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

Güler, A.T.

Intelligent Workflows for Automated Analysis of Mass Spectrometry-based Proteomics Data

Arzu Tuğçe Güler

Intelligent Workflows for Automated Analysis of Mass Spectrometry-based Proteomics Data

Arzu Tuğçe Güler

ISBN: 978-94-6423-731-3

Cover design: Arzu Tuğçe Güler
Printing: ProefschriftMaken.nl

Intelligent Workflows for Automated Analysis of Mass Spectrometry-based Proteomics Data

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 7 april 2022
klokke 11.15 uur

door

Arzu Tuğçe Güler
geboren te Ankara, Turkije in 1988

Promotor:	prof.dr. M. Wuhrer
Co-promotor:	dr. N.M. Palmblad
Leden promotiecommissie:	prof.dr. J.J. Goeman
	prof.dr. P.L. Horvatovich, Rijksuniversiteit Groningen
	dr. K.J. Wolstencroft
	dr. M. Roos

Ars longa, vita brevis

Table of Contents

CHAPTER 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].

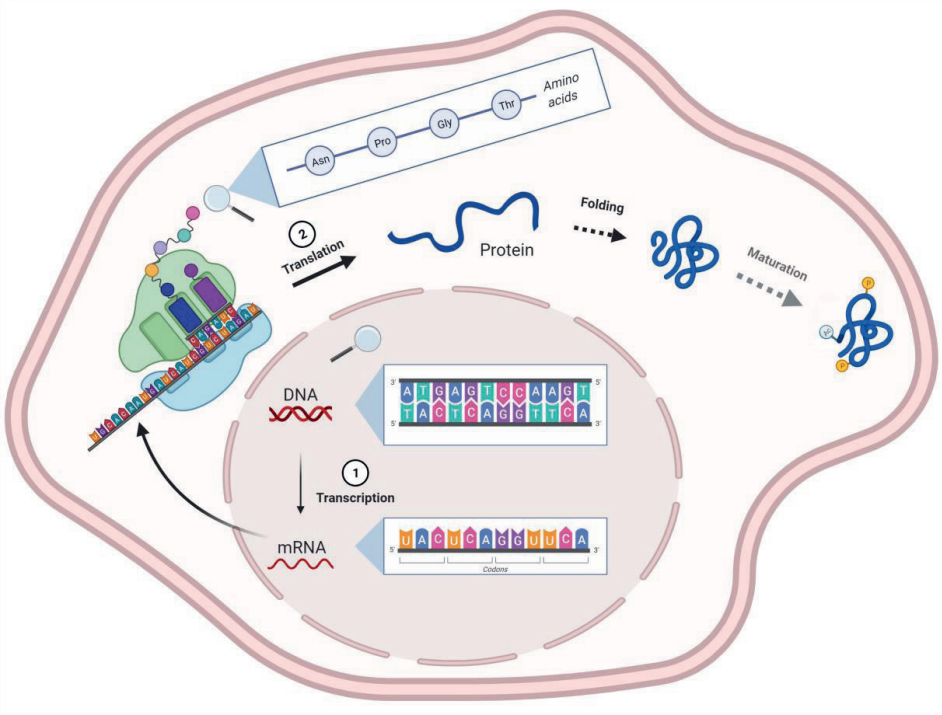


Figure 1.1. DNA is transcribed into complementary RNA in the nucleus during transcription. The mature RNA exported into the cytoplasm is translated into the final gene product: protein. The translated protein folds to its three-dimensional structure and further matures by post-translational modifications.

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].

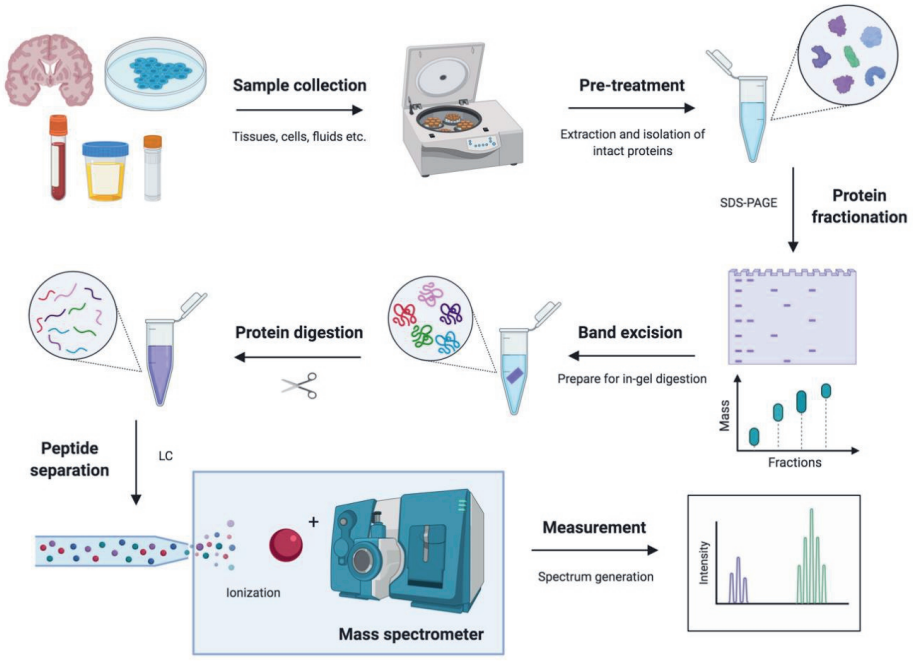


Figure 1.2. Following the sample collection, intact proteins are extracted and isolated. Proteins are fractionated to reduce complexity, and the bands excised from the gel are treated with protease to digest the proteins into peptides. The peptides are separated by liquid chromatography before entering 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.

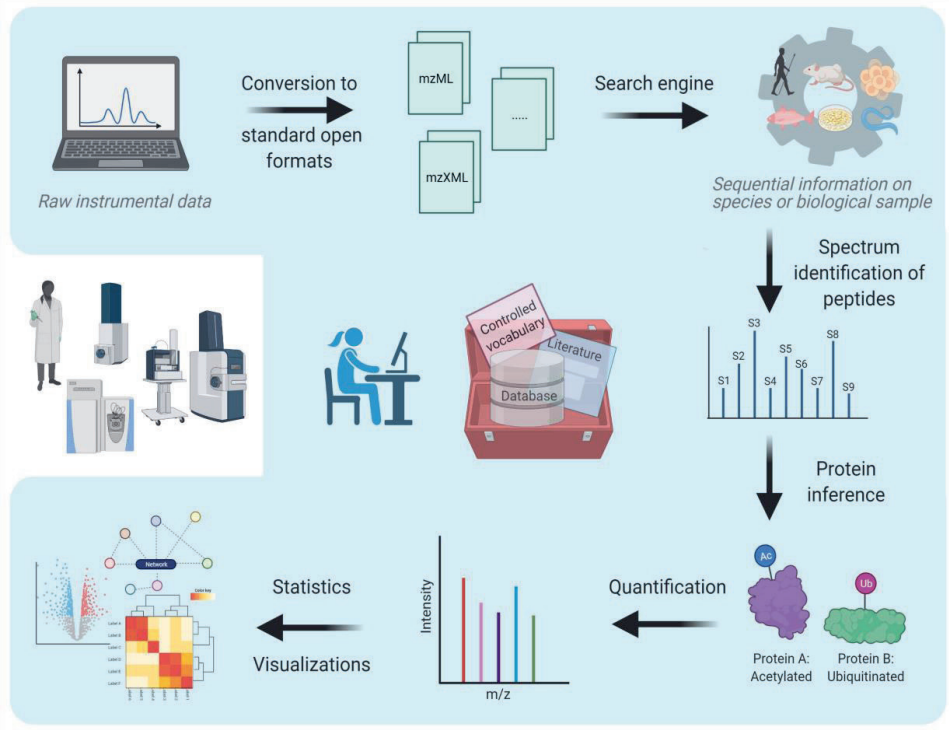


Figure 1.3. The raw data is converted to one of the standard open formats first. A sequence database search is performed to identify the peptides in the sample. Information on identified peptides is used to infer the proteins in the sample. The identified proteins can be quantified using various methods, 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].

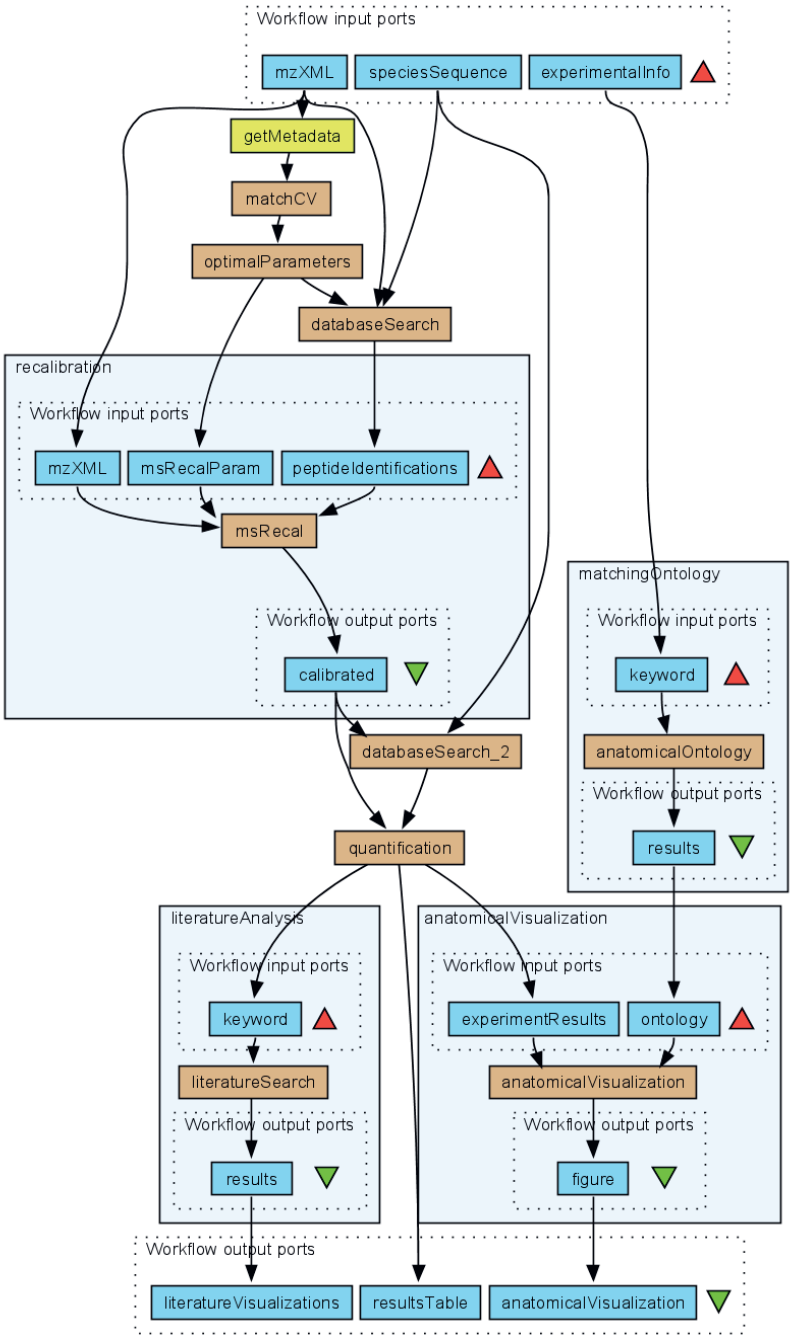


Figure 1.4. A schematic data analysis workflow that integrates modular tools for literature search, anatomical visualization, and mass recalibration.

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).

CHAPTER 2

Scientific Workflows for Bibliometrics

Arzu Tugce Guler[1], Cathelijn J. F. Waaijer[2], Magnus Palmblad[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

[2] Faculty of Social and Behavioural Sciences, Centre for Science and Technology Studies, Leiden University, The Netherlands

This chapter has been published in:
Scientometrics (2016) 107: 385-398 doi: 10.1007/s11192-016-1885-6

Abstract

Scientific workflows organize the assembly of specialized software into an overall data flow and are particularly well suited for multi-step analyses using different types of software tools. They are also favorable in terms of reusability, as previously designed workflows could be made publicly available through the myExperiment community and then used in other workflows. We here illustrate how scientific workflows and the Taverna workbench in particular can be used in bibliometrics. We discuss the specific capabilities of Taverna that makes this software a powerful tool in this field, such as automated data import via Web services, data extraction from XML by XPaths, and statistical analysis and visualization with R. The support of the latter is particularly relevant, as it allows integration of a number of recently developed R packages specifically for bibliometrics. Examples are used to illustrate the possibilities of Taverna in the fields of bibliometrics and scientometrics.

Introduction

Information processing permeates the scientific enterprise, generating and organizing knowledge about nature and the universe. In the modern era, computational technology enables us to automate data handling, reducing the need for human labor in information processing. Often information is processed in several discrete steps, each building on previous ones and utilizing different tools. Manual orchestration is then frequently required to connect the processing steps and enable a continuous data flow. An alternative solution would be to define interfaces for the transition between processing layers. However, these interfaces then need to be designed specifically for each pair of steps, depending on the software tools they use, which compromises reusability. Whether the data flow is automated or manually done by the researcher, the latter still has to deal with many detailed, low-level aspects of the execution process[1].

Scientific workflow managers connect processing units through data, control connections and simplify the assembly of specialized software tools into an overall data flow. They smoothly render stepwise analysis protocols in a computational environment designed for the purpose. Moreover, the implemented protocols are reusable. Existing workflows can be shared and used by other workflows, or they can be modified to solve different problems. Several general purpose scientific workflow managers are freely available, and a few more optimized for specific scientific fields[2]. Most of these managers provide visualization tools and have a graphical user interface, e.g. KNIME[3], Galaxy[4] and Taverna[5]. Not surprisingly, scientific workflows are now becoming increasingly popular in data intensive fields such as astronomy and biology.

In this paper, which builds on a recent ISSI conference paper[6], we describe the use of scientific workflows in bibliometrics using the *Taverna Workbench*. Taverna Workbench is an open source scientific workflow manager, created by the myGrid project[7], and is now being used in different fields of science. Taverna provides integration of many types of components such as communication with Web services (WSDL, SOAP etc.), data import and extraction (XPath for XML, spreadsheet import from tabular data), and data processing with Java-like Beanshell scripts or the

statistical language R[8]. Beanshell services allow the user to either program a small utility from scratch and towards a specific goal, or to integrate already existing software into the workflow. The R support is a particularly powerful feature of Taverna. Although R was initially developed as a language for statistical analysis, its widespread use has seen it adopted for many tasks not originally envisioned—a fate not unlike its commercial cousin, MATLAB. One such task is text mining. The R package "tm"[9] provides basic text mining functionality and is used by a rapidly growing number of higher-level packages, such as "RTextTools"[10], "topicmodels"[11] and "wordcloud"[12]. Similarly, there are many toolkits and frameworks for text mining in Java that could also be called from within a Taverna workflow. For geographic and geospatial analysis, e.g. using author affiliations, there are also a number of very powerful R packages. One such package is "rworldmap"[13], projecting scalar, numerical data onto a current map of the world using the ISO 3166-1 country names. rworldmap gives the user control of most aspects of the map drawing, and enables different map projections to be applied to the maps.

A simple example: comparing two authors

We designed a simple workflow, *Compare_two_authors* (Figure 2.1), to generate a histogram for the number of publications over time and a co-word map for the titles of the two authors' publications. The workflow takes as inputs PubMed results in XML, the names of two authors, a list of excluded words and a minimum number of occurrences.

The PubMed results are retrieved in an XML format, and the extraction of publication years, titles and author names are done by *XPath* services. XPath is a query language for selecting elements and attributes in an XML document. The XPath service in Taverna eases this process by providing a configuration pane to render an XML file of interest as a tree and automatically generate an XPath expression as the user selects a specific fragment from the XML (Figure 2.2). The results of the query can either be passed as text or as XML to other workflow components.

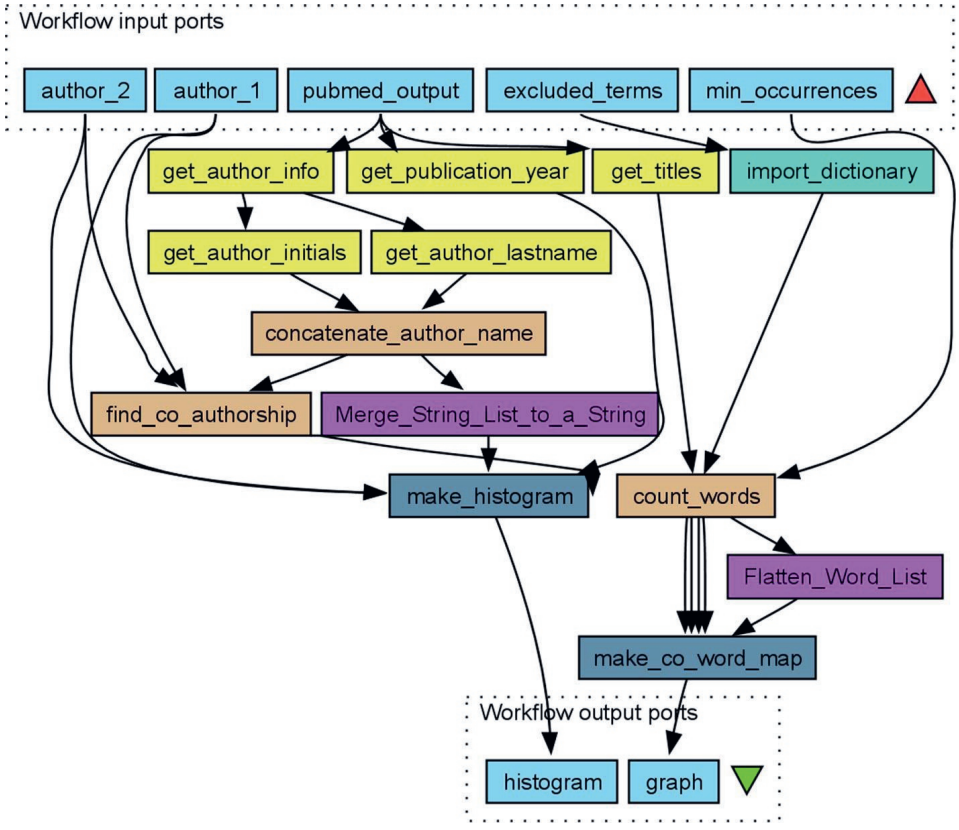


Figure 2.1. A workflow *Compare_two_authors* designed in Taverna for comparing the scientific output over time and word usages of two researchers (authors). Taverna uses *color* to indicate the type of service or tool. Although not performing a particularly sophisticated bibliometric analysis, this workflow demonstrates the use of Beanshells (*burly wood brown*), local services (*heliotrope violet*), spreadsheet import (t*urquoise*), XPaths (*laser lemon yellow*) and Rshells (*air force blue*). The inputs (sky blue) are some PubMed results in XML, the names of two authors, a dictionary of excluded terms and the minimum number of occurrences. Each execution of the workflow creates two outputs: a histogram of the publications in each year for the two authors and a co-word map comparing their research topics. Common words can be excluded for clarity. The *import_dictionary* spreadsheet import service is used to read a text file with one word per line containing words to be excluded.

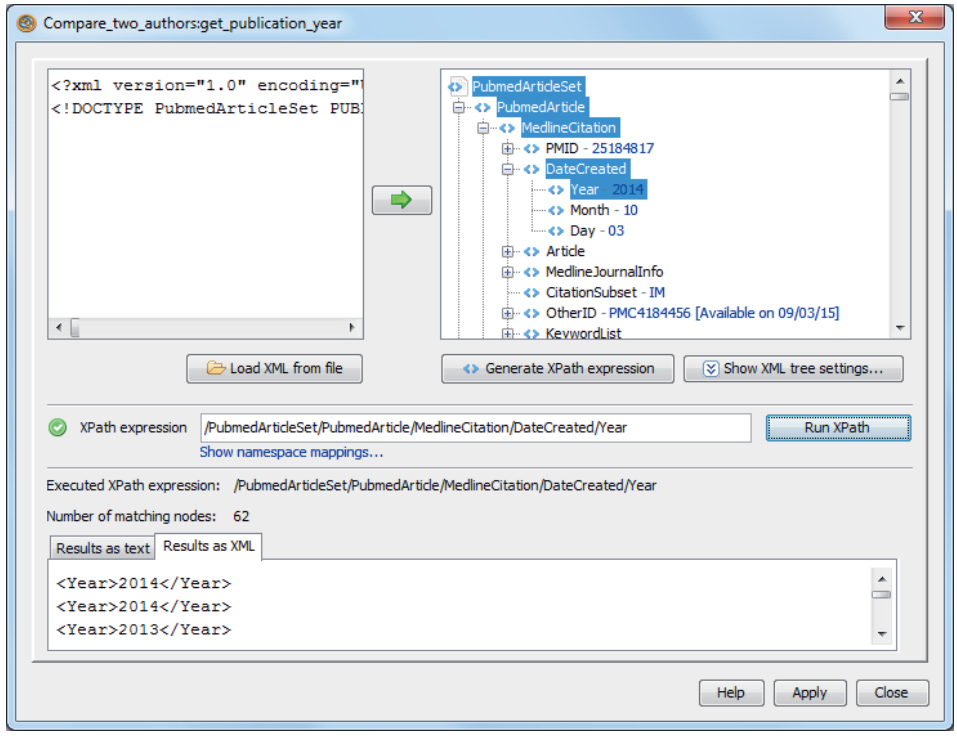


Figure 2.2. The XPath configuration pane provides a simple interface for extracting particular data fields from XML files, here publication years from PubMed search results in XML. There are several "years" in a PubMed entry, corresponding to the date-of-creation for the Medline citation, the article publication date or journal issue publication date. Only the Medline citation date is always present. The XPath/PubmedArticleSet/PubmedArticle/MedlineCitaTion/DateCreated/Year extracts the year from this date.

The data extracted by the spreadsheet import and XPath services is fed to a series of Beanshell components that find co-authorships and count co-occurrence of words in the extracted titles. Beanshell is a light-weight scripting language that interprets Java. In our workflow, the Beanshell services do simple operations on strings, such as concatenation of surnames and initials that are extracted separately using XPath (*concatenate_author_names*), matching strings to find co-authorships (*find_co_authorship*) and counting the number of words occurring in each title authored by one or both authors (*count_words*). The two authors' usage of the words, excluding *excluded_terms*, that appear at least *min_occurrences* times in total, are then used to draw a co-word map using the "igraph" R package[14]. Excluded terms may be

very common, non-informative words like articles and prepositions that would not carry any meaning in a co-word map. It is generally up to the workflow designer what part of the workflow to code in Java (Beanshell), in R, or in third language called via the *Tool* command-line interface. More types are available for data connectors between R components (logical, numeric, integer, string, R-expression, text file and vectors of the first four types) than between Beanshell components, where everything is passed as strings. Therefore, when dealing with purely numerical data, we recommend R over Beanshells within Taverna.

After all the necessary inputs are provided, the workflow is ready to be executed. In the Taverna Workbench Results perspective (Figure 2.3), each completed process is grayed out to show the progress of the workflow run. The execution times, errors and results are also visible in this perspective. We ran the workflow for two scientists active in our own field of mass spectrometry: Gary L. Glish and Scott A. McLuckey, whom we knew to have worked on similar topics over a long period of time and also co-authored a number of articles. However, the workflow will work on any two authors with publications indexed by PubMed. The co-word map in Figure 2.4 visualizes the co-occurrence of words in titles by the location and thickness of the connecting edge, while the relative frequency of usage by the two authors is indicated by color (here from red to blue). This is an example meant to illustrate the capabilities of scientific workflows, not to show a difficult or even particularly interesting bibliometric analysis, although we were surprised to see how strongly individual language preferences appear in these maps, even for two researchers who have a long history of collaboration. For example, one researcher (Glish) may have a strong preference to specify that a “quadrupole ion trap” was used in an experiment whereas another (McLuckey) may refer to the same apparatus as simply an “ion trap”.

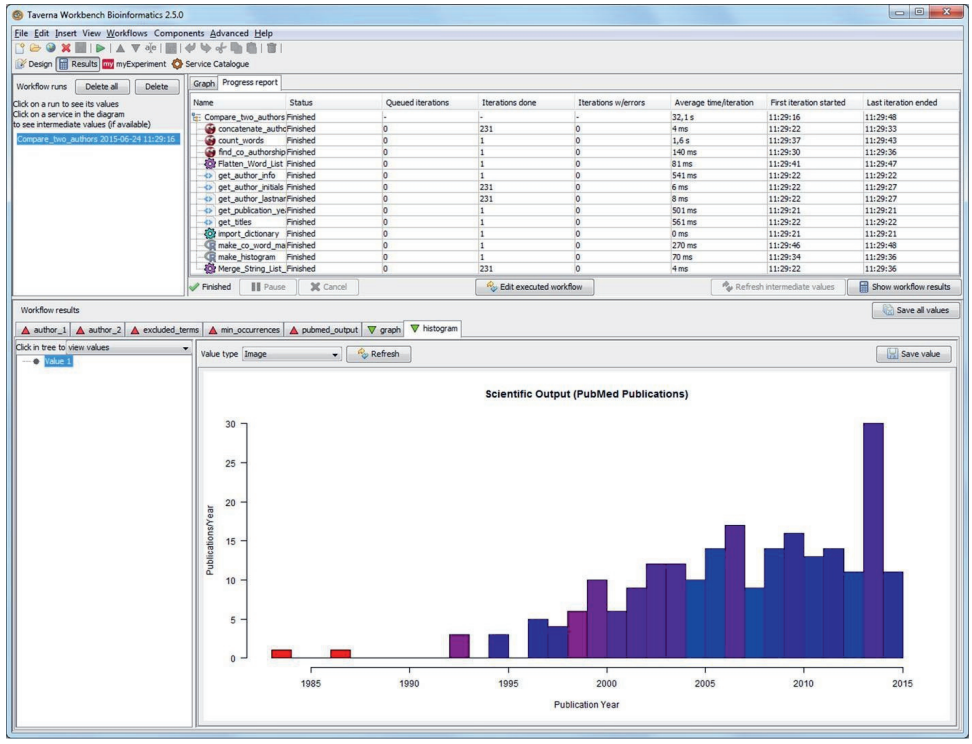


Figure 2.3. Workflow Progress report in the Taverna workbench Results perspective—here with a completed execution of the *Compare_two_authors* workflow in Figure 2.1. The ''histogram'' output is here captured by Taverna, allowing the user to browse the results and select what to save or export to a different data format. In this particular case, the histogram is colored according to relative author output, with *red* being Glish and *blue* McLuckey.

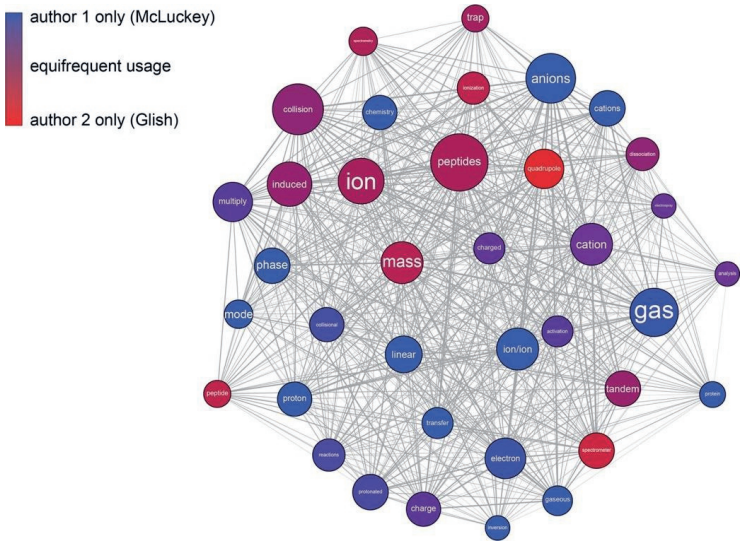


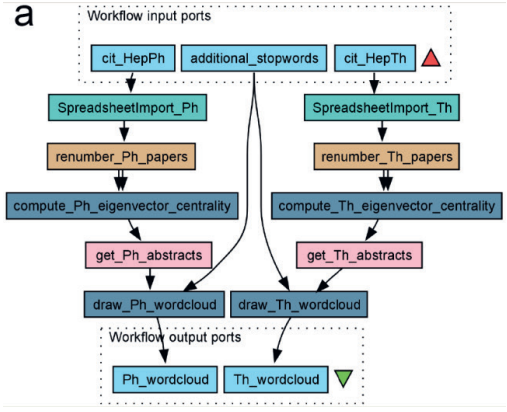
Figure 2.4. Co-word map output from the *Compare_two_authors* workflow. Graphical output from an Rshell can either be in PNG format to be captured by Taverna and browsed in the Results perspective or in a vector format such as SVG or PDF. A workflow can start from data available online, orchestrate all processing and analysis steps, and produce figures or charts in PDF, suitable for publication. Such workflows enable readers to replicate exactly what was published by another researcher with a few mouse clicks. As in the histogram in Figure 2.3, *red* stands for Glish and *blue* for McLuckey.

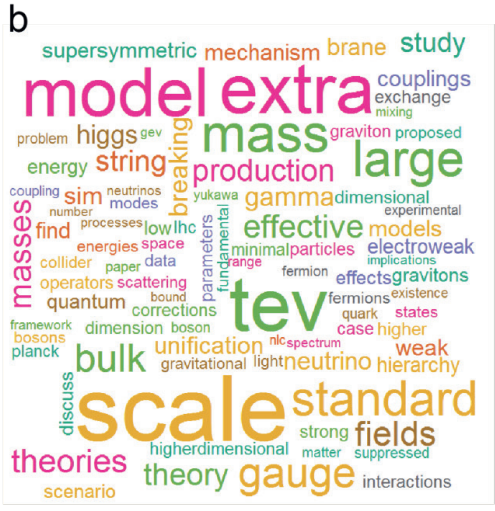
Citation analysis

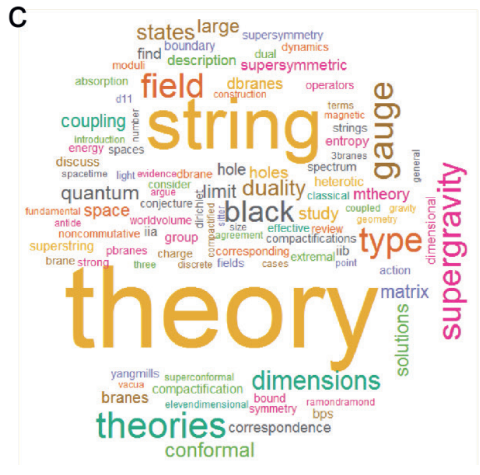
As a second example, we will use a Taverna workflow to analyze citation networks. Citation networks are widely used in bibliometrics to study patterns of who-cites-whom and to study associations between academic groups or areas of research. To simplify the example, we will start from existing networks and compare the citation networks cit-HepPh and cit-HepTh[15], which show the citation relations between all papers published in the e-print archive *arXiv* between January 1993 and April 2003 on high-energy physics phenomenology (cit-HepPh) and high-energy physics theory (cit-HepTh). Specifically, we compare the eigenvector centrality of the papers in both networks. The eigenvector centrality is a measure of the importance of a node in the network and depends on the centrality of the nodes it is linked to[16]. The workflow (Figure 2.5a) takes as input the citation graphs as edgelist files, as available from the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/).

SpreadsheetImport services are used to read the edgelist tables, skipping the first four header lines of each. The arXiv paper identifiers are renumbered consecutively, starting from zero, for improved compatibility with igraph. The indispensable components in the workflow are the Rshells compute_eigenvector_centrality, which calculate the eigenvector centrality using the igraph evcent() function. In this instance, directionality is ignored by specifying 'directed = FALSE'. The output of the Rshells are the original arXiv paper identifiers for the N papers with the highest eigenvector centrality in each network. Beanshell components are then used to create a query and fetch the abstract page of these papers directly from arXiv using the Taverna Get_Web_Page_from_URL service. Embedded in the abstract extraction services are also XPath components that extract the abstract texts from the HTML files. The corpora are then passed to a pair of Rshells for drawing wordclouds on common words in these two extreme sets of abstracts using the tm and wordcloud R packages. The output of the workflow shows the word clouds for the $N = 100$ most central papers in the cit-HepPh (Figure 2.5b) and cit-HepTh (Figure 2.5c) citation networks. The phenomenology word cloud includes physical units, such as TeV, and experimental facilities such as the LHC particle accelerator. The theory word cloud, perhaps unsurprisingly, is dominated by "string", "theory", and the related terms "M-theory", "supersymmetry", "elevendimensional", and so on. Using citation analysis and comparing measures of centrality in two citation networks distills the essential difference between two closely related fields—here two aspects of high-energy physics. Units of measurements have previously been shown to have the weakest co-occurrence coupling with terms such as "theory", "model" and "simulation" in the field of analytical chemistry[17].

Figure 2.5. Citation networks as defined by eigenvector centrality. Taverna workflow for citation analysis **(a)**, wordclouds for 100 core papers in the high-energy physics phenomenology **(b)** and theory **(c)**







Connecting to web services and external databases

As shown in the previous example, Taverna workbench can automatically analyze or generate networks directly from online data. Taverna can also invoke Web Services Description Language (WSDL) style Web services given the URL of the service's WSDL document. The WSDL is an XML-based interface description language often used together with a Simple Object Access protocol (SOAP) to access the functions and parameters of a service. Many bibliographic resources are available through Web services, such as Web of Science (WoS) or PubMed Central (PMC). Some services, including the WoS, require authentication. An entire bibliometric study can be contained inside a single Taverna workflow that authenticates the user, if needed, takes the user queries, or questions of the study, generates the Web service requests, executes these, retrieves the data and proceeds with further (local) statistical analysis and visualization.

A Taverna workflow that invokes WSDL services from WoS to automatically execute a query may look like in Figure 2.6. This Taverna workflow takes as input common search parameters and a generic WoS query string, and pass these to the Web service via the WoS WSDL interface. Values that have only one possible value, such as the language (English, "en") are here hard-coded in the workflow as *Text constants*.

A workflow that connects to the EBI Europe PubMed Central (PMC) SOAP Web service and maps the author affiliations article by article, ordered by publication year, is part of the workflow shown in Figure 2.7. The output of the entire workflow is a world map showing the geographic trends collaborative patterns of an individual researcher. The workflow can easily be adapted to show geographic trends in research topics, publications in a particular journal etc. All that needs to be modified are the PMC search query and the XPaths, and this can be done in a few mouse clicks without typing any code.

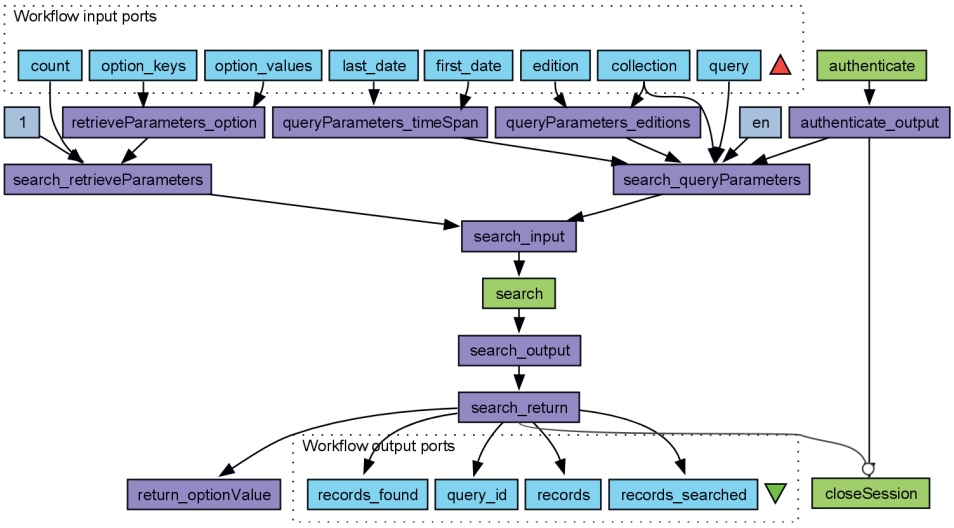


Figure 2.6. Taverna interface to the Thomson Reuters Web of Science Web services lite. This Web service has a relatively complex WSDL interface and also requires authentication. Taverna reveals the WSDL interface allowing the user to understand what is required by, and what can be retrieved from, the service. The port names are the same as in the Thomson Reuters Web service documentation.

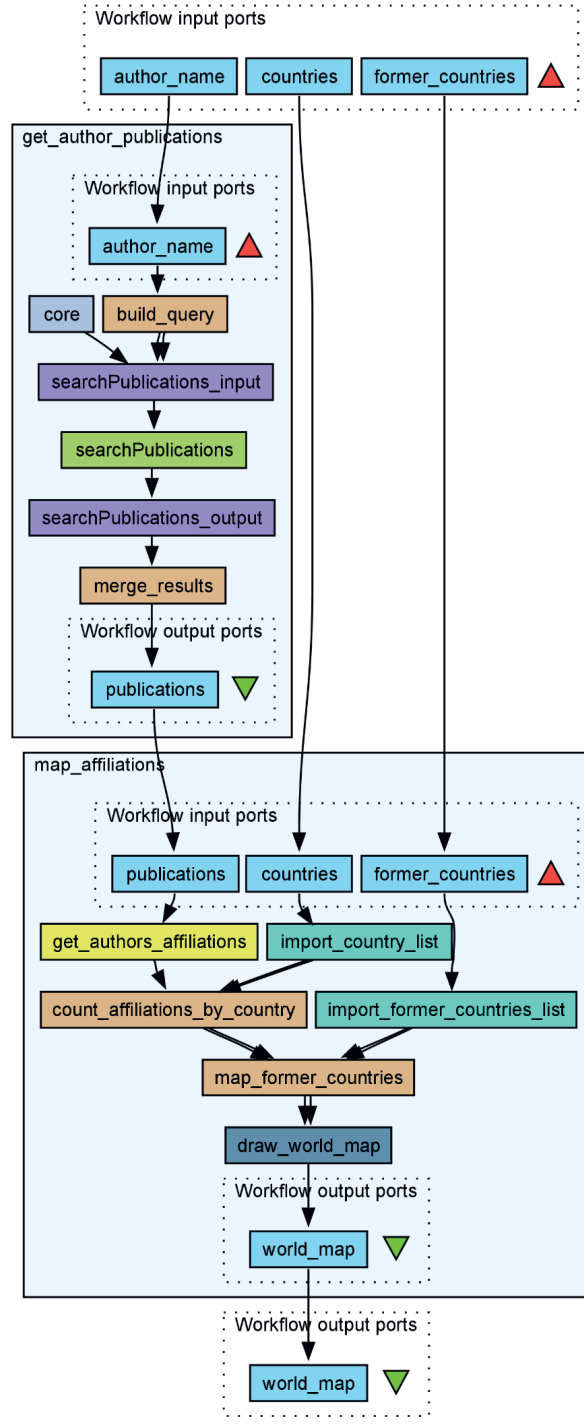


Figure 2.7. A simple workflow demonstrating retrieving bibliometric data using the EBI Europe PubMed Central WSDL Web service (*wild willow green*). Taverna handles the interface to the Web service, creating the SOAP request and retrieving selected results (*medium purple*). The output from this workflow is a world map showing the geographic distribution of collaborators. Alternatively, a time-lapse movie can be created using the "animation" package in R to show how the collaborations change over time.

Geographic analysis of publications

Using the rworldmap package described above, we constructed another simple example workflow, *Compare_pubmed_results_geographically*, to project author affiliations onto a map of the world, displaying the number of publications on a particular topic per country (Figure 2.8a). This example highlights how geographical (country) information can be extracted from the affiliation field in PubMed XML, matched to present-day countries in the ISO 3166-1 standard while transferring data from former countries (as defined in ISO 3166-3) to their successor states. This process works relatively well for publications later than ca. 1949, after we provided the workflow with a table linking former countries with their contemporary counterparts. The latter will obviously never be a perfect process, and some arbitrariness is unavoidable. For example, should research output from the former USSR be shared equally (on the map) between all fifteen independent states that emerged after the dissolution of the Soviet Union, or exclusively to the Russian Federation? Should it depend on where the authors were located at the time? Some borders, such as that between the former West and East Germany, have disappeared from the map in rworldmap. However, for visualization of research activity in the past two decades, rworldmap does the job well. rworldmap also allows some control of granularity and what area of the globe to plot. For example, Antarctica and small islands can be omitted without appreciable loss of accuracy. There are currently no human inhabitants on the Bikar Atoll in the Pacific, let alone research institutes.

The workflow in Figure 2.8a takes a PubMed XML, extracts all author affiliations and maps these to present-day countries in ISO 3166-1, tallies the publications and maps the total number per country onto a current map of the world. This workflow is also available on myExperiment[18]. The results from running this workflow on the topic defined as all articles matching “mass spectrometry” in their title or abstract published between 2010 and 2015 is shown in Figure 2.8b. As an alternative to starting from a PubMed XML file, we can connect the output from the PMC Web service as input to Compare_pubmed_results_geographically (Figure 2.7). This combined workflow is also available on myExperiment. In addition to producing static maps, it is also possible to export a series of author affiliation maps as a movie using the “animation” R package.

Journals covering in the same scientific field may have regional bias, with for example researchers based in the US preferentially publishing in an American journal and European researchers preferring a European journal. To investigate whether there is such a bias in the field of medicinal chemistry, we looked specifically at the *Journal of Medicinal Chemistry* (published by the American Chemical Society) and the *European Journal of Medicinal Chemistry*. To this end we assembled the workflow shown in Figure 2.9a. This workflow analyzes the geographical bias in author affiliation between any two journals, not just the two investigated here. The output is again a map generated by rworldmap, this time with a color gradient representing the relative number of publications in the two journals for each country (Figure 2.9b). It is clear from this analysis that authors from many Western European countries have a preference for the American journal. This may have something to do with this journal having a higher journal impact factor (as measured by the Thomson Reuters journal impact factor) and consequently being considered more prestigious in the field. On the other hand, other Western European countries such as France and Italy do not show this preference. may be explained by the fact that a sizeable share of the editorial board is comprised of researchers working in France or Italy.

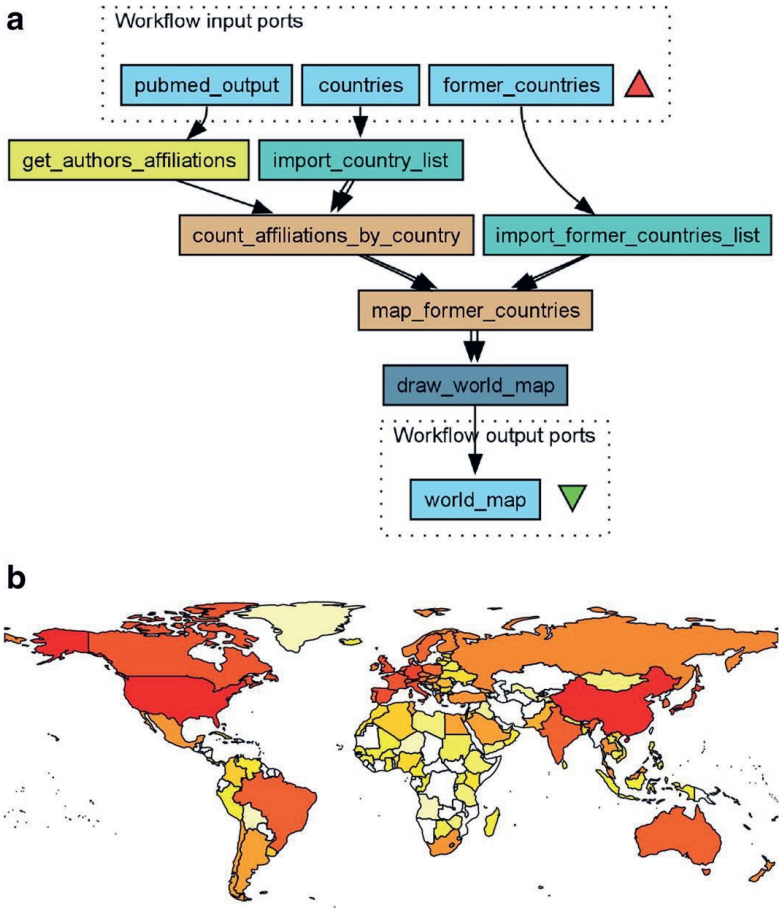


Figure 2.8. a. PubMed XML output contains information on author affiliations as provided by the authors themselves. This Taverna workflow extracts geographic information (here countries) and converts it to a standardized format (ISO 3166) from the PubMed XML output. The workflow counts the number of appearances of each country in the author affiliations in the XML file and uses the R package ''rworldmap'' to visualize them. rworldmap and similar tools require country names to be in a standard format, e.g. the three letter code from ISO 3166. The text mining component is therefore necessary to connect PubMed with geographic visualization. **b.** Output of the workflow for the search string ''(mass spectrometry[Title/Abstract]) AND (''2010/01/01''[Date-Publication]: ''2014/12/31''[Date- Publication])'' in PubMed, showing the geographic distribution of active (and actively publishing) researchers in the field of mass spectrometry in the past 5 years.

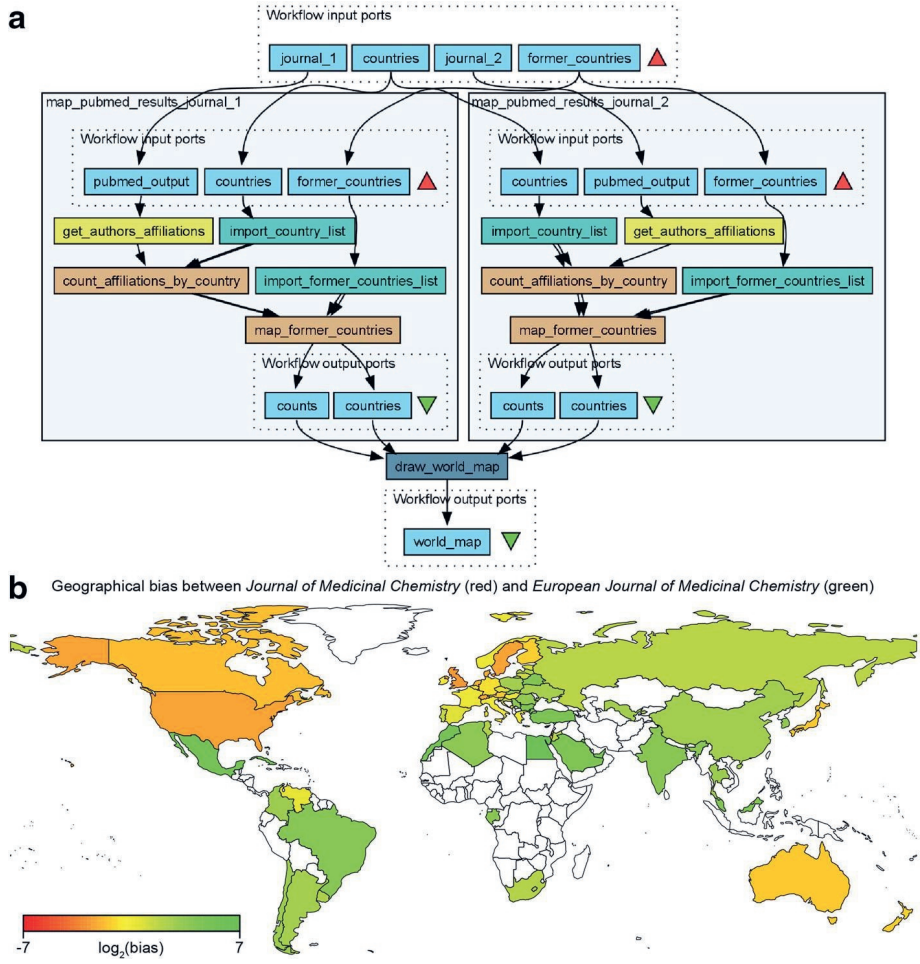


Figure 2.9. a. A workflow *Compare_two_journals_geographically* reusing the embedded *map_affiliations* workflow matching author affiliations with countries in Figure 2.6 for analyzing geographical bias in two medicinal chemistry journals: the American Chemical Society (ACS)-published *Journal of Medicinal Chemistry* and the European Journal of Medicinal Chemistry between 2000 and 2015. **b.** The results of the workflow above with publication bias shown as color from *red* to *green* representing a bias of a factor $2^7 = 128$ in publishing in the ACS over the European journal. The numbers of publications were normalized to the total number of articles in the two journals (11,219 articles in *Journal of Medicinal Chemistry* and 5842 articles in the *European Journal of Medicinal Chemistry* respectively). The most recent Thomson Reuters journal impact factor is 5.447 for the ACS journal (2014) and 3.432 for the European journal (2013), respectively.

Discussion and conclusions

The use of scientific workflows in bibliometrics is still in its infancy. The direct support of R inside Taverna workflows is particularly useful for bibliometrics and scientometrics. A number of R packages for bibliometric analysis have recently been released, ranging from simple data parsers such as the "bibtex" package[19] for reading BibTeX files to libraries or collections of functions for scientometrics, such as the *CITAN* package[20]. The latter package contains tools to pre-process data from several sources, including Elsevier's Scopus, and a range of methods for advanced statistical analysis. The igraph package itself comes with some functions specifically for bibliometric analysis, e.g. "cocitation" and "bibcoupling". Clustering or rearranging the graph spatially so that strongly connected words appear closer together is possible with igraph, but may also be assisted by other packages. We opted for showing a few simple but more or less representative examples here. Much more complex analyses can be designed based on or using the workflows and components here as a starting point. We did not include any advanced text mining functionality for homonym disambiguation or natural language processing. The "openNLP" R package currently in development provides an interface to openNLP[21] and may be used to extract noun phrases and refine the analyses.

In the examples here, we could show that individual language preferences can dominate when comparing two authors working in the same field. We could also show that the geographical bias between two medicinal chemistry journals, one European and one published by the American Chemical Society, probably has more to do with impact factor and perceived prestige than author location, based on the observation that researchers from the European countries usually ranking high in international research surveys, i.e. Denmark, the Netherlands, Sweden, Switzerland and the United Kingdom, also have the strongest preference for publishing in the higher-impact factor American journal. To the extent that such rankings are based on impact factors, this is of course in part a circular argument. We also observe that European countries well represented on the editorial board of the European journal, e.g. France and Italy, show no preference for the American journal. This is probably not a coincidence.

Scientific workflow managers are powerful tools for managing bibliometric analyses, allowing complete integration of online databases, Web services, XML

parsers, statistical analysis and visualization. Workflow managers such as Taverna eliminate manual steps in analysis pipelines and provide reusability and repeatability of bibliometrics analyses. All workflows for bibliometrics and scientometrics presented here can be found in the myExperiment group for Bibliometrics and Scientometrics (https://edu.nl/cag4d).

Acknowledgments

The authors would like to thank Thomson Reuters for granting access to the Web of Science Web services lite and Dr. Yassene Mohammed (LUMC) for technical assistance with Taverna workbench.

References

1. Gil, Y. From data to knowledge to discoveries: Artificial intelligence and scientific workflows. *Sci. Program.* **17**, 231–246 (2009).
2. de Bruin, J. S., Deelder, A. M. & Palmblad, M. Scientific Workflow Management in Proteomics. *Mol. Cell. Proteomics* **11**, M111.010595 (2012).
3. Berthold, M. R. *et al.* KNIME-the Konstanz information miner. in *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization* (eds. Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R.) (Springer Berlin, Heidelberg 2008).
4. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
5. Oinn, T. *et al.* Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054 (2004).
6. Guler, A. T., Waaijer, C. J. F. & Palmblad, M. Scientific Workflows for Bibliometrics. in *Proceedings of ISSI 2015 Istanbul: 15th International Conference on Scientometrics & Informetrics Conference, June 29-July 3, 2015, Istanbul, Turkey* 1029–1034 (2015).
7. Stevens, R. D., Robinson, A. J. & Goble, C. A. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **19**, i302–i304 (2003).
8. 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).
9. Feinerer, I., Hornik, K. & Meyer, D. Text Mining Infrastructure in R. *J. Stat. Softw.* **25**, 1–54 (2008).
10. Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E. & van Atteveldt, W. RTextTools: Automatic text classification via supervised learning. *R package version 1.4.2* https://cran.r-project.org/web/packages/RTextTools/index.html (2014).
11. Grün, B. & Hornik, K. topicmodels : An R Package for Fitting Topic Models. *J. Stat. Softw.* **40**, 1–30 (2011).
12. Fellows, I. wordcloud: Word Clouds. *R package version 2.4* https://cran.r-project.org/web/packages/wordcloud (2013).
13. South, A. rworldmap : A New R package for Mapping Global Data. *R J.* **3**, 35–43 (2011).

14. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695,** 1–9 (2006).

15. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graphs over time. in *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining, August 21-24, 2005, Chicago, Illinois, USA* 177–187 (2005).

16. Bonacich, P. Some unique properties of eigenvector centrality. *Soc. Networks* **29**, 555–564 (2007).

17. Waaijer, C. J. F. & Palmblad, M. Bibliometric mapping: Eight decades of analytical chemistry, with special focus on the use of mass spectrometry. *Anal. Chem.* **87**, 4588–4596 (2015).

18. Goble, C. A. *et al.* myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* **38**, W677–W682 (2010).

19. Francois, R. bibtex: bibtex parser. *R package version 0.4.0* https://cran.r-project.org/package=bibtex (2014).

20. Gagolewski, M. Bibliometric impact assessment with R and the CITAN package. *J. Informetr.* **5**, 678–692 (2011).

21. Hornik, K. openNLP: Apache OpenNLP Tools Interface. *R package version 0.2-3* https://cran.r-project.org/package=openNLP (2015).

CHAPTER 3

3

Automating Bibliometric Analyses Using Taverna Scientific Workflows

Arzu Tugce Guler[1], Cathelijn J. F. Waaijer[2], Yassene Mohammed[1], Magnus Palmblad[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

[2] Faculty of Social and Behavioural Sciences, Centre for Science and Technology Studies, Leiden University, The Netherlands

This chapter has been published in:
Journal of Informetrics (2016) 10, 830-841 doi: 10.1016/j.joi.2016.05.00

Abstract

Quantitative analysis of the scientific literature is a frequent task in bibliometrics. Several large online resources collect and disseminate bibliographic information, paving the way for broad analyses and statistics. The Europe PubMed Central (PMC) and its Web Services is one of these resources, providing a rich platform to retrieve information and metadata on scientific publications. However, a complete bibliometric analysis that involves gathering information and deriving statistics on an author, topic, or country is laborious when consuming Web Services on the command-line or using low level automation. In contrast, scientific workflow managers can integrate different types of software tools to automate multi-step processes. The Taverna workflow engine is a popular open-source scientific workflow manager, giving easy access to available Web Services. In this tutorial, we demonstrate how to design scientific workflows for bibliometric analyses in Taverna by integrating Europe PubMed Central Web Services and statistical analysis tools. To our knowledge, this is also the first time scientific workflow managers have been used to perform bibliometric analyses using these Web Services.

Introduction

As science becomes more data intensive, access to data and the process of generating meaningful information from them become the main vehicle in the scientific process. In this process, the primary challenge is moving from generated or retrieved data to information. As in most fields, typical bibliometric analysis workflows require several discrete steps, each employing different software tools. Frameworks that allow users to efficiently but easily connect data access points to information generation play a key role here. However, it is not always straightforward to use a generic framework or design custom workflows every time a new analysis protocol is to be implemented. In the absence of a framework, users have to manually connect the inputs and outputs of individual steps through the entire analysis. This risks introducing errors and makes analyses difficult to reproduce, especially for other researchers.

Scientific workflow managers integrate several processing units to automate a data analysis procedure. They are field-independent, so analysis on data from any field, including bibliometrics, can be automated. Scientific workflows typically have inputs and outputs, where series of operations are performed on the inputs in order to produce the outputs. Thus various atomic processing units can be assembled to produce an analysis protocol that can run without manual intervention[1]. On the other hand, reusability and reproducibility are also important for *in silico* experiments, facilitating collaboration and combining efforts. These are promoted by online scientific workflow repositories such as myExperiment[2]. However, deciphering the hierarchical composition of a workflow, its control and connections could be difficult in a larger-scale workflow[3]. Taking a modular approach and defining the scope of each module in the workflow eases this process. Most of the freely available scientific workflow managers have a graphical user interface that helps to visualize the overall protocol, both when designing and when executing the workflow. Galaxy[4], KNIME[5] and Taverna[6] are popular examples of such scientific workflow managers that also allow modular design. Automating an analysis consisting of several steps, such as in bibliometrics, using scientific workflow managers makes the process less laborious and decreases the risk of human errors. Scientific workflow managers follow a different paradigm than interactive software tools, such as the domain-specific (or

perhaps domain-limited) BibExcel[7], Publish or Perish[8] and Sci2[9] though Sci2 certainly provides some aspects of the modularity and tool integration of the workflow managers.

We have previously presented how scientific workflows can be used to solve simple bibliometrics problems, using Taverna Workbench[10]. Like any other scientific workflow manager, Taverna enables the user to integrate different types of components. What makes Taverna very useful for bibliometrics is that it already provides custom support for a number of tools and services that are easily adopted for performing such analyses, *e.g.* R tools and XPath, Beanshell and WSDL services. The programming language R is primarily developed for statistical computing and visualization. Specific R plug-ins or packages expands its capabilities to machine learning, text mining and natural language processing[11,12]. The XPath service is a user-friendly tool for creating XPath queries to parse XML documents by simply selecting nodes from an XML tree with a few mouse clicks. This is highly convenient, as most bibliometric databases can export information in XML format. For tabular formats, the Spreadsheet import service provides a similarly minimalistic tool for parsing tables. For general tasks, Beanshell services allow inclusion of scripts using a Java-like language. Last but not least, integrated support for Web Services allows Taverna workflows to directly communicate with remote databases using WSDL queries[13]. As most Web Services use XML as the preferred message format, the Taverna XPath service is typically used to parse the results returned from Web Service calls.

An important functional aspect of Taverna is that iterations over individual processes or parts of the workflows are done implicitly by list handling. This feature provides great flexibility if a process or a sub-workflow has more than one input port. The user can specify whether the inputs are subjected to a “cross product” (all list elements in one input against all list elements in the other input) or a “dot product” (element-wise), or for processors with more than two inputs a combination of both; all while being able to define the order and precedence of the workflow operations on these input lists. A core set of built-in features and services provides basic list handling, such as flattening, merging a list to a string and removing duplicates.

Here we present a tutorial on how to use Taverna to build workflows that interact with the Europe PubMed Central Web Services. In principle, Taverna could interact

with any Web Service that provide a SOAP or RESTful interface. The reason we are demonstrating the integration of Web Services in Taverna using PubMed rather than Scopus® or Web of Science™[14] is that, among these three, PubMed is currently the only that provides a free Web Service interface. PubMed is also the most used bibliographic resource in the life sciences. In this tutorial, we show how to retrieve information using Web Services, how to parse this information, and how to use the various built-in Taverna services and processors to calculate and visualize the results. In principle, the same approach could be taken using other resources, provided that the user has access to them. We also made an example Taverna interface for connecting to the Thomson Reuters Web of Science™ Web Services and made this available on myExperiment (https://edu.nl/gcxpg). We have built and tested the workflows in Taverna Workbench Bioinformatics 2.5.0, but in principle the workflows should run in any flavor of Taverna Workbench version 2.4.0 or later. For instructions on how to download and install Taverna, see https://edu.nl/6nhtk. For Rshells to be executable in Taverna, R, RServe and required R packages must be installed and deployed[15].

Getting started: connecting to Europe PMC Web Services

Europe PubMed Central, or PMC (http://europepmc.org) is one of the leading databases for peer-reviewed life science literature, providing access to 30.4 million abstracts and 3.3 million full-text articles and metadata (December 14, 2015). The goal of Europe PMC is to "build open, full-text scientific literature resources and support innovation by engaging users, enabling contributors and integrating related research data"[16]. This is achieved by providing access through a user-friendly Web interface, FTP, and SOAP and RESTful Web Service APIs. Here we will use the latter from within Taverna workflows. This is done as follows. First, the Europe PMC SOAP-based Web Services are imported into Taverna using "Import new services" in the Design pane using the WSDL https://www.ebi.ac.uk/europepmc/webservices/soap?wsdl. The available Web Services should now be listed as available in Taverna services menu. The 55-page Europe PMC SOAP Web Service Reference Guide[17] describes all details of the API to these newly imported services. Although strongly recommended, it is not absolutely necessary to read the entire manual before starting to integrate Europe PMC Web Services from within Taverna. A Web Service

component is simply added to a workflow by dragging it from the service menu and dropping it into the workflow whiteboard. To expose the component's inputs and outputs, we add XML splitters. These are found in the component Edit menu. For example, the *searchPublications* service currently has six input ports: *email*, *offset*, *pageSize*, *queryString*, *resultType* and *synonym*. Of these, only the *queryString* is mandatory. This string corresponds to what one would normally enter in the search field on the Europe PMC website. The *email* address registers the user with Europe PMC, the *pageSize* the number of entries to be retrieved in one page, the *offset* refers to which page of size *pageSize* to retrieve, *synonym* whether to expand the query using the MeSH and UniProt synonyms. The *resultType* is used to limit the retrieval to the data we want. It has three settings: *idlist*, *lite* and *core*. If we only want the PubMed IDs (PMIDs) for subsequent queries, *idlist* would be sufficient. The *lite* results contain key metadata such as the author list and basic bibliographic information, and *core* all metadata, including abstracts and full journal details. The full article, if in Europe PMC, is retrieved using another service, *getFulltextXML*. The workflow in Figure 3.1 illustrates the use of the *searchPublications* service with its input and output XML splitters. The workflow performs a single search similar to using a Web browser and the Europe PMC Website. The *results* is an XML tree with the first 100 results of the Europe PMC search defined by *query* where the number 100 is defined by *records_to_retrieve* constant. This workflow is available on myExperiment (https://edu.nl/wef8y). In Figure 3.1, all input and output ports of every workflow components are shown. In subsequent workflows, the ports details are hidden for simplicity. However, these can easily be displayed in Taverna workbench.

The Europe PMC results are retrieved in XML, and the extraction of the precise information we want are done by further XML output splitters or XPath services in Taverna. An XPath is a query written in the XPath language for selecting elements and attributes in an XML document. XPath allows postfix conditional statements within square brackets. For example, to restrict the results to PMIDs of cited papers (having a *citedByCount* larger than 0), the XPath */resultList/result[citedByCount>0]/pubYear* could be used on the output of the workflow in Figure 3.1 to retrieve the publication year (*pubYear*) for cited papers only. The XPath service in Taverna provides a configuration pane to automatically generate simple XPath expressions, which the

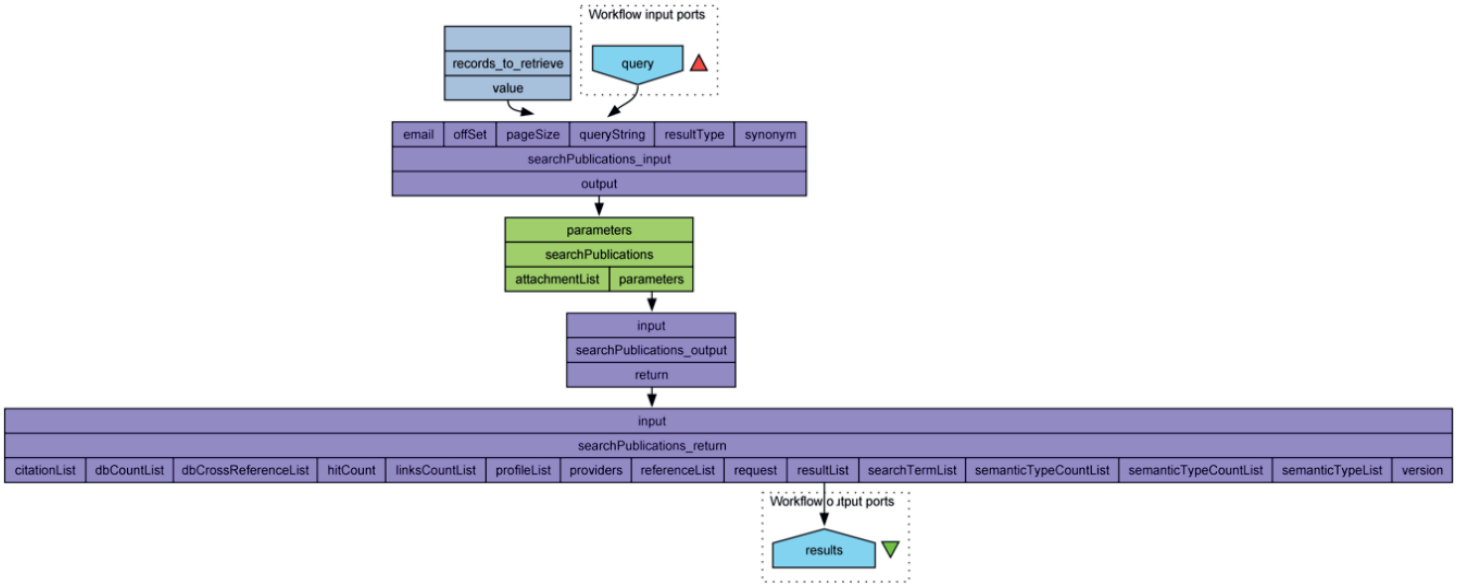


Figure 3.1. Basic workflow to access Europe PMC Web Services

user can then customize, for example by adding conditionals, or combining several expressions. The results of the Web Service and XML parsers can be passed to other workflow components either as text or as XML. A second workflow attaching an XPath statement to the workflow in Figure 3.1 is also available on myExperiment (https://edu.nl/pwyqv). The output of the workflow, after parsing the *searchPublications* output *resultList* with the XPath above, is a Taverna list of the publication years of the cited articles among the 100 first retrieved articles matching the search query *query*. As mentioned in the introduction, Taverna does iteration implicitly using lists. If a component for performing a certain task is given a list as input, the task will be performed on all elements in that list.

Publication records and citation networks

From these simple first steps, and using the same types of components, we will now construct more advanced workflows exploring the full power of the Europe PMC Web Services and Taverna. We do this using the notions of embedding and extensibility of scientific workflows. A simple workflow can be embedded in more complex workflows. Existing workflows, shared in the myExperiment community, can be accessed from the myExperiment pane in Taverna and modified or extended according to the user's needs.

The well-known Thomson Reuters Web of Science™ search provides a link to a "Citation Report" with two histograms, one over the number of published items in each year and one over the citations for these items in each year, based on the search results. In addition, the Citation Report provides simple statistics, such as average citations and the *h*-index for these search results (the *h*-index may be most relevant for a single author name search, but is calculated and reported for any set of publications). We can produce similar histograms based on the Europe PMC database using a Taverna workflow. For this, it is necessary to use two Web Services, *searchPublications* as before, and *getCitations* to get the publication year of papers citing the papers returned by the *searchPublications* query (for example on an author name). Figure 3.2 shows such a workflow, which is also available on myExperiment (https://edu.nl/uddhg). The workflow uses two Europe PMC Web Services:

searchPublications and *getCitations* to generate publication statistics in the form of a "citation report" for a particular author.

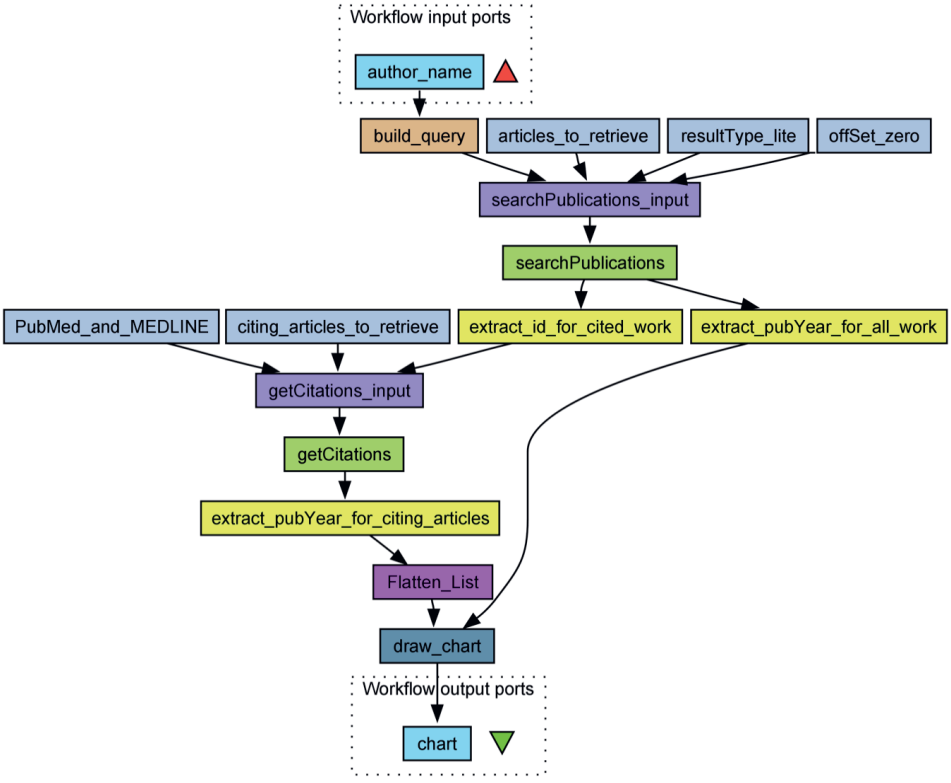


Figure 3.2. A workflow to generate simple statistics of the citation related to a specific author

The workflow in Figure 3.2 takes as input the full name of an author. This argument is passed to a Java BeanShell *build_query* that constructs the specific query "auth:\" " +author_name+"\" sort_date:y". Combined with a value 1,000 for the number of *articles_to_retrieve*, this will request the 1,000 most recent publications (sorted by date) for the author or authors matching *author_name*. The list of PMIDs returned by *searchPublications* is used as input to *getCitations*, which returns a list of lists of publication years for the papers citing the papers returned by *searchPublications*. In this workflow, XPaths are directly applied on the Web Services results. This skips the two XML output splitters and simplifies the visual appearance of the workflow. Whether to use output splitters and short XPaths, or longer XPaths directly on the

Web Service output, is mostly a matter of taste. The XPath extracting the *pubYear* for the citing articles produces a list of lists of publication years as output. In order to make a combined histogram over all citations to all papers from the author, we flatten this list of lists of publication years to a single list using the built-in *Flatten_List* local service. This single list of publication years is then passed to an Rshell component *draw_histogram* as data of the (semantic) type "integer vector", as specified in the input port to this workflow component. The integer type in R exists to pass data to programs written in strongly typed languages that expects them, and so that integer data can be represented "exactly and compactly"[18]. In this workflow, the publication years could just as well be passed as "numeric" vectors. The Rshell is very simple and uses the hist() function[19] to generate the two histograms. For authors having a unique identifier, such as an ORCID, the *build_query* can be changed to "authorid:\"" + author_id + "\" sort_date:y".

An output of this workflow for the author "Jonas Bergquist" (Professor Jonas Bergquist, Department of Chemistry - Biomedical Centre, Uppsala University, Sweden) is shown in Figure 3.3. An extended version of this workflow is available on myExperiment (https://edu.nl/ptexf) that combines the publications and citations records in a two-dimensional heatmap showing the delay, increase and decrease of citations for papers over time. The workflow can easily be extended to accept a list of authors rather than a single author, generating either combined statistics or individual citation reports for each author in the list.

Suppose instead we are interested in who is cited by whom or citing the work of a particular researcher and how these authors in turn cite each other. To put it more simply: we would like to construct and visualize a co-citation network based on one researcher. Any network consists of multiple items (vertices or nodes) and their underlying relationships (edges). In the case of our co-citation network, the vertices are the single researcher and the authors cited by or citing this particular researcher. The edges are all the citations to and from the researchers in the co-citation network.

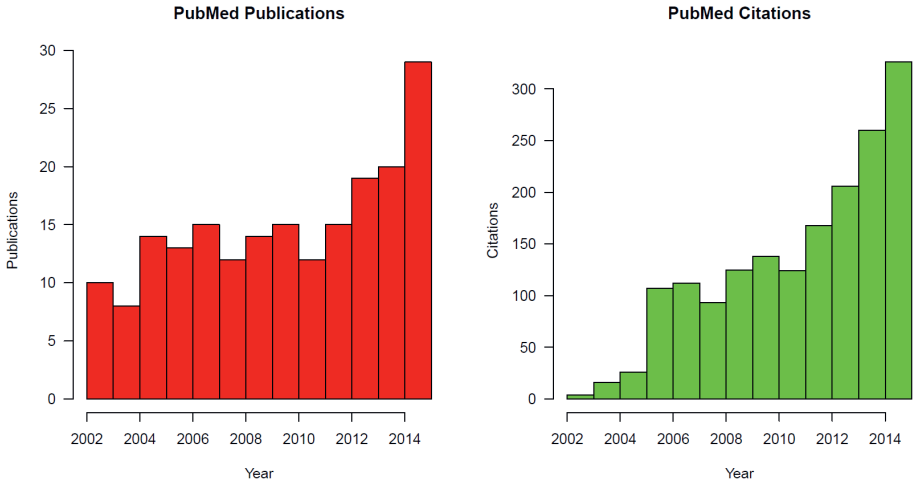


Figure 3.3. Citation report for an author ("Jonas Bergquist") generated by the workflow in Figure 3.2.

Constructing and visualizing this co-citation graph requires a slightly more elaborate workflow. Dividing up the task into smaller ones, we first look up all publications for the author using *searchPublications*. For each publication in the returned list, we then look up the references and citations in parallel using *getReferences* and *getCitations* respectively. These three requests returns all vertices in our co-citation graph, but does not retrieve citations, or edges, between papers by other authors. To retrieve these we combine all the vertices and call *getReferences* and *getCitations* again, once for each vertex. If the reference or citation is already represented by a vertex in the graph, we add a new edge to or from that vertex. To make the analysis more interesting, we can also have the workflow keep track of the author's own papers and self-citations. The best way to do this is by defining attributes to the edges and vertices in the co-citation graph. The workflow in Figure 3.4 does this by generating a description of the graph in Pajek[20] format using the BeanShell *combine_and_make_Pajek_file*. In most workflows, one would normally strive to use stream data between components or use simple tabular of XML file formats. When dealing with graphs, however, it is sensible to use a common format for defining graphs, such as GraphML[21], GML[22], LGL[23] or Pajek. The workflow in Figure 3.4 finds all papers citing and cited in articles published by an author, and all citations between them. The workflow generates a citation network graph that is captured by

and displayed inside Taverna. The workflow also generates a Pajek file incorporating the information on self-citations using different edge attributes (color) and labels the vertices differently for the author (last name and publication year) than for the other vertices (PubMed ID).

The Pajek content created by *combine_and_make_Pajek_file* and written to file by *Write_Text_File* is read by the Rshell *draw_graph* using the igraph R package[24]. The igraph package contains functions for reading and writing graphs in several formats, including those mentioned here. The outputs of the workflow are a simplified graph in Sugiyama layout[25] created by igraph using simplify() and layout.sugiyama(), and the corresponding Pajek file created by write_graph() after simplification. A static but visual representation of the graph is captured by Taverna as well as written to a PDF file. To interactively explore and analyze the graph, the Pajek file can be opened in Pajek or a tool such as the VOSviewer[26], which are both tools for the analysis and visualization of (bibliometric) networks. The Pajek file created by the Taverna workflow (run November 30, 2015) was opened in VOSviewer 1.6.3, showing the largest set of connected items (3,782) out of the 3,851 vertices in this citation graph (Figure 3.5; can also be opened as an interactive Java application by clicking on the https://edu.nl/avgwc). The clustering was performed with clustering resolution 0.05 and minimum cluster size 50. The author's own papers are annotated with first author, last name and year, other papers with PMID. The publication record in the example above, including citations, can be visualized as a several highly interconnected and overlapping clusters, the largest of which (red) is on proteomics. In the center of this large cluster is a highly cited review by Aebersold and Mann on mass spectrometry-based proteomics (PMID 12634793)[27]. The dark blue cluster covers work in psychophysiology and neuroscience, excluding proteomics but including new methods for analysis of cerebrospinal fluid[28,29]. In addition to this core of work in proteomics and neuroscience, we see a few protuberances representing collaborations with researchers in different disciplines, such example veterinary science applications[30] (light brown) and surface chemistry techniques[31,32] (magenta). The full VOSviewer map is also included as supplemental information.

Another way to view the research topics of a particular author is to count words and noun phrases in the titles and abstracts, visualizing the results as a graph or tag

cloud. A workflow using the *searchPublications* Web Service and the R packages tm[11] for text mining and wordcloud for visualization is also available on myExperiment (https://edu.nl/hgwkc).

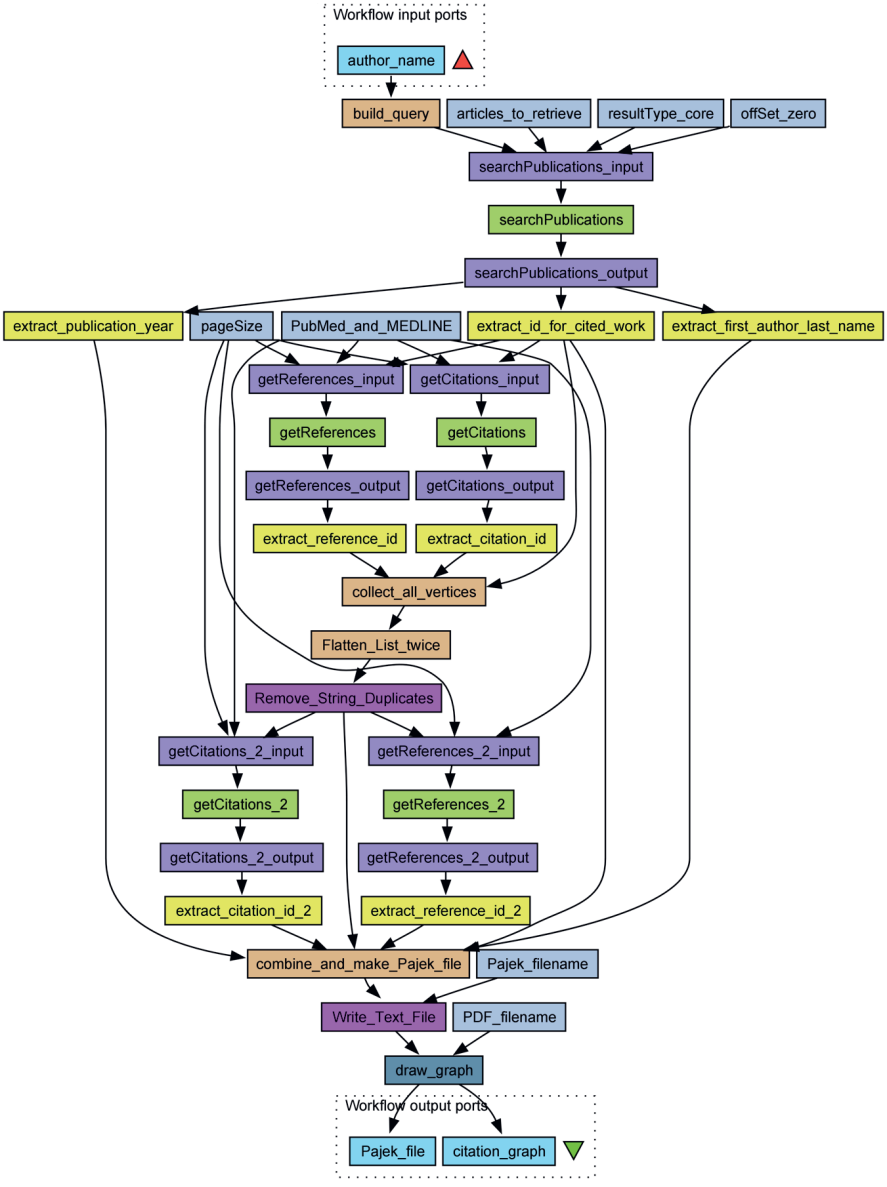


Figure 3.4. A scientific workflow for generating an author citation network

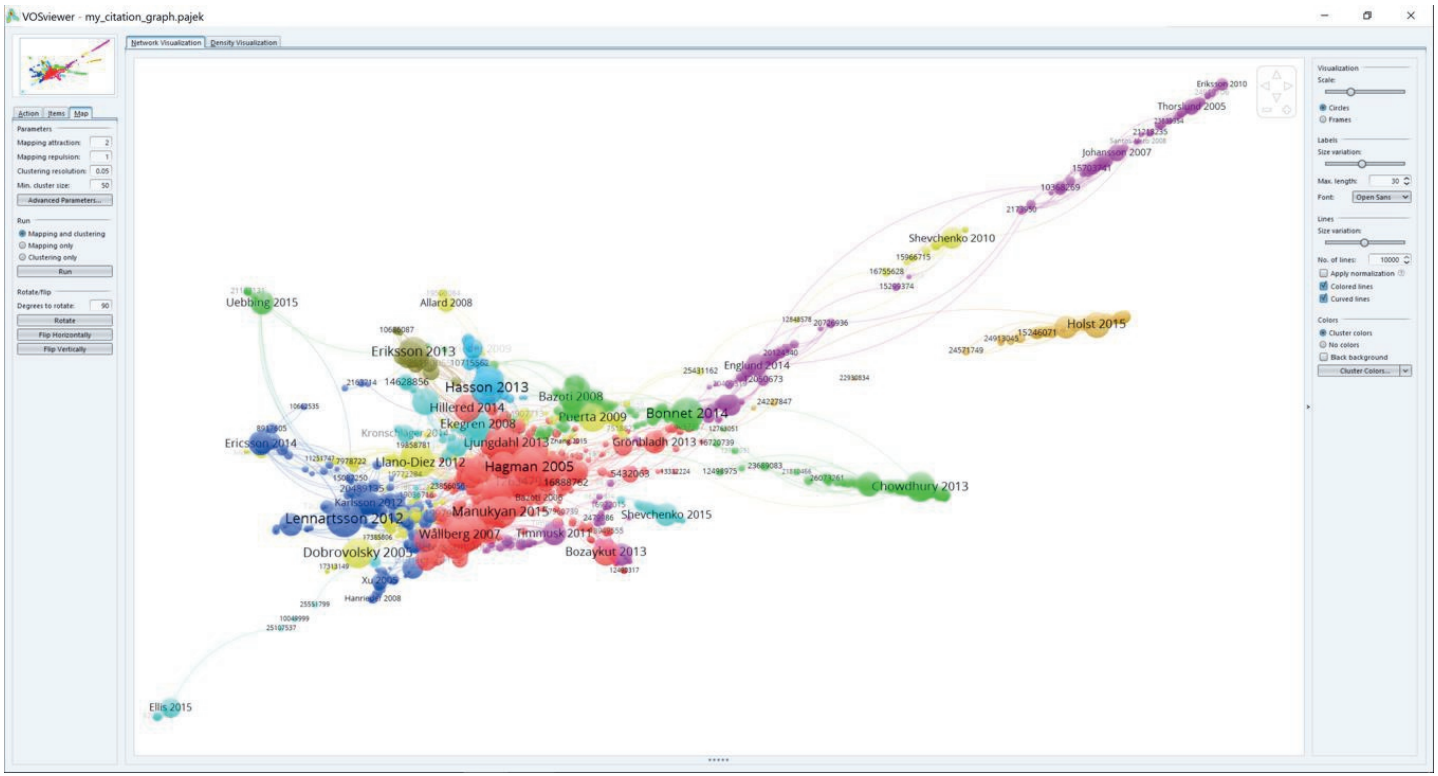


Figure 3.5. Visualization of the citation network for “Jonas Bergquist” using the VOSviewer

Biomolecular interactions

Resources such as UniProt[33], IntAct[34] and the RCSB PDB[35] provide a wealth of curated information on proteins, their functions, interactions and structures. Importantly, they always cite the original source of the information, which most commonly is a peer-reviewed scientific publication. Europe PMC is also cross-referenced to these and several other databases. The *getDatabaseLinks* Web Service is used to access the UniProtKB, IntAct or PDB records associated with an article. This may seem like a trivial service, but is in fact a programmatic access that allows us to explore the scientific literature, not only for bibliometrics, but also to investigate what the publications are about, *e.g.*, the chemical compounds, genes, proteins, diseases or biological species. For example, consider a researcher who is interested in protein *P* and would like to find all proteins mentioned in connection with this protein in the scientific literature in a specific context. This context could be a molecular interaction, being part of the same protein complex, one protein activating the other, etcetera. Most researchers would use databases such as IntAct or UniProtKB to search for this information under the entry for protein *P*. But suppose the researcher wants to look a bit more broadly at what has been reported in the scientific literature but not yet annotated in UniProtKB, IntAct, or any other database as a specific type of protein-protein interaction. This can be accomplished using *searchPublications* and *getDatabaseLinks* in the same workflow (Figure 3.6). The workflow looks up the proteins in UniProtKB most frequently co-occurring in the literature with a query protein and in a specified context, *e.g.*, type of protein-protein interaction or disease. The workflow then builds a network with the proteins as nodes and the weights of the edges corresponding to the number of co-occurrences in the literature.

For simplicity, the input and output port splitters are embedded with the Web Services as Taverna components in the workflow in Figure 3.6. The workflow builds a query string from user provided input to search for a particular UniProt identifier in the context of a certain phrase appearing in the title or abstract. The list of retrieved article identifiers (PMIDs) is then passed to *getDatabaseLinks*, which, like *getCitations*, returns a list of lists of all UniProt identifiers co-occurring in those publications. These may be very few, or number in the thousands for large proteomics studies. In general, we would expect a co-occurrence in a publication with few linked UniProt IDs to be

more relevant than a co-occurrence in a list of several thousand proteins. The results can be weighted using the returned *dbCountList*, or by limiting the number of UniProt IDs retrieved for each PMID to a small number to reduce the influence of proteomic studies. For example, a *searchPublications* query for UniProt ID P29083, or the Transcription factor IIE alpha subunit, with the phrase "complex" in the title or abstract returns a list of PMIDs for 9 publications (November 30, 2015). Passing this list of PMIDs to *getDatabaseLinks* and specifying a pageSize of 10 to retrieve at most 10 identifiers per PMID produces a list of lists with a total of 62 UniProt IDs, of which 53 are unique. The workflow in Figure 3.6 then counts the frequencies of these UniProt IDs and sort them in descending order using sort(table(UniProt_IDs), decreasing = TRUE) in the Rshell *count_frequencies*. The protein most frequently occurring in these lists is the query protein itself (5 occurrences). The runner-up is unsurprisingly UniProt ID P29084 or the beta subunit of the Transcription factor IIE with 3 appearances. Three other UniProt identifiers occur twice and the remainder once. Raising the *pageSize* limit to the maximum allowed 1,000 returns 2,948 identifiers, 2,550 of which are unique. Two sublists from two large-scale proteomics reports[36] reached the maximum of 1,000 UniProt IDs, reporting 2,932 and 5,159 identifiers respectively. The query protein is again in the top (13 occurrences), but the beta subunit is now only in 95th place, still with only 3 co-occurrences.

The network produced by the workflow in Figure 3.6 can be further analyzed in Cytoscape, a common tool for network visualization and analysis in bioinformatics[37]. The workflow output can be opened either directly as GML in Cytoscape. Figure 3.7 shows Cytoscape 3.3.0 with the output from the workflow in Figure 3.6 on Apolipoprotein A-I (UniProt ID P02647) and "complex" as before with the "Edge-weighted Spring Embedded" Cytoscape layout. The cluster of proteins associated with Apolipoprotein A-I was analyzed for enrichment of Gene Ontology biological processes by BiNGO 3.0.3[38]. In Figure 3.7, proteins frequently co-occurring with Apolipoprotein A-I and being involved in "macromolecular complex remodeling" (as well as "protein-lipid complex remodeling" and "plasma lipoprotein particle remodeling") are highlighted in yellow. Again, these results are not surprising given that Apolipoprotein A-I is the dominant protein component of high density lipoprotein

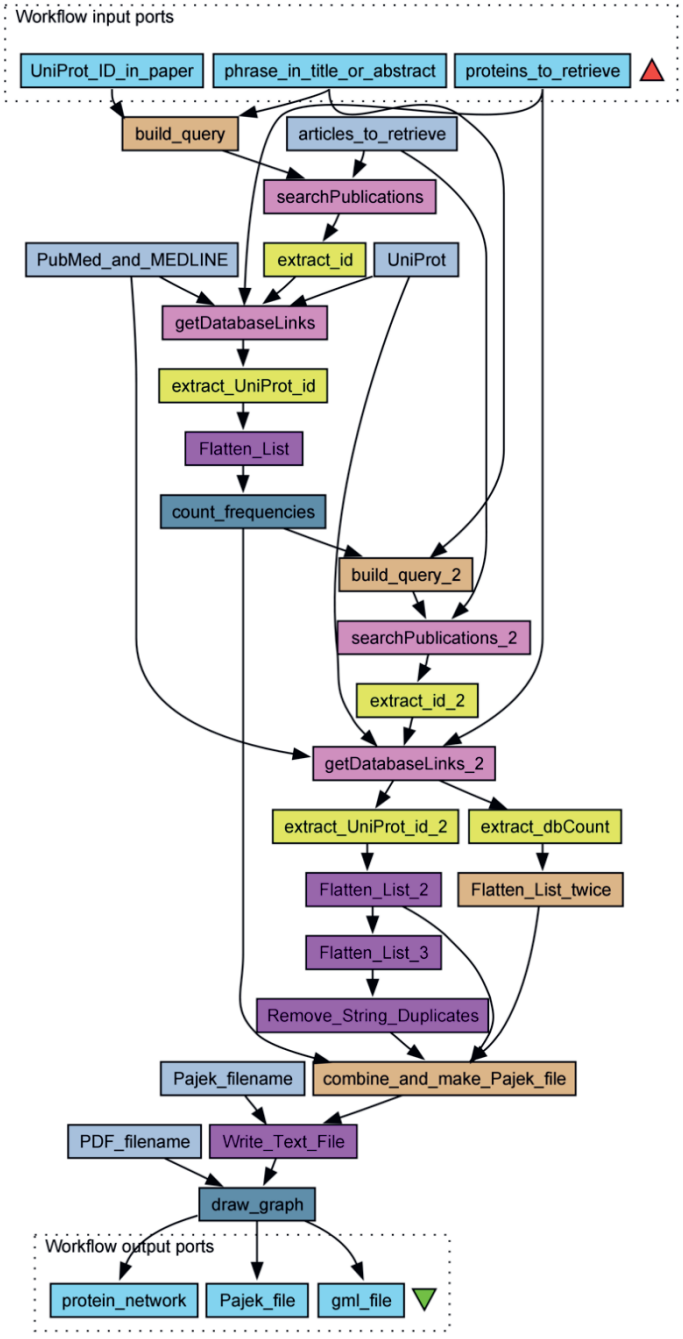


Figure 3.6. Workflow to generate a protein-protein network based on co-occurrence of UniProtKB accessions

(the “good cholesterol”) in plasma.

The relevance of a co-occurrence (of genes or proteins) decreases with the number of co-occurring genes or proteins in a particular publication, as long lists are the results of broad proteomics studies rather than specific experiments probing the interactions of a particular protein or dissecting a specific protein-protein complex. But to take all proteins into account, a simple trick to retrieve an arbitrary number of results from any Europe PMC Web Service in a Taverna workflow is to supply the Web Service call with a sufficiently long list of *offSet* values, counting from zero. This list can be created inside the *build_query* BeanShell to hide the details from the workflow view. For example, adding a simple piece of code defining a new output *offSets* as int [] offSets = new int [] { 0, 1, 2, 3, 4 }; to *build_query* and connecting the output port *offSets* to the *offSet* input port of the Web Service will retrieve at most 5 pages of *pageSize* results from a Europe PMC Web Service (5,000 results with the maximum *pageSize* of 1,000). This approach is generally fine for literature on genes and proteins, but the Europe PMC Web Services, or Web Services generally, are not intended for piecemeal retrieval of millions of records (or the entire Europe PMC). This can better be done using the FTP access. UniProt is just one of many molecular databases linked with Europe PMC. The API to access these database links is the same for all molecular databases, all using the *getDatabaseLinks* Web Service.

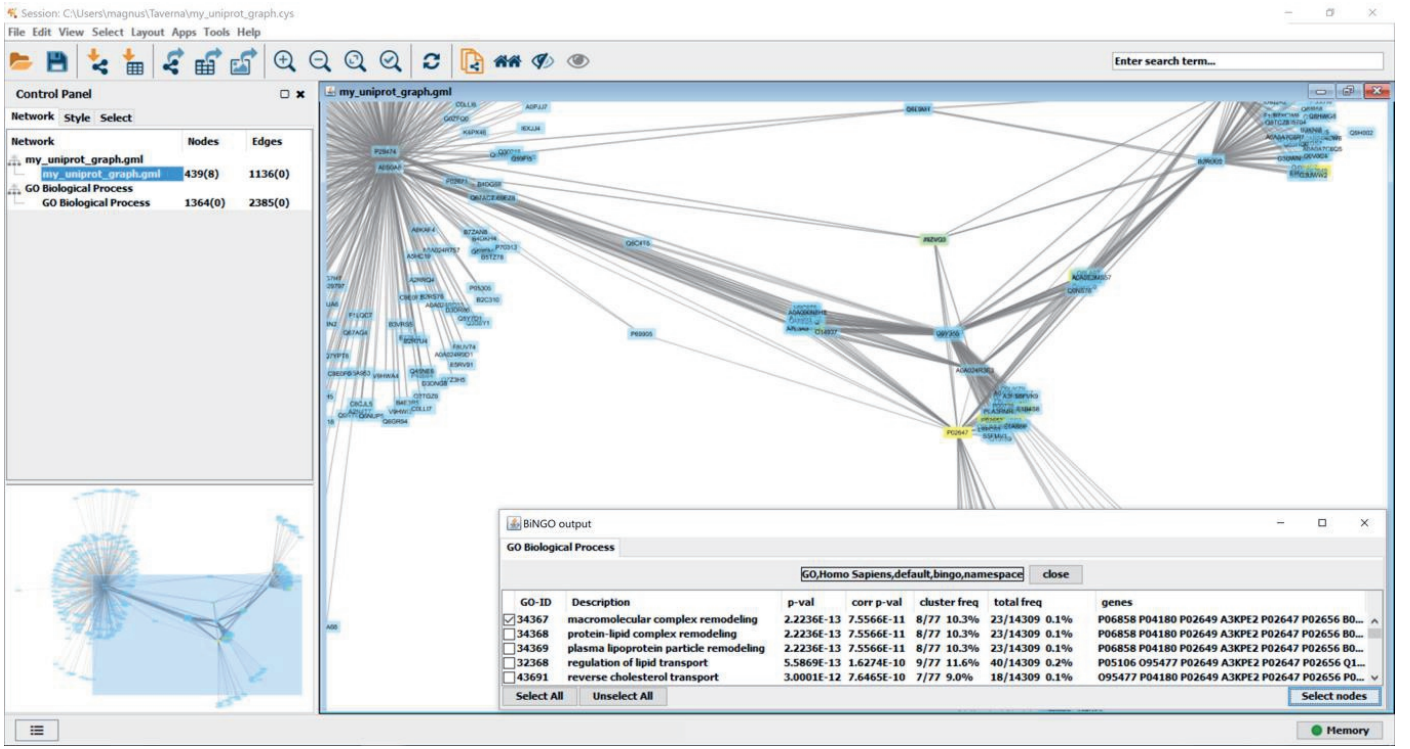


Figure 3.7. Workflow results visualized in Cytoscape with BiNGO Gene Ontology Annotations

Discussion

Most scientific workflow managers interact with the user through intuitive and visual interfaces. Like all software, Taverna also has some peculiarities. For example, to execute R scripts, a connection to RServe must first be established. The advantage is that this can be on a remote server just as easily as on the local machine, something that may be useful for computationally demanding tasks. Most workflow managers also support multiple scripting languages and types of workflow components. While this brings a lot of flexibility and power, it also makes it more difficult for those not familiar with all of these languages or services to understand the details of heterogeneous workflows. In this tutorial, we have used only Beanshell, Rshell, XPath and WSDL components. For simplicity, and because they were not needed, we did not include any local tools or shells, JSONPaths or REST services in these workflows. However, we have also uploaded a REST equivalent of the Figure 3.2 workflow to myExperiment (https://edu.nl/hw6wy).

Taverna is flexible, and can be used to organize the running of locally installed software, arrange a series of R scripts, shuffle data between external Web Services, or any combination thereof. Unlike KNIME, Taverna is free both as in 'speech' (open source) and as in 'beer' (gratis). Taverna's emphasis on Web Services makes it a perfect partner to bibliometric resources such as Europe PMC. The Taverna codebase is in Java, whereas Galaxy's is in Python. This is also reflected in the default scripting language in the workflow managers (Java in Taverna, Python in Galaxy). The programming paradigm is shared between all workflow managers however, and there have even been efforts to enact Taverna workflows through Galaxy[39].

Documentation is important, in particular for sharing or collaborative development of workflows. All elements (processors, data links, inputs and outputs) in Taverna workflows can be annotated individually. These annotations follow the components when imported from one workflow to another and are found under the "Details" tab in the Service panel in Taverna Workbench. Components and connections only have a generic "Description" field whereas inputs and outputs also have an "Example" field that can be used as a default value when executing the workflow. The workflow itself has "Author" and "Title" fields, in addition to a description. Workflows can be shared on myExperiment, as we have done. When uploading a Taverna workflow,

myExperiment attempts to extract workflow metadata such as title and description directly from these annotations. This works for Taverna, Galaxy and several other workflow managers. myExperiment also provide basic version control and allow users to comment on and discuss workflows. All this information can then be used to find workflows using a keyword search on the myExperiment website. Currently (May 2016), there are 3,752 workflows shared on myExperiment, so it is not practical to browse all workflows to find the one closest to what one needs (or the best starting point for one's workflow).

The examples in this paper do not use any nested structures which are otherwise common in large workflows. The legibility of complex workflows such as in Figure 3.4 may be improved by boxing the Web Service calls, hiding the details of the input/output splitters and XPaths and allowing the user to first grasp the overall logic of the workflow.

Conclusions

Bibliometric analyses often involve several steps that are carried out in different software tools. This requires much manual orchestration from one software tool to the other, which makes the process labor intensive and error prone. Scientific workflow managers, which are increasingly being used in other data intensive fields but have not yet seen widespread usage in bibliometrics, are useful tools to connect these different data retrieval and computational steps in an automated way. One such workflow manager is Taverna. In this study, we argue the direct support of Web Services, XML parsers and R in Taverna workflows make Taverna particularly useful for bibliometrics. With R comes direct access to a great number of powerful software packages such as igraph, wordcloud and rworldmap[40] for visualization, tm and openNLP for text mining and natural language processing. One limitation of using a scientific workflow manager such as Taverna, is that they are not meant for interactive exploration of large datasets. For this, it is more sensible to use domain-specific tools such as Pajek or VOSviewer for scientometrics, or Cytoscape for bioinformatics, as we demonstrated here.

In addition, software such as Taverna supplies repeatability and reusability to bibliometrics analyses. For example, all workflows discussed in this paper can be

found in the myExperiment group for Bibliometrics and Scientometrics (https://edu.nl/4r8x3) for anyone to open and run from within Taverna, using the exactly the same or any other input parameters to define the query. Other Taverna workflows in the Bibliometrics and Scientometrics group on myExperiment use rworldmap to map differences in the geographic distribution of author affiliations between two PubMed search results. Such workflows can for example look at geographical patterns of research on particular diseases, or geographical bias in different journals.

The bibliometric analyses illustrated in this study are exemplative of the kind of analyses we do in our research and here focused on the use of the Europe PMC Web Services. In a previous paper[10] we have used Taverna for other types of bibliometric analyses, such as geographic and temporal analyses of publication patterns, word usage and co-citation analysis. Here we have shown how to access Europe PMC through a Web Service API and how to perform bibliometric analyses using the Taverna scientific workflow manager, but, more importantly, how to combine the two.

Acknowledgments

The authors acknowledge Dr. Nees Jan van Eck at CWTS for advice on the VOSviewer and Professor Jonas Bergquist, Uppsala University for allowing us to use his publication record to demonstrate the workflows. The authors also wish to thank Michele Ide-Smith and Vid Vartak at the European Bioinformatics Institute for answering our questions regarding Europe PMC data.

References

1. de Bruin, J. S., Deelder, A. M. & Palmblad, M. Scientific Workflow Management in Proteomics. *Mol. Cell. Proteomics* **11**, M111.010595 (2012).
2. Goble, C. A. *et al.* myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* **38**, W677–W682 (2010).
3. Lu, S. & Zhang, J. Collaborative scientific workflows. in *2009 IEEE Int. Conf. Web Serv.* 527–534 (2009).
4. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
5. Berthold, M. R. *et al.* KNIME-the Konstanz information miner. in *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization* (eds. Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R.) (Springer Berlin, Heidelberg 2008).
6. Oinn, T. *et al.* Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054 (2004).
7. Persson, O. BibExcel. https://homepage.univie.ac.at/juan.gorraiz/bibexcel/ (2016).
8. Harzing, A. W. Publish or Perish. https://harzing.com/resources/publish-or-perish (2007).
9. Sci2 Team. Science of Science (Sci2) Tool. *Indiana University and SciTech Strategies* https://sci2.cns.iu.edu (2009).
10. Guler, A. T., Waaijer, C. J. F. & Palmblad, M. Scientific workflows for bibliometrics. *Scientometrics* **107**, 385–398 (2016).
11. Feinerer, I., Hornik, K. & Meyer, D. Text Mining Infrastructure in R. *J. Stat. Softw.* **25**, 1–54 (2008).
12. Hornik, K. openNLP: Apache OpenNLP Tools Interface. *R package version 0.2-4* https://cran.r-project.org/package=openNLP (2015).
13. 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).
14. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar : strengths and weaknesses. *FASEB J.* **22**, 338–342 (2008).
15. Wassink, I. *et al.* Using R in Taverna: RShell v1.2. *BMC Res. Notes* **2**, 138 (2009).

16. Gou, Y. *et al.* Europe PMC: A full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.* **43**, D1042–D1048 (2015).

17. Europe PMC. EBI Europe PMC SOAP Web Service 4.4 Reference Guide. http://europepmc.org/docs/EBI_Europe_PMC_Web_Service_Reference.pdf (2015).

18. Becker, R A, Chambers, J. M. *The New S Language: a Programming Environment for Data Analysis and Graphics*. (Wadsworth & Brooks/Cole, 1988).

19. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. *Technometrics.* **45,** 111 (2003).

20. Batagelj, V. & Mrvar, A. Pajek – program for large network analysis. *Connections* **21**, 47–57 (1998).

21. Eiglsperger, M., Brandes, U., Lerner, J. & Pich, C. Graph Markup Language (GraphML). in *Handbook of Graph Drawing and Visualization* 517–541 (Chapman & Hall/CRC, 2013).

22. Himsolt, M. *GML: A portable Graph File Format*. *Technical report* . University of Passau, 94030 Passau, Germany (1997).

23. Adai, A. T., Date, S. V., Wieland, S. & Marcotte, E. M. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol.* **340**, 179–190 (2004).

24. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).

25. Sugiyama, K., Tagawa, S. & Toda, M. Methods for Visual Understanding of Hierarchical System Structures. *IEEE Trans. Syst. Man Cybern.* **SMC-11**, 109–125 (1981).

26. van Eck, N. J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**, 523–538 (2010).

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

28. Wetterhall, M., Zuberovic, A., Hanrieder, J. & Bergquist, J. Assessment of the partitioning capacity of high abundant proteins in human cerebrospinal fluid using affinity and immunoaffinity subtraction spin columns. *J. Chromatogr. B* **878**, 1519–1530 (2010).

29. Dahlin, A. P. *et al.* Multiplexed quantification of proteins adsorbed to surface-modified and non-modified microdialysis membranes. *Anal. Bioanal. Chem.* **402**, 2057–2067 (2012).

30. Holst, B. S., Kushnir, M. M. & Bergquist, J. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) for analysis of endogenous steroids in the luteal phase and early pregnancy in dogs: A pilot study. *Vet. Clin. Pathol.* **44**, 552–558 (2015).

31. Thorslund, S., Lindberg, P., Andrén, P. E., Nikolajeff, F. & Bergquist, J. Electrokinetic-driven microfluidic system in poly(dimethylsiloxane) for mass spectrometry detection integrating sample injection, capillary electrophoresis, and electrospray emitter on-chip. *Electrophoresis* **26**, 4674–4683 (2005).

32. Eriksson, A. *et al.* Optimized protocol for on-target phosphopeptide enrichment prior to matrix-assisted laser desorption-ionization mass spectrometry using mesoporous titanium dioxide. *Anal. Chem.* **82**, 4577–4583 (2010).

33. Magrane, M. & UniProt Consortium. UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011**, 1–13 (2011).

34. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).

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

36. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325**, 834–840 (2009).

37. Shannon, P. *et al.* Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

38. Maere, S., Heymans, K. & Kuiper, M. Systems biology BiNGO : a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. **21**, 3448–3449 (2005).

39. Karasavvas, K. *et al.* Opening new gateways to workflows for life scientists. *Stud. Health Technol. Inform.* **175**, 131–141 (2012).

40. South, A. rworldmap : A New R package for Mapping Global Data. *R J.* **3**, 35–43 (2011)

CHAPTER 4

COMICS: Cartoon Visualization of Omics Data in Spatial Context Using Anatomical Ontologies

Dmitrii Travin[1,*], Iaroslav Popov[1,*], Arzu Tugce Guler[2], Dmitry Medvedev[1], Suzanne van der Plas-Duivesteijn[2], Monica Varela[3], Iris C. R. M. Kolder[3], Annemarie H. Meijer[3], Herman P. Spaink[3], Magnus Palmblad[2]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119234 Moscow, Russian Federation

[2] Center for Proteomics and Metabolomics, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands

[3] Institute of Biology, Leiden University, PO Box 9502, 2300 RA, Leiden, The Netherlands

[*] shared authors

This chapter has been published in:
Journal of Proteome Research (2018) 17, 739-744
doi: 10.1021/acs.jproteome.7b00615

Abstract

COMICS is an interactive and open-access Web platform for integration and visualization of molecular expression data in anatomograms of zebrafish, carp and mouse model systems. Anatomical ontologies are used to map omics data across experiments and between an experiment and a particular visualization in a data dependent manner. COMICS is built on top of several existing resources. Zebrafish and mouse anatomical ontologies with their controlled vocabulary (CV) and defined hierarchy are used with the ontoCAT R package to aggregate data for comparison and visualization. Libraries from the QGIS geographical information system are used with the R packages "maps" and "maptools" to visualize and interact with molecular expression data in anatomical drawings of model systems. COMICS allows users to upload their own data from omics experiments, using any gene or protein nomenclature they wish, as long as CV terms are used to define anatomical regions or developmental stages. Common nomenclatures such as the ZFIN gene names and UniProt accessions are provided additional support. COMICS can be used to generate publication-quality visualization of gene and protein expression across experiments. Unlike previous tools that have used anatomical ontologies to interpret imaging data in several animal models, including zebrafish, COMICS is designed to take spatially resolved data generated by dissection or fractionation and display this data in visually clear anatomical representations rather than large data tables. COMICS is optimized for ease-of-use, with a minimalistic web interface and automatic selection of the appropriate visual representation depending on the input data.

Introduction

Ontologies

For more than a decade, ontology-based data integration has been used to merge heterogeneous data in many domains, including bioinformatics[1]. In many disciplines, ontologies have to be actively maintained to keep up with the development or new discoveries in the field. This is particularly true for the more technical ontologies used to annotate datasets in genomics or proteomics, such as the PRIDE Controlled Vocabulary[2], and ontologies used to describe bioinformatics operations as well as data types, formats and identifiers, such as EDAM[3]. However, there are also examples of mature and essentially complete ontologies. These include the anatomical ontologies of well-studied organisms, the anatomies themselves being highly conserved over time (millions of years). Simpler controlled vocabularies (CVs) may be sufficient for some purposes, such as standardizing the way datasets in public repositories are annotated with metadata. However, when comparing or integrating heterogeneous (or heterogeneously annotated) data generated in different laboratories or using different experimental protocols, such CVs lack the necessary structure. A proteomics researcher may wish to find mass spectrometry datasets from an organism of interest generated using any "electrospray ionization" (CV term ID "MS:1000073") technique to build a spectral library of comparable data. But if some such datasets are annotated as having been acquired with "microelectrospray" (MS:1000397) and others as being derived from a "nanoelectrospray" (MS:1000398) experiment, how does the software know these all qualify as "electrospray ionization" mass spectrometry datasets? This information is provided by the relationships between the terms as defined in an ontology. In this case, both the specific "microelectrospray" and "nanoelectrospray" have a direct "is a" relationship with the more general or parent "electrospray ionization". One can therefore reason that they are all "electrospray ionization" datasets, and hence compatible for this researcher's defined purpose.

Common methods for generating deep proteomics datasets often involve separation or fractionation. These can be applied on the sample level, for example, by dissection[4], cell sorting[5] or organelle fractionation[6], each defining a spatial context of subsequently generated data. Fractionation on the protein level is also commonplace,

and provide a protein-level context for peptide-level data. When comparing two such large datasets in any -omics field, we cannot assume the two datasets have been acquired in exactly the same way. Depending on the laboratory, equipment, experimental protocol, skills of the experimentalists involved, or allocated effort, the dissection or fractionation may have been done differently, altering the spatial definition of the fractions of the dataset. To integrate such datasets for the purpose of comparison of spatial expression patterns, the datasets must be annotated using something like an anatomical or cellular ontology, with defined relationships between anatomical entities. Many such ontologies already exist, including the model-system specific *C. elegans* gross anatomy (WBBT)[7], the Drosophila gross anatomy (FBbt and FBdv), also referred to as the Drosophila anatomy ontology (DAO)[8], the Mouse Adult Gross Anatomy (MA)[9], Xenopus anatomy and development (XAO)[10] and Zebrafish anatomy and development (ZFA and ZFS)[11]. There are also the more general Anatomical Entity Ontology (AEO)[12], Biological Spatial Ontology (BSPO)[13] and the general vertebrate "Uber-anatomy" ontology (UBERON)[14] currently (20170415) containing 15,036 anatomical terms. The zebrafish ZFA and ZFS ontologies contain 3175 anatomical terms (20170627 release) and the mouse MA 3257 terms (20170207 version). For comparison, the two major ontologies covering human anatomy, the Foundational Model of Anatomy (FMA)[15] and SNOMED-CT[16], contain 75,019 and 30,933 anatomical concepts respectively[17].

Anatomical visualization

In their classic 1987 paper "Why a Diagram is (Sometimes) Worth Ten Thousand Words"[18], Larkin and Simon demonstrated how well-made figures or diagrams use location to group information, reduce the need for symbolic labels and enable a large number of conceptual inferences to be made, something the human brain is extremely good at. Larkin and Simon argued that the main advantages of diagrams are *computational* - diagrams are better representations not because they contain *more* information, but because the *indexing* of this information support extremely efficient computational processes, including those carried out in the human brain upon trying to grasp the contents of a research paper. Anatomical schemata or anatomograms are now used to interact with on-line databases, such as Reactome[19], the Human Protein Atlas[20], ProteomicsDB[21] and the EMBL-EBI Expression Atlas[22].

This paper describes a new stand-alone freeware, COMICS, with an interactive web-based interface designed to fit into a niche between existing tools for combined integration and visualization of molecular expression data in some vertebrate model organisms (zebrafish, carp and mouse). The software uses the existing anatomical ontologies to map arbitrary omics data across experiments and between one experiment and a particular visualization in a data-dependent manner. The method and software can be extended to other model systems, provided the relevant ontology and visual representation (picture). COMICS is designed for simplicity-of-use, and can generate custom, publication-quality, vector graphics mapping molecular expression (such as from transcriptomics, proteomics or metabolomics) data to anatomical diagrams. In addition to molecular expression levels, the locations in the diagram immediately convey information on similarity or dissimilarity between adjacent structures or parts of an organ, such as the eye or the brain, tissue specificity (one part against the whole) and differences in expression levels between genes/proteins or between animals.

Methods

COMICS takes as input a table of numerical data (e.g., gene or protein expression values) with each row corresponding to one CV term from an anatomical ontology, such as the ZFA[11] or MA[9], and each column to one particular gene or protein, with the CV terms as row names and gene or protein identifiers as column names. If the molecular identifiers and anatomical CV terms are swapped, then COMICS will automatically detect this and transpose the matrix. COMICS requires CV terms instead of common names of anatomical features to be able to match them correctly with parts of the picture. For carp, we also apply ZFA ontology CV-terms as there is no specific ontology for this species. Both species belong to a single Cyprinidae family and are quite close in terms of tissues and organs present[23].

First, the CV terms in the data uploaded by the user are matched to CV terms with a corresponding polygon defined in the shapefile for the selected species. This is performed using the R package ontoCAT[24], which enables extracting term parents and children (generalization/specialization) as well as terms with a part of/has part (whole/part) relationship with the given term from the anatomical ontology. This is a

key step that allows any correctly annotated data to be mapped by COMICS to the anatomical representations in the shapefiles. An example of the ontology-based pre-processing and aggregation of molecular expression data is shown in Figure 4.1. For computational efficiency, a lookup table for the mapping between the ontology and visualization shapefile is computed for each ontology. This lookup table is rebuilt once for each new version of the ontology or shapefile.

For the anatomical drawings, we used the mature QGIS open-source geographic information system[25] to create shapefiles. These shapefiles were constructed from simple polygons corresponding to anatomical structures such as organs or parts of organs in zebrafish and carp. These shapefiles can easily be extended to include other model systems or developmental stages for which anatomical ontologies are available. Inspiration for the anatomical illustrations was drawn from previously published work[26,27,28].

To visualize the numerical data obtained from the user on the anatomical shapefiles we used the existing *maps* and *maptools* R packages commonly used for working with maps and *gridSVG* to produce vector graphics in the SVG format. The Adobe PDF is supported by the pre-installed *grDevices* package. The range of numerical data is translated to a palette of colors forming a one-, two- or three-color gradient. COMICS has several options that enable the user to choose from several predefined color schemes or make a new one and choose the number of bins for the gradient and scaling (linear or logarithmic). In addition, the user can keep the gradient fixed across diagrams or scale it automatically for each visualization. The former option is used for comparing (absolute) expression across many diagrams. The latter automatically adapts to the minimum and maximum values in the data for each gene or protein and is optimal for looking at tissue specificity or relative expression of two genes or proteins. The expression of two entities can also be computed and compared directly in COMICS.

The cartoons can be saved individually or as a collection, as vector graphics in the PDF or SVG formats.

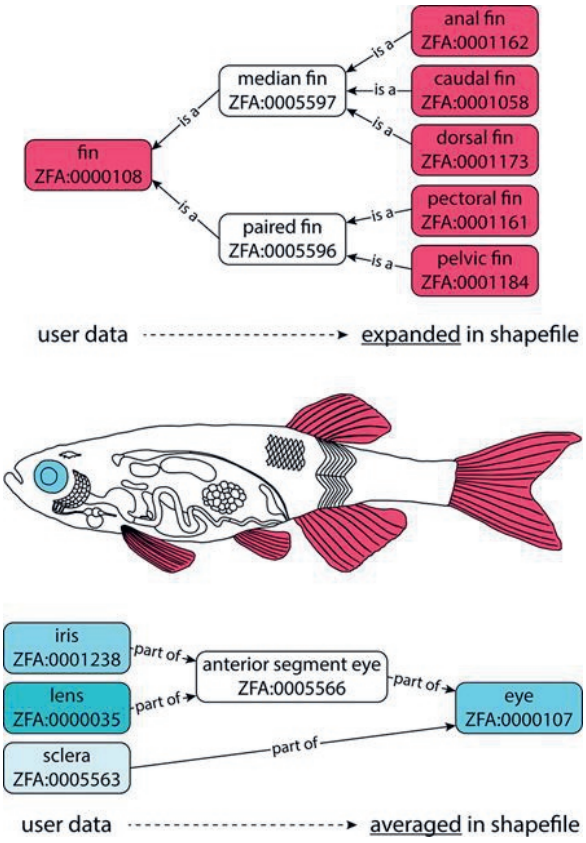


Figure 4.1. ZFA anatomical ontology is used to map scalar expression data to defined anatomical regions. This also provides a means to directly and visually compare data from different experiments and heterogeneous datasets. In this example, the user has provided data for "fin", which is then propagated to the five distinct fins visualized in the tool, through the parent-child (is a) relationships defined in ZFA. Because the fins are not distinguished in the user's dataset, the expression value provided by the user is mapped to all five visible fins. If the user provides information on a more detailed level than is visualized by COMICS, then the mean expression of all children or parts are mapped to the anatomical structure defined in the shapefile. Here, separate expression data for the iris, sclera and lens (all part of the eye) are averaged to the eye. The averaging is done once, for all parts, independent of intermediate levels in the ontological hierarchy (such as the anterior segment eye). The default shapefile corresponds to the organs and tissues that are easy to dissect for an omics experiment, although the shapefile can easily be modified to incorporate other experimental designs.

Technically, COMICS is a web application build around R scripts. For standalone usage it is containerized using Docker. The container includes all software, including source code, packages and scripts, making it very easy to install and run COMICS locally, independently of other installed software. The standalone mode enables the user to work with the application locally, without uploading datasets to any server. Links to the Docker container and locations where COMICS can be run remotely will be maintained on https://edu.nl/drrew.

To test COMICS, we used previously published data from the public domain. Wildtype gene expression data for zebrafish was taken from ZFIN, already annotated using the ZFA[29]. Protein expression data in adult zebrafish were taken from the zebrafish spectral library[4]. Expression data from carp were taken from a recent paper on the full-body transcriptome and proteome resource for this species[30]. Mouse gene expression data was downloaded from the Mouse Atlas of Gene Expression[31], and mouse protein data was generated in-house using the same method as for the zebrafish spectral library.

Results

The main product of this work is a software tool with a simple web interface as shown in Figure 4.2. The screenshot visualizes the gene expression of the carp ortholog of zebrafish cytokeratin-8, using the ZFA ontology mapped onto the anatomy of a carp, closely resembling that of zebrafish. The interface is divided into panels containing basic information about the underlying data, image controls, the image itself and links to cross-referenced databases (here UniProt, ZFIN and NCBI). The image is interactive: as the user hovers the mouse pointer over an anatomical region, the tooltip displays the name, ontology identifier and expression level (here for the dorsal fin). Clicking on the anatomical structure will lead to the web page for this part in the online version of the corresponding ontology. The image shapefiles annotated with the ZFA and MA anatomical ontologies are available as individual files for developers who would like to integrate them in their own software.

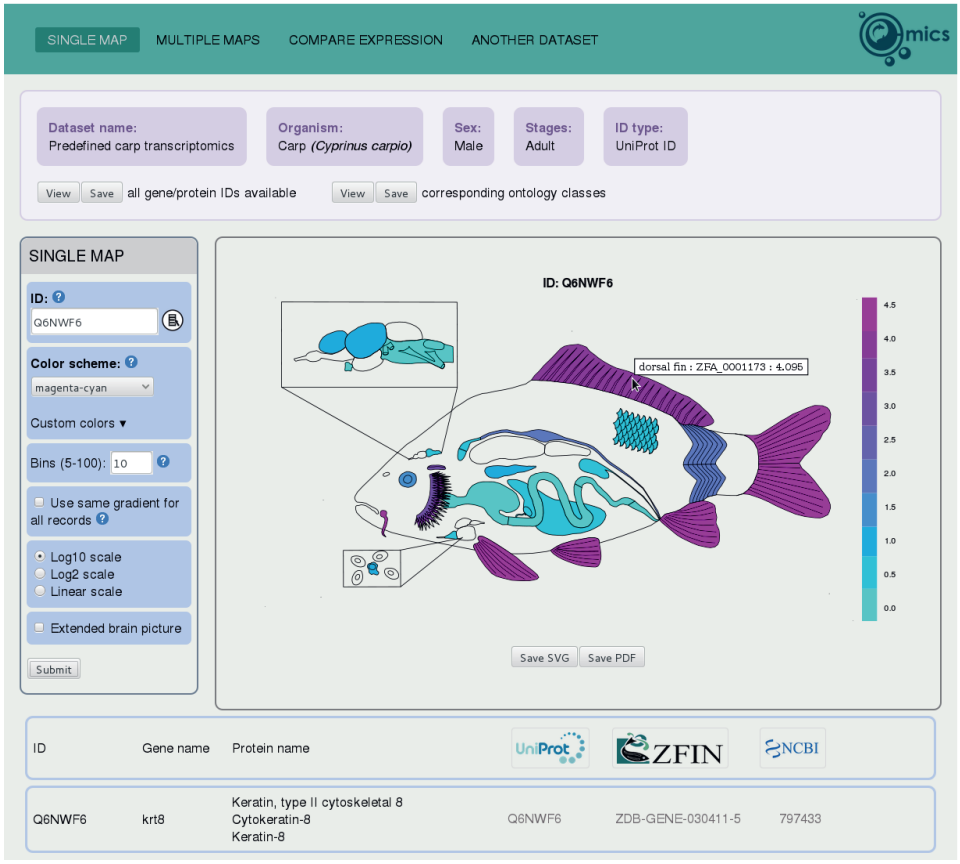


Figure 4.2. Screenshot of the COMICS interface, presenting the information about the selected dataset (top), a control panel with options and parameters for visualization (left), the generated output image (center, right) and a table containing the selected gene/protein description with links to the corresponding databases (bottom). Gene expression data[30] for the carp cytokeratin-8 (Q6NWF6) ortholog is here used as an example.

The COMICS tool is generic because it aggregates and displays any numerical data provided with anatomical ontology annotations linked to a shapefile. The tool can therefore be used to compare the expression of a few genes or proteins in one experiment and model system, look at the ratio of transcripts and the corresponding proteins, or compare the expression of orthologs across model systems. Figure 4.3 shows the expression of sarcosine dehydrogenase in zebrafish (*sardh* gene) and mouse (the sarcosine dehydrogenase protein), respectively, revealing the expression pattern for this pair of orthologs is conserved across the vertebrate subphylum (the

last common ancestor of the mouse and the two cyprinids lived over 400 million years ago[32]). As a final verification of the parsing of the anatomical ontology we looked at the expression of four genes with well-known spatial specificity in ZFIN (Figure 4.4). The four panels visualize gene expression, quantified as the number of experiments in which the transcript has been observed in wildtype fish and recorded by ZFIN, of four genes: rhodopsin (*rho*, ZDB-GENE-990415-271) in the eye (a), fatty acid binding protein 1a (*fabp1a*, ZDB-GENE-020318-3) in the liver (b), proopiomelanocortin a (*pomca*, ZDB-GENE-030513-2) in the brain, specifically the hypothalamus (c) and vitellogenin 2 (*vtg2*, ZDB-GENE-001201-2) in the liver and ovaries (d).

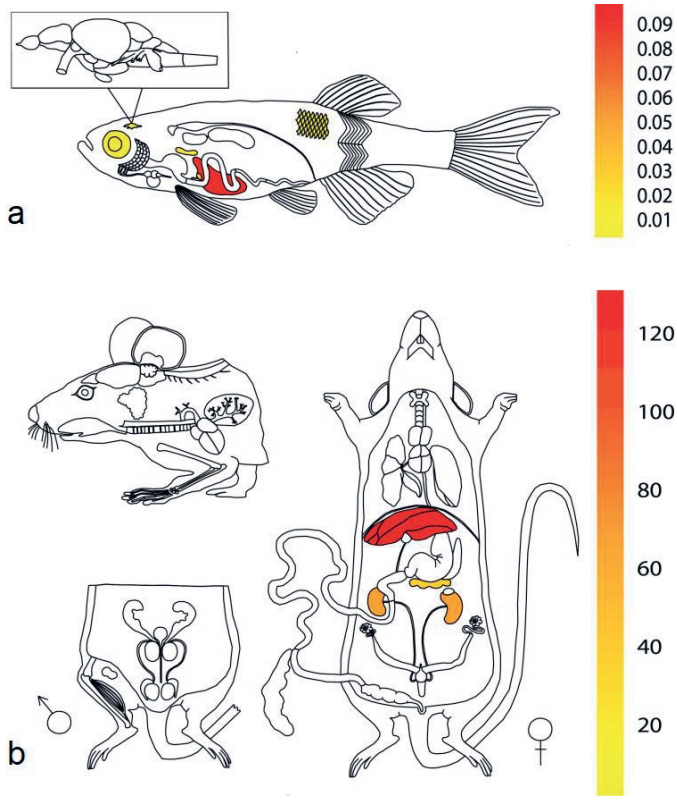


Figure 4.3. Publication-quality figures, showing the expression of Sarcosine dehydrogenase orthologs in zebrafish (*sardh* gene) **(a)**, mouse (Sarcosine dehydrogenase *protein*, UniProt accession number Q99LB7) **(b)**. The numbers on the color scales represent the fraction of experiments in ZFIN in which gene expression is observed (a) and absolute spectral counts (b)

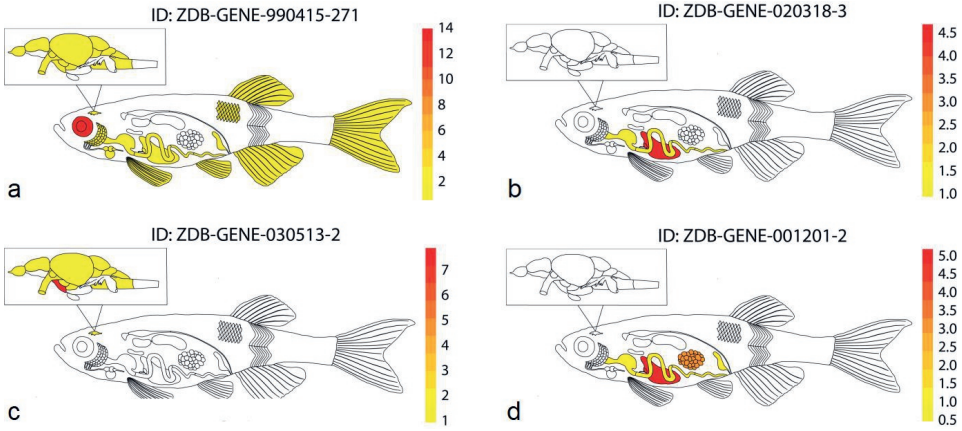


Figure 4.4. Organ-specific expression of four genes in zebrafish: rhodopsin **(a)**, fatty acid binding protein 1a **(b)**, proopiomelanocortin a **(c)**, vitellogenin 2 **(d)**, according to the number of registered detection of expression among all wildtype datasets in the ZFIN gene expression database. The color scale represent the number of experiments in ZFIN in which gene expression was observed in a particular organ or tissue.

If COMICS detects the presence of only male or female organ data, then the anatomical map will represent a single sex. If neither or both male and female organ annotations are included in the dataset, then a generic anatomical representation will be used. For mouse, a model with common superior and split inferior regions is also available.

Discussion

To summarize, COMICS is a simple, easy-to-use tool for generating visually clear, publication-quality vector graphics from arbitrary omics data using the mouse and zebrafish anatomical ontologies. COMICS should not be compared with resources pre-dating the development of these anatomical ontologies, such as the now off-line GEMS database[33], which was aimed at annotation of real images. COMICS can be used to compare the expression of a pair of genes or proteins, such as two isoforms, or the expression of a gene measured on the transcript and protein levels. In this way, one can visually inspect and quickly assess results from an ontology-based aggregation of two or more heterogeneous, spatially resolved, omics datasets. COMICS is not a tool to provide detailed and beautiful anatomical illustrations of an organism in the tradition of Vesalius[34]. Rather, we have deliberately compromised anatomical precision for

diagrammatic simplicity, ensuring the cartoons are clear also when viewed at a small scale, allowing quick side-by-side comparison of datasets. Future extensions of COMICS will include shapefiles of different embryonic and larval stages using the ZFS ontology as well as additional model systems.

Conclusions

We have here presented a simple software, COMICS, for mapping any numerical gene, protein or metabolomics data as choropleths in anatomical cartoons referred to as anatomograms. Unlike existing tools, COMICS makes full use of anatomical ontologies to integrate spatially or anatomically resolved data in several animal models, including zebrafish and mouse. COMICS is built on existing libraries and has a minimalistic web interface for selecting the appropriate visual representation and exporting publication-quality graphics. Additional model systems (as well as human anatomy or other developmental stages) are easy to add to the COMICS platform, provided an anatomical ontology in the OBO format and an organism-specific shapefile with mappings to the CV terms in the ontology are available. COMICS can be downloaded as a Docker image from https://edu.nl/drrew.

Acknowledgments

We gratefully acknowledge financial support from The Netherlands Organisation for Scientific Research (NWO) Vidi grant 917.11.398 (M.P.) and Olga Gancharova for valuable advice on the mouse anatomogram.

References

1. Wache, H. *et al.* Ontology-based Integration of Information - A Survey of Existing Approaches. in *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, August 4-5, 2001, Seattle, USA* 108–117 (2001).
2. Jones, P. *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34**, D659–D663 (2006).
3. Ison, J. *et al.* EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**, 1325–1332 (2013).
4. van der Plas-Duivesteijn, S. J. *et al.* Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J. Proteome Res.* **13**, 1537–1544 (2014).
5. Bernas, T., Grégori, G., Asem, E. K. & Robinson, J. P. Integrating cytomics and proteomics. *Mol. Cell. Proteomics* **5**, 2–13 (2006).
6. Lee, Y. H., Tan, H. T. & Chung, M. C. M. Subcellular fractionation methods and strategies for proteomics. *Proteomics* **10**, 3935–3956 (2010).
7. Lee, R. Y. N. & Sternberg, P. W. Building a cell and anatomy ontology of Caenorhabditis elegans. *Comp. Funct. Genomics* **4**, 121–126 (2003).
8. Costa, M., Reeve, S., Grumbling, G. & Osumi-Sutherland, D. The Drosophila anatomy ontology. *J. Biomed. Semantics* **4**, 1–11 (2013).
9. Hayamizu, T. F., Baldock, R. A. & Ringwald, M. Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. *Mamm. Genome* **26**, 422–430 (2015).
10. Segerdell, E., Bowes, J. B., Pollet, N. & Vize, P. D. An ontology for Xenopus anatomy and development. *BMC Dev. Biol.* **8**, 92 (2008).
11. van Slyke, C. E., Bradford, Y. M., Westerfield, M. & Haendel, M. A. The zebrafish anatomy and stage ontologies: Representing the anatomy and development of Danio rerio. *J. Biomed. Semantics* **5**, 12 (2014).
12. Bard, J. B. L. The AEO, an ontology of anatomical entities for classifying animal tissues and organs. *Front. Genet.* **3**, 18 (2012).
13. Dahdul, W. M. *et al.* Nose to tail, roots to shoots: Spatial descriptors for phenotypic diversity in the Biological Spatial Ontology. *J. Biomed. Semantics* **5**, 34 (2014).
14. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
15. Zhang, S. & Bodenreider, O. Aligning representations of anatomy using lexical and structural methods. *AMIA Annu. Symp. Proc.* **2003**, 753–757 (2003).
16. Côté, R. A. & Robboy, S. Progress in Medical Information Management. Systemized Nomenclature of Medicine (SNOMED). *J. Am. Med. Assoc.* **243**, 756–762 (1980).

17. Bodenreider, O. & Zhang, S. Comparing the representation of anatomy in the FMA and SNOMED CT. in *AMIA Annual Symposium Proceedings, November 11-15, 2006, Washington, DC, USA* 46–50 (2006).

18. Larkin, J. H. & Simon, H. A. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cogn. Sci.* **11**, 65–99 (1987).

19. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).

20. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

21. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

22. Petryszak, R. *et al.* Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* **44**, D746–D752 (2016).

23. Henkel, C. V. *et al.* Comparison of the exomes of common carp (Cyprinus carpio) and zebrafish (Danio rerio). *Zebrafish* **9**, 59–67 (2012).

24. Kurbatova, N., Adamusiak, T., Kurnosov, P., Swertz, M. A. & Kapushesky, M. ontoCAT: An R package for ontology traversal and search. *Bioinformatics* **27**, 2468–2470 (2011).

25. QGIS Association. QGIS Geographic Information System. https://www.qgis.org.

26. Davidson, A. J. & Zon, L. I. The 'definitive' (and 'primitive') guide to zebrafish hematopoiesis. *Oncogene* **23**, 7233–7246 (2004).

27. Wulliman, M. F., Rupp, B. & Reichert, H. *Neuroanatomy of the zebrafish brain: a topological atlas*. (Birkhäuser Verlag Basel, 1996).

28. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).

29. The Zebrafish Information Network. Expression Data for Wildtype Fish. https://zfin.org/downloads/wildtype-expression_fish.txt.

30. Kolder, I. C. *et al.* A full-body transcriptome and proteome resource for the European common carp. *BMC Genomics* **17**, 701 (2016).

31. Siddiqui, A. S. *et al.* A mouse atlas of gene expression: Large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18485–18490 (2005).

32. Broughton, R. E., Betancur-R., R., Li, C., Arratia, G. & Ortí, G. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Curr. Tree Life* **5** (2013).

33. Belmamoune, M. & Verbeek, F. J. Data integration for spatio-temporal patterns of gene expression of zebrafish development: the GEMS database. *J. Integr. Bioinform.* **5** (2008).

34. Vesalius, A. *De Humani Corporis Fabrica Libri Septem.* (Padua School of Medicine, Padua, Italy, 1543).

CHAPTER 5

Metadata–driven Calibration of Mass Spectrometry Data

Arzu Tugce Guler[1], Magnus Palmblad[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands

Manuscript in preparation

Abstract

Accurate determination of ion masses by the mass spectrometer increases the confidence of identifications and eventually leads to better identification and quantification. Although mass measurement accuracy and resolving power of mass spectrometers improved significantly throughout the years, there is a certain degree of systematic and random error in every data, depending on the instrument type. It is possible and beneficial to reduce mass measurement error after mass spectrometry analysis using computational methods. Here, we present a modular, command-line tool that performs automatic internal MS1 recalibration on mzXML files. msRecal selects the suitable calibration function based on the instrument type acquired from metadata and uses the calculated exact ion masses of high confidence identifications as calibrants.

Introduction

Advances in liquid chromatography – mass spectrometry have made the high throughput analysis of proteomics data more efficient and reliable. In bottom-up analyses, samples are very complex, and many peptides can elute at the same time. Thus, achieving high mass accuracy in MS1 measurements is important since precursor mass acts as an initial filter to identify peptides[1,2]. Due to instrumental factors, there is always a degree of deviation from the exact mass in measurements, affecting the accuracy and resulting in bias. Random errors are also present in measurements, affecting the precision. Taking repeated measurements of the same sample is not a practical solution to overcome these errors, as it is usually not feasible when working with biological samples, and yet, the systematic error remains an issue to tackle in any case[3]. Calibrating measured masses with a calibration function that uses calculated exact masses, i.e., theoretical masses, as calibrants is an efficient way to reduce systematic and random error[4,5]. Typically, calibration functions take the physics of the mass analyzer into account. There are several functions available in the literature for common mass analyzer types. Choosing the correct calibration function with suitable calibrants is essential for a good calibration[6]. Getting the instrument type from the metadata and choosing the correct calibration function and parameters according to this information is useful for automating mass calibration.

Open mass spectrometry data formats such as mzXML[7] and mzML[8] usually contain metadata containing details about the instrument type. Human Proteome Organization (HUPO) Proteomics Standards Initiative's controlled vocabulary for mass spectrometry (PSI-MS CV) defines mass spectrometry-related entities in a hierarchical manner, including mass analyzer type[9]. The PSI-MS CV directly supports open formats such as mzML, mzIdentML[10], and mzTab[11]; however, their standardized annotation is not enforced in the mzXML format[12,13]. Nevertheless, it is still possible to parse relevant information from the human-readable metadata present in mzXML files.

In principle, calibrants could be chosen among the peptides already identified with high confidence in the same analysis; however, it is also possible to use the identifications from a different MS run after additional steps if the analyzed samples are very similar or the same. Palmblad et al. showed that exact masses of peptides

identified by MS/MS in an ion trap instrument could be used to calibrate MS1 spectra from an FTICR instrument to reduce the overall mass measurement error after aligning the retention times[14]. Here, we focus on data from hybrid instruments, where the data is recalibrated by using peptides identified in the same MS run.

Methods

msRecal takes mzXML and pepXML[15] files as inputs and uses peptide identifications from the pepXML file to recalibrate the MS1 spectra and MS2 precursor masses in the mzXML file. The program outputs a recalibrated and reindexed mzXML file, ready to be used in different analysis pipelines. In principle, mzXML and pepXML files could be from different MS runs on similar samples. Retention times of different MS runs should be aligned before running msRecal. If identifications from the same MS run are used as calibrants, there is no need for this additional step.

msRecal is a command-line tool programmed in C. Dedicated libraries are used to read/write mzXML and pepXML files, and the GNU Scientific Library[16] is used to fit the calibration function. msRecal does not change the nature of the data; the input and output are of the same data type, mzXML. msRecal could be seamlessly incorporated into bottom-up MS analysis workflows that work with mzXML and pepXML, like the ones used by Bruin et al.[17] and Hussaarts et al.[18] An example workflow structure incorporating the msRecal module is shown in Figure 5.1. After recalibration with msRecal, the recalibrated mzXML file can be searched again with the same parameters for possible new identifications. However, since the MS2 precursor masses are updated with more accurate masses, it is also possible to do this search within a narrower error window than the initial search.

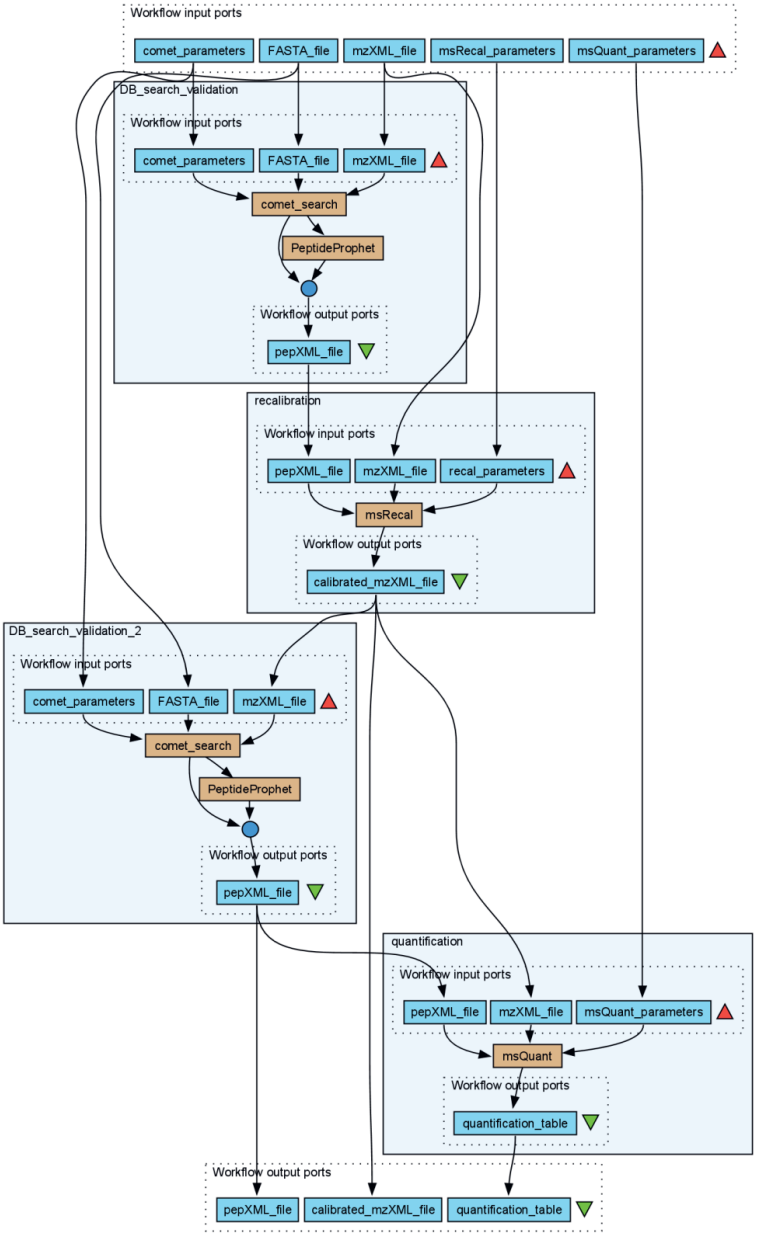


Figure 5.1. The mzXML file is recalibrated using the exact masses of the peptides from the pepXML file, obtained from the database search of the same mzXML file. The calibrated mzXML file is searched again for possible new identifications. The pepXML file with the results of the new search and the recalibrated mzXML can be used in other analysis modules downstream, e.g., in a quantification module.

The user can override the default values for parameters such as the minimum number of calibrants, maximum mass measurement error allowed for calibrants, internal mass measurement error target after calibration, threshold for background intensity, score type and threshold scores, retention time window for matching calibrants. It is highly recommended that certain parameters like 'threshold for background intensity', 'maximum mass measurement error allowed' are chosen by the user. The optimal values for these parameters vary from one data to another and may affect the calibration efficiency.

Application of the correct calibration function is the most critical step in the program. Currently, msRecal makes use of three calibration functions that are specific to instrument types[19,20,21,22].

Orbitrap $$\frac{m}{z} = \frac{A}{f^2} \quad (1)$$

FTICR $$\frac{m}{z} = \frac{A}{f+B} \quad (2)$$

TOF $$\frac{m}{z} = \frac{t-B}{A} \quad (3)$$

where *A*, *B*, and *C* are the calibration coefficients; *f* is the frequency; *t* is the time.

The calibration function is chosen according to the 'mass analyzer type' or 'instrument type' parameters. Normally, these parameters are parsed from the metadata in mzXML unless the user overrides them. The PSI-MS CV defines the three mass analyzer types that the program recognizes. (Figure 5.2) It is possible that the 'mass analyzer type' is missing, or sometimes even incorrect, in the mzXML metadata. However, in most cases, 'instrument type' is given correctly. If the 'mass analyzer type' is missing or deemed incorrect by the program, then the 'instrument type' is used to set the correct value for the former. The PSI-MS CV does not define a direct relationship between the children of 'instrument type' and 'mass analyzer type' entities, so we assume a hypothetical relationship to match them, as shown in Figure 5.2.

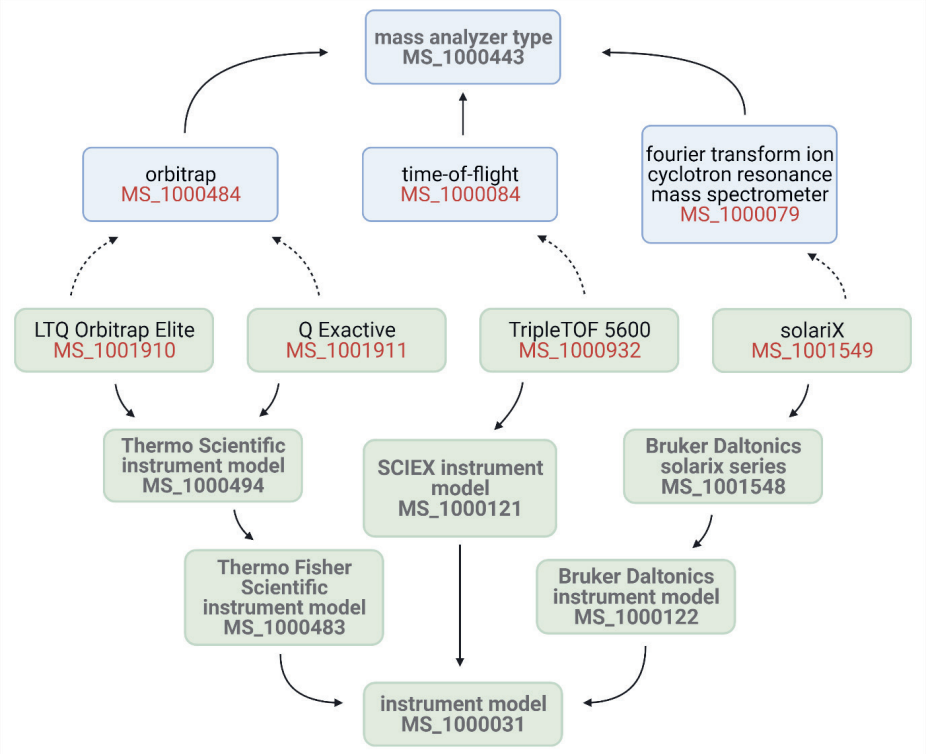


Figure 5.2. The PSI-MS controlled vocabulary groups the mass analyzer types and the instrument models separately. A direct match between the calibration function and the mass analyzer type is the most straightforward approach; however, it is also possible to make an indirect inference (shown with dashed lines) of the mass analyzer type if only the instrument model is provided. For instance, if the mass analyzer type is not given in the metadata, but the instrument model is stated as "LTQ Orbitrap Elite", then we use the function for Orbitrap.

msRecal uses the exact masses of peptides identified with high confidence to recalibrate the mass spectrometry data, thus first builds a peptide set from the pepXML file by selecting the peptides that fit the criteria, i.e., thresholds scores. In this version of msRecal, only unmodified peptides without isotope errors are used as calibrants, and the mass-to-charge ratios are calculated up to z = +4 charge state. The user could set the upper and lower score thresholds for selecting the high confidence peptides; by default, peptides with an expect score < 0.01 are selected. In addition to peptides, polydimethylcyclosiloxanes ($CH_3[Si(CH_3)_2O]_nSi(CH_3)_3$) are also added to the list of potential calibrants as they may be present when nanoelectrospray ionization is

used[23,24]. Within the matching scan/retention time window, by default [-30s,+90s], the maximum number of eligible calibrants are selected for each MS1. Only the peaks above the background intensity threshold are used, and the potential calibrants within the specified maximum mass measurement error window are matched to each peak. The suitable calibration function is used to calibrate each MS1 spectrum individually. This is done by taking the partial derivatives of the calibration function with respect to each calibration coefficient and then using the least-squares fit.
For instance, for the Orbitrap calibration function given in Eq. (1), the partial derivate with respect to its single coefficient is,

$$\frac{\partial(m/z)}{\partial A} = \frac{1}{f^2} \tag{4}$$

Next, a dummy unit for *f* is derived from the original calibration function, Eq. (1),

$$f = \frac{1}{\sqrt{m/z}} \tag{5}$$

The least-square minimization is applied first using all the measured calibrant *m/z* for an individual MS1 scan and their calculated *m/z* to find the optimal value for coefficient *A.* The calibration step is iterated several times while removing the calibrants that do not fit the function better than a given internal target, by default 2 ppm, as long as a specified minimum number of calibrants, by default 3, remain. Finally, the function in Eq. (1) is applied on the measured peak masses, using the calculated optimal value for coefficient *A,* and *f* in dummy units. Thus, the final equation used for calculating the calibrated *m/z* for an Orbitrap will be,

$$\left(\frac{m}{z}\right)' = A * \left(\frac{m}{z}\right) \tag{6}$$

where (m/z)' is the calibrated mass-to-charge ratio, *A* is the calculated calibration coefficient, and (m/z) is the measured mass-to-charge ratio.

It should be noted that the calibration coefficients of individual MS1 scans are used to calibrate MS1 peaks and the precursor masses of the corresponding MS2 scans. There is an option to exclude the uncalibrated scans in the output; otherwise, the original

masses of the calibrated scans are replaced with the calibrated masses in the mzXML file while the uncalibrated scans are left as is. The file is also reindexed so that the outputted mzXML is ready to be used.

The msRecal is demonstrated on different instruments to show the calibration performance. We used publicly available data from PRIDE with accession numbers PXD000563[25] for Orbitrap, PXD000071[26] for TOF, PXD004678[27] for FTICR. The datasets come from hybrid instruments, so we used the database search results of the same data to select the calibrants. The database searches and peptide validations were performed using Comet[28] version 2021.01 rev. 0 and PeptideProphet[29], respectively, in Trans-Proteomic Pipeline[30] v6.0.0. The Homo sapiens reference proteome downloaded from Uniprot[31] on October 2021, containing 78139 entries were used in the database search. It is, of course, possible to use other database search tools and pipelines that outputs the identifications in pepXML format. We used the default expect score < 0.01 in Orbitrap data and the PeptideProphet probability matching FDR < 0.01 in TOF and FTICR data as a threshold for high confidence peptides. Mass measurement error and background thresholds are chosen based on individual data. After the calibration, the outputted mzXML is searched again with Comet using the same parameters to check the improvement in mass measurement accuracy.

Results

The mass measurement error distributions of high confidence monoisotopic peptides before and after a single calibration are shown in Figure 5.3. The same thresholds used for selecting the calibrants were applied to select the high confidence peptides.

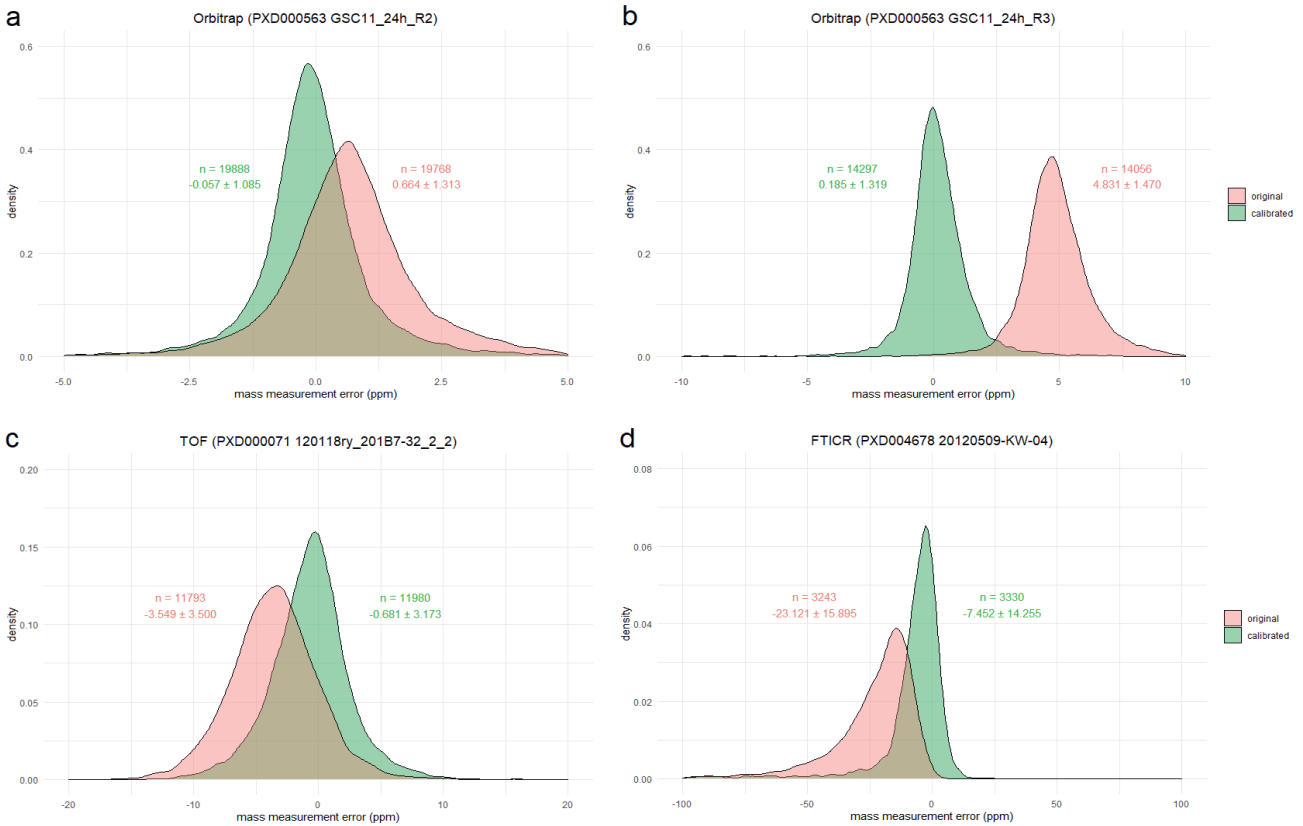


Figure 5.3. Mass measurement error distributions of original (pink) and recalibrated (green) Orbitrap (**a**, **b**), TOF (**c**) and FTICR (**d**) data. Only the high confidence peptides (expect value < 0.01 for a, b; FDR < 0.01 for c, d) without isotopic error are shown. The number of peptides, the mean, and standard deviation are also given.

As shown in figure 5.3, the mass measurement error distributions tend to center around zero and get narrower after recalibration. The mean mass measurement errors were < 1 ppm in Orbitrap and TOF data. The FTICR data already had a substantial residual bias to start with and did not have a high number of peptide identifications. Although the calibration improved both the accuracy and precision in the FTICR data, there was still some residual bias. Since the precursor masses are closer to their exact masses after calibration, searching the recalibrated data with the same parameters yielded more high confidence peptides in all.

Discussion and conclusions

The systematic and the random error decreases after recalibration, which is also the case with msRecal. Calibration performance, however, is dependent on many factors. Applying the correct calibration function is obviously the most important step, and msRecal tries to make this selection safe and automated by extracting relevant information from metadata. The number of high confidence peptides in the initial search is also a factor since having many potential calibrants increases the chances of good fits for the calibration function. On the other hand, significant mass deviation in the original data could have a negative impact on calibration performance. Even though this may already point to some issues in the original MS run, in most cases, msRecal still improves the mass error to a certain degree in such data. The improvement in peptide identifications could be observed better if the original and recalibrated data were searched in a narrower ppm range. The minimization of mass measurement error is beneficial for identification and should also improve quantification, as more peaks will be found within narrow mass measurement search windows in an MS1-based quantification. In this version of the software, only monoisotopic masses and unmodified peptides are used as calibrants. We plan to use them in future versions of the software as they could improve calibration performance in certain datasets.

The mzXML data format is still widely used, although mzML is (very) slowly replacing this format. However, since the PSI-MS CV annotation is not strictly enforced in mzXML, incomplete and even incorrect analyzer types are sometimes given in the metadata. For instance, the mass analyzer type for a QExactive instrument is

annotated as a quadrupole in some datasets, whereas the mass analyzer used to acquire the data is the Orbitrap, while the quadrupole is only used as a filter for selecting the precursors. For the time being, we try to come over this issue by resorting to the instrument model information. However, in the future, with extended vendor support of PSI-MS CV terms, this could be solved more easily. Marissen and Palmblad recently published a calibration method for mzML[5]. msRecal can be seen as a complement to their work since the mzXML format is still very popular and an automated calibration tool for this data type is very useful for reanalyzing publicly available data.

The msRecal tool can be incorporated into any mass spectrometry analysis workflow that analyzes data in mzXML format, as the output format is also an mzXML file. The recalibrated mzXML can be analyzed further downstream without readjusting the existing components of the pipeline. Automatic recalibration of data in public repositories using metadata facilitates reuse of this data consistent with the FAIR principles[32].

Acknowledgments

The source code is available on GitHub https://github.com/ArzuTugceGuler/msRecal and will be maintained with bug-fixes and future updates adding more functionalities. The authors would like to thank Rob Marissen for valuable advice on the project.

References

1. Mann, M. & Kelleher, N. L. Precision proteomics: The case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18132–18138 (2008).

2. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).

3. Brenton, A. G. & Godfrey, A. R. Accurate mass measurement: Terminology and treatment of data. *J. Am. Soc. Mass Spectrom.* **21**, 1821–1835 (2010).

4. Palmblad, M., Bindschedler, L. V, Gibson, T. M., Cramer, R. & Wiley, J. Automatic internal calibration in liquid chromatography / Fourier transform ion cyclotron resonance mass spectrometry of protein digests. **20,** 3076-3080 3076–3080 (2006).

5. Marissen, R. & Palmblad, M. mzRecal: universal MS1 recalibration in mzML using identified peptides in mzIdentML as internal calibrants. *Bioinformatics* **37**, 2768-2769 (2021).

6. Romson, J. & Emmer, Å. Mass calibration options for accurate electrospray ionization mass spectrometry. *Int. J. Mass Spectrom.* **467**, 116619 (2021).

7. 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).

8. Martens, L. *et al.* mzML - A community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133 (2011).

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

10. Jones, A. R. *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11**, 1–10 (2012).

11. Hoffmann, N. *et al.* MzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Anal. Chem.* **91**, 3302–3310 (2019).

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

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

14. Palmblad, M. *et al.* Improving mass measurement accuracy in mass spectrometry based proteomics by combining open source tools for chromatographic alignment and internal calibration. *J. Proteomics* **72**, 722–724 (2009).

15. Keller, A., Eng, J., Zhang, N., Li, X. jun & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005.0017 (2005).

16. Galassi, B. *et al. GNU Scientific Library Reference Manual - Third Edition.* (Network Theory Ltd., 2009).

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

18. 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).

19. Christian, N. P., Arnold, R. J. & Really, J. P. Improved calibration of time-of-flight mass spectra by simplex optimization of electrostatic ion calculations. *Anal. Chem.* **72**, 3327–3337 (2000).

20. Ledford, E. B., Rempel, D. L. & Gross, M. L. Space Charge Effects in Fourier Transform Mass Spectrometry. Mass Calibration. *Anal. Chem.* **56**, 2744–2748 (1984).

21. Shi, S. D. H., Drader, J. J., Freitas, M. A., Hendrickson, C. L. & Marshall, A. G. Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry. *Int. J. Mass Spectrom.* **195–196**, 591–598 (2000).

22. Gorshkov, M. V., Good, D. M., Lyutvinskiy, Y., Yang, H. & Zubarev, R. A. Calibration function for the orbitrap FTMS accounting for the space charge effect. *J. Am. Soc. Mass Spectrom.* **21**, 1846–1851 (2010).

23. Schlosser, A. & Volkmer-Engert, R. Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. *J. Mass Spectrom.* **38**, 523–525 (2003).

24. Haas, W. *et al.* Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell. Proteomics* **5**, 1326–1337 (2006).

25. Lichti, C. F. *et al.* Integrated chromosome 19 transcriptomic and proteomic data sets derived from glioma cancer stem-cell lines. *J. Proteome Res.* **13**, 191–199 (2014).

26. Yamana, R. *et al.* Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC-MS/MS analysis with meter-scale monolithic silica columns. *J. Proteome Res.* **12**, 214–221 (2013).

27. Worah, K. *et al.* Proteomics of Human Dendritic Cell Subsets Reveals Subset-Specific Surface Markers and Differential Inflammasome Function. *Cell Rep.* **16**, 2953–2966 (2016).

28. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

29. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–92 (2002).

30. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics - Clin. Appl.* **9**, 745–754 (2015).

31. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

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

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).

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.

Nederlandse Samenvatting

Massaspectrometrie is een krachtige techniek voor *proteomics* omdat het de benodigde hoge gevoeligheid en snelheid levert voor de analyse van het complexe en dynamische proteoom. Echter, het gebruik van deze techniek levert massaspectrometrische data op die ook zeer complex is, waarvoor een geavanceerde meerstappen analyse nodig is. Het gebruik van wetenschappelijke *workflows* voor de analyse van massaspectrometrie data is niet geheel nieuw. Desalniettemin zijn vernieuwende aanpakken hard nodig in deze tijd van grote (en complexe) data teneinde de *workflows* autonoom en intelligent te maken door middel van metadata, en flexibel en adaptief door een modulair ontwerp. De focus van dit proefschrift ligt op het creëren van intelligente *workflows* voor de analyse en het contextualiseren van massaspectrometrische *proteomics* data. De ontwikkelde *workflows* en *tools* kunnen eenvoudig worden aangepast voor additionele functionaliteiten en hergebruik in verschillende data analyse *workflows*. Het onderzoek beschreven in dit proefschrift beperkt zich niet tot de analyse van massaspectrometrie data, maar behandelt ook methoden voor literatuur onderzoek, welke cruciaal zijn voor het ontwerpen van experimenten, en de interpretatie en contextualiseren van de experimentele resultaten. **Hoofdstuk 1** geeft een overzicht van de kernconcepten van massaspectrometrische *proteomics* en computationele methoden die in de volgende hoofdstukken worden toegepast. Tevens worden uitdagingen en hoe deze momenteel worden aangepakt kort besproken.

Hoofdstuk 2 introduceert wetenschappelijke *workflows* en hun voordelen in meerstappen analyses. Zelfs in hetzelfde subveld kunnen verschillende auteurs verschillende termen gebruiken om te refereren aan dezelfde domein entiteit. Wanneer het literatuuronderzoek handmatig met specifieke trefwoorden uitgevoerd wordt, kunnen deze overlappende studies over het hoofd gezien worden. De *workflows* die hier gepresenteerd worden maken geautomatiseerde literatuuronderzoeken eenvoudiger zonder te verdwalen in complexe bibliometrische methoden. Het snel verkrijgen van een overzicht van een veld kan handig zijn voor auteurs bij het uitvoeren van interdisciplinaire studies en meta-analyses. Verder kunnen onderzoekers hiermee ook expertiselabs en andere onderzoekers vinden voor samenwerkingen. **Hoofdstuk 3** presenteert meer geavanceerde *workflows* voor

bibliometrie met het vermogen om gebruik te maken van *web services* en statistische analyses uit te voeren op de verzamelde data. Na presentatie van hoe de *web services* geïntegreerd kunnen worden in het Taverna workflowsysteem, worden workflows voor citatie netwerken en biomoleculaire interacties getoond. De output van deze processen kan gevisualiseerd worden in het populaire VOSviewer of Cytoscape, daarbij gebruikmakend van krachtige bestaande *tools*. Literatuuranalyses zijn zeer nuttig voor de interpretatie van de uitkomsten van een studie, deze binnen de context te visualiseren, en een routekaart te verkrijgen voor ontwerp van toekomstige experimenten; dit hoofdstuk bespreekt hoe deze processen geautomatiseerd kunnen worden.

Hoofdstuk 4 presenteert een robuuste tool voor integratie en anatomische visualisatie van kwantitatieve *omics* data van model organismen. Eén vernieuwing van deze *tool* is dat het gebruik maakt van anatomische ontologieën om automatisch anatomische informatie te visualiseren ongeacht de resolutie van de input data. De *tool* beweegt door de anatomische ontologie hiërarchie om het geschikte niveau van orgaan en weefsel details te visualiseren. Vervolgens wordt van de gebruiker een algemeen en simpel data formaat gevraagd voor visualisatie van de data. Aangezien dit data formaat geen op *omics*-gebaseerde aannames maakt, kan data van verschillende *omics* experimenten worden geïntegreerd, en probleemloos worden gevisualiseerd.

Hoofdstuk 5 presenteert een herkalibratie tool die de accuraatheid van de massabepaling verbetert door middel van geautomatiseerde interne kalibratie. Aan de hand van opnieuw gekalibreerde data werden meer peptiden geïdentificeerd met hogere zekerheid. De gemeten massa's zijn gekalibreerd met de accurate peptide identificaties uit de originele data of uit andere metingen van hetzelfde of vergelijkbare samples. Deze techniek maakt gebruik van mzXML metadata om automatisch de meeste geschikte kalibratiefunctie te selecteren voor de gebruikte massaspectrometer. Het output format is hetzelfde als de input dus deze techniek kan in principe eenvoudig worden verbonden met elke *bottom-up proteomics* data analyse *workflow* die gebruik maakt van mzXML.

Tenslotte, in **Hoofdstuk 6**, worden de methoden en concepten zoals data beschikbaarheid en hergebruik, automatisering en data integratie besproken in de

context van het onderzoek dat gepresenteerd wordt in dit proefschrift. Huidige en toekomstige *proteomics* data analyse wordt besproken waarbij inzicht wordt gegeven in wat noodzakelijk is voor het uitvoeren van geautomatiseerde analyses die gebruikmaken van beschikbare data uit publieke archieven met als doel bestaande kennis uit de literatuur te versnellen. Net als veel andere software en ondersteunende *tools* is het workflowsysteem, Taverna, dat wordt gebruikt in dit proefschrift, onderhevig aan veroudering. Echter, de concepten en methoden die in dit proefschrift gepresenteerd worden zorgen ervoor dat deze *tools* hiertegen bestand zijn en bieden een sterke basis voor de toekomst van *proteomics* data analyse.

Acknowledgments

Putting the seal on one of the most challenging episodes of my life is surely not easy. I would never have thought that writing the acknowledgments for my thesis would be one of the difficult tasks. There are so many amazing people that I have met that inspired me and helped me on the way, and I am grateful to all of them. It will be impossible to name them all, but I will give it a try.

First of all, I would like to thank my supervisor Magnus for giving me the opportunity to do a PhD at LUMC, which changed the course of my career and my life. Your patience and endless support made this thesis possible; thanks for all your teachings. I also would like to give a heartfelt thanks to my promotor Manfred. Your contribution to this thesis is immense; it wouldn't have been possible without you. I would like to remember and thank my initial promotor André, who is no longer with us. I know you were looking forward to my defense, and it breaks my heart that you won't be there to see it. I kept my promise to finish this thesis.

I cannot help but get emotional when I look back to the period I worked at the LUMC, during 2014-2018. So many great memories from the good old days. Anton, Simone, Suzanne, Dana, Linda, and Hans: how lucky I was to share an office with you; thanks for all the support. I have nothing but pleasant memories. Rob and Yassene, I still remember the stimulating and fun discussions we had during the time we shared an office. I was nearing the end of my contract, and your support and positivity really helped me back then; thanks a lot. Hulda, Isabelle, Bram, Frank, Guinevere, Clara, and Bart: thanks for being amazing colleagues and friends. All these fun times we had outside the LUMC, be it at Lemmy's, my place, or elsewhere, I'll always cherish. Manu, you have been a fantastic friend that helped me through the most difficult times; thank you very much for everything. Thank you, Yuri, Paul, Peter, Oleg, Martin, David, Sarantos, Aswin, Tao, and all the other members of the LUMC; it was great to get to know you and work in the same department with you.

I also would like to thank my current team at Amsterdam UMC; you have given me the opportunity to carry on my passion for science. Eric, thanks for having your trust in me and having me in your group. Karen, you have been a great colleague and a friend; thank you for all the support. Thanks for being amazing colleagues, Aleksandra,

Alicia, Jolien, Sabine, and Karlijne. I'm looking forward to the great work we will do as a team.

Linda, Anton, Suzanne, Hulda: How lucky I am to have you in my life. You know you will always be my precious friends.

Begüm, Cemre, Yasemin: Thank you for your endless emotional support during the most challenging times. I am blessed to have friends like you; one couldn't ask for more.

Last but not least, I would like to thank my brother, father, and mother. I hope I can make you proud. My dear brother Tolga, you know I wouldn't be where I am without your support; thank you very much for everything. My lovely nephew Demir, you came into my life during the first year of my PhD and have been a bundle of joy ever since. Thank you.

Curriculum Vitae

Arzu Tuğçe Güler was born on 24th of August, 1988 in Ankara, Turkey. She did her primary and secondary education in TED Ankara College and graduated from high school in 2005 as an honor student with a major in mathematics and natural sciences. She started studying Computer Engineering at TOBB Economy and Technology University (TOBB ETU) in Ankara in 2006. During the second year of her studies in 2008, she was selected by her department for the Erasmus Placement program and did a 4-month internship in Brussels, Belgium at the Turkish Research and Business Organizations Office, mainly working on tasks in relation to the EU 7th Framework Program and liaised with the national researchers. Between 2008 and 2009, she worked part-time at TOBB ETU's external affairs office as their website manager. During the third year of her studies, she did her second 4-month internship at a private information technology company where she worked on database management of the national healthcare system. She did her third and final internship during the last year of her bachelor studies at the same company and worked on a web application development project. Being always interested in bioinformatics, she did her graduation project on "algorithm development for p53 binding site prediction" and graduated with a Bachelor's degree in Computer Engineering from the Faculty of Engineering at TOBB ETU in 2010. In the fall of 2010, she started her master's studies in Bioinformatics at the Faculty of Bioscience Engineering at Katholieke Universiteit Leuven in Leuven, Belgium. She did her master's thesis at the Department of Electrical Engineering on "computational description and assessment of synonymous variations in the human genome". In 2013, she graduated with a Master's degree in Bioinformatics with a major in Bioscience Engineering. In January 2014 she moved to Leiden, The Netherlands to begin her PhD studies under the supervision of dr. Magnus Palmblad at the Center for Proteomics and Metabolomics at Leiden University Medical Center (LUMC), initially having emeritus prof. dr. Andre M. Deelder as the promotor till September 2017 and prof. dr. Manfred Wuhrer as the current promotor. During her research she attended various scientific conferences, gave talks and poster presentations, published in scientific journals, took several courses and also assisted with teaching short courses and (co-)supervised students at LUMC-Lomonosov Moscow State University Bioinformatics summer schools organized in 2014, 2015, and 2016. The papers that resulted from her PhD research are combined in this thesis entitled "Intelligent Workflows for Automated Analysis of Mass Spectrometry Based Proteomics Data". Since April 2018, she continues her research career as a Post-Doc in Reit's Group at Amsterdam UMC – AMC, focusing on proteomics of Huntington's Disease.

PhD Portfolio

Courses and certifications

Basic course for clinical investigators (BROK) 6 October 2016 - 31 March 2017
Course & National Exam NFU, The Netherlands

Introduction to 'omics data integration 7 - 10 March 2017
Course
EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Computational methods for the analysis and interpretation of mass-spectrometric high-throughput data 26 - 30 September 2016
Summer School
Leibniz-Zentrum für Informatik, Castle Dagstuhl, Wadern, Germany

Data Stewardship for Scientific Discovery and Innovation 10-15 July 2016
Summer School
LERU Doctoral Summer School, Leiden, The Netherlands

Advanced Proteomics Bioinformatics Workshop 21 June 2016
Workshop
Xth Annual Congress of the EuPA, Istanbul, Turkey

Basic Methods and Reasoning in Biostatistics 12-16 September 2016
Course
LUMC, Leiden, The Netherlands

Bioinformatics for Protein Identification 4-5 June 2016
Course
64th ASMS Conference, San Antonio, Texas, USA

Proteomics Informatics Course 22-25 September, 2015
Course
14th HUPO World Congress, Vancouver, B.C., Canada

Communication in Science 15 January – 12 February 2015
Hands-on Training
Leiden University, Leiden, The Netherlands

Fundamentals of Mass Spectrometry 23-24 August 2014
Short Course
20th IMSC, Geneva, Switzerland

Teaching and supervision experience

Teaching assistant 6 April 2017
Short course: *"Introduction to Cytoscape"*
LUMC, Leiden, The Netherlands

Lecturer 24 August 2016
Topic: "*Taverna workflows*"
Workshop on Computational Biology, Department for Biochemistry and Molecular Biology, University of South Denmark

Co-supervisor to two undergrad students July-August 2016
Project: *"Splitting tandem MS datasets into molecular classes using machine learning"*
LUMC-Lomonosov Moscow State University Bioinformatics Summer School, LUMC, Leiden, The Netherlands

Supervisor to two undergrad students July-August 2015
Project: *"Visualization of gene and protein expression data in spatial context"*
LUMC-Lomonosov Moscow State University Bioinformatics Summer School, LUMC, Leiden, The Netherlands

Teaching assistant 26-27 January 2015
Short course: *"Linux and basic scripting short course"*
LUMC, Leiden, The Netherlands

Co-supervisor to two undergrad students July-August 2014
Project: *"Barcoding of life using shotgun tandem mass spectrometry"*
LUMC-Lomonosov Moscow State University Bioinformatics Summer School. LUMC, Leiden, The Netherlands

Poster Presentations

Guler, A. T., Palmblad, M. Experimental setup independent automated internal calibration of LC-MS data for non-targeted proteomics analysis. *65th American Society for Mass Spectrometry Conference on Mass Spectrometry and Allied Topics*. Indianapolis, Indiana, USA, Jun 4 - 8, 2017.

Guler, A. T., Travin, D., Mohammed, Y., Palmblad, M. Scientific workflows for data analysis and visualization in quantitative proteomics. *64th American Society for Mass Spectrometry Conference on Mass Spectrometry and Allied Topics*. San Antonio, Texas, USA, Jun 5 - 9, 2016.

Guler, A. T., Palmblad, M. Intelligent workflows for proteomics data analysis. *Human Proteome Organization World Congress*. Vancouver, Canada, Sep 27 - 30, 2015.

Guler, A. T., Waaijer, C. J. F., Palmblad, M. Mapping scientific pedigrees and collaborative patterns using bibliometrics: six former presidents of the ASMS. *63rd American Society of Mass Spectrometry Conference on Mass Spectrometry and Allied Topics.* St. Louis, USA, May 31 - Jun 4, 2015.

Guler, A. T., Palmblad, M. Protein quantitation combining MS and MS/MS data with intelligent workflows. *63rd American Society of Mass Spectrometry Conference on Mass Spectrometry and Allied Topics.* St. Louis, USA, May 31 - Jun 4, 2015.

Guler, A. T., Mohammed, Y., Palmblad, M. Intelligent workflows for label-free, quantitative proteomics. *Dutch Techcentre for Life Sciences on track.* Amersfoort, The Netherlands, Dec 1, 2014.

Oral Presentations

Guler, A. T. Scientific workflows for combining MS and MS/MS data and improving mass measurement accuracy in proteomics. *2nd Annual Danish Bioinformatics Conference*. Odense, Denmark, Aug 25 - 26, 2016.

Guler, A. T. Integration and visualization of multi-omics data in animal models. *X Annual Congress of the European Proteomics Association*. Istanbul, Turkey, Jun 22 - 25, 2016.

Guler, A. T. Use of ontologies for automated data processing and their challenges: a bioinformatics view. *The Information Universe Conference*. Groningen, The Netherlands, Oct 7 - 9, 2015.

Guler, A. T., Waaijer, C. J. F., Palmblad, M. Scientific workflows for bibliometrics. *15th International Conference on Scientometrics and Informetric*. Istanbul, Turkey, Jun 29 - Jul 4, 2015.

Guler, A. T. Scientific workflows for proteomics. *11th Ardgour Symposium.* Ardgour, Scotland, Sep 8 - 12, 2014.

List of Publications

Sap, K. A.*, **Guler, A. T.***, Bury, A., Dekkers, D., Demmers, J. A. A., Reits, E. A. Identification of Full-Length Wild-Type and Mutant Huntingtin Interacting Proteins by Crosslinking Immunoprecipitation in Mice Brain Cortex. *J. Huntington's Dis.* **10**, 335-347 (2021). (*shared first authors)

Sap, K. A., **Guler, A. T.**, Bezstarosti, K., Bury, A. E., Juenemann, K., Demmers, J. A. A., Reits, E. A. Global Proteome and Ubiquitinome Changes in the Soluble and Insoluble Fractions of Q175 Huntington Mice Brains. *Mol. Cell. Proteomics* **18**, 1705-1720 (2019).

Travin, D., Popov, I., **Guler, A. T.**, Medvedev, D., van der Plas-Duivesteijn, S., Alvarez, M. V., Kolder, I. C. R. M., Meijer, A. H., Spaink, H., Palmblad, M. COMICS: Cartoon visualization of omics data with spatial context using anatomical ontologies. *Journal of Proteome Research,* **17,** 739-744 (2018). [This thesis]

Farnham, A., Kurz, C., Ozturk, M. A., Solbiati, M., Myllyntaus O., Meekes, J., Pham, T. M., Paz, C., Langiewicz, M., Andrews, S., Kanninen, L., Agbemabiese, C., **Guler, A. T.**, Durieux, J., Jasim, S., Viessmann, O., Frattini, S., Yembergenova, D., Benito, C. M., Porte, M., Grangeray-Vilmint, A., Curiel, R. P., Rehncrona, C., Malas, T., Esposito, F., Hettne, K. Early Career Researchers Want Open Science. *Genome Biology,* **18**, 221 (2017).

Hussaarts, L., Kaisar, M. M. M., **Guler, A. T**., Dalebout, H., Everts, B., Deelder, A. M., Palmblad, M., Yazdanbakhsh, M. Human dendritic cells with Th2-polarizing capacity: analysis using label-free quantitative proteomics. *International Archives of Allergy and Immunology,* **174**, 170-182 (2017).

Kolder, I. C. R. M., van der Plas-Duivesteijn, S. J., Tan, G., Wiegertjes, G. F., Forlenza, M., **Guler, A. T.**, Travin, D. Y., Nakao, M., Moritomo, T., Irnazarow, I., den Dunnen, J. T., Anvar, S. Y., Jansen, H. J., Dirks, R. P., Palmblad, M., Lenhard, B., Henkel, C. V., Spaink, H. P. A full-body transcriptome and proteome resource for the European common carp. *BMC Genomics,* **17**, 701 (2016).

Guler, A. T., Waaijer, C. J. F., Mohammed, Y., Palmblad, M. Automating bibliometric analysis using Taverna scientific workflows: A tutorial on integrating Web Services. *Journal of Informetrics,* **10,** 830-841 (2016). [This thesis]

Guler, A. T., Waaijer, C. J. F., Palmblad, M. Scientific workflows for bibliometrics. *Scientometrics,* **107**, 385-398 (2016766). [This thesis]