

Mass Spectrometry in the Illinois Bio-Grid

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

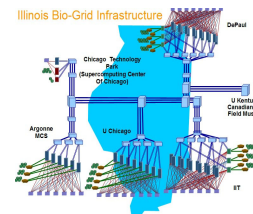
^[5]Argonne National Laboratory, gregor@mcs.anl.gov

Summary:

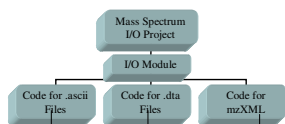
Mass spectrometers are used for proteomics research in the Illinois Bio-Grid; unfortunately each mass spectrometer manufacturer uses their own data format and software tools, thereby making the construction of unified, accurate databases difficult. These incompatibilities force researchers to use the software tools bundled with each mass spectrometer. This not only complicates database construction and analysis tool development, but also 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, as is the ability to modify the manufacturer provided software tools to fit individual needs. 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 using the Cactus framework.

How and why Mass Spectrometry is used in the Illinois Bio-Grid:

Mass spectrometry is used to determine the sequence of the amino acids that comprise a particular peptide or protein. With this information, it is possible to simulate complex biological events, such as protein folding. The use of grids is due to the computationally intensive nature of bioinformatics research. Complex simulations, such as protein folding, that would otherwise require months or years on a single computer can be completed in a matter of hours or days on the thousands of computers that comprise the Illinois BioGrid. These protein folding simulations, along with other computationally intensive tasks, can be used to aid the development of new drugs, and gain a better understanding of the biological processes of the world.

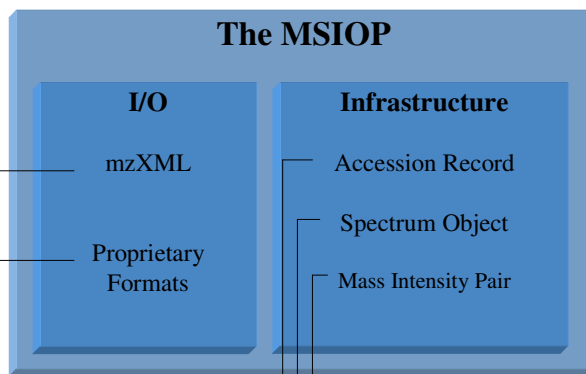


The Mass Spectrometry I/O Project (MSIOP)



I/O

Input and output within the MSIOP are handled by the I/O modules. These are responsible for both reading in 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. Users can select which modules are activated at runtime, increasing code modularity by allowing the individual I/O modules to be modified or replaced without necessitating changes to other sections of code.

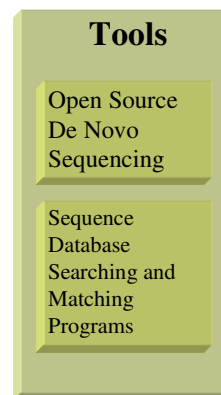


Infrastructure

The infrastructure of the MSIOP consists of three structures that are used to store data. 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, as well as a variety of other data. 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.

Tools

The MSIOP is currently utilized by several tools including an open source De Novo sequencing tool and by sequence database searching and matching. The MSIOP is ready to function as a foundation upon which other mass spectrometry analysis tools can be built, allowing the developers of these tools to focus on developing more efficient and effective tools instead of I/O and data structures.



mzXML

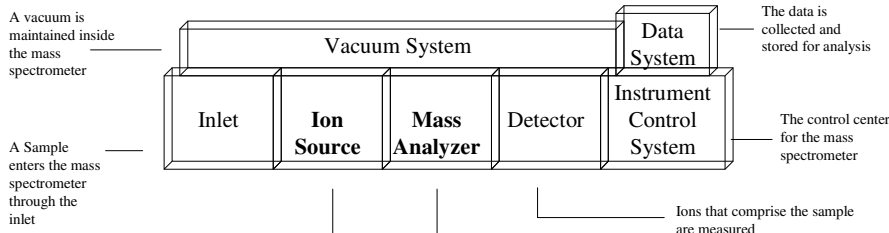
```
<msManufacturer category="msManufacturer" value="ThermoFinnigan" />
<msModel category="msModel" value="LCQ Deca" />
<msIonisation category="msIonisation" value="ESI" />
<msMassAnalyzer category="msMassAnalyzer" value="Ion Trap" />
<msDetector category="msDetector" value="EMT" />
<software type="acquisition" name="Xcalibur" version="1.3 alpha 6" />
<msInstrument>
  <dataProcessing centroided="1" />
  <software type="conversion" name="Thermo2mzXML" version="1" />
</dataProcessing>
<scan num="1" msLevel="1" peaksCount="780" polarity="+"
retentionTime="PT180.11S" lowMz="400" highMz="1800"
basePeakMz="1727.36" basePeakIntensity="6.75227e+007"
totIonCurrent="5.90564e+008">
  <peaks precision="32" byteOrder="network" pairOrder="m/z-
int">Q8hAfEKweKBDyMyyR7tyAEPJpA5JiZHIQ8o9hkkUZKBDy8sSPZTQEP
LY...
```

[Source: sashimi.sourceforge.net]

The mzXML standard is an XML based format for the storage of mass spectrometry data. This open format is designed to be easy to implement, feature rich, and because it is based on XML, mzXML can be extended to suit future needs. The Mass Spectrometry I/O Project and its related tools use mzXML for the storage of mass spectrometry data. Above is a small sample of mass spectrometry data in the mzXML format. The connections between the mxXML schema and the MSIOP are shown using arrows.

Mass Spectrometry

Mass spectrometers are used to determine the amino acid sequence that makes up a sample protein. This sample could, for example, be the result of research to determine the effectiveness of an experimental drug. The sample is placed in the mass spectrometer's inlet, and from there it is ionized. Next the sample passes through the mass analyzer and then the detector, where the ions that comprise the sample are measured. The data system of the mass spectrometer records the presence of the ions, and this information is output for later analysis. A mass spectrometers consist 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. Although all of the seven major components of the mass spectrometer are important, the ionization method and mass analyzer tend to determine the major attributes of a mass spectrometer, and are detailed below.



A vacuum is maintained inside the mass spectrometer

The data is collected and stored for analysis

A Sample enters the mass spectrometer through the inlet

The control center for the mass spectrometer

Ions that comprise the sample are measured

Ion Source:

Electrospray ionization involves placing a sample in a solution under conditions such that it ionizes while in the solution and then spraying that solution through the opening of a needle into a chamber with a large electric field. The charged sample is drawn out of solution and into the mass spectrometer where it is analyzed.

Matrix-assisted laser desorption/ionization (MALDI), is where the sample is dissolved in an ultra-violet absorbing solution (called the "matrix") and placed on a probe or stage where it dries and crystallizes. Once this solution has dried and crystallized, it is exposed to UV laser light and is vaporized, and then ionized by the acidic nature of the matrix compound, along with the addition of acid to the sample.

Mass Analyzers:

Quadrupole Mass Filters use a set of four charged rods that are configured in such a way that only ions with the desired m/z will stay "on course" through the instrument; all ions with a different m/z will be deflected away.

Ion Trap mass analyzers also use charged rods (shaped as rings), but instead of deflecting away the undesired ions, they initially trap all the ions that enter, and over time a RF-voltage is applied to the trapped ions, allowing the ions to exit the trap in order of increasing m/z and to reach the detector. This approach results in an instrument that can be very sensitive and selective.

Time-of-Flight Mass Analyzers (TOF) apply electric fields to the ions that comprise a sample, and depending on their m/z, the ions will spend a discrete amount of time "in flight", allowing the m/z of the ions to be determined. TOF mass analyzers have the added benefit of having virtually no upper limit on m/z, unlike the Quadrupole and Ion Trap methods.

Development Framework and Toolkit

The Mass Spectrometry I/O Project uses the Cactus framework, a high performance computing environment designed for use across multiple architectures. This framework allows the code written for the Mass Spectrometry I/O Project to compile and run on the heterogeneous mixture of hardware and software that comprise the Illinois Bio-Grid.

[Source: www.cactuscode.org]

The Globus toolkit is the foundation upon which the Illinois Bio-Grid is built. Globus provides the underlying architecture and security, along with implementing resource management. Through the use of the Globus toolkit, the Illinois Bio-Grid offers tremendous computational capability that can be securely accessed by its participants.

[Source: www.globus.org]