Published in final edited form as:
Nat Biotechnol. ; 30(2): 135–137. doi:10.1038/nbt.2112.

PRIDE Inspector: a tool to visualize and validate MS proteomics data

Rui Wang[1], **Antonio Fabregat**[1], **Daniel Ríos**[1], **David Ovelleiro**[1], **Joseph M. Foster**[1], **Richard G. Côté**[1], **Johannes Griss**[1,2], **Attila Csordas**[1], **Yasset Perez-Riverol**[1,3], **Florian Reisinger**[1], **Henning Hermjakob**[1], **Lennart Martens**[4,5], and **Juan Antonio Vizcaíno**[1,*]

[1]EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

[2]Department of Medicine I, Medical University of Vienna, Borschkegasse 8a, 1090 Vienna, Austria.

[3]Department of Proteomics, Center for Genetic Engineering and Biotechnology, Ave 31 e/158 & 190, Cubanacán, Playa, Ciudad de la Habana, Cuba.

[4]Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium.

[5]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium.

To the Editor:

Your editorial "Credit where credit is overdue"[1] aptly summarized the existing situation in the proteomics field, where full data disclosure remains very much a work in progress. Importantly, it also correctly pointed out that 'the software provided by the public repositories for searching and analysing proteomics data is not as efficient and as user friendly as it could be'. We therefore here introduce PRIDE Inspector (http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector), a user-friendly, freely available open source software tool that allows the user to efficiently browse and visualize mass spectrometry (MS) proteomics data. One of the key features of PRIDE Inspector is that it allows the user to perform an initial assessment on data quality and reliability. PRIDE Inspector can thus be used by researchers before their data submission is performed, by journal editors and peer reviewers during the manuscript review process, and by any interested user in the field after public release of the data in PRIDE (Figure 1).

Despite the increasing popularity of MS based proteomics, and the overall tendency in the life sciences towards open sharing of biological data, relatively little proteomics data is currently available in the public domain. This situation is however changing thanks to stricter data sharing guidelines by scientific journals and funding agencies. Some proteomics journals (for instance, *Proteomics* and *Molecular and Cellular Proteomics*, MCP) recommend, and in some concrete cases mandate public deposition of MS data in support of manuscripts. Journals from the *Nature* group also strongly recommend submission of

*Corresponding author. .

Author contributions RW did most of the programming of the core components and the GUI. AF was mainly responsible for the chart component. DR was the main developer behind the access component of the PRIDE MySQL instance. DO, JMF, RGC, JG, AC, YPR and FR contributed to multiple areas during the development of the tool and also participated in the writing of the documentation and testing process. LM had the original idea and started the project. HH and JAV supervised the whole process. JAV and LM wrote the manuscript.

The authors have no competing financial or commercial interests.

All authors have agreed to all the content in the manuscript, including the data as presented.

Europe PMC Funders Author Manuscripts

proteomics data to repositories like PRIDE[2], PeptideAtlas[3] and Tranche[4] (http://www.nature.com/authors/policies/availability.html).

Nevertheless, in practical terms, this public data-sharing policy can only succeed if reliable and user-friendly software tools exist to streamline the submission task. Therefore, the PRIDE Converter[5] application (http://code.google.com/p/pride-converter) was developed for data submissions to the PRIDE database[2]. Not only has PRIDE Converter rapidly become the most popular data submission path for PRIDE (accounting for 77% of all PRIDE experiments submitted since January 2009), its release also corresponded to the start of a very significant increase in the amount of deposited data in PRIDE (Supplementary Figure 1). Of course, the availability of data in public repositories is only a first step. The interpretation and validation of proteomics data remains controversial, especially for cases where proteins have been identified on the basis of one unique peptide-to-spectrum match, or if post-translational modifications (PTMs) are reported. The ability to inspect and validate reported results during the review process, as well as after publication, is therefore of paramount importance. Because of the amount of data involved, such inspections can only be undertaken efficiently with the help of suitable software tools that combine ease of access with effective visualizations. While viewers for MS proteomics data are already available[6, 7], they tend to suffer from different types of limitations. They may have been developed around a single proprietary and/or unique data format, fail to properly handle the very large files that are routinely produced, have only limited visualization and analysis functionality or be costly to license for smaller groups or individuals. We therefore developed PRIDE Inspector as a very user friendly, freely available tool to browse, inspect and analyse proteomics data from the PRIDE repository, or presented in standard formats.

PRIDE Inspector is a stand-alone Graphical User Interface (GUI) written in Java, released under the Apache2 open source license, which can be freely downloaded from http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector. Furthermore, PRIDE Inspector can also be started through a direct web link from the PRIDE homepage (http://www.ebi.ac.uk/pride). The main features of PRIDE Inspector are listed in the Supplementary Information, along with a description of its overall software architecture and other technical details.

PRIDE Inspector supports fast loading of PRIDE XML and mzML[8] (the community data standard for MS data) files, and provides direct access to all public PRIDE data through a direct MySQL database connection. Moreover, it includes an automated data download capability for private PRIDE experiments that allows journal editors and peer reviewers with the correct login credentials to assess the relevant experiment(s) during peer review. The Web Start version available at the PRIDE homepage furthermore adds the ability to start the application and access a particular dataset through a simple URL.

PRIDE Inspector presents different views to the users, each focusing on a specific aspect of the data (Figure 2). Depending on the type of information available for a file format or PRIDE dataset, some views can remain inactive (Supp. Figure 2). For that reason, an ‘Experiment Summary’ overview window is available in the bottom left part of the GUI. A context-sensitive ‘Help’ function is also included, providing tailored documentation for the current view. Currently, there are six views available in PRIDE Inspector. Firstly, the ‘Overview’ tab, which includes easily readable, uniform experimental metadata. The precise information displayed can vary slightly depending on the file format used and is split in three different views: ‘Experiment General’, ‘Sample and Protocol’, and ‘Instrument and Processing’ (Supp. Figures 3-6). The second view concerns proteins (Supp. Figures 7-8), and is possibly the most interesting view for biologists. For each identified protein, peptides, PTMs and corresponding spectra are displayed in a concise manner. Metadata related to protein identification (such as search engine or search database) are also provided here. A

powerful spectrum viewer is available as well, including an automatic annotation of the spectra based on submitted fragment ions. Combinations of up to three amino acids are indicated next to the mass differences between consecutive peaks (Supp. Figures 7 and 9). PRIDE Inspector also accesses some of the most popular protein databases (UniProtKB, UniParc, IPI, Ensembl and NCBI nr database) *via* a web service to retrieve the most up-to-date protein sequences and names for the reported identifiers. Using the PRIDE Inspector sequence viewer (Supp. Figures 8 and 11) it is possible to highlight different features in the protein sequence such as identified peptides and PTMs. The updated status of the protein identifier in the database (active, deleted, changed, unknown, merged or demerged, see Supp Information) is also provided, which can affect the reliability of the protein identification. In fact, it is then possible to find peptides that originally matched the sequence of the identified protein, but that no longer match the most recent version of the sequence in the database, The third view then focuses on the peptide identifications themselves. Metadata such as peptide score (adapted for the search engine used) and observed PTMs are displayed for each peptide (Supp. Figures 10-11). In both protein and peptide views, the difference between experimental and theoretical mass-over-charge ratio (delta m/z) is calculated for each peptide precursor and highlighted in the application, which can be useful as an indication for errors or inconsistencies. For both views, it is also possible to filter out the decoy matches and, as such, a straightforward estimation of the peptide FDR is also provided. The fourth view is aimed at accessing and visualizing all spectra in the data set, not only the identified ones (Supp. Figure 12). For mzML files, chromatograms are displayed here as well (Supp. Figure 13). Submitted metadata (e.g. precursor m/z and intensity) is shown for each entry, along with calculated information such as the number of peaks or the total peak intensity. Manual annotation of spectra is supported as well for quick *de novo* sequencing. The fifth view provides a collection of summary charts for assessing the overall properties of the dataset. At the time of writing, up to eight different charts can be generated per dataset, depending on the information available (Supp. Figures 14-18). These simple and easily understandable charts can provide a quick overview on data quality and reliability. Importantly, information in the spectrum-related charts can be shown for identified, unidentified or all spectra. Each chart is documented thoroughly in the supplementary information. Finally, a sixth tab focuses on the quantification information, where available (Supp. Figure 19). This kind of data is currently only present in a small number of PRIDE submissions but it is expected to become more and more popular. Apart from visualizing the quantification values for both protein and peptides, it is also possible to generate histograms where the expression values of up to ten proteins can be compared. Sample metatada for each reagent can also be easily visualized. Ratios can always be recalculated if the user decides to change the control sample.

Apart from the six main tabs, the ‘Search PRIDE’ panel gives access to all public data in PRIDE. It is then easy to search for particular experiments filtering by different types of metadata (Supp. Figures 20-21). In addition to data visualization and analysis functionality, PRIDE Inspector also provides various data export options (Supp. Figure 22). First of all, all spectra can be exported to Mascot Generic Format (mgf) files. In addition, details for all protein and/or peptide identifications (including PTMs), and the peptide to protein mappings can be output as tables in tab-delimited format. Finally, spectra and chromatograms (including annotations) can be saved as images in various formats.

PRIDE Inspector is fully supported and maintained by the PRIDE team. Moreover, it provides extra APIs/libraries, which can be reused independently by the scientific community: the PRIDE XML JAXB library (for rapid and memory-efficient reading of PRIDE XML files), and the PRIDE mzGraph Browser library (for the visualization and annotation of spectra and chromatograms). These libraries are described in the supplementary information. In addition, new features can be easily added to PRIDE

Europe PMC Funders Author Manuscripts

Inspector thanks to its modular software architecture and permissive open source licensing. Currently ongoing extensions include full support of the version 1.1 of the mzIdentML community standard for peptide and protein identifications[9], since this format has only just reached stability (v1.1 was released on September 2011). Once mzIdentML is fully supported, it will also be possible to check thoroughly the issues related to protein inference[10]. This means that researchers need to be aware of this limitation when interpreting protein identifications reported by non-ambiguous (or shared) peptides. The PRIDE XML format is limited for that aim in the sense that only one of the possible peptide-protein mappings is usually reported.

PRIDE Inspector thus provides a user-friendly, comprehensive tool for the browsing, inspection, and evaluation of data in the PRIDE database, or in a compatible standard file format. As such, we believe that PRIDE Inspector will substantially increase the ability of researchers, editors and peer-reviewers to explore, review, evaluate, and reuse proteomics data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Wellcome Trust [grant number WT085949MA] and EMBL core funding. RGC is supported by EU FP7 grant SLING [grant number 226073]. JAV is supported by the EU FP7 grants LipidomicNet [grant number 202272] and ProteomeXchange [grant number 260558]. AF was partially supported by the Spanish network COMBIOMED (RD07/0067/0006, ISCIII-FIS). LM would like to acknowledge support from the EU FP7 PRIME-XS grant [grant number 262067].

Abbreviations

API	Application Programming Interface
FDR	False Discovery Rate
GUI	Graphical User Interface
IPI	International Protein Index
JAXB	Java Architecture for XML Binding
MGF	Mascot Generic Format
PRIDE	PRoteomics IDEntifications (database)
PTM	Post-Translational Modification

References

1. Anonymous. Nat Biotechnol. 2009; 27:579. [PubMed: 19587644]
2. Vizcaino JA, et al. Nucleic Acids Res. 2010; 38:D736–742. [PubMed: 19906717]
3. Deutsch EW, Lam H, Aebersold R. EMBO Rep. 2008; 9:429–434. [PubMed: 18451766]
4. Hill JA, Smith BE, Papoulias PG, Andrews PC. J Proteome Res. 2010; 9:2809–2811. [PubMed: 20356086]
5. Barsnes H, Vizcaino JA, Eidhammer I, Martens L. Nat Biotechnol. 2009; 27:598–599. [PubMed: 19587657]
6. Searle BC. Proteomics. 2010; 10:1265–1269. [PubMed: 20077414]
7. Medina-Aunon JA, Carazo JM, Albar JP. Proteomics. 2011; 11:334–337. [PubMed: 21204261]
8. Martens L, et al. Mol Cell Proteomics. 2011; 10 R110 000133.

9. Eisenacher M. Methods Mol Biol. 2011; 696:161–177. [PubMed: 21063947]
10. Nesvizhskii AI, Aebersold R. Mol Cell Proteomics. 2005; 4:1419–1440. [PubMed: 16009968]

Europe PMC Funders Author Manuscripts

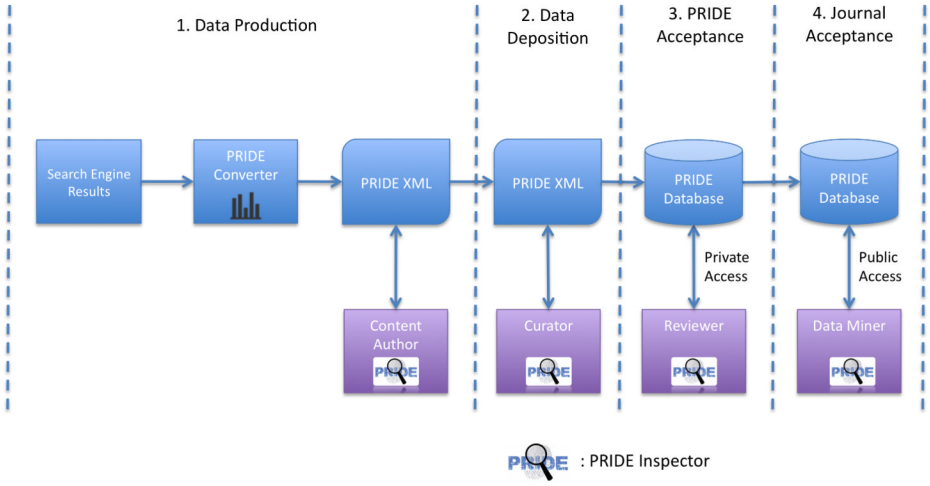

Figure 1.

PRIDE Inspector helps to perform every stage of the PRIDE submission workflow. The workflow consists of four stages: **1. Data Production**: Search engine output results are converted into PRIDE XML files using PRIDE Converter. Authors can then use the PRIDE Inspector to perform an initial assessment on data quality and check metadata annotation before submission to PRIDE. **2. Data Deposition**: the submitted PRIDE XML files are reviewed by PRIDE's in-house curators using the PRIDE Inspector. **3. PRIDE Acceptance**: the submission is accepted by PRIDE and the data is kept private. Journal's reviewers and editors can access these private PRIDE experiments using the PRIDE Inspector. **4. Journal Acceptance**: the submission is made public in the PRIDE database after journal acceptance. Data miners can extract, download or view PRIDE experiments using the PRIDE Inspector.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Figure 2.

Screenshots showing some of the graphical features of PRIDE Inspector: A) Section of the "Spectrum View' tab, B) 'Protein View' tab including the 'Spectrum Viewer' showing MS/MS fragment ion annotations (only b ion annotations are shown), C) 'Quantification View' D) 'Search PRIDE' panel, E) 'Number of Peptides Identified per Protein' chart, and F) 'Delta m/z' chart.