PRIDE Proteomes: a Condensed View of the Plethora of Public Proteomics Data Available in the PRIDE Repository

Noemi del Toro, Florian Reisinger, Joseph M. Foster, Javier Contell, Antonio Fabregat, Pau Ruiz Safont, Henning Hermjakob and Juan Antonio Vizcaíno

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

Abstract. The PRIDE (PRoteomics IDEntifications) database [1] (http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute (EBI, Cambridge, UK) is one of the main public repositories for mass spectrometry (MS)-based proteomics data. In PRIDE, the amount of data is constantly growing at a very significant pace. Apart from the data management component, the other main challenge in proteomics resources such as PRIDE is to provide an aggregated and quality-scored version of the peptide/protein identifications found across all the submitted projects, in order to decide which information is more reliable.

PRIDE Proteomes is a new resource providing a condensed, protein centric view of the MS data in PRIDE. As part of the new project a scheme for quality scoring is being developed at present. The information that is used for this aim is the experimental metadata annotation, the number of evidences and the resulting data after applying the 'PRIDE Cluster' [2] algorithm. This scoring will be then propagated to the peptide and protein level by using a set of defined rules.

A beta version of PRIDE Proteomes is now available for four species (human, rat, mouse and *Arabidopsis*) at http://wwwdev.ebi.ac.uk/pride/proteomes.

Keywords: proteomics, bioinformatics, mass spectrometry, database, repository, protein, peptide, psm, pipeline, web service, web application

1 Introduction

The PRIDE database (http://www.ebi.ac.uk/pride) at the EBI is a public data repository for MS-derived proteomics data. It stores peptide and protein identifications, the corresponding mass spectra, post-translational modifications (PTMs), protein/peptide expression values (if available) and experimental metadata.

PRIDE is the initial submission point for MS/MS data in the ProteomeXchange consortium [3] (http://www.proteomexchange.org). ProteomeXchange provides an internationally co-ordinated repository infrastructure compatible with community requirements for data deposition and dissemination. Data is made accessible in ProteomeCentral (http://proteomecentral.proteomexchange.org).

The study-centric view of the PRIDE repository does not provide the opportunity to analyse and compare the information available across all the public data submitted. The main purpose of the PRIDE Proteomes project is to provide a homogeneous and integrated view of the data stored in PRIDE together with a metric that allows the user to be confident in the reliability of the data.

The main challenge is how to determine the quality of the proteomics identifications in a highly heterogeneous resource like PRIDE. This still remains an issue in the field since results from different search engines, especially in different experimental settings, are not directly comparable. Our approach to the problem was the development of the PRIDE Cluster algorithm [2], based on spectral clustering. Data quality assessment is performed at different levels: experiment (based on metadata annotation), protein, peptide and peptide-spectrum match (PSM).

3 System Architecture

The PRIDE Proteomes project is structured in three different layers: database and pipeline, web service and web application. All the layers have been developed in the Java language and provide different levels of abstraction and methods to access the data.

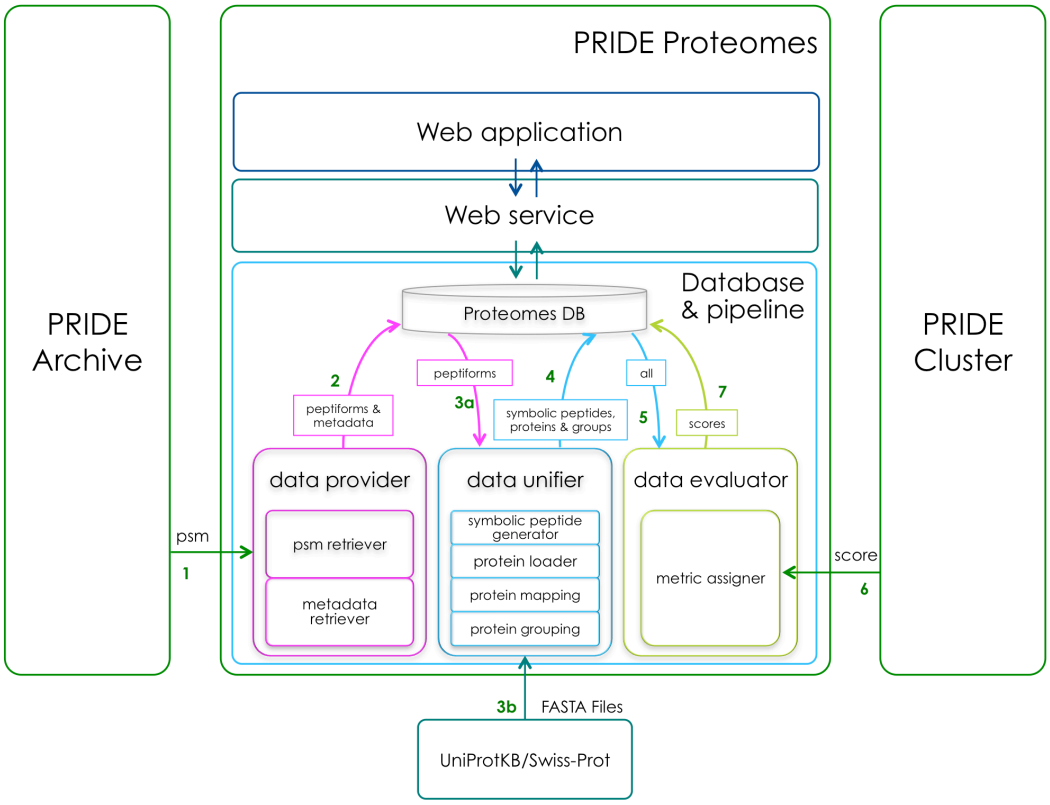

Fig. 1. Data flow of the PRIDE Proteomes project, from the PRIDE Archive, through the pipeline, to the web application available for the user.

3.1 Database & Pipeline

The database is the core of the PRIDE Proteomes project. Apart from the obvious role as a data store, the database is used as a means of communication between the different stages of the pipeline. The proteomes pipeline has been implemented with the Java Spring Framework. Inside the pipeline three sub-pipelines have been differentiated: data provider, data unifier and data evaluator. At the time of writing, the data provider and the data evaluator are still in early stages of development.

Data provider. The main goal of this sub-pipeline (Fig 1, steps 1 and 2) is to generate the peptides that represent the biological entity observed by MS experiments as the result of combining all the PSMs (Peptide Spectrum Matches) that share the same amino acid sequence and modifications (or their absence).

The PSMs are retrieved from the PRIDE Archive repository for the four species initially selected (human, rat, mouse and *Arabidopsis*) and are filtered by at least two criteria: the length of the peptide (between 6 and 100 amino acids) and all the amino acid letters that represent the sequence need to be unambiguous. Once the grouping stage is finished only the peptides created (called 'peptiforms', they represent a raw peptide sequence plus their modifications per species) are stored in the database (Fig 1, step 2).

Together with the peptiforms, the metadata from the projects associated with the PSMs are inserted in the database (Fig 1, step 2).

Data unifier. The pipeline behind the data unifier will be in charge of the next steps:

- Iterate through all the peptiforms generated in the previous stage (Fig. 1, step 3a) and extract from all of them the distinct peptide sequences (this set of sequences will be called 'symbolic peptides').
- In parallel to this process (Fig. 1, step 3b), the reference proteome is loaded into the database. For this purpose the corresponding UniProtKB FASTA files are downloaded, parsed and stored in the Proteomes database (Fig. 1, step 3).
- Once the proteins and the distinct sequences of the peptides are available in the database, an algorithm to map the peptide sequences to the protein sequences is executed. The corresponding start and end positions for the peptide in the protein sequence will be stored in the database, together with flags indicating whether the peptide is fully tryptic, and whether it is unique in the reference proteome (Fig. 1. step 4). .
- The last data unifier step allows the definition and generation of groups of proteins. By means of a simple comma-separated value (CSV) file, the pipeline can be configured with a specific criterion like the pertinence of the same gene and/or family of proteins to create the protein groups. After establishing the protein groups, the unique peptides for a group will be calculated and persisted. (Fig. 2, right, shows a representation of the groups in the web application).

Data evaluator. When all the information has been generated, the data evaluator phase starts. The evaluation of the quality of the evidences studied (Fig 1, steps 6 and 7) is propagated from the raw peptides or peptiforms, to the groups of proteins. Initially the peptiforms are associated a first value, or ranking, generated from the assessment of the PSMs that have helped to build them, and the associated metadata. This low level of quality checking has been implemented in previous prototypes of PRIDE Proteomes as a weighted function of the level of pertinence to one or several generated clusters from the 'PRIDE Cluster' algorithm [2]. In the future, reprocessed data by third parties will also be integrated.

3.2 Web Service

A RESTful web service (http://wwwdev.ebi.ac.uk/pride/ws/proteomes) exposes the content generated by the pipeline. It provides the communication between the data layer and the presentation layer enabling the decoupling among these components and provides a programmatic access for third parties.

Through this service users can perform several queries to the system to retrieve the peptides by specific metadata associated. It is possible to find out if specific peptides are unique across all of the chosen reference proteomes, which can be at the group or protein level. A summary of the main web service methods can be found in on-line documentation provided for the web service.

3.3 Web Application

The PRIDE Proteomes web application provides a rich and highly interactive web interface to analyse, in the context of the protein, the location of the peptides and the PTMs. Complementing the protein-centric view, a 'protein group' view is also available to simplify the comparison between different related proteins (e.g. products of the same gene). In Fig. 2 a screenshot of the current version of the web application is shown.

4 Future Work

PRIDE Proteomes is still in development; the main foundations and core functionality have been established. In the next iteration circle the highest priority for the project will be to achieve a stable confidence metric.

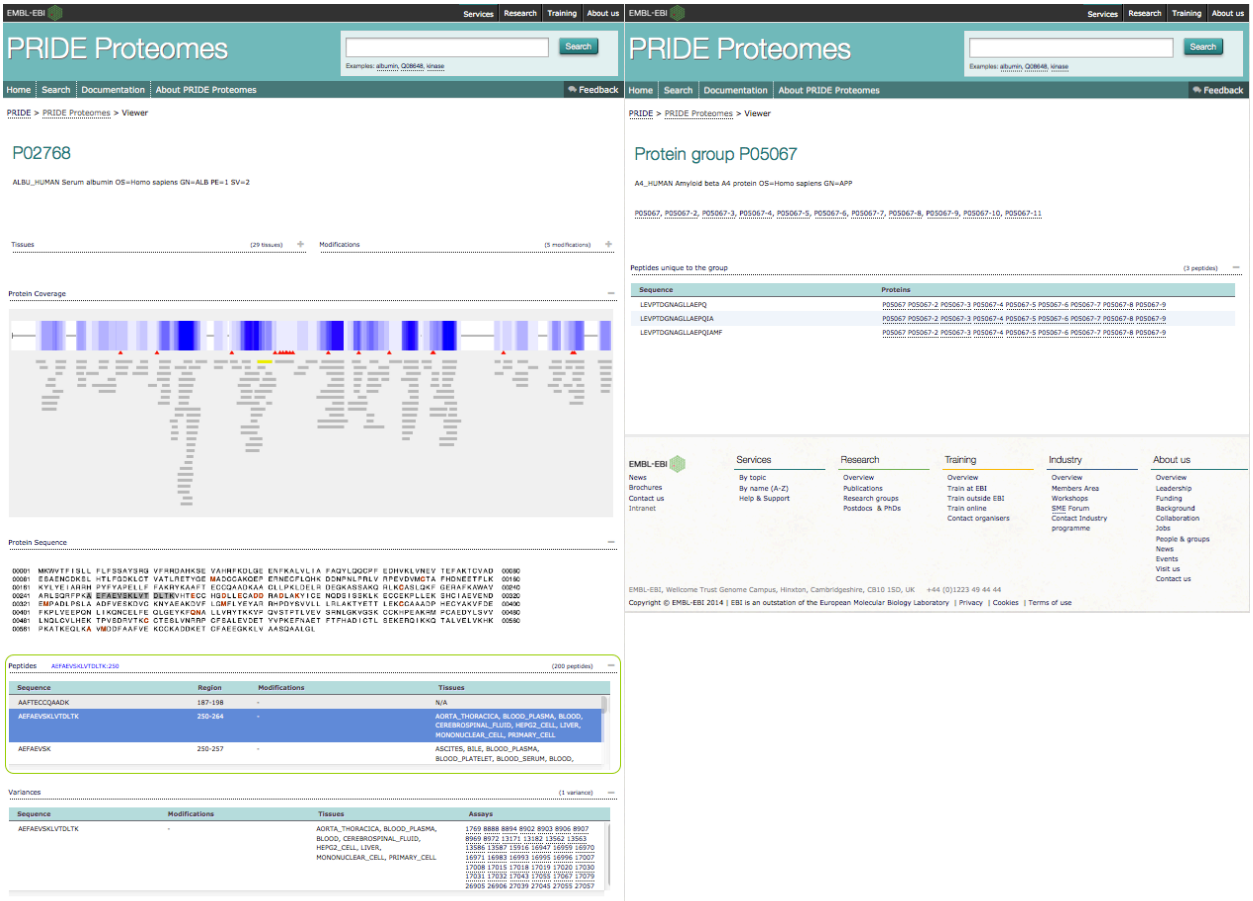

Fig. 2. Protein view (left) and protein group view (right) from the PRIDE Proteomes web application

5 Acknowledgments

We want to acknowledge funding from The Wellcome Trust [grant numbers WT085949MA and WT101477MA].

6 References

1. Vizcaíno J.A., Côté R.G., Csordas A., Dianes J.A., Fabregat A., Foster J.M., Griss J., Alpi E., Birim M., Contell J., O'Kelly G., Schoenegger A., Ovelleiro D., Pérez-Riverol Y., Reisinger F., Ríos D., Wang R. q, Hermjakob H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* Volume 41 p.d1063-9
2. Griss, J., Foster, J. M., Hermjakob, H., and Vizcaino, J. A. (2013) PRIDE Cluster: building a consensus of proteomics data. *Nat Methods* 10, 95-96.
3. Vizcaíno J.A., Deutsch EW2, Wang R., Csordas A., Reisinger F., Ríos D., Dianes J.A., Sun Z., Farrah T., Bandeira N., Binz P.A., Xenarios I., Eisenacher M., Mayer G., Gatto L., Campos A., Chalkley R.J., Kraus H.J., Albar J.P., Martinez-Bartolomé S., Apweiler R., Omenn G.S., Martens L., Jones A.R., Hermjakob H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* 32, 223–226