A Batch Import Module for an Empirically Derived Mass Spectral Database

Jennifer Van Puymbrouck, [1] David Angulo, [2] Kevin Drew, [3]
Lee Ann Hollenbeck, [4] Dominic Battre, [5] Alex Schilling, [6]
David Jabon, [7] Gregor von Laszewski, [8]

[1] DePaul University, jvanpuy@gmail.com
[2] DePaul University, dangulo@cti.depaul.edu
[3] The University of Chicago, kdrew@uchicago.edu
[4] DePaul University, lee@keynet.net
[5] DePaul University, dominic@battre.de
[6] University of Illinois at Chicago, aschilli@uic.edu
[7] DePaul University, djabon@depaul.edu
[8] Argonne National Laboratory, gregor@mcs.anl.gov

Abstract

Proteomic researchers who study mass spectrometry data have expressed a need for an accurate public database of empirically derived curated mass spectrum information. Lack of such a database limits proteomic researchers in their ability to identify and study proteins. 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. This paper will discuss the methods used to import the mass spectra into the Illinois Bio-Grid Mass Spectrometry Database, as well as the database requirements, motivation, use cases, design, and results.

Keywords: Mass spectrometry, proteomics, database search

1 Introduction

Mass Spectrometry (MS) in proteomics is a powerful analytical technique that is used to identify amino acid sequences and identify proteins. A mass spectrometer is an instrument that measures the mass-to-charge ratio of individual molecules that have been converted into electrically charged gas-phase molecules, or ions [1]. These ions are filtered in such a way as to produce an ordered separation of the ions as they pass through the instrument [2], ordered from lower to higher mass to charge ratios. The data is typically displayed as a plot of intensity vs. mass to charge ratio. The ionization techniques generally used with peptides in most proteomic analyses are Matrix Assisted Laser Desorption Ionization (MALDI) and Electrospray Ionization (ESI). [3]

Mass spectrometry is experiencing a period of rapid growth based on applications in proteomic analyses. As a consequence of this rapid growth, there is an urgent need to collect the large amounts of information produced and make it available to researchers worldwide. Not only must the information be available, it must also be organized and searchable.

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 pubic database of mass spectrometry data and implement a useful web interface that will allow researchers to access the data and perform a variety of tasks based on their individual needs.

The identification of proteins is becoming an easier and more accurate process due to the use of tandem mass spectrometry (MS/MS). MS/MS involves two stages of mass analysis in a single experiment. The first stage filters out the ions of interest from the sample. These ions are then passed into a collision cell that fragments the ions. The second stage then separates and detects the ions. [3] Although the generation of raw MS/MS spectra has become easier, the analysis and identification of that data is a difficult process to perform manually. For this reason, database searching is the most popular approach. [4]

There are three types of databases available for searching. The first type of database is the primary nucleotide sequence database, which contains genomic data or DNA base pairs. In order to search for a protein in this type of database, 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 form nucleotide databases. The third type of database is the curated protein database, which contains the sequence, function, and specific characteristics about the protein.

2 Requirements/Motivation

Currently, mass spectrometry proteomic data sets are analyzed with the same algorithms developed 5−10 years ago to interpret mass spectra. [5] The Sequest and Mascot algorithms are examples of mass spec database search algorithms. [6] [7] When a user submits her/his raw data to these types of search engines, peptides from a sequence database are compared to the raw data. More specifically, theoretical mass spectra are generated for a set of candidate peptides from the sequence databases and these spectra are then compared with the experimental spectra using a matching function.

Mascot is a search engine that uses mass spectrometry data to identify proteins from primary sequence databases. Mascot combines three types of searches. It uses experimental data of peptide molecular weights from the digestion of a protein by an enzyme, the use of tandem mass spectrometry (MS/MS) data from one or more peptides, and the combination of mass data with amino acid sequence data. The scoring algorithm is probability based, which has three important implications.

First, a statistical rule can be used to judge whether a result is significant or not. Second, scores can be compared with those from other types of searches, such as sequence homology. Finally, search parameters can be readily optimized by iteration. [6]

Sequest uses a complex cross-correlation scoring routine to match tandem mass spectra to model spectra derived from peptide sequences from a database. [3]

To assess a match, Sequest uses the difference between the first and second ranked sequences. This value is dependent on the database size, search parameters, and sequence

homologies. [7] The result of this type of algorithm is a more accurate search result, since several factors are taken into account in the scoring**.**

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. Using this type of software is a valuable tool if a researcher is unable to get valid results using a database search. If a protein is not listed in the database or has undergone a mutation, the database search will not return correct results. Since de novo sequencing does not use databases, researchers have an alternative means for protein identification, however the results are less accurate. PEAKS and Lutefisk are examples of this type of search software. [4] [8]

While these search algorithms have been implemented and in existence for a long period of time, they often give false positives, incorrect identifications, or improperly scored identifications, therefore a manual inspection of the results is often needed. [4] 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 a significant amount of accurate data in order to make it a viable resource. The usual way to submit spectra to the Mass spectrometry database is from a user through the web interface. If a user has what she/he believes to be an accurately identified protein, the user may submit the spectra for curation. The IBG-MSD administrator then sends the data to an independent curator for validation of the data, and 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 spectrum. This paper discusses the development of the batch entry module of the IBG-MSD.

3 Use cases

Several use cases have been identified for the IBG-MSD. For these use cases, the symbol **U** represents a user accessing the IBG-MSD web site, **W** represents the web interface and compute node, **M** represents the Mass Spectrometry database, **TR** represents the Tasks/Reports database, **S** represents the spectral comparison module, **A** represents an administrator, **C** represents a curator, and **I** represents the import module. These use cases are diagrammed in Figure 1.

i. A user registers at the IBG-MSD in order to gain access to the database.
ii. A user submits, edits, or deletes an accession record (described in Figure 3).
iii. A user submits, edits, or deletes a mass spectrum record (described in Figure 3).
iv. A user views public accession records.
v. A user views her/his private records.

vi. A user submits an accession record for curation.
vii. A user uploads a `.dta`, `.mgf`, or `.mzXML` file.
viii. A user submits unknown spectrum/spectra for identification.
ix. A user downloads database for local use.

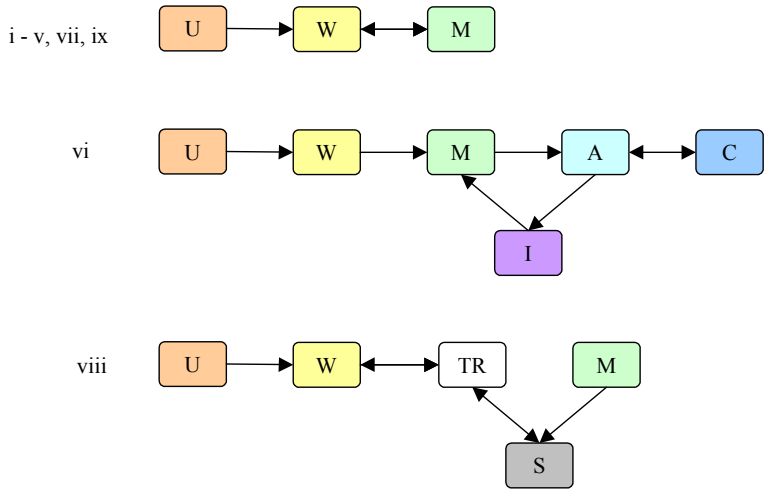

Figure 1: Use cases

U = User, **W** = Web interface and compute node, **M** = Mass spectrometry database, **TR** = Tasks/Reports database, **S** = Spectral comparison module, **A** = Administrator, **C** = Curator, **I** = Import module

4 Design

The IBG-MSD was designed based on the requirements and use cases. The 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 one to store the results of searches submitted by users. A service module was implemented that serves as a compute node and handles search queries submitted by users and spectral comparisons. When a user submits a spectrum for identification, a separate spectral comparison module compares the user's spectrum with each spectrum in the mass spectrometry database. Finally, a batch import module is used to import data to the database. These components and their relationship to each other are depicted in Figure 2.

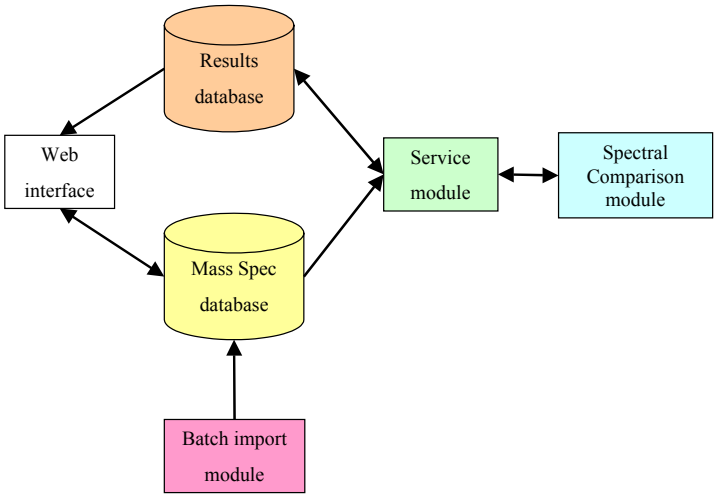

Figure 2: Components of the IBG-MSD

The focus of this paper is on the batch import module. This module communicates with the Mass Spectrometry database component, which consists of fifteen tables. The main tables are mass_spectrum, accession_record, curation_history, users, and annotations. These tables are described in Figure 5. The additional tables are static tables that contain 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 that relates 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 and 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.

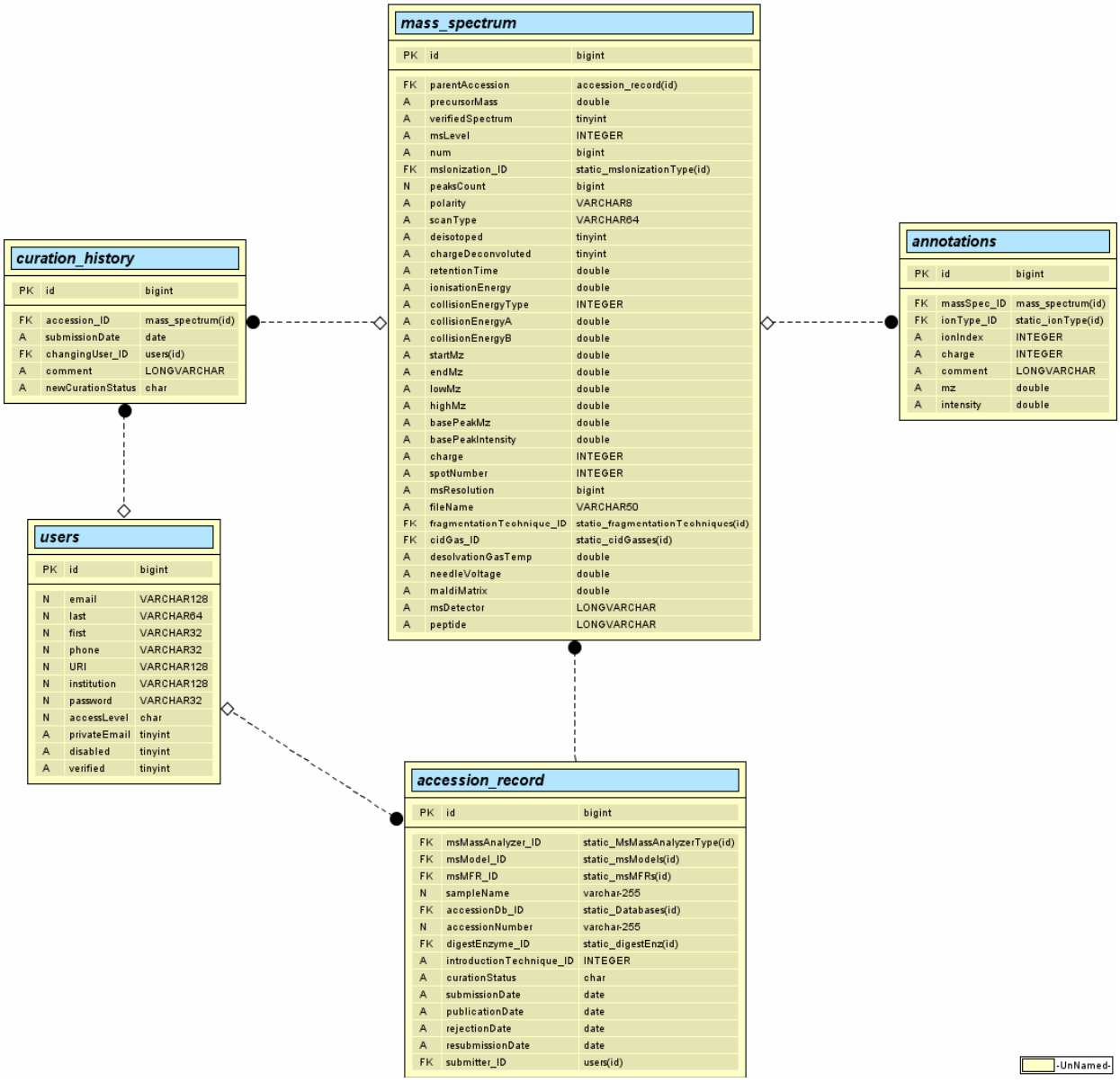

Figure 3: Main tables of Mass Spectrometry database

5 Implementation

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 functionality of the module is to read in 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).

6 Results

The batch import module was tested on data obtained in the Aurum data set from Proteome Commons [9] and multiple data sets from Peptide Atlas. [10] The Aurum data set contained approximately 3,000 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 over expressed in E. Coli and purified using a sequence tag by Genway. [11] The data sets from Peptide Atlas contained approximately 462,000 spectra. The data was in the form of `.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 certainly increase as more curated mass spectrum data is obtained, which will increase the functionality of the database.

7 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 spectrum of peptides for identification of proteins. By using the IBG-MSD, proteomic researchers will be able to more accurately identify unknown proteins.

Acknowledgements

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] Chiu, Chia M.; Muddiman, David C. *What Is Mass Spectrometry?* Web Page. Available from http://www.asms.org/whatisms. Last accessed July 26, 2005.

[2] Herbert C. G., Johnstone, R.A.W. *Mass Spectrometry Basics*. CRC Press, Boca Raton, Fl, 2003.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] http://www.genwaybio.com/