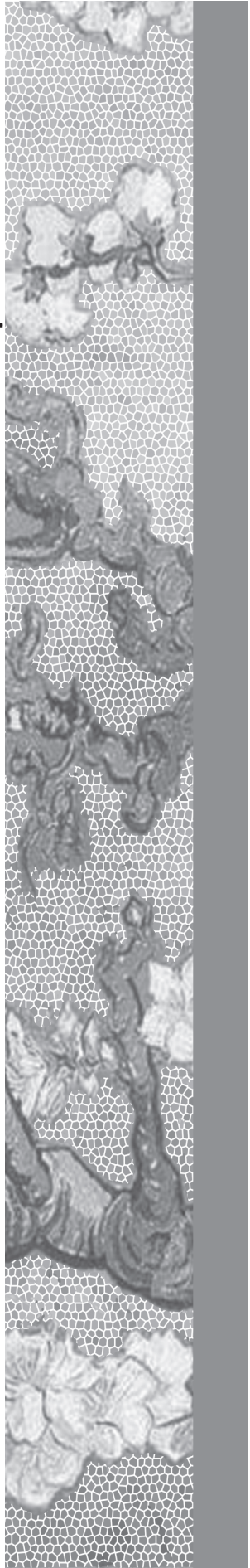
General Introduction



Proteomics

Most cellular processes are controlled by proteins. Their value in sustaining life is difficult to overestimate. Unlike the genome, the protein content of the cell – the proteome – is always changing: new proteins are continuously produced, modified, transferred from one subcellular compartment to another or degraded and removed. The proteome reflects the cellular state or external condition encountered by a cell, and therefore determines biological processes and pathways at a specific point in time.[1-3] The proteins are produced from the DNA blueprints via messenger RNA that represent the momentary "activity" of the genome. However, gene expression does not always correlate with the protein expression.[4-6] Even if the mRNA is expressed or expressed at a different rate, the protein may not be similarly expressed or changed in abundance, due to post-transcriptional regulation, protein transport or degradation. In order to be able to study and treat physiological changes in cells or the entire body, transcriptomics and proteomics have developed in parallel and complementary to each other. Unfortunately, the extrapolation from the genome is limited due to lack of (experimental) knowledge of gene (actually protein) function and incorrect annotations. Proteomics, on the other hand, can be viewed as an experimental approach to explain the information contained in genomic sequences in terms of their expression, subcellular localization, structure, interactions, biochemical functions and states of modification, all of which are interrelated. These aspects of proteins are all more or less amenable to a proteomics approach, the most difficult being structural determination and biochemical properties, which usually requires substantial amounts of the proteins to be purified. In the other words, proteomics attempts to study biological processes comprehensively by the systematic analysis of the proteins expressed in a cell or tissue.

Mass spectrometry (MS) has become a key tool for proteomic analysis and has made proteomics a powerful method for the identification, annotation and quantitation of proteins in large-scale studies. Over more than a century, mass spectrometry has been improving in separating chemical species of different mass. The development of electrospray ionization (ESI)[7, 8] and matrix assisted laser desorption/ionization (MALDI)[9] has revolutionized the way of protein analysis. These techniques introduced ways to generate ions from large, nonvolatile, molecules like proteins and peptides, and to transfer them directly into the gas phase for the MS analysis.[10] Unknown peptides or proteins can be routinely and automatically identified by data-dependent

tandem mass spectrometry (MS/MS). Recently described data-independent acquisition methods, such as SWATH MS,[11] also enables extensive proteome coverage in a fast, consistent and accurate manner. It cyclically records fragment ion spectra of all precursor ions contained within user-defined RT-*m/z* swath throughout the LC run. The resolution or peak capacity can then be increased further by coupling the mass spectrometer to a liquid chromatography (LC) system. However, the electrospray ionization is sensitive to presence of salts and even though LC does help in removing contaminants, additional purification and cleaning steps are often necessary.

The speed, accuracy and a large variety of mass spectrometry tools have brought proteomics to a new level, creating great possibilities for research applications. The area of proteomics dedicated to post-translational modifications (PTMs) is primarily concerned with the study of cell signaling and regulation which cannot be directly investigated by the genomic tools. The new field of clinical proteomics has emerged aiming at the protein profiling in large numbers of samples, and in drug and biomarker discovery studies. Biofluids commonly used for such investigations poses a great challenge due to their wide dynamic range of more than 10 orders of magnitude difference in concentration. Clinically relevant molecules are usually present at ng/mL levels and are near or below the detection limit of the currently available LC-MS/MS analysis tools.[12] Development of such approaches as selected reaction monitoring (SRM) for targeted MS measurements has made possible robust and sensitive measurements of protein biomarkers. Picotti *et al.* has shown that LC-SRM-MS improves the lower detection limit by up to 1000-fold and that the method therefore is suitable for the quantification of proteins over a large part of the range of cellular and body fluid concentrations.[13]

In parallel, one of the primary objectives of proteomics has become not only protein identification but also quantification of the differences between samples. Numerous labeling techniques are available for incorporating isotopic or fluorescent groups to the protein or peptide and are usually oriented only on the limited number of targets. Stable isotope standards and capture by anti-peptide antibodies (SISCAPA)[14] combines high sensitivity of SRM and high-throughput approaches, enabling rapid detection and quantitation of low abundant species in the biofluids.[15] Despite the high accuracy of these methods there are several obvious drawbacks. The sample preparation method is usually complex and involves additional labeling steps, the result is strongly dependent on the labeling efficiency and the

number of fractions which can be analyzed is always limited. It is also possible to quantify peptides/proteins directly from the mass spectrometry signal, so-called *label-free* quantitation. Even though undersampling in complex samples is still a limitation in label-free LC-MS or LC-MS/MS and spectral counting approaches, a recent study of Nagaraj *et al.* achieved nearly full coverage of the yeast proteome in a single-shot label-free analysis, illustrating these methods also harbor significant promise.[16] These methods are particularly useful for large clinical studies and time- or space-resolved proteomics in systems that are not easily or inexpensively labeled. Such approaches demand rapid, robust and highly reproducible sample preparation methods, preferably automated data analysis procedures and benefit from using high resolving-power mass spectrometers.

Design of experiment

As proteomic studies are usually complex and require many subsequent steps, it is crucial to have a clear overview of the experiment. Scientists have long realized the necessity to properly design experiments prior to their execution. Even though intuitively basic aspects of experimental design such as comparison, controls and repetition have been used since the beginning of science, the first formalized methodology for design of experiments (DOE) was suggested by Fisher in 1926.[17, 18] The aim is to design the most meaningful experiment which will provide clear and easily interpreted answers to the research question. DOE is a process which includes anticipation of the different variables and parameters that would influence the outcome of the experiment. From here there are two possible ways to go. The most common approach is to define the hypothesis which describes the theoretical or expected outcome, for instance the null hypothesis stating an absence of an effect, causality or correlation. Experiments are then performed to answer whether the hypothesis is true or false, or whether the null hypothesis can be rejected. Biomedical research in general still rests on this hypothesis-driven methodology, focusing on explaining the cause-effect relationship between controlled and observed variables. In this case, application of the established principles of DOE allows us to determine which number of samples or replicates is needed to observe a statistically significant effect, given the expected biological and technical sample-to-sample variability. It also informs us on which biological or technical controls or references are required. In protein

analyses, this implies a targeted approach, focused on small differences or changes in a small number of components.

With the rapid development of "omics" technologies, the amount of data that can be produced in a short time has increased tremendously, potentially generating large amounts of information. Today we are able to measure the expression of thousands of genes or proteins in a single analysis.[16] In such "omics" contexts, the goals of experimental design are radically different. The purpose of the experiment is more often than not to illuminate a biological system as completely as possible without knowing *a priori* which ones out of the thousands of analyzed components (transcripts, peptides, phosphorlylation sites) are most important. Compared to targeted analyses, we often have to sacrifice some analytical sensitivity. On the other hand, the data can guide the design further, targeted and more sensitive experiments to investigate a particular pathway or reaction in detail. To make an analogy, a targeted approach would look closely at the paintings of Vincent van Gogh and see only seemingly randomly and crudely applied strokes of paint. The "omics" approach is to take a step back and look at the paintings from a distance, from which you can see the entire picture with its story, composition and rich palette of colors. Similarly data-driven approaches in proteomics give an overview of the major characteristics of a biological system in a certain state, down to some but not infinitesimal level of refinement or detail. One could consider this lowest level of detail the "resolution" of the proteomics experiment. Subsequent experiment can of course be more targeted, looking into more detail but relinquishing the global level of analysis. This is essentially what is meant by data-driven, hypothesis generating experiments.

In a generalized sense, the objective of DOE in data-driven proteomics is to maximize the quality as well as quantity of data, such as the number of identified proteins, quantified peptides or detected PTMs, obtained from an experiment within time and economic constraints, or possibly even minimizing time and/or financial costs. There are a number of important variables or choices that need to be made for optimizing such experimental designs. When working in a purely data-driven mode, generating high-dimensional data to produce a base for formulating new hypotheses rather than answering the existing ones, we need to define the state of the system and what biological processes are to be illuminated. What magnitude of changes do we expect and on what time scale? This then determines how the biological system, *e.g.* cell culture or animal, should be sampled, how

much sample is needed, and how it needs to be prepared. In the time-course studies, how do we cultivate and harvest sufficient numbers of synchronized cells in a reproducible manner and under carefully controlled conditions while allowing sample collection at any given time without disturbing the culture(s)?

If we wish to fractionate the cells into compartments, how many cells are needed in the starting material to recover sufficient amount of analyte in each fraction? What subcellular or protein fractions should we enrich or purify, and how do we best set up the methods for analyzing the proteins or peptides using different fractionation techniques? How frequently does the system need to be sampled in order to follow rapid or oscillating processes, analogously to the Nyquist sampling theorem?[19, 20] What methods can be used for robust and easily parallelized (high throughput) sample preparation required in larger clinical studies? How do we achieve the best proteome coverage with the best possible accuracy and precision in identifying and quantifying the proteins? These are some of the most common considerations in planning proteomics experiments, and they can all be addressed by making conscious, systematic choices where each choice is dependent on other parameters. For example, opting for a bioreactor to cultivate and sample bacteria under controlled conditions for time-course studies, or the use of SDS-PAGE as an extra dimension of protein separation for increased proteome coverage for MS/MS-based analysis instead of digestion of a total cell extract.

Sometimes the dynamics of the system or the cellular population or subcellular component of most interest are unknown. A small, well-designed pilot experiment can then be useful in planning larger proteomics experiments. Typical parameters that one would determine from a pilot experiment are the amount of material needed, the time points to be sampled and the reproducibility of the measurements. Note that these may not require "omics" technologies – for instance system dynamics can be investigated using other, less expensive, readouts, such as microscopy or simple biochemical assays.

In the ideal situation, the data generated by a proteomics experiment would, with the proper analysis and interpretation, generate sufficient information to formulate single/few gene/protein-pathway hypotheses and a clear suggestion how to test these hypotheses. High-dimensional proteomics data, *e.g.* proteins and isoforms with an abundance varying in time and in cellular

localization, requires extensive mathematical and statistical analysis, including false identification/false discovery rate (FDR) estimation at different levels. In addition, as the field and experimental capabilities are rapidly growing, not seldom requiring development of new tools for automating and accelerating data processing from raw mass spectra to biological modeling and visualization (Figure 1).

When planning a proteomics study, the main question to be answered is what we want to achieve by the experiment. If properly posed, it can guide the design of the experiment to what cells, tissues or body fluids to use, when and how to collect samples, which separation and prefractionation techniques to exploit and what statistical and data analysis tools to apply.

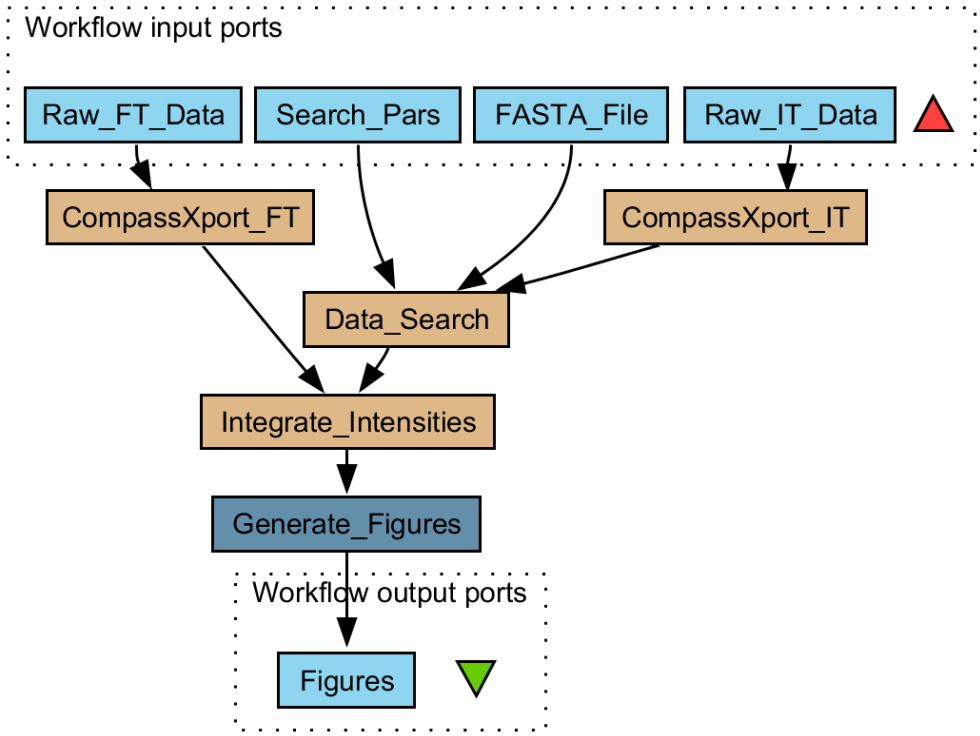


Figure 1. Simplified representation of the Taverna scientific workflow for the data processing. The raw ion trap data is converted to mzXML format and searched. Resulting library of accurate mass and time tags is then aligned with FT-MS data. Peptide identifications are grouped in proteins. For each match the intensities within retention time and m/z window are integrated. Final table is passed to Rshell where statistical analysis and plotting is performed.

A typical proteomics experiment

Most proteomics experiments consist of several stages: cell cultivation and/or sample collection, sample preparation, protein and/or peptide separation, data acquisition and data analysis. As there are no specific amplification techniques analogous to the polymerase chain reaction (PCR)[21] or sequencing by synthesis,[22] it is important to extract proteins efficiently, avoiding unnecessary losses during sample collection and preparation and rather enrich the sample in proteins of interest. Knowing at least approximately the optimal amount of peptides to load on the chromatographic column coupled directly to the mass spectrometer and estimating the protein or peptide yield, concentration or dilution in each step, it is possible to calculate backwards to the minimum appropriate amount of starting material required for an experiment. However, there are no universal rules on how to know in advance the amount of biological material, e.g. plasma volume, numbers of cells or mass required for a particular experiment. The choice is most often based on experience with similar samples or established by trial-and-error and dependent on which instrumentation is to be used for the analysis (different separation methods or direct infusion), the method of sample preparation (pre-fractionation or the total cell extract) and the nature of the sample (cells, organelles, tissues or biological fluid). For example, when comparing two different pre-fractionation techniques for increasing proteome coverage, our system for SDS-PAGE, separation of proteins does best with *ca.* 30 μg of protein, whereas our larger-volume isoelectric focusing device functions best with 100-250 μg peptides.[23]

To illustrate this using a real example, consider the following experiment: Bacteria, as most other organisms, are sensitive to environmental changes such as oxygen deprivation, high salt concentration or temperature alternation. If the temperature is raised quickly, such as would happen during high fever, the bacteria will suffer a *heat-shock*. Using the common and well-studied *Eschericia coli* model because of its simple culturing procedure and well-characterized and relatively small genome/proteome (~5,000 genes), we can investigate what happens at the proteome level in this bacterium during and after heat-shock. To observe these changes, cells need to be collected at the same time points from both normal/optimal growth conditions as well as at elevated temperature or stress conditions. Define time zero as the moment when half of the cell cultures, leaving the other half at the 37°C, are moved to a growth chamber set at 42°C (the heat-shock environment) (Figure 2a). To be extra careful, we can take all the

cultures out, before returning them to the growth chambers, to ensure they are handled in the same way except for the growth chamber temperature. As mentioned above, the culture volume we need to collect at each time point is dependent on the amount of protein needed, our estimated recovery during sample preparation and the cell density. The amount of protein extractable from a single cell obviously depend on the size of the cell. A prokaryote such as *E. coli* is only 0.5-5 μm^3 in volume, while human cells range from 100 to 100,000 μm^3.[24] It can not be assumed that the cell density or cell sizes will be the same in the heat-shocked cultures as in the control cultures (in fact, we know it will not be, as 42°C is not optimal for the growth of *E. coli*), so this needs to be taken into consideration when planning the experiment. To have a reference and simple readout of the experiment, it is useful to control the cell density, which can be seen as a marker for cell "well-being" and also gives an idea of the time scales of the processes involved and what time points should be sampled during the experiment. Before time "0", sampling can be done less frequently, as we assume cells are growing in the log phase, and a couple of time points are always useful to demonstrate that no significant changes, except for the rapid gain of the biomass, occur during this time. Even though changes at the protein level, essentially integrating gene expression over time (in simple systems), are less rapid than changes in gene expression, it is important to arrest cell growth and protein synthesis quickly immediately after collection, as the cells will otherwise keep growing and dividing, reacting to the new environment and leading to unwanted bias in the data. To ensure each sample is a "snapshot" of the cells, we quench all the cellular processes at the moment of sampling by instant cooling them by adding ice, then removing the growth medium and washing the cells in a sterile buffer.

When working with cells, we have to disrupt the cell wall to release proteins and get a high and reproducible yield. Depending on the biological system investigated and the compatibility of downstream sample preparation methods, we can opt for a mild lysis with detergent-free buffers or a harsher mechanical disruption with beads in a high concentration urea or extraction in a hot ultrasonic bath with SDS. For a label-free method, or any method that does not label the cell already in the culture or includes internal standard, the reproducibility of the protein extraction method is of paramount importance. Many commercial kits are available for protein extraction and each laboratory typically develops their own extraction protocols that work well in their lab with the available equipment and typically contain enzymes for breaking down the cell wall (lysozyme) and

DNA (an endonuclease). The latter is practical, as the extracts otherwise become extremely viscous, making pipetting and further sample preparation more difficult and likely less reproducible.

The "box standard" proteomic approach is bottom-up, operating on peptides obtained from protein extracts by proteolysis. Proteolysis by enzymatic digestion can be performed in free solution[25] as well as in-gel after protein separation by electrophoresis[26] or on filter[27] and is done in a few simple steps: reduction of disulfide bonds (cystines), alkylation of cysteins and finally enzymatic cleavage of the peptide bond by a more or less specific protease (Figure 2b). Even after the inherent sample cleanup during extraction and digestion, especially when using in-gel or in-filter digestion, the resulting peptide mixture is still too complex for direct analysis by mass spectrometry. At least one more separation step at the peptide level is required for deep proteome coverage. A comparison between different protein and peptide fractionation methods is found in **Chapter 1** of this thesis. In short, fractionation techniques are based on various physicochemical properties and aim to reduce sample complexity and/or enrich or deplete certain proteins or peptides and are often combined in multidimensional systems, connected off-line or on-line, with a final peptide separation by reversed-phase liquid chromatography (RPLC) introducing an orthogonal dimension of separation based on the hydrophobicity of the peptides.[28, 29] Obviously it makes no sense to combine similar separation techniques. In practice, most separations are oblique, *i.e*. not fully orthogonal, as fundamental properties such as size or charge always have some influence on the separation. The main reason RPLC is used last is that the mobile phase is fully compatible with electrospray ionization (or conversely, an ideal electrospray solvent still works as a mobile phase in RPLC).

Returning to the experiment, we have sampled the bacterial cultures a number of time points in replicate. Even in a simple and limited experiment such as this, we will have in the order of 100 samples that need to be prepared and analyzed. This is unavoidable if we want to study real biological processes, which are always dynamic, and need biological replicates to get meaningful results. In practice, this limits the number of dimensions of fractionation or separation to one or perhaps two. For *E. coli*, we can extract the proteins using a commercial lysis cocktail, such as BugBuster® from Novagen, and proceed directly with reduction, alkylation and digestion with trypsin. The digests are reasonably compatible with

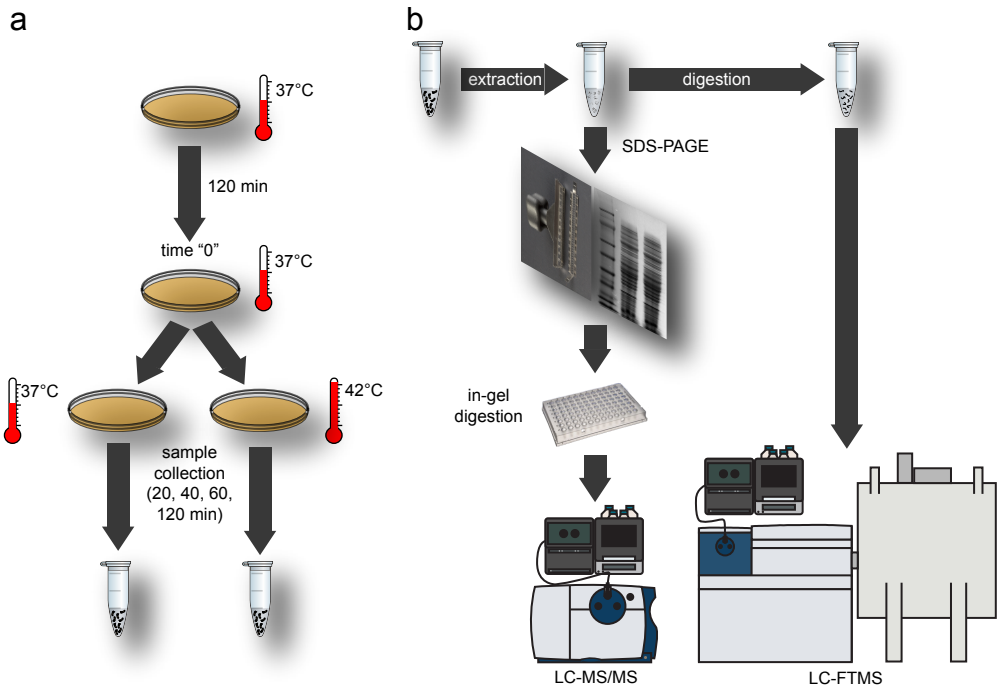


Figure 2. Experiment workflow. Sample collection (a) and sample preparation and mass spectrometric analysis (b). E. coli cells are incubated at two different temperature conditions (37°C and 42°C) and collected at times "0", 20, 40, 60 and 120 min after the splitting of cell culture. Proteins then extracted and split for SDS-PAGE and in solution digestion. SDS-PAGE is cut into 48 equal slices, placed in the 96-well plate and digested in gel. Resulting peptides are analyzed with LC-MS/MS. Peptides obtained after in solution digestion are analyzed with LC-FTMS.

RPLC, as long as trap columns are used when loading the sample. Each sample is then analyzed by LC-MS, ideally using a high-resolution mass spectrometer such as TOF or FTICR. The entire data acquisition workflow is described in **Chapter 3** of this thesis. Briefly, peptides are quantified from their intensity (peak height or peak area) in the LC-MS data while the peptide identification can be done on a different type of mass spectrometer, such as an ion trap (Figure 2b). This can be in the same sample using MS/MS, or, since especially close time points and biological replicates will contain many of the same proteins, albeit in different concentration, we can generate a small database of peptide identification with RPLC retention times and use the accurate mass and time (AMT) approach.[30, 31] We can even combine multiple ion traps and allow an extra dimension of fractionation to improve the identification of low-abundant peptides. In our experiment we used SDS-PAGE to create a library of identified peptides

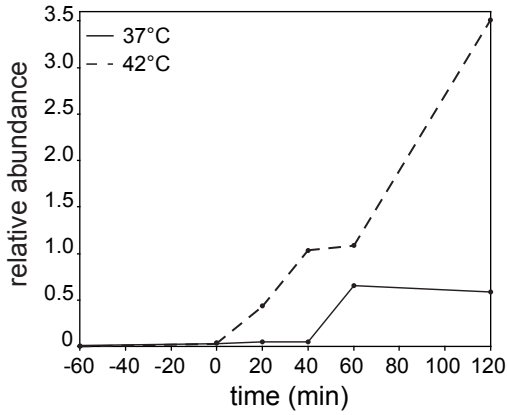


Figure 3. Protein expression profile of chaperon protein DnaK during the heat-shock. The abundance is calculated relative to changes in 30S ribosomal protein S1 which is essential for the growth[33] and has a more stable expression.

and proteins. We identified and quantified 616 proteins including Chaperone protein DnaK (UniProt accession number P0A6Y8), homologous to eukaryotic Heat-shock protein 70, Hsp70. This protein is known to be induced by the heat-shock (hence its name) and was clearly expressed more in cells which were shocked at 42°C (Figure 3). This finding is consistent with gene expression.[32] The protein expression can be mapped onto protein interaction or metabolic pathways for biological interpretation and hypothesis generation.

To summarize, proteomics is a powerful tool, both for describing biological systems in specific states as well as for quantifying differences between states or systems. However, the planning and execution of proteomics experiments remain complex and this thesis attempts to illuminate some of the most critical aspects of designing such experiments, including sample preparation, fractionation and enrichment, and data acquisition, analysis and visualization, in fundamental biological and clinical research.

REFERENCES

1. Wasinger, V. C.; Cordwell, S. J.; Cerpa-Poljak, A.; Yan, J. X.; Gooley, A. A.; Wilkins, M. R.; Duncan, M. W.; Harris, R.; Williams, K. L.; Humphery-Smith, I., Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **1995,** 16, (7), 1090-4.
2. Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J. C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphery-Smith, I.; Williams, K. L.; Hochstrasser, D. F., From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **1996,** 14, (1), 61-5.
3. Humphery-Smith, I.; Blackstock, W., Proteome analysis: genomics via the output rather than the input code. *J Protein Chem* **1997,** 16, (5), 537-44.
4. Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I., A sampling of the yeast proteome. *Molecular and Cellular Biology* **1999,** 19, (11), 7357-7368.
5. Griffin, T. J.; Gygi, S. P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R., Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Molecular & Cellular Proteomics* **2002,** 1, (4), 323-333.
6. Kolkman, A.; Daran-Lapujade, P.; Fullaondo, A.; Olsthoorn, M. M. A.; Pronk, J. T.; Slijper, M.; Heck, A. J. R., Proteome analysis of yeast response to various nutrient limitations. *Molecular Systems Biology* **2006,** 2.
7. Yamashita, M.; Fenn, J. B., Electrospray ion-source - another variation on the free-jet theme. *Journal of Physical Chemistry* **1984,** 88, (20), 4451-4459.
8. Alexandrov, M. L.; Gall, L. N.; Krasnov, N. V.; Nikolaev, V. I.; Pavlenko, V. A.; Shkurov, V. A., Ion extraction from solutions at atmospheric-pressure - a method of mass-spectrometric analysis of bioorganic substances. *Doklady Akademii Nauk Sssr* **1984,** 277, (2), 379-383.
9. Karas, M.; Bachmann, D.; Hillenkamp, F., Influence of the wavelength in high-irradiance ultraviolet-laser desorption mass-spectrometry of organic-molecules. *Analytical Chemistry* **1985,** 57, (14), 2935-2939.
10. Chait, B. T.; Kent, S. B., Weighing naked proteins: practical, high-accuracy mass measurement of peptides and proteins. *Science* **1992,** 257, (5078), 1885-94.
11. Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **2012,** 11, (6), O111 016717.
12. Surinova, S.; Schiess, R.; Huttenhain, R.; Cerciello, F.; Wollscheid, B.; Aebersold, R., On the development of plasma protein biomarkers. *J Proteome Res* **2011,** 10, (1), 5-16.
13. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009,** 138, (4), 795-806.

14. Anderson, N. L.; Anderson, N. G.; Haines, L. R.; Hardie, D. B.; Olafson, R. W.; Pearson, T. W., Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* **2004,** 3, (2), 235-44.

15. Hossain, M.; Kaleta, D. T.; Robinson, E. W.; Liu, T.; Zhao, R.; Page, J. S.; Kelly, R. T.; Moore, R. J.; Tang, K.; Camp, D. G., 2nd; Qian, W. J.; Smith, R. D., Enhanced sensitivity for selected reaction monitoring mass spectrometry-based targeted proteomics using a dual stage electrodynamic ion funnel interface. *Mol Cell Proteomics* **2011,** 10, (2), M000062-MCP201.

16. Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M., System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol Cell Proteomics* **2012,** 11, (3), M111 013722.

17. Fisher, R. A., The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **1926,** 33, 503-513.

18. Fisher, R. A., *The design of experiments*. Oliver and Boyd: 1935.

19. Nyquist, H., Certain topics in telegraph transmission theory. *Transactions of the A. I. E. E.* **1928**, 617–644.

20. Shannon, C. E., Communication in the presence of noise. *Proceedings of the I.R.E.* **1949,** 37, (1), 10-21.

21. Mullis, K. B.; Faloona, F. A., Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **1987,** 155, 335-50.

22. Ronaghi, M.; Karamohamed, S.; Pettersson, B.; Uhlen, M.; Nyren, P., Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **1996,** 242, (1), 84-9.

23. Hubner, N. C.; Ren, S.; Mann, M., Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics* **2008,** 8, (23-24), 4862-72.

24. Moran, U.; Phillips, R.; Milo, R., SnapShot: key numbers in biology. *Cell* **2010,** 141, (7), 1262-1262 e1.

25. Dong, M. W., Tryptic mapping by reversed phase liquid-chromatography. *Advances in Chromatography* **1992,** 32, 21-51.

26. Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **2006,** 1, (6), 2856-60.

27. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nat Methods* **2009,** 6, (5), 359-62.

28. Geng, X.; Regnier, F. E., Retention model for proteins in reversed-phase liquid chromatography. *J Chromatogr* **1984,** 296, 15-30.

29. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem* **2005,** 77, (19), 6426-34.

30. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y. F.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R., An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2002,** 2, (5), 513-523.

31. Strittmatter, E. F.; Ferguson, P. L.; Tang, K. Q.; Smith, R. D., Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2003,** 14, (9), 980-991.

32. Tilly, K.; McKittrick, N.; Zylicz, M.; Georgopoulos, C., The dnaK protein modulates the heat-shock response of Escherichia coli. *Cell* **1983,** 34, (2), 641-6.

33. Sorensen, M. A.; Fricke, J.; Pedersen, S., Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. *J Mol Biol* **1998,** 280, (4), 561-9.