TANDEM: matching proteins with tandem mass spectra

Robertson Craig[1] and Ronald C. Beavis[1,2,]*

[1]*Manitoba Centre for Proteomics, University of Manitoba, Winnipeg, MB, Canada R3T 2N2 and* [2]*Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637, USA*

ABSTRACT

Summary: Tandem mass spectra obtained from fragmenting peptide ions contain some peptide sequence specific information, but often there is not enough information to sequence the original peptide completely. Several proprietary software applications have been developed to attempt to match the spectra with a list of protein sequences that may contain the sequence of the peptide. The application TANDEM was written to provide the proteomics research community with a set of components that can be used to test new methods and algorithms for performing this type of sequence-to-data matching.

Availability: The source code and binaries for this software are available at http://www.proteome.ca/opensource.html, for Windows, Linux and Macintosh OSX. The source code is made available under the Artistic License, from the authors.

Contact: rbeavis@proteome.ca

A significant branch of research in peptide mass spectrometry in the last 30 years has focused on experimentally determining peptide sequences using tandem mass spectrometry (MS/MS). The fundamental idea is as follows: isolate a particular intact peptide parent ion with one mass spectrometer; add electronic and vibrational energy to the isolated ion; and observe the resulting fragment ions with another mass spectrometer (Aebersold and Mann, 2003). If the resulting fragment ion spectrum contains signals that can be correlated with bond-breaking reactions along the peptide backbone, then the sequence of the peptide should be calculable because of the nearly unique masses associated with each side chain (leucine and isoleucine are isobaric and cannot be distinguished by this type of experiment).

For a variety of experimental reasons, this type of MS/MS spectrum frequently does not produce enough interpretable ions to sequence a peptide completely. An alternative method for determining the sequence using the same information was developed that involved comparing the experimentally observed fragment ions against those that would be expected for every known peptide sequence that could be generated from the known proteome of the organism under investigation (Rappsilber and Mann, 2002). This approach has the advantage that it only requires enough information to rank peptides in a proteome. The necessity to be able to call even short stretches of contiguous sequence is removed, making it possible to use spectra that cannot be manually interpreted in the usual manner.

For any complete proteome, there are too many potential peptide sequences to consider performing this type of correlation manually. Several software implementations have been developed to carry out this process, but at the moment they are all proprietary and the source code is not available (e.g. Perkins *et al.*, 1999), which has limited the development of this type of technology. Our group has designed and implemented an open-source project that can be used for developing new algorithms to improve the results and efficiency of matching peptide sequences to MS/MS spectra. This project is called 'TANDEM'.

TANDEM was written to run from a command line, with an input XML file name as the only command line parameter. The code was created using a set of classes that perform the following tasks:

(1) read XML input parameter files;

(2) read protein sequences from FASTA files;

(3) read MS/MS spectra in common ASCII formats (DTA, PKL and Matrix Science);

(4) condition MS/MS spectra to remove noise and common artifacts;

(5) process peptide sequences with cleavage reagents, post-translational and chemical modifications;

(6) score peptide sequences; and

(7) create an XML output file capturing the best scoring sequences and some statistical distributions relevant to the scoring process.

The code for these objects was written in C++, using the Standard Template Library. The code was written so that it

---

*To whom correspondence should be addressed.

could cross-compile under Windows, Linux or OS X, with only minor differences, handled by preprocessor commands. The main difference between the platforms was the mechanism for starting worker threads, with specific calls made necessary because of the differences between the Windows and POSIX threading libraries.

The XML chosen for both input and output was BIOML (Fenyö, 1999), with mass spectra and other histograms represented using GAML (Duckworth, 2002; http://www.gaml.org/documentation.htm) as a namespace extension. The current implementation's API has 48 possible input parameters. To simplify the process of entering so many parameters, a two-step input parsing system was used. The input file specified on the command line can contain the name of a 'default file' that has values for all of the possible parameters. The input file parameters override the settings in the default file. By constructing a set of default files for common experimental situations, it is possible to create a very simple input file that only overrides the few parameters necessary for the experiment at hand. The sequence-containing FASTA files are specified in the input file by a 'taxon' name that is defined in a 'taxonomy' XML file. More than one FASTA file may be specified for any particular 'taxon' keyword. Examples for input parameter, default parameter and taxonomy files were included with the distribution release of the software. An example of how to use this software and a standard Common Gateway Interface to create an HTTP interface were also included.

TANDEM has been tested thoroughly, both by the authors and several other groups in academia and industry. It has also been used to generate new methods for carrying out common proteomics tasks (Craig and Beavis, 2003). The existence of this type of open-source project will hopefully give researchers a common platform for carrying out further exploration of the scientific issues currently outstanding in the application of large-scale proteomics to biological systems.

ACKNOWLEDGEMENTS

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Craig,R. and Beavis,R.C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.*, **17**, 2310–2316.

Duckworth,J. (2002) An XML Data Model for Analytical Instruments.

Fenyö,D. (1999) The Biopolymer Markup Language. *Bioinformatics*, **15**, 339–340.

Perkins,D.N., Pappin,D.J.C., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Rappsilber,J. and Mann,M. (2002) What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.*, **27**, 74–78.