PRIDE: Quick tour

Juan Antonio Vizcaíno [1]

- Proteins
- Beginner
- 0.5 hour

This quick tour provides a brief introduction to the EMBL-EBI's PRoteomics IDEntifications database, PRIDE.

Learning objectives:

- Have a basic understanding of PRIDE and how it can help you gain access to protein identifications from mass-spectrometry-based proteomics experiments.
- Have a basic understanding of the PRIDE submission pipeline.
- Knowing where to find out more about PRIDE.

What is PRIDE?

The PRIDE [2] (PRoteomics IDEntifications [3]) database is a publicly-accessible repository for proteomics data, which covers: protein and peptide identifications (including post-translational modifications), protein and peptide expression values and supporting evidence from mass spectrometry [4] (both peak lists and raw data [5]).

PRIDE is a core member in the ProteomeXchange (PX) consortium [6], which provides a standard framework for the submission and dissemination of mass spectrometry (MS)-based proteomics data to public-domain repositories. PRIDE acts as the initial submission point of MS/MS data. Datasets are then submitted to ProteomeXchange via PRIDE and are handled by expert biocurators. PRIDE also integrates the data into other resources such as UniProt [7].

Finally, the PRIDE team is responsible of the development and maintenance of several tools to help you to submit your data: PRIDE Converter 2, PRIDE Inspector and the ProteomeXchange submission tool.
What does PRIDE contain?

What does PRIDE store?

PRIDE [2] aims to represent the submitter's original view of the data. PRIDE stores:

- **Processed identification results** [8]: peptide and protein identifications, including post-translational modifications (PTMs);
- **Mass spectrometer output files** [9]: both raw data [5] and peak list [10] files;
- Quantification values: peptide and protein expression values (if available);
- **Metadata** [11]: technical and biological metadata;
- Other files related to one submission: gel images, scripts, etc

Why do we need PRIDE?

The PRIDE [2] database is one of the main public repositories for proteomics [12] data that have been generated by MS approaches. In addition, PRIDE is leading the ProteomeXchange consortium.

Various journals in the field strongly support, and some even mandate, deposition of MS-based proteomics data into a public proteomics repository such as PRIDE (or others from the ProteomeXchange consortium). This enables researchers to understand, and potentially reproduce, experiments described in a particular publication. It also allows you to reanalyse mass spectra data as protein sequence databases and search-engine toolkits improve.

In addition, PRIDE provides proteomics data to other data resources such as UniProt [13].
Searching and getting data from PRIDE

Searching using PRIDE Archive

You can use the main PRIDE Archive [14] website to search across datasets, and a list of project summaries will be shown as a result. (See Figure 2).

![Search results list](image)

Figure 2 An example search results list for the term 'stress' in the PRIDE Archive website.

Search terms (in the case of terms controlled by an ontology [15]) may not match directly with the original annotation [16] of a dataset, but a match can be inferred using the ontology hierarchy. For example a search for 'brain' will also retrieve datasets annotated with 'hippocampus' or 'cerebral cortex'. In those cases the search term will be shown in brackets behind the original annotation.

Project summary

A project summary includes all the metadata [11] for a dataset. For example:

- Project accession [17] (dataset identifier [18]);
- Title;
- Description.
A link to download the data files is present in the upper right-hand corner (see Figure 3).

Figure 3 An example of a project summary, corresponding to dataset PXD000442.

**Downloading data using PRIDE Archive**

For each project dataset in the PRIDE Archive there is a link to download all the available data files. (Figure 4).
File Download

8 Result Files

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>HTTP Download</th>
<th>Fast Download (Aspera)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRIDE_Exp_Complete_Ac_31011.xml</td>
<td>945.18 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31012.xml</td>
<td>582.541 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31013.xml</td>
<td>3,353.643 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31014.xml</td>
<td>1,086.579 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31015.xml</td>
<td>4,622.024 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31016.xml</td>
<td>1,169.444 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31017.xml</td>
<td>1,197.442 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
<tr>
<td>PRIDE_Exp_Complete_Ac_31018.xml</td>
<td>485.349 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
</tbody>
</table>

106 Search Files

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>HTTP Download</th>
<th>Fast Download (Aspera)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPHBR_CIB_MCEZ_GTU_CEL_P1_10_15.mzid.gz</td>
<td>52.350 MB</td>
<td>Download</td>
<td>Download</td>
</tr>
</tbody>
</table>

Figure 4 An example of a download page, corresponding to the dataset PXD000442.

Individual files of a dataset can be downloaded via [HTTP](http://example.com) [19], [FTP](http://example.com) [20] and, if possible, [Aspera](http://example.com) [21]. In the case of processed result files ([PRIDE XML](http://example.com) [22] or [mzIdentML](http://example.com) [23]), these files can also be opened directly using the PRIDE Inpsector Java webstart in order to visualise them.

Submitting data to PRIDE

PRIDE enables you to submit protein and peptide identification/quantification data, with the accompanying mass spectral evidence, which is due to be published in a peer-reviewed publication. The main focus of PRIDE is to support the deposition of shotgun MS/MS proteomics datasets.

The current way of submitting data to PRIDE is following the [PX consortium guidelines](http://example.com) [24]. For more information, see [Get help and support on PRIDE](http://example.com) [25].

Figure 5 (below) shows an overview of the submission process.
Figure 5. Overview of the submission to MS/MS datasets to ProteomeXchange via PRIDE.

PRIDE datasets are kept private by default, allowing journal editors and peer reviewers to access PRIDE without their data being made available to the general public. As a result, PRIDE is now the recommended submission point for proteomics data for several journals, including *Nature Biotechnology* [26], *Nature Methods* [27], *Molecular and Cellular Proteomics*, [28] and *Proteomics* [29].

**PRIDE tools**

**PRIDE Converter 2**

PRIDE Converter 2 is a Java open source tool that can be used to convert a variety of popular proteomics [12] data formats (search engine output files [30], such as *Mascot dat* [31], *X!Tandem* [32] *xml* [33]), into well-annotated PRIDE XML [22] files (Figure 6). It is freely available from the PRIDE Converter 2 download page [34].
Figure 6 The PRIDE Converter 2 tool.

PRIDE Converter 2 is available to download from the project website [35].

PRIDE Inspector

PRIDE Inspector is an integrated desktop application for visualising and assessing the quality of a dataset (it supports the following data formats: mzML [36], PRIDE XML [22] and mzIdentML [23]). It is freely available from this PRIDE Inspector zipped folder [37].
Figure 7 PRIDE Inspector showing a result file from dataset PXD000447.

PRIDE Inspector is available to download from the PRIDE Inspector project's website [38] (Figure 7).

**ProteomeXchange Submission Tool**

The ProteomeXchange Submission Tool is a Java open-source tool that can be used to submit data to PRIDE (as part of the ProteomeXchange consortium). It annotates and transfers the files to the EBI. It is freely available from the ProteomeXchange submission page [39].
The ProteomeXchange Submission Tool is available to download from the Proteome Central website (Figure 8).

PICR

The Protein Identifier Cross-Reference (PICR) service is a web application that provides web and programmatic (SOAP and REST) access to a mapping algorithm based on 100% sequence identity to proteins from over 98 distinct source databases (Figure 9).
Figure 9 Screenshot from the PICR web application

PICR is available from the [project website][40].

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on PRIDE

Support

- For more information on how to use PRIDE, see the [PRIDE Archive Help Overview][44].
- There is a more detailed course in Train online about how to do [ProteomeXchange submissions via PRIDE][49].
If you have any questions about PRIDE, email the pride-support [at] ebi.ac.uk (PRIDE support) team

References


Collaborators

PRIDE was originally developed through a collaboration between the EMBL-EBI, Ghent University in Belgium and the University of Manchester in the UK.

Funding

The PRIDE team is funded at present by the Wellcome Trust [grant number WT101477MA], the EU FP7 grants ‘ProteomeXchange’ [grant number 260558], and ‘PRIME-XS’ [grant number 262067], the BBRSC ‘PROCESS’ grant [reference BB/K01997X/1] and EMBL core funding.

In the recent past PRIDE has also been funded in the past by the Wellcome Trust [grant number WT085949MA] EU grant ‘LipidomicNet’ [grant number 202272], and the BBSRC grant ‘PRIDE Converter’ [reference BB/I024204/1].

Contributors
Juan Antonio Vizcaino [1]

EMBL-EBI
PRIDE Group Coordinator
Juan is the Group Coordinator of EMBL-EBI's PRIDE resource for mass spectrometry proteomics data, and he has authored more than 65 papers. He joined the PRIDE team as a postdoc in 2006 and has been involved in numerous projects that span research, development of tools, and training and support for PRIDE users. Since 2011, he has been managing the ProteomeXchange consortium of proteomics resources. He has a background in the life sciences, with a PhD in Molecular Biology from the University of Salamanca, Spain, a Masters in Microbiology and undergraduate degrees in Pharmacy and in Biochemistry.

Tobias Ternent [55]

EMBL-EBI
Bioinformatician - Hermjakob group
Tobias is a curator for the PRIDE team (2013), and previously a curator for the ArrayExpress team (2011). Before joining the EMBL-EBI, he studied a Bachelor's in Computer Science from the University of Manchester (2005), and later a Master's in BioHealth Informatics in the same department (2010) which involved collaboration with the KNH-Centre for Biomedical and Forensic Egyptology. Along the way, his working experience has included a few software companies, and in particular managing clinical trial data for AstraZeneca for translational informatics.

Source URL: https://www.ebi.ac.uk/training/online/course/pride-quick-tour-1

Links
[1] https://www.ebi.ac.uk/training/online/trainers/juan
[2] https://www.ebi.ac.uk/training/online/glossary/pride
[4] https://www.ebi.ac.uk/training/online/glossary/mass-spectrometry
[5] https://www.ebi.ac.uk/training/online/glossary/raw-data
[8] https://www.ebi.ac.uk/training/online/glossary/processed-identification-results
[9] https://www.ebi.ac.uk/training/online/glossary/mass-spectrometer-output-files
[10] https://www.ebi.ac.uk/training/online/glossary/peak-list
[11] https://www.ebi.ac.uk/training/online/glossary/metadata
[12] https://www.ebi.ac.uk/training/online/glossary/proteomics
[13] https://www.ebi.ac.uk/training/online/glossary/uniprot
[15] https://www.ebi.ac.uk/training/online/glossary/ontology
[16] https://www.ebi.ac.uk/training/online/glossary/annotation
[17] https://www.ebi.ac.uk/training/online/glossary/accession
[18] https://www.ebi.ac.uk/training/online/glossary/identifier
[19] https://www.ebi.ac.uk/training/online/glossary/http
[20] https://www.ebi.ac.uk/training/online/glossary/ftp
[21] https://www.ebi.ac.uk/training/online/glossary/aspera