

Proteomics data submission strategy for ProteomExchange

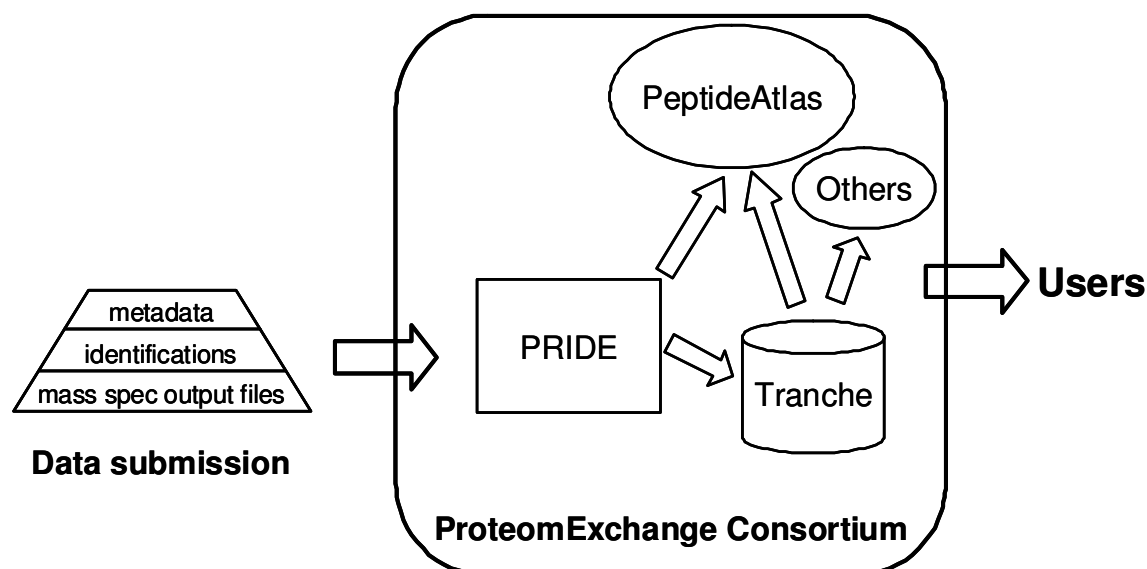
1. Summary

This document provides detailed guidelines for the submission of mass spectrometry-derived proteomics data to the ProteomExchange consortium databases PRIDE, PeptideAtlas, and Tranche. First the policy is summarized in this section; then in subsequent sections, definitions of terms, descriptions of the relevant resources, details on the submission path, and policies regarding data ownership and data privacy are provided. This policy has been adopted by the HUPO Plasma Proteome Project for the collection of its Phase II data; it is hoped that widespread adoption will follow.

Each submission shall consist of three major components: mass spectrometer output files, study metadata, and peptide/protein identifications (further details in section 4; definitions provided in section 2). All submissions will include all three components and will be made to the PRIDE repository using data sufficiency guidelines established by PRIDE as described below.

At the time when the submitted data are declared publicly available by the submitter, all mass spectrometer output files will be deposited in the Tranche repository. Hash keys required to download this information from Tranche and study metadata will be displayed in PRIDE and actively transmitted to PeptideAtlas and any other participating ProteomExchange repositories (see section 3 for information about the individual repositories) for further processing.

This insures that a simple one-time submission from a contributor is automatically distributed to all ProteomExchange repositories with sufficient information.



Summary Figure: data submissions are sent to the ProteomExchange Consortium via PRIDE. The ProteomExchange partners then ensure data is distributed internally, ultimately giving users the ability to access the data from any participating database.

2. Definitions

Proteomics data come in a variety of forms, which are defined here.

- Raw data: the binary, vendor-specific output files directly created by the instrument software. These files are typically large (several gigabytes) and require specialized software in order to be read.

Mass spectrometer output files: the data and metadata generated by mass spectrometers, usually one file per run (although some instruments put multiple runs per file). The data may be the original profile mode scans or may already have had some basic processing like centroiding applied. They may be raw data as described above, or peak list spectra in a standardized format (but not a plain text format). However, it is important that all of the scans that were generated are included with applicable metadata.

- Processed peaklists: heavily processed form of mass spectrometry data, usually derived from the raw data files through various (semi-)automatic steps, e.g.: centroiding, deisotoping, charge deconvolution. These files are formatted in plain text, with typical formats dta, pkl or mgf.
- Standardized data formats
There are currently three widely known mass spectrometry data formats in Proteomics: mzXML (developed at the Institute of Systems Biology (ISB), Seattle), mzData (developed by the HUPO Proteomics Standards Initiative (PSI)), and the successor to both of the above: mzML (jointly developed by the ISB and PSI, to be finalized and released in Q2 of 2008). These data formats can be used to represent processed peaklists, as well as raw data. In addition to the mass spectra, they contain detailed metadata (see below) that provide context to the measurements. These data formats may contain either the original profile mode spectra or centroided data.
- Identifications
Proteomics mass spectra can be matched to peptides or proteins, resulting in identifications for those spectra. Typically a spectrum is considered identified if the score attributed to a peptide or protein match qualifies against an *a priori* or *a posteriori* defined threshold. In the case of fragmentation spectra, the initial identification will consist of a peptide sequence; subsequent steps will derive a list of proteins from the identified peptides. The protein assembly step can be a discernible process with its own input and output files, or it can be implicit in the overall identification software.
- Metadata
Whereas mass spectra present the core output of any mass spectrometer, a simple collection of spectra does not provide sufficient information for confident interpretation. This lack of context can be solved by providing relevant metadata along with the spectra. Mass spectrometer output files (see above) typically accommodate this information in association with the spectra.

3. PRIDE, PeptideAtlas and Tranche

There are currently several proteomics data repositories, each with a different major focus. Here we discuss the Proteomics Identifications Database (PRIDE) at EBI, PeptideAtlas at ISB and ETHZ, and Tranche at the University of Michigan at Ann Arbor.

PRIDE is focused on submissions of identifications, usually presented in correlation with a manuscript or published paper. Individual MS or MS/MS spectra supporting the identifications are optional, but recommended. Mass spectrometer output files, although linked from PRIDE, are not stored inside the repository proper. Metadata are required as part of the submission. The datasets included in PRIDE can be searched by the metadata. Data submissions can be held privately within PRIDE, albeit allowing reviewers and journal editors access if desired, until an optional preset date is reached, or until the submitter chooses to release the data to the public.

The Tranche repository accepts any proteomics-related files, regardless of format and stores them in a large distributed file system on the internet. Uploading of mass spectrometer output files is encouraged. All deposited files have a unique hash key and each file is accessible to anyone with the hash key; if the hash key is publicly accessible, then the files are publically accessible. PRIDE already makes use of Tranche for the actual storage of raw data submitted to PRIDE, and allows users to retrieve these files again from Tranche using the hash keys.

PeptideAtlas accepts only mass spectrometer output files and associated metadata, but not identifications. Submitted files are reprocessed using multiple search strategies and the Trans Proteomic Pipeline. All experiments are then combined to form an inclusive view of all peptides and proteins observed for each species across all contributed data. PeptideAtlas contains associated software tools that support data analysis and mining, including spectral searching, proteotypic peptide selection for targeted proteomics approaches, and a general estimate of protein abundance.

Each of these repositories may thus present different views on an experiment, or contain different data components of an experiment. The ProteomExchange Consortium was formed to enhance communication and automated exchange of data among these and other proteomics repositories to ensure a more transparent access for the community to the available data .

Finally, it is important to note that all three repositories discussed in detail here support and actively promote the use of standard data formats.

4. Which data should be submitted to the ProteomExchange partners

As described above, proteomics repositories sometimes focus on or employ different types of data. As the objective of this document is to enable data sharing among the different repositories, and one-time submission to all three repositories, we here provide a minimal list of data types that must be made available with each submission.

- Mass spectrometer output files

PeptideAtlas relies uniquely on the information-rich mass spectrometer output files, which can be uploaded to and stored in Tranche. PRIDE captures only the more accessible peaklists in its repository. As such, a submission must include the mass spectrometer output files, as well as the peaklists (identified and unidentified) used for database searching.

- Identifications

While PeptideAtlas performs its own identification strategy directly on the submitted mass spectrometer output files, PRIDE faithfully represents the identifications obtained by the submitter. A list of the identified peptides and proteins must therefore be submitted, including clear links between the peptides and the proteins they identify. Researchers seeking the original datasets from published work would therefore use PRIDE.

- Metadata

Proteomics data are substantially enriched when sufficient metadata are provided. All three repositories are currently quite flexible in terms of the level of detail that they can accept for experimental metadata. The PRIDE repository currently presents the most stringent requirements for (semi-)structured metadata. PRIDE aims to have each dataset fulfill the Minimal Information About a Proteomics Experiment (MIAPE) guidelines for mass spectrometry data and proteomics identifications (both submitted for publication to Nature Biotechnology). From the strong support from the field at large (see the co-authors on Taylor *et al.*, 'The Minimal Information About a Proteomics Experiment (MIAPE)', Nature Biotechnology, 2007), it is clear that submitters should report the metadata required by MIAPE guidelines as soon as they are published. As a general rule prior to publication of these guidelines, metadata should be as detailed as possible, including detailed information on the original sample, experimental protocol, mass spectrometry instrumentation, software processing steps and associated operational parameters, identification software output (score(s), threshold(s) calculated or employed, etc.), and statistical methods and parameters.

An additional important type of metadata is literature references associated with the data, as these are prominently displayed in certain repositories and should therefore elicit due academic credit to the original authors when their submitted data are reused by others. Furthermore, the contact details of the data submitters can also be provided, allowing interested users to contact the original authors if desired. Finally, PRIDE also requires a clearly identified owner of the data, in the form of an encrypted user account on the PRIDE system (see below for data ownership issues).

5. Data submission workflow

All required data should be submitted initially to the PRIDE database (<http://www.ebi.ac.uk/pride/#submission>), which provides the ability to keep the submitted data private until the submitter actively chooses to make it publicly available (see below for data privacy issues).

PRIDE will capture the reported identifications, peaklists and associated metadata in its database, and will deposit the mass spectrometer output files in Tranche. In this

process, the peaklists will be converted into the PSI standard format for mass spectrometry data if not already submitted in this standard. The Tranche files will be referenced from the PRIDE experiment through their hash codes, allowing one-click downloading of the files from Tranche by PRIDE users after public availability of the data.

Upon public availability of the data (see below for details), PRIDE will notify PeptideAtlas of the raw data files in Tranche and the relevant metadata stored in PRIDE, and PeptideAtlas will subsequently obtain and process these files to provide its own view on the obtained mass spectrometry data.

This workflow assures that the exact results of a proteomics experiment will be made available to the community, along with a maximum of metadata and readily usable, standardized mass spectrometry data. At the same time, the mass spectrometer output files will be available to more specialized researchers through Tranche, and PeptideAtlas will be able to reprocess the findings to enrich its database on organismal proteomes and to further enhance the associated discovery tools.

6. Data ownership

The PRIDE database does not assume editorial control or ownership over the submitted data; it maintains the original author as owner of these data. As a result, PRIDE requires that an owner is explicitly identified for each dataset. This is done by associating a user account on the PRIDE system with a submitted (set of) experiment(s). Upon public availability of the data, the original data ownership is maintained in the database, although obviously dissemination and reuse of the released data are no longer restricted at that point.

7. Data privacy

PRIDE allows data to be kept private for any duration of time, until the owner of the data (as identified by the associated PRIDE user account, see above) gives explicit permission to release the data. A variant on this occurs when privately submitted data are associated with a manuscript submitted to a journal. The public availability of the submitted data will then be coordinated with the publication of the associated article in correspondence with the journal editor.

Data that are submitted privately to PRIDE can be shared with collaborators, however, provided that the data owner explicitly grants the associated user accounts on the PRIDE system access to the relevant experiments. Similarly, PRIDE can automatically provide reviewer accounts for each submitted experiment, which can be communicated to journal editors and referees in a submitted manuscript, thus allowing confidential reviewing of the privately submitted data.

The date of submission, as well as the date of public release, is archived in the PRIDE database system. After public release of the data, the PRIDE experiment and the corresponding raw data files in Tranche will be made available to the general public without further reservations. The original ownership of the data will remain asserted in the PRIDE database however. Any restrictions on data dissemination or reuse are obviously removed upon public availability of the data.