EuPA OPEN PROTEOMICS 8 (2015) 28–35

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Ten years of public proteomics data: How things have evolved, and where the next ten years should lead us

Kenneth Verheggen[a,b,c], *Lennart Martens*[a,b,c,*]

[a] *Medical Biotechnology Center, VIB, Ghent, Belgium*
[b] *Department of Biochemistry, Ghent University, Ghent, Belgium*
[c] *Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium*

ARTICLE INFO

Article history:
Received 9 July 2015
Received in revised form
14 July 2015
Accepted 21 July 2015
Available online 29 July 2015

Keywords:
Proteomics
Bioinformatics
Public data
PRIDE

ABSTRACT

Ten years ago, the first public proteomics repositories became available online. This anniversary is therefore an excellent occasion to look back on the past decade and evaluate what has changed in this time period. At the same time however, one should also dare to look forward, and therefore prepare for the next 10 years of proteomics data sharing.

The first published efforts at collecting and disseminating online proteomics data are traceable to 2004, with a report by Prince et al. [1] on the one hand, and the original Global Proteome Machine DataBase (GPMDB) publication by Craig et al. [2] on the other hand. It is worth considering the differences between these two efforts, as the former was aimed at a generic dissemination platform that held only high-level meta-information about the available data in a queryable format, while the latter provided a full online data ecosystem, complete with an integrated search engine (X!Tandem [3]) and a fully queryable data structure that captured the details of each spectrum and peptide-to-spectrum match (PSM). A similar concept to the GPMDB was also found in the PeptideAtlas system published shortly afterwards by Deseire et al. [4], building upon the Trans Proteomic Pipeline as a common reprocessing back-end for all contained mass spectrometry data [5]. Simultaneously, the Proteomics IDEntifications (PRIDE) database was published as a *bona fide* repository for mass spectrometry based proteomics data [6], similar in basic concept to the system proposed by Prince et al., but with the important extension that PRIDE captured all data in an experiment in a fully structured and queryable form. GPMDB, PeptideAtlas and PRIDE are still fully operational today, and the latter two have been founding members of the unifying ProteomeXchange consortium [7]. In the 10 years that have passed since the original publication of these resources, many

Abbreviations: PSM, peptide-to-spectrum match; PTM, post-translational modifications.

* Corresponding author at: Medical Biotechnology Center, VIB and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium.
E-mail address: lennart.martens@vib-ugent.be (L. Martens).

http://dx.doi.org/10.1016/j.euprot.2015.07.014

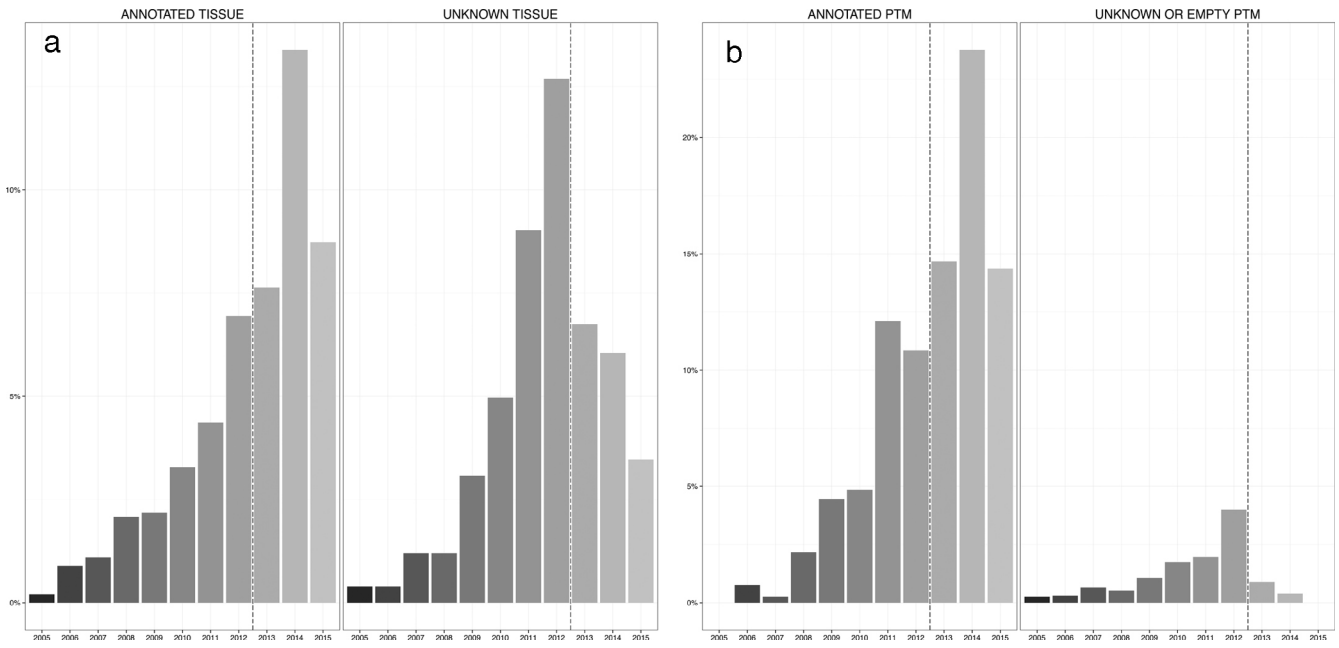

Fig. 1 – Evolution of tissue metadata and PTM data annotation in PRIDE over the past 10 years. Plots are based on public data submitted to PRIDE in each year from 2005 until 2015 (both inclusive). (a) Availability of tissue metadata information in PRIDE over the past 10 years. Number of experiments with available tissue information are shown at left, while number of experiments without this information are shown at right. (b) Availability of PTM annotations on identifications in PRIDE over past 10 years. Number of experiments with available PTM information are shown at left, number of experiments without this information are shown at right.

things have clearly evolved, and this Decennial of public proteomics data sharing is therefore a good moment to look back at what has changed in the field, and what has not. The findings also allow us to make some clear recommendations for the next 10 years of proteomics data sharing.

One of the most important developments in the past 10 years of public proteomics data sharing is certainly the development of numerous community standards for the various proteomics data types [8–11], coupled to minimal reporting guidelines that are all linked to the flagship Minimal Information About a Proteomics Experiment (MIAPE) standard [12]. In principle, these combined developments should have created a common minimal level of data formatting, and metadata annotation for public proteomics data. Unfortunately, when we consider the availability of tissue of origin metadata in PRIDE over the past 10 years as retrieved by the PRIDE web service [13], we can see that it took until 2013 before a trend change was observed in the relative availability of tissue annotation (Fig. 1a). Intriguingly, the main cause for the trend reversal is probably manual curatorial efforts by the PRIDE submissions team [14] (helped in no small part by the PRIDE Inspector software and its built-in quality control plots [15]) rather than the availability of submission software. Indeed, in 2009 the PRIDE Converter application [16] made submission from many search engines much easier thanks to a wizard-like interface, while a direct coupling to the Ontology Lookup Service (OLS) [17] allowed easier metadata annotation [18], but this development did not affect the relative abundance of experiments with missing annotation at the tissue level. Interestingly, the effect of PRIDE Converter on metadata can be spotted quite easily when we consider the reporting of detected post-translational modifications (PTMs) in PRIDE (Fig. 1b), as the number of annotated PTMs grows much more sharply than the number of unknown PTMs. The main reason for the relative success of conversion software here is that PTM annotation can be extracted automatically from the search engine output, thus eliminating the need for active user intervention. Even here however, the effect of stepped-up manual curation since 2013 is prominently visible as missing PTM annotation is almost entirely prevented. A similar picture about emerges when we consider the metadata about the instrument used in the analysis. For the sake of clarity, all reported instruments were grouped according to their mass analyzer types for MS and MS/MS analysis (Fig. 2). Some noteworthy trends are the emergence and subsequent dominance of the orbitrap family of mass spectrometers in 2007 [19], and the availability of fast MS/MS analysis in an orbitrap since 2012 [20]. Yet the most striking evolution is the overwhelming number of unannotated instruments, which here too declines from 2013 onwards.

It is obvious that adherence to the original MIAPE guidelines over the past 10 years has been haphazard at best, and that manual verification and active enforcement seems the most effective strategy to ensure compliance. This is highly regrettable because it saps valuable time and effort away from the repository's limited core staff, and because solutions already have been built that can automatically examine data sets for sufficient and semantically correct metadata