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

Güler, A.T.

CHAPTER 6

Discussion

The research presented in this thesis concerns common workflows and scalable tools in proteomics data analysis that minimize the need for human intervention and make the analyses as experiment-independent as possible. Three themes recur throughout the thesis: automation of mass spectrometry data analyses, FAIR data[1], and data integration. The themes are strongly interrelated and are almost impossible to disentangle.

Mass spectrometry in proteomics

Mass spectrometry is a very powerful tool for identifying, characterizing, and quantifying proteins. However, there is still room for improvement on the instrumental side to enable the analysis of complete proteomes in a manageable time with high sensitivity. Higher resolving power, sensitivity, and speed result in tremendous amounts of mass spectrometry data. Development and adoption of technologies like trapped ion mobility, as in the Bruker Daltonics timsTOF[2], increases the sequencing speed without losing sensitivity by taking advantage of parallel accumulation with serial fragmentation and introduces ion mobility as a fourth dimension into the data. Higher degrees of multiplexing, such as the TMTpro 16-plex[3] by Thermo Fisher Scientific, are now common in quantitative proteomics. All these trends suggest mass spectrometry data will continue to grow exponentially and become more complex. The ability to quickly analyze, document, and share data and results sometimes struggle to keep pace with developments on the instrument side. The methods and tools presented in this thesis make use of various practices such as scientific workflows, ontologies, FAIRification of data and software to help in this endeavor.

Mass spectrometry-based proteomics data analysis

Analysis of proteomics samples with mass spectrometry is becoming more accessible to researchers and has, without a doubt, established itself as an essential analysis method in the field. The technology has developed in recent years in terms of speed and flexibility, and there has been an increase in the number of core facilities performing these analyses for researchers. As vendor software tools are not readily and freely available for the research groups that do not own the equipment but

instead get their mass spectrometry data from core facilities or public repositories, it is quite common to use academic tools that are usually free and open. Academic tools and method developments are usually initial ideas or alternatives that eventually end up in vendor software and constitute a rich ecosystem for analyzing mass spectrometry-based proteomics data. Although there are exceptions, academic software is prone to decay, as most of the time, update and management efforts fade after the project is finished or runs out of funding. There are initiatives to support the management of existing software, such as "Essential Open Software for Science", which aims to fund the further development and management of software with proven impact[4]. Hopefully, in the era of Open Science, more of these initiatives will help open software to reach the level of vendor software in terms of service quality, maintenance, and bug-fixing.

Different techniques and experimental procedures require different analyses. There are more tools available than common operations in proteomics data analysis, creating a burden for the researcher to find the right tool for their experimental setup, let alone the most appropriate tool for the job[5,6]. Apart from the experiments, input/output formats are also important when selecting a tool. Software registries with functional annotations such as Elixir bio.tools[7] make finding the right tool for a specific task easier[8] and facilitate building workflows[9].

Automation of data analysis

Terabytes of mass spectrometry data are being generated every day. Analyzing them becomes an enormous burden for data scientists, given the time and resources available. Complex data requires multiple steps of analysis that need extra effort for channeling the data flow through different steps. Each step usually employs different data analysis modules that are not readily interoperable with each other's input and output. This issue can be managed to a certain extent using command-line "shims". However, these solutions are not particularly user-friendly. There is no doubt that scientific workflow management systems are gaining popularity since they are very efficient for combining modules that are not readily compatible for data flow while remaining easy to use and share[10,11,12,13].

The recalibration tool presented in **Chapter 5**, msRecal, improves the mass measurement accuracy through internal calibration. As a result, the number of high confidence identifications is increased. The output format is the same as the input, so this module can be easily plugged into a bottom-up label-free analysis workflow, such as the one that we used to analyze the data in Hussaarts et al.[14], as demonstrated for ion trap-FTICR data by de Bruin et al.[15]

Managing the flow of data through interoperating tools is a good starting point; however, automation of data analysis also requires semantic interoperability within and across these modules. In mass spectrometry-based proteomics, experimental attributes such as instrument type, sample preparation methods, biological species, etc., are important as different parameters are required for different set-ups when performing data analysis. Controlled vocabularies and ontologies are frequently used for this purpose, as they are easier for machines to interpret, and they also solve the ambiguities in semantics to a certain extent[12]. Open data formats such as mzXML, mzML, and mzData support embedded metadata[16]. The data elements are annotated as free text descriptions in mzXML, while mzML and mzData rely heavily on controlled vocabularies for this purpose. Commonly, the data elements in these files are annotated with high-level terms, or sometimes even with incorrect terms, since the raw vendor files usually do not contain information at a sufficient level of detail in the first place. Having the annotation at the correct hierarchy level can help choose a better suiting analysis method or visualization. Vendors should provide sufficient metadata using controlled vocabularies with the raw output, and open software developers should use the same vocabularies in the tools they develop. The tools that use metadata to select analysis or visualization methods should be flexible to traverse between different levels of abstraction. This is demonstrated with the anatomical ontology visualization tool presented in **Chapter 4** and the recalibration tool in **Chapter 5**.

A literature study is an essential first step when designing an experiment or data analysis in any field, and mass spectrometry-based proteomics is no exception. Comprehensive manual literature analysis is prohibitively time-consuming. Bibliometrics emerged in the first half of the 20th century and was concerned with measuring various aspects of books and different forms of publications. As a field, it

has developed its own methods and practices[17]. Nevertheless, it is possible to design compact and reproducible field-specific literature analysis workflows without getting lost in the details of advanced bibliometrics methods. In **Chapters 2 and 3**, some examples of bibliometrics analysis applicable in mass spectrometry and proteomics research are presented. The bibliometrics workflows in these chapters could be used before designing or conducting an experiment. The information in the literature could also guide choosing the settings and parameters for certain steps in data analysis, like recalibration in **Chapter 5**, where data from different experimental set-ups typically require different settings. Bibliometrics analysis also comes in handy at the end of a study to map or contextualize experimental results relative to the literature to expand existing knowledge. Scientific workflows such as those presented in **Chapter 2** can guide users and help them find relevant publications, or even potential collaborators, on a particular topic, especially when different authors use slightly different vocabulary. The use of different terms by authors working in the same field is also explored in this chapter. This ambiguity in naming terms is one of the reasons why common nomenclatures and controlled vocabularies of species, chemicals, genes, proteins, and methods are necessary.

Data availability and reusability

In increasing numbers of proteomics and mass spectrometry journals, the researchers are required to submit their raw data and analysis results. There are several public mass spectrometry repositories, with PRIDE being the largest and most popular repository of mass spectrometry-based proteomics data[18,19]. Each dataset uploaded to PRIDE is linked to a publication. The publications using new or already existing data available on PRIDE also have links to the datasets; thus, the data and the publication are accessible in both ways.

Although data analysis is one of the final steps in a proteomics experiment, how it is done can have tremendous effects on the results and how much can be inferred from the experimental data. An inadequate analysis can easily squander an otherwise well-designed and conducted experiment. Making data FAIR prevents poorly annotated good data from going to waste. Usually, it is easier to comply with the first two principles of FAIR, findable and accessible, than the last two, interoperable and

reusable, as it requires more than trivial effort to make them such. FAIR data can be retrieved by other groups and reanalyzed to draw new biological conclusions. The anatomical visualization tool in **Chapter 4** and the mass recalibration tool in **Chapter 5** are meant to analyze new experimental data and existing data from public repositories. The scientific workflows presented in **Chapters 2 and 3** are useful for searching published studies and retrieving findings.

In addition to the data itself, it is crucial to make the metadata FAIR as well, since they are essential when analyzing data on public repositories. The standardization of the metadata is a relatively new concept in the field; recently European bioinformatics community has initiated an open-source project called Sample to Data file format for Proteomics for this purpose[20]. Without a doubt, such efforts will make data reanalysis easier in the future.

Data integration

To comprehensively study the biological mechanism, integrating heterogeneous sources of data is practically necessary for omics research. None of the omics fields exist in a vacuum; they all complement each other[21]. However, integrating data across different omics levels is only one side of the story. Data across similar experiments are also integrated to minimize experiment-dependent variations[22]. Inherently, such integration is more straightforward than integrating data from different omics levels; however, the metadata remains a crucial component since even the slightest difference in sample preparation, or instrumental setup that is overlooked can lead to a greater diversion from reality. The importance of vendor support of open data formats remains central for the feasibility of data integration as it is the melting pot for data from different sources. Data integration will be quite complex or even impossible if the data on public resources are not FAIR.

The anatomical visualization tool from **Chapter 4**, COMICS, uses metadata to automate the selection of anatomical abstraction levels. It requires only one standard input for any omics experiment, smoothly integrating data across different omics experiments. The msRecal tool presented in **Chapter 5** can also be used to integrate data from different mass spectrometers and experiments if the analyses are

performed on similar samples. The msRecal tool works with an open format, mzXML, and it can analyze data from many different types of mass spectrometers.

Future perspectives

Increasing acceptance of FAIR and open science notably improves the logistics of conducting scientific research in mass spectrometry-based proteomics. These efforts take the field one step closer towards achieving automated, wide-coverage robust analyses that can reuse and integrate existing data. Open data repositories such as PRIDE[18] and MassIVE[23] have existed for some time already. Although they provide invaluable data resources for reanalysis, the requirements for uploading data on these repositories still have room for improvement. The data on these repositories are usually linked to their respective publications explaining the original experiment, data analysis methods, and results. However, most of the essential information is not readily machine-readable, and some datasets have incomplete or missing metadata. As a result, extensive manual labor is still needed to get the data ready for reanalysis. For the time being, automated literature search and information extraction methods like the ones presented in this thesis could make these steps manageable to a certain extent. As the requirements for uploading data to these repositories become stricter in the future, automated literature search could be employed beyond its horizon rather than making up for the missing bits that should have already been there. One foreseeable use case scenario would be using web services integrated with machine learning for advanced, direct, and manual-labor-free reanalysis of data.

The generation of good quality data undoubtedly needs a lot of technological resources (i.e., mass spectrometers and other lab equipment) and human resources for operating this high-end instrumentation and analyzing the results. The resources and efforts needed for designing and developing efficient data analysis tools are often overlooked in biological sciences. Time-wise and funding-wise, data analysis tools should get their fair share in bioscience research. As much as the FAIRness of data is crucial, applying these principles for data analysis tools is also essential and well worth the investment in the long term. There is already a rich ecosystem for finding and sharing data analysis tools, such as GitHub[24] for version control and source code management, WorkflowHub[25] and MyExperiment[26] for sharing scientific workflows,

Galaxy Community[27] for sharing Galaxy workflows and deploying Galaxy Servers, and ELIXIR bio.tools[7] for a comprehensive registry of bioinformatics software. The importance of these communities for data and tool sharing, bug reporting is evident, and a broader audience in bioscience research should support them.

Initiatives for developing community standards, such as HUPO PSI[28] and EDAM ontologies[29], are vital for achieving the goals listed here. Standard open data formats are also fundamental, although they need more vendor support to reach their full potential in data automation. Some of the workflow managing software used today for automation may become obsolete in the future. However, the concept of scientific workflows with scalable components is here to stay, and most likely, there lies the future of proteomics data analysis.

References

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

2. Bruker. timsTOF. https://www.bruker.com/en/products-and-solutions/mass-spectrometry/timstof/timstof.html

3. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).

4. Chan Zuckerberg Initiative. Essential Open Source Software for Science. https://chanzuckerberg.com/eoss/

5. Tsiamis, V. *et al.* One Thousand and One Software for Proteomics: Tales of the Toolmakers of Science. *J. Proteome Res.* **18**, 3580–3585 (2019).

6. Weintraub, S. T., Hoopmann, M. R. & Palmblad, M. 2021 Special Issue on Software Tools and Resources: Finding the Right Tools for the Job. *J. Proteome Res.* **20**, 1819–1820 (2021).

7. ELIXIR. bio.tools: Bioinformatics Tools and Services Discovery Portal. https://bio.tools/

8. Ison, J. *et al.* The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol.* **20**, 164 (2019).

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

10. da Silva, R. F. *et al.* Workflows Community Summit: Bringing the Scientific Workflows Community Together. *Zenodo* (2021)

11. Deelman, E. *et al.* The future of scientific workflows. *Int. J. High Perform. Comput. Appl.* **32**, 159–175 (2018).

12. Bowers, S. Scientific Workflow, Provenance, and Data Modeling Challenges and Approaches. *J. Data Semant.* **1**, 19–30 (2012).

13. Damevski, K., Khan, A. & Parker, S. Scientific Workflows and Components : Together at Last! in *Proceedings of the 3rd Workshop on Component-Based High-Performance Computing, October 16-17, 2008*, *Karlsruhe, Germany* (2008).

14. Hussaarts, L. *et al.* Human Dendritic Cells with Th2-Polarizing Capacity: Analysis Using Label-Free Quantitative Proteomics. *Int. Arch. Allergy Immunol.* **174**, 170-182 (2017).

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

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

17. Godin, B. On the origins of bibliometrics. *Scientometrics* **68**, 109–133 (2006).

18. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

19. Chen, T., Zhao, J., Ma, J. & Zhu, Y. Web resources for mass spectrometry-based proteomics. *Genomics, Proteomics Bioinforma.* **13**, 36–39 (2015).

20. Perez-Riverol, Y. & European Bioinformatics Community for Mass Spectrometry. Toward a sample metadata standard in public proteomics repositories. *J. Proteome Res.* **19**, 3906–3909 (2020).

21. Santiago-Rodriguez, T. M. & Hollister, E. B. Multi 'omic data integration: A review of concepts, considerations, and approaches. *Semin. Perinatol.* **45**, 151456 (2021).

22. Zhang, B. & Kuster, B. Proteomics is not an island: Multi-omics integration is the key to understanding biological systems. *Mol. Cell. Proteomics* **18**, S1–S4 (2019).

23. Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* **7**, 412-421.e5 (2018).

24. GitHub Inc. GitHub: Where the world builds software. https://github.com/

25. WorkflowHub. https://workflowhub.org/

26. myExperiment - Home. https://www.myexperiment.org/home

27. The Galaxy Community. https://galaxyproject.org/community/

28. Taylor, C. F. *et al.* The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS A Journal of Integrative Biology* **10**, 145-151 (2006).

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