HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2013 April 10.

Published in final edited form as:

Nat Biotechnol. 2012 October ; 30(10): 918–920. doi:10.1038/nbt.2377.

A Cross-platform Toolkit for Mass Spectrometry and Proteomics

Matthew C. Chambers[1,*], **Brendan Maclean**[2,*], **Robert Burke**[3,*], **Dario Amodei**[16], **Daniel L Ruderman**[3], **Steffen Neumann**[6], **Laurent Gatto**[7], **Bernd Fischer**[15], **Brian Pratt**[4], **Jarrett Egertson**[2], **Katherine Hoff**[3], **Darren Kessner**[4], **Natalie Tasman**[4], **Nicholas Shulman**[2], **Barbara Frewen**[2], **Tahmina A. Baker**[2], **Mi-Youn Brusniak**[14], **Christopher Paulse**[14], **David Creasy**[5], **Lisa Flashner**[3], **Kian Kani**[3], **Chris Moulding**[3b], **Sean L. Seymour**[8], **Lydia M. Nuwaysir**[8], **Brent Lefebvre**[8], **Frank Kuhlmann**[9], **Joe Roark**[9], **Paape Rainer**[10], **Suckau Detlev**[10], **Tina Hemenway**[11], **Andreas Huhmer**[11], **James Langridge**[12], **Brian Connolly**[13], **Trey Chadick**[13], **Krisztina Holly**[3b], **Josh Eckels**[13], **Eric W. Deutsch**[14], **Robert L Moritz**[14], **Jonathan E. Katz**[3], **David B. Agus**[3], **Michael MacCoss**[2], **David L. Tabb**[1], and **Parag Mallick**[3,16]

[1] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37212-8575, USA [2] Department of Genome Sciences, University of Washington, Seattle, WA 98195 [3] Center for Applied Molecular Medicine, University of Southern California, Los Angeles, CA 90033, USA [3b] USC Stevens Institute for Innovation, University of Southern California, Los Angeles, CA 90089, USA [4] Insilicos, Seattle WA 98109 [5] Matrix Science, Boston, MA 02110 [6] Department of Stress & Developmental Biology, Leibniz Institute for Plant Biochemistry, Halle (Saale), Germany [7] Proteomics Services, Cambridge Centre for Proteomics, Cambridge, England [8] AB SCIEX, Foster City, CA 94404 [8b] AB SCIEX, Concord, Ontario, Canada [9] Agilent Technologies, Santa Clara, California 95051 [10] Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany [11] Thermo Fisher Scientific, San Jose, CA 95134 [12] Waters Corporation, Manchester, UK, M22 5PP [13] LabKey Software, Seattle, WA 98102 [14] Institute for Systems Biology, Seattle, WA 98109 [15] Genome Biology, EMBL Heidelberg, Germany [16] Canary Center for Cancer Early Detection, Stanford University, Stanford, CA 94024

Abstract

Mass-spectrometry-based proteomics has become an important component of biological research. Numerous proteomics methods have been developed to identify and quantify the proteins in biological and clinical samples[1], identify pathways affected by endogenous and exogenous perturbations[2], and characterize protein complexes[3]. Despite successes, the interpretation of vast proteomics datasets remains a challenge. There have been several calls for improvements and standardization of proteomics data analysis frameworks, as well as for an application-programming interface for proteomics data access[4,5]. In response, we have developed the ProteoWizard Toolkit, a robust set of open-source, software libraries and applications designed to

Corresponding Author: Parag Mallick, Ph.D. Stanford University 1501 S California Avenue, Room 2212 Palo Alto, CA 94304 paragm@stanford.edu.

*These authors contributed equally

facilitate proteomics research. The libraries implement the first-ever, non-commercial, unified data access interface for proteomics, bridging field-standard open formats and all common vendor formats. In addition, diverse software classes enable rapid development of vendor-agnostic proteomics software. Additionally, ProteoWizard projects and applications, building upon the core libraries, are becoming standard tools for enabling significant proteomics inquiries.

Historically, the development of proteomics software tools has been hindered by three factors: 1) the numerous file formats, ranging from vendor-specific mass spectrometry data formats to software application-specific formats, used for processing mass spectrometry data and storing analysis results; 2) the time-consuming and error-prone development of code implementing common, but critical algorithms, such as protein digestion, mass computation, peak integration, charge state detection, and isotope deconvolution; and 3) the complexity of comparing and validating analysis algorithms. Together, these three impediments create a significant bottleneck in the development of new proteomics software applications. Beyond slowing the pace of proteomics software development, these impediments have also hampered the field of proteomics by interfering in the meaningful comparison, sharing, and exchange of data analyses obtained on different platforms or by different laboratories.

Efforts to mitigate these issues led initially to the development of several ‘open’ interchange formats[6, 7] and a series of software tools that extracted data from vendor formats into open formats. The majority of MS vendors also now provide approaches to export their data to open formats. Though an important step forward, both the academic and commercial tools suffer from a few limitations. For example, despite extensive conversion tools, a robust code-base that allowed developers to easily extract data from datafiles for use in their own applications did not exist. Efforts by our group and by the OpenMS team attempted to address this issue[8, 9]. In addition, early converters depended upon instrument control software libraries; consequently, users without instruments could neither access nor convert vendor datafiles. Furthermore, each vendor format had its own converter (e.g., MassWolf for Waters Files and ReAdW for Thermo Fisher files) thus complicating software maintenance. Lastly, despite the amazing success of these open formats and the proliferation of tools that use them, the converter-centric, common-format approach did not address the issue of direct access to primary raw data. Most native vendor formats encode valuable, but vendor-specific, meta-data including details of instrument settings and instrument readouts.

Direct access to raw, primary data can critically affect the comparability of experimental platforms because common computational processing steps associated with export, such as centroiding, may impact benchmarking results. The comparison challenge is even more significant for data analysis approaches; a bioinformatics approach could easily appear inferior because of unintended (possibly error-derived) upstream data processing steps. Lastly, cross-platform comparison of workflows (both computational and experimental) is hampered when tools are developed to read files from a particular vendor but cannot be applied in data from other instrument types. As the field of proteomics attempts to become more robust, the need for integrated pipelines for processing and analyzing complex proteomics data sets in a platform-agnostic manner has become critical.

With version 3.0 of the ProteoWizard Toolkit[8], we attempt to mitigate these challenges through open-source, permissively licensed, cross-platform software. The Toolkit has two components: 1) a suite of libraries that facilitate the development and comparison of tools for proteomics data analysis and 2) a set of tools, developed using these libraries, that perform a wide array of common proteomics analyses. The Toolkit has been developed under modern design principles in the C++ language and supports a variety of platforms with native compilers (GCC on Linux, MSVC on Windows, and XCode on OSX). The toolkit was released under the Apache 2.0 license[10] to ensure that it can be used in both academic and commercial projects. New to ProteoWizard 3.0 and unlike previous efforts, vendor reader libraries are now directly distributed with the Toolkit independently of instrument control libraries (a further description of new features can be found in Supplemental Text 1). Furthermore, ProteoWizard employs a single converter and access interface for all formats; this singular point of maintenance allows a more stable and optimized set of tools. Additional robustness comes from ProteoWizard's use of a continuous integration and testing environment. Though common in commercial projects, this scale of quality assurance is uncommon in traditional academic projects.

As shown in Figure 1A, ProteoWizard is built upon a modular framework of many independent libraries grouped in dependency levels. Each library only depends on libraries in lower levels of the hierarchy. The data layer provides a unified access interface to mass spectrometry data, independent of the format-specific details associated with a given source file. The underlying data model of the data layer directly translates HUPO-PSI data elements to C++ data structures. In Supplemental Text 2, we show this mapping for a piece of the msData module that implements mzML[11]; equivalent mappings exist for mzIdentML[12] and TraML[7].

Field-standard open formats (e.g., mzML, mzXML, MGF, pepXML, and mzIdentML) and vendor proprietary formats are handled with a plug-in reader interface (Figure 1B). In partnership with proteomics standards bodies and instrument and software vendors, we have developed a series of adapters that translate between input files and the core msData data structures to support a wide range of formats (see Supplementary Tables 1 and 2 for supported proprietary and open formats). These adapters bridge between vendor-provided libraries that read proprietary formats and the fully open ProteoWizard data layer. Through a series of generous licenses, the ProteoWizard Software Foundation has permission to distribute vendor-provided libraries from AB SCIEX, Agilent, Bruker, Thermo Fisher Scientific, and Waters with the ProteoWizard Toolkit. Consequently, bioinformatics developers are not required to have direct access to an instrument to develop software that can analyze data generated by it.

Furthermore, any application built upon the ProteoWizard framework is significantly format-agnostic for the dominant formats in the field. By writing their software using ProteoWizard's msData API, developers can focus on algorithmic challenges, rather than on the complex details of the wide array of formats prevalent throughout the field of proteomics Furthermore, the use of the ProteoWizard API has the potential to improve the robustness and reliability of other proteomics software efforts. As vendors frequently change their file formats to accommodate new instruments and public standards evolve rapidly, software

tools can rapidly become unusable unless significant resources are devoted to continually update data-reader code. The robust upkeep of ProteoWizard, in concert with its widespread use, will effectively reduce the investment that the public has to make in maintaining the longevity of open-source software.

Supplementary Example 1 illustrates how the mass spectral data from a mass spectrometer data file can be browsed and printed. Also highlighted in Supplementary Example 1 are the benefits of ProteoWizard's Common Language Infrastructure bindings, which allow the library to be accessed from diverse languages including C#, IronPython, and Visual Basic. Supplementary Example 2 illustrates how peptide and protein identification data can be browsed and printed. In Supplemental Example 3, we illustrate how the mzR library enables ProteoWizard-based data access within the R statistical analysis toolkit. Notably, mass spectrometry data can be used for a variety of applications other than proteomics investigation. The data layer does not impose any restrictions that inhibit its use for any mass-spectrometry-based problem. ProteoWizard is already used in metabolomics applications[13] and should find utility in analysis of glycomics data.

Below the data layer is the Utility Layer (Figure 1A). The Utility Layer contains applications that perform computations such as binary to text encoding, XML parsing, and mathematical calculations that are common in data analysis. A list of available utility classes is provided in Supplementary Table 3. Though the majority of computations available in these classes are straightforward, their implementation can be time consuming. By using ProteoWizard, developers are able to focus on developing novel algorithms rather than on redundant implementation of requisite parsing and data handling code, thus accelerating the development timeline.

The Analysis Layer further builds upon the data layer and provides common proteomics-centric analysis modules. A significant bottleneck in proteomics software development can arise from the time required to implement the vast array of standard operations routinely required of a proteomics algorithm such as computing the mass of a peptide (Supplementary Example 4) or performing an *in silico* digest of a protein read from a FASTA file (Supplementary Example 5). There are also independent modules for handling chemical formulas, peptide calculations, and isotope envelopes. All these computations are contained in reusable, platform-independent modules in the Analysis Layer. A list of available analysis classes is provided in Supplementary Table 3.

Additional analysis modules are currently in development with an emphasis on establishing standard interfaces for common proteomics computations such as peak picking, isotope deconvolution, and precursor estimation[14]. Our goal is to work collaboratively to create a modular analysis infrastructure in which experts will be able to contribute a module that can then be plugged into various software tools. This will allow, for example, an expert in signal processing to contribute a peak picker without having to handle details of file formats, operating systems, or command-line configurations. The ProteoWizard Toolkit also includes a number of small, useful applications, listed in Supplementary Table 4, that are built upon the libraries. These applications support data conversion (msConvert, msConvertGUI,

idConvert), data visualization (msPicture, seeMS), data access (msAccess, msCat, idCat, msPicture), data analysis (peekaboo, msPrefix[14]), and basic proteomics utilities (chainsaw).

Beyond the ProteoWizard Toolkit, the ProteoWizard Software Foundation has built several Projects on top of the ProteoWizard Toolkit that provide useful end-user applications. The most widely known example, Skyline[15], is becoming the standard tool for targeted proteomics investigation. A second project, Topograph, is focused on measuring protein turnover in metabolic labeling time-course experiments. Other projects are underway. To be included in ProteoWizard, projects must demonstrate broad applicability within the field and active ownership within the contributing organization. They must also adopt non-restrictive licensing[1] and continue to develop new features in open source. Project contributors must provide thorough automated testing and participate in the ProteoWizard build and continuous integration processes.

The ProteoWizard Toolkit and Projects attempt to provide useful analytic tools to the proteomics community while simplifying the process of software development and bioinformatics for mass spectrometry and proteomics. Our hope is that a standardized toolkit will enable rigorous development and assessment of diverse computational approaches to significantly accelerate proteomics research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by the Wunderkinder Foundation, Redstone Family Foundation, and National Cancer Institute and National Institutes of Health grants and contracts P41 RR011823, CCNE-TR 5U54CA119367, CCNE-T 1U54CA151459, PSOC-MCSTART 5U54CA143907, R01CA126218 and U24CA126479. RM and CP have been supported by National Science Foundation MRI grant No. 0923536. LG has been supported by the European Union 7th Framework Program PRIME-XS project, grant agreement number 262067. The authors also thank the ProteoWizard developer community and TPP developer community for their contributions to the project and manuscript.

Citations

1. Schwanhausser B, et al. Nature. 2011; 473:337–342. [PubMed: 21593866]
2. Raj L, et al. Nature. 2011; 475:231–234. [PubMed: 21753854]
3. Bouwmeester T, et al. Nat Cell Biol. 2004; 6:97–105. [PubMed: 14743216]
4. Patterson SD. Nature biotechnology. 2003; 21:221–222.
5. Askenazi M, Parikh JR, Marto JA. Nat Methods. 2009; 6:240–241. [PubMed: 19333238]
6. Pedrioli PG, et al. Nature biotechnology. 2004; 22:1459–1466.
7. Orchard S, et al. Proteomics. 2010; 10:1895–1898. [PubMed: 20623474]
8. Kessner D, Chambers M, Burke R, Agus D, Mallick P. Bioinformatics. 2008; 24:2534–2536. [PubMed: 18606607]
9. Sturm M, et al. BMC bioinformatics. 2008; 9:163. [PubMed: 18366760]
10. Apache.Software.Foundation. 2004. http://www.apache.org/licenses/LICENSE-2.0.html
11. Martens L, et al. Mol Cell Proteomics. 2011; 10:R110. 000133. [PubMed: 20716697]

[1]http://www.apache.org/licenses/LICENSE-2.0.html

12. Eisenacher M. Methods Mol Biol. 2011; 696:161–177. [PubMed: 21063947]
13. Benton HP, Wong DM, Trauger SA, Siuzdak G. Anal Chem. 2008; 80:6382–6389. [PubMed: 18627180]
14. Luethy R, et al. J Proteome Res. 2008; 7:4031–4039. [PubMed: 18707148]
15. MacLean B, et al. Bioinformatics. 2010; 26:966–968. [PubMed: 20147306]

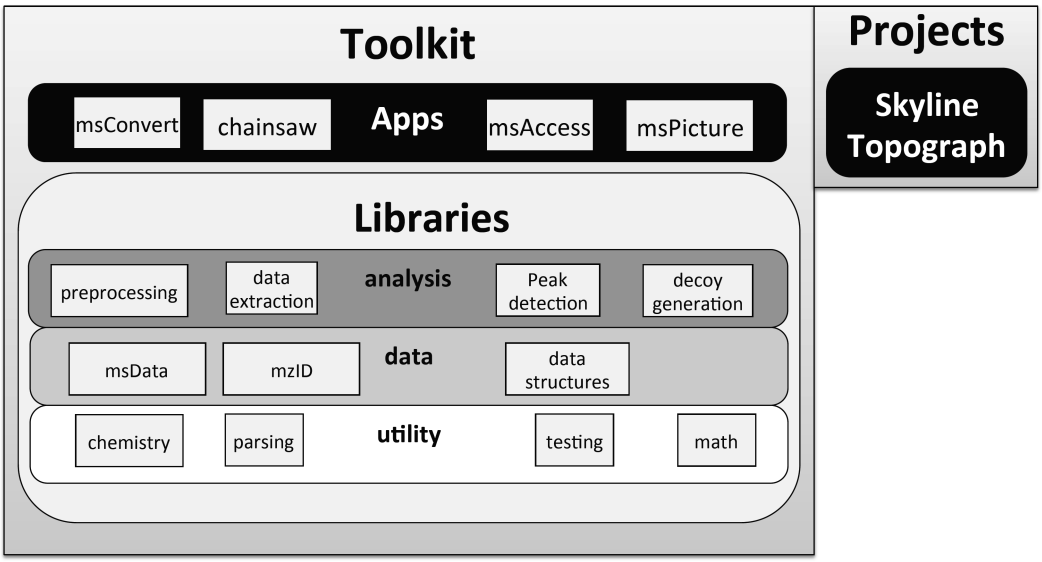

Figure 1a.

ProteoWizard uses modern design principles to implement a modular framework of many independent libraries grouped in dependency levels with strict interfaces. This allows extensive development at each level while enforcing stability.

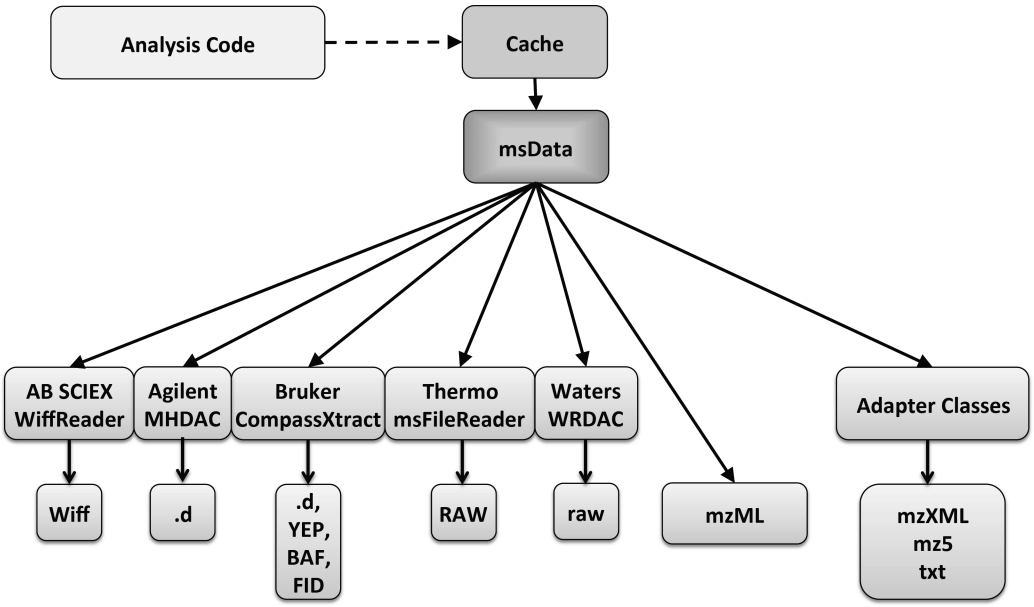

Figure 1b.

The data layer presents a unified access interface to mass spectrometry data. The modular framework allows additional readers for diverse file-types to be easily added via plug-in adapter classes. Developers only need interact with the primary interface to access data, agnostic to the details of an input source file.