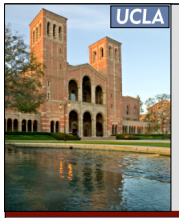

New Developments in ProteoWizard

Darren Kessner[1], Matt Chambers[2], Robert Burke[1], Kate Hoff[1], Brian Pratt[3], Natalie Tasman[4], Parag Mallick[1]

[1]Spielberg Family Center for Applied Proteomics & UCLA & USC, Los Angeles, CA

[2]Vanderbilt University, Nashville, TN [3]Insilicos Software, Seattle, WA [4]Institute for Systems Biology, Seattle, WA

USC UNIVERSITY OF SOUTHERN CALIFORNIA

Abstract

ProteoWizard is a modular and extensible set of open-source, cross-platform tools and libraries for proteomics data analysis.

The libraries enable rapid tool creation by providing a robust, pluggable development framework that simplifies and unifies data file access, and performs standard chemistry and LCMS dataset computations.

Since its initial release a year ago, the project has undergone significant development, with major contributions from the Spielberg Family Center for Applied Proteomics at Cedars-Sinai Medical Center, the Vanderbilt University Mass Spectrometry Research Center, Insilicos Software, and the Seattle Proteome Center at the Institute for Systems Biology.

The software is freely available under the Apache open source license.

Features

- full support for the HUPO-PSI mzML 1.1 standard mass spectrometry data format
- uses modern C++ techniques and design principles
- cross-platform with native compilers (gcc on Linux, Visual C++ on Windows, XCode on OSX)
- has modular design, for testability and extensibility
- facilitates rapid development of data analysis tools
- open source license suitable for both academic and commercial projects (Apache v2)

Data Formats

ProteoWizard provides a cross-platform data file abstraction layer that allows the scientific programmer to focus on the actual data analysis, without having to deal with specific details about the source file format or the operating system.

ProteoWizard currently supports the following data formats:

- mzML 1.1
- mzXML
- MGF
- Thermo RAW *
- Waters RAW *
- Bruker FID/YEP/BAF *

* *Vendor formats supported with vendor-supplied DLLs on Windows only*

mzML 1.1

The Institute for Systems Biology (ISB) and the Proteomics Standards Initiative (PSI) have collaboratively designed and implemented a new single format, mzML, taking from the best aspects and requirements of both original formats. The format is implemented in XML using schema definitions in conjunction with a controlled vocabulary implemented within the Open Biological Ontologies framework. Semantic validation, in addition to standard XML validation, is achieved via a special mzML semantic validator tool.

Version 1.0 of mzML was released at ASMS in June 2008. The format has undergone much internal and external review and we are ready to release a significant update to the schema: version 1.1. Several software tools are available to utilize the new format, including writers, readers, and validators (e.g. ProteoWizard, TPP, TOPP, jmzML, Phenyx, Pride, etc.). Details can be obtained on the HUPO-PSI website.

The ProteoWizard development team has been active in the development of the mzML standard, producing the reference implementation and first example files.

Data Layer

ProteoWizard uses an internal data model that is a one-to-one translation from mzML data elements to C++ data structures.

The data layer has the following features:

- a virtual interface for accessing spectrum lists, to allow lazy evaluation when accessing the spectra contained in a data file
- a plug-in Reader interface to allow reading of vendor proprietary data formats
- built-in diff calculation, for comparison of two data files, useful for validation after data format conversion or preprocessing
- iostream serialization to/from mzML and mzXML

ProteoWizard provides simple conversion functionality from supported data formats to mzML, mzXML, and MGF.

```
<fileDescription>
  <fileContent>
    <cvParam cvLabel="MS" accession="MS:1000580" name="MSn spectrum" value=""/>
  </fileContent>
  <sourceFileList count="1">
    <sourceFile id="rawfile" name="data01.RAW" location="c:/data/raw">
      <cvParam cvLabel="MS" accession="MS:1000563" name="Xcalibur RAW file" value=""/>
      <cvParam cvLabel="MS" accession="MS:1000569" name="SHA-1 "value="6bde97 [...]"/>
    </sourceFile>
  </sourceFileList>
</fileDescription>
```

Figure 1a. mzML fragment

```
struct SourceFile : public ParamContainer
{
    string id;
    string name;
    string location;
};
```

```
struct FileDescription
{
    FileContent fileContent;
    vector<SourceFilePtr> sourceFilePtrs;
    vector<Contact> contacts;
};
```

Figure 1b. Corresponding ProteoWizard data structures

Architecture

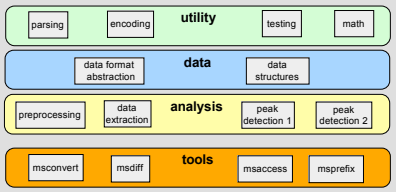

Figure 2. High level architecture of ProteoWizard

ProteoWizard is built from many independent libraries, grouped together in dependency levels. Each library is independently testable, and depends only on libraries in lower levels of the hierarchy. The utility layer contains independent classes that perform computations applicable in a wide variety of situations. The data layer abstracts the source data file, hiding any format-specific details. The analysis layer contains all scientific computation, in reusable modules. The tools layer code is responsible only for regulating interaction between the user and the analysis modules.

SeeMS

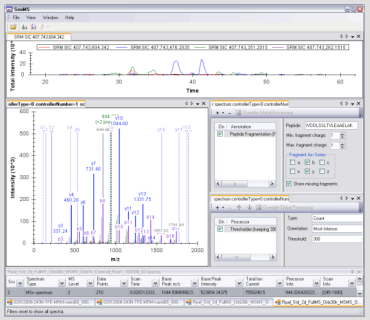

Figure 3. Screenshot of SeeMS

SeeMS is a Windows .NET application written by Matt Chambers at Vanderbilt University. SeeMS is a general-purpose spectrum and chromatogram visualization utility. It reads all MS formats that ProteoWizard supports and works natively with the mzML data model. It allows the user to add and arrange data processing layers (charge state calculation, intensity thresholding, etc.) onto individual spectra or an entire file. It also supports annotation of spectra with a peptide fragmentation model and will soon support other annotations types (like peptide mass fingerprinting). The interface allows a heavy level of customization of the window layout, including nested docking, tabbing, floating, and auto-hiding panels. SeeMS is free for non-commercial use.

Tools

msConvert: data format conversion from vendor proprietary formats to mzML, mzXML, and MGF.

msDiff: comparison of two data files, for validation of conversion and preprocessing

msAccess: command line access to mass spec data files, including spectrum binary data and metadata, selected ion chromatograms, and pseudo-2D gel image creation.

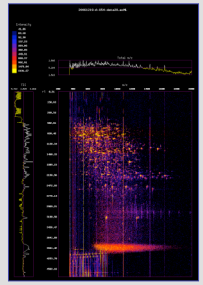

Figure 4a. Pseudo 2D gel image generated by *msPicture*

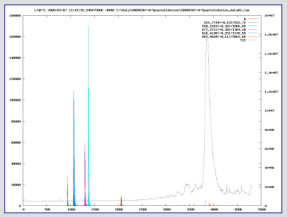

Figure 4b. gnuplot image generated by script from data extracted by *msaccess*

Current Status

- The Trans-Proteomic Pipeline tools from the Institute for Systems Biology use ProteoWizard for mzML support.
- The Insilicos Viewer from Insilicos Software uses ProteoWizard for mzML support.
- ProteoWizard is being actively used and extended by an international community of developers.

More Information

Darren Kessner
darren@proteowizard.org

ProteoWizard
http://proteowizard.sourceforge.net

Spielberg Family Center for Applied Proteomics
http://sfcap.cshs.org

HUPO Proteomic Standards Initiative
http://www.psidev.info