References

[1] Kaiser J. Proteomics. Public-private group maps out initiatives. Science 2002;296:827.

[2] Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jr., Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25:887-93.

[3] Martínez-Bartolomé S, Medina-Aunon JA, Jones AR, Albar JP. "Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports". Proteomics (in press) 2009.

The Proteomics Identifications database (PRIDE), its associated tools and the ProteomeXchange consortium

Juan Antonio Vizcaíno, Florian Reisinger, Richard Côté, Henning Hermjakob

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Abstract

The Proteomics Identifications Database (PRIDE, http://www.ebi.ac.uk/pride) has become one of the main repositories of mass spectrometry derived proteomics data. In this communication we will summarize the main capabilities of the PRIDE system, including its associated tools. Finally, we will introduce the ProteomeXchange consortium, as a collaborative approach to share proteomics data between the most important proteomics repositories.

The PRIDE Proteomics Identifications database (http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute (EBI) provides users with the ability to explore and compare mass spectrometry (MS) based proteomics experiments that reveal details of the protein expression found in a broad range of taxonomic groups, tissues and disease states [1]. PRIDE stores three different kinds of information: peptide and protein identifications derived from MS or MS/MS experiments, MS and MS/MS mass spectra as peak lists, and any and all associated metadata. PRIDE is now the recommended submission point for proteomics data for several journals such as *Nature Biotechnology, Nature Methods, Molecular and Cellular Proteomics*, and *Proteomics*.

1. PRIDE associated tools

PRIDE relies heavily on two additional tools: the Ontology Lookup Service [2] (OLS, http://www.ebi.ac.uk/ols), and the Protein Identifier Cross-Referencing system [3] (PICR, http://www.ebi.ac.uk/Tools/picr). OLS provides convenient and powerful access to a large number of biomedical ontologies and controlled vocabularies (CVs). PRIDE takes advantage of OLS to store, structure, and present any and all metadata annotations on experiments, proteins, peptides and mass spectra.

The PICR tool on the other hand, is built to overcome one of the most recurrent problems in proteomics: the existence of heterogeneous and changing identifiers or accession numbers referring to the same protein in different databases. PICR is used to map all the submitted protein identifications in PRIDE to all known accession numbers for those proteins in the most important protein databases (including UniProt, IPI, Ensembl and RefSeq, among others). Therefore, protein identifications in PRIDE that were originally derived from different databases, or from different time points of the same database, thus become fully comparable. In addition to these two established tools, a new application called Database on Demand [4] (DoD, http://www.ebi.ac.uk/pride/dod) has recently been added to the PRIDE toolkit. This tool allows custom sequence databases to be built in order to optimize the results from search engines for gel-free proteomics experiments.

2. ProteomeXchange Consortium

One of the reasons why proteomics data sharing is not a universal fact yet is the heterogeneity of the

existing proteomics repositories, each repository having a major focus. This is why the ProteomeXchange consortium was founded [1]. The current members and the way they interact are represented in Figure 1. Guidelines for ProteomeXchange submissions are being finalized (http://www.proteomexchange.org), and include three mandatory data types that will have to be included per submission: instrument output files (raw data, peak lists), associated metadata and peptide/protein identifications. A large scale ProteomeXchange pilot submission has already been performed containing data from the HUPO Plasma Proteome Project 2 (PPP 2) [5]. Therefore, certain experiments in PRIDE (experiment accession numbers 8172-8544, http://www.ebi.ac.uk/pride/ppp2_links.do) contain links to files stored in the Tranche repository in the "Experiment View" page. For these experiments, it is therefore already possible to get the original raw data or search engine output files, which are not stored as such in the PRIDE system.

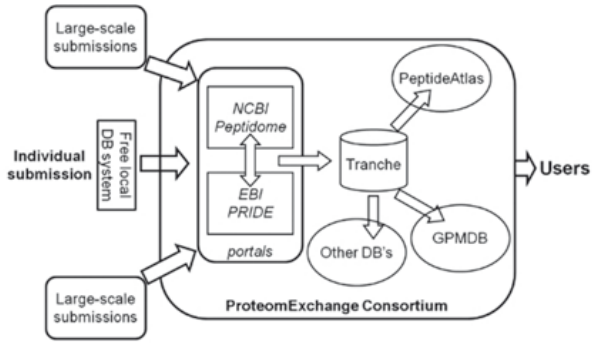

Figure 1. *Summary figure of the ProteomeXchange consortium data flow. Data submissions are sent to the consortium via PRIDE or NCBI Peptidome. The ProteomeXchange partners then ensure data are distributed internally, ultimately giving users the ability to access the data from any participant database.*

References

[1] Vizcaíno JA, Côté R, Reisinger F, Foster J, Mueller M, Rameseder J, *et al*. A guide to the Proteomics Identifications Database proteomics data repository. Proteomics 2009;9:4276-83.

[2] Côté RG, Jones P, Martens L, Apweiler R. and Hermjakob H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. Nucleic Acids Res 2008;36:W372-6.

[3] Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, *et al*. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics 2007;8:401.

[4] Reisinger F. and Martens L. Database on Demand – an online tool for the custom generation of FASTA formatted sequence databases. Proteomics 2009;9:4421-4.

[5] Omenn GS, Aebersold R. and Paik YK. 7(th) HUPO World Congress of Proteomics: launching the second phase of the HUPO Plasma Proteome Project (PPP-2) 16-20 August 2008, Amsterdam, The Netherlands. Proteomics 2009;9:4-6.