JOURNAL OF PROTEOMICS 73 (2010) 2136–2146

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Review

Proteomics data repositories: Providing a safe haven for your data and acting as a springboard for further research

Juan Antonio Vizcaíno[a,1], *Joseph M. Foster*[a,1], *Lennart Martens*[b,c,*]

[a]*EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
[b]*Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium*
[c]*Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium*

ARTICLE INFO

Keywords:
Proteomics
Databases
Bioinformatics
Data standards
Repositories

ABSTRACT

Despite the fact that data deposition is not a generalised fact yet in the field of proteomics, several mass spectrometry (MS) based proteomics repositories are publicly available for the scientific community. The main existing resources are: the Global Proteome Machine Database (GPMDB), PeptideAtlas, the PRoteomics IDEntifications database (PRIDE), Tranche, and NCBI Peptidome. In this review the capabilities of each of these will be described, paying special attention to four key properties: data types stored, applicable data submission strategies, supported formats, and available data mining and visualization tools. Additionally, the data contents from model organisms will be enumerated for each resource. There are other valuable smaller and/or more specialized repositories but they will not be covered in this review. Finally, the concept behind the ProteomeXchange consortium, a collaborative effort among the main resources in the field, will be introduced.

Contents

Abbreviations: CV, Controlled Vocabulary; HGNC, HUGO Gene Nomenclature Committee; MCP, Molecular and Cellular Proteomics; MRM, Multiple Reaction Monitoring; NIH, National Institutes of Health; OLS, Ontology Lookup Service; PICR, Protein Identifier Cross-Referencing; PSI, Proteomics Standards Initiative; QC, Quality Control; SRM, Selected Reaction Monitoring; SBEAMS, Systems Biology Experiment Analysis Management System; TPP, Trans Proteomics Pipeline.

* *Corresponding author*. Department of Medical Protein Research, Universiteit Gent-VIB A. Baertsoenkaai 3 B-9000 Ghent, Belgium. Tel.: +32 9 264 93 58; fax: +32 9 264 94 84.
E-mail address: lennart.martens@UGent.be (L. Martens).

[1] These authors contributed equally to this work.

1874-3919/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2010.06.008

1. Introduction

Public availability of data has been of paramount importance in the fast development of most of the life sciences, and has become one of the foundations of modern biology. Indeed, researchers can now freely access DNA sequence information [1], microarray and gene expression data [2,3], and small molecules and chemicals [4]. At the protein level, well-annotated protein sequences can be accessed in UniProt [5], protein structures in the Protein Data Bank (PDB) [6], protein modifications in UniMod [7] and RESID [8], and protein interactions in the various resources forming the IMEX consortium [9].

This public availability of data is particularly interesting for model organisms, as they have been most vigorously researched using various high-throughput 'omics' analytical methodologies over the last two decades. Indeed, a large proportion of the wealth of data thus obtained has become publicly available to the research community *via* various resources, including the ones mentioned above. While the field of MS proteomics is therefore currently ahead of certain other "omics" approaches (e.g. metabolomics and glycomics) in terms of public data availability, it is still trailing other well-established "omics" disciplines such as genomics and transcriptomics in this respect. Indeed, compared to these more mature fields, relatively few MS proteomics data are currently available in the public domain, despite the increasing popularity of the approach. As a consequence, mandatory full data disclosure in the field of MS proteomics remains an important work in progress [10].

This unfortunate situation is all the more regrettable since there is some biological information that is uniquely available through MS proteomics data. For instance, transcriptomics approaches cannot predict accurately changes in active, mature protein levels in a quantitative way. Indeed, several studies have shown that a simple deduction of protein concentrations from mRNA transcript analyses is not appropriate [11–13].

Another topic that can only be well studied using MS proteomics approaches is the detection and quantification of co- or post-translational protein modifications [14]. A last point to consider here is the use of proteomics as a valuable tool for clinicians to develop new diagnostic methods or to identify biomarkers. While microarray-based approaches have also been used for this purpose, their usefulness may ultimately be more limited. The main reason is that proteomics, unlike transcriptomics, also has access to secreted, circulating proteins in different proximal body fluids such as blood plasma, serum, or urine, which present highly convenient targets for detection and quantification [15,16].

Despite the absence of a universal directive to make published MS proteomics data publicly available, several suitable repositories have been established to address the demand for storage and availability of proteomics data in the public domain. In parallel to the intrinsic complexity of the field, proteomics repositories are quite heterogeneous and have different interests and focus. Clearly, this variation is one of the many reasons why data sharing in proteomics remains limited: the current situation is simply too confusing for researchers in the field. Therefore, it is important to mention here that no single proteomics data resource will be ideally suited to all possible use cases and all potential users. As a matter of fact, existing resources already display a remarkable complementarity in that respect.

The main publicly available databases for proteomics data are the Global Proteome Machine Database (GPMDB) [17], PeptideAtlas [18], the PRoteomics IDEntifications database (PRIDE) [19], Tranche (http://www.tranche.proteomecommons.org), and the most recent addition to the list, NCBI Peptidome [20]. Additionally, there are other very valuable proteomics resources such as Human Proteinpedia [21], PepSeeker [22], the Genome Annotating Pipeline (GAPP) [23], MAPU [24], OPD [25],