Dataset Submission Guidelines for the Human Proteome Project

Version 1.0 (2012 October 30): Based on discussions at the HUPO Proteomics Standards Initiative (PSI) workshop (2012 Mar 12-14), the ProteomeXchange workshop (2012 Mar 15-16), the World HUPO Congress in Boston (2012 Sept 9-13), follow-on emails, and review and approval by the HPP-EC.

Summary

The many participants of the Human Proteome Project (HPP) are expected to produce large amounts of heterogeneous datasets in pursuit of the characterization of the human proteome. The raw data must be captured in a way that makes it accessible and useful to the community. As these data will serve as the basis for the reliable re-measurement of the proteome in many and diverse research projects, all claimed contributions to the HPP based on experimental data must be supported by well-annotated raw data and results successfully stored in an approved data repository at the time of submission of the claim. More detailed requirements by data type are described herein.

Introduction

The Human Proteome Project is composed of three primary resource pillars (antibody, mass spectrometry, and knowledgebase) and two programs, the chromosome-centric HPP (C-HPP) and biology/disease-driven HPP (B/D-HPP). The many participants in these programs are expected to produce large amounts of heterogeneous datasets in pursuit of the characterization of the human proteome. The deposited data and generated data and reagent resources, in turn, will constitute the basis for the exploration of the human proteome in many and diverse research projects and thus significantly contribute to the generation of high quality proteomic datasets for basic and translational research. In the same way that mandatory deposition of raw data became a primary tenet of the Human Genome Project, the deposition of raw data contributing to the HPP will maximize the impact of these contributions, enable reuse of the data, and enhance credibility with the broader biomedical research community.

At present, these guidelines lay out requirements for which types of files must be submitted where, and by implication, the minimum amount of metadata describing the generation and handling of the data, since a minimum amount of information is required to be accepted by the repositories. However, these guidelines do not specify data quality metrics that must be met, as imposed by the MCP Guidelines, for example. Such data quality metrics may become a future addition to these guidelines.

There are several repositories already set up to receive data depositions. The HPP will not require a new repository to be set up. Further, the repositories are working together under the banner of the ProteomeXchange Consortium (http://www.proteomexchange.org/) to ensure simplicity of data submission, automatic distribution of data to the various repositories, and clarity in presentation of datasets across the repositories. The primary participating repositories are the PRoteomics IDEntifications Database (PRIDE) hosted at the European Bioinformatics Institute (EBI) and PeptideAtlas hosted at the Institute for Systems Biology (ISB). Other repositories, such as the distributed file sharing system Tranche will be reviewed from time to time. Together, these repositories and other partners are building

the infrastructure to process data depositions for published journal articles and for large-scale projects such as the HPP.

The HPP mandated the HUPO-PSI to discuss and propose dataset guidelines for the HPP. These guidelines were discussed and elaborated during the PSI 2012 annual workshop in San Diego, March 12-14. PSI working group members as well as stakeholders such as journals representatives, MS vendor representatives, bioinformatics developers and other PSI workshop participants have contributed to the elaboration of this document. These guidelines have subsequently been refined during wider discussions with the HPP community.

The dataset guidelines for the HPP are described below, first in general terms and then with specific details by data type.

General precepts

All claimed contributions to the HPP based on experimental data must be supported by well-annotated raw data, and results must be submitted to an approved data repository at the time of submission of the claim. The claim shall be accompanied by the repository accession number(s). Claims of contribution to the HPP shall be made either in a journal article or at the HPP portal http://thehpp.org/. For the most common data types, the primary data submission point shall be to the appropriate ProteomeXchange repository (based on the data type) and the confirmation of submission shall be in the form of a ProteomeXchange identifier. Further data type-specific extensions are described below.

Data type specifics

MS/MS datasets

Tandem mass spectrometry (MS/MS) datasets must be submitted to PRIDE via the ProteomeXchange submission tool/toolkit, and accepted by ProteomeXchange, following ProteomeXchange guidelines, including raw data, results, and sufficient study metadata. Contribution claims must be accompanied by a valid PXD identifier. See details in the last section of this document. Once the dataset has been accepted by ProteomeXchange, its availability will be announced via RSS, and it will be automatically accessed by automated reprocessing resources such as PeptideAtlas and GPMdb. PeptideAtlas will reprocess data using Trans-Proteomic Pipeline (TPP), applying a stringent FDR threshold of 1% at the protein level (not peptide level).

SRM datasets

Selected Reaction Monitoring (SRM) datasets must be submitted to PeptideAtlas (PASSEL component) via the PeptideAtlas data submission mechanism, and accepted by ProteomeXchange, following ProteomeXchange guidelines, including raw data, results, and sufficient study metadata. Contribution claims must be accompanied by a valid PXD identifier.

Molecular interactions datasets

Molecular interactions datasets must be submitted to IMEx (http://www.imexconsortium.org/), checked by an IMEx curator, and accepted following IMEx guidelines. Contribution claims must be accompanied by a valid IMEx identifier.

MS1 datasets

Single stage mass spectrometry (MS1) datasets come in two primary types, peptide mass fingerprinting (PMF) data and MS1 mapping (i.e. shotgun without MS/MS triggering). PMF datasets analyzed with Mascot are supported by ProteomeXchange and should be submitted to PRIDE. MS1 mapping datasets are not currently accepted by ProteomeXchange repositories in a manner in which a PXD identifier can be assigned. The dataset could be uploaded to the PeptideAtlas PASS system or to the EBI/PRIDE raw data repository accompanying the MS/MS or SRM data that usually will accompany the MS1 mapping data. However, there is not currently a mechanism to ensure that the submission is complete.

Gel images

Gel images accompanying MS/MS datasets should be uploaded as supporting raw data with the MS/MS dataset. There is a possibility that Swiss2DPage or ProteoRed may be able to serve as a Gel image repository, but this is not yet confirmed.

Imaging MS

For imaging MS data, the guidelines for MS/MS or MS1 data should be followed, as appropriate. The vendor raw formats or imzML files must be uploaded.

Immunohistochemistry images and annotations

Immunohistochemistry (IHC) imaging datasets are not currently accepted by ProteomeXchange repositories in a manner in which a PXD identifier can be assigned. Such datasets could be uploaded to the PeptideAtlas PASS system if an immediate resource is necessary, but there is not currently a mechanism to ensure that the submission is complete or useful. The Human Protein Atlas is a primary resource of IHC images and annotations for the HPP, but there is not a mechanism to receive external submissions of data and characterizations there. It is hoped that in the near future there will be a suitable public repository for IHC data and results, as is currently under development by the Euro-BioImaging Consortium (http://www.eurobioimaging.eu/), but this has not yet been confirmed. Meanwhile, we encourage all HPP imaging datasets to meet the MISFISHIE standard (http://scgap.systemsbiology.net/standards/misfishie/) for specification of results and use the OME data model (http://www.openmicroscopy.org/site/support/file-formats/) for image metadata.

Genomics data (RNA-Sequencing, DNA sequencing, et al.)

Genomics or transcriptomics data that accompany HPP contribution claims must be deposited in a public centralized repository that is designed for this kind of data. Currently this is the NCBI/EBI Sequence Read Archive (SRA). Transcript abundance information must be submitted to NCBI's GEO or EBI's ArrayExpress. SRA, GEO, or ArrayExpress accession numbers must be provided with HPP contributions.

Support of formats

Given the great variability of existing data processing and analysis pipelines, not all data formats are handled by the repositories. Some workflows may rely on data files that cannot be accepted as input for the repositories. As proteomics workflows, file formats, and repositories develop rapidly, data producers are advised to check for the latest status at http://www.proteomexchange.org, and contact prospective repositories early by email. There are two types of data submission workflows to PRIDE/ProteomeXchange, depending on whether it is possible to achieve full processed data representation or not.

A) Full data representation of processed results in PRIDE/ ProteomeXchange is possible

These datasets will get a PXD and a DOI identifier. The data submission must have two mandatory components: raw data and processed results. As of October 2012, the PRIDE/ProteomeXchange submission tool will accept these MS/MS related formats:

1- Raw data:

- mzML, mzXML, mzData. These files must not be heavily processed to be considered 'raw'.
- Thermo .RAW, ABSCIEX .wiff, .wiff.scan, Agilent .d/, Waters .raw/
- imzML, Shimadzu .run/, Bruker .yep

All peak lists formats (mgf, dta, ms2, pkl) can be supported but they will not be considered raw data. They will be considered as 'peak list files'. Peak list data encoded in mzML, mzXML or mzData formats will also be considered as "peak list files".

- Processed result files. At present, PRIDE XML is the only file format that can be accepted by PRIDE to allow full representation of the data. The PRIDE team is working towards getting full support for the PSI standard mzIdentML (together with the different peak list formats). Until this is done, different search engine output files need to be converted to PRIDE XML using existing tools like PRIDE Converter 2 (http://code.google.com/p/pride-converter-2/). Formats supported:

- Tandem XML, OMSSA .csv.
- Mascot .dat
- Sequest Crux .txt
- SpectraST .xls
- Thermo Proteome Discoverer .msf files.
- All the accompanying peak lists formats.

Other types of files are optional and can be supported by the PX submission tool, such as additional search engine output files, quantitation results (in formats such as mzQuantML, mzTab).

B) Full data representation of processed results in PRIDE/ProteomeXchange is not possible

In this case, a PXD identifier will be assigned to the data submissions, but in principle not a DOI (although there could be exceptions). PRIDE will always handle this way the formats that cannot be converted to PRIDE XML using the existing converters or to mzIdentML in the near future. The data submission must have at least two mandatory components:

- Raw data:
    - mzML, mzXML, mzData. These files must not be heavily processed.
    - Thermo .RAW, ABSCIEX .wiff, .wiff.scan, Agilent .d/, Waters .raw/
    - imzML, Shimadzu .run/, Bruker .yep

All peak lists formats (mgf, dta, ms2, pkl) can be supported but they will not be considered raw data. They are considered as 'peak list files'. Peak list data encoded in mzML, mzXML or mzData formats will also be considered as "peak list files".

- Search engine output files: Only those that cannot be converted to PRIDE XML are considered to be 'unsupported formats' and can use this alternative approach. As of October 2012, there are no reliable converters to PRIDE XML for the following formats:
    - pepXML, protXML.
    - SEQUEST .out, SEQUEST SRF, SQT.
    - Scaffold format.
    - MaxQuant output files, DTASelect, and many others.

Other types of files are optional and can be supported. It is the case of quantitation results (mzQuantML, mzTab), images, etc.