Bioinformatics Support for Farm Animal Proteomics

Aivett Bilbao and Frédérique Lisacek

Abstract In this chapter, we attempt to compile information published in the most recent reviews and regular publications highlighting the use of bioinformatics in the field of veterinary proteomics. We present a summary of the data resources and popular end user-oriented computational tools that do not require advanced informatics skills.

1 Introduction

The application of proteomics in veterinary science is lagging behind in comparison to studies that have explored the potential of advanced proteomic technologies in human research. The situation is particularly acute in clinical medicine (Ceciliani et al. 2014). This slow start may turn out an advantage, as the recent boost in veterinary proteomics is contemporary with technological development (e.g. targeted proteomics or data-independent acquisition) and improvement of method accuracy and coverage. This progress is in turn challenging the design of automation procedures necessary to cope with ever-increasing amounts of data, thereby justifying a dedicated chapter on bioinformatics in this book.

Several generic reviews summarise the advent of shotgun proteomics (Aebersold 2003; Nesvizhskii 2010) that provides the original context for software development. The corresponding methodology mainly focused on protein identification based on sequence database search engines is now mature with its recognised shortcomings such as a strong bias towards abundant proteins, and the present chapter is centred on more recent efforts applicable to veterinary science.

A. Bilbao
Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
e-mail: bilbao.aivett@gmail.com

F. Lisacek (✉)
Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
e-mail: frederique.lisacek@sib.swiss

A.M. de Almeida et al. (eds.), *Proteomics in Domestic Animals: from Farm to Systems Biology*, https://doi.org/10.1007/978-3-319-69682-9_18

One of the remarkable paradigm shifts from shotgun proteomics with data-dependent acquisition (DDA) is data-independent acquisition (DIA). It introduces the systematic fragmentation of precursor ions as opposed to a limited selection based on high-intensity peaks. Within successive and overlapping sliding windows along the *m*/*z* dimension, all precursor ions are fragmented resulting in a very populated tandem MS collection of redundant data (Chapman et al. 2014; Bilbao et al. 2015a). The sheer amount of mass spectra requires smart and accurate software for data analysis (Bilbao et al. 2015a). DIA is now well established for its ability to monitor detectable peptides with high sensitivity and reproducibility across multiple samples. Specifically, since DIA fragmentation of all detectable ions within a wide *m*/*z* range is systematically carried out regardless of intensity, then extracted ion chromatograms (XICs) can be generated at the fragment ion level. It is therefore particularly suitable to perform more consistent and accurate quantification. A direct consequence of using DIA methods is the expanding use of spectral libraries in particular as an alternative approach to sequence database search. Matching experimental to reference spectra is considerably faster and less error-prone than checking all possible theoretical spectra of the tryptic digest of proteins. Corresponding software tools for spectral library searching are reviewed in Griss (2016). However, the lack of a standardised file format and the possible incomplete coverage of spectral libraries are still limiting expansion. At present, library search is recognised as a helpful complementing approach to database search.

The second major development in recent proteomics is quantification and in particular targeted proteomics (Picotti and Aebersold 2012). Selecting the appropriate peptides, optimising fragment prediction and integrating these steps in a pipeline are the main bioinformatics challenges as summarised in Reker and Malmström (2012). These are definitely usable across the board, irrespective of the application.

Finally, a crucial point remains that of data and processing quality. Indeed, data should be appropriately processed with robust software ensuring reproducible and accurate results. Even with robust software and optimised settings, low-quality data yield poor and questionable results. Reproducible and high confidence results strongly rely on software usability and the ability to choose the most appropriate parameters, since different parameters could lead to different results and wrong interpretations or conclusions. Within all the steps in the proteomics workflow, from sample collection to data processing, "Mount Bioinformatics" remains the last and the highest peak to climb (Aebersold 2009). The challenges of bioinformatics in software for quantitative proteomics have been previously described (Cappadona et al. 2012). Important issues that directly impact the effectiveness of proteomic quantitation and common tasks in computational solutions to correct for the occurrence of these factors are well depicted. This chapter surveys the different resources and software tools that are currently in use for data reference and analysis in the field of proteomics and tackles veterinary proteomics issues in this context either referring to published work or to prospective solutions.

2 Data Resources

As presented by Perez-Riverol et al. (2015), the information generated in proteomics experiments is organised in three levels:

(1) Raw MS data
(2) Processed experimental data
(3) Interpreted biological results

Level (1) corresponds to MS data collection through the ProteomeXchange protocol (Vizcaıno et al. 2014; Deutsch et al. 2017). In the past few years, this international initiative has allowed channelling all reported/published experimental data into three main repositories through standardised submission and dissemination pipelines.

Level (2) encompasses a significant number of databases storing peptide/protein identification and quantification whenever available.

Level (3) is associated with the concept of "knowledgebase" in which protein information is curated and recorded and can be searched, compared or mined.

2.1 Generic Proteomics Databases in a Nutshell

Several recent reviews comprehensively cover the topic of proteomics databases (Martens 2010; Perez-Riverol et al. 2015). Suffice to say that the collection of identified proteins is steadily growing and this broadens the extent of comparative or integrative approaches of data analysis. This is made possible through the generalised use of shared data formats and standards acknowledged by the Proteomics Standard Initiative (PSI; http://psidev.info). Table 1 summarises the range of formats in use in the current proteomics databases.

Martens and Vizcaıno (2017) very recently praised the "golden age" of proteomic data sharing precisely based on the availability and broad usage of standards.

2.1.1 Mass Spectrometry Data

File formats commonly used in MS-based proteomics are reviewed by Deutsch (2012). Mass spectra are stored in multiple repositories: PeptideAtlas (Farrah et al. 2013), GPMDB (Craig and Beavis 2004), Massive (https://massive.ucsd.edu) and PRIDE (Vizcaıno et al. 2015) to cite the most renowned. The vast majority of these publicly available datasets are generated for human and the main model organisms (*S. cerevisiae*, *D. melanogaster*, *C. elegans*, etc.). However, some dedicated resources have been created for farm animal proteomics as detailed in Sect. 2.2.1. One of the obvious advantages of such raw data availability is the potential for reuse and reanalysis. Reanalysis can be done individually, but note that PeptideAtlas and

Table 1 Formats in use in the current proteomics databases

Encoding purpose	Standard name	References
Mass spectra	mzXML	Pedrioli et al. (2004)
	mzML	Martens (2010)
Peptide/protein	pepXML/protXML	Keller et al. (2005)
identifications	mzIdentML	Jones et al. (2012)
Quantitative analysis	mzQuantML	Walzer et al. (2013)
SRM transitions	TraML	Deutsch et al. (2012)

GPMDB, for instance, routinely reprocess many datasets with in-house bioinformatics pipelines.

The second remarkable feature of these repositories is the option of defining a characteristic spectral library. Despite the existence of institutions such as the National Institute of Standards and Technology (NIST) where reference spectral libraries are collected, it may be of interest to constitute a specialised library. As mentioned in the introduction, the current shift from DDA to DIA approaches emphasises the need for quality spectral libraries.

It is important to stress that minimal quality checks are undertaken so that the resulting data resources contain uneven quality spectra and subsequent more or less reliable identifications of peptides and proteins.

2.1.2 Integrated Data

UniProt (www.uniprot.org) and protein data collected at NCBI (https://www.ncbi.nlm.nih.gov/protein) are the main sequence sources used in database search engines. UniProt however has a greater level of data integration and contains a wealth of information beyond sequence features including protein expression and structure. It is, as such, the most popular resource used for characterising proteins as finely as possible.

Recently proteogenomics has become a popular approach for merging information originating from genomics, transcriptomics and proteomics studies. Indeed, DNA sequence and RNA expression data accumulation over the past three decades provides rich sources to be combined with ever-increasing proteomics data. Proteogenomics currently significantly contributes to the identification of sequence variants, especially in humans, and the tools developed in this context could easily benefit farm animal proteomics studies. At the time of writing this chapter, no farm animal proteogenomics study has been published.

2.2 Farm Animal Proteomics Dedicated Databases

Farm animal dedicated databases are usually focused on genetic mapping information. Though apparently discontinued, ArkDB (http://www.thearkdb.org/arkdb) still hosts genome mapping data from farmed and other animal species spanning genetic linkage and QTL (quantitative trait locus). A related website called ResSpecies (http://www.resspecies.org/resspecies/) gathers tools for displaying or exporting genotyping and phenotype data as well as population coverage and markers. Along the same lines, QTLdb [http://www.animalgenome.org/QTLdb, (Hu et al. 2015)] collects publicly available trait mapping data for a smaller range of animal and is associated with CorrDB, the animal trait correlation database (http://www.animalgenome.org/cgi-bin/CorrDB). In order to complete this abundant genetic information, some effort was invested into developing proteomic-centred resources especially in capturing mass spectrometry data.

2.2.1 Farm Animal Mass Spectrometry Data

As mentioned in Sect. 2.1.1, mass spectrometry data repositories have recently imposed a paradigm shift in considering published data. PeptideAtlas pioneered in collecting datasets dedicated to farm animals though admittedly, these are not as often updated or enriched with new data as the human collection is, as shown in the homepage where all newly included sets are listed.

The frequent use of mass spectrometry in studying milk and its constituents naturally led to the first initiative devoted to collecting MS data in bovine milk and mammary gland (Bislev et al. 2012). Then the Equine PeptideAtlas (Bundgaard et al. 2014) and the Pig PeptideAtlas were introduced (Hesselager et al. 2016), and finally the PeptideAtlas for the domestic chicken (McCord et al. 2017) is the most recent addition to the collection. This data can be queried from the generic interface (“Queries” tab), and corresponding data can be downloaded.

2.2.2 Farm Animal Integrated Data

The success of proteomic-based investigations largely depends on the availability of complete and annotated databases containing the gene and protein sequence information for the animal species of interest (Soares et al. 2012). Nonetheless, once proteins are identified by matching mass spectrometry data with sequence data, according to various strategies detailed in Sect. 3, stored and in the best of cases integrated knowledge of protein properties is very useful to potentially rationalise the content of the studied sample. There are very few such resources for FAP, and the following three cover the range:

1. ProteINSIDE specifically supports the annotation of farm animal proteomics experiments (http://www.proteinside.org; Kaspric et al. 2015). It stores

information on bovine, sheep and goat. An in-built workflow processes lists of proteins input by a user to extract information on protein function, subcellular location (secreted/cytoplasmic) and interacting partners.

2. AgBase is an alternative resource for functional annotation though not proteomics-oriented (http://www.agbase.msstate.edu; McCarthy et al. 2010). It covers more farm animal organisms.
3. paxDB stores proteome-wide protein abundance information across organisms and tissues (http://pax-db.org; Wang et al. 2015b). Pig, bovine, horse and chicken are included.

3 Data Analysis Software

The use of two-dimensional gel electrophoresis (2DE) approaches combined with MS already allowed the characterisation of several distinct proteomes in different fields of animal science (Soares et al. 2012). However, the complexity of most proteomic samples challenges the separation power of such traditional techniques. More importantly, even with the current improvements in 2DE, it is still a manual and time-demanding process. In proteomics studies, liquid chromatography (LC) has been increasingly used as a replacement technique for gel electrophoresis, since it can be employed to analyse large numbers of samples in a faster, automated and more repeatable fashion. This trend can be observed in Fig. 1.

As one of the core technologies routinely used in advanced proteomics research, we focus this section on software and computational methods for analysing LC-MS data. We particularly describe the data and associated algorithms from a

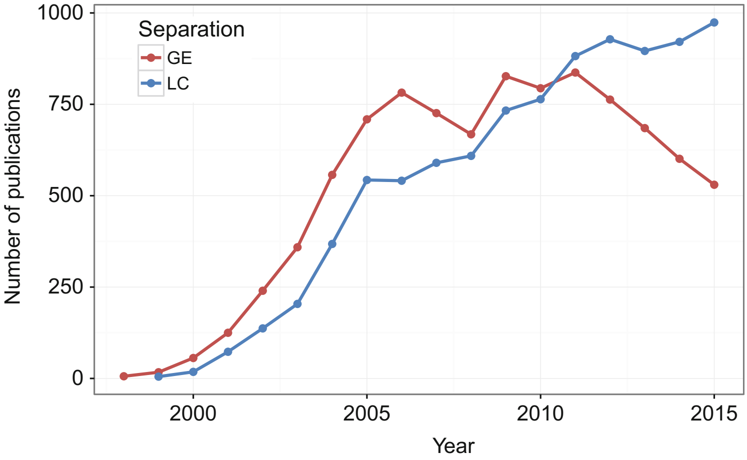

Fig. 1 Estimated number of MS-based proteomics publications per year: gel electrophoresis (GE, *red*) vs. liquid chromatography (LC, *blue*). PubMed queries: “proteomics gel electrophoresis mass spectrometry” and “proteomics liquid chromatography mass spectrometry”

perspective of the acquisition methods by which it is generated, because understanding the process of data generation may turn out to be critical for successful data analysis and result interpretation.

Prior to LC separation, complex protein samples are digested into peptides. Since the chemistry of peptides is more uniform than the chemistry of proteins, tandem MS methods to sequence peptides are nowadays robust and mature (Yates III 2015). The most popular enzyme used for protein digestion is trypsin, which leads to peptides with C-terminally protonated amino acids, providing an advantage in subsequent MS-based peptide sequencing (Aebersold 2003). As peptides elute from the LC column, they are ionised by electrospray ionisation (ESI), and resulting ions are analysed by the mass spectrometer.

The MS analysis can be performed with different instruments and different operation modes. Broadly speaking, they can be classified as targeted and shotgun methods. The characteristics of the spectra and informatics approaches for these methods are discussed in the following subsections.

As illustrated in Fig. 2, we emphasise examples of chromatogram-based MS quantification workflows using the Skyline software tool (MacLean et al. 2010), developed and maintained by the MacCoss lab. Skyline is a Windows client, versatile and robust platform that can be used to analyse the different types of MS data here described. A key feature of Skyline is the extraction of MS data directly from many instrument vendor formats, that is, a conversion to open file formats is usually not required. Skyline is freely available and open source, with an interactive and rather intuitive graphical user interface (GUI) for visualisation, and several tutorials describing the usage are available. Results can be exported as text reports that researchers can further process using other tools.

Moreover, several tools are available as plug-in modules within Skyline. A framework called "external tools" allows researchers to integrate their tools into Skyline without modifying the Skyline codebase (Broudy et al. 2014). With a uniform interface for installation into Skyline, the external tools can be easily accessed by all users for downstream statistical analyses.

3.1 Targeted MS Methods

Selected reaction monitoring (SRM)—also referred to as multiple reaction monitoring (MRM)—is a targeted MS technique whereby a predefined series of transitions (precursor/fragment ion pairs) are selected by the two mass filters of a triple quadrupole instrument and monitored over chromatographic elution for precise quantification (Lange et al. 2008; Picotti and Aebersold 2012).

Since SRM strictly targets a predetermined set of peptides, it is particularly useful when only a handful of proteins from a complex background, such as those constituting a cellular network or a set of candidate biomarkers, need to be measured across multiple samples in a consistent, reproducible and quantitatively precise manner.

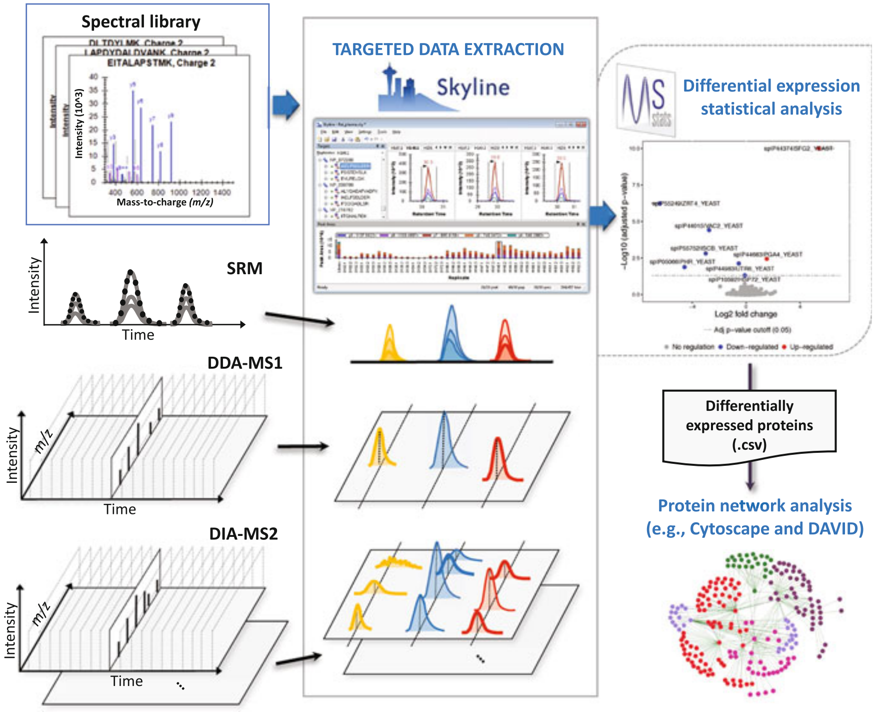

Fig. 2 Chromatogram-based MS quantification workflows in proteomics. Raw MS data generated by LC-MS analysis have different complexities depending on the purpose of the study. Targeted MS methods such as SRM require more effort prior to the actual LC-MS analysis, but the data contains single chromatogram traces for each targeted peptide transition and therefore requires less complex data processing. Shotgun MS methods acquire a full mass spectrum within each cycle. For DDA, MS1 data is used for quantification and MS2 data for identification. For DIA, the MS1 data is optional, and quantification is usually performed with the MS2 data, since the systematic fragmentation acquires continuous data for all fragments across the complete elution profile. The *m*/*z* range is typically divided in several precursor isolation windows, and thus DIA spectra contain several fragment ion maps. The spectral library provides information such as *m*/*z* and expected retention time used by Skyline as seeds for chromatogram extraction. The spectral library is not required for processing SRM data, but it might be used for designing the acquisition method, e.g., for selecting the best peptides and fragment ions. For DDA, the spectral library is used to extract the MS1 chromatograms for each identified peptide. For DIA, the spectral library is used to extract the MS2 chromatograms for the fragment ions of each identified peptide, from the spectra corresponding to the precursor isolation window. Statistical analysis to find differentially expressed proteins can be performed using MSstat, either as one of the external tools available within the Skyline GUI or independently using the exported results for the input of the R MSstat package. Further downstream analysis can be performed with tools such as Cytoscape and DAVID

A useful discussion of the best practices for targeted MS measurements in biology and medicine, also applicable in animal science, can be found in Carr et al. (2014). The authors discussed the analytical goals and the experimental evidence needed to properly describe developed assays according to the required levels of performance, as well as the computational and statistical tools useful for the analysis of targeted MS.

3.1.1 Design of Targeted Acquisition Methods

In SRM-based proteomics, a significant amount of time is required for the design of the acquisition method or assay, and several informatics tools are available to assist this process. For instance, software to select proteotypic peptides, transitions and best acquisition settings include SRMCollider (Röst et al. 2012), MRMOptimizer (Alghanem et al. 2017) and PREGO (Searle et al. 2015).

We highlight the PNNL Biodiversity Plugin as one of the external tools available in Skyline (Degan et al. 2016). The tool summarises available mass spectrometry data in a pathway-centric view and facilitates querying it from a biological perspective to design quantitative experiments. Selected proteins and their underlying mass spectra are imported to Skyline for further assay design (transition selection). The PNNL Biodiversity Library catalogues MS/MS spectra from over 3 million peptides and 230,000 proteins from 118 distinct organisms across the tree of life all cross-referenced to KEGG pathways for intuitive biological interpretation.

To maximise the number of peptides that can be monitored in a single LC-MS analysis, scheduled SRM methods can be designed (Stahl-Zeng et al. 2007). Using information of the expected peptide elution time in the target list, computer programs automatically generate SRM acquisition methods where the transitions of a specific peptide are only targeted during a time window around its elution time. In this way, the number of peptides measured in a LC-MS run is increased without compromising the limit of detection or the quantitative accuracy.

3.1.2 Data Processing for SRM Quantification

Unlike in other MS-based proteomic techniques, no full mass spectra are recorded in SRM analysis. SRM data consist of a set of chromatographic traces with the retention time and signal intensity for each of the monitored transition. This non-scanning nature translates into an increased sensitivity by one or two orders of magnitude compared with conventional “full-scan” techniques. The two levels of quadrupole mass selection with narrow mass windows result in a high selectivity, as co-eluting background ions are filtered out very effectively.

Integration of the chromatographic peaks for each transition supports the relative or, if suitable heavy isotope-labelled reference standards are used, absolute quantification of the targeted peptides, which are used as a surrogate measure of the proteins of interest. Isotope labelling increases the complexity and costs of an experiment with the benefit of more precise quantification.

Other targeted methods, in which high-resolution full MS/MS spectra are acquired for each target peptide, such as parallel reaction monitoring (PRM) (Peterson et al. 2012), generate data for all detectable fragment ions. The third quadrupole of a triple quadrupole is substituted with a high-resolution and accurate mass analyser.

3.2 Shotgun MS Methods

In contrast to targeted MS methods, the so-called shotgun or bottom-up approach does not require predefined information about the proteins of interest or analytes in the sample. The MS instrument is operated to record the spectra of many peptides as possible, as they elute from the LC column and are ionised by electrospray.

3.2.1 Data-Dependent Acquisition and MS1 Quantification

Within each DDA cycle, ion signals are recorded in a MS1 or survey scan (precursor ion signals), and the top-*N* most abundant ions are then selected and serially isolated for fragmentation (MS/MS, MS2, or tandem MS) to generate structural information. Since fragmentation models are well characterised for amino acid sequences, theoretical spectra can be generated according to factors including peptide sequence and type of fragmentation, typically collision-induced dissociation (CID). Typically, most of the MS/MS data is highly pure (each spectrum contains fragments from mainly one peptide) and can be annotated with peptide sequences using search tools such as Mascot (Perkins et al. 1999), SEQUEST (Eng et al. 1994), Andromeda (Cox et al. 2011), X!Tandem (Craig and Beavis 2004) and MS-GF+(Kim and Pevzner 2014). A search tool in silico digests the protein sequences into peptides and generates theoretical spectra to score the observed tandem mass spectra against the predicted fragmentation (Nesvizhskii 2010).

While search tools produce a match for almost every input MS/MS spectrum, only a fraction of those peptide to spectrum matches (PSMs) are true. The most commonly used and accepted statistical confidence measure is the false discovery rate (FDR) (Benjamini and Hochberg 1995) adapted in proteomics (Elias and Gygi 2007)—also known as "target-decoy approach" (TDA)—as a summary statistics for the entire collection of PSMs. A decoy database is generated by reversing or shuffling the amino acids in the sequences of the reference database and thus included in the search to estimate the FDR as the expected proportion of incorrect PSMs among all accepted PSM (Nesvizhskii 2010). FDR analysis is typically included within the identification software; however, it can also be performed and refined using other tools such as MAYU (Reiter et al. 2009), Percolator (Käll et al. 2007; The M et al. 2016), iProphet (Shteynberg et al. 2011), PeptideProphet (Keller et al. 2002; Choi and Nesvizhskii 2007) and ProteinProphet (Nesvizhskii et al. 2003).

Properly estimating and controlling the FDR are essential steps in the computational pipeline for preventing subtle but profound errors in high-throughput science.

It is necessary to place less emphasis on the number of identifications achieved and instead to value the work as a whole (Serang and Käll 2015).

After FRD analysis, confidently identified peptides can be used to build a library, which can be imported into Skyline to perform label-free relative quantification. Skyline extracts the precursor ion signals of each peptide from the MS1 raw data and computes the area under the peak or ion chromatogram from each peptide elution profile (Schilling et al. 2012).

Another popular method that has been traditionally used for label-free quantification in shotgun proteomics is the spectral count, which is based on the number of MS/MS spectra identified for each peptide sequence (Liu et al. 2004). Despite providing a rapid and semi-quantitative measure of abundance, spectral count-based quantification is affected by sample complexity; it has been found to often give irreproducible results and being unsuitable for quantifying low-abundance proteins (Cappadona et al. 2012; Ahrné et al. 2013). In contrast, chromatogram-based MS quantification based on integration of the peptide elution profile provides a level of accuracy comparable to labelling approaches.

3.2.2 Data-Independent Acquisition and MS2 Quantification

With recent developments in MS instrumentation, application of alternative MS operation modes such as DIA has become feasible (Chapman et al. 2014; Bilbao et al. 2015a). In contrast to DDA and by means of systematically parallelising the fragmentation, DIA avoids the selection of individual peptide ions during LC-MS analysis, therefore providing several advantages for characterising complex protein digests.

In a single injection or LC-MS analysis, DIA generates a comprehensive and permanent digital record of the sample (Liu et al. 2013). Because of the systematic sampling process, there is no need to reinject the sample for LC-MS analysis, as opposed to DDA and SRM. Acquired once and mineable forever, DIA spectra can be used to test for new hypothesis or reprocessed when a better-quality library is available (e.g. new genome available) or an updated or new processing software tool is released.

At the same time, in DIA, convoluted or multiplexed MS/MS spectra are generated without explicit association between each single precursor and its corresponding fragments. As a result, DDA search engines are not appropriate for processing DIA spectra, and several informatics tools have been developed recently to effectively process these complex datasets for identification: DIA-Umpire (Tsou et al. 2015, 2016), Group-DIA (Li et al. 2015) and MSPLIT-DIA (Wang et al. 2015a).

Here we focus on the targeted data extraction strategy related to the SWATH methodology (Gillet et al. 2012), where a list of peptide transitions (also called assay) built from previous DDA/SRM libraries is required. These libraries can be either collected from public repositories such as PeptideAtlas (www.peptideatlas.org), SRMAtlas (www.srmatlas.org) and SWATHAtlas (www.swathatlas.org) or generated by analysing the studied sample also in DDA mode to generate reference libraries (for a detailed protocol to generate high-quality reference libraries, see Schubert et al. (2015a)). A multiplexed variant of the SWATH methodology, termed MSX (Egertson et al. 2013, 2015), is also implemented within the Skyline software, supporting both acquisition method design and data processing for quantification.

As for MS1 quantification, Skyline extracts the ion signals of each peptide, but in this case fragment ion signals are extracted from the MS2 raw DIA files. Quantification is therefore performed using the elution profile of the peptide fragments, like for SRM, with the area under the peak or ion chromatogram as the abundance measure. Fragment ion abundances are subsequently aggregated into the corresponding peptides and proteins. A statistical measure of detection confidence is also computed for each peptide, using a similar version of the mProphet algorithm (Reiter et al. 2011) implemented within Skyline.

Other software tools for targeted data extraction are also available: Spectronaut (Bernhardt et al. 2012) (Biognosys proprietary software with free license for academics) and OpenSWATH (Röst et al. 2014) (open-source standalone tool or integrated module into the proteomics software OpenMS) (Sturm et al. 2008; Röst et al. 2016), including tools without GUI such as DIANA (Teleman et al. 2014) and SWATHProphet (Keller et al. 2015).

Recently, the performance of several of these tools was compared (Navarro et al. 2016). The authors observed similar reliable performances after software and parameter optimisation and concluded that targeted data extraction is a valid alternative to isotope-labelling-based methods.

Another consideration related to DIA data processing is the fact that the concurrent fragmentation of peptides has the drawback of increasing the likelihood of interference due to the overlap of fragment ions from different precursors. Several computational strategies can tackle this issue to further expand the benefits of DIA (Zhang et al. 2015; Bilbao et al. 2015b, 2016).

3.3 Statistical Analysis of Quantitative Results

Based on the quantification data, the next step is to determine candidate proteins showing significant differences across several sample types or conditions. The R statistical package MSstats (Choi et al. 2014) can be used as a standalone or as one of the external tools available within Skyline. MSstats can be used to interpret SRM, DDA and DIA quantification results.

The external set of QuaSAR tools (Abbatiello et al. 2010; Mani et al. 2012) automate and assist quantification of stable isotope dilution experiments. QuaSAR produces tabulated results for every peptide for essential statistics such as coefficient of variation, regression slope and intercept (with confidence intervals) and limits of detection and quantification as well as figures summarising their distribution and variation.

3.4 *Automated PTM Detection*

It is now well established that posttranslational modifications (PTMs) act in isolation or in combination with proteins for modulation and regulation purposes. In recent years, this field of investigation has led to intense bioinformatics development.

3.4.1 Main PTM Bioinformatics Resources

Two main databases are considered as references for storing PTM-related information. UniMod (http://www.unimod.org) is a comprehensive list of protein modifications for mass spectrometry applications. dbPTM (http://dbPTM.mbc.nctu.edu.tw) describes substrate specificity of PTM sites and provides functional annotation of PTM-related substrates and known interacting proteins. Neither specifically distinguishes between species as both aim at increasing numbers of their respective statistical tables. However, with this concern for broad coverage, dbPTM maintains a comprehensive list of databases and prediction software dedicated to individual or groups of PTMs. To complement the summary information associated with each resource, useful and more detailed comments can be found, for instance, in Kamath et al. (2011).

Over the past decades, the accumulation of sequence data led to implementation of PTM site prediction software based on amino acid patterns in aligned sequences. Each method was usually designed to identify individual PTMs. Many of these methods were developed with web interfaces and are hosted on the ExPASy (www.expasy.org/proteomics/post-translational_modification) and the CBS (www.cbs.dtu.dk/databases/PTMpredictions) servers whose creators pioneered in this field. More recently, the accumulation of mass spectrometry data to support PTM detection contributed to refining the reliability of prediction based on more comprehensive experimental data. This is, for instance, the case of phosphorylation sites through the use of resources such as PhosphoSite (http://www.phosphosite.org) that collects published mass spectrometry data for site annotation. In fact, understanding PTM occurrence goes along with studying the corresponding modifying enzyme(s). In many cases, these enzymes are not known, or their target is not precisely defined. For phosphorylation, KinBase (http://kinase.com/kinbase/) is the kinome reference, and the combined use of MS-validated sites and kinase specificity helps in refining site prediction as further explained in Sect. 3.4.2.

Although not fully considered as a PTM, protein cleavage should nonetheless be part of the modification landscape, and proteases have long been collected and classified in the MEROPS database (http://merops.sanger.ac.uk). The connection between proteolysis and PTMs is brought out in TOPFind, the N-/C-terminal modification database (http://clipserve.clip.ubc.ca/topfind). With a strong focus on human and mouse data, TOPFind also attempts to merge protein cleavage information with protein-protein interactions with the PathFinder tool (Fortelny et al. 2014) through mapping and modelling a protease interaction network.

Nonetheless, at present, besides published articles, UniProt remains the main source compiling information on alternative effectors of PTMs. In fact, examples of protein-protein interaction networks integrating PTM knowledge (occurrence + specific effector) are rare. They tend to be devoted to mapping data collected in eukaryotes such the yeast methylome (Erce et al. 2012) or the phospho-tyrosine interaction network in human (Grossmann et al. 2015).

3.4.2 Predicting PTM and Their Associated Enzymes

The association between a phosphorylated site and the kinase that actually performed the attachment on a serine, threonine or tyrosine residue is far from being obvious to predict, despite the clear need for getting a fuller picture of phosphorylation. So far, the most known tool that combines several sources of data to suggest site-enzyme associations for phosphorylation is networKIN (http://networkin.info). The method first uses a predictor to label a given phosphosite sequence with a kinase or kinase family. This predictor is trained with experimental data to ascertain the relationship between a site and a kinase. For instance, Scansite (http://scansite3.mit.edu) mostly relies on peptide library screening, phage display and mass spectrometry experiments to get enough examples of labelled sites and identify characteristic sequence patterns for a site in definite association with a known kinase. Nonetheless a high level of ambiguity persists and as a second step in order to narrow down the options, networKIN includes contextual information by extracting knowledge of protein-protein interactions centred on the kinases of interest from a database of interactions. By calculating the proximity of the substrate to all kinases in a network of functional relationships, networKIN infers the most likely candidate kinase for each site.

The knowledge of phosphorylation is by far more advanced than that of other common PTMs such as glycosylation. Despite the possible mapping between a glycan structure and the set of enzymes that are required for its synthesis, the characterisation of intact glycopeptides remains a definite challenge. As it is, most glycan structures have been solved after being cleaved off their natural support, while protein glycosylation sites are identified after removing the attached glycans. In the end, key information on the glycoconjugate is lost. The correlation between glycan structures and glycoproteins can be restored manually through literature searches that are both labour- and time-consuming. This is, for example, the purpose of UniCarbKB (www.unicarbkb.org). In this context, the design of prediction tools linking a glycosite with the appropriate glycosyltransferases may happen in a not too distant future. Information on these enzymes has accumulated in the CAZy database (www.cazy.org) over the past two decades.

3.4.3 PTM Discovery

Mass spectrometry is the method of choice for detecting PTMs. Since the early days of software development for analysing mass spectra, the concern for identifying possible mass shifts corresponding to the addition or removal of chemical groups of known masses has been shared by bioinformaticians. Then it appeared that, conversely, the occurrence of regular and identical mass shifts of unknown origin in MS2 data could be a source of new knowledge, and a range of tools was then developed to perform the so-called open modification search, that is, PTM search with no a priori (Ahrné et al. 2010). This approach was scaled to process high-throughput proteome data with the prospect of discovering unexpected modifications (Na et al. 2011; Horlacher et al. 2015). However, data interpretation remains a challenge, and findings require experimental validation. Nonetheless, large-scale processing supports scientists in investigating and discovering new leads.

It is worth noting that the top-down proteomic approach is very promising for generating mass information on PTMs (Smith et al. 2013). The ProSight software that is commonly run to analyse this particular type of data is adapted to the identification of a broad range of PTMs (Fellers et al. 2015).

Recently, the potential of label-free by PRM was demonstrated for targeted phosphoproteome analysis (Lawrence et al. 2016). The authors also created a web-based assay development application that queries the database for optimal peptide selection and retention time scheduling (phosphopedia.gs.washington.edu).

3.4.4 PTM Combination

The next challenge is to identify the constraints that rule PTM cooperative and/or antagonist effects as pointed in Venne et al. (2014). Indeed, PTMs can be mutually exclusive such as the phosphorylation and the O-GlcNAcylation of serine and threonine of signalling proteins presented very early on as the “yin-yang hypothesis” (Hart et al. 1995), but they can also be cooperative as in the well-studied case of histones (Schwammle et al. 2014). Data has accumulated on a few other proteins such as tubulins (Verhey and Gaertig 2007), the FoxO regulator (Calnan and Brunet 2008) or chaperones (Cloutier and Coulombe 2013). However, a critical mass of experimental data is still missing to design appropriate bioinformatics tools supporting the discovery of PTM co-occurrence rules that could potentially explain corresponding effects on protein function. In the meantime, simple co-occurrence is collected, for instance, in the PTMCode database (Minguez et al. 2014) that shows potential trends. This field is likely to blossom in the years to come.

3.5 Interactomics and Data Interpretation

A list of identified proteins is not sufficient for characterising a sample. The challenge is to understand why some proteins are co- or differentially expressed and what are the underlying processes explaining cooperative or concurrent activity.

Detailed information about how MS-based proteomics has been applied to network biology, to detect and quantify perturbation-induced network changes and to correlate network dynamics with cellular phenotypes can be found in Bensimon et al. (2012).

The most common approach is to extract protein information from the gene ontology (GO) that assigns relevant terms from controlled vocabularies to specify the protein subcellular location, its function and its contribution to one or more processes (see www.geneontology.org). Then, when proteins share terms characterising a location or a function or a process, this type of similarity helps in shaping a hypothesis. Alternatively, the knowledge of protein-protein interactions stored in databases such as IntAct (http://www.ebi.ac.uk/intact/) or BioGrid (http://www.biogrid.org) can be used to build an interaction network. Note that readily available interaction networks can be visualised and queried in STRING (http://string-db.org). In all cases, integrated tools in the cited data resources provide support for protein data interpretation. One of the most popular open-source software for integration, visualisation and analysis of biological networks is Cytoscape (Shannon et al. 2003). This popularity has been driven by the ability of extending Cytoscape functionality through plugins (Saito et al. 2012), yielding a powerful and heterogeneous set of tools and enabling a broad community of scientists to contribute. This part is illustrated further in the following section with concrete examples.

4 Applications

There are roughly three types of applications where high-throughput proteomics methods are used and therefore bioinformatics is necessary. In most instances detailed below, resources tailored for the study of model organisms spanning mainly humans, mouse, drosophila and yeast show some limitation in supporting the interpretation of experimental results.

4.1 Animal Health

In much the same way as in medical studies, one of the goals of using proteomics is to identify reliable disease biomarkers for diagnosis or prognosis. For example, serum protein profiles have been used to detect infectious disease in pigs (Koene et al. 2012). However, the prevalence of genomics approaches remains as illustrated in a recent comprehensive review on bioinformatics tools available to study parasites of veterinary significance (Cantacessi et al. 2012). Omics data integration is still a prospect at this point in time.

As highlighted in Sect. 3.5, interactomics provides the most attractive ground for the interpretation of proteomics data. The identification of protein complexes in a sample is a first step in rationalising and understanding protein co-expression. For example, cellular proteins of the host-forming complexes with specific proteins of the porcine reproductive and respiratory syndrome virus (PRRSV) essential in virus replication were identified by pull-down experiments in Dong et al. (2016). Binding partners of viral proteins were then systematically mapped on known pathways with a piece of software (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) to shortlist those that involve the mostly expressed partners. This study along with further experimental work led to identify and validate the role of the HSP70 chaperone as key to transcription and replication of PRRSV. More generally, the investigation of host-pathogen interactions is obviously suited to proteomics-based approaches, and the wealth of information stored in bioinformatics databases provides useful support. Proteomics is increasingly introduced in veterinary medicine (Ceciliani et al. 2014), and data accumulation will soon popularise the use of bioinformatics to allow for information-rich comparative studies.

Furthermore, Bundgaard et al. (2016) applied SRM to investigated levels of eight inflammatory acute phase proteins in interstitial fluid from wounds in horses. Selection of protein-specific peptides was performed using the equine PeptideAtlas website.

Packialakshmi et al. investigated proteomics differences in the plasma of healthy and femoral head necrosis-affected chickens using shotgun MS methods (Packialakshmi et al. 2016). MS/MS data was converted from proprietary format to mzXML files using the instrument vendor software and submitted to global proteome machine (GPM; http://www.thegpm.org) for identification with X! Tandem.

Proteins with at least one unique peptide and 5% FDR were considered true for protein identifications. The results were downloaded as *.xml files for Skyline software. After MS1 signal extraction, label-free quantitation was performed using MSstats as one of the external tools directly available in Skyline. Group comparison function was used for the label-free quantitation and to generate the volcano plot that shows the differentially expressed proteins.

The list of proteins was mapped to the corresponding ensemble gene IDs using Biomart and analysed for relative enrichment, clustering and GO annotations using DAVID (Huang et al. 2009a, b).

4.2 Adipose Tissue and Muscle Studies

The lack of farm animal dedicated resources in proteomics motivated the development of the ProteINSIDE database cited in Sect. 2.2.2. It was used in a study of bovine adipogenesis and myogenesis as well as the balance between these two processes (Kaspric et al. 2015). Previously analysed mass spectrometry data of adipose (Taga et al. 2012) and muscle (Chaze et al. 2009) foetal bovine tissues led to identify proteins, which were poorly annotated. Data and tool integration of ProteINSIDE supported the interpretation of protein lists. In particular protein-protein interactions, as collected from various sources as cited in Sect. 3.5, were used as the main piece of information for identifying clusters of functionally interconnected proteins. In the muscle they were associated with four processes, muscle development, cell proliferation, energetic complex and respiratory chain, and in the adipose tissue with seven, cell proliferation, proteasome complex, complexes I and III of the respiratory chain, redox activity and differentiation and metabolism of adipose tissue. The overlap between the two tissue types led to suggest possible crosstalk mechanisms.

Other authors have used bioinformatics sequence analysis tools to identify proteotypic peptides in an attempt to define biomarkers of meat authenticity following a targeted proteomics approach (Orduna et al. 2015). In the same vein, Stella and co-workers used SRM to quantify 12 potential protein markers of skeletal muscle and detect anabolic treatments with dexamethasone (Stella et al. 2016). The listed proteins were markers identified in a previous study applying a two-dimensional difference gel electrophoresis proteomics approach. A scheduled SRM method was developed using Skyline software to monitor 24 signature peptides from the 12 considered protein markers (two peptides per protein). For each peptide, 3 precursor-to-product ion transitions were targeted. Peptide quantification was achieved using a spike-in dedicated internal standard for each target. To this end, 13C/15N isotopically labelled peptides sharing the same sequence but with a defined mass shift were used. Peptide quantification was achieved using Skyline software integrating the area of the chromatographic peak of each peptide and the corresponding labelled internal standard.

Using R, protein abundances were graphically described using box plots, and potential differences of protein concentration values among different animal groups were explored performing one-way analysis of variance (ANOVA) on the two animal sets.

4.3 Milk Proteome and Glycoproteome

As mentioned in Sect. 2.2.1, the analysis of protein content in farm animal milk has long been the focus of veterinary and biological science. It is not surprising that several groups already undertook global studies such as a more detailed pathway mapping of 106 human milk proteins (D'Alessandro et al. 2010). In this in silico study, pathway analysis software based on knowledge of metabolism is used to identify cell proliferation and differentiation pathways on top of the usual nutrition and immune functions known to characterise milk proteins. This provides evidence of tissue growth and organ development capacities of milk proteins based on collecting public data and using pathway analysis tools.

To account for quantitative aspects, mass spectrometry and bioinformatics tools were used in two recent studies (Tacoma et al. 2016; Zhang et al. 2016). While the former reference conventionally relies on gene ontology to track functional features of differentially expressed proteins in two dairy cow breeds, the latter brings the comparison of the human and bovine milk proteomes over lactation further. In this study, MS data was first differentially quantified using the MaxQuant software, and interactions between the most co-expressed proteins were then derived from STRING (see Sect. 3.5). Results reveal the interconnected roles of milk proteins in nutrition and protection to the neonate.

Finally, the importance of glycosylation in milk needs to be acknowledged, and quantitative studies are also undertaken in a systematic way. For example, Huang et al. (2016) used SRM to quantitate seven human milk proteins and their glycoforms.

5 Conclusion

The application of proteomics in veterinary studies has been moving from the initial qualitative description towards the quantification stage, where experiments to identify and quantify protein changes in different samples of particular tissues or fluids become more common (Ceciliani et al. 2014). Application of untargeted label-free quantification methods is rapidly increasing in the proteomics field in general, fuelled by the advantages of DIA methods and development of software tools and innovative algorithms, which already have been shown implementations to estimate absolute cellular protein concentrations (Schubert et al. 2015b). In this context, further improvements and developments of new computational strategies for quantification are expected.

Numerous studies have been published to date in domestic and farm animal proteomics; among several challenges and limitations, we highlight the lack of detailed information of MS-based proteome informatics tools in the context of farm animal reviews/resources. Only general notions are described but not specific tools and very few references.

The previously described difficulty to use MS proteomics software (Cappadona et al. 2012), associated with the lack of appropriate documentation or with a poor graphical user interface, is still an issue. Fortunately, this is changing with tools like Skyline. Skyline is an active project and continues expanding, for instance, it currently supports chromatogram extraction from MS1 and MS2 spectra with the additional ion mobility separation recently available in commercial instruments

(Baker et al. 2015). An increasing number of available tools within open-source and collaborative projects can be expected, and we encourage development of more robust MS tools that can be used by researchers nonspecialised in mass spectrometry or informatics.

References

Abbatiello SE, Mani D, Keshishian H, Carr SA (2010) Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem 56:291–305

Aebersold R (2003) Mass spectrometry-based proteomics. Nature 422:198–207

Aebersold R (2009) A stress test for mass spectrometry-based proteomics. Nat Methods 6:411–412

Ahrné E, Müller M, Lisacek F (2010) Unrestricted identification of modified proteins using MS/MS. Proteomics 10:671–686

Ahrné E, Molzahn L, Glatter T, Schmidt A (2013) Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics 13:2567–2578

Alghanem B, Nikitin F, Stricker T, Duchoslav E, Luban J, Strambio-De-Castillia C, Muller M, Lisacek F, Varesio E, Hopfgartner G (2017) Optimization by infusion of multiple reaction monitoring transitions for sensitive peptides LC-MS quantification. Rapid Commun Mass Spectrom 31(9):753–761

Baker ES, Burnum-Johnson KE, Ibrahim YM, Orton DJ, Monroe ME, Kelly RT, Moore RJ, Zhang X, Théberge R, Costello CE et al (2015) Enhancing bottom-up and top-down proteomic measurements with ion mobility separations. Proteomics 15:2766–2776

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300

Bensimon A, Heck AJ, Aebersold R (2012) Mass spectrometry-based proteomics and network biology. Annu Rev Biochem 81:379–405

Bernhardt OM, Selevsek N, Gillet LC, Rinner O, Picotti P, Aebersold R, Reiter L (2012) Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. In: 60th ASMS Conference on mass spectrometry and allied topics. Vancouver, Canada, pp 20–24

Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, Müller M, Lisacek F (2015a) Processing strategies and software solutions for data-independent acquisition in mass spectrometry. Proteomics 15:964–980

Bilbao A, Zhang Y, Varesio E, Luban J, Strambio-De-Castilla C, Lisacek F, Hopfgartner G (2015b) Ranking fragment ions based on outlier detection for improved label-free quantification in data-independent acquisition LC-MS/MS. J Proteome Res 14:4581–4593

Bilbao A, Lisacek F, Hopfgartner G (2016) Dedicated software enhancing data-independent acquisition methods in mass spectrometry. CHIMIA Int J Chem 70:293–293

Bislev SL, Deutsch EW, Sun Z, Farrah T, Aebersold R, Moritz RL, Bendixen E, Codrea MC (2012) A Bovine PeptideAtlas of milk and mammary gland proteomes. Proteomics 12:2895–2899

Broudy D, Killeen T, Choi M, Shulman N, Mani DR, Abbatiello SE, Mani D, Ahmad R, Sahu AK, Schilling B et al (2014) A framework for installable external tools in Skyline. Bioinformatics 30:2521–2523

Bundgaard L, Jacobsen S, Sørensen MA, Sun Z, Deutsch EW, Moritz RL, Bendixen E (2014) The Equine PeptideAtlas: a resource for developing proteomics-based veterinary research. Proteomics 14:763–773

Bundgaard L, Bendixen E, Sørensen MA, Harman VM, Beynon RJ, Petersen LJ, Jacobsen S (2016) A selected reaction monitoring-based analysis of acute phase proteins in interstitial fluids from experimental equine wounds healing by secondary intention. Wound Repair Regen 24:525–532

Calnan DR, Brunet A (2008) The FoxO code. Oncogene 27:2276–2288

Cantacessi C, Campbell BE, Jex AR, Young ND, Hall RS, Ranganathan S, Gasser RB (2012) Bioinformatics meets parasitology. Parasite Immunol 34:265–275

Cappadona S, Baker PR, Cutillas PR, Heck AJ, van Breukelen B (2012) Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino Acids 43:1087–1108

Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, Deutsch EW, Grant RP, Hoofnagle AN, Hüttenhain R, Koomen JM et al (2014) Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. Mol Cell Proteomics 13:907–917

Ceciliani F, Eckersall D, Burchmore R, Lecchi C (2014) Proteomics in veterinary medicine. Vet Pathol 51:351–362

Chapman JD, Goodlett DR, Masselon CD (2014) Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom Rev 33:452–470

Chaze T, Meunier B, Chambon C, Jurie C, Picard B (2009) Proteome dynamics during contractile and metabolic differentiation of bovine foetal muscle. Animal 3:980–1000

Choi H, Nesvizhskii AI (2007) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 7:254–265

Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, Vitek O (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30:2524–2526

Cloutier P, Coulombe B (2013) Regulation of molecular chaperones through post-translational modifications: decrypting the chaperone code. Biochim Biophys Acta Gene Regul Mech 1829:443–454

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794–1805

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467

D'Alessandro A, Scaloni A, Zolla L (2010) Human milk proteins: an interactomics and updated functional overview. J Proteome Res 9:3339–3373

Degan MG, Ryadinskiy L, Fujimoto GM, Wilkins CS, Lichti CF, Payne SH (2016) A skyline plugin for pathway-centric data browsing. J Am Soc Mass Spectrom 27:1752–1757

Deutsch EW (2012) File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 11:1612–1621

Deutsch EW, Chambers M, Neumann S, Levander F, Binz P-A, Shofstahl J, Campbell DS, Mendoza L, Ovelleiro D, Helsens K et al (2012) TraML—a standard format for exchange of selected reaction monitoring transition lists. Mol Cell Proteomics 11:111–15040

Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, Campbell DS, Bernal-Llinares M, Okuda S, Kawano S et al (2017) The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res 45: D1100–D1106

Dong S, Liu L, Wu W, Armstrong SD, Xia D, Nan H, Hiscox JA, Chen H (2016) Determination of the interactome of non-structural protein12 from highly pathogenic porcine reproductive and respiratory syndrome virus with host cellular proteins using high throughput proteomics and identification of HSP70 as a cellular factor for virus replication. J Proteomics 146:58–69

Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V et al (2013) Multiplexed MS/MS for improved data-independent acquisition. Nat Methods 10:744–746

Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ (2015) Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat Protoc 10:887–903

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214

Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5:976–989

Erce MA, Pang CNI, Hart-Smith G, Wilkins MR (2012) The methylproteome and the intracellular methylation network. Proteomics 12:564–586

Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang C-Y, Moritz RL (2013) The state of the human proteome in 2012 as viewed through PeptideAtlas. J Proteome Res 12:162–171

Fellers RT, Greer JB, Early BP, Yu X, LeDuc RD, Kelleher NL, Thomas PM (2015) ProSight Lite: graphical software to analyze top-down mass spectrometry data. Proteomics 15:1235–1238

Fortelny N, Cox JH, Kappelhoff R, Starr AE, Lange PF, Pavlidis P, Overall CM (2014) Network analyses reveal pervasive functional regulation between proteases in the human protease web. PLoS Biol 12:e1001869

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. https://doi.org/10.1074/mcp.O111.016717

Griss J (2016) Spectral library searching in proteomics. Proteomics 16:729–740

Grossmann A, Benlasfer N, Birth P, Hegele A, Wachsmuth F, Apelt L, Stelzl U (2015) Phospho-tyrosine dependent protein-protein interaction network. Mol Syst Biol 11:794–794

Hart GW, Greis KD, Dong L-YD, Blomberg MA, Chou T-Y, Jiang M-S, Roquemore EP, Snow DM, Kreppel LK, Cole RN, Comer FI, Arnold CS, Hayes BK (1995) O-linked N-Acetylglucosamine: the "yin-yang" of ser/Thr Phosphorylation? In: Alavi A, Axford JS (eds) Glycoimmunology, Advances in experimental medicine and biology. Springer Nature, Boston, MA, pp 115–123

Hesselager MO, Codrea MC, Sun Z, Deutsch EW, Bennike TB, Stensballe A, Bundgaard L, Moritz RL, Bendixen E (2016) The Pig PeptideAtlas: a resource for systems biology in animal production and biomedicine. Proteomics 16:634–644

Horlacher O, Lisacek F, Müller M (2015) Mining large scale tandem mass spectrometry data for protein modifications using spectral libraries. J Proteome Res 15:721–731

Hu Z-L, Park CA, Reecy JM (2015) Developmental progress and current status of the animal QTLdb. Nucleic Acids Res 44:D827–D833

Huang DW, Sherman BT, Lempicki RA (2009a) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57

Huang DW, Sherman BT, Lempicki RA (2009b) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13

Huang J, Kailemia MJ, Goonatilleke E, Parker EA, Hong Q, Sabia R, Smilowitz JT, German JB, Lebrilla CB (2016) Quantitation of human milk proteins and their glycoforms using multiple reaction monitoring (MRM). Anal Bioanal Chem 409:589–606

Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, Selley JN, Searle BC, Shofstahl J, Seymour SL et al (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics 11:111–14381

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4:923–925

Kamath KS, Vasavada MS, Srivastava S (2011) Proteomic databases and tools to decipher post-translational modifications. J Proteomics 75:127–144

Kaspric N, Picard B, Reichstadt M, Tournayre J, Bonnet M (2015) ProteINSIDE to easily investigate proteomics data from ruminants: application to mine proteome of adipose and muscle tissues in bovine foetuses. PLoS One 10:e0128086

Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392

Keller A, Eng J, Zhang N, Li X, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1(2005):0017

Keller A, Bader SL, Shteynberg D, Hood L, Moritz RL (2015) Automated validation of results and removal of fragment ion interferences in targeted analysis of data independent acquisition MS using SWATHProphet. Mol Cell Proteomics. https://doi.org/10.1074/mcp.O114.044917

Kim S, Pevzner PA (2014) MS-GF + makes progress towards a universal database search tool for proteomics. Nat Commun. https://doi.org/10.1038/ncomms6277

Koene MG, Mulder HA, Stockhofe-Zurwieden N, Kruijt L, Smits MA (2012) Serum protein profiles as potential biomarkers for infectious disease status in pigs. BMC Vet Res 8:32

Lange V, Picotti P, Domon B, Aebersold R (2008) Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 4:1–14

Lawrence RT, Searle BC, Llovet A, Villén J (2016) Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nat Methods 13(5):431–434

Li Y, Zhong C-Q, Xu X, Cai S, Wu X, Zhang Y, Chen J, Shi J, Lin S, Han J (2015) Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat Methods. https://doi.org/10.1038/NMETH.3593

Liu H, Sadygov RG, Yates JR III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201

Liu Y, Hüttenhain R, Collins B, Aebersold R (2013) Mass spectrometric protein maps for biomarker discovery and clinical research. Expert Rev Mol Diagn 13:811–825

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–968

Mani D, Abbatiello SE, Carr SA (2012) Statistical characterization of multiple-reaction monitoring mass spectrometry (MRM-MS) assays for quantitative proteomics. BMC Bioinformatics 13:S9

Martens L (2010) Proteomics databases and repositories. In: Methods in molecular biology. Springer Nature, Boston, MA, pp 213–227

Martens L, Vizcaıno JA (2017) A golden age for working with public proteomics data. Trends Biochem Sci 42(5):333–341

McCarthy FM, Gresham CR, Buza TJ, Chouvarine P, Pillai LR, Kumar R, Ozkan S, Wang H, Manda P, Arick T, Bridges SM, Burgess SC (2010) AgBase: supporting functional modeling in agricultural organisms. Nucleic Acids Res 39:D497–D506

McCord J, Sun Z, Deutsch EW, Moritz RL, Muddiman DC (2017) The PeptideAtlas of the domestic laying Hen. J Proteome Res 16:1352–1363

Minguez P, Letunic I, Parca L, Garcia-Alonso L, Dopazo J, Huerta-Cepas J, Bork P (2014) PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. Nucleic Acids Res 43:D494–D502

Na S, Bandeira N, Paek E (2011) Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11:M111.010199

Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, Tate SA, Tsou C-C, Reiter L, Distler U et al (2016) A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol 34(11):1130–1136

Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 73:2092–2123

Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658

Orduna AR, Husby E, Yang CT, Ghosh D, Beaudry F (2015) Assessment of meat authenticity using bioinformatics, targeted peptide biomarkers and high-resolution mass spectrometry. Food Addit Contam Part A 32:1709–1717

Packialakshmi B, Liyanage R, Jackson O, Lay J, Okimoto R, Rath NC (2016) Proteomic changes in the plasma of broiler chickens with femoral head necrosis. Biomark Insights 11:55–62

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R et al (2004) A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 22:1459–1466

Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcauno JA (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 15:930–950

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol Cell Proteomics 11:1475–1478

Picotti P, Aebersold R (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9:555–566

Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 8:2405–2417

Reiter L, Rinner O, Picotti P, Hüttenhain R, Beck M, Brusniak M-Y, Hengartner MO, Aebersold R (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods 8:430–435

Reker D, Malmström L (2012) Bioinformatic challenges in targeted proteomics. J Proteome Res 11:4393–4402

Röst H, Malmström L, Aebersold R (2012) A computational tool to detect and avoid redundancy in selected reaction monitoring. Mol Cell Proteomics 11:540–549

Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinovi SM, Schubert OT, Wolski W, Collins BC, Malmström J, Malmström L et al (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol 32:219–223

Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich H-C, Gutenbrunner P, Kenar E et al (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13:741–748

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to Cytoscape plugins. Nat methods 9:1069–1076

Schilling B, Rardin MJ, MacLean BX, Zawadzka AM, Frewen BE, Cusack MP, Sorensen DJ, Bereman MS, Jing E, Wu CC et al (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in Skyline application to protein acetylation and phosphorylation. Mol Cell Proteomics 11:202–214

Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, Lam H, Amodei D, Mallick P, MacLean B et al (2015a) Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc 10:426–441

Schubert OT, Ludwig C, Kogadeeva M, Zimmermann M, Rosenberger G, Gengenbacher M, Gillet LC, Collins BC, Röst HL, Kaufmann SHE, Sauer U, Aebersold R (2015b) Absolute proteome composition and dynamics during dormancy and resuscitation of mycobacterium tuberculosis. Cell Host Microbe 18:96–108

Schwammle V, Aspalter C-M, Sidoli S, Jensen ON (2014) Large scale analysis of co-existing post-translational modifications in histone tails reveals global fine structure of cross-talk. Mol Cell Proteomics 13:1855–1865

Searle BC, Egertson JD, Bollinger JG, Stergachis AB, MacCoss MJ (2015) Using data independent acquisition (DIA) to model high-responding peptides for targeted proteomics experiments. Mol Cell Proteomics 14:2331–2340

Serang O, Käll L (2015) Solution to statistical challenges in proteomics is more statistics, not less. J Proteome Res. https://doi.org/10.1021/acs.jproteome.5b00568

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 10:111–7690

Smith LM, Kelleher NL et al (2013) Proteoform: a single term describing protein complexity. Nat Methods 10:186–187

Soares R, Franco C, Pires E, Ventosa M, Palhinhas R, Koci K, de Almeida AM, Coelho AV (2012) Mass spectrometry and animal science: protein identification strategies and particularities of farm animal species. J Proteomics 75:4190–4206

Stahl-Zeng J, Lange V, Ossola R, Eckhardt K, Krek W, Aebersold R, Domon B (2007) High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. Mol Cell Proteomics 6:1809–1817

Stella R, Barrucci F, Angeletti R, James P, Montesissa C, Biancotto G (2016) Targeted proteomics for the indirect detection of dexamethasone treatment in bovines. Anal Bioanal Chem 408:8343–8353

Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K et al (2008) OpenMS–an open-source software framework for mass spectrometry. BMC Bioinform 9:163

Tacoma R, Fields J, Ebenstein DB, Lam Y-W, Greenwood SL (2016) Characterization of the bovine milk proteome in early-lactation Holstein and Jersey breeds of dairy cows. J Proteomics 130:200–210

Taga H, Chilliard Y, Meunier B, Chambon C, Picard B, Zingaretti MC, Cinti S, Bonnet M (2012) Cellular and molecular large-scale features of fetal adipose tissue: Is bovine perirenal adipose tissue Brown1685. J Cell Physiol 227:1688–1700

Teleman J, Röst H, Rosenberger G, Schmitt U, Malmström L, Malmström J, Levander F (2014) DIANA-algorithmic improvements for analysis of data-independent acquisition MS data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu686

The M, MacCoss MJ, Noble WS, Käll L et al (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. J Am Soc Mass Spectrom 27:1719–1727

Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras A-C, Nesvizhskii AI (2015) DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics. Nat Methods 12:258–264

Tsou C-C, Tsai C-F, Teo G, Chen Y-J, Nesvizhskii AI (2016) Untargeted, spectral library-free analysis of data independent acquisition proteomics data generated using Orbitrap mass spectrometers.16(15–16):2257–2271

Venne AS, Kollipara L, Zahedi RP (2014) The next level of complexity: crosstalk of posttranslational modifications. Proteomics 14:513–524

Verhey KJ, Gaertig J (2007) The Tubulin Code. Cell Cycle 6:2152–2160

Vizcaıno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rıos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz P-A, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus H-J, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32:223–226

Vizcaıno JA, Csordas A, del Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu Q-W, Wang R, Hermjakob H (2015) 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 44:D447–D456

Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW et al (2013) The mzQuantML data standard for mass spectrometry–based quantitative studies in proteomics. Mol Cell Proteomics 12:2332–2340

Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C (2015a) Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163–3168

Wang J, Tucholska M, Knight JD, Lambert J-P, Tate S, Larsen B, Gingras A-C, Bandeira N (2015b) MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat Methods. https://doi.org/10.1038/nmeth.3655

Yates JR III (2015) Pivotal role of computers and software in mass spectrometry–SEQUEST and 20 years of tandem MS database searching. J Am Soc Mass Spectrom 26:1804–1813

Zhang Y, Bilbao A, Bruderer T, Luban J, Strambio-De-Castillia C, Lisacek F, Hopfgartner G, Varesio E (2015) The use of variable Q1 isolation Windows improves selectivity in LC–SWATH–MS acquisition. J Proteome Res 14:4359–4371

Zhang L, van Dijk ADJ, Hettinga K (2016) An interactomics overview of the human and bovine milk proteome over lactation. Proteome Sci. https://doi.org/10.1186/s12953-016-0110-0