DATA AND TEXT MINING

PDV: an integrative proteomics data viewer

Kai Li[1], Marc Vaudel[2,3], Bing Zhang[4,5], Yan Ren[1*], Bo Wen[4,5*]

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway. [3]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway. [4]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.

ABSTRACT

Summary: Data visualization plays critical roles in proteomics studies, ranging from quality control of raw data to validation of peptide identification results. Herein, we present PDV, an integrative proteomics data viewer that can be used to visualize a wide range of proteomics data, including database search results, *de novo* sequencing results, proteogenomics files, MS/MS raw data, and data from public proteomics repositories. PDV is a lightweight visualization tool that enables intuitive and fast exploration of diverse, large-scale proteomics datasets on standard desktop computers in both graphical user interface and command line modes.

Availability: PDV software and the user manual are freely available at http://pdv.zhang-lab.org. The source code is available at https://github.com/wenbostar/PDV and is released under the GPL-3 license.

Contact: bo.wen@bcm.edu; reny@genomics.cn.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS) is the leading technique in proteomics, and it generates large volumes of data. Analyzing the raw data requires many different tools that produce additional data in various formats. These formats include protein identification results from database searching or *de novo* sequencing, and proteomic data from public databases such as PRIDE (Jones, et al., 2006) or spectral libraries such as PeptideAtlas (Desiere, et al., 2006). Visualization at each stage of the analysis is critical in exploiting and understanding the information derived from these data (Oveland, et al., 2015).

Currently, several tools are available to visualize proteomics data, such as TOPPView (Sturm and Kohlbacher, 2009), PRIDE Inspector (Wang, et al., 2012), MS-Viewer (Baker and Chalkley, 2014), and BatMass (Avtonomov, et al., 2016). TOPPView provides a graphical user interface (GUI) for visualizing raw MS/MS data. PRIDE Inspector is a Java-based tool for visualizing protein identification results in PRIDE XML or mzIdentML formats. MS-Viewer is a web-based tool for visualizing protein identification results from database searching. However, it cannot handle de novo sequencing results or raw MS/MS data. BatMass is a recently published tool for visualizing both raw proteomics and metabolomics LC-MS data. All of these tools are designed for the visualization of a specific data type. However, a proteomics study commonly requires visualizing data at different stages and in multiple formats, requiring researchers to use multiple tools. Furthermore, most tools are only available as a GUI and do not feature a batch mode. This makes the efficient generation of high quality figures for large scale datasets and implementation in other pipelines difficult.

Here, we propose PDV, a standalone software tool programmed in Java that can be used to visualize different kinds of proteomics data in both GUI and command line modes. The functions of PDV include visualization of raw MS/MS data, database search results, *de novo* sequencing results, proteogenomics files, and data from public databases such as PRIDE and PeptideAtlas. In addition, PDV provides a function to generate various quality control figures and tables for assessment of proteomic data quality.

2 FEATURES AND IMPLEMENTATION

PDV is platform independent and written in Java 1.8. It can be run in GUI or command-line mode. The functions of PDV can be divided into the following modules as illustrated in Figure 1.

2.1 MS/MS raw data visualization module

PDV supports input of raw MS/MS data in mzML or mzXML format for visualization. It supports visualizing multiple files at once. PDV displays a table that includes meta information about the MS/MS data and a panel with the total ion current (TIC) chromatogram enabling the assessment of the chromatographic performance by visualizing the intensity distribution over the retention time (Figure S1). The MSDK library (https://github.com/msdk/msdk) is used to quickly extract TIC data from mzML or mzXML files.

2.2 Database searching result visualization module

PDV accepts identification result files in the mzIdentML standard format (Jones, et al., 2012), the pepXML format, or a tab-delimited text format. The tab-delimited text file requires three columns: spectrum index, peptide sequence, charge state and modification information. In order to generate a spectrum annotation figure, an MS/MS data file in MGF, mzXML, or mzML format is also needed. The identification result format from the widely used MaxQuant (Cox and Mann, 2008) software is also supported. In order to quickly visualize the results, multithreading technology and SQL database technology are used. If a user has an MS/MS spectrum from a synthetic peptide which is also identified by a spectrum in the input identification file, PDV provides a function to support direct visual comparison of the two spectra matching to the same peptide (Figure S2). The spectrum visualization and annotation is mainly based on the compomics-utilities library (Barsnes, et al., 2011).

2.3 De novo sequencing result visualization module

De novo sequencing is a popular technique in proteomics for identifying peptides from tandem mass spectra without relying on a protein sequence database. PDV can import the identification results from pNovo+ (Chi, et al., 2013), Novor (Ma, 2015) and DeepNovo (Tran, et al., 2017). The basic module of this function is developed based on DeNovoGUI (Muth, et al., 2014). The result visualization panel includes a table presenting the result for each

*To whom correspondence should be addressed.

spectrum and a panel to present the spectrum annotation figure (Figure S3).

2.4 Single PSM visualization module

In order to visualize a single PSM, PDV provides a separate panel allowing users to input a peptide sequence and an MS/MS spectrum in the MGF format (Figure S4). In this panel, users can manually set modifications that occurred in the peptide sequence. All modifications from Unimod are available (Figure S5). In addition, the user can tune the spectrum annotation.

2.5 Proteogenomics data visualization module

Two new standard formats, proBAM and proBED, were recently released to facilitate the integration of genomics, transcriptomics, and protoemics data in proteogenomics studies (Menschaert, et al., 2018). PDV supports input of proBAM and proBED files thanks to the Htsjdk library (https://github.com/samtools/htsjdk) (Figure S6).

We have developed a fast and easy-to-use software named PDV for visualization of different kinds of proteomics data. The PDV GUI enables users lacking programing experience to visualize proteomics data interactively while the command line interface allows users to generate annotated spectra or TIC figures in batch mode for large-scale datasets. Furthermore, the use of multi-threading and SQL database technologies in PDV enables efficient processing of large datasets. For example, loading a 2 GB mzIdentML file takes less than 15 seconds and loading 20 raw files (20GB) takes about 30 seconds with a Windows computer (Intel Core i3-7100, 8 GB of RAM and 256 GB of SanDisk X400 SSD). We anticipate that researchers from the proteomics community will benefit from PDV for the interpretation and validation of proteomics data.

Conflict of Interest: none declared.

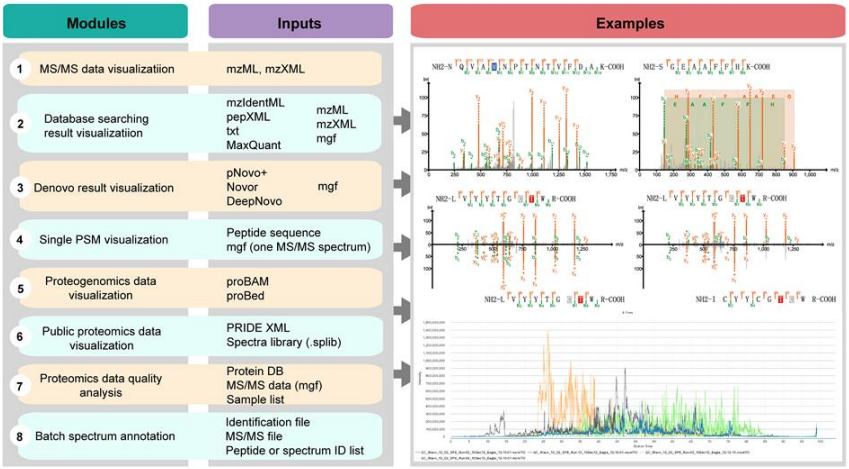

Figure 1. An overview of PDV.

2.6 Public proteomics data visualization module

Proteomics data from the PRIDE repository in the PRIDE XML format and PeptideAtlas in splib format are supported. The visualization panel is similar to the database searching visualization panel (Figure S7).

2.7 Data quality analysis module

PDV provides a GUI utilizing proteoQC (bioconductor.org/packages/proteoQC) for proteomics data quality assessment (Figure S8). It accepts MS/MS data (MGF, mzML or mzXML) and a protein database and generates an HTML-based report that includes identification-free (ID-free) metrics and identification-based (ID-based) metrics within a single experiment, as well as across multiple experiments.

2.8 Batch spectrum annotation module

PDV provides a command line module to produce figures of annotated spectra or TIC in batch mode. It can be used to generate figures according to a list of peptide sequences or a list of spectrum indexes. This function is especially useful when requiring the generation of a large number of high quality figures for publication (Wang, et al., 2017).

3 RESULTS AND DISCUSSION

REFERENCES

Avtonomov, D.M., Raskind, A. and Nesvizhskii, A.I. BatMass: a Java Software Platform for LC-MS Data Visualization in Proteomics and Metabolomics. *J Proteome Res* 2016;15(8):2500-2509.

Baker, P.R. and Chalkley, R.J. MS-viewer: a web-based spectral viewer for proteomics results. *Mol Cell Proteomics* 2014;13(5):1392-1396.

Barsnes, H., *et al.* compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* 2011;12:70.

Chi, H., *et al.* pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J Proteome Res* 2013;12(2):615-625.

Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 2008;26(12):1367-1372.

Desiere, F., *et al.* The PeptideAtlas project. *Nucleic Acids Res* 2006;34(Database issue):D655-658.

Jones, A.R., *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* 2012;11(7):M111 014381.

Jones, P., *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* 2006;34(Database issue):D659-663.

Ma, B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* 2015;26(11):1885-1894.

Menschaert, G., *et al.* The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data. *Genome Biol* 2018;19(1):12.

Muth, T., *et al.* DeNovoGUI: An Open Source Graphical User Interface for de Novo Sequencing of Tandem Mass Spectra. *Journal of Proteome Research* 2014;13(2):1143-1146.

Oveland, E., *et al.* Viewing the proteome: How to visualize proteomics data? *Proteomics* 2015;15(8):1341-1355.

Sturm, M. and Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *J Proteome Res* 2009;8(7):3760-3763.

Tran, N.H., *et al.* De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences* 2017;114(31):8247-8252.

Wang, R., *et al.* PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* 2012;30(2):135-137.

Wang, X., *et al.* Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Mol Cell Proteomics* 2017.