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

Güler, A.T.

APPENDICES

Summary

Mass spectrometry is a powerful technique that provides the high sensitivity and throughput needed for analyzing the complex and dynamic proteome. However, this power comes with a price; the data generated by mass spectrometers are quite complex and require advanced multi-step analysis. Using scientific workflow systems for the analysis of mass spectrometry data is not entirely new. However, innovative solutions are needed to make these workflows autonomous, intelligent, flexible, and adaptable in the era of big (and complex) data. This thesis focuses on building intelligent workflows and modular tools for analyzing, integrating, and contextualizing mass spectrometry-based proteomics data. These workflows and tools could be easily adjusted for additional functionalities and be reused in different data analysis workflows. The scope of this thesis is not limited to the downstream analysis of mass spectrometry data; methods for automated literature search that would be useful for designing experiments and interpreting experimental results are covered in detail. **Chapter 1** gives an overview of the core concepts related to mass spectrometry-based proteomics and data analysis in line with the content presented in subsequent chapters. The challenges and how they are currently being addressed are also explained briefly.

Chapter 2 gives a detailed introduction to scientific workflows and their advantages in multi-step analyses through bibliometrics analysis. The bibliometrics analyses show that different authors could refer to the same domain entity using different terms, even in the same subfield. If the literature search is performed manually using specific keywords, some overlapping studies may be overlooked. The workflows presented here make it easy to perform automated literature searches without getting lost in advanced bibliometrics methods. Getting a quick overview of a field may come in handy for authors when conducting interdisciplinary studies and meta-analyses. Furthermore, researchers can find expert labs and other researchers for collaborations. **Chapter 3** presents more advanced workflows for bibliometrics with the ability to use web services and perform statistical analyses on the data retrieved. After presenting how the web services can be integrated with the Taverna workflow manager, workflows for citation networks and biomolecular interactions

are shown. The output of the workflows could be visualized using existing powerful tools, i.e., the popular VOSviewer or Cytoscape. Literature analyses are favorable when interpreting the findings of a study, visualizing them in context, and getting a roadmap for designing future experiments; this chapter presents how these processes could be automated.

Chapter 4 presents a robust tool for integration and anatomical visualization of quantitative omics data from model organisms. One novelty of this tool is that it uses anatomical ontologies to automatically visualize anatomical information regardless of the resolution of the input data. The tool moves through the anatomical ontology hierarchy to select the appropriate level of organ and tissue details. Second, a simple standard data format is required from the user to visualize their data. Since this data format does not make any omics-based assumptions, data from different omics experiments can be integrated and visualized smoothly without further ado.

Chapter 5 presents a recalibration tool that improves mass measurement accuracy in mass spectrometry data through automated internal calibration. As a result, more peptides could be identified with higher confidence from the recalibrated data. The measured masses could be calibrated using accurate peptide identifications from the original data or from different measurements of the same or similar samples. This tool uses mzXML metadata to select the most suitable mass analyzer-dependent calibration function automatically. The output format is the same as the input, so this tool can easily be plugged into any bottom-up proteomics data analysis workflow working with mzXML in principle. This tool can be used to analyze new experimental data and also existing public repository data.

Finally, in **Chapter 6**, the methods and concepts such as data availability and reusability, automation, data integration are discussed, referring to how they apply to the research presented in this thesis. The present and future of proteomics data analysis are also discussed, shedding light on what needs to be done to accelerate achieving the goal of fully automated, wide-coverage analyses that can reuse publicly available data and knowledge from the literature. Like most software, the workflow manager used in this thesis, Taverna, and other supporting tools, are prone to decay. However, the concepts and methods presented with the help of these tools are here to stay, and the future of proteomics data analysis will build upon them.