The handle http://hdl.handle.net/1887/21012 holds various files of this Leiden University dissertation

**Author:** Mostovenko, Ekaterina  
**Title:** Towards high throughput and spatiotemporal proteomics: analytical workflows and quantitative label-free mass spectrometry  
**Issue Date:** 2013-06-25
General Discussion
The application of proteomics has expanded substantially in the past decade, promoting the development of new experimental techniques, instrumentation and data analysis methodology. Each step from raw samples to lists of identified peptides/proteins and their biological interpretation always involves a choice, conscious or not, of methodology, which is often dependent on decisions made at other stages of the analysis. This thesis emphasizes and explores the importance of these choices with regard to sample preparation, instrumental setup and data analysis.

The whole course of most experiments, also in proteomics, is driven by a biological question. The smallest living biological unit which translates genomic information into proteins is the cell. The protein content of the cell defines and describes its type, biological state and function. Therefore it is logical to turn towards cell-centered proteomics. Unlike the genome, which is considered to be constant throughout the life of the cell, the proteome is variable in both time and space. Detailed studies including protein localization provide additional molecular information for a more comprehensive analysis. In different ways, both large-scale immunohistochemical efforts like the Human Protein Atlas and mass spectrometry-based proteomics can produce such knowledge. The Human Protein Atlas project uses antibodies to generate accurate and high-resolution information on protein localization. Ideally, one would want a specific reagent for each protein and for each major isoform, but this goal would be very difficult to achieve. On the other hand, mass spectrometry offers various techniques to generate information on protein localization. One of these is MS-imaging, which can be used to visualize compounds in biological tissues and may eventually be useful for clinical diagnostics. However, so far, commercially available instrumentation does not allow ‘omics’ scale applications and has a limited spatial resolution capability to visualize proteins or other molecules at the cellular or subcellular level. Both strategies mentioned above provide an actual image of the cell/tissue at the specific state of its development. Organellar proteomics can also provide knowledge about protein localization, and can be more easily applied in time course studies providing multidimensional data on protein spatiotemporal dynamics.

When approaching a particular biological question, careful choices of sample preparation and pre-fractionation methods are needed. In addition, for label-free quantitative proteomics, the choice of appropriate instrumentation is essential to perform the experiments well and to generate
data of meaningful quality. Label-free quantitation has some obvious advantages compared to metabolic or chemical labeling techniques: no artificial breakpoints in the number of the analyzed samples and no additional labeling step. However, it is more critically dependent on robust sample preparation and analytical instrumentation, and requires different data processing tools. State-of-the-art mass spectrometry, as described in Chapter 3 of the thesis, combines fast and sensitive MS/MS of ion traps and high accuracy and resolving power of an FTICR allowing for the parallel and broadband identification and quantitation of peptides. This method was demonstrated in a ‘textbook’ experiment of glucose-lactose diauxie in *E. coli* described in Chapter 5. The increasing popularity of label-free proteomics also calls for the development of better and reproducible sample preparation methods suitable for high-throughput work and large sample cohorts. This is discussed in Chapters 1 and 2 of the thesis.

In the classical Design of Experiment theory, the biological question determines or influences the experimental methodology and instrumental setup. In data-driven ‘omics’ approaches, the dependency is just as strong, but oriented in the other direction. High-information content and high-throughput approaches can be used for the generation of large multidimensional datasets which are then used as a base for forming new hypotheses. The dependency between sample and data handling is also unequivocal. The information derived from the sample preparation and fractionation methods themselves can be of further use during data processing for learning more about the proteins or (more mundanely) finding and removing erroneous or uncertain identifications. The amount of data produced in one experiment is often measured in gigabytes, and often involves hundreds or thousands of individual files, demanding new tools and methods for efficient data processing. A simple method for accelerating processing of such ‘big data’ is to use virtual machines and clouds, temporarily acquiring the necessary computing power for a very reasonable cost and without physical access to the computer hardware. One method for this is described in Chapter 4.

The process of extracting the biologically relevant information is now much faster, but nonetheless still challenging. Mass spectrometry-based proteomics experiments are no longer imaginable without a strong bioinformatic and systems biology analysis. The data interpretation is one of the most critical parts of the experiment. As an illustration of a new
framework for this purpose, the Taverna scientific workflow manager has been used throughout this thesis, in Chapters 1, 2, 4 and 6. Taverna implements automated data processing pipelines and analyses that are fully controllable by the researcher, but also supports and simplifies remote processing of large datasets on a cloud or grid. Additionally, all workflows can be shared online, enabling other researchers to completely repeat or reuse parts of the workflow for their analysis, leading to unified and more transparent data analysis.

Only a few years ago, the possibility of fast, accurate and quantitative measurement of thousands of proteins across many samples in one experiment seemed to be a utopia and microarray was the only reasonably comprehensive method to compare two systems at different states. Nowadays, MS-based proteomics is a widely-accepted universal methodology providing miscellaneous information on molecular mechanisms regulating cellular systems from the point of protein function, localization, modification and interactions. Looking into the future, it is logical to expect that a large part of the increased understanding of the life of the cell will rest on system-wide data collection, including at the protein level. With time, data driven approaches in science will mature into more robust, more quantitative, more high-throughput and more integrative methods. Another and not mutually exclusive direction would probably involve minimization of different aspects of proteomics such as analysis time and required material quantity, in ideal situations enabling single-cell or few-cells in-depth analyses. This will require improvements in both hardware and software and a close partnership between different scientific communities. For example, for the most part, mass spectrometry hardware today is developed commercially. Software connects instrumentation and applications, and the most innovative developments are unsurprisingly driven by academic research groups working directly with the scientific end users. This seems to be a fairly natural division, which also helps standardization of protocols, data formats, publication requirements and collaboration through open-source software and shared workflows.

Proteomics harbors significant promise for medicine and human health, which is illustrated by a large demand for focused clinical applications and research areas such as regenerative medicine and cancer. Novel targeted, quantitative, high-throughput methods have emerged, enabling screening of cells for many proteins in one experiment. Affinity-based proteomics such as SISCAPA™ is a quickly arising strategy unlocking the next generation,
quantitative biomarker discovery. However, by going into a detailed analysis of the molecular mechanisms, perhaps not only a marker of the problem, but also more information related to its causes could be revealed. For this, comprehensive, spatiotemporal or organellar proteomics and protein-protein interaction networks research will be important. Complexes of multiple proteins and proteins with RNA, metabolites or other molecules play key roles in regulatory processes, signaling pathways and therefore in the functioning of the cell.

Cell-centered approaches are focused in that they remove most issues of tissue or sample heterogeneity. However, any cell-based biological question is a complex task and requires integration of different perspectives, combining disciplines such as genomics, transcriptomics, proteomics, metabolomics and other ‘omics’ sciences. Understanding and thus influencing the cross-communication and dependencies between different domains of molecular biology within one cell or between different cells might lead to better diagnostics, disease prognostics and personalized medicine.

This thesis clearly demonstrates that each part of a proteomics experiment involves important choices which influence the outcome and the information gained from the experiment, and that these choices are also interdependent. It should be emphasized that only a well-tuned combination of such method parameters will lead to valuable and meaningful results. Mass spectrometry-based proteomics has rightfully established itself in molecular diagnostics and fundamental research providing complementary information to other ‘omics’ disciplines.
REFERENCES


