Crossroads
The ACM Student Magazine

Achieving I/O Improvements in a Mass Spectral Database

by ***Eric Puryear***, ***Jennifer Van Puymbrouck***, ***David Sigfredo Angulo***, ***Kevin Drew***, ***Lee Ann Hollenbeck***, ***Dominic Battre***, ***Alex Schilling***, ***David Jabon***, ***Gregor von Laszewski***

Abstract

Research in proteomics has created two significant needs: the need for an accurate public database of empirically derived mass spectrum information and the need for managing the I/O and organization of mass spectrometry data in the form of files and structures. Lack of an empirically derived database limits the ability of proteomic researchers to identify and study proteins. Managing the I/O and organization of mass spectrometry data is often time-consuming due to the many fields that need to be set and retrieved. As a result, incompatibilities and inefficiencies are created by each programmer handling this in his or her own way. Until recently, storage space and computing power has been the limiting factor in developing tools to handle the vast amount of mass spectrometry information. Now the resources are available to store, organize, and analyze mass spectrometry information.

The Illinois Bio-Grid Mass Spectrometry Database is a database of empirically derived tandem mass spectra of peptides created to provide researchers with an organized and searchable database of curated spectrum information to allow more accurate protein identification. The Mass Spectrometry I/O Project creates a framework that handles mass spectrometry data I/O and data organization, allowing researchers to concentrate on data analysis rather than I/O. In addition, the Mass Spectrometry I/O Project leverages several cross-platform and portability-enhancing technologies, allowing it to be utilized on a variety of hardware and operating systems.

Introduction

Mass Spectrometry

Mass Spectrometry is used in proteomics to determine the mass and quantity of amino acids in a given sample. The knowledge of which amino acids are present and their concentrations allows researchers to determine which peptides and proteins comprise the sample. This is done using a mass spectrometer, which consists of the following seven major components: the sample inlet, ion source, mass analyzer, detector, vacuum system, instrument-control system, and data system.

The Illinois Bio-Grid Mass Spectrometry Database

The analysis and identification of data produced from MS/MS spectra is a difficult process to perform manually. For this reason, database searching is the most popular approach [6]. There are three types of databases available for searching. The first is the primary nucleotide sequence database, which contains genomic data or DNA base pairs. In order to search for a protein, database search programs must convert nucleotide data to amino acid data. The second type of database is the comprehensive protein sequence database, which is derived from nucleotide databases. The third type of database is the curated protein database, which contains the sequence, function, and specific characteristics of the protein.

The Illinois Bio-Grid Mass Spectrometry Database (IBG-MSD) is a public database of curated and annotated empirically derived mass spectra of peptides. The goal of the database is to address the need for a public database of mass spectrometry data and implement a useful web interface that allows

researchers to access the data and perform a variety of tasks based on their individual needs [15].

The Mass Spectrometry I/O Project (MSIOP)

The purpose of the MSIOP is to provide an I/O and data structure framework for mass spectrometry analysis tools and separate biologically significant programming from I/O. Handling the I/O and organization of data inside our framework and exposing a simple API allows researchers to spend their development time on code that accomplishes their research objectives.

First envisioned in the spring of 2004 and implemented during the summer of that same year, the MSIOP was designed for use with mass spectrometry data from a variety of mass spectrometers and file formats [12]. The MSIOP stores an abundance of metadata describing the experiment's conditions, in addition to the peak lists, allowing researchers to annotate their data to facilitate collaboration and curation of the data. The original MSIOP stores data in three structures. The top-level structure is called an accession record and contains information about the researcher, the mass spectrometer, and other metadata. The intermediate-level structure is the mass spectrum, which holds metadata that pertains to the lowest-level structure, which contains mass intensity pairs. The structure of the MSIOP is depicted in Figure 1.

The I/O portion of the MSIOP consists of modules for reading proprietary file formats, such as the .dta and .mgf formats, as well as the community standard mzXML format. These modules create the appropriate MSIOP data storage structure from a mass spectrometry data file [13].

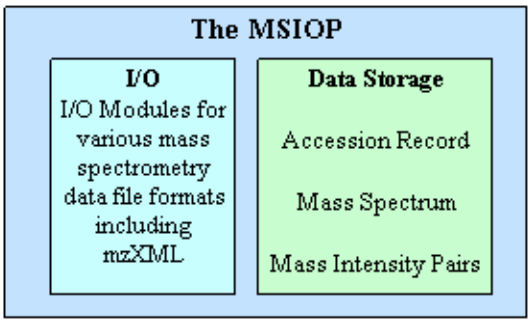

Figure 1: The structure of the MSIOP, showing the three main structures and the I/O module.

Since it was first developed, the structure of the MSIOP has closely followed that of mzXML, a community standard for the storage of mass spectrometry data. The similarities between mzXML and the MSIOP allow for easy and efficient conversion between the MSIOP's internal data structures and an mzXML file. Figure 2 shows a small sample of an mzXML file that highlights similarities to the MSIOP. For example, the first ten lines are equivalent to part of an accession record in the MSIOP, while line 11 represents data that is stored in a mass spectrum. In an actual mzXML file, there would be an element named peak, which stores the peak list; this element maps to the mass intensity pairs structure in the MSIOP.

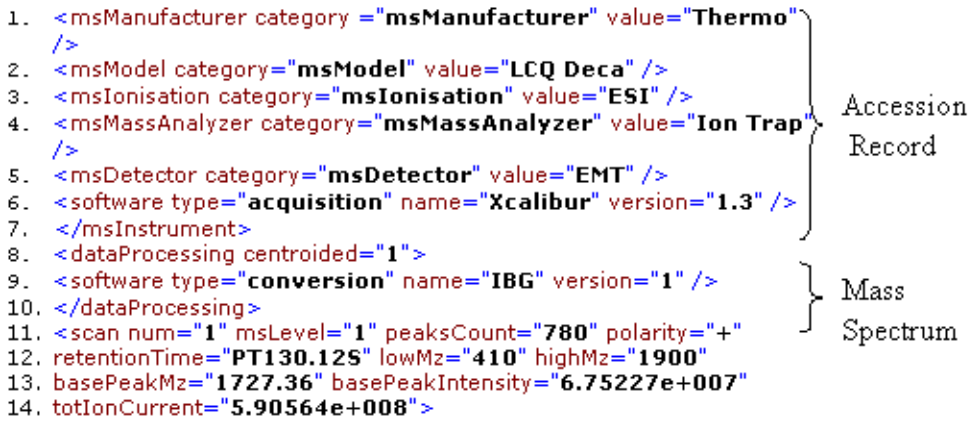

Figure 2: Data in the mzXML format.

One of the most important features of the MSIOP is mzXML support. The mzXML format is quickly becoming the preferred method of storing mass spectrometry data and is supported by companies like Matrix Science. Because mzXML is platform-independent and stores ample amounts of derived data as well as metadata related to the experiment's parameters, mzXML is being integrated into an increasing number of mass spectrometry toolkits, including the MSIOP--especially since The Institute for Systems Biology began promoting the mzXML standard [7].

Motivation

IBG Database and Import Module

Currently, mass spectrometry proteomic data sets are analyzed with the same algorithms developed five to ten years ago to interpret mass spectra [10]. The Sequest and Mascot algorithms are examples of mass spectra database search algorithms [9] [3]. When a user submits raw data to these search engines, theoretical mass spectra are generated for a set of candidate peptides from the sequence databases. These spectra are then compared with the experimental spectra using a matching function.

De novo sequencing is a protein identification tool that serves as an alternative to database searching. It involves analyzing a spectrum to determine the sequence of the peptide that is represented by that spectrum. This software is a valuable tool for a researcher unable to get valid results using a database search, but the results are less accurate. PEAKS and Lutefisk are examples of this type of search software [6] [14].

While these search algorithms have existed for a long time, they often give false positives, incorrect identifications, or improperly scored identifications. A manual inspection of the results is often needed [6]. In addition, proteins often undergo modifications and are mutants of a wild type, have a post-translational modification, or have gaps in ion sets. While these modifications will be present in the mass spectra of the protein, the database searches and de novo sequencing may not return correct results. Furthermore, the search algorithms do not take intensity data into account, as intensity data is very difficult to obtain theoretically.

The motivation of the IBG-MSD is to address the shortcomings of current database searches and de novo sequencing discussed above and make peptide and protein identification more accurate. To achieve this, the IBG-MSD uses empirical data rather than theoretical data derived from a protein sequence. The use of empirical data allows for more accurate protein identification, especially in cases of post-translational modifications.

An important step in the development of the database is to populate it with accurate data in order to make

it a viable resource. The usual way to submit spectra to the mass spectrometry database is through the web interface. If users have what they believe to be an accurately identified protein, they may submit the spectra for curation. The IBG-MSD administrator then sends the data to an independent curator for validation of the data. If it is validated, the data is added to the database. This process is often too tedious for users, and the need arose for batch entry of mass spectra, so a batch entry module for the IBG-MSD was developed.

In order to populate the mass spectrometry database with an ample amount of data, a batch import program was written in Java. Part of the functionality of the module is that it can take in metadata from a comma-delimited file and match this metadata with spectra contained in .mgf files. The comma-delimited file is read and parsed to obtain the metadata, and the .mgf file names are parsed. The metadata is then matched to each .mgf file based on unique identifiers found in the comma-delimited file and .mgf file name, such as the target plate, target well, and precursor mass of each mass spectrum. The .mgf file is then converted to a .dta file and stored in the service module. The metadata, including the corresponding file name of the .dta file in the service module, is then stored in the mass spectrometry database and can be viewed through the web interface.

Another function of the module is to read data in the form of mzXML files. The metadata and spectra from mzXML files are obtained using classes generated by the Java Architecture for XML Binding framework (JAXB) [15].

Spectral Comparison

Spectral comparison implements two different methods of comparing mass spectra. These are the similarity index method and the cosine similarity method [16]. Spectral comparison uses bins of a user-defined size to hold one or more peaks, which are then compared. If more than one peak resides in a bin, spectral comparison averages the peaks to get a single value, which is then compared to the value in the corresponding bin of the other spectrum. Given the spectra in Figure 3, it would be difficult to get an accurate comparison between the spectra without variable bin size, despite their obvious similarities.

Spectrum 1 (m/z, intensity): 100, 1	200, 2	300, 5
Spectrum 2 (m/z, intensity): 103, 1	202, 2	301, 5

Figure 3: An example of how bin size can be used to compensate for slight variances in the mass to charge ratio (m/z).

With spectral comparison's variable bin size, the spectra in Figure 3 are recognized as being similar when a bin size of five Daltons is set. This allows peaks to shift due to noise and the differences between mass spectrometers without adversely affecting the ability to compare spectra.

Cosine Similarity

Cosine similarity is the preferred method of comparing spectra in spectral comparison. This method implements the formula in Equation 1, which represents each of the spectra as a vector in n-dimensional space--where n is the number of peaks--and compares the cosine of the angle between the vectors [12], as seen in Figure 4.

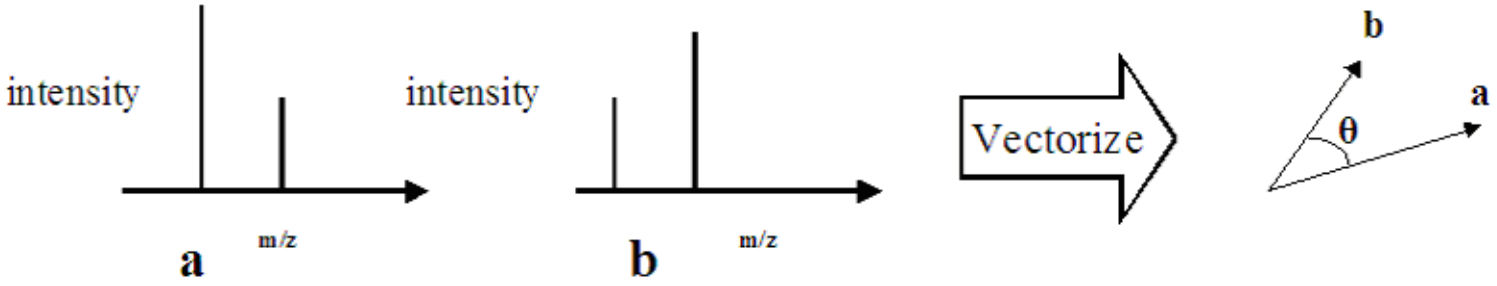

Figure 4: A simple example of how spectra are vectorized, and the resulting vectors.

$$\text{COS}\,\Theta = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2 b_i^2}}$$

Equation 1: The formula used by Spectral Comparison for Cosine Similarity, with a and b representing the two spectra to be compared [16].

Similarity Index

The similarity index method, as developed by Wan et al. [16], compares the difference in signal intensity by the smaller intensity for peaks that fall within a given mass range. The formula for the similarity index formula is shown in Equation 2.

$$SI = \Sigma \frac{\sqrt{I - I_0}}{I_0} \times 100$$

Equation 2: The formula used by Spectral Comparison for Similarity Index, with I and I_0 representing the two spectra to be compared [16].

Design

The IBG-MSD design includes six main components. The user component is a web interface that allows access to the IBG-MSD. Two databases were constructed, one to store mass spectrometry data and another to store the results of searches submitted by users. A service module was implemented that serves as a compute node, and handles spectral comparisons and search queries submitted by users. When a user submits a spectrum for identification, a separate spectral comparison module compares the user's spectrum to each spectrum in the mass spectrometry database. Finally, a batch import module imports data into the database. These components and their relationship to each other are depicted in Figure 5.

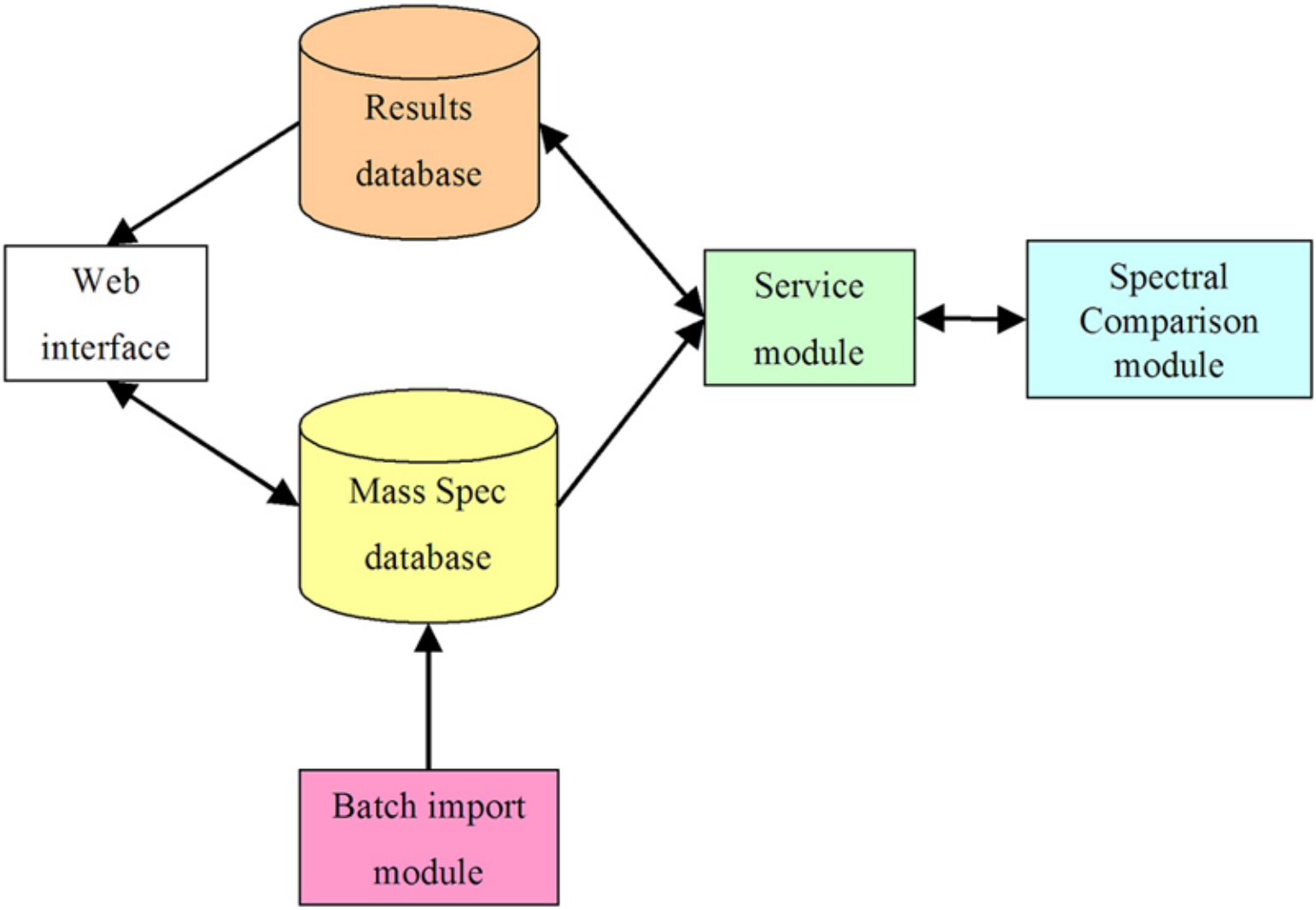

Figure 5: IBG-MSD main components.

The batch import module communicates with the mass spectrometry database, which consists of fifteen tables. The main tables are mass_spectrum, accession_record, curation_history, users, and annotations. These tables are shown in Figure 6. The additional tables are static tables containing default values for metadata fields in the five main tables.

The mass_spectrum table contains metadata about each spectrum. The accession_record table is a child of the mass_spectrum table and contains information relating to the mass spectrum record. There may be multiple mass spectrum records that relate to the same accession record. The curation table contains information regarding the curation status of a mass spectrum record. The annotations table contains fields with additional comments about a mass spectrum record. Finally, the users table holds information about all users registered with the IBG-MSD.

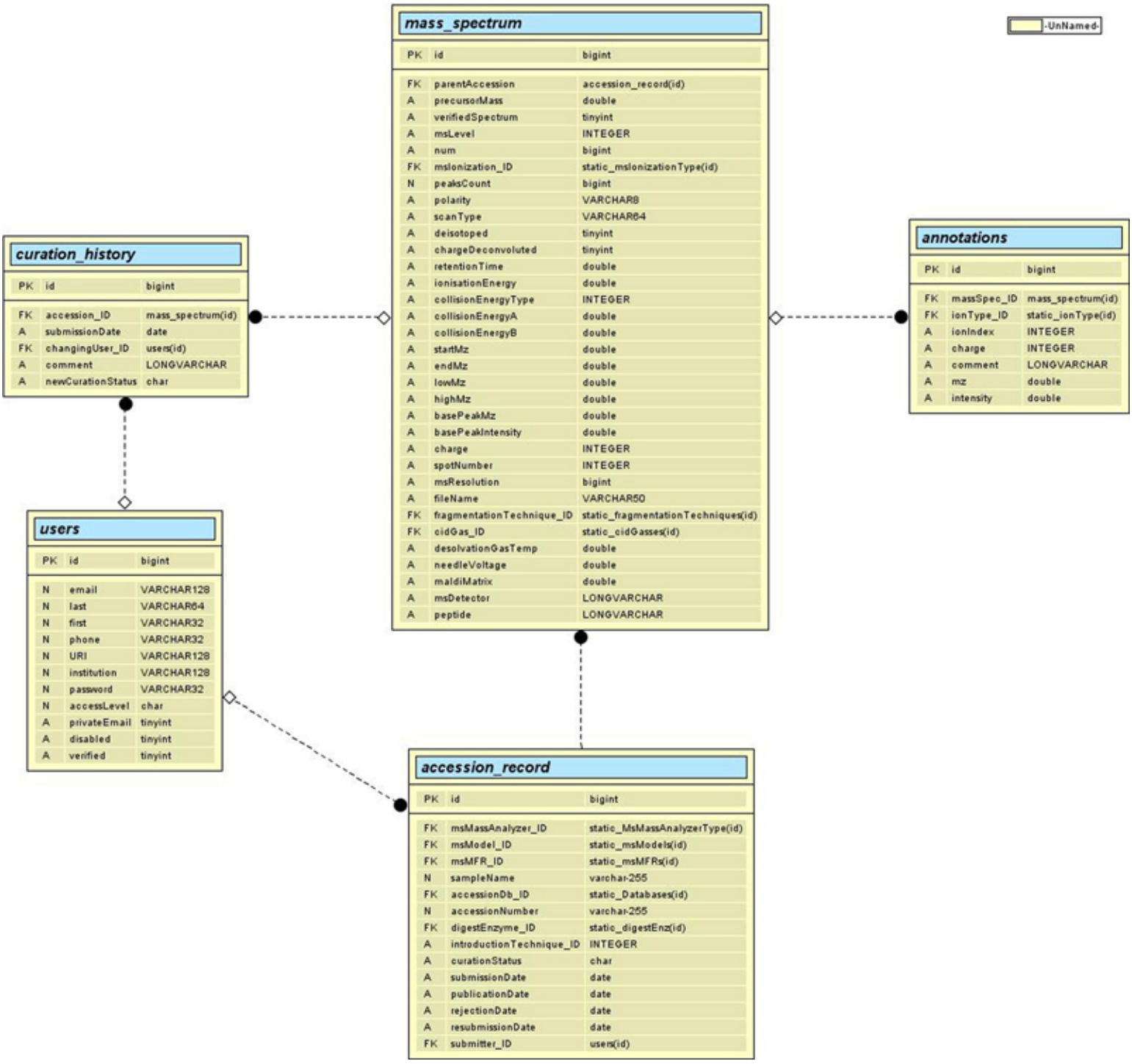

Figure 6: Main tables of the Mass Spectrometry database.

Import Module

The batch import module was tested on data obtained in the Aurum data set from Proteome Commons [11] and multiple data sets from Peptide Atlas [8]. The Aurum data set contained approximately 3000 spectra. The metadata was contained in a comma-delimited file and the spectra were in .mgf files. Proteins in this data set were digested with trypsin and analyzed using an ABI 4700 (MALDI Tof/Tof) mass spectrometer. The data set includes over 250 known proteins, and each has been checked for purity using 1D gel analysis. The proteins were overexpressed in E. Coli and purified using a sequence tag by Genway [4]. The data sets from Peptide Atlas contained approximately 462,000 spectra. The data was stored in mzXML files, and the proteins were digested with trypsin.

Currently, there are 118,955 spectra in the IBG-MSD. These spectra can be searched, downloaded,

and mined. These spectra represent proteins from humans and yeast. The number of spectra will increase as more curated mass spectrum data is obtained, which will increase the usefulness of the database [15].

MSIOP Performance

The performance of the MSIOP has been evaluated and found to be satisfactory. Table 1 shows read- and write-speed performance statistics for the mzXML I/O modules when running on an Intel Pentium 4 3.2GHz processor with 2GB of RAM under Linux 2.4 with a 7200 RPM ATA100 Hard Drive.

Read Or Write	User Time	System Time	File Size (KB)	KB/s
Read	0.87s	0.090s	14216K	16340
Write	7.46s	0.10s	14216K	1905

Table 1: Performance results for the MSIOP mzXML I/O modules.

The MSIOP is currently utilized by several tools. These include Spectral Comparison, HDXRates, and SpectralMatch [2]. Each of these tools uses the MSIOP framework to handle file I/O and internal data management.

Conclusion

Since the Mass Spectrometry database is now populated with a significant number of spectra, it can be made available to proteomic researchers who are seeking a unique database of empirically derived mass spectra of peptides for identification of proteins. By using the IBG-MSD, proteomic researchers can more accurately identify unknown proteins.

Future Work

Having created a foundation for the I/O framework and import module for the IBG-MSP database, future tools could be built to utilize these components. These may include database query tools and database download tools. In addition, continual importation of mass spectrometry data is necessary to make the IBG-MSP a viable resource for researchers.

Acknowledgments

This research was performed by members of the Illinois Bio-Grid. The research was funded by the National Science Foundation under Grant No. 0353989 and supported by Argonne National Laboratory.

References

[1] Beazley, David M. Tcl and SWIG as a C/C++ Development Tool. Department of Computer Science,University of Chicago. 1998.

[2] Drew, Kevin; Angulo, David; Schilling, Alex; Freeman, Tim. Mass Spectra Analysis on the Illinois Biogrid. Proceedings of 2004 Midwest Software Engineering Conference, Chicago. June 2004. (Poster). (Electronically published http://facweb.cs.depaul.edu/bioinformatics/publications/MSEC-MassSpec-Poster 2004.ppt).

[3] Eng, Jimmy K.; McCormack, Ashley L.; Yates III, John R. An Approach To Correlate Tandem Mass Spectral Data Of Peptides With Amino Acid Sequences In A Protein Database. J Am Soc Mass Spectrum, 1994, Number 4, pages 976-989.

[4] Web address: http://www.genwaybio.com/. Last accessed August 18, 2005.

[5] Kinter, M., Sherman, N. E. Protein Sequencing and Identification Using Mass Spectrometry. Wiley-Interscience, N.Y., 2000.

[6] Ma, Bin; Zhang, Kaizhong; Hendrie, Christopher; Liang, Chengzhi; Li, Ming; Doherty-Kirby, Amanda; Lajoie, Gilles. PEAKS: Powerful Software For Peptide De Novo Sequencing By Tandem Mass Spectrometry. Rapid Communications in Mass Spectrometry, 2003, Num. 17, page 2337.

[7] Pedrioli, Patrick G. A common open representation of mass spectrometry data and its application to proteomics research. Nature Biotechnology. Volume 22, issue 11, pages 81-92, November 2004.

[8] Web address: http://www.peptideatlas.org. Last accessed August 18, 2005.

[9] Perkins, DN; Pappin, DJ; Creasy, DM; and Cottrell, JS. Probability-based Protein Identification By Searching Sequence Databases Using Mass Spectrometry Data. Electrophoresis, 1999, Vol. 20, Num. 18, page 3551.

[10] Prince, John T.; Carlson, Mark W.; Wang, Rong; Lu, Peng; Marcotte, Edward M. The Need For A Public Proteomics Repository. Nature Biotechnology, April 2004, Volume 22, Number 4, page 471.

[11] Web address: http://www.proteomecommons.org/archive/1122567790437/index.html. Last accessed August 18, 2005.

[12] Puryear, Eric; Angulo, David; Drew, Kevin; Schilling, Alex; von Laszewski, Gregor. Developing a Distributed and Scalable Foundation for Mass Spectrometry Data. DePaul University CTI Tech Report, pages 1-7. April 2005.

[13] Puryear, Eric; Angulo, David; Drew, Kevin; Schilling, Alex; von Laszewski, Gregor. Comparing Mass Spectra. DePaul University CTI Tech Report, pages 1-7. March 2006.

[14] Taylor, Alex J.; Johnson, Richard S. Sequence Database Searches Via De Novo Peptide Sequencing By Tandem Mass Spectrometry. Rapid Communications in Mass Spectrometry, 1997, Volume 11, pages 1067-1075.

[15] Van Puymbrouck, Jennifer; Angulo, David; Battre, Dominic; Drew, Kevin; Hollenbeck, LeeAnn; Jabon, David; Schilling, Alex; von Laszewski, Gregor. A Batch Import Module for an Empirically Derived Mass Spectral Database. DePaul University CTI Tech Report, March 2006.

[16] Wan, Katty X, Vidavsky, Ilan, and Gross, Michael L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle.

Biography

Eric Puryear (**epuryear@gmail.com**) is a first year DePaul University Law Student who will begin his MS in Computer Science next year. He has a BS in Computer Science from DePaul University.

Jennifer Van Puymbrouck (**jvanpuy@gmail.com**) is currently a Master's student in Computer Science at DePaul University. She has a BS in Scientific Data Analysis and Visualization from DePaul University.

David Sigfredo Angulo: DePaul University, (**dangulo@cti.depaul.edu**).

Kevin Drew: The University of Chicago, (**kdrew@uchicago.edu**).

Lee Ann Hollenbeck: DePaul University, (**lee@keynet.net**).

Dominic Battre: DePaul University, (**dominic@battre.de**).

Alex Schilling: University of Illinois at Chicago, (**aschilli@uic.edu**).

David Jabon: DePaul University, (**djabon@depaul.edu**).

Gregor von Laszewski: Argonne National Laboratory, (**gregor@mcs.anl.gov**).