Developing a Distributed and Scalable Foundation for Mass Spectrometry Data

Eric Puryear[1], David Angulo[2], Alex Schilling[3], Kevin Drew[4], Gregor von Laszewski[5]

[1] DePaul University, epuryear@students.depaul.edu
[2] DePaul University, dangulo@cs.depaul.edu
[3] The University of Chicago, aschilli@uchicago.edu
[4] The University of Chicago, kdrew@uchicago.edu
[5] Argonne National Laboratory, gregor@mcs.anl.gov

Abstract

Mass spectrometers are tools in the field of proteomics. Each mass spectrometer manufacturer uses its individual data format and software tools, making the creation of additional software tools and databases difficult and incompatible with one another. The Mass Spectrum I/O Project (MSIOP) addresses this problem, allowing for storage and analysis of mass spectrometer data from multiple manufacturers across various platforms while providing a framework upon which mass spectrometry software tools can be constructed.

Keywords: Proteomics, Mass Spectrometry, mzXML, Parallel Computing, Grid Computing, Grid Technology.

1 Introduction

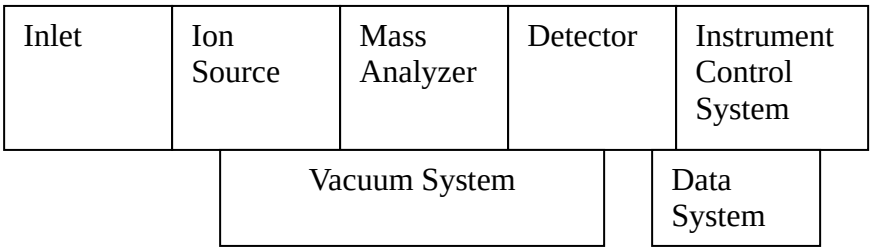
Mass spectrometers are used in Proteomics to determine the mass and abundance of the various amino acids that comprise a particular peptide or protein. The basic operation of a mass spectrometer involves placing a sample in the inlet, ionizing the sample, separating the ionized sample by mass, and detecting the separated particles [4]. There are various types of mass spectrometers such as matrix-assisted laser desorption/ionization (MALDI), and electrospray ionization (ESI), each of which takes a different approach to ionizing and separating the particles that comprise a sample. Each type of mass spectrometer is best suited for particular tasks as each method of ionization and detection has its own advantages and disadvantages. Regardless of the specifics of the mass spectrometer(s), there is a fundamental problem of incompatible data formats, and a lack of infrastructure to handle mass spectrometry data. The Mass Spectrum I/O Project (MSIOP) was designed to address this problem, allowing for the use of data from a multitude of mass spectrometers, while providing a sound infrastructure upon which mass spectrometry tools can be built.

1.1 Mass Spectrometry

Researches use mass spectrometers to determine what amino acids are present and their concentration allows researchers to determine what peptides and proteins comprise the sample. This is done using a mass spectrometer, which is composed of the following seven major components: the sample inlet, ion source, mass analyzer, a detector, a vacuum system, instrument control system, and finally, the data system [4]. Although all of the components of a mass spectrometer are important, the first three of these tend to determine the major attributes of the mass spectrometer [4].

Basic Mass Spectrometer Components

Fig



ure 2

Although detailed information about the specific operation of the ion source and mass analyzer is beyond the scope of this paper, differences in these components can affect the results that a mass spectrometer can generate. For example, MALDI ionization produces only singly charged ions, while other ionization methods can produce multiply charged ions, drastically affecting the data output as a doubly charged ion has twice the mass to charge ratio as an equally massive, singly charged ion. These differences, coupled with the wide variety of data formats, lead to difficulties when creating software that can interpret data from the various formats. To overcome these problems, the MSIOP stores meta-data about the mass spectrometry run and the instrument that it was conducted on, allowing for adjustments to be made by analysis tools, based on the mass spectrometer type.

2 Problem

Mass Spectrometer Manufactures and Data Formats

Manufacturer	**File Format**
Thermo Instruments	RAW
Agilent Technologies	HP
ABI - Qstar QTOF	SCIEX
ABI - 4700 TOF-TOF	Oracle DB
Waters	MassLynx
Bruker Daltonics	Datastar

Figure 1

The utility of mass spectrometers is hindered by a variety of incompatible data formats that are the result each manufacturer's implementation of their own data format [Figure 1]. These incompatibilities force researchers to use the software tools bundled with each mass spectrometer [1]. This also complicates database construction and makes comparisons between results obtained from mass spectrometers difficult as different software packages interpret the data differently. Additionally, support for various computer platforms such as Linux and Solaris is lacking, while the closed-source nature of the manufacturers' software tools limit the ability of users to modify the source code of these tools to suit their individual needs.

The Mass Spectrum I/O Project (MSIOP) seeks to solve these problems by providing an open source means of converting data files from various mass spectrometer manufacturers into mzXML and a framework upon which mass spectrometry tools can be built [5]. The mzXML format is an XML-based standard for the storage of mass spectrometry data, designed to be flexible enough for use by all mass spectrometry researchers, and extendable to allow for future changes. In order to ensure cross-platform compatibility, the Cactus framework and C programming are used. The Cactus framework [2] enables cross platform

portability for the MSIOP by providing transparent type conversion when running on various platforms, along with other infrastructural support. The cross-platform nature of programs that utilize the Cactus framework allows the MSIOP to be deployed on a wide variety of hardware architectures and software environments, including a heterogeneous grid [7], without modifications.

3 Design

The MSIOP was designed to be a robust, efficient, and portable infrastructure for mass spectrometry data. These considerations resulted in the use of mzXML, the C programming language, and the Cactus framework. More information on mzXML and Cactus are provided below.

3.1 mzXML

As stated above, mzXML is an XML based format for the storage of mass spectrometry data. This open and extendable format is designed to be easy to implement and feature rich. Since it is based on XML, mzXML can be extended to suit future needs. Figure 2 shows a small sample of mass spectrometry data in the mzXML format [6].

3.2 Cactus Framework

A high priority for the MSIOP is cross-platform portability. This emphasis on portability is responsible for the use of the Cactus framework, as detailed in section 3.2. The Cactus framework allows for cross platform portability by replacing data types that vary by platform with Cactus variables that maintain their properties regardless of hardware architecture or software environment [2]. Another major benefit of the Cactus framework is that the user is able to quickly and easily change the program parameters on-the-fly, such as changing from reading in one file format to another file format without the need to recompile the program. This is done by activating or deactivating individual modules of code.

Example of Data in the mzXML format

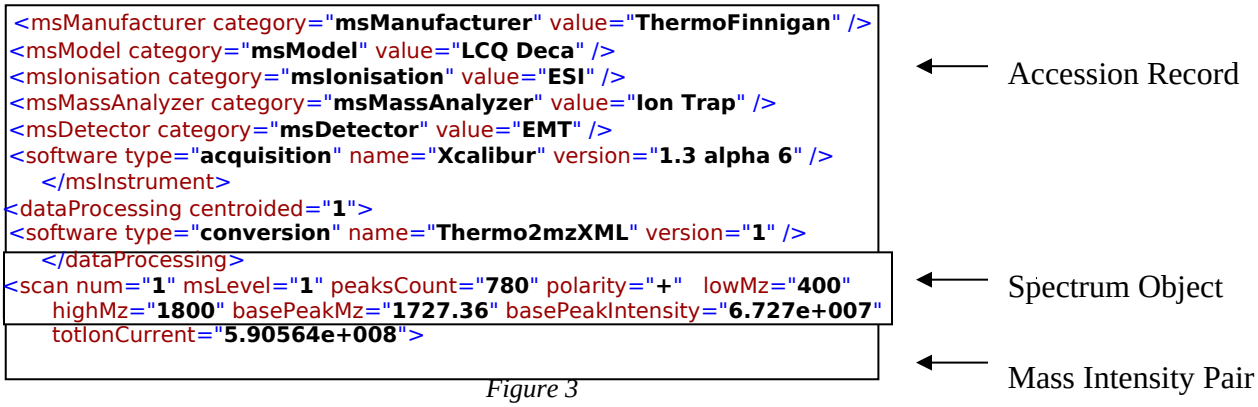


Figure 3

3.3 Structure of the MSIOP

The MSIOP consists of several code modules. Although some modules used by the MSIOP project are stand-alone, most depend on one or more additional modules to provide data types and functions. These modules and their specific functions are detailed below.

3.4 Spectrum Object

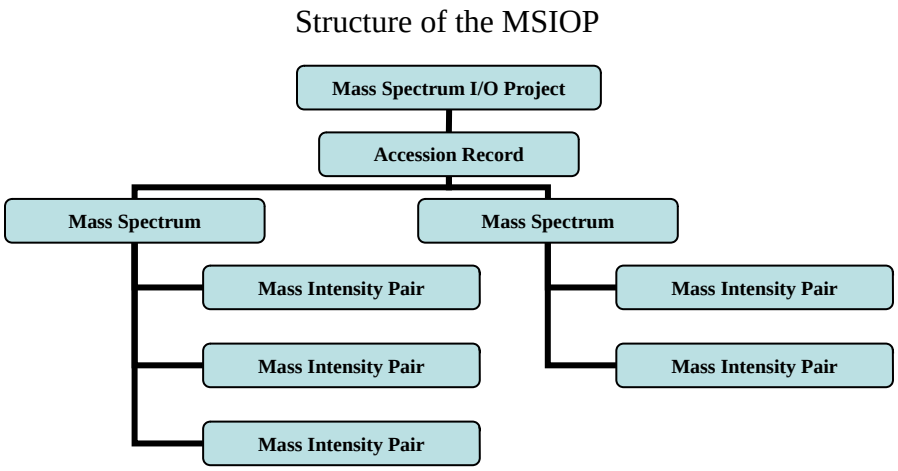


Figure 4

The Spectrum Object contains three substructures that are used to store the various data fields as defined by mzXML [6]. These are Accession Record, Mass Spectrum, and Mass Intensity Pair. Accession Record is used to store meta data about the mass spectrometer run that was performed, ranging from the email address of the operator, to the manufacturer and model number of the mass spectrometer itself. Accession Records have child Mass Spectrum, which contain data including whether the data has been manually verified by a researcher, or if it has been centroided. As an Accession Record has one or more child Mass Spectrum, Mass Spectrum has Mass Intensity Pair records as children. Mass Intensity Pair contain a mass and an intensity that represent a peak in the mass spectrometer output. This structure is shown in figure 4.

3.5 I/O

Input and output within the MSIOP are handled by the I/O modules. These are responsible for both parsing data files from various manufacturers and storing the values in a Spectrum Object; they are also used for writing the contents of a Spectrum Object to an mzXML file. Cactus allows the user to select which modules are activated at runtime, increasing code modularity by allowing the individual I/O modules to be modified or replaced without requiring changes to other sections of code. The structure of the I/O portion of the MSIOP is diagrammed in Figure 5.

Structure of the I/O Modules

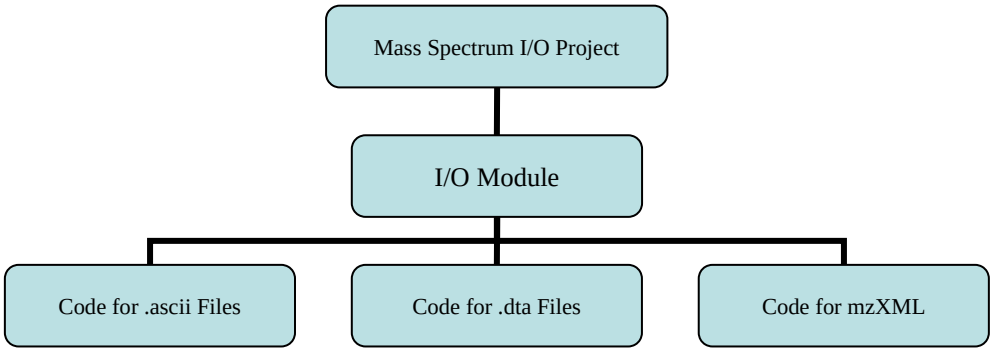


Figure 5

4 Result

The MSIOP is currently utilized by several tools including an open source De Novo sequencing tool. De Novo sequencing is the attempt to find the amino acid sequence of a given sample, usually using tandem mass spectrometry. Another tool that utilizes the MSIOP is a sequence database searching and matching program, shown in Figure 5. These programs and the other programs that will be developed take advantage of the flexible, portable nature of the MSIOP, allowing their developers access to a robust infrastructure, while minimizing the need to focus on I/O

5 Discussion

The MSIOP is intended to be utilization by researchers and programmers requiring I/O and a ready-made data structure for mass spectrometer data in either the numerous incompatible formats used by mass spectrometer manufacturers, or the mzXML standard. The ability to convert these proprietary data files into mzXML allows developers to focus on the functionality of their software instead of concerning themselves with I/O and file conversions. In addition, the structures used in the MSIOP are designed to be used as an infrastructure upon which other mass spectrometry tools can be constructed. Due to the MSIOP's modular nature, new sections of code can be integrated with ease when the need to read additional types of mass spectrometer data arises. This same modularity, combined with the freely available source code, allows users to modify the MSIOP to better suit their specific needs.

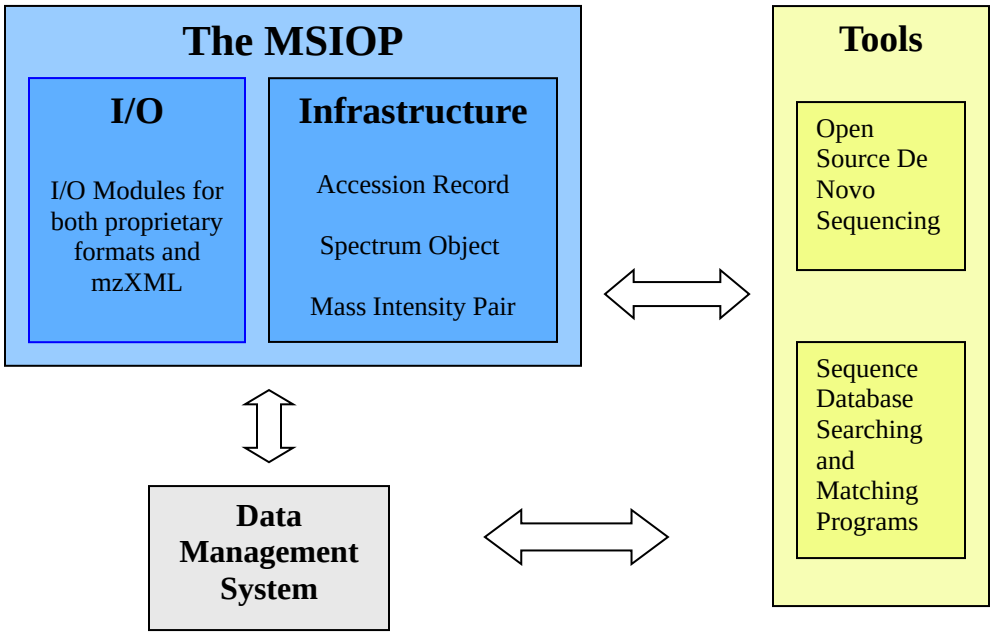


Figure 6

6 Further Work

Having created a foundation for the storage and I/O of mass spectrometry data, the next step will be to implement the MSIOP into additional programs that will perform analysis of mass spectrometry data. A data management system is currently being designed that will make use of the MSIOP for I/O and as a translation layer between the database and the various analysis tools, and a foundation upon which these tools will be built, such as the De Novo sequencing and sequence database searching and matching programs. As the need arises, new I/O modules will be added, allowing the MSIOP to read and write additional file types.

7 Conclusion

Mass Spectrometers are tools in the field of Proteomics, and unfortunately each mass spectrometer manufacturer uses their own data file format. This complicates data analysis and makes the use of the manufacturer's software tools necessary. The MSIOP allows mass spectrometer data to be converted from various formats into the universal mzXML format, and its modular and open source nature allows for new file types to be read and written as the need arises. In addition to providing I/O, the MSIOP is a solid foundation upon which mass spectrometry tools can be built, utilizing the already existing Accession Record, Mass Spectrum, and Mass Intensity Pair structures.

References

1. Drew, Kevin; Angulo, David; Schilling, Alex; Freeman, Tim. *Mass Spectra Analysis on the Illinois BioGrid.* Proceedings of 2004 Midwest Software Engineering Conference, Chicago.

2. Goodale, Allen, Lanfermann, Massó, Radke, Seidel, Shalf. *The Cactus Framework and Toolkit: Design and Applications.*
(http://www.cactuscode.org/Papers/VecPar_2002.pdf)

3. Krane, Dan and Raymer**,** Michael **.** *Fundamental Concepts of Bioinformatics.* Benjamin/Cummings, 2000.

4. Kinter, M., and Sherman N. Protein Sequencing and Identification Using Tandem Mass Spectrometry. Wiley-Interscience: New York, 2000.

5. Lesk, Arthur. *Introduction to protein architecture: the structural biology of proteins*. Oxford University Press. 2001.

6. Pedrioli, Patrick et al. *A common open representation of mass spectrometry data and its application to proteomics research*. Nature Biotechnology. 2004

7. Puryear, Eric; Angulo, David; Drew, Kevin; Schilling, Alex; Von Laszewski, Gregor. *Mass Spectrometry in the Illinois BioGrid*. (Poster) Proceedings of 2004 DePaul Natural Sciences, Mathematics & Technology Showcase. November 2004.

8. Steele Adam, Angulo, David S. *The Illinois BioGrid: A Prototype for Industry-Academe Collaboration*. Proceedings of 2003 Midwest Software Engineering Conference, Chicago.