

Data and text mining

## ProteomeCommons.org JAF: reference information and tools for proteomics

J. A. Falkner, J. W. Falkner and P. C. Andrews\*

Department Biological Chemistry, University of Michigan, Ann Arbor, MI 48104, USA

Received on October 30, 2005; revised on December 5, 2005; accepted on December 15, 2005

Advance Access publication January 24, 2006

Associate Editor: Jonathan Wren

### ABSTRACT

**Summary:** Analysis of proteomics data, specifically mass spectrometry data, commonly relies on libraries of known information such as atomic masses, known stable isotopes, atomic compositions of amino acids, observed modifications of known amino acids and ion masses that directly correspond to known amino acid sequences. The Java Analysis Framework (JAF) for proteomics provides a freely usable, open-source library of Java code that abstracts all of the aforementioned data, enabling more rapid development of proteomics tools. The JAF also includes several user tools that can be run directly from a web browser.

**Availability:** The current version and an archive of all older versions of the Java Analysis Framework for Proteomics is freely available, including complete source-code, at <http://www.proteomecommons.org/current/511/>

**Contact:** [jfalkner@umich.edu](mailto:jfalkner@umich.edu)

### PROTEOMECOMMONS.ORG JAVA ANALYSIS FRAMEWORK

The ProteomeCommons.org Java Analysis Framework (JAF) provides a library of freely usable, open-source Java code supported by ProteomeCommons.org (Falkner and Andrews, 2005) that abstracts information regarding commonly used atoms, stable isotopes of atoms, residues and modifications to residues. The code initially started as a programmer's application programming interface (API) for accessing standardized, public information and speeding up development of tools that relied on calculations such as the masses of peptides and proteins, single nucleotide polymorphisms (SNPs) encoded in a protein sequence, theoretical isotope distributions of ions observed by mass spectrometry and references for atomic weights and residue compositions. The JAF currently provides both the aforementioned programmer's API and several user tools.

The complete JAF project includes documentation that references the various API components and user tools available with the project. The abbreviated list below highlights popular tools and includes links.

- Java API Components

- Configurable atomic mass library, including default atomic mass values, stable atomic isotopes and abundancies of atomic isotopes based on published NIST values (Coplen, 2001, Rosman and Taylor, 1998).

- Configurable amino acid residue library, including all of the common amino acids (Voet and Voet, 2004).
- Configurable library of known modifications to the standard amino acids, including many of the modifications observed in tandem mass spectrometry.
- Isotopic distribution tool that predicts the isotope distribution based on either a peptide sequence or a *m/z* value that is assumed to be comprised of amino acids.
- Memory efficient, rapid lookup library of amino acid combinations based on mass, including combinations of modified amino acids.
- SNP component to identify possible SNPs for a given protein or peptide sequence.

- User Tools

- HTML reference of atomic masses <http://www.proteomecommons.org/current/511/docs/atom-reference.html>
- HTML reference of amino acids and known modified amino acids <http://www.proteomecommons.org/current/511/docs/residue-reference.html>
- HTML reference of amino acid combinations, including combinations of modified amino acids and mass shifts associated with N-terminus or C-terminus residues. <http://www.proteomecommons.org/current/511/docs/residue-combination-reference.html>
- Online peptide *m/z* and pI calculator. A tool that will calculate the mass of a peptide, its approximate pI and possible proteolytic fragments, optionally accounting for potential amino acid modifications. <http://www.proteomecommons.org/current/511/PeptideCalculator.jnlp>
- Online residue combination calculator. Finds all combinations of amino acid residues that match a given *m/z* with a specified tolerance. <http://www.proteomecommons.org/current/511/ResidueCombinationCalculator.jnlp>

The JAF is actively developed and maintained as a project on ProteomeCommons.org. All of the API components and tools follow the ProteomeCommons.org guidelines for being easily accessible and well documented. Anyone may use the JAF code or join in development of this project or others on the ProteomeCommons.org website; however, knowledge of Java programming is required in order to edit the code associated with the JAF. ProteomeCommons.org maintains archives of all formal releases of the JAF code along with a public source-code repository so that developers may download the current source-code at any time.

\*To whom correspondence should be addressed.

## BACKGROUND AND COMPARISON AGAINST OTHER TOOLS

The JAF is designed so that it allows very fine control over atoms, amino acids, peptides and commonly observed, mass spectrometry-related modifications of amino acids and peptides. This design allows for the most accurate masses that can be determined for peptide and protein sequences, limited by the accuracy of published NIST standards and rounding errors inherent to how numbers are represented by computers. Other projects exist that provide similar functionality to parts of the JAF, particularly some of the user tools provided by the JAF, but the JAF is unique in that it provides a robust framework that enables other tools to preserve very high peptide and protein mass accuracy even when accounting for incorporation of stable isotopes or amino acid modifications.

The ProteomeCommons.org IO framework (<http://www.proteomecommons.org/current/531>) and the Peptide Finite State Machine (PFSM) (Falkner and Andrews, 2005) projects are excellent examples of related tools that leverage the JAF for high mass accuracy. The IO framework can read and write various proteomics-related file formats, including MGF, T2D, DTA, mzData, mzXML and FASTA. The PFSM project is a *de novo*-like tool that uses tandem mass spectra and known amino acids to rapidly identify protein sequences that may account for the given tandem mass spectra. In both projects, representations of amino acids, peptides and proteins are handled by the JAF, which enables the projects to provide very high mass accuracy (10 p.p.m. or less) even if the data include amino acids that have incorporated isotopes of atoms or that have been modified.

Compared with existing tools, particularly tools designed to work with relatively lower mass accuracy instruments such as an LCQ, frameworks like the JAF are commonly replaced by lists of the known amino acids. Modifications of amino acids are handled by allowing users to assign arbitrary mass shifts. X!Tandem (Craig and Beavis, 2004) and subsequently the GPM (Craig *et al.*, 2004) and Mascot (Perkins *et al.*, 1999) are examples of such tools. In general these tools are intended to compare mass spectra with protein datasets and identify peptides and proteins that best account for the observed spectra, a task that is very different from what the JAF does; however, the tools rely on being able to accurately identify peptide masses. The accuracy in these tools is limited by a simple list of masses, optionally user edited, that is correlated to known amino acids and known modifications of amino acids. Errors or rounding that occurs in this list will alter the mass accuracy and performance of the overall tool, and changes such as mass shifts owing to incorporation of stable isotopes are often non-trivial to account for. For these types of tools the JAF is intended as a replacement for the list of masses, ideally by having the tool use the JAF API or using the JAF to generate the appropriate list of masses that the tool requires.

Other tools such as the DBToolKit (Martens *et al.*, 2005) and BioJava ([biojava.org](http://biojava.org)) serve as frameworks that are intended to be used in development of more tools. Both DBToolKit and BioJava

happen to also be coded in the Java programming language; however, they are significantly different than the JAF framework. The DBToolKit is primarily focused on manipulation of protein and peptide datasets, functionality most similar to the ProteomeCommons.org IO framework mentioned previously. For most cases accurate masses that the JAF can provide are not needed by the DBToolKit as the code treats protein sequences as collections of characters. The BioJava framework is similar. It too is open-source and Java based; however, the framework is much more grand in scale, providing functionality applicable to most bioinformatics tasks. Much of BioJava treats protein and peptide sequences similar to DBToolKit, as a collection of symbols. The JAF differs from BioJava primarily in its focus on proteomics and mass spectrometry. The JAF is better suited for coding a mass spectrometry related applications and where very accurate masses are important.

Summarizing the JAF in comparison to other tools, the JAF is coded in the Java programming language, completely free to use, open-source and highly specific to proteomics and high mass accuracy mass spectrometry. The Java programming language is reasonably popular with proteomics tools, but many Perl, Python, C/C++ and other programming language libraries still exist. Being completely free to use and open-source is a feature of the JAF that makes it well-suited for general use, whereas other tools and frameworks often provide restricted access. And finally, the mass spectrometry specificity and high mass fidelity tools provided are the key features that differentiate it from other existing tools.

## ACKNOWLEDGEMENTS

This project is part of the National Resource for Proteomics and Pathways funded by NCRR grant P41-RR018627.

*Conflict of Interest:* none declared.

## REFERENCES

- Falkner, J.A. and Andrews, P.C. (2005) ProteomeCommons.org: code and data archive and dissemination for the proteomics community. *Am. Biotechnol. Lab.* (in Press).
- Falkner, J. and Andrews, P. (2005) Fast tandem mass spectra-based protein identification regardless of the number of spectra or potential modifications examined. *Bioinformatics*, **21**, 2177–2184.
- Coplen, T.B. (2001) Atomic Weights of the Elements 1999. In *Pure Appl. Chem.*, **73**, 667–683.
- Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
- Craig, R. *et al.* (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.*, **3**, 1234–1242.
- Martens, L. *et al.* (2005) DBToolKit: processing protein databases for peptide-centric proteomics. *Bioinformatics*, **21**, 3584–3585.
- Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
- Rosman, K.J.R. and Taylor, P.D.P. (1998) Isotopic compositions of the elements (1997). *Pure Appl. Chem.*, **70**, 217–235.
- Voet, D. and Voet, J. (2004) *Biochemistry*. 3rd edn. Wiley, Indianapolis, IN, ISBN: 047119350X.