleotide alignments, many of the generic features can also be applied to nucleotide sequences.

Results and Discussion

The alignment viewer is shown in Figure 1. The alignment can be edited in a number of ways. A gap can be inserted with the space bar or by dragging residues to the right while holding the *SHIFT* key down. A gap can be deleted with the *BACKSPACE* button or by dragging residues to the left while holding the *SHIFT* key down. Selecting multiple sequence names with the *CTRL* or *SHIFT* key allows one to edit a collection of sequences simultaneously. One can easily delete alignment columns that are composed entirely of gaps (*Edit -> Delete -> All Gap Columns*).

Double clicking on any of the three Name Panels will display the sequence annotations dialog box. Double click-
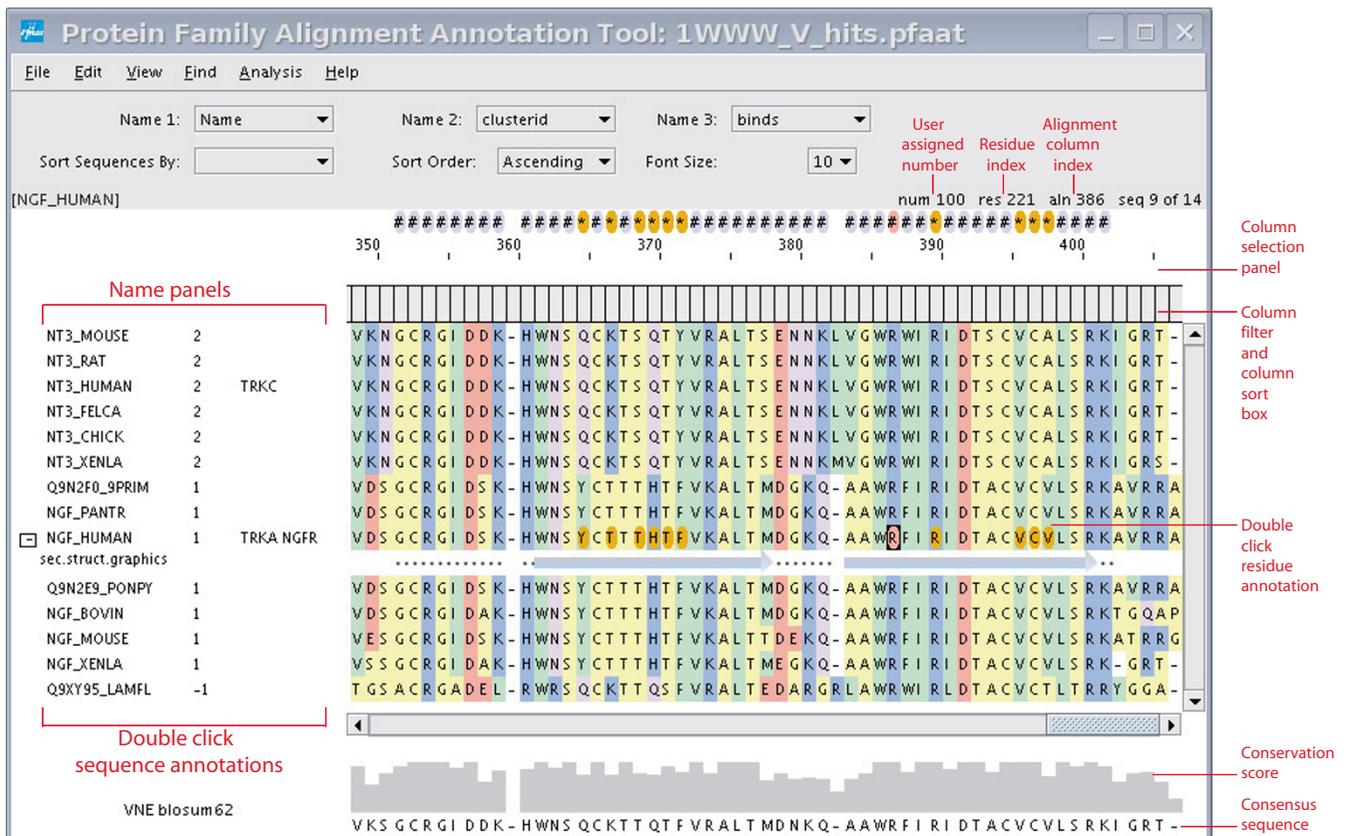
Figure 1
A screen shot of the alignment viewer. The red text highlights annotations along with various other features. The top row of drop-down menus was used to set the displayed sequence annotations in the name panels.

ing on a residue will display the residue annotation dialog box. The Tree viewer and structural viewer can be launched from the Analysis and File menus respectively. The tool bar (Figure 1) contains several drop-down menus that change the alignment view. The top row of drop-down menus changes the displayed annotation in Name Panels 1–3. The bottom row of drop-down menus sort sequences by annotation value and changes the font size.
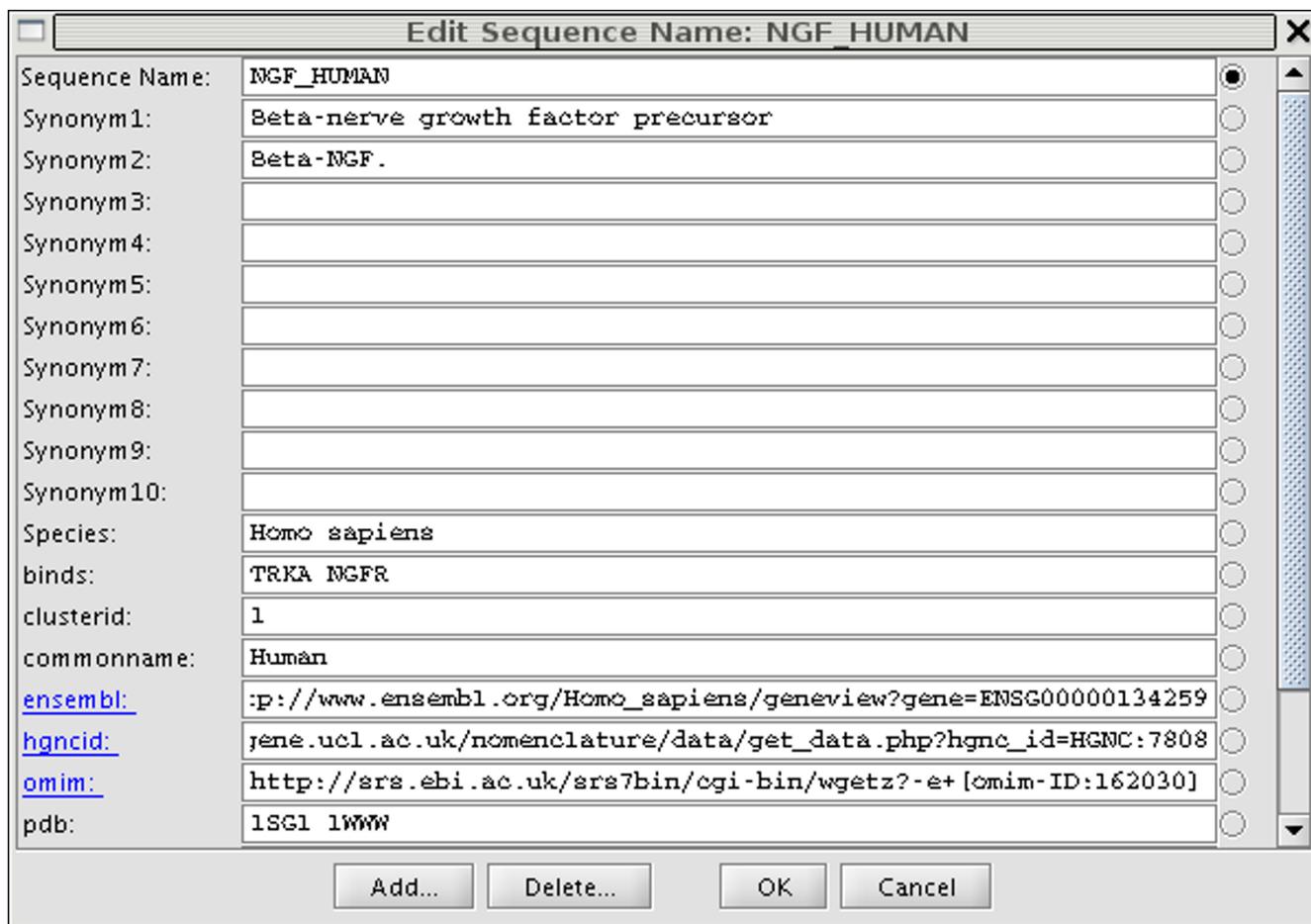
Sequence Annotations

A sequence annotation provides a convenient way to assign a name and a value to one or more sequences in the alignment. Figure 2 shows all sequence annotations for a sequence after double clicking on Name Panel 1. The currently displayed annotation is indicated by the radio button and the text field allows editing of the annotation values. Sequence annotations include but are not limited to synonyms, species, and cluster IDs. The easiest way to create a new annotation name and value is to click on the *Add* button. We recommend adding a UniProt entry name, as it can be used later to automatically extract sequence annotations from UniProt [19]. PFAAT mines several UniProt fields which include synonyms, species names, PDB codes [20], and hyper-linked ENSEMBL [21] IDs. The user documentation on the PFAAT home page describes other ways to create sequence annotations easily.

When working with a large number of sequences, sequence annotations facilitate rapid sorting and triaging of sequences. For example, the *Find* menu allows one to find and select sequences that match one or more search terms (e.g. species *equals* Homo sapiens *AND* Pdb *is not empty*). The selected sequences can then be moved to the top using *View -> Sort Sequences by -> selection*.

Residue Annotations

Residue annotations are a useful way to flag residues of interest. They are typically used to specify binding sites, SNPs and post-translational modifications. Figure 3

Figure 2
Sequence annotation dialog for NGF_HUMAN.
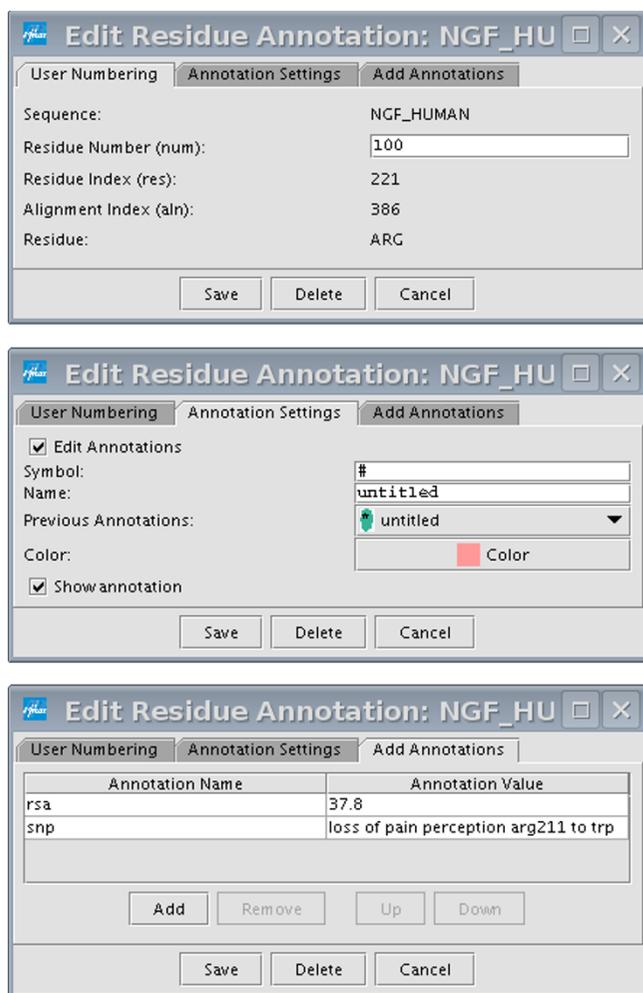
Figure 3
Residue annotation dialog for a residue in NGF_HUMAN.
There is a screen shot for each tab.

shows the residue annotation dialog box that appears after one double clicks on a residue. The first tab displays various numbering for the residue and allows users to assign their own numbering. This field is particularly useful when the user omits domains or segments of sequence that will offset the default numbering. The second tab allows the user to specify how a residue annotation will appear. Each residue annotation must have a color and a symbol that is displayed above the alignment. There is also a check box that allows the user to show or hide the residue annotation. The third tab is where all annotation names and values are specified. After clicking on the *Save* button, the annotation will appear as a residue with a colored oval. The annotations can be viewed by mousing over the residue or double clicking the residue again.

Residue annotations provide a gateway for several types of subsequent analysis. For example, the *Find* menu allows

one to quickly find and select residue annotations that match one or more search terms. The residue selection can be extended to the alignment column, and there is an option to hide all other columns. As a next step, one might apply one of the many features that can be applied to selected columns, including sorting by percent identity and most of the features in the *Analysis* menu.

Phylogenetic Analysis
PFAAT reconstructs phylogenetic trees using an implementation of the neighbor joining algorithm [22]. An option to perform bootstrap analysis is also provided. Trees can be reconstructed using selected sequences or selected columns. PFAAT recognizes various tree formats (nh, nhx, nexus) and can display tree files generated by other software.

The tree viewer (Figure 4) borrows much of its code from ATV [23]. We have since added a number of new features that are tightly linked to the alignment viewer. Sequence annotations can be accessed by double clicking on the terminal nodes. The tree viewer has the same advanced searching capabilities as the alignment viewer. Sequence selection is synchronized between the tree viewer and the alignment viewer, allowing the user to quickly triage sequences based on evolutionary relationships. Another popular feature is the assignment of the cluster IDs in the tree viewer. Cluster IDs can be assigned based on gene duplication events or by dragging a vertical bar from left to right so it cuts the tree into clusters of a desired average size. As the cluster ID is a sequence annotation, it can also be used to sort sequences in the alignment viewer or to compute conservation scores for particular clusters.

Structural Analysis
PFAAT uses Jmol [24] to display three-dimensional protein structures. A sequence needs to be associated with PDB file to ensure that the sequence numbering is consistent with the PDB numbering. The PDB sequence is automatically aligned to the PFAAT sequence and the user has the option of fixing any alignment errors. Once the alignment is approved, the PDB numbering is applied to the aligned residues. Secondary structure elements are drawn below the sequence and Jmol is launched. By default, residue selection is synchronized between the alignment viewer and Jmol. This is particularly useful when binding site residues need to be mapped from structure to sequence or vice versa. Once the structure is associated, one can compute residue solvent accessibilities (rsa) [25] which are stored as residue annotations. This feature is useful, if one wants to focus analysis on exposed residues. For example, conservation scores can be computed for a cluster of sequences and mapped to surface residues on the structure (Figure 5) in a similar fashion to the Evolutionary Trace method [26]. A standard color spectrum is
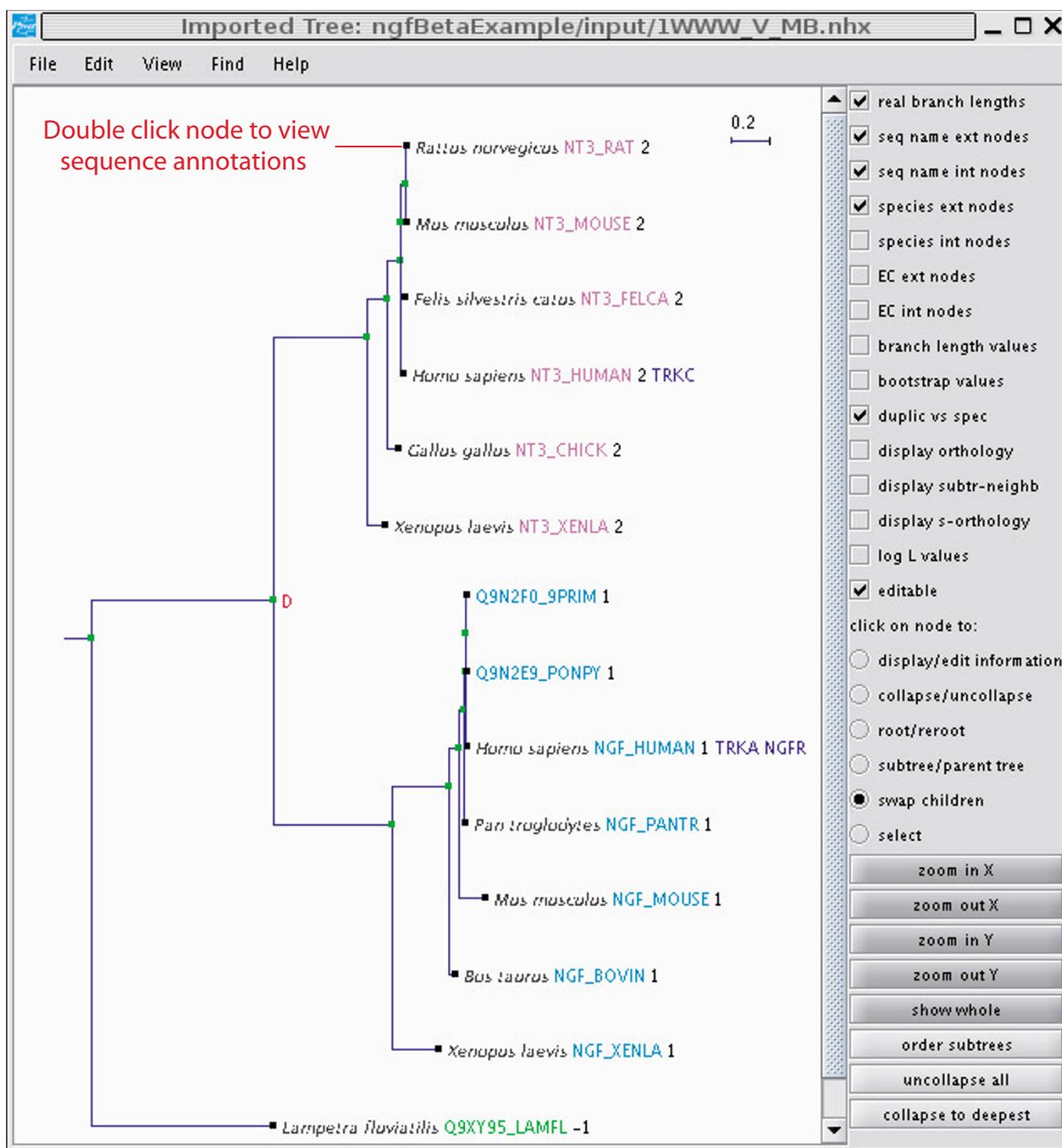
Figure 4
The tree viewer. Cluster IDs (black text) are assigned based on the gene duplication event and names are colored by cluster ID. The "binds" sequence annotation is displayed in blue text.
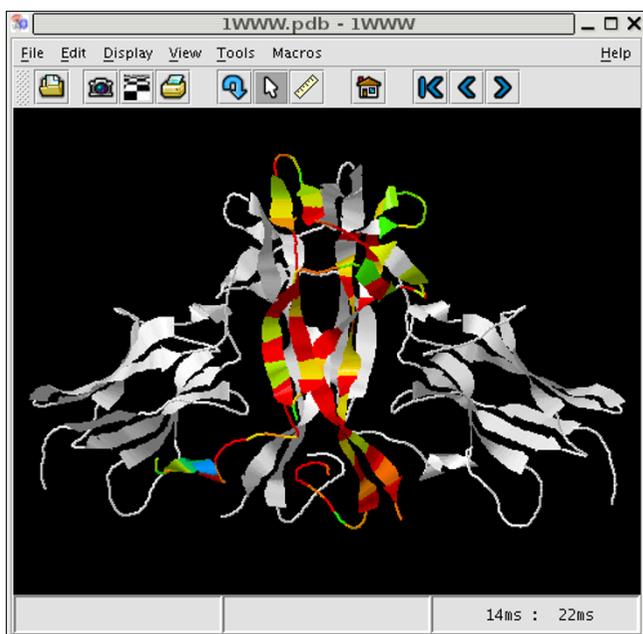
Figure 5
Conservation scores mapped to a structure.

applied to the structure, where red is for invariant columns and violet is for the most variable columns.

Sequence Analysis
There are number of sequence analysis tools that are primarily found under the *Analysis* menu. For example, amino acid percent identities can be computed between all sequences or a subset of their columns. There is also an identity count, which reports the number of sequences that have a residue that is identical to a particular sequence at each column. There are a variety of conservation scores, the default being a von Neumann Entropy based score (described below) that can be applied to selected sequences as well as selected columns. The Conservation scores can be mapped to a 3D structure as discussed above. The PLSR method allows one to identify sequence trends that best correlate with numerical experimental measurements (e.g. binding data that is stored as a sequence annotation). Immediately above each alignment column is a gray box. A single click on a box will show the number and type of residues that are found at a column. In sort mode, the user can select a residue type that will determine how the sequences are sorted. For example, one might be interested in moving all sequences that have a lysine or arginine at column 100 to the top. In filter mode, all sequences that do not have a lysine or arginine would be hidden. The sort mode is often used for mutagenesis experiments as it provides a nice summary of residues that are tolerated at a given position. The filter mode can be used when designing selective drugs for a

large gene family. Several other features are described in the documentation on the PFAAT home page.

Von Neumann Entropy
Although Shannon Entropy is a popular measure of residue conservation, it incorrectly treats amino acids as being orthogonal. Von Neumann Entropy overcomes this shortcoming and is the default measure of residue conservation in PFAAT. Shannon Entropy is described in equation 1, where i enumerates each mutually exclusive entity, $\lambda_i > = 0$ and $\Sigma \lambda_i = 1$. The λ_i are a measure of the probability of encountering the entity i in the collection.

$$\text{Entropy} = -\Sigma \lambda_i \log(\lambda_i) \quad (1)$$

As the 20 amino acids are non-orthogonal (overlapping) vectors, the set must be expressed in terms of an equivalent orthogonal basis set. The mutual overlap of the distinct amino acid vectors in each column is described by a matrix ρ encoding the pairwise similarities between these non-orthogonal vectors. We have found that the following simple 20×20 matrix, also called the density matrix, works well for amino acid conservation:

$$\rho = FS \quad (2)$$

where F is a diagonal matrix of amino acid 'counts' or frequencies and S is an appropriate amino acid similarity matrix (e.g BLOSUM 62).

Now ρ can be naturally expressed in terms of an orthogonal basis through diagonalization, i.e. by calculating its eigenvectors E and eigenvalues $\Lambda = \text{diag}(\lambda_i)$ [27]:

$$\rho = E \Lambda E^{-1} \quad (3)$$

The eigenvectors can be interpreted as 20 orthonormal amino acid properties spanning 'amino acid space'. If ρ is normalized such that Trace $(\rho) = 1$ (i.e. $\Sigma \lambda_i = 1$), the eigenvalues λi can be interpreted as the probabilities of encountering each of these 20 orthogonal eigenvector properties in the column. Inserting the eigenvalues λ_i into the formula (1) now gives the entropy measure we desire. The entropy measure can in fact be written directly in terms of ρ itself

$$\text{Von Neumann Entropy} = -\text{Trace}(\rho \log \rho) \quad (4)$$

as can be seen by inserting (3) into (4) to recover (1). Equation (1) is computationally more efficient than equation (4) and is implemented in PFAAT.

Conclusion
A MSA provides valuable information about a protein family. Additional knowledge is provided by the user in

the form of annotations. By combining these annotations with sophisticated analysis, PFAAT allows researchers to test hypothesis that relate to sequence, structure and function. This release of PFAAT marks a significant improvement in functionality over version 1.0. The major improvements are described in the *What's new?* section of the user documentation. We eagerly anticipate user feedback and a 'request features' link is provided on the project home page. Future areas of development might include the extraction of sequence annotations from additional databases (e.g. GO, KEGG, and PFAM) and employing mechanisms to propagate annotations to other sequences [17].

Availability and Requirements
Project name: PFAAT

Project home page: http://pfaat.sourceforge.net/

Documentation: http://pfaat.wiki.sourceforge.net/

Operating Systems: Platform independent

Programming language: Java 1.4.2 or higher

License: GNU General Public License

Any restrictions to use by non-academics: None

Abbreviations
PFAAT – Protein Family Alignment Annotation Tool

MSA – Multiple Sequence Alignment

PLSR – Partial Least Squares Regression

Competing interests
The author(s) declares that there are no competing interests.

Authors' contributions
DRC wrote Java code, implemented algorithms, specified and prioritized features, wrote user documentation, and contributed to the writing of the manuscript. VM wrote Java code and re-architected the package. PHD wrote Java code and re-architected the package. MO wrote Java code. SS specified algorithms for implementation and contributed to the writing of the manuscript. YW wrote Java code. EJH wrote Java code and implemented algorithms. SP wrote Java code and implemented algorithms. BEC wrote Java code and implemented algorithms. ESH contributed to the writing of the manuscript. All authors read and approved the final manuscript.

References
1. Edgar RC, Batzoglou S: **Multiple sequence alignment.** *Curr Opin Struct Biol* 2006, **16(3):**368-373.
2. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs.** *Nucleic Acids Res* 1999, **27(13):**2682-2690.
3. **BioEdit**   [http://www.mbio.ncsu.edu/BioEdit/bioedit.html]
4. **MPSA**   [http://mpsa.ibcp.fr/]
5. Bentz J, Baucom A, Hansen M, Gregoret L: **DINAMO: interactive protein alignment and model building.** *Bioinformatics* 1999, **15(4):**309-316.
6. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20(3):**426-427.
7. Deleage G, Combet C, Blanchet C, Geourjon C: **ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities.** *Comput Biol Med* 2001, **31(4):**259-267.
8. Galtier N, Gouy M, Gautier C: **SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny.** *Comput Appl Biosci* 1996, **12(6):**543-548.
9. Nicholas KB: **GeneDoc: Analysis and Visualization of Genetic Variation.** *EMBNET NEWS* 1997, **4:**14.
10. Parry-Smith DJ, Payne AW, Michie AD, Attwood TK: **CINEMA–a novel colour INteractive editor for multiple alignments.** *Gene* 1998, **221(1):**GC57-63.
11. **BELVU**   [http://www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html]
12. Barton GJ: **ALSCRIPT: a tool to format multiple sequence alignments.** *Protein Eng* 1993, **6(1):**37-40.
13. Goodstadt L, Ponting CP: **CHROMA: consensus-based colouring of multiple alignments for publication.** *Bioinformatics* 2001, **17(9):**845-846.
14. Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, McMahan L, Sali A: **ModView, visualization of multiple protein sequences and structures.** *Bioinformatics* 2003, **19(1):**165-166.
15. Li W, Godzik A: **VISSA: a program to visualize structural features from structure sequence alignment.** *Bioinformatics* 2006, **22(7):**887-888.
16. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH: **Cn3D: sequence and structure views for Entrez.** *Trends Biochem Sci* 2000, **25(6):**300-302.
17. Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O: **MACSIMS: multiple alignment of complete sequences information management system.** *BMC Bioinformatics* 2006, **7:**318.
18. Johnson JM, Mason K, Moallemi C, Xi H, Somaroo S, Huang ES: **Protein family annotation in a multiple alignment viewer.** *Bioinformatics* 2003, **19(4):**544-545.
19. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic Acids Res* 2006:D187-191.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1):**235-242.
21. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al.*: **Ensembl 2007.** *Nucleic Acids Res* 2007:D610-617.
22. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4):**406-425.
23. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17(4):**383-384.
24. **Jmol**   [http://www.jmol.org/]
25. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility.** *J Mol Biol* 1971, **55(3):**379-400.
26. Morgan DH, Kristensen DM, Mittelman D, Lichtarge O: **an application for predicting and visualizing functional sites in protein structures.** *Bioinformatics* 2006, **22(16):**2049-2050.
27. Strang G: **Orthogonality.** In *Introduction to Linear Algebra* Wellesley-Cambridge Press, Wellesley, MA; 1993.