A Software Shell for MS Data Conversion and Database Submission of MS Data

Yang Su[1,5], Sequin Huang[1,3,6], Hua Huang[1,3], David H. Perlman[1,3] Catherine E. Costello[1,2,4], Mark E. McComb[1,3]

1. Cardiovascular Proteomics Center, 2.Mass Spectrometry Resource, 3.Department of Medicine, 4.Department of Biochemistry, Boston University School of Medicine, 5.Bioinformatics Program, Boston University, Boston, MA, 6.Present address: Waters Corp, Milford, MA

Introduction: With a multitude of different MS instrumentation and data analysis software platforms available for MS and proteomics it becomes difficult to manipulate and manage various data sets. Recently we presented a software application that will allow the conversion of processed MS data files obtained on a variety of instruments into several common formats accepted by different software applications. We have further developed the program to add support for the mzXML[1] format and incorporate a front end interface which may be linked to several web based database searching engines including Mascot, ProteinProspector and BUPID[2] (a peptide mass fingerprinting program based on a log-likelihood ratio model developed here at BUSM).

Methods: The data processing software was developed using Microsoft Visual Basic 6.0. To add support for mzXML format, we used MSXML 4.0 as an XML parser and built a visual C++ library to decode Base64 encoded peak list data in the mzXML file. Other supported data formats are intermediate files converted from raw data files using manufacturers' software: LC MS/MS data is processed with Analyst QS (ABI/Sciex), MassLynx/PLGS2.1 (Waters); and MALDI MS data with MOverZ (Proteometrics LLC); and FTMS data with BUDA[3] (BUSM). The BUPID program was developed in C under Linux and made accessible to the main program through a CGI based web interface.

Results: The shell data conversion program was written to implement a user friendly GUI interface which may be operated in an unattended batch processing mode. Testing of the program was performed on existing MALDI-TOF MS, MALDI-FT MS and LC MS/MS data sets obtained in house. The program allowed the conversion of large volumes of data obtained on different instruments to the formats of several commercially and publicly available search engines. Files were then submitted for protein identification to the search engines with the search settings specified by the user. For a batch of files, the search setting only needs to be specified once, thus allowing unattended operation. Results files are automatically saved in HTML format and can then be viewed directly inside the program. Our recent implementation of the mzXML format introduced by the Institute for Systems Biology affords the benefits of a common data format for summation of results obtained on different MS platforms, comparative analysis of MS methodology and archiving of data such that it may be analyzed at a later date in-house or at a different facility.

Conclusions: The software provides an easy-to-use graphical interface for automatic MS data conversion and database searching. It can also be easily be expanded for more MS data types and linked to more database search engines.

Reference:

1. Pedrioli, P.G. et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004 Nov;22(11):1459-66
2. Tong, W. et al. BUPID: Probability-Based Protein Identification by Searching Sequence Databases Using Peptide Mass Fingerprint Data. ASMS 2005 WP Slot: 388
3. By O'Connor, P.B. http://www.bumc.bu.edu/FTMS
4. Kaur, P. et al. MASSPIKE (Mass Spectrum Interpretation and Kernel Extraction) for Biological Samples. ASMS 2005 TP Slot: 402

Acknowledgement:
This project was funded by NIH NHLBI contract N01-HV-28178 and NIH NCRR grants P41-RR10888 and S10-RR15942.

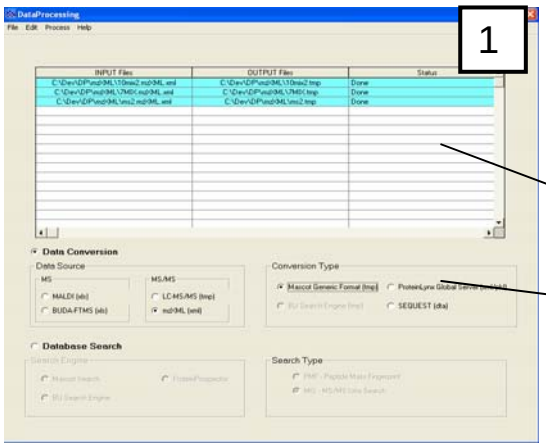

Fig 1. Front page of the software showing and example of the batch processing of the conversion of LC MS/MS mzXML[1] files to Mascot generic format.

File grid shows the files being converted or searched.

Selection panel lets user choose between different functions.

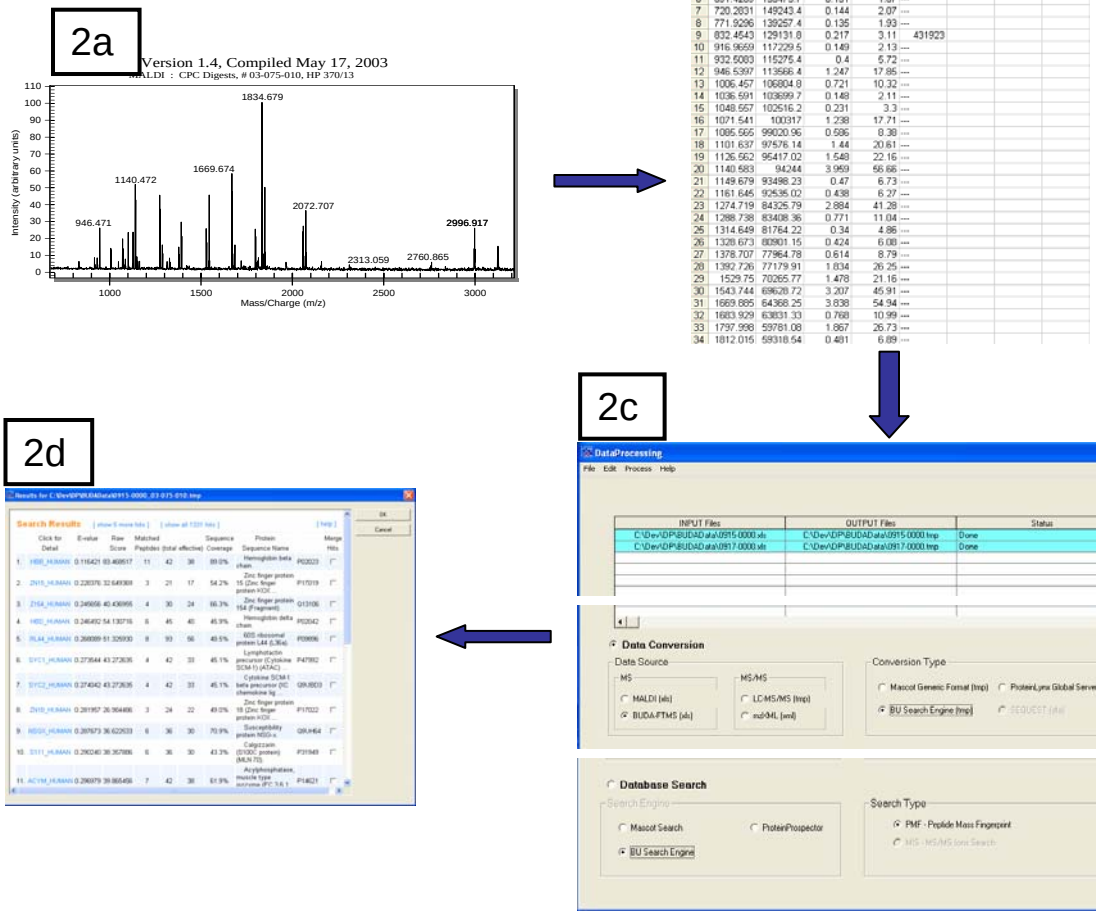

Fig 2.The pipeline of MS Data Analysis. (a). MALDI-FTMS spectrum shown in BUDA[3], (b) Processed peak list file produced by MassPike[4], (c) Conversion to BUPID[2] format and batch mode BUPID database search (d) Search result. All programs are developed in-house at BUSM.