Intelligent workflows for automated analysis of mass spectrometry-based proteomics data

Güler, A.T.

CHAPTER 1

1

General Introduction

Mass spectrometry is a powerful and comprehensive technique for studying proteomics. Instrumentation and techniques for generating and analyzing mass spectrometry data are constantly evolving. This first chapter introduces the essential topics and concepts that the research in this thesis is built upon, such as; proteins and proteomics, use of mass spectrometry in proteomics, analysis of proteomics data, and scientific workflows. The scope of the thesis and the content of the following chapters are also briefly introduced.

Proteins – the building blocks of life

Proteins are the executive molecules in cells. They interact with many other molecules, and their structure and behavior affect how cells function. Being key players in cellular mechanisms and disease pathologies, the study of proteins remains one of the main interests of biomedical research[1].

The functional properties of a protein are determined by its structure, which is directly influenced by its amino acid sequence[2]. "The central dogma of molecular biology" refers to the unidirectional transfer of sequence information from nucleic acids to proteins. This transfer of information is first carried out in a process called transcription from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) and then from RNA to protein in a process called translation[3]. In eukaryotic cells, these processes take place in different compartments of the cell. Transcription happens in the nucleus where the DNA is located; then, the synthesized mRNA is transferred to the cytoplasm for translation[4,5]. The synthesized protein folds to its three-dimensional conformation and can be further matured by post-translational modifications that alter its function[6]. (Figure 1.1) Proteins can be transported to various compartments in the cell or secreted[7,8,9].

Genes are essentially the blueprints for proteins that are synthesized in the cell. The collection of all the genes in the genetic material of an organism, the genome, is highly similar in virtually all the cells of the organism. The set of expressed proteins, on the other hand, varies extensively depending on time and condition[10,11]. The name proteome refers to the set of proteins, as they are the complements expressed from the genome[12]. Since a protein can have many different forms and take part in higher-order complexes, the term proteomics is expanded to all the processes that follow protein synthesis. Proteomics processes are inherently very dynamic[13]. Due to the proteome's dynamic nature, it is not possible to determine the biological function of the genes and their expressed proteins merely from the genomic sequences. That is why proteomics is essential to make sense of genomics data, as both fields support and complement each other in every sense[14,15]. In addition, knowing the amounts of proteins synthesized and present in a cell and how they change throughout different timepoints and conditions provides valuable insights into cellular processes[16].

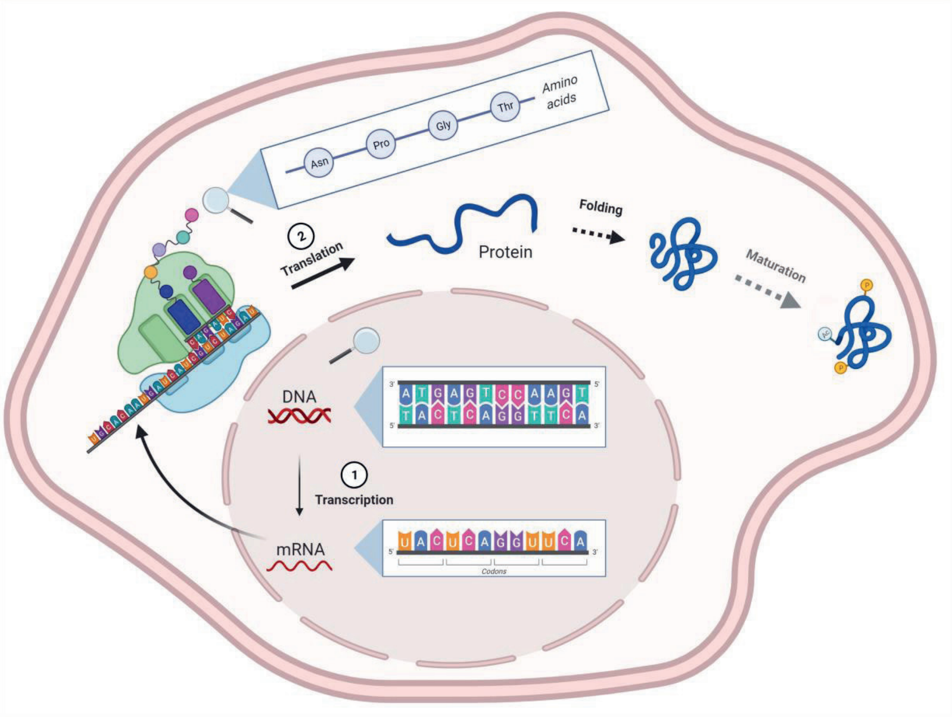

Figure 1.1. DNA is transcribed into complementary RNA in the nucleus during transcription. The mature RNA exported into the cytoplasm is translated into the final gene product: protein. The translated protein folds to its three-dimensional structure and further matures by post-translational modifications.

Proteomics aims at developing an understanding of how the different proteins in an organism function. In order to accomplish this with vast amounts of data, efforts are being made to catalog and organize existing knowledge on proteins with initiatives like the Gene Ontology[17]. Various technologies are being developed, advanced, and used to compare proteins and their abundances in different samples and to unravel the roles of post-translational modifications, localization of proteins, and protein interactions[18].

Mass spectrometry-based proteomics

Proteomics is a big omics domain following its older sibling genomics, providing new insights into biological functions[15]. Proteomics complements genomics since proteins are generally the end products of genomic expression. However, in contrast to the genome, which is relatively static, the proteome is a dynamic entity and quite complex. Therefore, higher throughput and sensitivity are absolute necessities in proteomics analyses[13]. Mass spectrometry is a highly sensitive analytical technique that is commonly preferred in proteomics as it can detect even very low abundance proteins in a complex sample[19].

Mass spectrometry measures the mass-to-charge ratio of analytes to characterize and identify them, even in complex mixtures. Analytes may be further fragmented, and the ionized fragments can be measured and used in further characterization and identification. In the top-down sequencing approach, the analytes are the intact proteins; in bottom-up sequencing, the analytes are the peptides of digested proteins[20]. Top-down approaches are beyond the scope of this thesis. A typical bottom-up analysis usually starts with isolated, intact proteins digested into peptides using a protease. In complex samples, proteins may be fractionated prior to digestion using a protein fractionation technique such as SDS-PAGE. The proteins are then digested in-gel following band excision. The resulting peptides are separated using an appropriate separation method. There are different separation methods used for this purpose; this thesis focuses on high-performance liquid chromatography. The separation reduces the complexity of the peptide mixture introduced to the mass spectrometer at any given time, and it is directly coupled to a mass spectrometer[21]. In data-dependent acquisition, the mass spectrometer first analyzes the ionized intact peptides. The peptides are then isolated by their mass to charge ratio and selected for further fragmentation and measurement[22]. The identification of peptides is based on the mass-to-charge ratio of intact peptides and their fragmentation patterns[23]. An example experimental workflow is shown in Figure 1.2. The ionization step and mass analyzers are central to the mass spectrometry technology, and they have a strong influence on the output and are usually taken into consideration when analyzing performance and interpreting results[24].

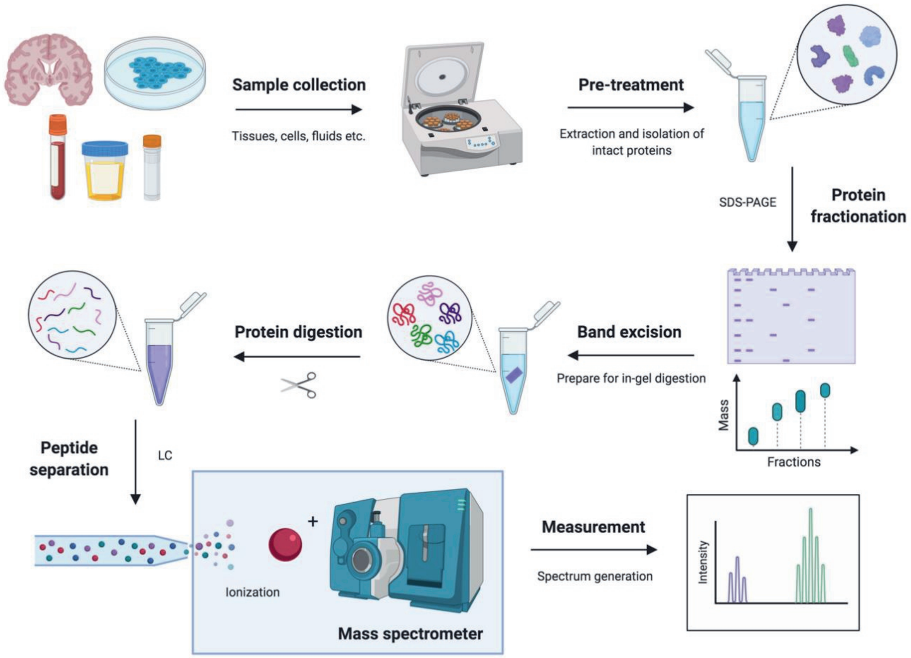

Figure 1.2. Following the sample collection, intact proteins are extracted and isolated. Proteins are fractionated to reduce complexity, and the bands excised from the gel are treated with protease to digest the proteins into peptides. The peptides are separated by liquid chromatography before entering the mass spectrometer. The intact peptides are ionized, and mass-to-charge ratios are measured. Elution times of the intact peptides, along with the mass-to-charge ratios and intensities of the ionized peptides, are all recorded by the mass spectrometer.

Analysis of proteomics data

Analysis of complex and large datasets, such as the ones acquired by mass spectrometry, requires a range of different tools with different functionalities from start to finish. The computational analysis workflow is often not straightforward and requires interventions by the user to channel the output from one tool to another[25]. Even though this means flexibility to a certain extent, it usually comes with a price, too. First of all, computational analyses requiring manual labor are impractical and time-consuming[26]. Second, leaving the user with too many options may be confusing, and they could take incorrect approaches concerning the nature of data. Overall, selecting an appropriate set of tools and optimal parameters is crucial, as any failure to do so often results in accumulation of error or loss of quality[27].

Nevertheless, even when a complete sequence of analysis is performed successfully on mass spectrometry-based proteomics data, it is usually not enough on its own to make biological derivations. Comparison among different analyses and datasets is the starting point for finding meaningful explanations for certain behaviors and characteristics[28]. However, differences among samples cannot be attributed to biological variations only. Technical variability in sample preparation, instrument settings, computational analysis, and statistics performed on the data also affect the results[29,30,31]. In order to assess the outcomes of an experiment objectively, relevant information about these factors should be known. The Proteomics Standards Initiative proposed a "Minimum Information About a Proteomics Experiment (MIAPE)" as a guideline on how relevant contextual data should accompany proteomics data[32]. The contextual data accompanying the measurement data itself is often referred to as the metadata. Providing relevant information about the experiment, measurement, and data analysis in the metadata is necessary for objectively evaluating the study and also contributes to experimental repeatability and reproducibility[33]. It is essential that the metadata is presented in a semantically unambiguous manner[34]. Unsurprisingly, the rapid evolution of mass spectrometry-based proteomics techniques also brought along the same need that genomics once had, and still has, with the explosion of advanced sequencing technologies: a common language for nomenclature[17]. This goal is generally accomplished by "scientific ontologies" or "controlled vocabularies" that represent knowledge in a formalized manner by using the hierarchies and relationships among the domain entities as a backbone.

Bioinformaticians need information on biological species, sample preparation, and instrumentation to choose suitable data analysis methods. Formalized representation of the metadata, on the other hand, paves the way for using the ultimate potentials of computers in interpreting knowledge and making decisions[35]. Another primary benefit of formalized knowledge is semantic interoperability, as integration from different sources across different tools becomes more manageable when the vocabulary of a domain is standardized[36]. All in all, the use of formalized vocabularies and ontologies makes automated decision-making in computational analysis a reality[37]. Since the initiation of Gene Ontology, bio-ontologies have come a long way. There are many controlled vocabularies and ontologies for different life sciences domains, the majority of which can be found in the Open Biological and Biomedical Ontologies (OBO) foundry[38]. The most relevant one for mass spectrometry-based proteomics is the Human Proteome Organization (HUPO) Proteomics Standards Initiative Controlled Vocabulary (PSI-CV), which consists of all the terms used in mass spectrometry pipelines for proteomics[39]. Primarily, sub-branches below 'spectrum generation information' provide annotations regarding instrumentation, sample, and scans that are invaluable for optimizing parameters in automated analysis. The EDAM ontology is helpful in describing and constructing data analysis workflows for mass spectrometry-based proteomics, having a comprehensive definition of bioinformatics methods, data types, and operations[25,40].

Data types are another aspect that should be taken into account while performing mass spectrometry data analysis. There are many different types of mass spectrometers manufactured by different vendors, and almost all of them have their unique raw data format. As a result, too many, not necessarily novel, software are being developed for analyzing data in different formats, while integration and interoperation between them are almost impossible. Standard data formats that can be used across different tools and platforms are necessary to tackle this issue[41]. There are several open data formats for mass spectrometry data, e.g., mzXML[42] and mzML[43]. Each takes advantage of XML's portability, while mzML, being developed by the HUPO PSI team, uses the PSI-CV terms to represent metadata and goes with a more flexible format for new annotations in the future[39].

Reproduction and reuse of scientific data, building upon previous work, are fundamentally important to the progress of science[44,45]. The most established guidelines for reaching these goals are outlined in the FAIR principles, where FAIR stands for Findability, Accessibility, Interoperability, and Reusability[46]. Deposition of mass spectrometry data to online repositories fulfills these principles to some extent, and some journals even have this as a requirement for the publication of findings from mass spectrometry-based proteomics data. Besides the raw data itself, the software and methods used for the data analysis and data formats for input and output are also "data" in a different domain, i.e., bioinformatics. Thus, the community standards should be followed for the management of research software as well[47,48]. The adoption of standard open data formats is an important step towards this aim[46,49,50]. There are efforts to popularize these formats by providing tools for conversions from raw vendor formats; this is one of the many aims of the ProteoWizard project[51]. Newly developed open-source bioinformatics tools should work with common open data formats, and existing tools should be integrated into analysis pipelines when possible to avoid reinventing the wheel[52]. The Trans-Proteomic Pipeline (TPP) is a perfect example in that respect, as it contains a variety of tools that work with common open data formats[53]. It is also possible to incorporate TPP components into other pipelines, as shown in Chapter 5 of this thesis. A conceptual framework of the mass spectrometry-based proteomics data analysis ecosystem is presented in Figure. 1.3.

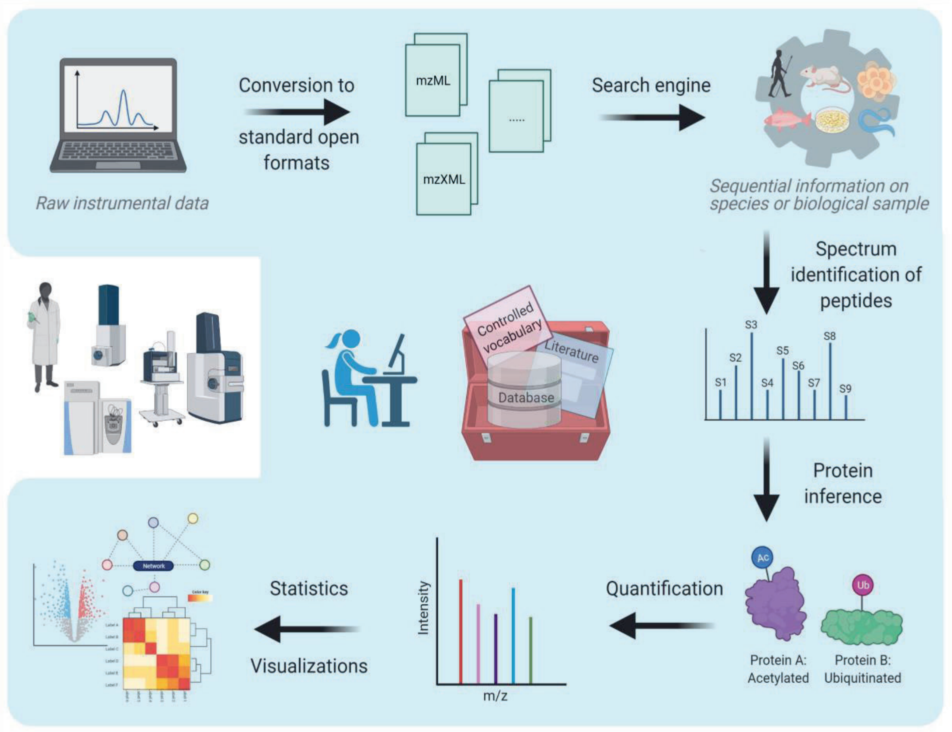

Figure 1.3. The raw data is converted to one of the standard open formats first. A sequence database search is performed to identify the peptides in the sample. Information on identified peptides is used to infer the proteins in the sample. The identified proteins can be quantified using various methods, such as spectral counting or intensity-based calculation. Further statistics can be performed to show the differences among the samples or to visualize the findings. Existing literature, controlled vocabularies, and databases are essential tools in bioinformatician's toolbox; they can be employed at any step of the analysis.

Scientific workflows

Analysis workflows that use many different software components face challenges since different software components have different requirements and use different data types[26]. Scientific workflow managers are developed to overcome this challenge as they assemble different software units by controlling and directing data inputs and outputs in a workflow[54].

Scientific workflows are helpful for automating mass spectrometry-based proteomics data analyses that typically require several different tools[55,56]. However, connecting different modules is not sufficient for the complete automation of analyses. Most multi-step processes require data-dependent decisions. Knowledge level information processing with conditional constructs is one approach for automating data analyses. Formalized knowledge, such as ontologies and controlled vocabularies, can make such condition-based decision-making feasible within scientific workflows[26]. Furthermore, using scientific workflows rather than taking the data through a traditional step-by-step analysis minimizes user interference and supports modularity and reusability[55]. A simple scientific workflow schema for integrating different tools in mass spectrometry-based proteomics is shown in Figure 1.4.

There are many scientific workflow managers publicly available for orchestrating complex data analyses. Some are suitable for analyzing data from a wide range of research disciplines, such as the KNIME Analytics Platform[57] and the Kepler System[58], while some are designed towards a specific field, like the Galaxy Workflow System[59] used in biomedical research. Taverna[60] was one of the pioneering workflow managers and the first that saw widespread use in bioinformatics. The Taverna workflow management system is particularly suitable for high-throughput omics analysis as access to many popular life science tools and services are readily provided[61].

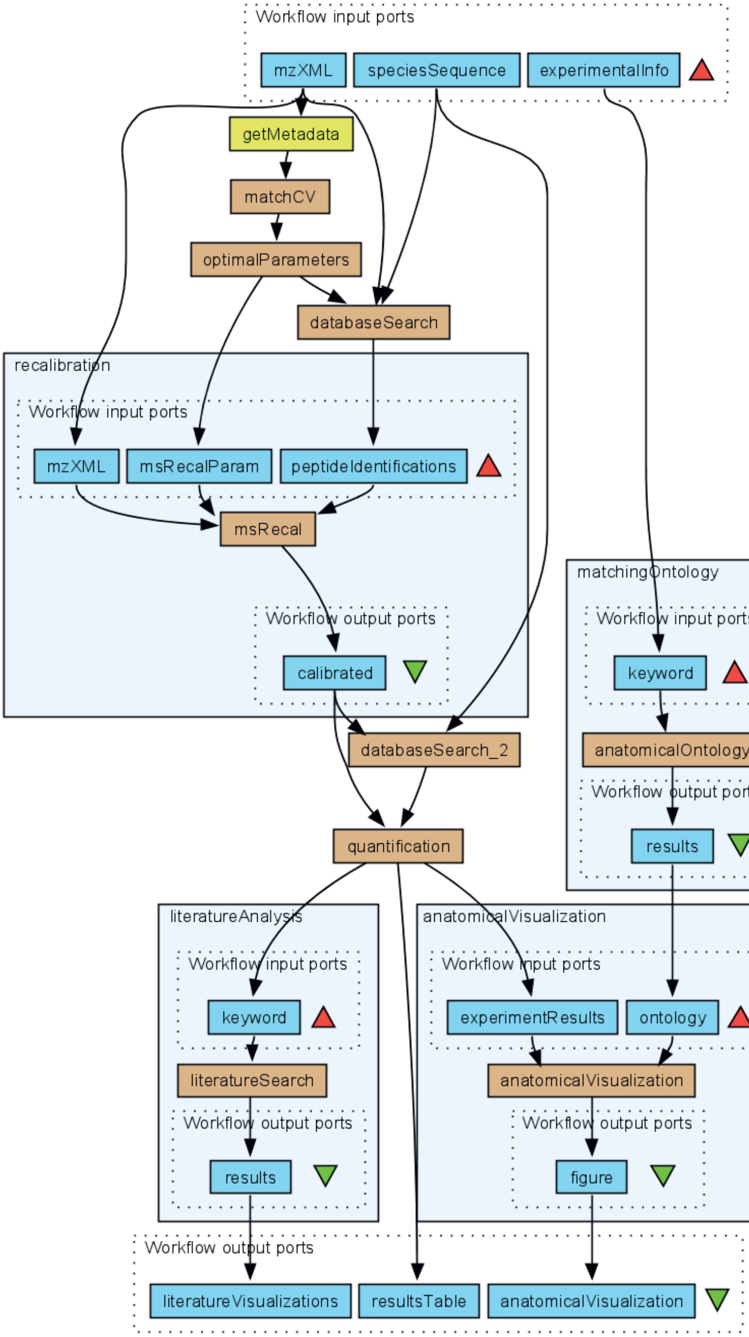

Figure 1.4. A schematic data analysis workflow that integrates modular tools for literature search, anatomical visualization, and mass recalibration.

Scope of the thesis

The aim of this thesis is to build capable scientific workflows for mass spectrometry-based proteomics research and create modular, interchangeable, and interoperable tools for different steps of data analysis. The thesis is concerned with bottom-up, data-dependent proteomics experiments; however, the concepts and methods are applicable to all types of LC-MS/MS-based proteomics.

The Taverna workflow suite is used to demonstrate the capabilities of scientific workflow managers (Chapters 2 and 3). If not all, most of the things demonstrated in this study could also be reproduced on other platforms, as they share the same basic principles. The workflows and tools presented in this thesis are deposited online and could be used in their original version or modified according to the user's needs. The programming languages used mainly throughout the study are as follows: R for statistics and visualizations (Chapters 2, 3. 4, and 5), C for open format reader/writer libraries and mass spectrometry data recalibration tool (Chapter 5). All the data used for testing the tools were retrieved from public databases. The ontologies and controlled vocabularies mentioned are also publicly available in the OBO foundry.

Chapter 2 introduces scientific workflows and how they could be used to assemble different tools for multi-step analyses. A traditional scientific study starts with a literature review, and this chapter sets the foundations needed for initiating an experiment or analysis from scratch. The capabilities and services demonstrated here are general; however, each workflow could easily be adjusted to retrieve field-specific or even topic-specific bibliometrics data that would be crucial to come up with a clear goal and a valid hypothesis for the prospective study.

Chapter 3 builds on the study presented in Chapter 2, advancing bibliometric analyses by integrating Web services in Taverna. One of the many advantages of this functionality is that it enables the analysis of curated information online, for instance, from UniProt[62] or PDB[63]. This functionality broadens the horizon from general bibliometric analyses to exploring protein-disease associations, biomolecular interactions, and more. This type of exploration is valuable for researchers when beginning a study and interpreting findings at the end to see where it fits in the context of annotated or published information.

Chapter 4 presents an interactive tool for the anatomical visualization and exploration of omics data in several model organisms. In omics research, integration of genomics, transcriptomics, and proteomics data is becoming a common practice. However, this usually complicates the data analysis as there are not many generic tools for anatomical visualization capturing spatial information at arbitrary levels of detail. The tool presented here uses anatomical ontologies to map data at different levels of anatomical detail and is compatible with any type of omics data.

Chapter 5 focuses on a recalibration component that is generally applicable in most mass spectrometry-based proteomics data analysis workflows. This calibration tool, msRecal, improves mass measurement accuracy through automated internal calibration. Accurate determination of precursor ion masses increases confidence in identifications and also improves quantitative precision in label-free proteomics. This version of msRecal can recalibrate data from FTICR, Orbitrap, and TOF instruments. The calibration mode is chosen based on the mass analyzer type retrieved from the metadata. Notably, the msRecal component does not change the type or format of the data in any way. Thus, it can easily be plugged into virtually any bottom-up proteomics data analysis workflow.

Finally, **Chapter 6** offers a general discussion on the methods and concepts presented in this thesis. The current issues regarding them are reflected, and future perspectives are discussed.

Acknowledgments

Figures 1.1, 1.2, and 1.3 are created with BioRender.com; figure 1.4 is created with Taverna Workflow Manager.

References

1. Kumar, C. & Mann, M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Letters* **583**, 1703–1712 (2009).
2. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 1309–1314 (1961).
3. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
4. Roeder, R. G. 50+ Years of Eukaryotic Transcription: an Expanding Universe of Factors and Mechanisms. *Nat. Struct. Mol. Biol.* **26**, 783–791 (2019).
5. Köhler, A. & Hurt, E. Exporting RNA from the nucleus to the cytoplasm. *Nat. Rev. Mol. Cell Biol.* **8**, 761–773 (2007).
6. Wang, Y. C., Peterson, S. E. & Loring, J. F. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* **24**, 143–160 (2014).
7. Palade, G. Intracellular Aspects of the Process of Protein Synthesis. *Science.* **189**, 347-358 (1975).
8. Van Vliet, C., Thomas, E. C., Merino-Trigo, A., Teasdale, R. D. & Gleeson, P. A. Intracellular sorting and transport of proteins. *Prog. Biophys. Mol. Biol.* **83**, 1–45 (2003).
9. Benham, A. M. Protein secretion and the endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **4**, a012872 (2012).
10. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
11. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5**, 1512-1526 (2009).
12. Wilkins, M. R. *et al.* From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14**, 61-65 (1996).
13. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
14. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
15. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat. Methods* **7**, 681–685 (2010).
16. Gerster, S. *et al.* Statistical approach to protein quantification. *Mol. Cell. Proteomics* **13**, 666–677 (2014).
17. Ashburner, M., Bal, C. A., Blake, J. A. & Botstein, D. Gene Ontology : tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

18. Colinge, J. & Bennett, K. L. Introduction to computational proteomics. *PLoS Comput. Biol.* **3**, 1151–1160 (2007).

19. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

20. Wysocki, V. H., Resing, K. A., Zhang, Q. & Cheng, G. Mass spectrometry of peptides and proteins. *Methods* **35**, 211–222 (2005).

21. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews* **113**, 2343–94 (2013).

22. Hu, A., Noble, W. S. & Wolf-Yadlin, A. Technical advances in proteomics: New developments in data-independent acquisition. *F1000Research* **5**, 1–12 (2016).

23. Noble, W. S. & MacCoss, M. J. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.* **8**, e1002296 (2012).

24. Cristoni, S. & Bernardi, L. R. Bioinformatics in mass spectrometry data analysis for proteomics studies. *Expert Rev. Proteomics* **1**, 469–483 (2004).

25. Palmblad, M., Lamprecht, A. L., Ison, J. & Schwämmle, V. Automated workflow composition in mass spectrometry-based proteomics. *Bioinformatics* **35**, 656–664 (2019).

26. Gil, Y. From data to knowledge to discoveries: Artificial intelligence and scientific workflows. *Sci. Program.* **17**, 231–246 (2009).

27. Holl, S., Mohammed, Y., Zimmermann, O. & Palmblad, M. Scientific workflow optimization for improved peptide and protein identification. *BMC Bioinformatics* **16**, 1–13 (2015).

28. Fay, D. S. & Gerow, K. A biologist’s guide to statistical thinking and analysis. *WormBook* **9***,* 1–54 (2013).

29. Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T. & VanBogelen, R. A. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **3**, 1912–1919 (2003).

30. Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2010).

31. Pursiheimo, A. *et al.* Optimization of Statistical Methods Impact on Quantitative Proteomics Data. *J. Proteome Res.* **14**, 4118–4126 (2015).

32. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**, 887–893 (2007).

33. Kolker, E. *et al.* Toward More Transparent and Reproducible Omics Studies Through a Common Metadata Checklist and Data Publications. *Omi. A J. Integr. Biol.* **18**, 10–14 (2014).

34. Canham, S. & Ohmann, C. A metadata schema for data objects in clinical research. *Trials* **17**, 1–11 (2016).

35. Soldatova, L. N. & King, R. D. An ontology of scientific experiments. *J. R. Soc. Interface* **3**, 795–803 (2006).

36. Bodenreider, O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.* 67–79 (2008).

37. Mayer, G. *et al.* Controlled vocabularies and ontologies in proteomics: Overview, principles and practice. *Biochim. Biophys. Acta - Proteins Proteomics* **1844**, 98–107 (2014).

38. Smith, B. *et al.* The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).

39. Mayer, G. *et al.* The HUPO proteomics standards initiative mass spectrometry controlled vocabulary. *Database* **2013**, 1–13 (2013).

40. Ison, J. *et al.* EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**, 1325–1332 (2013).

41. Deutsch, E. W. *et al.* Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J. Proteome Res.* **16**, 4288–4298 (2017).

42. Pedrioli, P. G. a *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459–66 (2004).

43. Martens, L. *et al.* mzML - A community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, 1–7 (2011).

44. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 1–9 (2017).

45. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–25 (2015).

46. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).

47. Lamprecht, A.-L. *et al.* Towards FAIR principles for research software. *Data Sci.* **3**, 37–59 (2019).

48. Harjes, J., Link, A., Weibulat, T., Triebel, D. & Rambold, G. FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results. *Database* **2020**, baaa059 (2020).

49. Katayama, T. *et al.* The 2nd DBCLS BioHackathon: Interoperable bioinformatics Web services for integrated applications. *J. Biomed. Semantics* **2**, 1–18 (2011).

50. Deutsch, E. W. File formats commonly used in mass spectrometry proteomics. *Mol. Cell. Proteomics* **11**, 1612–1621 (2012).

51. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

52. Farnham, A. *et al.* Early career researchers want Open Science. *Genome Biol.* **18**, 221 (2017).

53. Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10**, 1150–1159 (2010).

54. Talia, D. Workflow Systems for Science: Concepts and Tools. *ISRN Softw. Eng.* **2013**, 404525 (2013).

55. de Bruin, J. S., Deelder, A. M. & Palmblad, M. Scientific workflow management in proteomics. *Mol. Cell. Proteomics* **11**, M111.010595 (2012).

56. Perez-Riverol, Y. & Moreno, P. Scalable Data Analysis in Proteomics and Metabolomics Using BioContainers and Workflows Engines. *Proteomics* **20**, 1–12 (2020).

57. Berthold, M. R. *et al.* KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* **11**, 26–31 (2009).

58. Ludäscher, B. *et al.* Scientific workflow management and the Kepler system. *Concurr. Comput. Pract. Exp.* **18**, 1039–1065 (2006).

59. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).

60. Hull, D. *et al.* Taverna: A tool for building and running workflows of services. *Nucleic Acids Res.* **34**, 729–732 (2006).

61. Wolstencroft, K. *et al.* The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **41**, W557–W561 (2013).

62. Magrane, M. & UniProt Consortium. UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011**, bar009 (2011).

63. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).