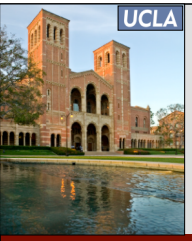


New Developments in ProteoWizard

Darren Kessner[1], Matt Chambers[2], Brendan MacLean[3], Robert Burke[1], Kate Hoff[1], Brian Pratt[4], Natalie Tasman[5], Parag Mallick[1]

[1]USC and UCLA, Los Angeles, CA [2]Vanderbilt University, Nashville, TN [3]University of Washington, Seattle, WA
[4]Insilicos Software, Seattle, WA [5]Institute for Systems Biology, Seattle, WA

USC UNIVERSITY OF SOUTHERN CALIFORNIA

Abstract

The ProteoWizard software project consists of:

1)a set of C++ libraries to facilitate cross-platform proteomics tools development, and
2)a set of cross-platform data analysis tools created using those libraries.

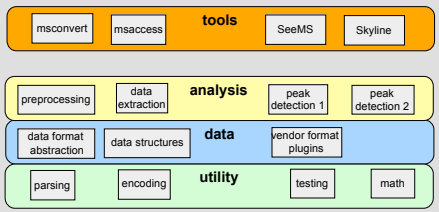
Since its initial release 18 months ago, the project has undergone significant development, with major contributions from the Center for Applied Molecular Medicine at University of Southern California, the Vanderbilt University Mass Spectrometry Research Center, the MacCoss Lab at University of Washington, Insilicos Software, LabKey Software, and the Institute for Systems Biology.

ProteoWizard is available for both commercial and non-commercial use, under the permissive Apache v2 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)

Architecture



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.

Open Data Format Standards

The ProteoWizard development team has been active in the development and implementation of HUPO Proteomics Standards Initaitive (HUPO-PSI) data format standards.

ProteoWizard provided the reference implementation and example files for the mzML format for storing raw mass spectrometry output data. mzML 1.1 was recently released, in June 2009.

ProteoWizard has also recently added an implementation of mzIdentML, the HUPO-PSI format for storing peptide/protein identification information.

HUPO-PSI: http://www.psidev.info

Data Layer

ProteoWizard uses a data abstraction layer to insulate the programmer from details of the input data format and compiler/platform.

A plug-in Reader interface allows reading of vendor proprietary data formats, as well as simple conversion to open data formats.

ProteoWizard includes CLI bindings, which allows usage in Windows .NET applications.

The ProteoWizard data layer is currently used by many tools, including:

- Trans-Proteomic Pipeline (TPP) tools from the Institute for Systems Biology: http://tools.proteomecenter.org
- Insilicos Viewer from Insilicos Software: http://insilicos.com

Data Formats

ProteoWizard currently supports the following data formats on all platforms (Linux, OSX, Windows):

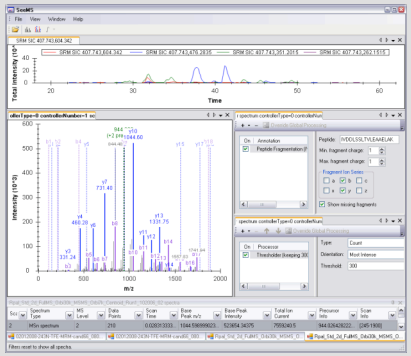
- mzML 1.1
- mzXML
- MGF
- mzIdentML

ProteoWizard also supports the following vendor proprietary formats on Windows, through the use of vendor software libraries:

- Agilent* (MassHunter .d)
- Applied Biosystems (WIFF)
- Bruker (Compass .d, FID, YEP, BAF)
- Thermo Fisher* (RAW)
- Waters (MassLynx .raw)

* *vendor has agreed to have their software libraries distributed as part of the ProteoWizard download*

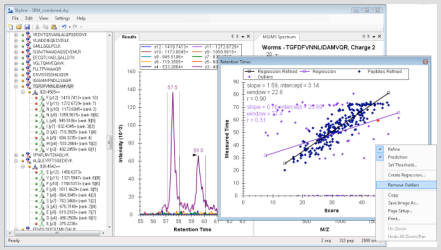
SeeMS



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.

http://www.mc.vanderbilt.edu/msrc

Skyline



Skyline is a Windows .NET application written by Brendan MacLean at the MacCoss Lab at Unviversity of Washington in Seattle. It is a tool for building Selected Reaction Monitoring (SRM) / Multiple Reaction Monitoring (MRM) methods and analyzing the resulting mass spectrometer data. It aims to employ cutting-edge technologies for creating and iteratively refining SRM methods for large-scale proteomics studies.

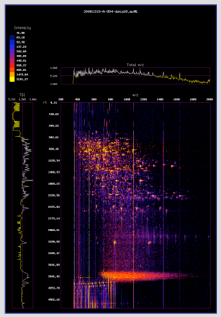
http://proteome.gs.washington.edu/software/skyline

Tools

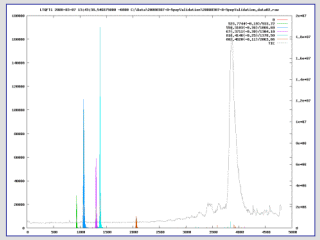
msconvert: data format conversion from vendor proprietary formats to mzML, mzXML, and MGF.

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

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



Pseudo 2D gel image generated by *mspicture*



gnuplot image generated by script from data extracted by *msaccess*

Testing

ProteoWizard was designed to be testable, and unit tests for each code module are an integral part of the project.

LabKey Software provides continuous integration servers and automatic builds for the ProteoWizard project.

http://www.labkey.com

More Information

Darren Kessner
darren@proteowizard.org

ProteoWizard
http://proteowizard.sourceforge.net

Center for Applied Molecular Medicine
http://camm.usc.edu