TOWARDS HIGH THROUGHPUT AND SPATIOTEMPORAL PROTEOMICS.

Analytical workflows and quantitative label-free mass spectrometry

Ekaterina Mostovenko

ISBN: 978-94-6182-295-6

Cover design: Bart van Engeldorp Gastelaars

Cover is based on: Vincent van Gogh, Almond Blossom, 1890, Van Gogh Museum, Amsterdam.

Printing: Off Page, Amsterdam, www.offpage.nl

TOWARDS HIGH THROUGHPUT AND SPATIOTEMPORAL PROTEOMICS.

Analytical workflows and quantitative label-free mass spectrometry

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus Prof. Mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op 25 Juni 2013
klokke 10:00 uur

door

Ekaterina Mostovenko

geboren te Tver, Rusland in 1987

PROMOTIECOMMISSIE

Promotor: Prof. Dr. A.M. Deelder

Co-promotor: Dr. M. Palmblad

Overige leden: Prof. Dr. L. Maartens
VIB Department of Medical Protein Research, Ghent University, Ghent, Belgium

Prof. Dr. J. Bergquist
Department of Chemistry – BMC Analytical Chemistry, Uppsala University, Uppsala, Sweden

Prof. Dr. J. Kok
Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands

Prof. Dr. A.E. Gorbalenya

Prof. Dr. Christine Mummery

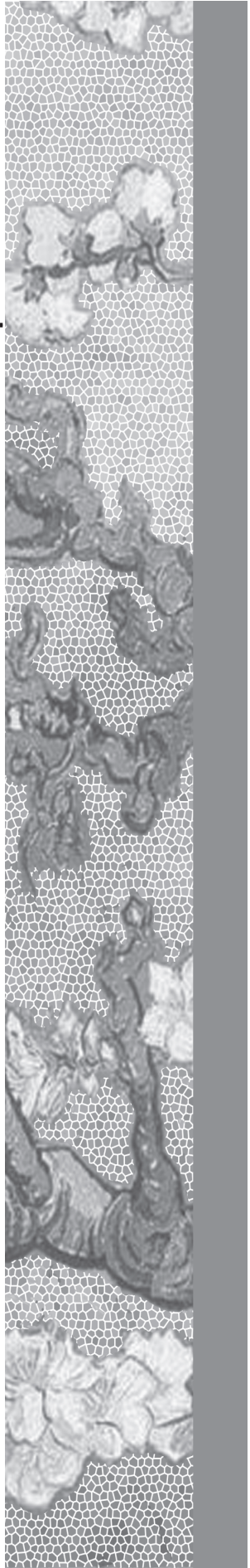
Ducunt volentem fata, nolentem trahunt

(The fates will lead who is willing, the unwilling they drag)

Seneca

TABLE OF CONTENTS

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.

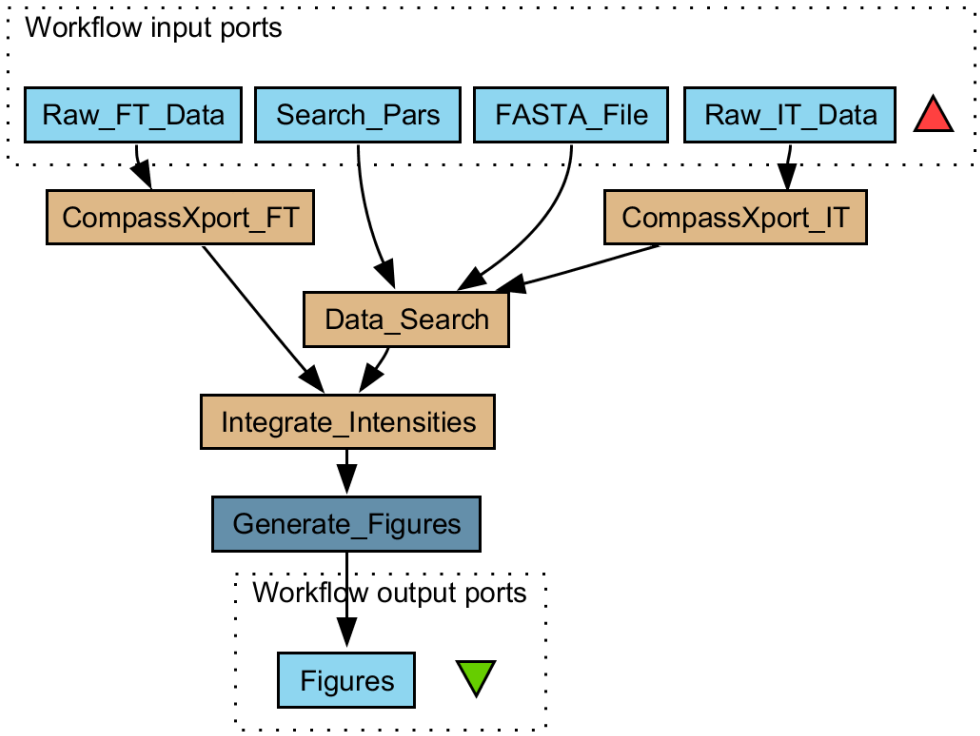


Figure 1. Simplified representation of the Taverna scientific workflow for the data processing. The raw ion trap data is converted to mzXML format and searched. Resulting library of accurate mass and time tags is then aligned with FT-MS data. Peptide identifications are grouped in proteins. For each 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

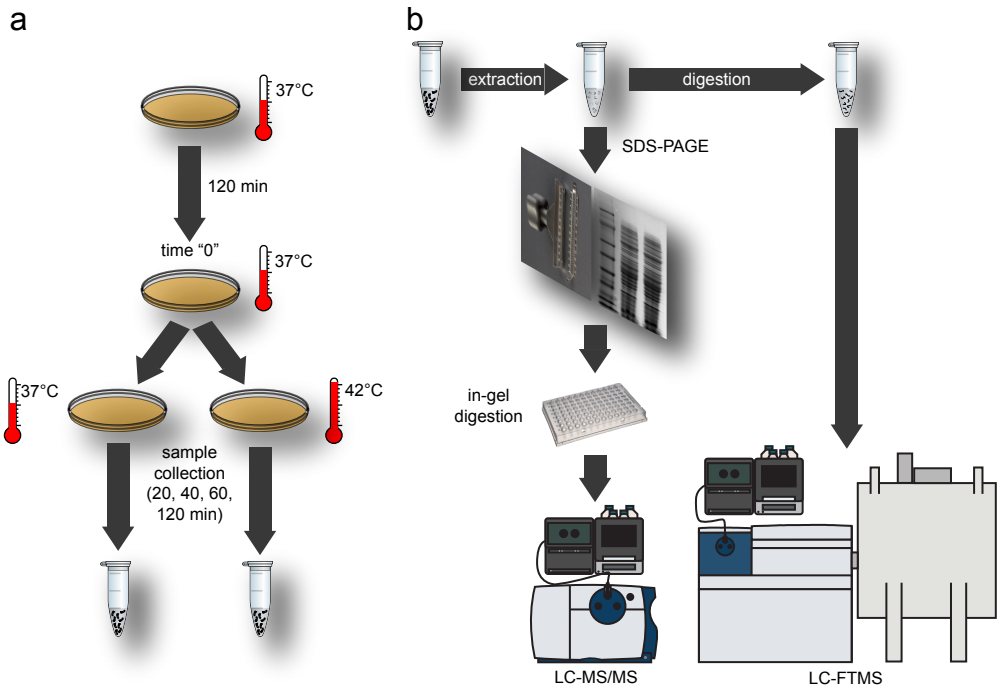


Figure 2. Experiment workflow. Sample collection (a) and sample preparation and mass spectrometric analysis (b). E. coli cells are incubated at two different temperature conditions (37°C and 42°C) and collected at times "0", 20, 40, 60 and 120 min after the splitting of cell culture. Proteins then extracted 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

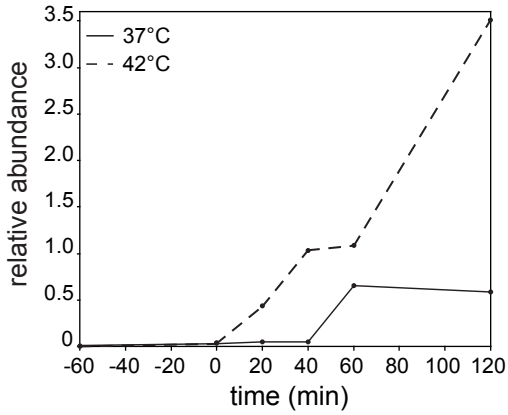


Figure 3. Protein expression profile of chaperon protein DnaK during the heat-shock. The abundance is calculated relative to changes in 30S ribosomal protein S1 which is essential for the growth[33] and has a more stable expression.

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.

1

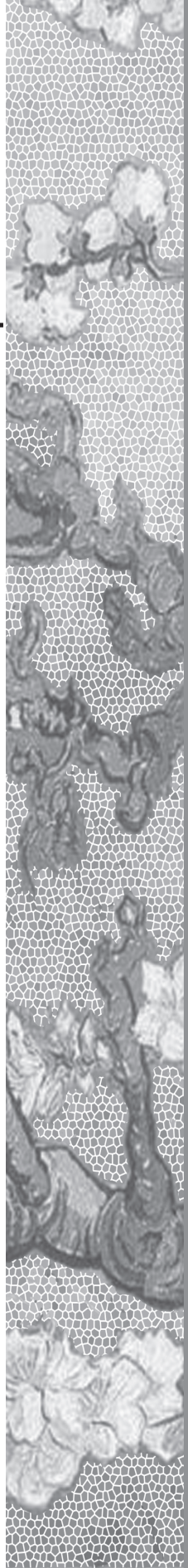
Comparison of Peptide and Protein Fractionation Methods in Proteomics

Ekaterina Mostovenko,[1] Chopie Hassan,[2] Janine Rattke,[1] André M. Deelder,[1] Peter van Veelen[2] and Magnus Palmblad[1]

[1] Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

[2] Department of Immunology and Blood Transfusion, Leiden University Medical Center, The Netherlands

Manuscript submitted to EuPA Open Proteomics Journal pending revisions



ABSTRACT

Multiple fractionation or separation methods are often combined in proteomics to improve signal-to-noise and proteome coverage and to reduce interference between peptides in quantitative proteomics. Furthermore, a given fractionation method provides additional information on the analytes, such as molecular weight, hydrophobicity or isoelectric point that can be used to improve identification, and to discover protein splice variants or large post-translational modifications. Finally we describe a Taverna scientific workflow for analysis and comparison between strong cation exchange chromatography (SCX), peptide isoelectric focusing (pIEF) and SDS-PAGE performed using robust capillary LC and ion trap tandem mass spectrometry.

INTRODUCTION

Even with the recent improvement in speed and sensitivity of tandem mass spectrometry and performance of liquid chromatography systems, loading capacity and ion suppression still limit the coverage of complex samples, such as in proteomics. Thus, the prefractionation or reduction of complexity of samples is still beneficial in most analyses, when sufficient amounts of material are available. In general, each fraction contains a "simplified" mixture of peptides/proteins enabling identification and possibly quantitation of more peptides and proteins, including those of lower abundance. At the same time, fractionation adds information about the analytes without any additional analytical effort. This information can be used together with the tandem mass spectrometry data in the validation of peptide-spectrum matches.

A wide range of fractionation strategies for peptides and proteins are generally available, often combined in multidimensional methods or systems. Any type of chromatographic separation can be used at the protein level, including ion exchange, [1] reversed phase,[2] hydrophobic interaction[3] or size exclusion,[4, 5] prior to digestion. Ion exchange chromatography is frequently combined with reversed-phase chromatography, also at the peptide level, either off-line or on-line in the same column (MudPIT).[6] Other popular methods include Gelfree® fractionation system and SDS-PAGE.[7] The last involves protein fractionation according to molecular weight, slicing the entire gel lane containing the proteins and then digesting the proteins in the gel. Isoelectric focusing of peptides or proteins can be done in capillaries,[8, 9] segmented tubes,[10-12] gels[13] or liquid compartments connected by a gel.[14]

In this work we have attempted to compare, with as little bias as possible, three very different and commonly used fractionation methods for two very different types of samples. We compared SDS-PAGE fractionation at the protein level,[7] with Off-Gel™ isoelectric focusing, fractionating according to their isoelectric point,[13] and strong cation exchange (SCX) chromatography, separating based on size and charge at a fixed pH,[1] both at the peptide level.

Several previous studies have already been published for comparing these and other fractionation methods.[15-17] However, the choice of the best method likely also depends on the sample. We therefore compared the same three methods using exactly the same protocols for two different biological samples – an *Escherichia coli* whole cell lysate and human plasma. The *E. coli* cell lysates are easy to work with and not dominated by a few proteins. Human plasma on the other hand, is dominated by a small number of proteins, with albumin making up 45-50% of the total protein content, immunoglobulin G and transferrin another 8-20% and 3-7% respectively.[18] The 20 most abundant proteins constitute more than 99% of the total protein content in plasma.[18] Both samples are easily obtained in large (even gram) quantities, making it possible to use almost any method for fractionation, from preparative scale chromatography to microfluidic methods coupled directly to the mass spectrometer.

The three compared methods each contribute information about a different peptide or protein property. This information can be used by some algorithms and pipelines to validate peptide and/or protein identification and remove erroneous identifications. In SDS-PAGE, the position of the protein on the gel has direct relationship with its molecular weight. When the measured protein molecular weight is compared with that predicted from the genome and used for the peptide identification, splicing events or post-translational processing could be detected. In IEF, the distribution of the peptides corresponds to their pI, which can also be predicted, albeit not with perfect accuracy. Finally, in SCX, the elution time (*i.e.* fraction number) depends on the size and charge of the peptides at the system pH,[19] which may also be possible to predict from the peptide sequence. The Trans-Proteomic Pipeline[20] (TPP) already provides a standard score (also known as Z-score) for peptides based on their pI, and the same can in principle also be used for SCX chromatography. Indeed, the use of pI information to decrease the false discovery rate for IEF fractionated samples has been already demonstrated by other groups.[21, 22] As part of the work presented here we also developed a general data analysis method and implemented this in a Taverna scientific workflow. The workflow compares multiple fractionation methods with respect to peptide and protein coverage while also extracting additional information on the peptides and proteins from each fractionation method. This information can be used for validation of peptide identifications and detection of splicing or post-translational events. We used this workflow to perform and visualize the comparison between the three different fractionation techniques for the two different types of

samples, and briefly discuss the applicability of each method for each type of sample.

MATERIALS AND METHODS

For this study we compared three different separation approaches for two types of samples (human plasma and *Escherichia coli*). Both groups of samples were treated similarly to enable comparison between methods to determine their suitability for different kinds of samples.

Sample preparation

Human plasma from healthy volunteers was collected into BD Vacutainer® tubes with 18.0 mg K_2:EDTA (K2E, REF 367525, BD Vacutainer Systems, Plymouth, UK) and immediately spun down at 1,300×*g* for 10 minutes at 21˚C then aliquoted and stored at -80˚C until use.

Escherichia coli K12 strain MG1655 (ATCC® Number 47076, ATCC, Manassas, VA); was grown overnight in 4×25 mL Luria-Bertani (LB) medium in 50 mL Falcon tubes. The optical density at 600 nm (OD600) was 2.1. Then all cells were spun down and the supernatant removed. The pellets were resuspended in 10 mL warm (37°C) PBS to pool all cells and gently spun down at 194×*g* at 37°C for 5 min. After the supernatant was removed, all pellets were rinsed with 1ml PBS, transferred to a 1.5-mL Eppendorf tube and spun down again for 10 min at maximum speed (16,100×*g*) at 4°C. The wet pellet was weighed and 5 mL of the BugBuster® Master Mix (Novagen, Merck KGaA, Darmstadt, Germany) was added per gram cell paste. Cells were incubated at room temperature on a shaking platform at low speed for 20 min. After the insoluble cell debris was removed by centrifugation at 16,100×*g* for 20 min at 4°C, the supernatant was stored at -80°C until used.

In-solution digestion

Two mg of each sample were digested using trypsin. To each sample DTT in 25 mM ammonium bicarbonate (ABC) was added to its final concentration 10 mM and incubated for 45 min at 56°C to reduce cystines. After alkylation for 1h at room temperature with 25 mM iodoacetomide also in 25 mM ABC trypsin (sequencing grade, Promega, Madison, WI) was added in the ratio 1:100 (trypsin:sample) and kept for 10 h at 37°C. Digestion was quenched with 10% TFA with the final concentration of TFA 0.1-1.0%. Resulting samples were desalted using Oasis HLB cartridges and aliquoted in 100 and 200 µg for IEF and SCX respectively.

Desalting and solid phase extraction

Prior to fractionation both samples were desalted using Oasis HLB cartridges (Waters, Milford, MA). Cartridges were first activated with methanol and equilibrated with 50% acetonitrile (ACN) in water according to the manufacturer's protocol. The sample was applied and washed 4 times with 500 µL water. The peptides were eluted into a fresh Eppendorf tube with 800 µL 50% ACN.

Fractions collected after the separation were desalted with solid-phase extraction (SPE) using C18 OMIX tips (Agilent Technologies, Waldbronn, Germany). Tips were first wetted with 50% ACN in water, washed and equilibrated with water containing 0.1% TFA. Samples were acidified with TFA, washed again and eluted with 50 µL 50% aqueous ACN containing 0.1% TFA.Acetonitrile was evaporated after each cleaning step.

Strong cation exchange

SCX was performed on a Dionex UltiMate 3000 (Thermo Fischer Scientific, Waltham, MA) at a flow rate of 200 µL/min. Tryptic peptides (200 µg) were loaded onto a 100×2.1 mm PolySULFOETHYL A™ (PolyLC, Columbia, MD) column with 3 µm packing material and eluted with a linear gradient using ACN/potassim phosphate buffers (buffer A – 20% ACN /80% 10 mM potassium phosphate, pH 2.9; buffer B – 20% ACN /80% 10 mM potassium phosphate, 500 mM potassium chloride, pH 2.9). The elution program was 100% buffer A for 10 min, continued by a short (1 min) gradient of 0 to 3% of buffer B, followed by a gradient of 3%-15% for 19 min, a 15%-45% gradient for 15 min and a 45%-100% gradient for 2 min. At the end of the gradient the column was kept at 100% buffer B for 7

min and then for 10 min in buffer A. Flow-through fractions (48 in total) were collected into a 96-well plate from 5 to 55 min. Adjacent fractions were combined pairwise to obtain 24 fractions and then desalted with SPE (described above).

Isoelectric focusing

For peptide IEF separations, the Off-Gel Agilent 3100 fractionator (Agilent Technologies) was used. A modified method was applied by addition of 1 M urea to the buffer sample and rehydration buffer, instead of 5% glycerol only. Tryptically digested and desalted peptides (100 µg in total) were resuspended in a modified IPG buffer that contained 1M urea in addition to the 3–10 pH linear IPG buffer (GE Healthcare, Uppsala, Sweden). Sample volumes of 150 µL/well were loaded onto a commercially available 24-cm IPG strips with a linear 3-10 pH gradient (GE Healthcare) after rehydration of the gel for 20 min in 40 µL/well rehydration solution. Cover fluid (mineral oil, Agilent Technologies) was applied to both ends of the gel strip. The focusing method OG24PE01, as supplied by the manufacturer, was used for 24-well fractionations. Fractions were recovered in separate Eppendorf tubes, cleaned by SPE as described above and store at -80°C till use.

SDS-PAGE and in-gel digestion

Protein concentration was measured by the bicinchoninic acid (BCA) protein assay kit (Thermo Fischer Scientific) and 30 µg of proteins per sample was loaded on a 1-mm 10-well 4-12% NuPAGE® Bis-Tris gel (Invitrogen, Carlsbad, CA). Proteins were separated in the gel for 1 h at 180 V, after which the gel was stained in NuPAGE® Colloidal Blue (Invitrogen) overnight at room temperature and destained with milli-Q water until the background was transparent.

The gel lane with separated proteins was cut into 48 identical 1.5×5-mm slices using a MEE1.5-5-48 disposable gel cutter (Gel Company Inc., San Francisco, CA). Each gel piece was placed into one well in a 96-well polypropylene PCR plate (Greiner Bio-One, Frickenhausen Germany). Destaining of the gel pieces, DTT reduction and IAA alkylation were performed according the previously published protocol.[23] In-gel tryptic digestion was performed in 30 µL of 25 mM ABC containing 5 ng/µL

trypsin (sequencing grade, Promega, Madison, WI) for 6 h at 37°C. The resulting peptides were TFA-extracted according to the previously described protocol.[23] The extracts were pooled pairwise to obtain 24 total fractions as for SCX.

LC-MS/MS analysis

Prior to LC-MS/MS analysis all samples were dried down and reconstituted in 25 µL 0.1% TFA. The analysis was performed using a splitless NanoLC-Ultra 2D plus (Eksigent, Dublin, CA) for parallel ultra-high pressure liquid chromatography (UHPLC) with an additional loading pump for fast sample loading and desalting. The UHPLC system was configured with 300 µm-i.d. 5-mm PepMap C18 trap columns (Thermo Fischer Scientific) and 15-cm 300 µm-i.d. ChromXP C18 columns (Eksigent). Peptides were separated by a 45-minute linear gradient from 4 to 33% acetonitrile in 0.05% formic acid with 4 µL/min flow rate. The UHPLC system was coupled on-line to an amaZon ETD speed high-capacity 3D ion trap with CaptiveSpray source (Bruker Daltonics, Bremen, Germany). After each MS scan, up to ten abundant multiply charged species in the *m/z* 300-1300 range were automatically selected for MS/MS but excluded for one minute after having been selected twice. The UHPLC system was controlled using HyStar 3.4 with a plug-in from Eksigent and the amaZon ion trap by trapControl 7.0, all from Bruker.

Data analysis

All acquired tandem mass spectrometry data were processed in one batch using the Taverna workbench.[24] Taverna can invoke a number of services, including local Java Beanshell scripts, R (using an R server) and a wide range of Web services, enabling combination of sequence database search, analysis and visualization in a single workflow. Built-in tools for parsing XML- files simplify information retrieval andlarge datasets can be remotely processed on a grid or cloud using the Taverna Engine.[25] The workflow used here converts raw data to mzXML[26] using compassXport 3.0 (Bruker) and passes this, along with the sequence database and search parameters to X!Tandem[20, 27] in the TPP.[20] The X!Tandem scores are converted to pepXML,[20] modelled and converted to probabilities for each peptide-spectrum match by PeptideProphet.[28] The X!Tandem search was here done against the UniProt human reference proteome set (2012_02, canonical sequences only) and the UniProt *Escherichia coli* reference set (2010_01) with the monoisotopic mass error (±0.5 Da), carbamidomethylation as fixed

modification, the k-score plug-in[20] and allowing for isotope error. After PeptideProphet analysis, the resulting lists of peptide/protein identifications with 0.95 probability cut-off (<1% FDR) were analyzed and compared in the Rshell script in the same workflow. For each peptide within one IEF fraction, pI values predicted by attached function in TPP (based on pK values from Bjellqvist *et al.*[29]) were extracted and the pI Z-scores were calculated as a distance in standard deviation from the mean. To compare the pI of true and false matches, a search was also done against a decoy database generated by randomizing the *E. coli* database with *make_random* (http://www.ms-utils.org/make_random.html). For SDS-PAGE, the protein molecular weight was calculated from the sequences downloaded from the UniProt website directly in Taverna workflow as these are not kept in the pepXML results. The entire processing workflow is available in myExperiment (http://www.myexperiment.org/workflows/3486.html).

RESULTS

In this work we compared SDS-PAGE, SCX and IEF separation strategies for two different types of samples. For both samples the highest proteome coverage we observed with SCX (Figure 1.1) identifying 1,645 peptides in plasma and 6,731 peptides in *E. coli*. While the number of protein identifications for the *E. coli* sample, was approximately the same with the three methods, the number of identified peptides varied from 4,221 for IEF to 6,231 and 6,731 for SDS-PAGE and SCX respectively. For the plasma sample, SCX was clearly better compared to 1,015 peptides identified with SDS-PAGE and 831 with IEF. In the recent work of Hassan *et al*. [30] SCX was also demonstrated to be better than IEF, as measured by the number of identified peptides. When comparing the number of identified proteins, SDS-PAGE gave the lowest coverage for plasma, similarly to a previous comparison using HeLa cells.[31]

To define the quality of the separation we looked on the distribution of the number of peptides identified per fraction (Figure 1.2). When separated with SCX, most peptides were found in one fraction. For IEF, the majority of peptides is still determined only in one fraction, however the number of

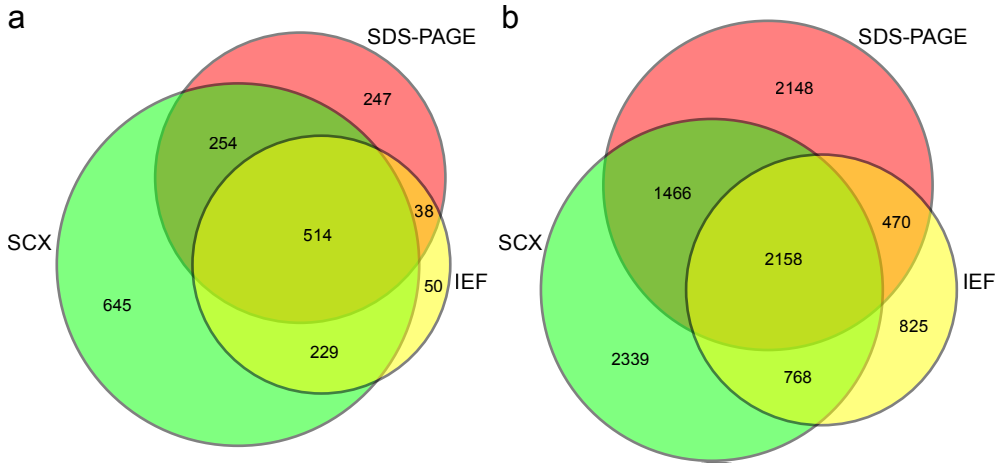


Figure 1.1. Comparison of the number of unique peptide identifications from SDS-PAGE, SCX and IEF datasets for human plasma (a) and *E. coli* (b).

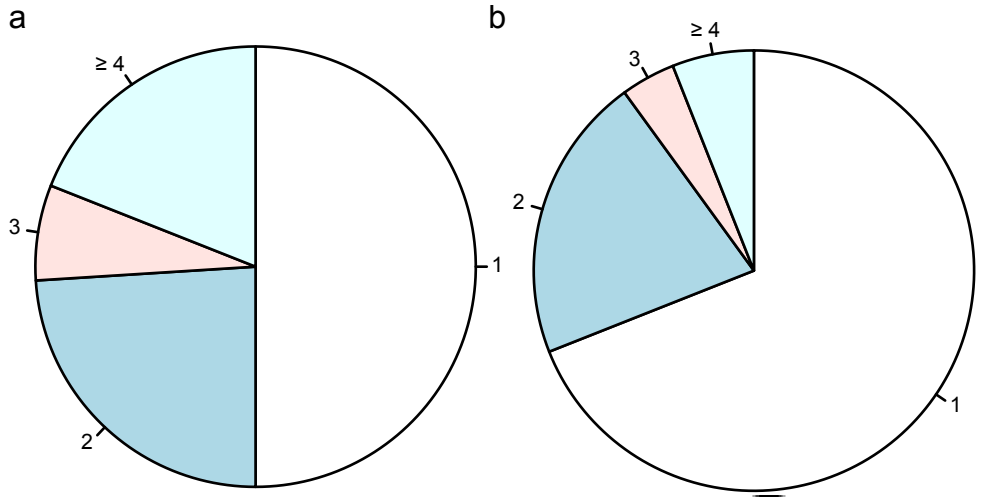


Figure 1.2. Pie charts illustrating the percentage of peptides identified in one or more fractions after separation of human plasma by IEF (a) or SCX (b). During the SCX chromatography 48 fractions from 65 min gradient were collected and every two consecutive ones pooled together.

peptides found in two and more fractions is much higher compared to those for SCX. Presented in Figure 1.2 pie charts illurate the peptide distribution for human plasma sample. For *E. coli* the observation is consistent (data not presented).

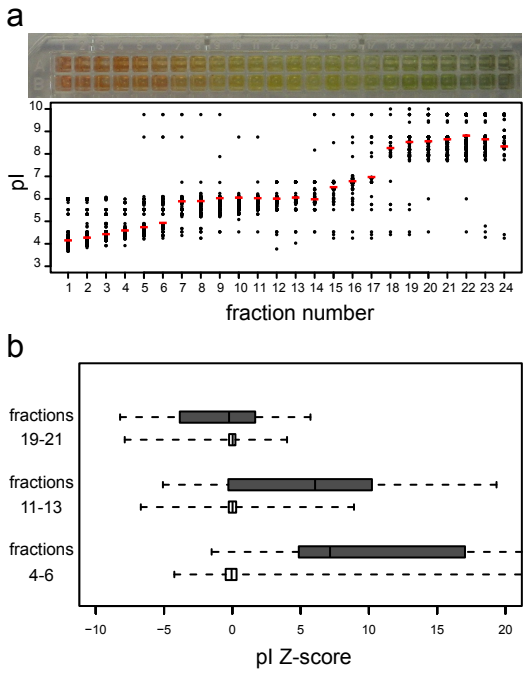


Figure 1.3. Fractionation of human plasma and *E. coli* peptides by isoelectric focusing. The photograph shows the pH indicator from pH ~3 to pH ~10 in the fractions of plasma (top) and *E. coli* (bottom). The pH gradient appears reproducible and independent from the sample. The fraction yielding the largest number of spectrum matches for a peptide can be plotted against the predicted pI for the peptide (here showing only the IEF fractions in *E. coli*). The mean pI of the peptides in each fraction is marked with red bars. The pI information can be used to weed out false identifications. The box plot (b) illustrates the distribution of pI Z-scores with putatively correct matches in white and decoy matches in grey.

A further motivation behind this study was to produce, from the same samples, similar datasets using the three different peptide and protein fractionation techniques to illustrate the value of the additional information on the analytes that can be automatically obtained from a particular method. Using a pH indicator, we observed that the peptides in IEF separate more or less linearly in the pH gradient independent of the nature of the sample (Figure 1.3a, top). Thus the calculated pI can be plotted against the actual fraction number and eventual outliers would most likely be false identifications (Figure 1.3a, bottom). Predicted pI appear to change in more discrete steps compared to the smooth transitions of the pI indicator. Another way to represent this information is to calculate pI Z-score and visualize their distribution for each fraction separately using histogram or box-plot (Figure 1.3b). The decoys have a wide distribution in Z-score (the unit determined by the standard deviation in predicted pI of the matches from the correct database) and as expected with bias towards higher pI for fractions of low pI and towards lower pI for fractions of high pI, whereas the correct identifications are focused near the average pI of all peptides identified in the fraction.

SDS-PAGE, on the other hand, provides direct information about the proteins rather than the peptides. Predicted, based on the sequence, protein molecular weight plotted against its location on gel (fraction number) shows

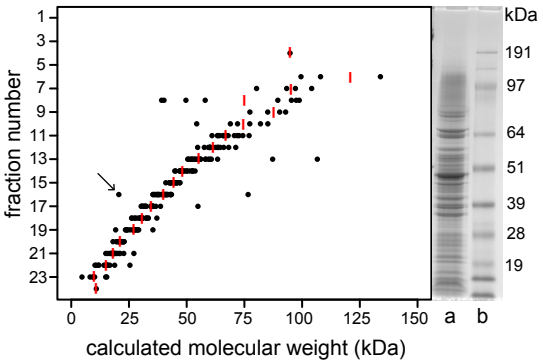


Figure 1.4. Protein molecular weight distribution against fraction number in the *E. coli* sample (left). Median molecular weight per fraction is marked in red. SDS-PAGE of 30 µg of *E. coli* sample (a) and molecular weight marker (b) on the right. Alkyl hydroperoxide reductase subunit C (UniProt accession number P0AE08) is marked with an arrow.

a clear correlation (Figure 1.4). However, a number of outliers can still be identified for closer examination or discarding as false discoveries. As an example, the 20 kDa Alkyl hydroperoxide reductase subunit C (UniProt accession number P0AE08), in native conditions disulfide-linked homodimer, was observed at ~40 kDa (Figure 1.4, arrow).

DISCUSSION

The wide range of available fractionation techniques makes it challenging to choose the one best suited for a particular sample or biological research question. We performed in this work comparison of the described above techniques at the level of proteins (SDS-PAGE) and peptides (IEF and SCX). The major challenge in setting up such a study is to make the comparison "fair", given the differences in scale and practical implementation of the techniques, *i.e.* sensitivity levels, system volumes/flow rates and fraction collection. It is especially difficult to use the same amount of starting material for each method without diluting the sample or overloading one or more of the systems. In the case of SDS-PAGE, the maximum amount of protein that can be applied on the standard, commercially available, gel without overloading is around 30 µg. For the IEF and SCX methods, the equivalent amount of peptides would be too low due to the minimal volumes involved. The work of Hubner *et al.* [31] demonstrated that the best separation with IEF could be achieved with 50 µg material, while the maximum number of protein identifications was achieved with 250 µg. From our experience, the optimal condition for Off-Gel IEF (balance between good separation and wide proteome coverage)

has been obtained when loading 100 µg. Even though SCX gave reasonable separation when loaded 100 µg of material, the system was far from its maximum loading capacity, leading us to increase the amount of proteins injected to limit sample dilution. For this reason we compromised and loaded different amounts to allow each fractionation technique to operate near its maximum capacity, taking into account the significant dilution in the IEF and SCX as compared to SDS-PAGE. Robustness and stability of the liquid-chromatography-mass spectrometry analysis is also important for the method comparisons in absence of internal standards or labels. To balance sensitivity and robustness, the choice was made to use the new CaptiveSpray (Bruker) source, accommodating higher flow rates and therefore more robust chromatography than the more sensitive but less stable nanoelectrospray.

One would expect SDS-PAGE to be a good choice for samples dominated by a small number of abundant proteins, such as plasma, as these abundant proteins can be confined to a few bands or fractions. In contrast, when performing the fractionation at the peptide level, peptides from the abundant proteins will be present in most if not all fractions. However, in this comparison, we demonstrated that SCX was clearly better in both peptide and protein yield. This proves that the loading capacity can be more important than the separation method or whether the fractionation is done at the protein or peptide level. The IEF approach gave the smallest number of identifications and showed the largest overlap with the other two techniques for both samples. Even though the work of Hubner *et al.* [31] showed that Off-Gel IEF gives higher number of protein identifications compared to SDS-PAGE in human cell lines, other recent work comparing SDS-PAGE, SCX, IEF and organelle fractionation have showed the opposite.[32] For the IEF system it is known that near the edges of the gel (at pH 3 and pH 10) it is common to see diffuse bands if the gel is stained, indicating less sharp separation. Consequently, peptides can be found in more than one fraction near the edges, increasing the redundancy of the data and reducing the number of different peptides that can be identified. During the SCX chromatography, fractions were collected every 60 seconds and subsequent fractions were pooled for the analysis to keep the number of fractions similar to those obtained with IEF and mass spectrometry analysis time constant throughout the whole experiment. The number of collected fractions with SCX is determined by the fraction collection method, which can easily be adjusted to any number, as long as the vials or wells can hold

the volume and there are enough physical vials or wells in the fraction collector. Similarly for SDS-PAGE, any reasonable number of pieces can be cut, as long as the slices are not too thin to handle in a practical manner. Generally, more fractions lead to wider proteome coverage if the mass spectrometry time per fraction is constant. As the numbers of SCX fractions and gel slices are easy to vary, the defining factor for the number of collected fractions was the IEF system. For SDS-PAGE, we used an already existing and commercially available cutter enabling slicing the gel into 48 equal slices at once. Similarly, we used an existing method for collecting 48 SCX fraction and then pool the adjacent ones to obtain 24 total fractions for each separation method. Most peptides identified with SCX were found in a single fraction, showing that peptides elute in narrow peaks (maximum 2 minutes in a 65-minute gradient). Compared to the studies conducted by Slebos *et al.* [17] and Elschenbroich *et al.* [33] demonstrating that IEF is superior to SCX in resolution, we used longer and better analytical column for SCX, with smaller bead size. Not surprisingly, in our experimental set-up, SCX had better resolution than IEF, defined as peptide overlap between fractions. The fractionation settings and the design of the comparison have more influence on the result than the nature of the sample.

For any scheme that uses information from the fractionation prior to the chromatographic separation on-line with the tandem mass spectrometer it is crucial that this information is preserved throughout the data analysis. This is easily accomplished by a systematic naming of files or by loading fractions in sequence into a microtiter plate. From each dataset, specific protein or peptide information could be extracted and used for filtering out spurious identifications. The theoretical model used for pI calculation is based on the peptide sequence and does not take influences of nearby residues into account, leading to a discrete rather than smooth distribution of pI in the IEF-separated samples. However, this information is used in the calculation of pI Z-scores for each peptide-spectrum match, assuming they derive from a fraction with a narrow pI distribution, and is already implemented in the TPP. Random, false (decoy) peptide-spectrum-matches can derive from peptides of any pI therefore having a wide span, whereas the correct identifications are concentrated around 0. For a perfect Gaussian distribution, the lower and upper quartiles, *i.e*. the “box”, would be between Z-score -0.68 and 0.68. In the pI box plots in Figure 1.3b, the lower and upper quartiles of the putatively correct peptide identifications span a slightly smaller interval. This is likely due to a number of outliers caused by

very abundant peptides being identified in many fractions and differences between calculated and real (experimental) pI.

Although some success has been reported in the prediction of peptide retention times in SCX[34, 35], this has so far only been achieved with machine-learning techniques such as artificial neural networks, requiring tens or hundreds of thousands of peptide identifications to train the model. This makes the approach feasible only when very large collections of datasets are available. A simpler model could be plugged into the workflow as available on myExperiment. Since both SCX and IEF are primarily based on charge (SCX on the charge at a particular pH) it may be tempting to use a similar model for SCX prediction as for pI prediction in IEF. However, for the datasets used in this work, this did not produce a useful model.

Protein information derived from SDS-PAGE can indirectly indicate whether peptide identifications correspond to a protein that is likely to be present in the fraction from which the spectrum was acquired. However, as there are many reasons why the calculated and measured protein molecular weights may differ significantly, it is probably more sensible to use the protein level information to learn something about the proteins. Proteins located far above a curve fitted to the predicted molecular weights are larger than predicted (Figure 1.4), which might be due to an incomplete sequence in the database, a large post-translational modification or a covalent protein complex. Hits below the curve indicate that the observed protein is only part of the predicted (database) protein sequence. If both explanations are implausible and the number of confident peptide-spectrum matches for a protein is small (given the total number of spectra acquired), the protein identification is likely incorrect. This assumption is supported by the relatively low probabilities for the peptide-spectrum matches for these proteins. In prokaryotic organisms, there is little post-transcriptional activity, such as splicing, that leads to multiple protein isoforms from the same gene or entry in the searched FASTA file. There are also fewer post-translational events decorating proteins with adducts large enough to be noticeable by SDS-PAGE. Therefore, the outliers are most likely false identifications, and their number is very small compared to those in eukaryotic samples. A few exceptions, such as covalent complexes, can still be identified though, as shown in Figure 1.4. Using the SDS-PAGE information, false positives can be weeded out at the protein- rather than the peptide level, without influencing the probabilities assigned to the peptide-spectrum matches.

CONCLUSIONS

In shotgun proteomics, good coverage of complex samples still requires more than one dimension of fractionation or separation. However, not all separation methods are equally suitable for all types of samples and research questions. Here we compared three of the most commonly used techniques, SDS-PAGE, SCX and IEF, for two different and "typical" samples. The fractionation methods are based on different physicochemical properties and were performed at different levels – at the protein level with SDS-PAGE and at the peptide level with SCX and IEF. When comparing such different separation techniques, it is difficult to make a "fair" comparison. We kept the final number of fractions collected equal and the total mass spectrometry analysis time constant, but decided to compromise on the amount of protein used, performing the fractionation near the optimal conditions/highest capacity of each method. The number of collected fractions and MS instrument time were kept the same for the comparisons, even though the SDS-PAGE and SCX would likely have performed better if more fractions had been collected. Under the studied conditions, IEF showed the lowest coverage for both samples, which may be partly due to the dilution occurring during the run but also to suboptimal number of fractions in IEF. The extracted pI information gives an easily implemented method to filter out false peptide-spectrum matches. The SDS-PAGE approach resulted in better coverage of the proteome, while also providing molecular weight information on the proteins. We compromised the resolving power of the gel by pooling consecutive pairs of gel slices to keep the total number of fractions the same as for the IEF. There is no strict reason to believe that combining adjacent fractions is the most optimal way to reduce the number of fractions. By pooling two neighbouring fractions where most likely similar proteins are dominant, there will be suppression of the less represented ones. It is possible that it would be better to combine gel slices containing large and small proteins, even though the results would be more difficult to interpret manually.

Strong cation exchange provided the best coverage of both peptides (especially for *E. coli*) and proteins (particularly for plasma). The information of SCX retention times which could be used to improve sensitivity and lower the false discovery rate was not implemented. Although SCX is a very efficient separation technique for peptides and orthogonal to reversed-phase, it is most likely that it was the larger amount of sample that could be loaded on the SCX column, compared to the SDS-

PAGE and IEF that contributed the most to the higher number of identifications.

The data analysis, from raw data to the graphs as they appear in the paper, could be performed entirely within one Taverna workflow, facilitating sharing not only raw data but also executable workflows. This allows other researchers to reproduce the analysis while varying input parameters or apply the same analysis workflow on their own data. Additionally, separate components of the workflow could be reused in different analyses or adopted for other tasks. The workflow executed local commands and took the data through the Trans-Proteomic Pipeline interfacing data analysis of three separate datasets in parallel using one parameter and one FASTA files piped to different processes assuring exactly the same conditions for each search. This workflow also fetched information from on-line databases, performed statistical analyses in R and plotted the results.

ACKNOWLEDGEMENTS

The authors wish to thank Hans Dalebout and Oleg Klychnikov for technical assistance and useful discussions, and Yassene Mohammed for help with the Taverna workflows.

REFERENCES

1. Choudhary, G.; Horvath, C., Ion-exchange chromatography. *Methods Enzymol* **1996,** 270, 47-82.
2. Howard, G. A.; Martin, A. J. P., The Separation of the C-12-C-18 Fatty Acids by Reversed-Phase Partition Chromatograhy. *Biochemical Journal* **1950,** 46, (5), 532-538.
3. Hjerten, S., Some General Aspects of Hydrophobic Interaction Chromatography. *Journal of Chromatography* **1973,** 87, (2), 325-331.
4. Lathe, G. H.; Ruthven, C. R., The separation of substances on the basis of their molecular weights, using columns of starch and water. *Biochem J* **1955,** 60, (4), xxxiv.
5. Porath, J.; Flodin, P., Gel filtration: a method for desalting and group separation. *Nature* **1959,** 183, (4676), 1657-9.
6. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd, Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999,** 17, (7), 676-82.
7. Shapiro, A. L.; Vinuela, E.; Maizel, J. V., Jr., Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochem Biophys Res Commun* **1967,** 28, (5), 815-20.
8. Chen, J.; Lee, C. S.; Shen, Y.; Smith, R. D.; Baehrecke, E. H., Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation. *Electrophoresis* **2002,** 23, (18), 3143-8.
9. Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S., Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. *Anal Chem* **2003,** 75, (13), 3145-52.
10. Hjerten, S., Free zone electrophoresis. *Chromatogr Rev* **1967,** 9, (2), 122-219.
11. Margolis, J.; Corthals, G.; Horvath, Z. S., Preparative reflux electrophoresis. *Electrophoresis* **1995,** 16, (1), 98-100.
12. Horvath, Z. S.; Gooley, A. A.; Wrigley, C. W.; Margolis, J.; Williams, K. L., Preparative affinity membrane electrophoresis. *Electrophoresis* **1996,** 17, (1), 224-6.
13. Michel, P. E.; Reymond, F.; Arnaud, I. L.; Josserand, J.; Girault, H. H.; Rossier, J. S., Protein fractionation in a multicompartment device using Off-Gel (TM) isoelectric focusing. *Electrophoresis* **2003,** 24, (1-2), 3-11.
14. Xiao, Z.; Conrads, T. P.; Lucas, D. A.; Janini, G. M.; Schaefer, C. F.; Buetow, K. H.; Issaq, H. J.; Veenstra, T. D., Direct ampholyte-free liquid-phase isoelectric peptide focusing: application to the human serum proteome. *Electrophoresis* **2004,** 25, (1), 128-33.
15. Gan, C. S.; Reardon, K. F.; Wright, P. C., Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of Synechocystis sp. PCC 6803. *Proteomics* **2005,** 5, (9), 2468-78.
16. Essader, A. S.; Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr., A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in

shotgun proteomics. *Proteomics* **2005,** 5, (1), 24-34.

17. Slebos, R. J.; Brock, J. W.; Winters, N. F.; Stuart, S. R.; Martinez, M. A.; Li, M.; Chambers, M. C.; Zimmerman, L. J.; Ham, A. J.; Tabb, D. L.; Liebler, D. C., Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res* **2008,** 7, (12), 5286-94.
18. Putnam, F. W., Alpha, beta, gamma, omega - the roster of the plasma proteins. In *The Plasma Proteins*, 2 ed.; Putnam, F. W., Ed. Academic Press: New York, 1975; Vol. 1, pp 57-130.
19. Peng, J. M.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *Journal of Proteome Research* **2003,** 2, (1), 43-50.
20. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **2005,** 1, 1-8.
21. Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *Journal of Proteome Research* **2004,** 3, (1), 112-119.
22. Krijgsveld, J.; Gauci, S.; Dormeyer, W.; Heck, A. J., In-gel isoelectric focusing of peptides as a tool for improved protein identification. *J Proteome Res* **2006,** 5, (7), 1721-30.
23. Mostovenko, E.; Deelder, A. M.; Palmblad, M., Protein expression dynamics during Escherichia Coli glucose-lactose diauxie. *BMC Microbiol* **2011,** 11, 126.
24. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P., Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004,** 20, (17), 3045-3054.
25. Mohammed, Y.; Mostovenko, E.; Henneman, A. A.; Marissen, R. J.; Deelder, A. M.; Palmblad, M., Cloud Parallel Processing of Tandem Mass Spectrometry Based Proteomics Data. *Journal of Proteome Research* **2012,** 11, (10), 5101-5108.
26. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004,** 22, (11), 1459-66.
27. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.
28. 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* **2002,** 74, (20), 5383-92.
29. Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D., The Focusing Positions of Polypeptides in

Immobilized Ph Gradients Can Be Predicted from Their Amino-Acid-Sequences. *Electrophoresis* **1993,** 14, (10), 1023-1031.

30. Hassan, C.; Kester, M. G.; Ru, A. H.; Hombrink, P.; Drijfhout, J. W.; Nijveen, H.; Leunissen, J. A.; Heemskerk, M. H.; Falkenburg, J. H.; Veelen, P. A., The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol Cell Proteomics* **2013**.

31. 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.

32. Antberg, L.; Cifani, P.; Sandin, M.; Levander, F.; James, P., Critical Comparison of Multidimensional Separation Methods for Increasing Protein Expression Coverage. *Journal of Proteome Research* **2012,** 11, (5), 2644-2652.

33. Elschenbroich, S.; Ignatchenko, V.; Sharma, P.; Schmitt-Ulms, G.; Gramolini, A. O.; Kislinger, T., Peptide Separations by On-Line MudPIT Compared to Isoelectric Focusing in an Off-Gel Format: Application to a Membrane-Enriched Fraction from C2C12 Mouse Skeletal Muscle Cells. *Journal of Proteome Research* **2009,** 8, (10), 4860-4869.

34. Alpert, A. J.; Petritis, K.; Kangas, L.; Smith, R. D.; Mechtler, K.; Mitulovic, G.; Mohammed, S.; Heck, A. J., Peptide orientation affects selectivity in ion-exchange chromatography. *Anal Chem* **2010,** 82, (12), 5253-9.

35. Petritis, K.; Kangas, L.; Jaitly, N.; Monroe, M. E.; Lopez-Ferrer, D.; Maxwell, R. A.; Mayampurath, A.; Petritis, B. O.; Mottaz, H. M.; Lipton, M. S.; Camp, D. G.; Smith, R. D. In *Strong cation exchange LC peptide retention time prediction and its application in proteomics*, American Society for Mass Spectrometry (ASMS), Denver, CO, USA, 2008; Denver, CO, USA, 2008.

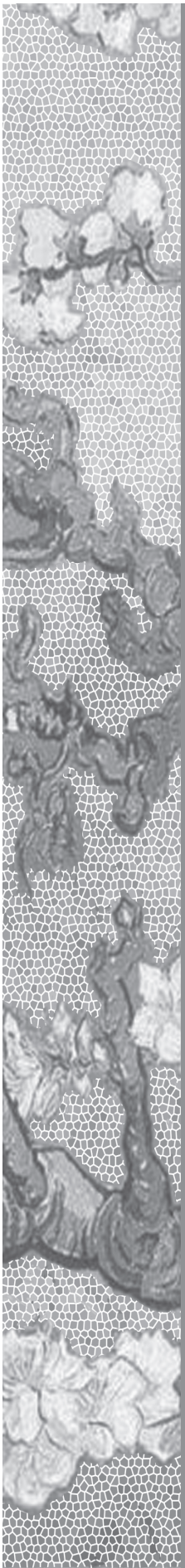
2

Protein Fractionation for Quantitative Plasma Proteomics by Semi-Selective Precipitation

Ekaterina Mostovenko, Hannah C. Scott, Oleg Klychnikov, Hans Dalebout, André M. Deelder and Magnus Palmblad

Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

Journal of Proteomics and Bioinformatics, 2012, 5, 217-221



ABSTRACT

Blood plasma is a highly complex mixture of proteins, metabolites and lipids, and a rich source of potential biomarkers for a range of diseases and conditions. The wide range in protein abundance poses a tremendous challenge for plasma proteomics. However, as a relatively small number of proteins makes up most of the total protein pool, the concentration range can be compressed by depletion of abundant proteins, such as albumin. Although many commercial solutions are available for depleting one or more abundant proteins, general enrichment of low-abundant proteins or specific enrichment of selected peptides after enzymatic digestion, none of these solutions is simultaneously robust, high-throughput, inexpensive and suitable for label-free analysis. We have explored a method for binary partition with partial depletion of albumin for quantitative plasma proteomics based on semi-selective precipitation with acetonitrile at different pHs. The method is simple, reproducible and easily parallelised (high throughput), and may be well suited to fractionate plasma proteins for label-free quantitative proteomics.

INTRODUCTION

Plasma contains carbohydrates, lipids, salts, vitamins, amino acids, nucleic acids, hormones and around 75 mg/mL protein.[1] The carrier-protein albumin dominates with 45-50% of the total protein concentration, while immunoglobulin G and transferrin contribute 8-20% and 3-7%, respectively.[2] These and other highly abundant, large proteins mask less abundant ones by decreasing their relative concentration and through effects such as ion suppression in electrospray ionization mass spectrometry. Although changes in the abundant proteins may also be indicative of the physiological status of the organism,[3] low-abundant proteins, for instance from tissue leakage, may mark an early state of a disease such as cancer.[4, 5] Although plasma is easily sampled, the concentration range of proteins, spanning from picogram to microgram per millilitre, is a major challenge in clinical proteomics.

Numerous techniques have been suggested and employed to reduce the complexity of the plasma proteome, including depletion of abundant proteins,[6] nonspecific enrichment of low-abundant proteins via combinatorial peptide libraries[7] and specific enrichment of targeted peptides after enzymatic digestion.[8] Complexity reduction can be performed by classical methods such as centrifugation or extraction with organic solvents[9] or by immunodepletion[10]. A range of depletion columns, spin cartridges and affinity capture beads for removal of albumin, IgG[11] and many other abundant proteins are commercially available. Several of these commercial kits have previously been compared by Chromy *et al*.[12] and Björhall *et al*.[13] for their utility in plasma proteomics. Immunoaffinity is efficient in depleting selected, abundant proteins, but in significantly reducing the concentration range of proteins in plasma, many different antibodies are needed. As the immunoaffinity depletion is carried out under native conditions, other, less abundant proteins may still be bound to one of the abundant proteins being depleted, for instance albumin in plasma. Typically, commercial affinity columns use avian IgY antibodies against the most abundant ("top") plasma proteins and remove from 50% (anti-albumin only) to 99% (top-20) of total plasma protein. In theory, assuming a 100% recovery, low abundance proteins would then be enriched by a factor 2 to 100, respectively. However, both reproducibly manufacturing and applying columns with a large number of different antibodies is not trivial. For instance, we have previously observed a significant column-to-column variation in commercial affinity depletion columns (unpublished results). Although this may not be a serious problem in a general exploration of the

plasma proteome or in studies where proteins have been isotopically (or otherwise) labelled prior to the depletion/enrichment step, poor reproducibility obviously poses a serious problem for label-free studies.

Many of the abundant proteins in plasma have molecular weights exceeding 60 kDa (*e.g.* albumin, transferrin, fibrinogen, IgA, α-2-antitrypsin, apolipoproteins, and acid-1-glycoprotein). A simple and semi-selective depletion of many of these large and highly abundant plasma proteins is possible by precipitation using organic solvents such as acetonitrile and this has indeed been demonstrated in plasma and serum from several species with reproducible results.[14-18] This procedure results in a separation, wherein most of the more soluble low molecular weight proteins are left in the supernatant and the larger proteins precipitate. Acetonitrile has also been shown to release albumin-bound proteins, which could be potential biomarkers.[5] Protein solubility is also affected by pH, ionic strength and temperature,[19] and by adjusting one or more of these parameters, the precipitation may be optimized to efficiently remove as much of the abundant proteins, such as albumin, as possible in a single step, while maintaining low-abundant proteins in solution. Alternatively, several precipitation steps can be combined for a more efficient depletion of abundant proteins and increased recovery of low-abundant proteins. Semi-selective precipitation may also be tuned to partition the proteome in two or more complementary fractions with limited overlaps for increased combined coverage of the proteome. In this work we focused on the effect of pH on the plasma depletion by acetonitrile and its suitability for clinical applications. Such a simple precipitation procedure is attractive for large scale studies as they are inexpensive, scalable, easy to parallelize, potentially robust and reproducible and not dependent on expensive affinity separations with concomitant batch-to-batch or column-to-column variation, problematic for label-free methods.

MATERIALS AND METHODS

Sample preparation and organic precipitation

Human plasma from healthy volunteers was collected into BD Vacutainer® tubes with 18.0 mg K_2 EDTA (K2E, REF 367525, BD Vacutainer Systems, Plymouth, UK) and immediately spun down at 1,300×g for 10 minutes at 21°C, and 50 µl aliquotes were stored at -80°C until use. Samples were

thawed at 4℃ and then centrifuged at 16,100×g at 4℃ for 1 minute. The pH was adjusted in three identical aliquots to 5.0, 7.0 and 9.0 by adding acetic acid and ammonium hydroxide directly to the sample. Three other aliquots were diluted 1:10 (v:v) with 100 mM ammonium acetate buffer with corresponding pH's to investigate the effect of protein concentration. For protein precipitation, acetonitrile was mixed with the samples in 1:1 (v:v) ratio and the samples were vortexed three times at 1,000 rpm for 5 s, and then incubated for 10 minutes in an ultrasonic bath at room temperature. Vortexing and sonication steps were repeated twice before the samples were centrifuged at 16,100×g at 4℃ for 10 minutes. The supernatants after precipitation were collected in fresh Eppendorf tubes and both the pellets and the supernatants were lyophilized. The precipitates were vigorously vortexed and sonicated in 100 µl BugBuster Master Mix (Novagen, Merck KGaA, Germany) for pellets and 30 µl for supernatants. The pellet precipitates were resuspended in a Bullet Blender (Next Advance Inc., Averill Park, NY) with 0.1 mm glass beads which were then removed by centrifugation through 30 µm pore size micro-spin columns (Thermo Fisher Scientific, Waltham, MA) at the lowest speed. The protein concentration was then defined using a bicinchoninic acid (BCA) protein assay kit (Thermo Fisher Scientific). This protein extraction reagent has been developed for the lysis and protein solubilisation from bacteria, but is routinely used in our laboratory and directly compatible with BCA analysis, SDS-PAGE, tryptic digestion and samples are easily cleaned up for analysis by liquid chromatography-mass spectrometry (LC-MS).

SDS-PAGE and in-solution digestion

Thirty micrograms of protein (BCA) per sample were loaded on a 1-mm 10-well 4-12% NuPAGE® Bis-Tris gel (Invitrogen, Carlsbad, CA). All samples were diluted in 2X NuPAGE® Sample Buffer (Invitrogen). Proteins were separated in the gel for 1 h at 180 V. The gel was stained in NuPAGE® Colloidal Blue (Invitrogen) overnight at room temperature and distained with milli-Q water until the background was transparent.

For in-solution tryptic digestion, 20 µg of each sample was used. The digestion was performed after DTT reduction (10 mM, 56℃ for 45 min) and IAA alkylation (25 mM, 1 h in the dark at room temperature) in 25 mM ABC with protein to trypsin ratio 20:1 for 12 h at 37℃. The reaction was then quenched with 5 µL of 10% TFA. The samples were stored at -35 ℃ until analysis.

Liquid chromatography-mass spectrometry

Peptides derived from all protein digests were separated by splitless parallel reversed phase C18 NanoLC-Ultra 2D plus (Eksigent, Dublin, CA) ultra-high pressure liquid chromatography (PepMap trap columns C18 5-mm, 300 µm-i.d., Dionex Sunnyvale CA; ChromXP analytical C18 columns 15-cm, 300 µm-i.d., Eksigent) with an additional loading pump for fast sample loading and desalting. Samples were analyzed for 120 min using a linear gradient from 4 to 33% acetonitrile in 0.05% formic acid with flow rate 2 µl/min. The MS and MS/MS (CID-only) spectra were recorded on an amaZon ETD high-capacity 3D ion trap with CaptiveSpray source (Bruker Daltonics, Bremen, Germany). The ten most abundant multiply charged species in the *m*/*z* range 300-1300 were automatically selected for MS/MS with one minute dynamic exclusion after having been selected twice.

Data analysis

The complete experiment was analyzed in a single Taverna scientific workflow [20] (Figures 2.1 and 2.2) with all external software installed in their default locations. For each sample the raw LC-MS/MS files were first converted to mzXML [21] using compassXport 3.0.5 (Bruker). The mzXML files were then processed as in the open source Trans-Proteomic Pipeline (TPP) [22] using both the X!Tandem [22, 23] database search engine and the SpectraST spectral library search. With X!Tandem we used the UniProt human reference proteome set (2012-02-05, canonical sequences only), carbamidomethylation as the only and fixed modification, the k-score plug-in [22] and a monoisotopic mass error ±0.5 Da, including also the first and second isotopic peaks. For SpectraST, the NIST human spectral library from 2011-05-26 was searched with default settings except for carbamidomethylation ("CAM") of cysteines. All search results (in pepXML[22]) were analyzed by PeptideProphet, [24] then refined and combined by InterProphet. Peptide-spectrum matches with a PeptideProphet probability $p \geq 0.95$ corresponding to approximately a 1% false discovery rate (FDR) were included in the analysis. For each protein sequence in the FASTA file, a BeanShell component in the comparison workflow (Figure 2.2) calculated molecular weight using average masses of amino acids, GRAVY score (using amino acid hydrophathy information from Kyte and Doolittle [25]) and pI (using pK values from Bjellqvist *et al.* [26]). The protein spectral counts (number of peptide-spectrum matches per protein) in the different fractions were then compared with respect to this information and visualized using an Rshell. Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner

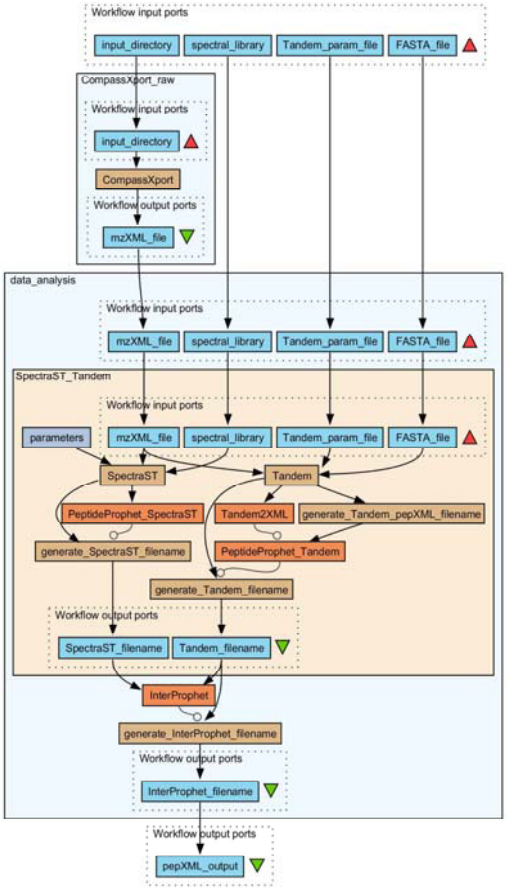
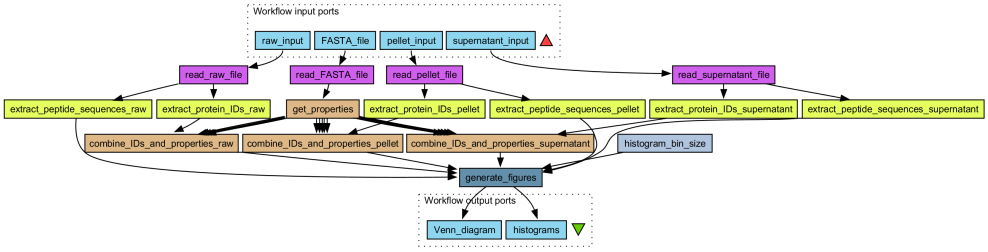


Figure 2.1 (left). Taverna scientific workflow for the proteomics data processing based on the TPP. The raw data is converted to the mzXML format by CompassXport and each file is then separately searched by X!Tandem and SpectraST. Only peptides with PeptideProphet probabilities ≥95% were taken for the further analysis. Separate search results were combined by InterProphet. The workflow allows parallel sample processing.

Figure 2.2 (below). Taverna workflow to produce a Venn diagram and charts of spectral counts as function of protein molecular weight, pI and GRAVY score as seen in Figures 2.4 and 2.5. All inputs are provided from the outputs of the workflow in Figure 2.1 and the two workflows may be combined into a single, complete workflow.



repository [27] with the dataset identifier PXD000042. The full workflow is freely available via www.myExperiment.org ("Plasma Precipitation Analysis").

RESULTS AND DISCUSSION

The method for protein fractionation explored here was designed to partition the proteins in the sample, reducing the relative abundance of the dominating proteins, and if possible simultaneously remove contaminants that might interfere with protein quantitation and biomarker detection in body fluids such as plasma. However, at high protein concentrations, such as in plasma, there is always a high risk of co-precipitating otherwise soluble proteins. Experimentally, we indeed found the preparation of diluted samples to be more robust, less time-consuming and the results were highly reproducible (Figure 2.3). This method therefore could be more easily applied in larger studies. The fractionation of proteins in plasma by acetonitrile is expected to be correlated with the molecular weight and hydrophobicity (at a given pH) of the proteins.[28] It was possible to influence the solubility of different plasma proteins by alternating the pH of the buffer. For example, proteins with pI 5-6 such as albumin could be expected to readily precipitate at a pH of 5 or 7. Pellets obtained at pH 5 or 7 were relatively easy to resuspend, but precipitates at pH 9 were very hard to dissolve and required additional use of ultrasonication. The reproducibility of protein extraction from pH 9 pellets was also poor, with notable changes in the abundance distribution of the proteins. Plasma pH in the sample usually varies between 7.5 and 8.5 and not surprisingly its precipitation profile is most similar to pH 9 where the pellet fraction is not much enriched and the supernatant is still highly dominated by albumin (data not shown).

The combination of X!Tandem and SpectraST identified 8,418 spectra (672 unique peptides) in the LC-MS/MS analysis of raw plasma, 6,751 spectra (568 unique peptides) in pellet fraction and 8,799 spectra (463 unique peptides) in the supernatant. As expected, the largest difference, or smallest overlap, was observed between the precipitate and the supernatant (Figure 2.4). The total proteome coverage discovered in pellet and supernatant fractions was 25% higher compared to crude plasma. A few peptides and proteins were only identified in the raw plasma and not in either the pellet or the supernatant fraction. However, relative spectral counts clearly show that most of the abundant proteins precipitate at pH 5 and remain in the pellet, while small proteins are enriched in the supernatant fraction (Figure 2.5a). Examples of such small proteins include several apolipoproteins (*e.g.* A1, A2, A4, C1 and C3) as previously shown by Anderson and Hunter.[29] Some mid-range (40-60 kDa) molecular weight proteins also increased in relative

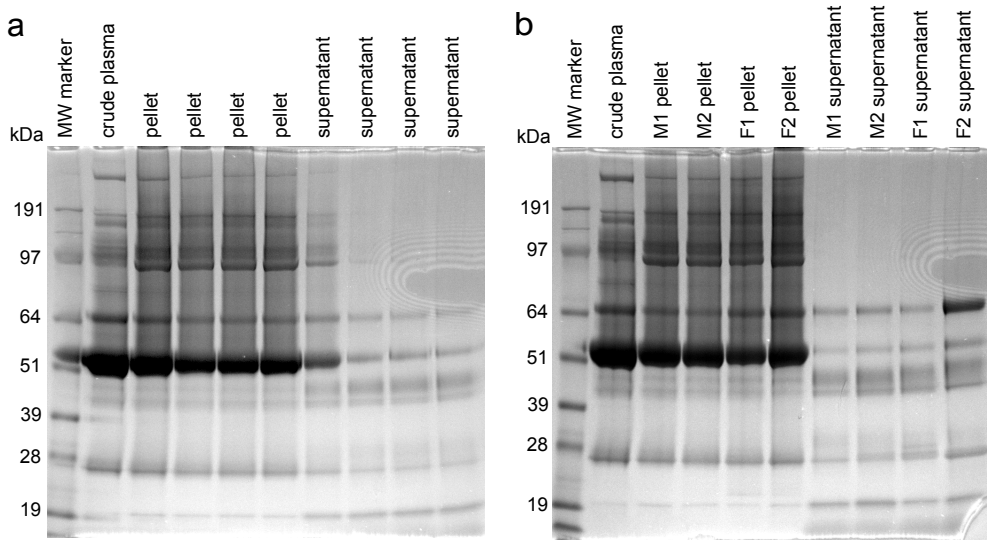


Figure 2.3. Representative SDS-PAGE gels illustrating the reproducibility of fractionation by acetonitrile precipitation at pH 5 of 20 μg human plasma protein from the same (a) and different (b) healthy volunteers. The pH of crude plasma samples was adjusted by adding 100 mM ammonium acetate buffer of corresponding pH in a 1:10 ratio and then precipitated with an equal volume of acetonitrile. M1 and M2 – plasma from male volunteers, F1 and F2 – plasma from female volunteers.

abundance in the supernatant. The spectral counts for proteins between 60 and 80 kDa are primarily due to albumin (90% in the raw plasma). The fraction of identified spectra assigned to albumin peptides in the entire raw plasma dataset was close to 60%. In the supernatant sample, only 5% of the identified spectra were from albumin peptides, indicating a depletion of ~90%. Also a number of other large and highly abundant proteins, such as α-2-macroglobulin and complement C3, were found to be significantly depleted. The relative abundance of albumin in the pellet fraction was approximately the same as for crude plasma.

The protein pI can be used as an additional criterion for fraction comparison and method evaluation (Figure 2.5b). Proteins are known to precipitate at the pH close to their pI values, and therefore most proteins including albumin were expected to precipitate at pH 5. However, more proteins with pI 5.0-5.5 were identified in the supernatant fraction and very few with pI 6.0-6.5. Interestingly, despite the peaks at pI 6.5-7.0, 8.0-8.5 and 9.0-9.5 there were only minor differences between the precipitates generated at different pH. The histogram for raw plasma showed a similar distribution to the sum of the pellet and supernatant fractions, if produced at the same pH (Figure 2.5b).

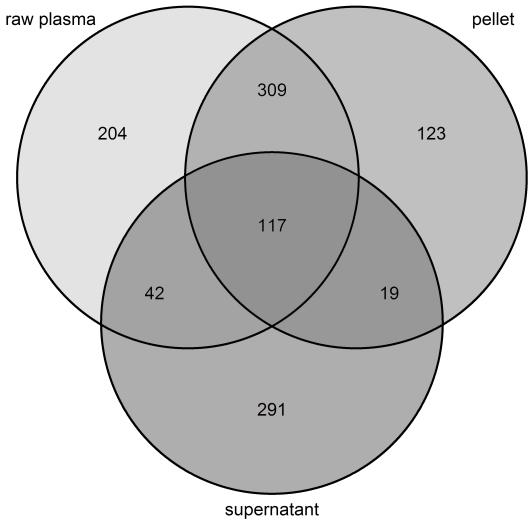


Figure 2.4. Venn diagram illustrating shared and uniquely identified peptides in the pellet, supernatant and crude plasma samples. The diagram was generated by the workflow described in Figure 2.2. In total, 6,700-8,800 spectra were identified in each fraction.

Additional information such as the isoelectric point of a protein or its molecular weight can be used as an extra dimension to filter out the erroneous identifications in samples fractionated in a pI or molecular weight-dependent manner. The Trans-Proteomic Pipeline already implements this for pI at the level of the peptides.

The workflow also calculated the protein hydrophobicity or GRAVY score. When comparing protein abundance in the pellet and the supernatant fractions with respect to GRAVY score and protein molecular weight, we see - somewhat surprisingly – that the hydrophobicity has a very small effect on the precipitation in comparison with molecular weight (Figure 2.5c).

CONCLUSIONS

Although blood plasma is one of the most popular sample sources in biomarker discovery, the large dynamic range of the protein concentration provides a serious challenge. As was shown by Kay *et al.* [28], albumin can be precipitated by simply adding acetonitrile. We have shown that adjustment of the pH prior to precipitation and addition of equal volume of acetonitrile was sufficient to remove approximately 90% of albumin and many other large proteins from the supernatant extracts. Moreover the proteome coverage has been increased by 25%. The procedure is simple, reproducible, can be quickly performed with common laboratory chemicals and

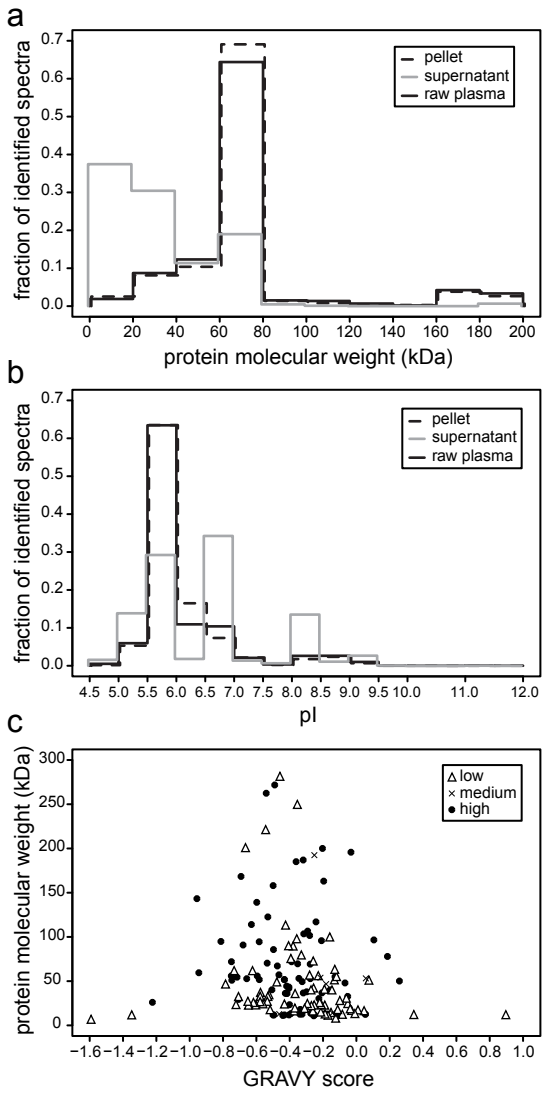


Figure 2.5. Histograms of the molecular weight (a) and predicted pI (b) distributions of proteins identified in the crude sample (solid, black), pellet (dashed, black) and supernatant (gray), accompanied by a graph with calculated GRAVY score plotted against protein molecular weight (c). In the latter, proteins marked by black circles have pellet to supernatant spectral count ratio ≥2, by triangles ≤0.5, and by crosses more than 0.5 and less than 2.

equipment, and is compatible with standard techniques such as SDS-PAGE and LC-MS/MS. The method may be applicable in many types of proteomic analyses of plasma and other samples. For instance, optimised organic precipitation can not only be used for the sample decomplexification but also to concentrate target proteins which might be an advantage in biomarker discovery. This method has been successfully implemented in urine proteomics.[30]

Although the gain in protein coverage is lower than what can be achieved with immunoaffinity procedures, it should be emphasized that the present technique is robust and can easily be applied in large clinical studies.

Further improvements or adaptation of experimental protocols may focus on specific enrichment for protein modifications (sulfation, phosphorylation, glyco- or lipoproteins), as well as providing some constraints for the peptide/protein identification algorithms, such as limits on pI, molecular weight or post-translational modifications. Further optimization may also aim at improving the quality and albumin depletion of the pellet fraction.

As an additional remark, the Taverna scientific workflow used in this study contains in a single workflow and interface all the steps from raw mass spectrometry data through format conversion, peptide identifications, statistical evaluation, data mining to visualization in figures essentially as they appear in this paper, completely automated and without any interactive manual input. The workflow and the data discussed here are available on-line, enabling anyone to repeat the analysis or adapt the workflow for any other experiment comparing two or more tandem mass spectrometry datasets with respect to physico-chemical protein properties.

REFERENCES

1. Van Slyke, D. D.; Hiller, A.; Phillips, R. A.; Hamilton, P. B.; Dole, V. P.; Archibald, R. M.; Eder, H. A., The Estimation of Plasma Protein Concentration from Plasma Specific Gravity. *The Journal of Biological Chemistry* **1950,** 183, (1), 331-347.
2. Putnam, F. W., Alpha, beta, gamma, omega - the roster of the plasma proteins. In *The Plasma Proteins*, 2 ed.; Putnam, F. W., Ed. Academic Press: New York, 1975; Vol. 1, pp 57-130.
3. Roche, M.; Rondeau, P.; Singh, N. R.; Tarnus, E.; Bourdon, E., The antioxidant properties of serum albumin. *Febs Letters* **2008,** 582, (13), 1783-1787.
4. Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A., Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **2002,** 359, (9306), 572-577.
5. Tirumalai, R. S.; Chan, K. C.; Prieto, D. A.; Issaq, H. J.; Conrads, T. P.; Veenstra, T. D., Characterization of the low molecular weight human serum proteome. *Molecular & Cellular Proteomics* **2003,** 2, (10), 1096-1103.
6. Chen, Y. Y.; Lin, S. Y.; Yeh, Y. Y.; Hsiao, H. H.; Wu, C. Y.; Chen, S. T.; Wang, A. H. J., A modified protein precipitation procedure for efficient removal of albumin from serum. *Electrophoresis* **2005,** 26, (11), 2117-2127.
7. Guerrier, L.; Claverol, S.; Fortis, F.; Rinalducci, S.; Timperio, A. M.; Antonioli, P.; Jandrot-Perrus, M.; Boschetti, E.; Righetti, P. G., Exploring the Platelet Proteome via Combinatorial, Hexapeptide Ligand Libraries. *J. Proteome Res.* **2007,** 6, (11), 4290-303.
8. 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). *Journal of proteome research* **2004,** 3, (2), 235-44.
9. Cohn, E. J., The Properties and Functions of the Plasma Proteins, with a Consideration of the Methods for their Separation and Purification. *Chem Rev* **1941,** 28, (2), 395–417.
10. Ofosu, F.; Cassidy, K.; Blajchman, M. A.; Hirsh, J., Immunodepletion of Human-Plasma Factor-Viii. *Blood* **1980,** 56, (4), 604-607.
11. Eliasson, M.; Olsson, A.; Palmcrantz, E.; Wiberg, K.; Inganas, M.; Guss, B.; Lindberg, M.; Uhlen, M., Chimeric Igg-Binding Receptors Engineered from Staphylococcal Protein-a and Streptococcal Protein-G. *Journal of Biological Chemistry* **1988,** 263, (9), 4323-4327.
12. Chromy, B. A.; Gonzales, A. D.; Perkins, J.; Choi, M. W.; Corzett, M. H.; Chang, B. C.; Corzett, C. H.; McCutchen-Maloney, S. L., Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. *J Proteome Res* **2004,** 3, (6), 1120-7.
13. Björhall, K.; Miliotis, T.; Davidsson, P., Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* **2005,** 5, (1), 307-317.

14. Michael, S. E., Isolation of Albumin from Blood Serum of Plasma by Means of Organic Solvents. *Biochemical Journal* **1962,** 82, (1), 212-218.

15. Polson, C.; Sarkar, P.; Incledon, B.; Raguvaran, V.; Grant, R., Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography-tandem mass spectrometry. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **2003,** 785, (2), 263-275.

16. Alpert, A. J.; Shukla, A. K. In *Precipitation of large, high-abundance proteins from serum with organic solvents*, ABRF, Denver, 2003; Denver, 2003; pp Poster# P111-W.

17. Chertov, O.; Biragyn, A.; Kwak, L. W.; Simpson, J. T.; Boronina, T.; Hoang, V. M.; Prieto, D. A.; Conrads, T. P.; Veenstra, T. D.; Fisher, R. J., Organic solvent extraction of proteins and peptides from serum as an effective sample preparation for detection and identification of biomarkers by mass spectrometry. *Proteomics* **2004,** 4, (4), 1195-1203.

18. Chertov, O.; Simpson, J. T.; Biragyn, A.; Conrads, T. P.; Veenstra, T. D.; Fisher, R. J., Enrichment of low-molecular-weight proteins from biofluids for biomarker discovery. *Expert Review of Proteomics* **2005,** 2, (1), 139-145.

19. Jameson, E., A Phase Rule Study of the Proteins of Blood Serum: The Effect of Changes in Certain Variables. *J Gen Physiol* **1937,** 30, (6), 859–877.

20. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P., Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004,** 20, (17), 3045-3054.

21. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004,** 22, (11), 1459-66.

22. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **2005,** 1, 1-8.

23. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.

24. 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* **2002,** 74, (20), 5383-92.

25. Kyte, J.; Doolittle, R. F., A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology* **1982,** 157, (1), 105-132.

26. Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D., The Focusing Positions of Polypeptides in Immobilized Ph Gradients Can Be Predicted from Their Amino-Acid-Sequences. *Electrophoresis* **1993,** 14, (10), 1023-1031.

27. Vizcaino, J. A.; Cote, R.; Reisinger, F.; Barsnes, H.; Foster, J. M.; Rameseder, J.; Hermjakob, H.; Martens, L., The

Proteomics Identifications database: 2010 update. *Nucleic Acids Res* **2010,** 38, (Database issue), D736-42.

28. Kay, R.; Barton, C.; Ratcliffe, L.; Matharoo-Ball, B.; Brown, P.; Roberts, J.; Teale, P.; Creaser, C., Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. *Rapid Commun Mass Spectrom* **2008,** 22, (20), 3255-60.
29. Anderson, L.; Hunter, C. L., Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Molecular & Cellular Proteomics* **2006,** 5, (4), 573-588.
30. Thongboonkerd, V.; Chutipongtanate, S.; Kanlaya, R., Systematic evaluation of sample preparation methods for gel-based human urinary proteomics: quantity, quality, and variability. *J Proteome Res* **2006,** 5, (1), 183-91.

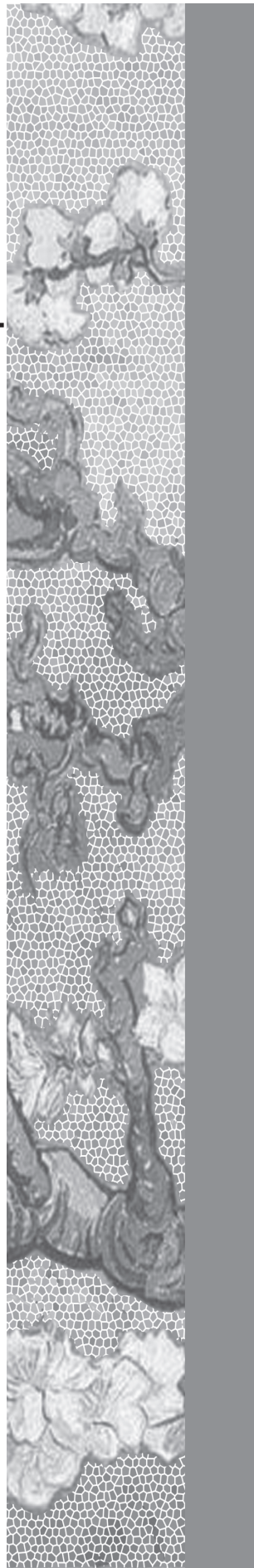
3

A Novel Mass Spectrometry Cluster for High-Throughput Quantitative Proteomics

Magnus Palmblad, Yuri E. M. van der Burgt, Ekaterina Mostovenko, Hans Dalebout and André M. Deelder

Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

J. Am. Soc. Mass Spectrom. 2010, 21, 1002–1011



ABSTRACT

We have developed and implemented a novel mass spectrometry (MS) platform combining the advantages of high mass accuracy and resolving power of Fourier transform ion cyclotron resonance (FTICR) with the economy and speed of multiple ion traps for tandem mass spectrometry. The instruments are integrated using novel algorithms and software and work in concert as one system. Using chromatographic time compression, a single expensive FTICR mass spectrometer can match the throughput of multiple relatively inexpensive ion trap instruments. Liquid chromatography (LC)-mass spectrometry data from the two types of spectrometers are aligned and combined to hybrid datasets, from which peptides are identified using accurate mass from the FTICR data and tandem mass spectra from the ion trap data. In addition, the high resolving power and dynamic range of a 12 tesla FTICR also allows precise label-free quantitation. Using two ion traps in parallel with one LC allows simultaneous MS/MS experiments and optimal application of collision induced dissociation and electrontransfer dissociation throughout the chromatographic separation for increased proteome coverage, characterization of post-translational modifications and/or simultaneous measurement in positive and negative ionization mode. An FTICR-ion trap cluster can achieve similar performance and sample throughput as multiple hybrid ion trap-FTICR instruments, but at a lower cost. We here describe the first such FTICR-ion trap cluster, its performance and the idea of chromatographic compression.

INTRODUCTION

In recent years, mass spectrometry (MS) has become a popular method for identification and quantitation of proteins and metabolites in complex biological matrices. The reasons for this are at least 2-fold: mass spectrometry can separate a very large number of chemical species of different mass in a complex sample and, secondly, unknown peptides or proteins can be routinely and automatically identified by data-dependent tandem mass spectrometry (MS/MS). The resolution or peak capacity is increased further by coupling the mass spectrometer to a liquid chromatography (LC) system. Analytical challenges from the rapidly expanding field of proteomics have pushed the development of mass spectrometers in general and led to further optimization of systems for peptide and protein analysis. The workhorses in MS-based proteomics are ion traps.[1] These are perfectly suited for on-line coupling to LC via electrospray ionization (ESI) and are capable of analyzing complex peptide mixtures by rapid MS/MS. Although ion traps are sensitive and versatile, they have relatively low resolving power ($<10^4$) and mass measurement accuracy (~100 ppm). In modern mass spectrometers, quadrupoles or ion traps are often used in hybrid configuration with a second mass analyzer, such as time-of-flight[2] (TOF), Orbitrap,[3] or FTICR,[4, 5] within a single vacuum system. In these hybrid instruments, the ion trap provides rapid, sensitive MS/MS, or at least precursor ion selection, and the second analyzer accurate mass and high resolving power. The hybrid linear ion trap-Fourier transform ion cyclotron resonance (FTICR) mass spectrometer[4] can be considered to represent the current state-of-the-art in commercially available mass spectrometry instrumentation. An alternative paradigm to combining two mass analyzers in one physical instrument is to merge data from two different instruments analyzing the same sample. Several analytical strategies have been developed based on this general idea, notably the accurate mass and time tag (AMT) approach of Smith *et al.*,[6] wherein MS/MS data from ion traps are used to validate peptide identifications based on accurate mass and to train a predictor of chromatographic retention times.[7, 8] Accurate MS and ion trap MS/MS data can also be directly combined using chromatographic alignment.[9, 10] All published hybrid instrument designs and previously published data fusion schemes are inefficient, at least timewise, in their use of the accurate mass detector.

This is because the latter invariably has much higher resolving power than the ion trap used for MS/MS and thus is able to detect most species selected

for MS/MS on a considerably shorter chromatographic time scale. Given the difference in cost and size of ion traps (small and relatively inexpensive) and Orbitrap or FTICR mass spectrometers (larger and more expensive), it also makes economic sense to combine multiple ion traps with a single FTICR mass spectrometer. We have therefore designed and implemented an integrated FTICR-ion trap cluster, a system of mass spectrometers that work together as a single entity to analyze one sample stream using differential chromatographic gradients for LC-MS and LC-MS/MS. The use of a dedicated instrument platform is novel, but it is the concept and use of different chromatographic time scales that is important, rather than the exact configuration or number of ion traps. We will therefore describe the performance and consequence of chromatographic compression in detail and exemplify this approach with an application in a quantitative proteomics study.

MATERIALS AND METHODS

Test Samples

For evaluation of the FTICR-ion trap cluster, we repeated the classic glucose/lactose diauxie experiment by Jacob and Monod.[11] *E. coli* K12 strain MG1655 was acquired from ATCC and cultured in 1 L MOPS minimal medium with 0.5 g/L glucose and 1.5 g/L lactose in a 3 L fermentor (Applikon Biotechnology, Schiedam, The Netherlands), duplicating as closely as possible the recent glucose/lactose diauxie gene expression study by Traxler *et al*.[12] The culture was monitored by spectrophotometric OD600 measurement, glucose concentration measured using a glucose oxidase assay kit (Sigma-Aldrich Chemie B.V., Zwijndrecht, The Netherlands) and lactose concentration followed using a galactosidase/lactose kit (BioVision, Mountain View, CA, USA). Three replicate cultures were sampled at approximately –100, –50, –10, 0, 10, 20, 30, 40, 50, and 60 min relative to the diauxic shift. Proteins were extracted using the Novagen "BugBuster" kit (Merck KGaA, Darmstadt, Germany), following the manufacturer's recommended protocol with 5 mL of lysis buffer per gram of wet cell weight.

For identification, 25 μg protein from two time points, one before and one after the diauxic shift, was fractionated using SDS-PAGE (NuPAGE 8%–12%; Invitrogen, Carlsbad, CA, USA) by cutting the gel lane into 26 two-

mm bands, reduced (10 mM DTT, 56 °C, 45 min), alkylated (iodoacetamide, room temperature, 1 h in dark) and digested in-gel using trypsin (sequencing grade; Promega, Madison, WI, USA). For quantitative measurement, 250 μg of protein from each individual sample was digested as above but in solution and 2 μg of each digest injected on column.

To compare the sensitivity and quantitative precision, bovine serum albumin (BSA) was spiked into an *E. coli* protein extract from a post-diauxie time point at BSA-to-*E. coli* ratios of 0, 0.01%, 0.1%, 1%, and 10%. The digestion protocol was the same as for the gel slices except for the omission of the destaining and washing steps. The spiked samples were analyzed in an iterated sequence from low to high concentration BSA, with a blank after the highest concentration.

FTICR-Ion Trap Cluster

All LC systems in the FTICR-ion trap cluster are parallel, splitless NanoLC-Ultra 2D plus (Eksigent, Dublin, CA, USA) for ultra-high-pressure parallel LC with an additional loading pump for fast sample loading and washing. For this work, all LC systems were configured with 300 μm-i.d 5-mm PepMap C18 trap columns (Dionex, Sunnyvale, CA, USA), 15-cm 300 μm-i.d. ChromXP C18 columns supplied by Eksigent and running linear gradients, all from 4% to 44% acetonitrile in 0.05% formic acid, but of different lengths.

The FTICR is a solariX 12 T FTICR (Bruker Daltonics, Bremen, Germany) equipped with an Apollo II ESI source and external quadrupole for precursor ion selection and/or MS/MS outside the cell. In the FTICR ion trap cluster, this quadrupole is only used as an ion guide with transmission optimized for *m*/*z* 400–1000, which includes most doubly- and triply charged tryptic peptides. Typically 2^{20} (~10^6) data points are acquired per spectrum, and one spectrum is acquired every 2–3 s.

The ion traps in the particular instrument cluster used to generate all data shown here were of two models, both from Bruker Daltonics, with one HCT ultra PTM Discovery system for collision-induced dissociation (CID) and electron-transfer dissociation (ETD), and one standard HCT ultra system exclusively for CID combined in pairs and connected to a single LC system. After each MS scan, up to five abundant multiply charged species in *m*/*z* 300–1300 were selected for MS/MS and excluded for 1 min after being

selected twice. For spectral counting no active precursor exclusion was used.

Each mass spectrometer is controlled by a dedicated computer, but all instruments are monitored from a single desk with two monitors using dual 4-port KVM switches. The LC systems are controlled using the HyStar 3.2-3.4 with a plug-in from the LC manufacturer, the ion traps by esquireControl 6.2 and the FTICR by apexControl 3.0, all from the instrument manufacturer. The acquired data from each mass spectrometer is automatically transferred to a dedicated server and processed as described below.

Data Analysis

Automation of data analysis tasks is essential for the easy operation of the instrument cluster. All data is continuously copied over a gigabyte/s Ethernet connection to a dedicated data processing server. By using a convention with delimited LC mass spectrometer species and unique sample identifiers in the HyStar sample lists, and consequently the resulting filenames, the ion trap data can be automatically searched against a species-specific sequence database using a local installation of X!Tandem.[13] The identified peptides are then used to align each ion trap dataset with the corresponding FTICR dataset from the same sample in the "hybrid instrument emulation mode," as previously described.[10] Additionally and optionally, identified and aligned peptides can also be used to internally calibrate the FTICR mass spectra,[14] generating a hybrid peak list with sub-ppm precursor mass measurement uncertainty. The hybrid peak lists are then automatically searched against the same database but with a narrow precursor (peptide) mass tolerance window. The data analysis so far is performed in the background without any user input. The processing scripts and all software used for the alignment of LC-MS and LC-MS/MS datasets and for integrating peak areas in the LC-MS data will be freely available as open source on http://www.ms-utils.org/cluster. All analyses can also be performed in batch-mode and off-line, allowing the use of other search engines such as Mascot[15] or Phenyx[16] running on separate servers. For the data presented in this paper, we exclusively relied on Mascot, as it is the most common of the search engines available in our lab. All quantitative analyses are currently only performed in batch-mode. All FTICR datasets are searched for all identified peptides in a narrow retention time window and a very narrow m/z window (typically ±2 ppm, as low signal-to-noise

peaks have larger mass measurement errors) to retrieve quantitative information of all identified peptides in each biological replicate. The quantitation is done by adding all signals in this narrow *m*/*z* and time window. This only requires that the last chromatographic dimension used with the ion traps for identification can be aligned with that used with the FTICR for quantitation. All peptide intensities are then summed to total protein intensity.

RESULTS

FTICR-Ion Trap Cluster

The basic working principle of the FTICR-ion trap cluster (Figure 3.1) is that each compound in each sample is analyzed twice, once on the FTICR for accurate mass determination and once on an ion trap for MS/MS. The high resolving power of the 12 tesla (T) FTICR mass spectrometer allows many simultaneous accurate mass determinations also in very complex spectra. All mass spectrometers in the cluster are coupled on-line to parallel ultra-high-pressure LC systems for efficient use of the mass spectrometers and high chromatographic peak capacity. The new and enabling idea behind the FTICR-ion trap cluster is the use of short and long chromatographic gradients, respectively, with a single accurate mass/high-resolution mass spectrometer, here a 12 T FTICR, and multiple rapid and sensitive MS/MS instruments, here 3 × 2 state-of-the-art ion traps. These two independent LC separations are performed in such a way that the single FTICR can keep up with the throughput of and provide accurate MS data to the MS/MS data from the multiple ion traps, for instance by compressing the gradient a factor equal to the number of ion traps or ion trap "modules" in the system, where a module is defined as one or more ion traps coupled to a single LC (see Figure 3.1). The specific ion trap cluster described here combines one ion trap for ETD[17] with one ion trap for CID in each of three such modules in the system. The use of separate instruments allows different configurations, the most straightforward being one LC system per ion trap.

However, there are some advantages in the use of two ion traps in parallel to one LC. In these ion traps, there is significant dead time in the switching between CID and ETD. Using two ion traps where one is dedicated to CID and one to ETD not only improves duty cycles, but in principle allows the instruments to be tuned and optimized for a particular dissociation method,

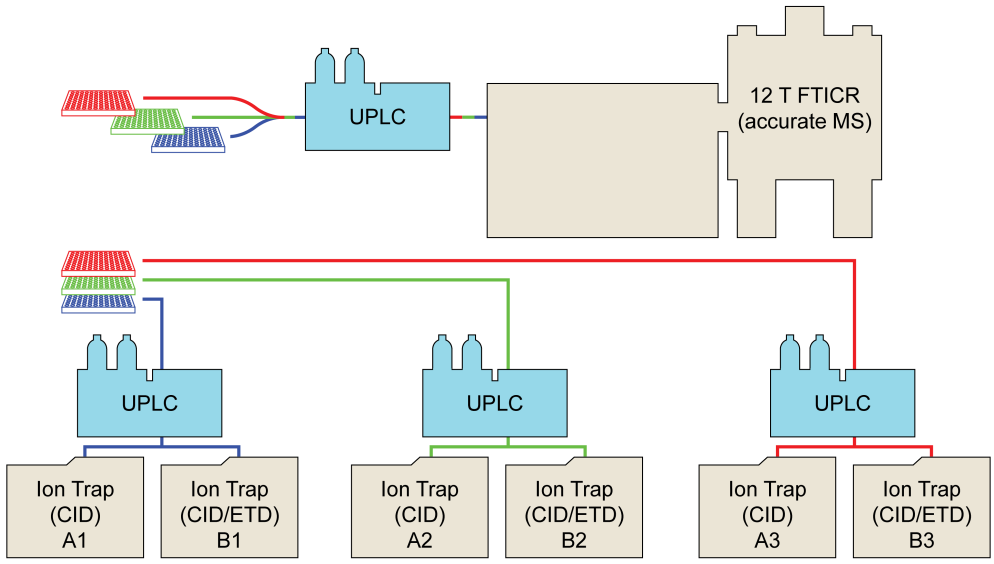


Figure 3.1. The FTICR-ion trap cluster consists of six ion traps grouped in three pairs. In each pair, the eluent from one capillary LC system is split in two, with half going to a CID-only ion trap and half to an ETD-capable ion trap. The ion traps are used for fast MS/MS acquisition. The FTICR component consists of an identical capillary LC system as used with the ion traps and a 12 T solariX Qq-FTICR. Although capable of MS/MS and even MS*n*, the FTICR is used exclusively for accurate MS and quantitation in this instrument cluster.

including precursor ion selection criteria. For complex samples, it is also well known that ion traps cannot sample all detected peptides for MS/MS on any reasonable chromatographic timescale. It is also feasible to combine data from split samples acquired on two or more LC-ion trap systems with one LC per ion trap. The data acquisitions then need not be concurrent, but the results from one acquisition can be used to make an exclusion list for the next.

Automated integration of accurate mass determinations from the FTICR with a large number of MS/MS spectra from the ion traps is key in the instrument cluster, and alignment and combination of accurate MS data from FTICR and MS/MS data from ion trap mass spectrometers improves confidence in peptide identifications and makes it possible to identify MS/MS spectra of lower quality, resulting in more peptide and protein identifications at a given false discovery rate.[10] The ion trap and FTICR data acquisitions are physically and timewise independent which allows the instrument cluster to be operated in different modes. We call two basic

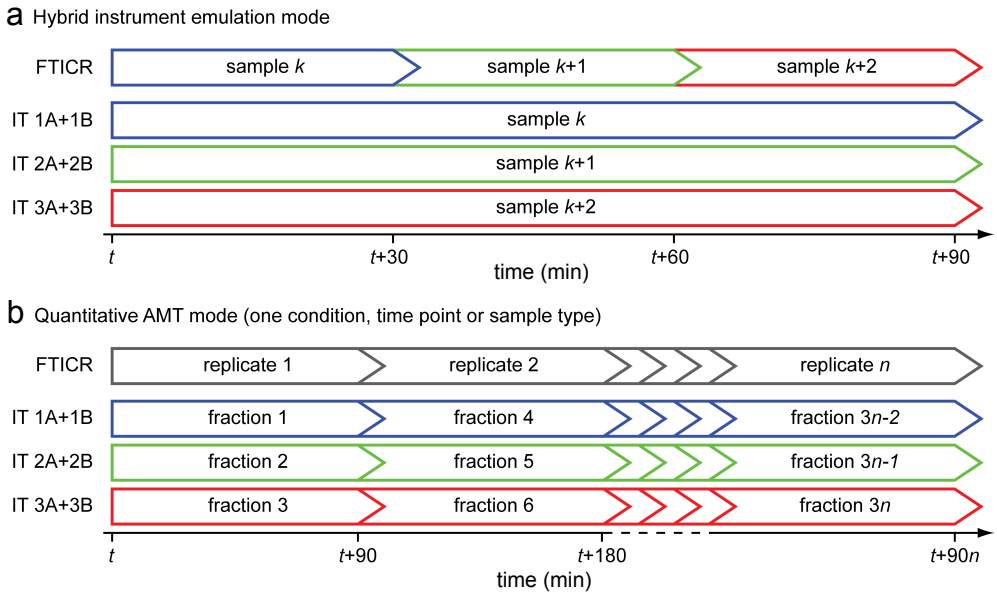


Figure 3.2. Repeating scheduling blocks for the FTICR-ion trap cluster with 3 _ 2 ion traps in the hybrid instrument emulation mode (a) and quantitative accurate mass and time mode (b). If these schedules are kept, the FTICR and ion trap cluster will require exactly the same time for the analysis of a batch of samples. It is not necessary to strictly adhere to such time schedules, but they are helpful in maximizing sample throughput and obtainable data quality.

modes the "hybrid instrument emulation" mode and the "quantitative accurate mass and time tag" (QAMT) mode, respectively (Figure 3.2).

In the hybrid instrument emulation mode, each sample is analyzed on both the ion trap and FTICR platforms, and the data aligned and combined to hybrid datasets with accurate precursor ion masses from FTICR and MS/MS from two ion traps, for instance both CID or one CID and one ETD (a detailed comparison of different CID/ETD schemes was recently published by Leinenbach *et al.*[18] These hybrid datasets are complete as each tandem mass spectrum is supplied with an accurate precursor mass, and similar in quality to what would be obtained from a hypothetical ETDcapable 12 T ion trap-FTICR instrument. If the FTICR dataset is acquired first, inclusion lists can be made based on the accurate mass for subsequent MS/MS in the ion traps, just as in a hybrid instrument. The latest version of the ion trap control software allows for scheduled precursor lists with *m/z* as well as elution time windows of precursors to be selected for MS/MS. The QAMT mode uses a different set of samples, or fractions of a representative or pooled

sample for identification of peptides quantified in individual biological replicates, time points or experimental conditions by the FTICR. This is similar to the AMT scheme by Smith *et al.* but the emphasis here is on the FTICR providing both quantitation and accurate MS to increase confidence in peptide identifications. The FTICR-ion trap cluster also allows the on-the-fly generation of the AMT database.

Dataset Alignment and Chromatographic Compression

An *E. coli* whole-cell lysate obtained as described in the Methods section was analyzed using linear chromatographic gradients of 9, 10, 11.3, 12.9, 15, 18, 22.5, 30, 45, and 90 min corresponding to compression ratios of 10:1 down to 1:1. The previously described alignment algorithm[10] had been designed to be as general and robust as possible, and was found to be insensitive to chromatographic time scales. Without modification, the algorithm correctly aligned datasets from the different chromatographic time scales (Figure 3.3). The elution times depend linearly ($R^2 > 0.99$) on compression ratio, as expected (Figure 3.4a). However, the alignment is more robust, or reproducible, for lower compression ratios, as indicated by the larger variation in the slope of the piecewise linear alignment. This is caused by the software having to look for the precursor *m*/*z* in a larger relative time window at higher compression ratios. The alignment is very robust up to at least a compression ratio of 3:1, which is the number of ion trap modules and the highest compression ratio in our setup.

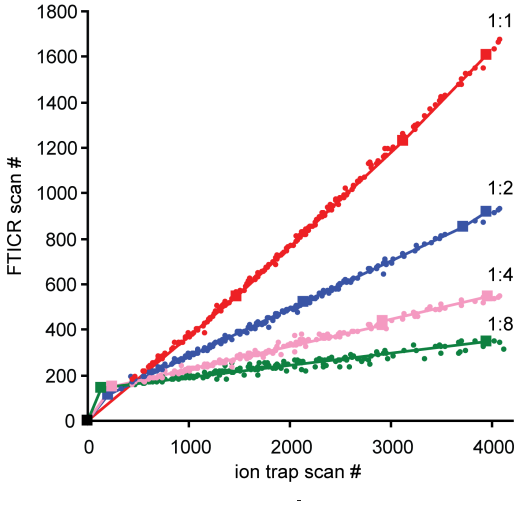


Figure 3.3. Automatic alignments (lines) of FTICR and ion trap datasets with chromatographic compression FTICR:ion trap 1:1, 1:2, 1:4, and 1:8, showing the peptide features contributing to the fitness function in the genetic algorithm used for alignment (dots). All alignments were performed allowing a 25 scan residual standard error, mass measurement error tolerance ±1 ppm and 262 unique peptides identified with a Mascot ion score cutoff 30. The squares represents the breakpoints in the piecewise linear alignments. No parameters needed to be adjusted to align chromatograms of different time scales.

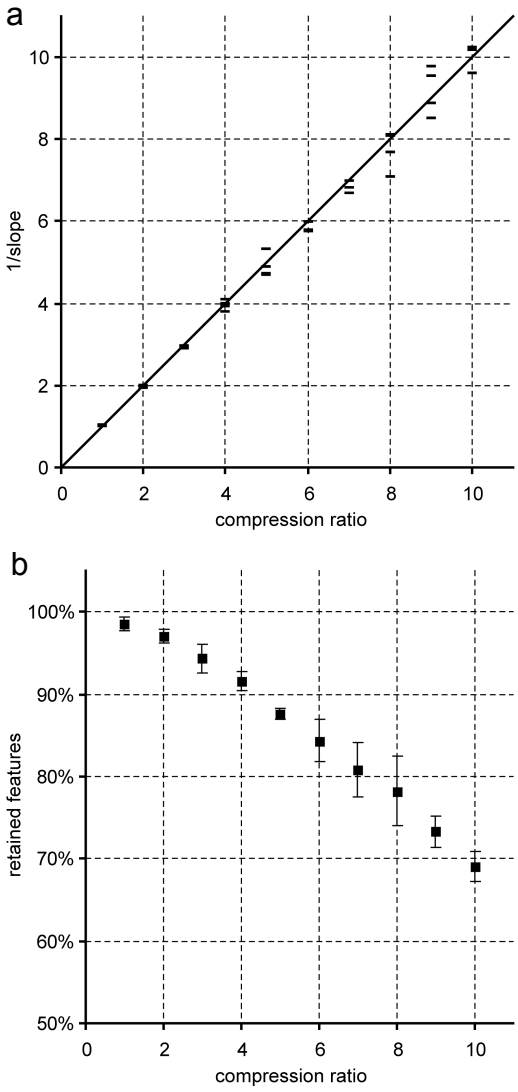


Figure 3.4. The slope of the main segment of the piecewise linear alignment as a function of chromatographic compression (a). The plotted slope for four replicates at each compression ration was normalized to compensate for the different data acquisition rates in the FTICR and ion traps. The retention times of matched peptide features have a linear dependence on the length of the chromatographic gradient, but alignment is more difficult at higher compression ratios. More features are also lost, or not observed, within ±1 ppm in the expected elution time window, the more the chromatographic gradient is compressed on the FTICR (b). These numbers are derived from the relatively abundant peptides identified by MS/MS. For lower abundant species not identified by MS/MS, the fraction lost is likely larger.

Another cost of compression of chromatographic gradients is that fewer features (here peptides) are observed at higher compression ratios. This can be quantified as the fraction of species tentatively identified by MS/MS in the ion traps and above the detection limit and within ±1 ppm of the theoretical mass in the FTICR. The fraction of identified peptides retained as a function of chromatographic compression is shown in Figure 3.4b. Note that the percentage of preserved features decreases smoothly as a function of chromatographic compression, from around 98% with no compression to just below 70% with compression from 90 to 9 min. At 3-fold compression

from 90 to 30 min, 94% of the peptides selected for MS/MS are retained. This limited loss is compensated with nearly 200% gain in throughput.

There will probably always be some tradeoff between coverage and measurement throughput in proteomics. The FTICR-ion trap cluster is no exception, even though it is designed to provide a relatively deep coverage using high-field FTICR with a high throughput using chromatographic compression and multiple ion traps. It is important to keep in mind that these numbers refer to peptides selected for MS/MS and producing good CID spectra in the ion traps, and not all features detectable by MS. The 2% "lost" peptides with identical chromatographic gradients is comparable to what would be expected between repeated analyses on the same system and are comprised of erroneous peptide identifications from the ion trap data alone, peptides measured outside the tolerated and searched *m*/*z* window in the FTICR, and peptides falling below the detection limit in the FTICR. Gygi and coworkers have reported "losses" of a similar magnitude between LTQonly and LTQ-FT datasets, where slightly more MS/MS spectra are acquired and a few more (unmodified) proteins identified using only the LTQ rather than the hybrid LTQ-FT,[19, 20] so this is not a phenomenon unique to the FTICR-ion trap cluster. In the FTICR-ion trap cluster, the MS/MS data is acquired in the ion traps completely independently from the FTICR, without any time loss.

High Throughput Quantitative Proteomics

Label-free quantitation using FTICRMS is more precise and covers a larger dynamic range in relative protein abundance than label-free quantitation using only ion traps, for instance with the emPAI spectral counting method[21] (Figure 3.5). Good agreement with calculated isotopic distributions and precise relative quantitation using ^{15}N-labeling and FTICRMS has also been reported previously.[22] The FTICR-ion trap cluster is ideally suited to study proteome dynamics, analyzing large cohorts of similar samples. We have chosen to illustrate the throughput and applicability of the instrument cluster with a time-course study of the glucoselactose diauxie[11] in *E. coli*. A subset of the data is shown in Figure 3.6 and compared with a recently published gene expression study.[12] The diauxie experiment serves as a positive control, as we expect to see β-galactosidase to be the most up-regulated protein during the glucose-lactose shift. A 10-fold increase in abundance of this protein could also be observed in each of the three replicate time series (Figure 3.6). This quantity and quality of data can be routinely generated in less 24 h using the FTICR-ion

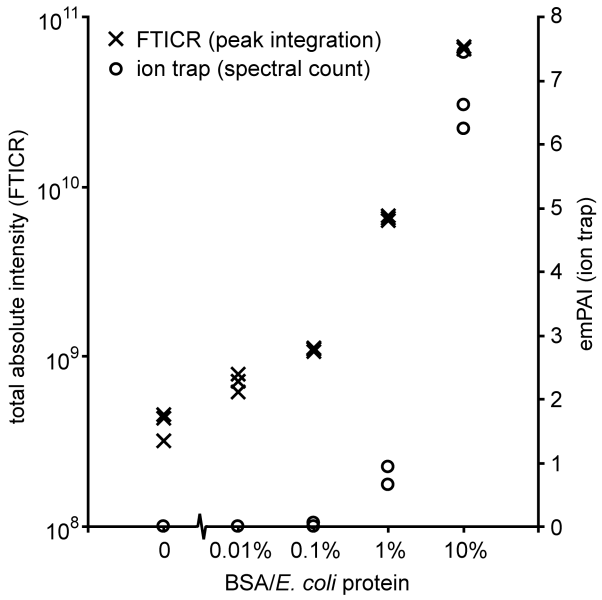


Figure 3.5. Comparison of technical reproducibility (quantitative precision) and dynamic range between FTICR and the emPAI[21] method for label-free quantitation using three replicate measurements of BSA spiked in a constant background of *E. coli* cell lysate at each of five relative abundance levels: 0, 0.01%, 0.1%, 1, and 10% of the background of *E. coli* protein. The FTICR signal abundance is the total peak intensity integrated (40 scans around the chromatographic maximum in a region ±5 ppm and ±25 scans from the *m*/*z* and retention time predicted from the ion trap MS/MS data) over the chromatographic peaks of 28 identified BSA peptides, without normalization. No BSA peptides were identified in the ion trap at the 0 or 0.01% spike level, and only a single peptide in one of the replicates at the 0.1% level. At the 1% and 10% levels, spectral counting works well, with relative standard error of 19% at the 1% spike level and 9% at the 10% spike level. Conversely, the relative standard error of the FTICR measurement is around 3% at and above the 0.1% level, and 12% at the 0.01% level. Both the limit of detection and limit of quantitation, as functions of relative abundance, are approximately two orders of magnitude lower in the FTICR method than with spectral counting from the same sample. In this comparison, the same spiked samples and 90-min gradients were used for the ion trap and FTICR analyses. The figure shows data from two columns, one used with the FTICR and one with the ion trap.

trap cluster, illustrating the power of the system in large-scale proteomic studies.

With sufficient protein or peptide fractionation or enrichment, the number of identified peptide features may eventually exceed the number of observable peptide features in the FTICR data. The more identified peptides, the larger the risk of false matches between FTICR and ion trap data. However, the

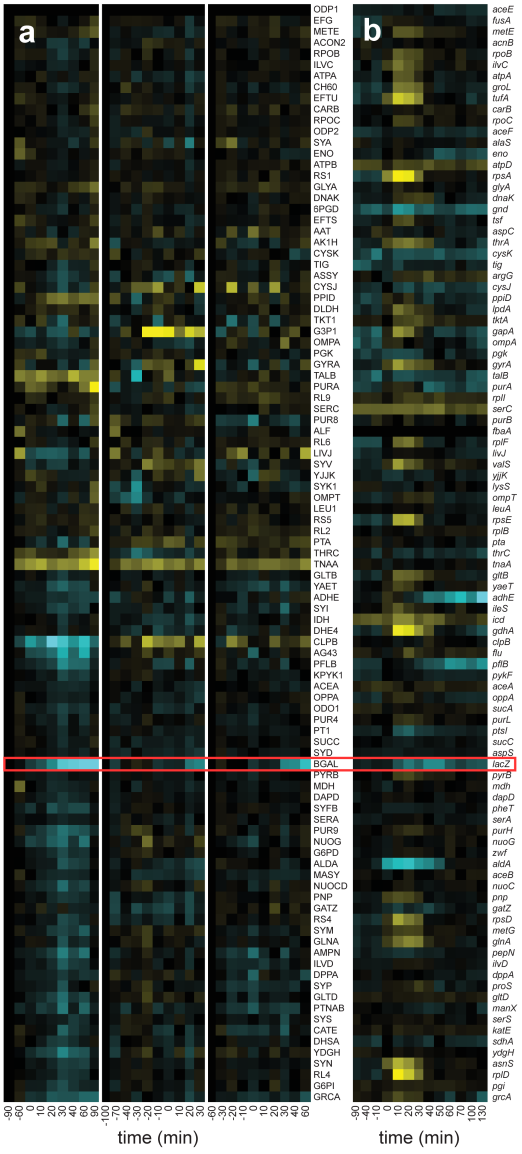


Figure 3.6. Protein (a) and gene (b) expression in *Escherichia coli* K12 during glucose-lactose diauxie, showing 100 out of 948 proteins (630 with more than one peptide) quantified at 10 time points across three biological replicates – a total of 27,492 (18,270 with multiple peptides) quantitative protein abundance measurements by the FTICR-ion trap cluster. Cyan color denotes increase, yellow decrease, and black no change. The times are relative to the onset of diauxic shift as observed by OD600 measurement. These data can be obtained in less than 24 h using the instrument cluster. The figure shows ~10% of the total proteomic dataset (see Supplemental Information, which can be found in the electronic version of this article) compared with the corresponding gene expression data from a similar experiment by Traxler *et al.*[12] The protein with the strongest increase in abundance during the glucose-lactose diauxic shift is β-galactosidase/*lacZ*, as expected (marked in red).

situation may be improved by good mass measurement accuracy (<0.5 ppm standard error in the 12 T FTICR after automatic internal calibration) chromatographic reproducibility and restricting the number of peptides of similar mass and elution times, which is essentially the idea behind the accurate mass and time tag approach. Most importantly, the false matches between the FTICR and ion trap data is expected to increase with the difference between the individual samples analyzed by the FTICR and the sample or samples used for fractionation and peptide identification.

Quantifying samples from individual time points or experimental conditions using one or more pooled samples from the same experiment for identification is likely to reduce the risk of false matches. In the pipeline described here, we also strive to use all available information, which includes using at least all peptides with a sequence unique to a protein for protein identification and quantitation. This way, the false positive rate can be much lower on the protein level than on the peptide level, the confidence generally increasing the more peptides are available for identification and quantitation per protein.

DISCUSSION

The first implementation of an FTICR-ion trap cluster described here was designed for robustness and optimal instrument performance using state-of-the-art ion trap and 12 T FTICR mass spectrometers, capillary ultra-high-performance liquid chromatography systems and standard ESI source, operating at 2 μL/min. We have chosen a 12 T system as it provides highperformance at a reasonable cost. The chromatographic alignment is more robust the more features are used for alignment, i.e., the more information that is available for alignment. However, the time required for evaluation of the fitness function in the genetic algorithm is proportional to the number of features or peptides used for alignment. In our experience, it is a good practice to limit the number of features to at most a few thousand.[23] For peptides, this is easily done by raising the search engine score threshold. At the other end, the algorithm does not need more than 30–40 matched peptides distributed over the chromatographic separation to produce a good alignment.[10]

Fragmentation by CID is the most commonly used method for MS/MS in general as well in the ion trap cluster. However, peptides with labile post-translational modifications such as phosphorylation or glycosylation often lose these before producing sequence-specific backbone fragments by CID, thus preventing the exact localization of the post-translational modification. The recently introduced ETD in linear and three-dimensional ion traps is therefore extremely useful for primary structure determination of peptides containing posttranslational modifications, and has been rapidly implemented and accepted in the field. Electron-transfer dissociation, like electron capture dissociation[24] in Fourier transform ion cyclotron resonance (FTICR) mass spectrometers, cleaves N–Cα bonds of a peptide backbone

more or less evenly and cleavages are less dependent on amino acid sequence than in CID. Often labile post-translational modifications are retained after ETD of the backbone, making it possible to localize the modified amino acid residue. Each ion trap module in the FTICR-ion trap cluster therefore contains one ion trap equipped for ETD.

The resolving power and dynamic range of the FTICR is taken full advantage of for quantitative peptide and protein measurements. The system is ideal for large-scale quantitative proteomic studies, using labelfree quantitation or stable-isotope labeling methods such as SILAC[25] or multiplexed ^{15}N-labeling.[22] Despite its performance, the cluster has some limitations. For instance, it is not ideally suited for iTRAQ[26, 27] measurements, as these require MS/MS for the relative quantitation. The use of capillary rather than nanoflow LC reduces absolute sensitivity, but we have chosen this option as the chromatography systems and ESI sources require considerably less maintenance than in typical nanoflow systems. This robustness is essential in large-scale studies. Data-dependent precursor ion selection based on accurate mass in real time is also not possible in the cluster. However, it is possible to first perform the FTICR analysis and then use the accurate mass information to construct so-called scheduled precursor lists (inclusion criteria based on time as well as *m/z*) for one or more of the ion traps. For instance, species suspected to be phosphorylated based on accurate mass measurement[28] could be targeted for MS/MS using ETD, or a combination of CID and ETD. Moreover, the loose coupling of the instrumentation and nonconcurrent (or not necessarily concurrent) acquisition of MS and MS/MS data allows great freedom in constructing and exploring novel schemes for datadependent acquisition, as the analyses of the MS data dos not have to be performed in real-time and integrated into the instrument control software.

To appreciate the QAMT analysis mode, one can consider the following experiment. Assume n replicates of b biological samples or experimental conditions are collected, in total $n \times b$ samples to be analyzed on a cluster with c ion trap modules. Each sample and replicate is digested by trypsin and analyzed by LC-FTICR using reversed-phase chromatography only. In parallel, the proteins or digests of the replicates of each type of biological sample are pooled and fractionated by, for instance, SDS-PAGE (proteins) or strong cation exchange chromatography (peptides), into $c \times n$ fractions, for a total of $c \times n \times b$ fractions. For three ion trap modules and four biological replicates, 12 fractions would be collected for each of the b types

of biological sample. These fractions are subsequently analyzed on one of the ion trap modules in the cluster. In such a case, the same length chromatographic gradients can be used with the FTICR and the ion traps, with the total analysis time on the FTICR exactly matching that of the ion traps. For example, with 60-min LC methods and three ion trap modules, the independent quantitative analysis of four biological replicates of six different conditions or time points with 12 SCX peptide fractions collected for each takes 24 h, generating 24 LC-FTICR analyses for quantitation, and 72 ion trap LC-MS/MS datasets for identification. The peptides are quantified in a similar manner as in the AMT protocol[6] by integrating the area under the LC-MS peak in the FTICR data for the major peak in a narrow m/z range and the predicted retention time window. All "AMT tags" are confirmed by MS/MS on one of the ion traps on at least one similar sample from the same study, although not necessarily for every individual sample, treatment, or time point. This also means that peptides (and consequently, proteins) can be quantified at much lower levels than are needed for confident identification by MS/MS, which is also illustrated by the BSA measurements summarized in Figure 3.5. Under ideal LC conditions, the elution time of a peptide in the last, reversed phase dimension does not depend on which protein or peptide separation or separations were used in the prior dimensions, e.g., which SCX fraction is analyzed, and the elution times in the last (reversed-phase) dimension on the LC-ion trap systems can be aligned with those in the only (reversed-phase) LC separation with the FTICR. This QAMT scheme is feasible as the 12 T FTICR has sufficient resolving power and dynamic range for the $c \times n$-fold higher sample complexity compared to the ion trap LC-MS/MS. Analogous schemes can be constructed for any number and type of peptide or protein fractionation before the final reversed-phase separation, for instance, proteins can be fractionated by SDS-PAGE and digested for identification by LCMS/MS while individual samples are analyzed by LC-FTICR MS only.

As illustrated by the *E. coli* diauxie example, the QAMT mode is particularly useful for comparison of relatively similar samples, such as a series incorporating different time points, treatments, or experimental replicates. It is then feasible to use a two-dimensional separation before MS/MS, where the second dimension is of the same type, e.g., reversed-phase, as the one used as the only LC dimension for FTICR-MS. A high-field FTICR mass spectrometer is capable to resolve and detect more than 10,000 peptides in a relative short chromatographic separation, and the

additional dimension of separation aids the identification in the ion traps. Complex peptide mixtures have even been analyzed by direct infusion and high-field FTICR, i.e., without prior separation.[29, 30] For instance, we recently demonstrated that even by direct infusion, it is possible to detect and resolve most peptides in a combinatorial library with more than 1000 unique elemental compositions spanning a factor 36 in concentration.[31] For a very large number of biological replicates, the scheme becomes similar to the AMT tag protocol developed by Smith *et al.*,[7, 32, 33] where the "AMT tags" are verified at least once by MS/MS, placed in a database, and aligned through normalization of the retention times with the accurate mass data from an FTICR.[9] A particular advantage of the AMT or QAMT mode is that it is still possible to quantify peptides and proteins in samples where they are present at a lower concentration than would be required to produce good MS/MS data, which is required to generate any quantitative information in many other methods such as iTRAQ or spectral counting.[21, 34]

The FTICR-ion trap cluster provides quantitative proteomics data of a similar quality with comparable throughput to that of multiple hybrid ion trap-FTICR or ion trap-Orbitrap instruments at lower cost and infrastructure requirements. The instrument cluster has a few limitations, but in turn opens up additional possibilities for data-dependent MS/MS acquisition, and can serve as a test bed for the design and development of hybrid instruments with a single accurate mass analyzer and multiple ion traps for MS/MS. The cluster design and idea of chromatographic compression between LC-MS and LC-MS/MS is not limited to the use of a high-field FTICR and six ion traps, but may also be applicable to the combination of one MS-only TOF or Orbitrap with one or more MS/MS-capable mass spectrometers, albeit with lower mass accuracy and resolving power than any high-field FTICR instrument. The cluster scheme provides an inexpensive means to adding accurate mass capability in laboratories already operating one or more ion trap instruments for LC-MS/MS. All components in the cluster are individually exchangeable, which provides a high degree of flexibility that can be used to continuously upgrade the system. For instance, during the first year of operation, the 12 T FTICR front-end was upgraded from a previous apex ultra model to the recently introduced solariX.

ACKNOWLEDGMENTS

The authors thank Bart Schoenmaker, Rico Derks, Hannah Scott, and Rene van Zeijl for technical assistance, and Paul Hensbergen for helpful comments on the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2010.02.001.

REFERENCES

1. March, R. E., An introduction to quadrupole ion trap mass spectrometry. *Journal of Mass Spectrometry* **1997,** 32, (4), 351-369.
2. Morris, H. R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R. S.; Hoyes, J.; Bateman, R. H., High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry* **1996,** 10, (8), 889-896.
3. Makarov, A.; Denisov, E.; Kholomeev, A.; Baischun, W.; Lange, O.; Strupat, K.; Horning, S., Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry* **2006,** 78, (7), 2113-2120.
4. Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F., Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **2004,** 3, (3), 621-6.
5. O'Connor P, B.; Pittman, J. L.; Thomson, B. A.; Budnik, B. A.; Cournoyer, J. C.; Jebanathirajah, J.; Lin, C.; Moyer, S.; Zhao, C., A new hybrid electrospray Fourier transform mass spectrometer: design and performance characteristics. *Rapid Commun Mass Spectrom* **2006,** 20, (2), 259-66.
6. Pasa-Tolic, L.; Masselon, C.; Barry, R. C.; Shen, Y.; Smith, R. D., Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* **2004,** 37, (4), 621-4, 626-33, 636 passim.
7. Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D., Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Analytical Chemistry* **2003,** 75, (5), 1039-48.
8. Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W. J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; Camp, D. G., 2nd; Smith, R. D., Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Analytical Chemistry* **2006,** 78, (14), 5026-39.
9. Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R.; Adkins, J. N.; Smith, R. D., Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Analytical Chemistry* **2006,** 78, (21), 7397-409.
10. Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R., Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J Am Soc Mass Spectrom* **2007,** 18, (10), 1835-43.
11. Jacob, F.; Monod, J., Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **1961,** 3, 318-56.
12. Traxler, M. F.; Chang, D. E.; Conway, T., Guanosine 3',5'-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in

Escherichia coli. *Proc Natl Acad Sci U S A* **2006,** 103, (7), 2374-9.

13. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.

14. Palmblad, M.; Bindschedler, L. V.; Gibson, T. M.; Cramer, R., Automatic internal calibration in liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry of protein digests. *Rapid Commun Mass Spectrom* **2006,** 20, (20), 3076-80.

15. www.matrixscience.com

16. Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J., OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003,** 3, (8), 1454-63.

17. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **2004,** 101, (26), 9528-33.

18. Leinenbach, A.; Hartmer, R.; Lubeck, M.; Kneissl, B.; Elnakady, Y. A.; Baessmann, C.; Muller, R.; Huber, C. G., Proteome analysis of Sorangium cellulosum employing 2D-HPLC-MS/MS and improved database searching strategies for CID and ETD fragment spectra. *J Proteome Res* **2009,** 8, (9), 4350-61.

19. Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Molecular & Cellular Proteomics* **2006,** 5, (7), 1326-1337.

20. Bakalarski, C. E.; Haas, W.; Dephoure, N. E.; Gygi, S. P., The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics. *Analytical and Bioanalytical Chemistry* **2007,** 389, (5), 1409-1419.

21. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics* **2005,** 4, (9), 1265-72.

22. Palmblad, M.; Mills, D. J.; Bindschedler, L. V., Heat-shock response in Arabidopsis thaliana explored by multiplexed quantitative proteomics using differential metabolic labeling. *J Proteome Res* **2008,** 7, (2), 780-5.

23. Nevedomskaya, E.; Derks, R.; Deelder, A. M.; Mayboroda, O. A.; Palmblad, M., Alignment of capillary electrophoresis-mass spectrometry datasets using accurate mass information. *Analytical and Bioanalytical Chemistry* **2009,** 395, (8), 2527-33.

24. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **1998,** 120, (13), 3265-3266.

25. Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* **2002,** 1, (5), 376-386.

26. Ross, P. D., S.; Pillai, S.; Daniels, S.; Williamson, S.; Guertin, S.; Minkoff,

M.; X., C.; Purkayastha, B.; Pappin, D. *Proceedings of the 54th ASMS Conference on Mass Spectrometry*, Seattle, WA, May 28–June 1, 2006; Seattle, WA, 2006.

27. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics* **2004,** 3, (12), 1154-1169.

28. Spengler, B.; Hester, A., Mass-based classification (MBC) of peptides: highly accurate precursor ion mass values can be used to directly recognize peptide phosphorylation. *J Am Soc Mass Spectrom* **2008,** 19, (12), 1808-12.

29. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* **2000,** 11, (4), 320-32.

30. Palmblad, M.; Wetterhall, M.; Markides, K.; Hakansson, P.; Bergquist, J., Analysis of enzymatically digested proteins and protein mixtures using a 9.4 Tesla Fourier transform ion cyclotron mass spectrometer. *Rapid Commun Mass Spectrom* **2000,** 14, (12), 1029-34.

31. Palmblad, M.; Drijfhout, J. W.; Deelder, A. M., High resolution mass spectrometry for rapid characterization of combinatorial peptide libraries. *J Comb Chem* **2010,** 12, (1), 65-8.

32. Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D., Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J Am Soc Mass Spectrom* **2003,** 14, (9), 980-91.

33. Zimmer, J. S.; Monroe, M. E.; Qian, W. J.; Smith, R. D., Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* **2006,** 25, (3), 450-82.

34. Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G., Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **2005,** 4, (10), 1487-502.

4

Cloud Parallel Processing of Tandem Mass Spectrometry-based Proteomics

Yassene Mohammed,[1,2] Ekaterina Mostovenko,[1] Alex A. Henneman,[1] Rob J. Marissen,[1] André M. Deelder,[1] and Magnus Palmblad[1]

[1] Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

[2] Distributed Computing Security Group and L3S, University of Hannover, Germany

Journal of Proteome Research 2012, 11 (10), 5101–5108

ABSTRACT

Data analysis in mass spectrometry-based proteomics struggles to keep pace with the advances in instrumentation and the increasing rate of data acquisition. Analyzing this data involves multiple steps requiring diverse software, using different algorithms and data formats. Speed and performance of the mass spectral search engines are continuously improving, although not necessarily as needed to face the challenges of acquired big data. Improving and parallelizing the search algorithms is one possibility, data decomposition presents another, simpler strategy for introducing parallelism. We describe a general method for parallelizing identification of tandem mass spectra using data decomposition that keeps the search engine intact and wraps the parallelization around it. We introduce two algorithms for decomposing mzXML files and recomposing resulting pepXML files. This makes the approach applicable to different search engines, including those relying on sequence databases and those searching spectral libraries. We use cloud computing to deliver the computational power and scientific workflow engines to interface and automate the different processing steps. We show how to leverage these technologies to achieve faster data analysis in proteomics and present three scientific workflows for parallel database as well as spectral library search using our data decomposition programs, X!Tandem and SpectraST.

INTRODUCTION

Mass spectrometry (MS), particularly tandem mass spectrometry (MS/MS), is currently the most used method for identifying unknown proteins present in biological samples. Advances in instrumentation have reduced acquisition time and increased resolution and sensitivity, which in combination with complementary fragmentation mechanisms[1, 2] and high resolving-power mass analyzers in both MS and MS/MS[3-5] have led to very complex data. This has brought new challenges to proteomics, i.e. how do we store and process these large data volumes. Standard desktop computers often cannot process data at the rate it is being generated, creating an additional bottleneck in the analysis pipeline. The analysis of the mass spectrometry data typically involves several steps. One essential and computationally expensive step is peptide identification, i.e. the mapping of each spectrum to a unique peptide or one or more peptides. In this manuscript we describe a method of handling mass spectrometry "big data" by outsourcing computationally intensive tasks using off-the-shelf open source tools and in-campus cloud resources. We introduce a method for parallelizing common search engines like X!Tandem and SpectraST that are part of the Trans-Proteomic Pipeline (TPP)[6], which can also work for most other available search engines. We show how peptide identification speed using workflow engines, cloud computing, and a new data decomposing/recomposing algorithm can easily be improved by a significant factor. In our tests we reached more than 30-fold speed improvement comparing X!Tandem running locally (one core) with the same program running on the cloud and a 7-fold improvement for the SpectraST spectral library search.

METHODS

One important step in the processing pipeline of mass spectrometry data is associating a particular (tandem) mass spectrum with a peptide sequence. There are three types of search engines for peptide identification, i.e. database, library, and *de novo*. Database search engines, like Mascot[7], SEQUEST[8], or X!Tandem[9], compare each spectrum obtained from the sample with theoretical spectra generated from a list of predicted peptides. The predicted peptides list is ideally derived from all of the protein sequences that could be expressed in the experiment sample. Library search

engines, like SpectraST[6] or X!Hunter[10] assume that the fragmentation of a particular molecule in a mass spectrometer is partially reproducible between analyses and instruments. One can therefore generate a library of ion fragmentation spectra with each spectrum being associated with a corresponding molecular structure. A library search engine assigns a specific structure to an experimental spectrum by comparing it with the entries in the library.

Peptide identification using a search engine is a main processing bottleneck in mass spectrometry based proteomics. A normal search of 30,000 spectra could take up to 40 minutes on common modern desktop with a 4-core processor, depending on the search parameters. In many cases this is impractical for scientists, especially if they want to include more modifications in their searches, which can significantly increase the search space. Enhancing search engine speed besides developing search algorithms for high performance computing environment are continuously under development.[11-14] While making faster algorithms is a main objective of several groups[14-17], we are only targeting the data itself leaving the search engine intact. This makes the approach applicable to many search engines. Search engines are legacy software that have gained acceptance and usability in the proteomics community and we therefore prefer to consider them as black boxes and not modify them in any way, but instead wrap the parallelization around them. In the rest of the section we describe the data formats used, the new decomposition and recomposition algorithms, the processing pipelines and how to scale these up to scientific workflows.

Data formats

To build on other efforts, such as the TPP[6], we chose to use common XML formats such as mzXML[18, 19] for input and pepXML[20] as output. mzXML and pepXML are two *de facto* open format standards still used for mass spectrometry data. Converters from almost any other format to mzXML or pepXML can be obtained.[19-21] Extension of our method to mzML[22] is also feasible, as only the logic in the data decomposition algorithm needs to be modified with no further changes. Using open standard formats maintains compatibility with other efforts and existing pipelines and avoids making this work an isolated solution.

Data decomposition and recomposition

Decomposition involves breaking down a complex system into smaller pieces. It is the basis for finding the tasks that can run concurrently in parallel applications. There are two major decomposition methods in parallel programming, i.e. functional and data decomposition.[23] Data decomposition is used more often and it depends mainly on the developer's knowledge about the data and how an algorithm processes the data. In order to facilitate parallelism of peptide identification of mass spectrometry data we developed two algorithms for decomposing and recomposing the inputs and outputs of an arbitrary search engine. The only assumption made is that each spectrum will be processed by the search engine independently from other spectra. This is true for many search algorithms, but not subsequent validation steps, such as PeptideProphet[24] and Percolator.[25] However, the latter are not nearly as computationally expensive as the initial peptide-spectrum matching. The search engines we used to demonstrate our parallelization approach, i.e. X!Tandem without model refinement[26] and SpectraST process each spectrum independently. OMSSA[27], MS-GFDB/MS-GF+[1] and Crux/Tide[16, 17] are other search engines that could also be parallelized in this way.

Processing Pipelines and Scientific Workflows

There are multiple software packages that allow stepwise processing of mass spectrometry data, such as TPP[6], Proteomatic[28] and ProteoWizard.[29] In this sense, processing pipelines and workflows are overloaded terms, and sometimes used synonymously. We use processing pipelines to refer to a multistep sequential processing of one dataset at a time, in which transitions from one step to the next happen with some manual interaction as in the TPP. Scientific workflows involve concurrency and parallel processing capabilities, in which the transition between the processing steps can happen automatically or with breakpoints according to the workflow design. Scientific workflow engines like Galaxy[30], Moteur[31], Kepler[32], and Taverna[33] were introduced in the last decade to facilitate interfacing modular processing steps, automating analysis pipelines, scaling them up to workflows, and make analyses reproducible and sharable. We have previously described[34] how Taverna can be used to automate analysis workflows in mass spectrometry based proteomics on a local machine. We also demonstrated how workflow and data decomposition can scale up processing pipelines to run in high performance computing environments.[23]

Here we use Taverna 2.4 to build our processing workflows and to perform job orchestration, i.e. to manage data and software transfer to and from the cloud. In this respect, we use Taverna not only as a workflow manager, but also as a technical enabler to build our *adhocratic*[35, 36] experiment oriented distributed computing environment using in-campus clouds. Taverna offers various kinds of processors.[33, 37] Scientists can chose between WSDL web services, Beanshell processors, REST Web services, Rshell processors, Tools and XPath processors. Details about these processors and how to use them can be found in literature[33] as well as in the Taverna documentation.[37] In the following we highlight the two processor types that are important for our implementation.

Beanshell processors enable executing small Java code snippets as part of a workflow. Typically they are used for small tasks like simple file and data manipulation, parsing and formatting, saving to a local directory, calling local program, interacting with the user, etc. Tool processors are very suitable to call commands in a shell on any machine, to which Taverna can obtain an SSH connection - including the local machine. We mainly use Beanshell processors to launch software with their correct inputs locally, and Tool processors to interact with the cloud resources, upload data, and retrieve results. We used cloud resources based on the open source cloud middleware OpenNebula.[38] These cloud resources are freely available for academic research users in the Netherlands. Such resources are common in various universities. A cloud environment in regard to our method can include any machine, to which Taverna could have an SSH connection.

Used Datasets for Testing

In order to profile our method and compare it with the local run of the used search engines we ran multiple tests from realistic database search scenarios. For these tests we used two ion trap datasets; the first consisted of 5 LC-MS/MS datasets from tryptically digested human serum samples and the second of LC-MS/MS data from 20 fractions of one *E. coli* whole cell lysate, also digested with trypsin. All data was acquired on amaZon ion trap mass spectrometer (Bruker Daltonics, Bremen, Germany). The five human datasets each contains around 27,000 spectra whereas the 20 *E. coli* datasets each contains around 10,600 spectra (see Table 4.1 and Figure 4.6). In the X!Tandem search, strict tryptic cleavage specificity were assumed (C-terminally or R and K, not N-terminally of P), the precursor mass measurement error tolerance -0.5 to 2.5 Da, 2 missed enzymatic cleavage

allowed, and carbamidomethylation as the only and fixed modification. Phosphorylation as variable modification and semi-tryptic cleavage were also considered in the performance tests. In the SpectraST search, average masses instead of monoisotopic masses were used and precursor mass measurement error tolerance of 3 Th. All other parameters for X!Tandem and SpectraST were as the defaults in the TPP package. For the X!Tandem

	One sample (human)		5 samples (human)		20 samples (*E. coli*)	
Size of file(s)	113.8 MB		565.8 MB		1,540 MB	
Number of spectra	27,436		139,211		212,141	
Search engine	**X!Tandem**	**SpectraST**	**X!Tandem**	**SpectraST**	**X!Tandem**	**SpectraST**
Size of database/ library	35.6 MB[40]	2,123 MB[41]	35.6 MB[40]	2,123 MB[41]	1.75 MB[39]	303 MB[42]
Number of protein/spectra entries	70,254[40]	310,688[41]	70,254[40]	310,688[41]	4,303[39]	50,369[42]
Wall time monolithic running locally[1] (1 core / 4 cores) in min:sec	39:32 / 10:17	27:43	191:53 / 55:13	52:32	40:16/ 28:09	43:22
Wall time[2] parallel running on the cloud in min:sec	2:15	4:17	5:42	7:09	10:34	11:31
Speedup in fold	18 / 4.6	6.8	34 / 10	7	3.8/2.7	3.8

[1] The used system to run all the local experiments was an HP Elite 8200 computer with Windows® 7 Enterprise 64bit operating system, Intel® i7-2600 processor running at 3.40 GHz, and 8 GB of RAM.

[2] Wall time here refers to the actual time experienced by the user, i.e. the time needed to decompose, transfer, analyze and recompose data, starting with the spectra in mzXML file(s) on the user local computer and ending with the peptide identification in pepXML format stored in the same directory as the mzXML file.

Table 4.1. Performance tests of the described method comparing elapsed time for analyzing multiple input datasets.

search, the used databases for the human serum and for *E. coli* datasets were retrieved from UniProt.[39, 40] The spectral libraries for human and *E. coli* from National Institute of Standards and Technology (NIST) were used for the SpectraST searches.[41, 42]

Related Work

Duncan et al. have developed a parallel version of a X!Tandem for Message Passing Interface (MPI) enabled cluster.[11] It is beneficial to run a search engine on a cluster using MPI in terms of speed; this demands anyhow the availability of an MPI enabled server/cluster to the scientist. Pratt et al.[13] developed a cloud parallel peptide identification using parallel X!Tandem[11], Hadoop[43, 44], and MapReduce.[45, 46] They used a similar approach to X!!Tandem[12] in extending X!Tandem's threading onto a network, but used Hadoop and MapReduce instead of MPI. Their implementation is meant for Amazon Elastic Cloud and they achieved speedup of 31-fold using 200 Amazon cloud instances (corresponding to processing unit or a core). The current TPP version allows outsourcing X!Tandem searchs to Amazon Elastic Cloud to run multiple searches at the same time. We are not aware of any parallel implementation of SpectraST, but Baumgardner et al. have recently implemented their own spectral library search algorithm for GPUs using CUDA.[14] Our goal is to achieve data parallelism to accelerate peptide identification while preserving the search engine without any modification to its code. In principle, this makes the solution compatible also with closed-source algorithms.

RESULTS AND DISCUSSION

The employed technologies can be divided into three categories: data decomposition, cloud computing and scientific workflow engines. Data decomposition/recomposition is the parallelization *enabler*. The virtual and physical computers in the cloud delivers the processing and storage power. Finally, scientific workflows are used to imbed the logic of the data analysis into interfaced processing steps, to scale analysis pipelines up to workflows, and to orchestrate the parallel processing. In the following we explain how we are leveraging these technologies in our implementation.

Data decomposition and recomposition algorithms

Our decomposition algorithm splits an mzXML file into multiple smaller syntactically correct mzXML files. Syntactically correct here means that each daughter file is itself a valid mzXML file according to the mzXML schema.[19] The requested number of daughter files is passed to the algorithm as an input. Typically, LC-MS or LC-MS/MS datasets incorporate many low quality (information-poor) spectra; particularly at the beginning and near the end of the chromatographic gradient, while the good (information-rich) spectra are concentrated in the middle of the chromatographic run. Simply dividing the data in equal and sequential time intervals would therefore be suboptimal, as the early and late time intervals contains many spectra that would be immediately filtered out by the search engine. These data subsets would therefore process much faster than subsets from the middle of the gradient. To avoid this, we designed the algorithm to distribute the spectra from the original mzXML file randomly to all daughter files. This is an ad hoc approach to distribute good and bad spectra in order to divide the computational load evenly over the processing nodes. This also makes the method scaleable and independent on the chromatographic gradient and experimental design. Our data recomposition algorithm takes multiple pepXML files and composes them into one pepXML file. The algorithm takes into account the different original naming of the file and corrects the scan numbers to make the composed pepXML file schematically correct.[20] Both algorithms are written in Java and are available on ms-utils.org/decomposition.

Cloud computing

We used a dedicated infrastructure for cloud computing at SARA.[47] The infrastructure runs on OpenNebula cloud middleware. The instances/workers we used were minimal Ubuntu 11.04 server 64-bit virtual machines with Oracle (Sun) Java 6 build 1.6.0_26 installed. Depending on the workflow, a number of identical images can be initiated and used. For our tests we always used 8 instances, each of 8 virtual CPUs. Currently starting the workers from the workflow using OpenNebula Cloud Computing Interface services[48] is not permitted due to the security policy of the provider. All necessary software to run a workflow, for instance the search engines, will be deployed on the target machine from within the workflow. This keeps the cloud instances lightweight and the workflows easier to update and adjust to the target cloud architecture. In case one

prefers another version of the search engine, or using a 32-bit server, only the corresponding executable has to be provided as an input to the workflow.

Scientific workflows

The minimal workflow consists of three main processors: the mzXML decomposer, a search engine, and the pepXML composer (see Figure 4.1 and 4.2). One extra processor is needed to uncompress (unzip) the downloaded data from the cloud. Moving compressed (zipped) data between the cloud and the local machine and vice versa reduces the latency regarding the network speed. This is very helpful when the data is in ASCII format and can be compressed down to 68% of its original size like in mzXML and pepXML formats. The NIST spectral libraries can be compressed down to 32% of their original size. Each workflow processor includes the needed logic to run the corresponding program from the command line. The firing mechanism in Taverna is the availability of the data on the inputs of each processor. Taverna takes care of transferring the data between the processors. The data decomposing/recomposing processors are Beanshell

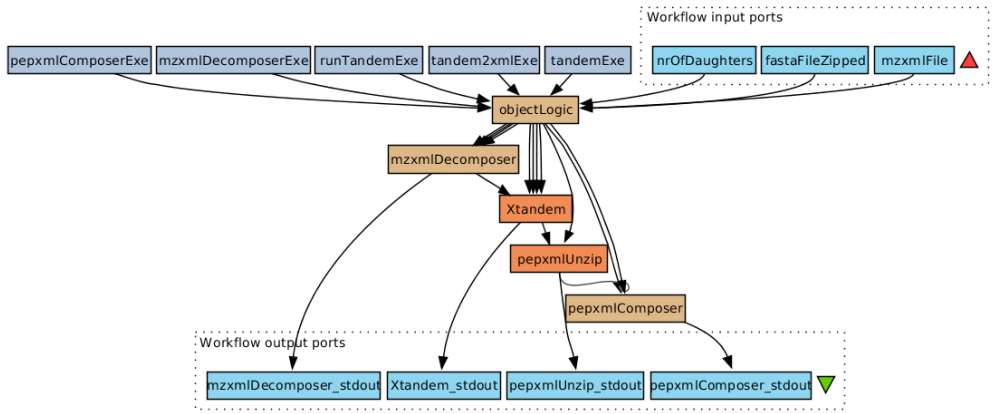


Figure 4.1. A scientific workflow for searching LC-MS/MS mass spectrometry data using X!Tandem on the cloud. The workflow consists of 5 processors. The *objectLogic* processor prepares all inputs in the right format, i.e. keeping or converting strings into file object according to the following processor. The *mzxmlDecomposer* and *pepxmlComposer* run the decomposing/recomposing algorithms. *objectLogic*, *mzxmlDecomposer* and *pepxmlComposer* are Beanshell processors and they run locally. *Xtandem* runs X!Tandem on a remote machine and *pepxmlUnzip* unzip the pepXML files to a local directory; both are Tool processors.

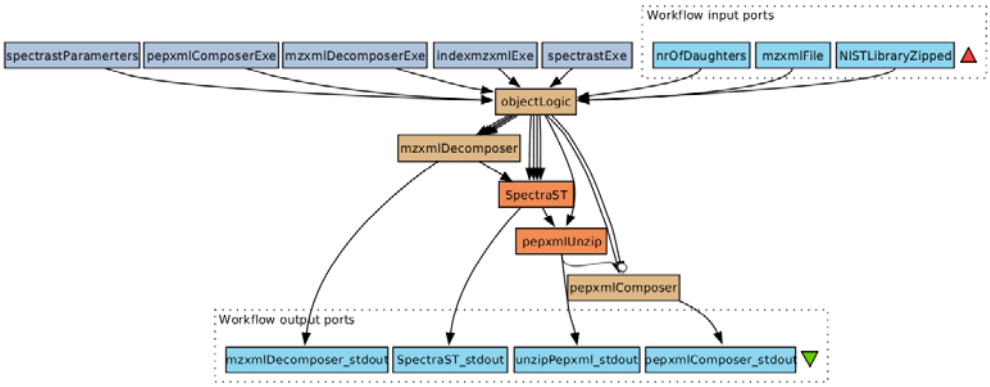


Figure 4.2. A scientific workflow for searching LC-MS/MS data using SpectraST on the cloud. The processor *mzxmlDecomposer*, *pepxmlUnzip* and *pepxmlComposer* are identical to the one in the X!Tandem workflow (Figure 5.1). The only difference is that the *Xtandem* processor is exchanged with the *Spectrast* processor and the constant inputs are adjusted to SpectraST. This approach is also possible for other search engines as described in the *Data decomposition and recomposition* paragraph

processors and run locally. The search engine is a tool processor and runs on the cloud. Taverna stores the IP addresses and passwords of the cloud worker nodes in its credential management. The password repository is protected with a master password, i.e. the user need to authenticate only once when starting Taverna.

Figure 4.1 shows a workflow to run X!Tandem on the cloud. The workflow takes the mzXML file(s), zipped search data base file in FASTA format and the number of the daughter mzXML files as inputs. Ideally the number of the daughter files is an integer factor of the available cloud workers. The search engine parameters are included in the *runTandemExe* processor. Figure 4.2 shows a simple scientific workflow to run SpectraST on the cloud. Similarly, the workflow takes the mzXML file(s), the zipped search library files (including the .splib, .spidx and .pepidx files) and the number of daughter mzXML files as inputs. SpectraST search parameters are included in the *spectrastParameters* processor, which is a string and is adjustable for different experiments. The processing logic of both workflows is very similar. The decomposition, recomposition and unzip pepXML processors are identical. The search engine calling processors are adjusted to each search engine, but are still logically very similar. This processor can be readjusted for other search engines. It is sometimes beneficial to separate the preprocessing/decomposing of the mzXML files from the logic of

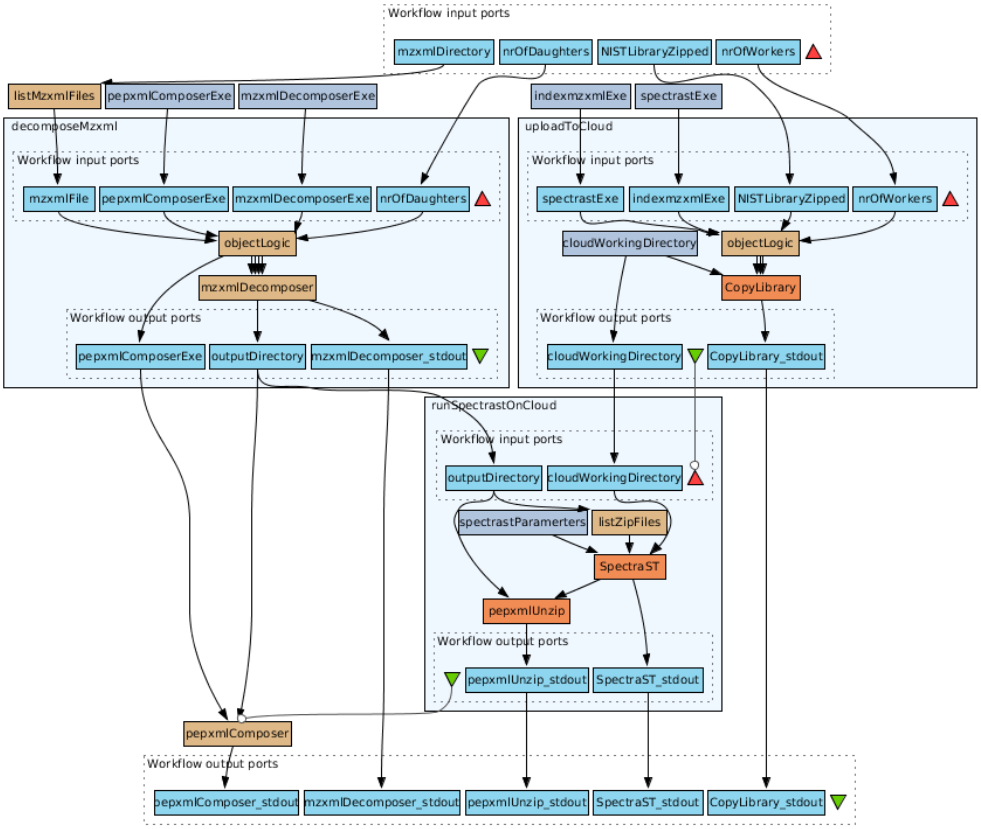


Figure 4.3. An advanced scientific workflow for searching LC-MS data using SpectraST on the cloud. Uploading the libraries is optimized to achieve better performance, which makes this workflow more suitable for processing mzXML spectra files from human samples, as the corresponding NIST library needed by SpectraST is larger than 2 GB. Here we connect 3 nested workflows, in which the first 2, i.e. *decomposeMzxml* and *uploadToCloud* run in parallel while the third nested workflow, i.e. *runSpectrastOnCloud* will start only if *uploadToCloud* finished all iteration. *runSpectrastOnCloud* and *decomposeMzxml* can still run in parallel.

uploading big data like the NIST human spectral libraries[41, 42] for SpectraST. Figure 4.3 illustrates an advanced workflow for SpectraST, in which the needed library and executables for the processing are uploaded simultaneously while decomposing the input mzXML files. The three workflows are available from ms-utils.org/cloud.

Speed performance comparison

We compared the elapsed wall clock time needed to analyze one file of the human dataset, the whole human data set, and the whole *E. coli* dataset on a local workstation and on the cloud using our method. The results are summarized in Table 4.1 and Figure 4.6. In the simplest scenario of analyzing one mzXML file we achieved speedup of 11-fold in case of X!Tandem running on single core and of 6-fold in case of SpectraST.

In order to exploit our implementation and test it for possible future big data challenges, we used spectra from fractioned sample to construct a single large mzXML file of 213,788 spectra. We profiled the number of cores in relation to the elapsed wall clock time needed to process these spectra and the results are illustrated in Figure 4.4. We were able to perform the peptide identification using X!Tandem and 8 cloud machines each of 8 processors within 12 min, a 26-fold faster than running it on a single core machine. When allowing phosphorylation as a variable modification, it was possible to obtain identical results within 72 minutes using 64 CPUs, or 5 hours on a single 8 CPU cloud node (comparable to an 8 CPU local machine).

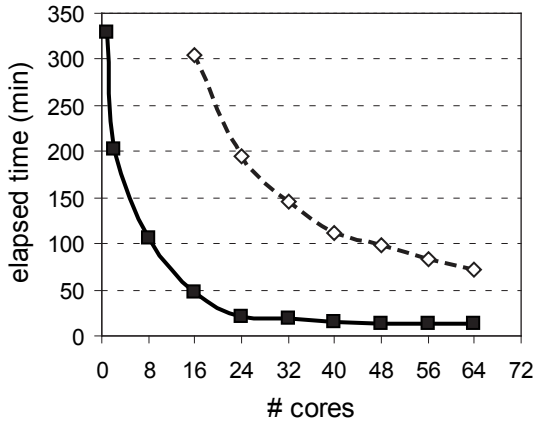


Figure 4.4. The wall time needed to search a large input file against a human sequence database[40] as a function of the number of cores with only fixed modifications (solid, squares) and with phosphorylation as variable modification (dashed, diamonds). By implementing data parallelism on the workflow level, it was possible to process 213,788 spectra in a 1.3 GB mzXML file with a search window of -0.5 to 2.5 Da in 12 minutes with only fixed modifications, a job which would take more than 5 hours on a desktop computer. When allowing phosphorylation (on serine, threonine and tyrosine) as a variable modification, the search took 72 minutes using 64 processors.

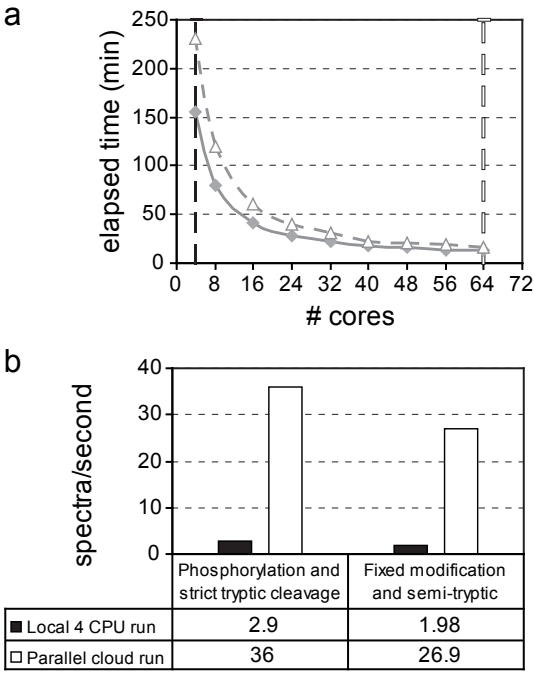
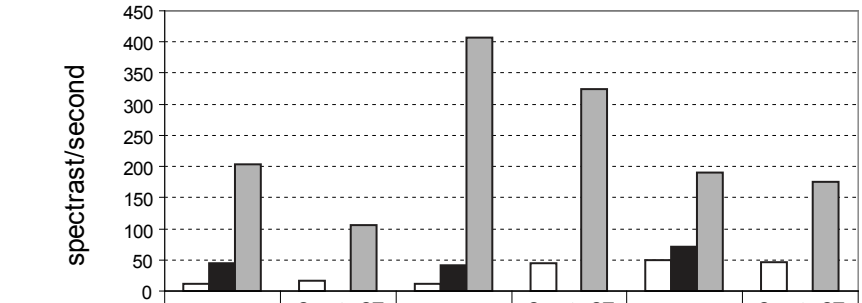


Figure 4.5. Comparison of different search times with either variable modification or semi-specific cleavage (a). The X!Tandem workflow was used to search a human dataset of 27,436 spectra against the human sequence database,[40] with strict tryptic cleavage and allowing for phosphorylation (dashed gray, triangles), and with semi-tryptic cleavage and only fixed modifications (solid gray, diamonds) as variable modification. Speed improvement of parallel processing in the 64-CPU cloud compared to 4 local CPUs of the same searches (b).

To evaluate the performance in common practice, where the enzyme fidelity is not known a priori or other proteases may have been active, we ran a series of tests with semi-specific cleavage on a smaller set of 27,000 spectra and measured a 27-fold increase in speed (see Figure 4.5). When allowing three variable PTMs, the 64 CPU cloud finished searching these spectra 36 times faster than a 4-core local machine.

CONCLUSIONS

In data mining, data decomposition is considered the "most useful form of transformation of datasets".[49, 50] With our approach of wrapping data parallelism via decomposition and recomposition around the search engine, we were able to parallelize more than one peptide identification software. We demonstrated this using a common database search engine - X!Tandem - and a spectral library search engine - SpectraST. We achieved the parallelism by an ad hoc approach using off-the-shelf open source software for scientific workflow, i.e. Taverna workbench, and OpenNebula for cloud computing. We believe that such adhocratic cloud implementations can scale with future needs to handle big data in mass spectrometry based



	One sample (human)		5 samples (human)		20 samples (*E. coli*)	
	X!Tandem	SpectraST	X!Tandem	SpectraST	X!Tandem	SpectraST
Local run 1 CPU	11	16	12	44	50	46
Local run 4 CPUs	44	n/a	42	n/a	71	n/a
Parallel cloud run	203	106	407	324	191	175

Figure 4.6. The performance of the cloud compared to local runs for the same search engines and data. Three experiments are compared, the details for which are listed in Table 5.1.

proteomics. A single large mzXML file of 1.3 GB containing 213,788 spectra was searched using our cloud parallel X!Tandem in 12 min. Compared to other parallel implementation of search engines like the Message Passing Interface (MPI) enabled parallel X!!Tandem, or the Hadoop MapReduce deployment on Amazon web services – MR-Tandem, our method does not require dedicated MPI hardware or rewriting of the search algorithm. We designed our method to be generally applicable to any software that searches spectra independently and demonstrated this with X!Tandem and SpectraST. The method can possibly be used in combination with the other parallel programs. The decomposition/recomposistion algorithms and slightly modified workflows can then be used to distribute an mzXML file to multiple machines with the Hadoop MapReduce X!Tandem deployment or multiple machines with the MPI-parallel X!!Tandem. In this case the workflow can be modified to use the already installed search engine. In comparing speed performance by running identical searches on local machines and in parallel, our method achieved more than twice the increase in speed reported by the MPI-parallel X!!Tandem. Compared to the 31-fold speedup on 200 processors reported by the Hadoop MR-Tandem implementation, our method achieved 36-fold speedup on 64 processors.

To make the implementation useful to the research community, we used common standards for input and output. Furthermore, cloud instances from providers like Amazon or in-campus clouds can be used as long as they are associated with public IP addresses. In such cases our implementation can be used without modification. Where the instances have private IP

addresses, one can still launch the workflow from one of these instances without modification. Researchers can also build their own cloud environment by accessing accounts on different Linux machines without the need to install additional software; only Java is required, which is available for nearly all platforms. The decomposition and recomposition algorithms can be used in other scenarios, with or without clouds. For instance, when using a computer cluster or a computer with a multi-core CPU, the researcher can still use the data decomposition and recomposition with single-threaded algorithms such as SpectraST to gain parallelism.

We are currently working with the developers of scientific workflow managers and cloud providers to address different issues including starting and shutting down the virtual machines on the cloud entirely from within the workflow, using certificates authentication, and enhancing the security on the cloud. We are convinced the 36-fold speedup reported here is still not exploiting the available resources to their full potential and also work to further improve the acceleration of these algorithms using cloud.

ACKNOWLEDGMENTS

This work was supported by the Dutch Organization for Scientific Research (De Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO), grants NRG-2010.06, BG-043-11 and VI-917.11.398. Used cloud resources are part of the Dutch e-Science Grid – "BigGrid".

SUPPLEMENTARY MATERIAL

The used data decomposition and recomposition algorithms are written Java and are available from www.ms-utils.org/decomposition. The Taverna workflows are available from www.ms-utils.org/cloud and on www.myExperiment.org.

REFERENCES

1. Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J.; Pevzner, P. A., The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol Cell Proteomics 2010, 9, (12), 2840-52.
2. Swaney, D. L.; McAlister, G. C.; Coon, J. J., Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat Methods 2008, 5, (11), 959-64.
3. Resemann, A.; Wunderlich, D.; Rothbauer, U.; Warscheid, B.; Leonhardt, H.; Fuchser, J.; Kuhlmann, K.; Suckau, D., Top-down de Novo protein sequencing of a 13.6 kDa camelid single heavy chain antibody by matrix-assisted laser desorption ionization-time-of-flight/time-of-flight mass spectrometry. Anal Chem 2010, 82, (8), 3283-92.
4. Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S., Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 2011, 10, (9), M111 011015.
5. Frese, C. K.; Altelaar, A. F.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J.; Mohammed, S., Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J Proteome Res 2011, 10, (5), 2377-88.
6. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 2005, 1, 2005 0017.
7. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification bu searching sequence databases using mass spectromery data. Electrophoresis 1999, 20, (18), 3551-2567.
8. Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 1995, 67, (8), 1426-36.
9. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004, 20, (9), 1466-7.
10. Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C., Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 2006, 5, (8), 1843-9.
11. Duncan, D. T.; Craig, R.; Link, A. J., Parallel tandem: A program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem. Journal of Proteome Research 2005, 4, (5), 1842-1847.
12. Bjornson, R. D.; Carriero, N. J.; Colangelo, C.; Shifman, M.; Cheung, K. H.; Miller, P. L.; Williams, K., X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. J Proteome Res 2008, 7, (1), 293-9.
13. Pratt, B.; Howbert, J. J.; Tasman, N. I.; Nilsson, E. J., MR-Tandem: parallel X!Tandem using Hadoop MapReduce on Amazon Web Services. Bioinformatics 2012, 28, (1), 136-7.

14. Baumgardner, L. A.; Shanmugam, A. K.; Lam, H.; Eng, J. K.; Martin, D. B., Fast parallel tandem mass spectral library searching using GPU hardware acceleration. J Proteome Res 2011, 10, (6), 2882-8.

15. Pratt, B., GPU-ACCELERATED PEPTIDE SEARCH. In Funded by Department of Health and Human Services, 1R43HG006414-01: 2011.

16. Park, C. Y.; Klammer, A. A.; Kall, L.; MacCoss, M. J.; Noble, W. S., Rapid and accurate peptide identification from tandem mass spectra. J Proteome Res 2008, 7, (7), 3022-7.

17. Diament, B. J.; Noble, W. S., Faster SEQUEST searching for peptide identification from tandem mass spectra. J Proteome Res 2011, 10, (9), 3871-9.

18. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004, 22, (11), 1459-66.

19. Seattle Proteome Center/Institute for Systems Biology mzXML Format. http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML (June 13),

20. Seattle Proteome Center/Institute for Systems Biology pepXML Format. http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML (June 13),

21. List of free software for analysis of mass spectrometry data. www.ms-utils.org (June 13),

22. Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W., mzML--a community standard for mass spectrometry data. Mol Cell Proteomics 2011, 10, (1), R110 000133.

23. Mohammed, Y.; Shahand, S.; Korkhov, V.; Luyf, A. C. M.; Schaik, B. D. C. v.; Caan, M. W. A.; Kampen, A. H. C. v.; Palmblad, M.; Olabarriaga, S. D., Data Decomposition in Biomedical e-Science Applications. In IEEE 7th International Conference on E-Science, e-Science 2011, Workshop Proceedings, Stockholm, Sweden, 2011.

24. 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 2002, 74, (20), 5383-92.

25. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007, 4, (11), 923-5.

26. Craig, R.; Beavis, R. C., A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 2003, 17, (20), 2310-6.

27. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search

algorithm. J Proteome Res 2004, 3, (5), 958-64.

28. Specht, M.; Kuhlgert, S.; Fufezan, C.; Hippler, M., Proteomics to go: Proteomatic enables the user-friendly creation of versatile MS/MS data evaluation workflows. Bioinformatics 2011, 27, (8), 1183-4.

29. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008, 24, (21), 2534-6.

30. Goecks, J.; Nekrutenko, A.; Taylor, J., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010, 11, (8), R86.

31. Maheshwari, K.; Montagnat, J. In Scientific Workflow Development Using Both Visual and Script-Based Representation, Services (SERVICES-1), 2010 6th World Congress on, 5-10 July 2010, 2010; 2010; pp 328-335.

32. Altintas, I.; Berkley, C.; Jaeger, E.; Jones, M.; Ludascher, B.; Mock, S., Kepler: An Extensible System for Design and Execution of Scientific Workflows. In Proceedings of the 16th International Conference on Scientific and Statistical Database Management, IEEE Computer Society: 2004; p 423.

33. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P., Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004, 20, (17), 3045-3054.

34. de Bruin, J. S.; Deelder, A. M.; Palmblad, M., Scientific Workflow Management in Proteomics. Mol Cell Proteomics 2012.

35. Waterman, R. H., Jr., 'Adhocracy': lessons from the changemasters. Hospitals 1991, 65, (1), 56.

36. Waterman, R. H., Jr., Adhocracy W. W. Norton & Company: 1993; p 128.

37. Taverna Website. www.taverna.org.uk/ (June 13),

38. OpenNebula Website. www.opennebula.org/ (June 13),

39. Uniprot canonical sequence in FASTA format, obtained from www.uniprot.org on June 18, 2012 with the search string: “organism:Escherichia AND coli AND keyword:181 AND keyword:1185 AND reviewed:yes”.

40. Uniprot canonical sequence in FASTA format, obtained from www.uniprot.org on June 4, 2012 with the search string: “organism:"Homo sapiens" AND keyword:181”.

41. Eds. S.E. Stein and P.A. Rudnick, NIST Peptide Tandem Mass Spectral Libraries. E. coli Peptide Mass Spectral Reference Data, E. coli, ion trap, Official Build Date: April 20, 2012. National Institute of Standards and Technology, Gaithersburg, MD, 20899. Downloaded from http://peptide.nist.gov on June 18, 2012. In.

42. Eds. S.E. Stein and P.A. Rudnick, NIST Peptide Tandem Mass Spectral Libraries. Human Peptide Mass Spectral Reference Data, H. sapiens, ion trap, Official Build Date: May 26, 2011. National Institute of Standards and Technology, Gaithersburg, MD, 20899. Downloaded from http://peptide.nist.gov on June 6, 2012. In.

43. Apache Hadoop. http://hadoop.apache.org/ (June 13),

44. White, T., Hadoop: The Definitive Guide. O'Reilly Media, Inc.: Sebastopol, CA 95472., 2009.

45. Dean, J.; Ghemawat, S., MapReduce: simplified data processing on large clusters. Commun. ACM 2008, 51, (1), 107-113.

46. Taylor, R. C., An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 2010, 11.

47. SARA cloud. www.cloud.sara.nl (May 30),

48. Open Grid Forum, Open Cloud Computing Interface Specification. In 2009.

49. Kusiak, A., Decomposition in data mining: An industrial case study. Ieee Transactions on Electronics Packaging Manufacturing 2000, 23, (4), 345-354.

50. Maimon, O.; Rokach, L., Decomposition Methodology for Knowledge Discovery and Data Mining. In DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK, Maimon, O.; Rokach, L., Eds. Springer: New York, 2005; pp 981-1003.

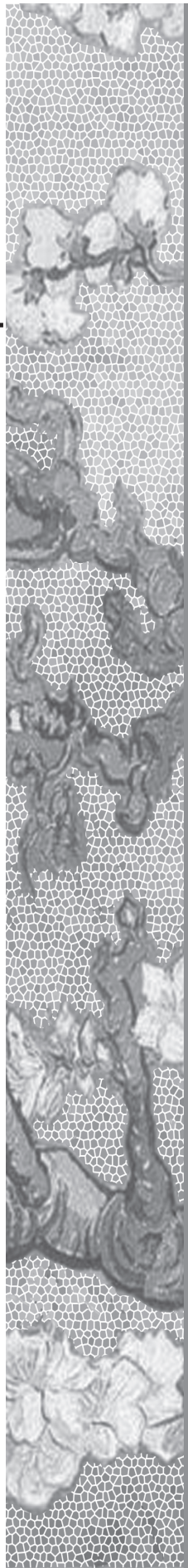
5

Protein Expression Dynamics during Escherichia coli Glucose-Lactose Diauxie

Ekaterina Mostovenko, André M. Deelder and Magnus Palmblad

Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

BMC Microbiology 2011, 11(1), 126



ABSTRACT

Background: *Escherichia coli* is a well-studied anaerobic bacteria which is able to regulate metabolic pathways depending on the type of sugar presented in the medium. We have studied the glucose-lactose shift in *E. coli* at the protein level using a recently developed mass spectrometry platform.

Method: Cells were grown in minimal medium containing two sugars (glucose and lactose) and analyzed using novel mass spectrometry cluster. The cluster combines the high resolving power and dynamic range of Fourier transform ion cyclotron resonance (FTICR) for accurate mass measurement and quantitation with multiple ion traps for fast and sensitive tandem mass spectrometry. The protein expression profile was followed in time across the glucose-lactose diauxic shift using label-free quantitation from the FTICR data.

Results and Conclusion: The entire dataset was interrogated by KEGG pathway analysis, mapping measured changes in protein abundance onto known metabolic pathways. The obtained results were consistent with previously published gene expression data, with β-galactosidase being the most strongly induced protein during the diauxic shift.

INTRODUCTION

Bacteria, such as *Escherichia coli*, provide "simple" biological models due to a relatively small genome/proteome size (less than 5,000 genes/proteins) and are easy to culture. When the growth medium is rich in glucose, *E. coli* uses glycolysis to convert glucose into pyruvate, requiring adenosine diphosphate (ADP) and oxidized nicotinamide adenine dinucleotide (NAD^+) as cofactors. But *E. coli* is also able to use many other sugars, including lactose, as the main carbon source.[1] The genetic mechanism of metabolic switch from glucose to lactose was first described in the pioneering work of Jacob and Monod fifty years ago.[2] The operon model that they suggested[3] can be described as follows: In the absence of any regulation, the expression of three structural genes (*lacZ*, *lacY*, *lacA*) is inhibited by a repressor molecule, the protein product of *lacI* gene. If present, lactose is taken up from the medium and allolactose, formed from lactose, releases the repressor from the operator. In absence of glucose, cAMP concentration is high and cAMP binds to the catabolite activator protein (CAP), allowing the latter to bind to the promoter and initiate mRNA synthesis. This kind of double control causes the sequential utilization of the two sugars in discrete growth phases. According to this model, the operator region is not essential for operon activity, but rather serves as a controlling site superimposed on a functioning unit.[4]

While previous studies were focused on discovery of genetic mechanisms of metabolic switches, we used a new label-free proteomic approach to study the dynamics of protein expression during the metabolic switch. Proteomics is a powerful and rapidly developing field of research, increasingly expanding our detailed understanding of biological systems. It can be used in basic studies on protein dynamics, localization, and function[5] but also to discover potential biomarkers for diseases and response to pharmaceuticals.[6] Proteomics aims to be *comprehensive* - quantifying "all" proteins present in an organism, tissue or cell. This is a non-trivial task, as there are no amplification methods akin to the polymerase chain reaction available, and proteins in a complex sample typically vary over many orders of magnitude in concentration. Common solutions to overcome this problem include fractionation of proteins, *e.g.* by SDS-PAGE[7] or chromatography, and depletion of abundant proteins.[8, 9] The accurate quantitation of changes in protein expression in or between different samples or states is one of the primary objectives in proteomics.[10] Several methods for labeling proteins

metabolically (in cell cultures) or after extraction are widely applied in "shotgun" proteomics. The labels either incorporate heavy, stable isotopes or a fluorescent group. Nonetheless, it is also possible to quantify peptides and proteins in individual samples directly from the mass spectrometer signal, the so-called "label-free" quantitation. This type of quantitation demands reproducible sample preparation and protein digestion, and benefits from using a mass spectrometer with a wide dynamic range and resolving power, such as an FTICR instrument. Despite these prerequisites, label-free quantitation holds a few advantages over the use of labels. For instance, the sample workup procedure is simpler as there is no labeling step, and the number of samples is not in any way limited by number of labeling reagents and can be used in large studies or for analyzing a large number of time points. Methods based on labeling, on the other hand, have a built-in maximum number of samples that can be analyzed in parallel, beyond which multiple analyses has to be made by bridging between them (which requires one sample or reference to be shared between at least two analyses). Label-free methods seek to reduce potential interferences, for instance by increasing resolving power, and improving accuracy, *e.g.* through data normalization.[11] In our study we used a novel FTICR-ion trap cluster which combines the high mass accuracy of FTICR with fast and relatively inexpensive ion traps for MS/MS[12] making it ideally suited for large-scale, label-free proteomic studies.

MATERIALS AND METHODS

Escherichia coli Glucose-Lactose Diauxie Experiment

Previous work has shown that glucose-lactose diauxie involves activation of the *lac* operon and high expression of β-galactosidase, but also of many other genes and proteins. To compare with gene expression data we reproduced the experiment of Traxler *et al*. using *E. coli* K12 strain MG1655 (ATCC® Number 47076, ATCC, Manassas, VA, USA); this strain was grown overnight in 25 mL Luria-Bertani (LB) medium in 50-mL Falcon tubes. When optical density at 600 nm (OD600) reached 5.0, the cell culture from each Falcon tube was spun down in an Eppendorf 5810 centrifuge at 194×g and 37°C. The supernatants were removed, the pellets

resuspended in warm (37°C) sterile PBS, pooled together and spun down again with the same parameters. After the PBS was removed, 10 ml of 1X. MOPS minimal medium (Teknova, Hollister, CA, USA) was added and the OD600 measured. This culture was then used to inoculate a 3-L bioreactor (Applikon, Schiedam, Netherlands) with 1 L 1X MOPS minimal medium containing 0.5 g/L glucose and 1.5 g/L lactose as the only carbon sources. The temperature was kept at 37°C, dissolved oxygen maintained above 20% and the growth of cells monitored by sampling 1.5 mL of culture for OD600 measurement. The concentration of glucose and lactose were assayed using enzymatic kits (Sigma-Aldrich, St. Louis, MO, USA and BioVision, Mountain View, CA, USA, respectively). Samples were drawn from the culture every 30 minutes before and after diauxie and every 10 minutes near and during the diauxic shift. Cells were spun down at 4°C and 3,500 rpm, transferred to a fresh tube and frozen at -20°C. After collection of all time points, all pellets were thawed, rinsed with ice cold PBS, transferred to a 1.5-mL Eppendorf tube and spun down again for 10 min on maximum speed (16,100×g) at 4°C.

Protein Extraction, In-solution and In-gel Digestion

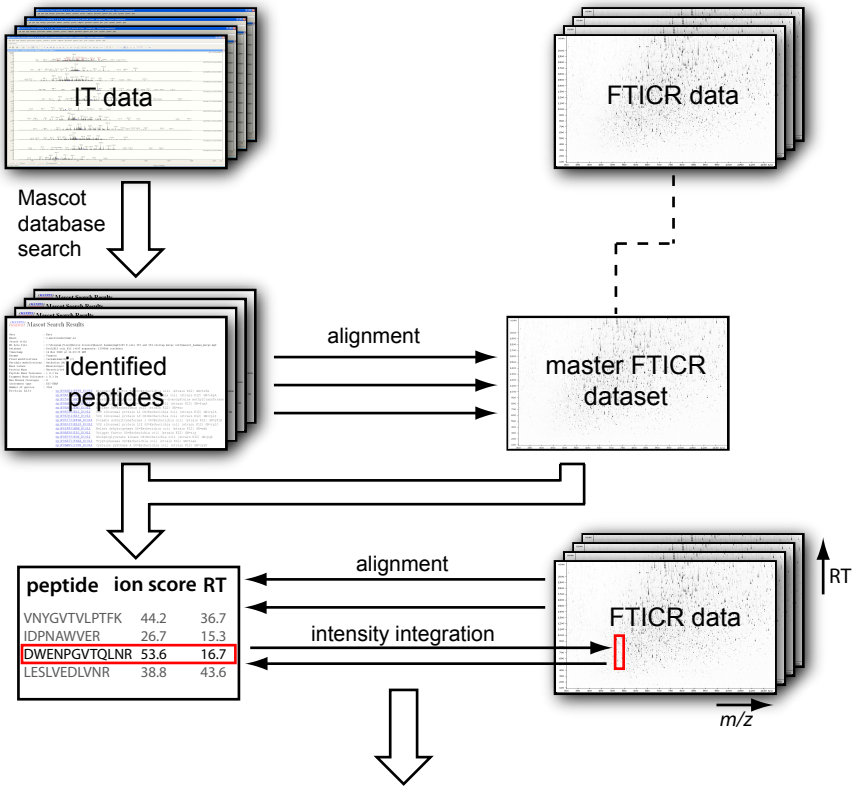
The pellets were weighed and 5 mL of the BugBuster® Master Mix (Novagen, Merck KGaA, Germany) was added per gram cell paste. Cells were incubated at room temperature on a shaking platform at slow settings for 20 min. After the insoluble cell debris was removed by centrifugation at 16,100×*g* for 20 min at 4°C, the supernatant was transferred to a fresh tube. Proteins extracted from the pooled sample of one early and one late time point were used for SDS-PAGE protein separation and in-gel digestion for peptide and protein identification. The rest of the proteins were used for in-solution digestion and peptide and protein quantitation. The extracted proteins for each time point and replicate were digested using trypsin. To each 50 µL of protein extract (approximately 0.25 mg protein) 10 µL 60 mM DTT in 25 mM ammonium bicarbonate (ABC) was added, followed by incubation for 45 min at 56°C to reduce cystines. After 45 minutes, 100 mM iodoacetamide (IAA) in ABC was added to a final IAA concentration 25 mM and the samples kept in dark for 1 h at room temperature to alkylate and protect the cysteins. The proteins were then digested for 5 hours at 37°C by adding 10 µL 100 ng/µL sequencing-grade trypsin (sequencing grade, Promega, Madison, WI, USA) in ABC. The digestion was quenched by

adding 5 µL 10% TFA to lower the pH. The peptide digests were stored at -20°C until analysis.

For MS/MS peptide identification, 25 µg of proteins from two time points, one before and one after the diauxic shift, were fractionated using 8-12 % acrylamide SDS-PAGE (NuPAGE™ 8-12%, Invitrogen, Carlsbad, CA, USA). The gel was stained overnight (12 h) in staining solution (Invitrogen) with 5% methanol and was then washed with milli-Q water until cleared. The gel lanes were cut into twenty-six 2 mm bands and transferred to 96-well plate. Each band was de-stained using 25 mM ABC and acetonitrile, reduced (75 µL 10 mM DTT, 56°C, 30 minutes), alkylated (75 µL 55 mM iodoacetamide, room temperature, 20 min in dark) and digested in-gel using trypsin (20 µg in 20 µL) 12 h at 37°C. The supernatant from each well was transferred to a fresh plate. The digestions were quenched by adding 4 µL 5% TFA (first extraction). The gel pieces were then incubated for 1 hour at 37°C in 0.1 % TFA, after which the second supernatant was pooled with the first extraction and frozen.

FTICR – Ion Trap Cluster

The novel FTICR – ion trap cluster[12] consists of a refrigerated solariX™ 12 T FTICR (Bruker Daltonics, Bremen, Germany) and six ion traps. In this study, CID data from an HCT ultra ion trap (Bruker Daltonics) was used for peptide identification by MS/MS. All mass spectrometers in the cluster were coupled on-line to parallel, splitless NanoLC-Ultra 2D plus systems (Eksigent, Dublin, CA, USA) with additional loading pumps for fast sample loading and washing, which resulted efficient use of the mass spectrometers and high chromatographic peak capacity. All LC systems were configured with 15-cm 300 µm-i.d. ChromXP C18 columns supplied by Eksigent and linear 90 minute gradients from 4 to 44% acetonitrile in 0.05% formic acid were applied. The LC systems were controlled by HyStar 3.2-3.4 with a plugin from the LC manufacturer, the ion traps by esquireControl 6.2 and the FTICR by apexControl 3.0, all from Bruker. The acquired data from each mass spectrometer was automatically transferred to a dedicated server and processed as described below.



protein	peptide	ion score	intensity	RT	m/z	scan number
SODM_ECOLI	EFWNVVNWDEAAAR	76.7	22848000	36.3	853.8997	1860
PHSM_ECOLI	EWANDLDQLINLEK	62.7	5552352	46.3	850.9281	1969
AAT_ECOLI	NYLGIDGIPEFGR	75.4	50948096	35.1	725.8699	1704
AAT_ECOLI	SVFNSAGLEVR	69.0	50687232	24.3	589.8118	1032
AK1H_ECOLI	IAEVDGNDPLFK	48.7	39209856	30.5	659.3379	1148
BGAL_ECOLI	APLDNDIGVSEATR	93.2	1194616	21.5	729.3652	877
BGAL_ECOLI	GDFQFNISR	55.8	11462864	25.0	542.2645	1123
BGAL_ECOLI	IGLNCQLAQVAER	82.3	889548	28.5	787.8925	1091
BGAL_ECOLI	LPSEFDLSAFLR	54.9	11384224	53.9	697.8694	2115
BGAL_ECOLI	LWSAEIPNLYR	57.7	725716	36.0	681.3642	1543
BGAL_ECOLI	VDEDQPFPAVPK	72.5	31464704	18.9	671.3379	982

Figure 5.1. Data processing workflow. The data obtained from the FTICR-ion trap cluster was processed using the workflow illustrated here. First, the LC-MS/MS datasets from the ion trap were searched against the *Escherichia coli* protein sequence database using Mascot. Each individual result was aligned to a single master LC-MS dataset and then merged into one file with aligned retention times. Each separate FTICR LC-MS dataset was aligned against the merged LC-MS/MS data (and hence the master FTICR dataset). Intensities of the identified peptides were then extracted from each FTICR LC-MS dataset by taking the maximum signal in a window of defined *m/z* and retention time relative to the identified peptide. The resulting list contained the protein name, peptide sequence, maximum observed ion score, and absolute intensities for each peptide. This information from each sample could then easily be collapsed into a single, uniform sample/data matrix with the total absolute intensities for all identified proteins and samples.

Data analysis

Each individual MS/MS dataset provided by the ion traps was converted to MGF files using DataAnalysis (Bruker Daltonics). The datasets were separately searched using Mascot 2.1 and converted to the pepXML[13] format. Using the identified peptides, each LC-MS/MS dataset was aligned against a master FTICR LC-MS dataset using msalign[14] and merged. All identified peptides with a best Mascot ion score of at least 25 were then aligned against each individual FTICR LC-MS dataset, one for each biological replicate and time point. Using these alignments, the peaks corresponding to the identified peptides were integrated over the duration of the chromatographic peak. The data analysis workflow is illustrated in Figure 5.1. Only peptide identifications confirmed by accurate mass measurement were thus used. The peptides were then grouped into proteins, using only peptides attributable to a single protein, and the sum of all peptide intensities used as a measure of protein abundance. The data was normalized against the most abundant protein and the earliest time point. The resulting relative protein intensities were $\log_2$–transformed and visualized using the *gplots* package in R. In the same package we created hexadecimal color codes corresponding to the average values over all expression ratios for each protein. An expression ratio of +2.5 thus corresponded to #00FF00, 0 to #FFFF00 and -2.5 to #FF0000. The color codes were then mapped onto metabolic pathways available in the Kyoto Encyclopedia of Genes and Genomes (KEGG).[15]

RESULTS AND DISCUSSION

The glucose-lactose diauxie is a classical *Escherichia coli* experiment which has been repeated many times, including recent studies on gene expression using microarrays.[16] In our experimental setup, the growth rate and glucose concentration allowed precise determination of onset of glucose-lactose (Figure 5.2). The onset of diauxie occurred when cell suspension reached OD600 of ~0.6 or a density of approximately 5×10^8 cells/mL.[17] This was reproducible in each experiment (OD600 of 0.64, 0.60, and 0.55 respectively) and the OD600 could be used as a predictor during the experiment to optimize the sampling of the culture before and during the diauxic shift. The cell density at the onset of diauxic shift was

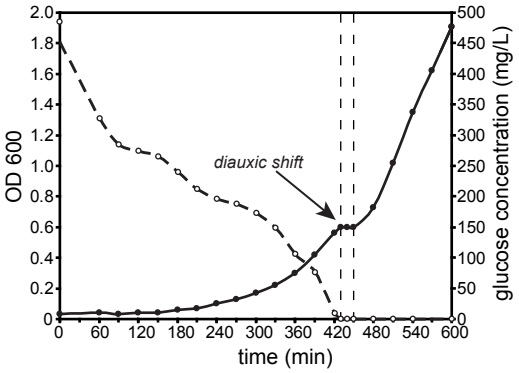


Figure 5.2. Measured cell growth (OD600, solid) and glucose concentration (dashed) in one glucose-lactose diauxie experiment. The onset of the diauxic shift is easily determined from the 20-30 minute plateau in the growth curve, which coincides with the depletion of glucose in the medium. After about +200 minutes, both sugars are exhausted and the growth stops ($OD600_{max}$ = 2.2-2.4).

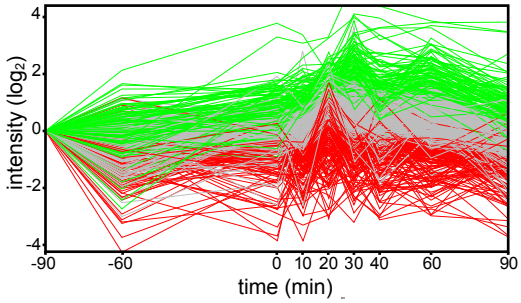


Figure 5.3. Glucose-lactose diauxie protein expression. The proteins expressions were visualized using R and clustered in three groups (green – upregulated , red – downregulated, gray – no change). For subsequent analyses, the time scales of all replicates were aligned with time t = 0 at the observed onset of diauxic shift.

approximately one quarter of the final density, which is consistent with previous observations, and depends on the glucose-lactose ratio [18]. The time scales for all protein expression measurements could thus be aligned to the growth curve for each replicate culture experiment, facilitating discrimination consistent observations and measurement noise. From the LC-MS/MS data of 52 SDS-PAGE slices, 4,333 peptides from 948 proteins were identified (see the additional file 1) with a false discovery rate of 6.75% of the peptide level (Figure 5.3). During the diauxie, we observed rapid changes in protein expression (see the additional file 2). However the magnitude of those changes was not as drastic as gene expression. Comparing with the publicly available gene expression data from Traxler *et al.*,[16] many similar expression patterns can be recognized, especially for strongly upregulated genes/proteins. Not surprisingly, β-galactosidase expression increased strongly, almost 16-fold, during diauxic shift and followed the dynamics of gene expression (Figure 5.4) with a small lag expected by the delay between gene activation and accumulated protein. The genetic response occurred immediately after glucose exhaustion but protein synthesis is typically delayed between 20 seconds and several minutes in *E. coli*.[3] Small relative changes in concentration of already abundant proteins are difficult to detect immediately and need to be

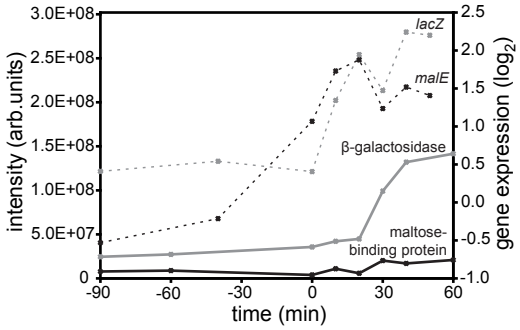


Figure 5.4. Expression of *lacZ* and *malE*. mRNA (dashed) and protein (solid) dynamics for periplasmic maltose-binding protein/*malE* (black) and β-galactosidase/*lacZ* (gray) upregulated during glucose-lactose diauxie (time 0).

accumulated for some time before they can be observed. Nevertheless, we noticed that the most significant changes in protein abundance took place within 40 minutes after onset of diauxic shift, which is consistent with published gene expression data and the observed resuming of growth. Since the gene expression data was derived from that published by Traxler *et al.*, the alignments of the time-scales are not perfect and minor discrepancies between the sampling of the gene and protein expression could be expected. The protein expression measurements were with a few exceptions reproducible, albeit not always in perfect agreement with the published gene expression data. This could be explained by noise in the data and the fact that gene and protein expression were not measured in the same cell culture. For instance, the change in gene expression of *malE* is almost the same as for *lacZ*, but at the proteomic level we observed only slight changes in abundance of the maltose-binding protein coded for by *malE* (Figure 5.4). (The maltose-binding protein is a periplasmic component of the maltose ABC transporter which is capable of transporting malto-oligosaccharides up to seven glucose units long.[19])

Using the clustering function for large datasets, *clara*, from the R *cluster* package[20], the dataset could be broadly divided into groups of up- and downregulated proteins, along with proteins that do not change measurably as a function of the diauxic shift. The FTICR-ion trap cluster provided comprehensive label-free quantitative proteomic data with sufficient throughput for an arbitrary number of conditions or time points and biological replicates (here about 30), allowing a global study of protein expression dynamics in *E. coli*. With this instrument platform, proteomics data such as that presented here can be routinely generated in less than 48 h.

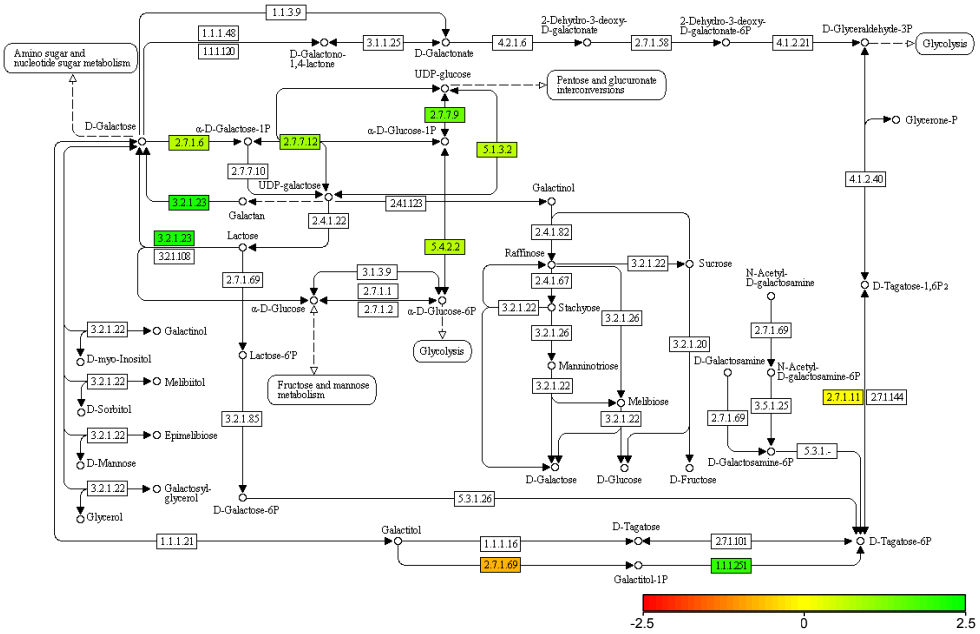


Figure 5.5. The protein expression profiles mapped onto the galactosidase metabolic pathways highlights changes in metabolism when shifting from glucose to lactose as primary carbon source. The measured changes in enzyme (protein) abundance were converted to color and mapped onto KEGG pathways. Upregulated proteins are marked in green, downregulated in red, and unchanged in yellow.

To illustrate changes in metabolic pathways, the protein expression data was mapped onto KEGG metabolic pathways and changes in level of expression indicated by color (Figure 5.5). Most proteins in the same pathways as β-galactosidase were also markedly upregulated, leading to a global activation of the galactose pathway responsible for channeling lactose into the glycolytic pathway. Other metabolic pathways changed to a lesser degree, as measured by protein (enzyme) abundance.

CONCLUSIONS

We have reproduced the textbook glucose-lactose diauxie experiment in *E. coli* using a state-of-the-art method for quantitative proteomics using a novel mass spectrometry platform, the FTICR-ion trap cluster. In each of

three experiments the onset of diauxie occurred at approximately the same cell density and the duration of diauxic shift was also similar. The identified and individually quantified peptides were collected into quantitative protein measurements, which were visualized and compared using tools developed in-house. Through kind assistance from KEGG it is now possible to upload color codes for a whole list of quantified proteins on any metabolic pathway overview (the R program for generating the color codes from protein abundance ratios is available from the authors). We could confirm that the most strongly induced enzymes belong to the pathway responsible for glucose and lactose metabolism.

The FTICR-ion trap cluster in combination with the appropriate visualization tools makes an efficient approach for investigation of protein expression dynamics. The new instrument configuration and software proved robust in acquiring and processing data, allowing label-free quantitation of ~1,000 identified proteins over ~30 time points in a 24 h measurement. Furthermore, the high dynamic range and resolving power of FTICR made label-free quantitation accurate and precise, at least for a label-free method.[21] Finally, as expected, key aspects of the proteome dynamics were indeed bound to reflect gene expression under the glucose-lactose metabolic switch.

ACKNOWLEDGMENTS

The authors wish to thank René van Zeijl, Hans Dalebout, Hannah Scott for technical assistance and Mao Tanabe for kind help with the KEGG pathway "mapper".

SUPPLEMENTARY MATERIAL

Supplementary files 1 and 2 associated with this article may be found in the online version at http://www.biomedcentral.com/1471-2180/11/126/additional.

REFERENCES

1. Lewis, I. M., Bacterial Variation with Special Reference to Behavior of Some Mutabile Strains of Colon Bacteria in Synthetic Media. *J Bacteriol* **1934,** 28, (6), 619-39.
2. Jacob, F.; Monod, J., Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **1961,** 3, 318-56.
3. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P., *Molecular Biology of the Cell*. 4 edn. ed.; Garland Science Publishing: 2002.
4. Beckwith, J. R., Regulation of the lac operon. Recent studies on the regulation of lactose metabolism in Escherichia coli support the operon model. *Science* **1967,** 156, (3775), 597-604.
5. James, P., Protein identification in the post-genome era: the rapid rise of proteomics. *Q.Rev.Biophys.* **1997,** 30, (4), 279-331.
6. Mullner, S.; Neumann, T.; Lottspeich, F., Proteomics--a new way for drug target discovery. *Arzneimittelforschung.* **1998,** 48, (1), 93-95.
7. Laemmli, U. K., Cleavage of Structural Proteins during Assembly of Head of Bacteriophage-T4. *Nature* **1970,** 227, (5259), 680-&.
8. Boschetti, E.; Righetti, P. G., The ProteoMiner in the proteomic arena: A non-depleting tool for discovering low-abundance species. *Journal of Proteomics* **2008,** 71, (3), 255-264.
9. Echan, L. A.; Tang, H. Y.; Ali-Khan, N.; Lee, K.; Speicher, D. W., Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics* **2005,** 5, (13), 3292-3303.
10. Ong, S. E.; Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology* **2005,** 1, (5), 252-262.
11. Elliott, M. H.; Smith, D. S.; Parker, C. E.; Borchers, C., Current trends in quantitative proteomics. *J.Mass Spectrom.* **2009,** 44, (12), 1637-1660.
12. Palmblad, M.; van der Burgt, Y. E.; Mostovenko, E.; Dalebout, H.; Deelder, A. M., A Novel Mass Spectrometry Cluster for High-Throughput Quantitative Proteomics. *J.Am.Soc.Mass Spectrom.* **2010**.
13. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **2005,** 1, 2005 0017.
14. Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R., Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J. Am. Soc. Mass Spectrom.* **2007,** 18, (10), 1835-43.
15. Kanehisa, M.; Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000,** 28, (1), 27-30.
16. Traxler, M. F.; Chang, D. E.; Conway, T., Guanosine 3',5'-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** 103, (7), 2374-9.
17. Brown, T. A., *Gene Cloning and DNA Analysis: An Introduction*. 6 ed.; John Wiley and Sons Ltd: Chicester, UK, 2010; p 336.

18. Loomis, W. F., Jr.; Magasanik, B., Glucose-lactose diauxie in Escherichia coli. *J Bacteriol* **1967,** 93, (4), 1397-401.

19. Ferenci, T., The recognition of maltodextrins by Escherichia coli. *Eur J Biochem* **1980,** 108, (2), 631-6.

20. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Cluster Analysis Basics and Extensions. In 2005.

21. Mann, M.; Kelleher, N. L., Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci U S A* **2008,** 105, (47), 18132-8.

6

Spatiotemporal Proteomics of Human embryonic Stem Cells Differentiating to Cardiomyocytes

Ekaterina Mostovenko,[1] Harsha D. Devalla,[2] Yassene Mohammed,[1] André M. Deelder,[1] Robert Passier[2] and Magnus Palmblad[1]

[1] Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, The Netherlands

[2] Department of Anatomy and Embryology, Leiden University Medical Center, The Netherlands

Manuscript in preparation

ABSTRACT

Directed differentiation of human pluripotent stem cells (hPSCs) to specialized cell types opens new possibilities for tissue replacement in regenerative therapies. Although efficient protocols have been developed to generate cardiomyocytes *in vitro*, much remains to be learned about the changes that take place in the transcriptome, proteome and methylome of the differentiating cells. Studies aimed at deciphering these dynamic changes during differentiation will lead to better understanding of the multiple events involved in this process. With rapidly increasing sensitivity and speed of mass spectrometry, large-scale, spatiotemporal analyses of the proteome are becoming feasible. Here, we show how a fast and simple sample preparation method can be used to obtain enriched subcellular fractions. Combined with samples collected at different stages of differentiation, the changes in the proteins can be studied in the context of their localization and expression dynamics. In total, we identified 41,626 peptides from 6,936 proteins over four time points and five subcellular localizations. Known cardiomyocyte markers were identified, and proteins associated with motor and structural activity increased over time, including known proteins responsible for cell-cell communication and contractile properties. We describe a method and scientific workflow to analyze, visualize and browse the data, incorporating mass spectrometry and gene ontology information The data as well as all methods are made publically available.

INTRODUCTION

Human embryonic stem cells (hESC), derived from the inner cell mass of blastocyst-stage embryos, have the capacity to self-renew indefinitely and to differentiate to all cell types of the human body[1]. In the presence of the appropriate signaling cues ESCs can form tissues of all three primary germ layers[2] and all types of somatic cells *in vitro* and *in vivo*[3, 4] In recent years, detailed differentiation protocols have been developed for the production of specialized cell types *in vitro*. However, many details of the differentiation process are still unclear and identification or isolation of specific (subtype) cell populations is challenging. A better understanding of the different steps during stem cell differentiation may lead to a better control of the differentiation process and to a potentially unlimited source material for tissue replacement in regenerative therapies[5]. This is even more relevant following the demonstration of the groundbreaking technology for reprogramming human (patient-derived) somatic cells to induced pluripotent stem (iPS) cells,[6] which share similar properties with ESCs.

Recently, several groups, including ours, have reported defined differentiation protocols leading to efficient cardiomyocyte production using human ESC end iPS cells.[7] In general, stage-specific modulation of different proteins, such as Activin A, Bone Morphogenetic Protein (BMP) and Wnt, are important for consistent and efficient production of functional cardiomyocytes,[8, 9] which are phenotypically and electrophysiologically comparable to primary human fetal cardiomyocytes in culture.[10]

Although whole genome transcriptome analysis of various stages during cardiomyocyte differentiation of hESC has been performed,[11] studies of the human proteome of hESC during cardiomyocyte differentiation are either at very early stages of differentiation or in hESC-derived cardiomyocytes.[12, 13] The level of mRNA expression cannot be directly correlated with amount of the protein it codes for, due to post-translation regulation, post-translation modification, protein export or degradation. In addition, gene expression information is very hard to interpret in terms of biochemical function, structure and subcellular localization of its protein product and mostly derived from *in silico* prediction models. Quantitative spatially and temporally resolved proteomics enables us to gain insight into the mechanisms driving the differentiation towards a specific cell type. Differentiating ESCs undergo massive transformations in the proteome on

their path to mature cardiomyocytes in order to alter their metabolism and drastically change their morphology, including the synthesis and assembly of sarcomers, the smallest contractile units of cardiomyocytes. Most regulatory proteins exist in cells at low concentrations and are localized at very specific subcellular compartments. Additionally, due to the high complexity of eukaryotic cells, extra fractionation step is commonly beneficial for the whole proteome studies. Here, we couple simple organellar fractionation with standard proteomics techniques, which allow us, in principle, to measure and follow changes in protein distribution by quantifying proteins in individual organellar fractions separately.

One of the major challenges here, especially for label-free proteomics, is the separation of enriched subcellular partitions (cytosol, cytoskeleton, membranes, nucleus and nucleolus) from a limited number of cells in a reproducible manner. To analyze enriched bur not necessarily pure fractions, additional computational and statistical tools may be needed. In this work we describe a simple protocol for sample preparation for label-free spatiotemporal proteomics in hESC differentiating towards the cardiac linage, including an automated workflow to analyze and visualize the results.

MATERIALS AND METHODS

Cell cultivation, collection and protein extraction

For differentiation of human pluripotent stem to the cardiac lineage we used a previously described a cardiac reporter line, hESC Nkx2.5-GFP, which expresses GFP under control of the cardiac transcription factor Nkx2.5.[8] At different stages during differentiation we collected cells for protein extraction. For cardiomyocyte differentiation cells were resuspended in defined differentiation medium containing BPEL and a cocktail of growth factors BMP4 (20 ng/mL), Activin A (20 ng/mL), VEGF (30 ng/mL) and SCF (40 ng/mL) and plated at a density of 3000 cells per well in V-shaped 96-well plates. After 3 days cells were washed in BPEL only and after 7 days cell aggregates (so-called embryoid bodies) were plated on matrigel-coated wells.[8] Cells were collected at different time points (0, 3, 7, and 15 days) recapitulating different embryonic/cardiac developmental stages. Day

0 represents the "pluripotent undifferentiated" stage, day 3 "mesoderm", day 7 "cardiac progenitor" and day 15 "beating cardiomyocyte" stages. For collection of cells, medium was removed following centrifugation (3 min at 100×g) and, washed twice with PBS, and stored at -80°C. Proteins were isolated following cell lysis in 1% SDS in a hot (70°C) ultrasonic bath (VWR Ultrasonic Cleaner) for 2×10 min. To remove DNA, the protein extract was incubated on ice 15 min with 12.5 U benzonase (Novagen, Merck KGaA, Darmstadt, Germany) in presence of 2 mM $MgCl_2$. Thirty µg of proteins of each sample (bicinchoninic acid (BCA) protein assay kit; Thermo Fischer Scientific, Waltham, MA) was diluted in water and 4x LDS running Buffer (Invitrogen, Carlsbad, CA) for SDS-PAGE.

Subcellular fractionation

For the organellar fractionation 10^6 cells (equivalent to 30 µg of protein in total cell extract) we used Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Fischer Scientific). From the day 7 time point, collected cells were first resuspended in ice-cold Cytoplasmic Extraction Buffer and incubated on ice for 10 min. Cytoplasmic fraction was separated by centrifugation at 500×*g* for 5 min and supernatant was collected to a fresh pre-chilled tube. Obtained pellet was mixed with Membrane Extraction Buffer and vortexed for 5 sec. After the incubation on ice for 10 min the sample was centrifuged at 3,000×*g* for 5 min, the membrane extract was transferred to a fresh pre-chilled tube. Subsequently, the nuclear fraction was acquired by incubation of the remaining pellet with Nuclear Extraction Buffer on ice for 30 min and further centrifugation at 5,000×*g* for 5 min. The same buffer containing 100 mM $CaCl_2$ and Micrococcal Nuclease was used to extract chromatin-associated proteins. Finally, Pellet Extraction Buffer was added to the remaining sample and incubated for 10 min at room temperature. The supernatant after centrifugation at 16,000×*g* for 5 min containing cytoskeletal extract was transferred to a new tube. To prevent protein degradation all buffers contained Protease Inhibitor Cocktail (included in the kit). Resulting fractions (approximate protein content 3-5µg) were lyophilized and reconstituted in 15 µL of water and diluted with 5 µL 4x LDS running Buffer (Invitrogen) for SDS-PAGE.

SDS-PAGE and in-gel digestion

Each sample was loaded on a 1 mm 10x well 4-12% NuPAGE® Bis-Tris gel (Invitrogen) and separated for 1 h at 180 V. The gel was stained in NuPAGE® Colloidal Blue (Invitrogen) overnight on the shaking platform at

room temperature and distained with milli-Q water until cleared. Each lane was cut into 48 identical 1.5×5-mm slices using a MEE1.5-5-48 disposable gel cutter (Gel Company Inc., San Francisco, CA) and placed into a 96-well polypropylene PCR plate (Greiner Bio-One, Frickenhausen Germany). The gel pieces were distained using consecutively 25 mM ammonium biocarbonate (ABC) and 100% acetonitrile. DTT reduction and IAA alkylation were performed according to the previously published protocol.[14] In-gel tryptic digestion was done in 30 µL of 25 mM ABC containing 5 ng/µL trypsin (sequencing grade, Promega, Madison, WI) for 6 h at 37°C. The resulting peptides were extracted with TFA according to the previously described protocol.[14]

LC-MS/MS analysis

The analysis was performed using a splitless NanoLC-Ultra 2D plus (Eksigent, Dublin, CA) for parallel ultra-high pressure liquid chromatography (UHPLC) with an additional loading pump for fast sample loading and desalting. The UHPLC system was configured with 300 µm-i.d. 5-mm PepMap C18 trap columns (Dionex, Sunnyvale, CA) and 15-cm 300 µm-i.d. ChromXP C18 columns (Eksigent). Peptide separation was performed running 90 min linear gradients from 4 to 33% acetonitrile in 0.05% formic acid. The UHPLC system was coupled on-line to an amaZon ETD speed high-capacity 3D ion trap with the standard 2-50 µL/min electrospray source (Bruker Daltonics, Bremen, Germany). After each MS scan, up to ten abundant multiply charged species in *m*/*z* 300-1300 range were automatically selected for data-dependent MS/MS but excluded for one minute after being selected twice. The LC-MS/MS system was controlled using HyStar 3.4 with an Eksigent plug-in and trapControl 7.0, all from Bruker.

Data analysis

All acquired tandem mass spectrometry data was processed in one batch using Taverna workbench.[15] The workflow (Figure 6.1) converts raw data .yep files to mzXML[16] using compassXport 3.0.5 (Bruker) and passes this, along with the sequence database to SpectraST in the Trans-Proteomic Pipeline (TPP).[17] The NIST human spectral library from 2011-05-26 was searched with default settings except for allowing carbamidomethylation ("CAM") of cysteines. Search results in pepXML[17] format were then fit by a mixture model and the SpectraST discriminant score converted to

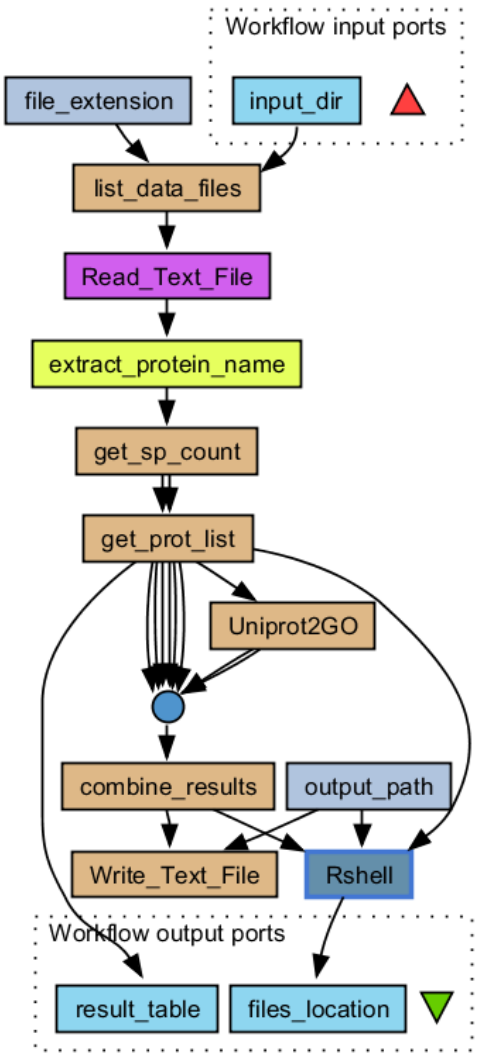


Figure 6.1. Taverna scientific workflow for downloading GO information and visualization of protein expression over time and as function of subcellular fraction. The workflow also compares measured with predicted protein localization and calls *goa_gubbar* to summarize the protein spectral counts in gene ontology categories. The output result table with the whole list of identified proteins and their localization is presented in the supplementary material.

probability for each peptide-spectrum match by PeptideProphet[18]. Results were reported with 2% FDR and for each protein identification its relative abundance was calculated based on the spectral count. The identifications/abundance table was then passed to an Rshell script where it was analyzed, comparing the spatial distribution of each protein with protein localization from a gene ontology annotation database (http://www.ebi.ac.uk/QuickGO/). However, the obtained fractions were impure due to imperfect fractionation or protein trafficking and to remove the possible bias the data from each organellar was first normalized against

total spectral count. For each protein the data was anew normalized across the maximum spectral count per protein, creating the distribution between 0 and 1 which was then converted to a color intensity and mapped onto a symbolic representation of the cell with only these subcellular compartments. Two instances of such representations can combine all protein measurements and the protein localization from gene ontology and included in tables for browsing the data on the protein level, similar to ProteinProphet in the TPP. Gene ontology information for all or specific subsets of the identified proteins and their corresponding spectral counts were summarized using the *goa_gubbar* tool (http://www.ms-utils.org/goa_gubbar/index.html).[19]

RESULTS AND DISCUSSION

Time-resolved proteomics

The aim of this study was to develop a fast, simple and high-throughput organellar fractionation method to generate a library of peptide identifications. Secondly, to establish a simple computational tool to analyze, combine and visualize the multidimensional data and thirdly to apply this technology at various stages of hESC differentiating to cardiomyocytes. To demonstrate feasibility of the spatiotemporal proteomics, two experiments were performed separately. The mass spectrometric analysis methods have been described previsouly[14] and are based on LC-FTMS and label-free quantitation of peptides identified by MS/MS in separate ion trap measurements, similar to the accurate mass and time method[20] and Corra.[21] From the collected undifferentiated pluripotent stem cells at day 0, early mesodermal cells at day 3, cardiac progenitor cells at day 7 and functional beating cardiomyocytes at day 15. Over all time points we identified 39,067 peptides and 5,184 proteins with 2% FDR on the PSM level and approximately similar number of identified peptides/proteins per sample (see Figure 6.2a and supplementary material). By simple spectral count, we observed known sarcomeric proteins induced at the later stages, when cardiomyocytes display contractile activity and are to express high levels of these proteins (Figure 6.3). As expected, proteins involved in motor activity/contractility or cytoskeletal architecture are present in higher abundancy at day 7 and 15 (Figure 6.4). The functional classification of proteins was here based on the gene ontology "slim" terms, with myosin being the major contributor to the motor and structural

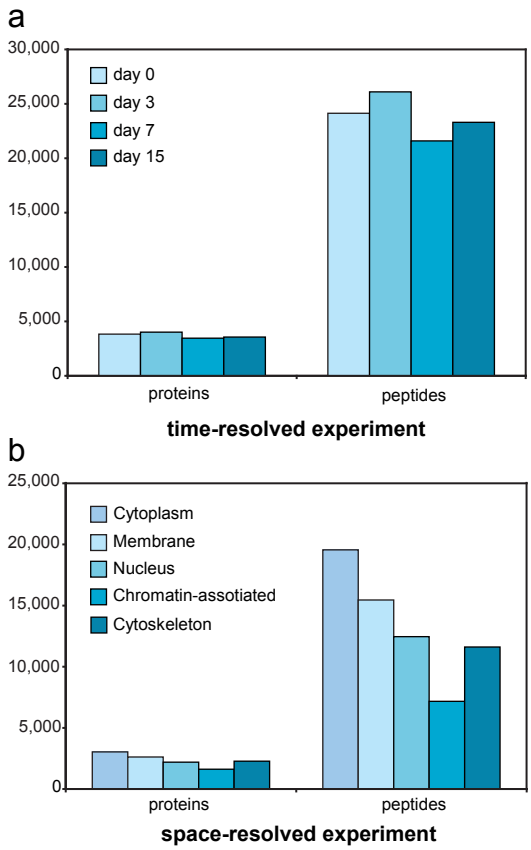


Figure 6.2. SpectraST search results per time point (a) and per organellar (b) in number of identified peptides and proteins.

molecular activity. Laminin, the protein responsible for intercellular communication and unified contraction in cardiomyocytes,[22] was also found upregulated. However, translation regulator activity decreased with time, finding is not unexpected, given that mature cardiomyocytes are tightly interconnected via gap-junctions and have only limited motility. At day 7 of differentiation, GFP fluoresence, representing activity of the transcription factor Nkx2.5, is visible for the first time. At this stage immature cardiomyocytes or cardiac progenitors are present.

Organellar proteomics

Subcellular fractionation is a classical sample refinement method and a wide range of protocols and several commercial kits are available. At first the cellular structure is disrupted using either mild detergent-free buffers to

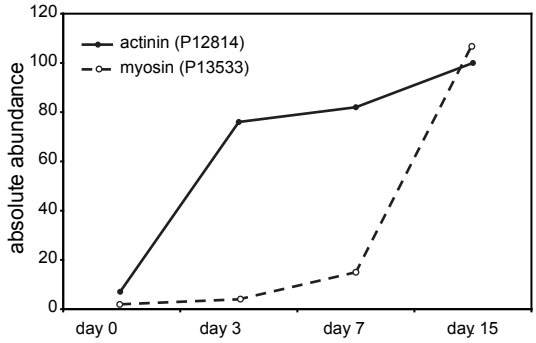


Figure 6.3. Protein expression of the cardiomyocyte markers (actinin – P12814 and myosin – P13533) based on the absolute spectrum count per protein.

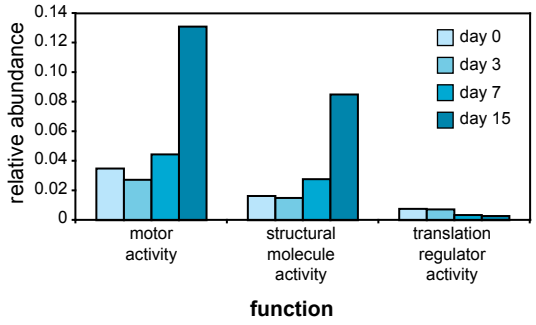


Figure 6.4. Functional distribution of proteins during cardiomyocyte differentiation. Proteins are grouped based on the GeneOntology "slim" terms using *goa_gubbar* tool. The relative abundance is calculated as the percent of total spectral count.

keep the membranes intact or homogenized in presence of more aggressive reagents and then fractionated by ultracentrifugation into different populations of organelles. However, different populations share similar physical properties and cosediment to some extent. Although these organellar fractions are not absolutely pure, they are highly enriched in the targeted organelle. By measuring *all* fractions of a cell, it is possible to account for *all* of the protein and determine where a particular protein is most abundant. This information then either supports or contradicts predicted or previously observed functions and interactions of that protein. If applied to data from several time points, the obtained information enables us to trace the protein dynamics in time as well as in space, producing two-dimensional expression matrix for each protein.

The availability of starting material (cells) is often limiting the use of classical methods such as homogenization followed by ultracentrifugation in a sucrose gradient. We used a simple commercial kit directly compatible with SDS-PAGE, thereby avoiding buffer exchange with concomitant sample loss. Resulting spectral count after SDS-PAGE and LC-MS/MS give

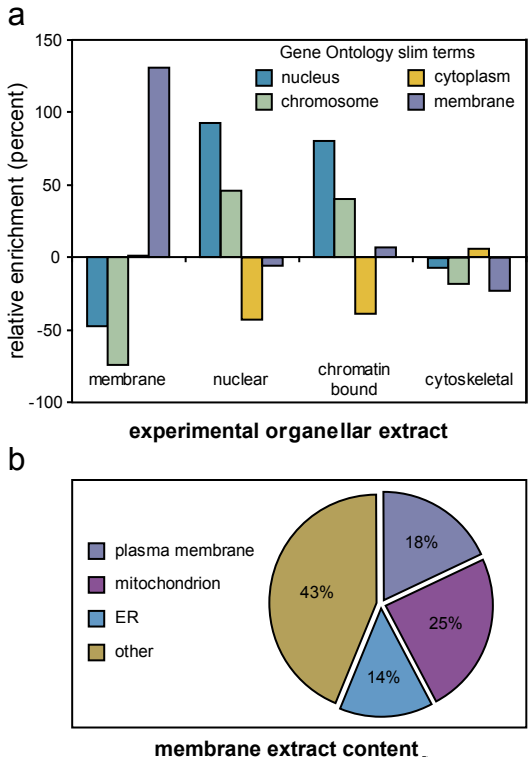


Figure 6.5. Organellar enrichment relative to the experimental cytoplasmic extract compared to the GeneOntology-derived protein localization (a) and the protein content of the membrane extract (b).

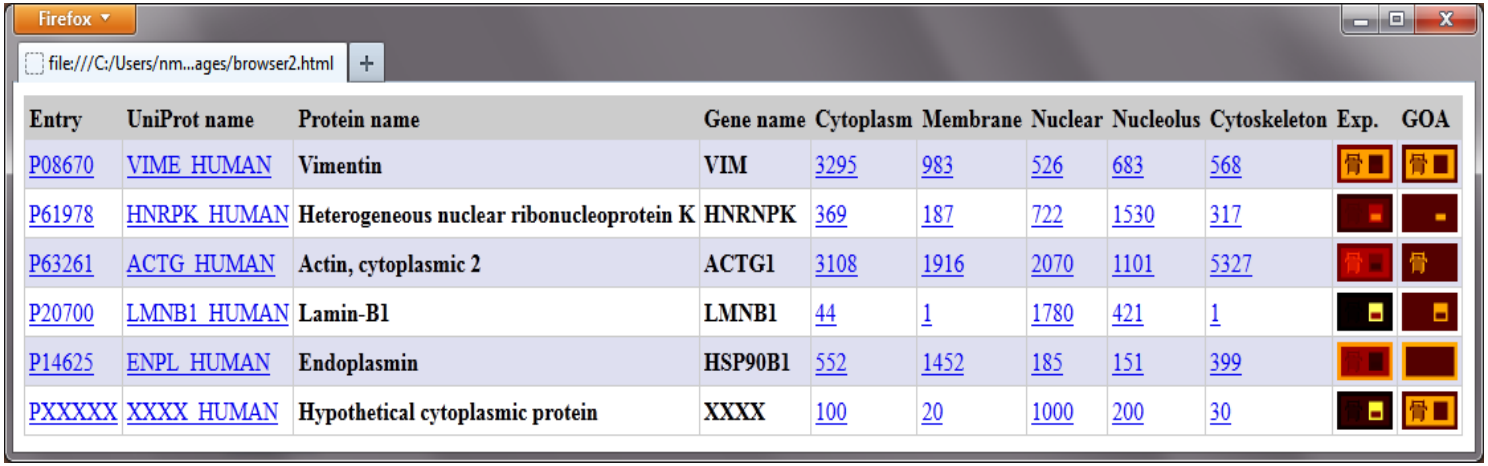
approximate information on protein abundancy. From one time point (day 7) across five cellular components 31,023 peptides from 4,264 proteins were identified (with 2% FDR on the PSM level). Not surprisingly, the number of identified peptides/proteins varied greatly between subcellular fractions (Figure 6.2b and supplementary material). To determine the quality of the partitioning, we compared measured localization with that from gene ontology (Figure 6.5a). The changes are expressed relative to the cytoplasmic fraction, as this is collected first and contains some material derived from all other cellular compartments as well as the extracellular matrix. The membrane extract is strongly enriched with known or predicted membrane proteins compared to all other fractions, and also includes many endoplasmic reticulum and mitochondrial proteins (Figure 6.5b). The ontological protein localization patterns in the nuclear and chromatin-bound nuclear extracts were similar, and strongly enriched in nuclear and chromosomal proteins. However, the gene ontology vocabulary does not correspond to mutually exclusive categories and is often imprecise. For example, most nucleolar proteins are also assigned to the nucleus. The

cytoskeletal fraction is collected last and is enriched in structural proteins, most of which are annotated as cytoplasmic, and contains less membrane and nuclear proteins than the initially collected cytoplasmic fraction. Despite its obvious shortcomings, this simple gene ontology breakdown may be useful for comparing and optimizing subcellular fractionation methods.

Spatiotemporal proteomics

In this preliminary work we only included by one trace along the temporal and one trace along spatial axes. In general, and in future studies, both variables will be varied, producing an abundance matrix for each protein (or protein). For browsing such data, it would be beneficial to analyze and simultaneously visualize spatial and temporal aspects of the data – here the subcellular localization and time point. By translating the protein abundance into a color code and map it to a simplified but intuitive graphical representation of a cell, the data can be visualized statically (Figure 6.6).

The protein expression dynamics can be visualized by a series of such glyphs, or as an animation, for instance as embedded animated PNG images, which can also be generated by R from within a Taverna workflow. This provides a primitive, protein-level browser for inspecting individual proteins or groups of proteins. Links to other, freely available, resources with information on protein localization, such as ProteinAtlas (http://www.proteinatlas.org) can then be used to compare findings from the proteomics experiment with high-resolution immunohistochemical microscopy for that protein, bridging mass spectrometry and affinity based proteomics.



Entry	UniProt name	Protein name	Gene name	Cytoplasm	Membrane	Nuclear	Nucleolus	Cytoskeleton	Exp.	GOA
P08670	VIME_HUMAN	Vimentin	VIM	3295	983	526	683	568		
P61978	HNRPK_HUMAN	Heterogeneous nuclear ribonucleoprotein K	HNRNPK	369	187	722	1530	317		
P63261	ACTG_HUMAN	Actin, cytoplasmic 2	ACTG1	3108	1916	2070	1101	5327		
P20700	LMNB1_HUMAN	Lamin-B1	LMNB1	44	1	1780	421	1		
P14625	ENPL_HUMAN	Endoplasmin	HSP90B1	552	1452	185	151	399		
PXXXXX	XXXX_HUMAN	Hypothetical cytoplasmic protein	XXXX	100	20	1000	200	30		

Figure 6.6. The screenshot of a simple HTML file comparing the measured protein localization with their predicted or or previously determined subcellular compartment. Quantitative data is normalized against the total spectral count in the sample and then divided by the maximum spectral count per protein. The resulting values are mapped to a heat map color scheme from black to red to yellow to white.

CONCLUSIONS

The ability of hESCs to differentiate into all cell types provides many possibilities for clinical applications. However, the lack of detailed, molecular information on changes in the proteome during differentiation still limits our understanding of the underlying mechanisms and processes. In this work we followed the changes in protein expression during the entire differentiation process from embryonic stem cells to mature cardiomyocytes and demonstrated the feasibility of combining temporally and spacially resolved proteomics to shed light on intracellular processes at the protein level and their role in the organellar formation and evolvement. Subcellular fractions could be compared using gene ontology annotations and were indeed found to be enriched in proteins from the corresponding subcellular compartment.

The combined peptide and protein identifications from four time points across five subcellular fractions could be used as an accurate mass and time library in combination with LC-MS from a high resolving power mass spectrometer to obtain high-throughput quantitative data. As yet one more dimension, phosphorylation or other post-translational events could be quantified as functions of time and space by enriching subcellular fractions for phosphoproteins before or phosphopeptides after enzymatic digestion. Due to the destructive nature of the measurement and trade-off between robustness and sensitivity, the single most limiting factor in such multidimensional analyses are the availability of cells.

In this work, all data analysis and visualization was performed within one Taverna workflow, allowing automated processing of large datasets and sharing of the complete analysis method itself. In the most cases, the experimentally defined protein localization (the maximum of spectral count) agreed with the gene ontology annotation. However the GeneOntology database is not complete, and for many proteins the localization information has only been determined *in silico*. A major challenge of spatiotemporal proteomics is the visualization and statistical analysis of the multidimensional data, with a set of quantified peptides for each protein as a function of two coordinates (time and subcellular fraction) and comparison of this with other mass spectrometry or imaging data. Undoubtedly, these challenges will be addressed in the near future, as more and more researchers now undertake this type of studies and increasing numbers of

datasets of this type become publicly available. All data presented here is available in PRIDE and the Taverna workflow used for the data analysis is available on myExperiment.

ACKNOWLEDGEMENTS

The authors wish to thank Chantal Schreurs for help with cell cultivation and harvesting, Hans Dalebout and Oleg Klychnikov for technical assistance and useful discussions.

REFERENCES

1. Thomson, J. A.; Itskovitz-Eldor, J.; Shapiro, S. S.; Waknitz, M. A.; Swiergiel, J. J.; Marshall, V. S.; Jones, J. M., Embryonic stem cell lines derived from human blastocysts. *Science* **1998**, 282, (5391), 1145-7.
2. Bradley, A.; Evans, M.; Kaufman, M. H.; Robertson, E., Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **1984**, 309, (5965), 255-6.
3. Keller, G., Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev* **2005**, 19, (10), 1129-55.
4. Reubinoff, B. E.; Pera, M. F.; Fong, C. Y.; Trounson, A.; Bongso, A., Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol* **2000**, 18, (4), 399-404.
5. Vats, A.; Bielby, R. C.; Tolley, N. S.; Nerem, R.; Polak, J. M., Stem cells. *Lancet* **2005**, 366, (9485), 592-602.
6. Takahashi, K.; Tanabe, K.; Ohnuki, M.; Narita, M.; Ichisaka, T.; Tomoda, K.; Yamanaka, S., Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **2007**, 131, (5), 861-872.
7. Mummery, C. L.; Zhang, J. H.; Ng, E. S.; Elliott, D. A.; Elefanty, A. G.; Kamp, T. J., Differentiation of Human Embryonic Stem Cells and Induced Pluripotent Stem Cells to Cardiomyocytes A Methods Overview. *Circulation Research* **2012**, 111, (3), 344-358.
8. Elliott, D. A.; Braam, S. R.; Koutsis, K.; Ng, E. S.; Jenny, R.; Lagerqvist, E. L.; Biben, C.; Hatzistavrou, T.; Hirst, C. E.; Yu, Q. C.; Skelton, R. J. P.; Oostwaard, D. W. V.; Lim, S. M.; Khammy, O.; Li, X. L.; Hawes, S. M.; Davis, R. P.; Goulburn, A. L.; Passier, R.; Prall, O. W. J.; Haynes, J. M.; Pouton, C. W.; Kaye, D. M.; Mummery, C. L.; Elefanty, A. G.; Stanley, E. G., NKX2-5eGFPw hESCs for isolation of human cardiac progenitors and cardiomyocytes. *Nature Methods* **2011**, 8, (12), 1037-40.
9. Kattman, S. J.; Witty, A. D.; Gagliardi, M.; Dubois, N. C.; Niapour, M.; Hotta, A.; Ellis, J.; Keller, G., Stage-Specific Optimization of Activin/Nodal and BMP Signaling Promotes Cardiac Differentiation of Mouse and Human Pluripotent Stem Cell Lines. *Cell Stem Cell* **2011**, 8, (2), 228-240.
10. Mummery, C.; Ward-van Oostwaard, D.; Doevendans, P.; Spijker, R.; van den Brink, S.; Hassink, R.; van der Heyden, M.; Opthof, T.; Pera, M.; de la Riviere, A. B.; Passier, R.; Tertoolen, L., Differentiation of human embryonic stem cells to cardiomyocytes: role of coculture with visceral endoderm-like cells. *Circulation* **2003**, 107, (21), 2733-40.
11. Beqqali, A.; Kloots, J.; Ward-van Oostwaard, D.; Mummery, C.; Passier, R., Genome-wide transcriptional profiling of human embryonic stem cells differentiating to cardiomyocytes. *Stem Cells* **2006**, 24, (8), 1956-1967.
12. Van Hoof, D.; Munoz, J.; Braam, S. R.; Pinkse, M. W.; Linding, R.; Heck, A. J.; Mummery, C. L.; Krijgsveld, J., Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell* **2009**, 5, (2), 214-26.
13. Van Hoof, D.; Dormeyer, W.; Braam, S. R.; Passier, R.; Monshouwer-Kloots,

J.; Ward-van Oostwaard, D.; Heck, A. J.; Krijgsveld, J.; Mummery, C. L., Identification of cell surface proteins for antibody-based selection of human embryonic stem cell-derived cardiomyocytes. *J Proteome Res* **2010**, 9, (3), 1610-8.

14. Mostovenko, E.; Deelder, A. M.; Palmblad, M., Protein expression dynamics during Escherichia coli glucose-lactose diauxie. *BMC Microbiol* **2011**, 11, 126.

15. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M. R.; Wipat, A.; Li, P., Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004**, 20, (17), 3045-3054.

16. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004**, 22, (11), 1459-66.

17. Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **2005**, 1, 1-8.

18. 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* **2002**, 74, (20), 5383-92.

19. Palmblad, M.; Bindschedler, L. V.; Cramer, R., Quantitative proteomics using uniform (15)N-labeling, MASCOT, and the trans-proteomic pipeline. *Proteomics* **2007**, 7, (19), 3462-9.

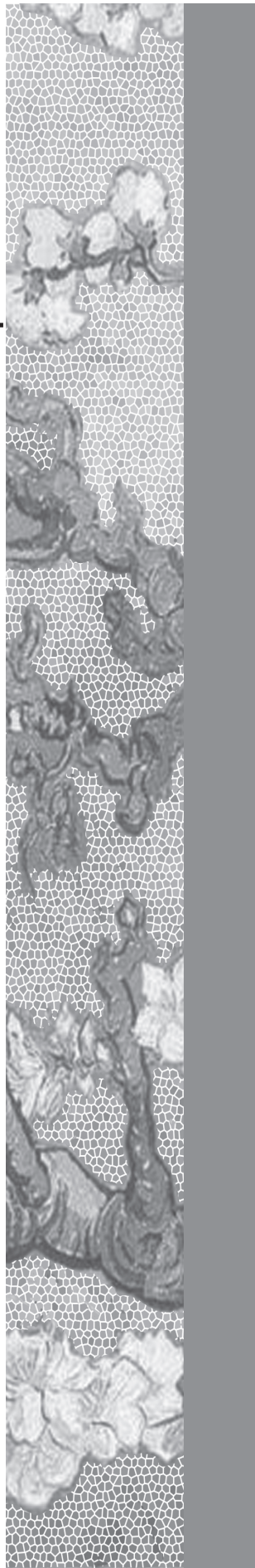
20. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y. F.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R., An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2002**, 2, (5), 513-523.

21. Brusniak, M. Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D., Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics* **2008**, 9, 542.

22. van Dijk, A.; Niessen, H. W.; Zandieh Doulabi, B.; Visser, F. C.; van Milligen, F. J., Differentiation of human adipose-derived stem cells towards cardiomyocytes is facilitated by laminin. *Cell Tissue Res* **2008**, 334, (3), 457-67.

23. Woods, A. J.; Kantidakis, T.; Sabe, H.; Critchley, D. R.; Norman, J. C., Interaction of paxillin with poly(A)-binding protein 1 and its role in focal adhesion turnover and cell migration. *Mol Cell Biol* **2005**, 25, (9), 3763-73.

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.[1] 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[2, 3] 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 multi-dimensional 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® [4] 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,[5] metabolites[6] 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

1. Kelleher, N. L., A cell-based approach to the human proteome project. *J Am Soc Mass Spectrom* **2012,** 23, (10), 1617-24.
2. Uhlen, M.; Bjorling, E.; Agaton, C.; Szigyarto, C. A.; Amini, B.; Andersen, E.; Andersson, A. C.; Angelidou, P.; Asplund, A.; Asplund, C.; Berglund, L.; Bergstrom, K.; Brumer, H.; Cerjan, D.; Ekstrom, M.; Elobeid, A.; Eriksson, C.; Fagerberg, L.; Falk, R.; Fall, J.; Forsberg, M.; Bjorklund, M. G.; Gumbel, K.; Halimi, A.; Hallin, I.; Hamsten, C.; Hansson, M.; Hedhammar, M.; Hercules, G.; Kampf, C.; Larsson, K.; Lindskog, M.; Lodewyckx, W.; Lund, J.; Lundeberg, J.; Magnusson, K.; Malm, E.; Nilsson, P.; Odling, J.; Oksvold, P.; Olsson, I.; Oster, E.; Ottosson, J.; Paavilainen, L.; Persson, A.; Rimini, R.; Rockberg, J.; Runeson, M.; Sivertsson, A.; Skollermo, A.; Steen, J.; Stenvall, M.; Sterky, F.; Stromberg, S.; Sundberg, M.; Tegel, H.; Tourle, S.; Wahlund, E.; Walden, A.; Wan, J.; Wernerus, H.; Westberg, J.; Wester, K.; Wrethagen, U.; Xu, L. L.; Hober, S.; Ponten, F., A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **2005,** 4, (12), 1920-32.
3. Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Bjorling, L.; Ponten, F., Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **2010,** 28, (12), 1248-50.
4. 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.
5. Castello, A.; Fischer, B.; Eichelbaum, K.; Horos, R.; Beckmann, B. M.; Strein, C.; Davey, N. E.; Humphreys, D. T.; Preiss, T.; Steinmetz, L. M.; Krijgsveld, J.; Hentze, M. W., Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **2012,** 149, (6), 1393-406.
6. Li, X.; Gianoulis, T. A.; Yip, K. Y.; Gerstein, M.; Snyder, M., Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell* **2010,** 143, (4), 639-50.

Addendum

SUMMARY

A large part of modern biology is dedicated to the functional annotation and interpretation of genetic information and its influence on the subject's phenotype. In an attempt to obtain comprehensive information at different levels a number of 'omics' fields have emerged. Genomic and transcriptomic analyses provide direct knowledge of the activities of the genes, but the genetic content is practically constant throughout an organism's body and its lifespan. The proteome on the other hand, is a translation of sequences encoded in the genome and the protein content is continuously changing, reflecting the current internal and environmental conditions of an organism. Proteomics describes the state of the system from the perspective of expression, structure, localization, interaction and function of the proteins. This makes proteomics research possibilities widely applicable. However, the absence of amplification techniques for proteins demands careful experiment planning with efficient and reproducible sample preparation procedures, especially so for a quantitative high-throughput label-free approach. The main challenges and difficult choices of the design of such experiments have been addressed in the **General Introduction**.

Proteomics is a complex field which requires a combination of various disciplines such as cellular biology and biochemistry for sample preparation, analytical chemistry for sample measurement and bioinformatics for the processing of data. Different chapters of this thesis illustrate various aspects of the proteomics pipeline and emphasize the importance and connection between them. Therefore the information in this dissertation can be divided into three major parts depicting these aspects of a proteomics experiment. First, sample preparation for proteomics has been addressed through two studies aimed at decreasing sample complexity and increasing proteome coverage. The second part is dedicated mostly to the technical step of the mass spectrometry measurement and the analysis of obtained spectra. The final part describes example- and proof-of-principle applications.

Mass spectrometry (MS) is a powerful tool for protein analysis. However, to improve proteome coverage of complex samples additional pre-fractionation and purification steps need to be performed prior to the measurement. **Chapter 1** presents a comparison between three pre-fractionation techniques performed at protein or peptide level; namely strong cation

exchange chromatography (SCX), isoelectric focusing (IEF) and SDS-polyacrylamide gel electrophoresis (SDS-PAGE). All three methods were applied to *Escherichia coli* and human plasma to assess the suitability of each method for a particular sample, as we assume that the choice of the method is likely to depend on the nature the sample. In addition, each method provides extra information on peptide or protein properties which can be both measured and calculated (protein molecular weight for SDS-PAGE, peptide pI for IEF and peptide charge at the system pH for SCX). These characteristics were used for the validation of peptide/protein identification enabling filtering for false discoveries. The whole data analysis from the raw data files to comparison and visualization of the results is interfaced in an automated manner within one pipeline using Taverna scientific workflow manager.

Chapter 2 introduces a different sample de-complexation approach for blood plasma proteomics. Plasma is a common clinical sample containing countless proteins, metabolites and lipids, and is of huge interest for biomarker discovery. However, the large protein abundance range found in blood plasma is still a hurdle for proteomics analysis. The method described in **Chapter 2** is a fast and robust depletion procedure that can be easily parallelized and applied in large clinical studies. It is based on a simple pH-adjusted organic solvent precipitation, which removes up to 90% of albumin from the sample and increases proteome coverage by at least 25% due to enrichment of lower-abundant proteins, including clinically relevant apolipoproteins. In comparison with existing commercial solutions, the technique is inexpensive, reproducible, high-throughput and suitable for quantitative label-free proteomics. This method can also be applied to other samples dominated by one or more proteins by adjusting the pH to match their respective pI values.

Protein quantitation is an important aspect of proteomics. **Chapter 3** describes a novel MS platform for high-throughput quantitative label-free proteomics using a Fourier transform ion cyclotron resonance (FTICR)-ion trap cluster. By combining high mass accuracy and resolving power of FTICR for quantitation with sensitive, fast and inexpensive MS/MS analysis through multiple ion traps for the peptide identification, similar performance and throughput as multiple hybrid ion trap-FTICR instruments can be achieved at a lower cost. The challenges in merging data from different instruments based on chromatographic alignment is also discussed.

Although the tool for an automated method for data analysis was already introduced in earlier chapters (**Chapters 1** and **2**), mass spectral searching of large amounts of data acquired during the high-throughput 'omics' experiments is still limited by the computational power. Typically such data processing involves multiple steps using various software and data formats. The peptide-spectrum assignment step is especially computationally demanding and increases the analysis time tremendously when performed on standard desktop computers. **Chapter 4** demonstrates the use of Taverna workflows for parallelized identification of tandem mass spectra through data decomposition algorithms applicable for publicly available database (X!Tandem) and spectral library (SpectraST) search tools. By outsourcing these processes, and thereby increasing the computational power, the analysis time of 5 combined human plasma datasets was reduced 30-fold for X!Tandem and 7-fold for SpectraST.

The acquired knowledge and developed methods for sample preparation, measurement and data analysis can be applied to a large variety of biological questions involving different types of samples. Following a well-studied *Escherichia coli* glucose-lactose diauxie experiment, in **Chapters 3** and **Chapter 5** the protein expression was matched with publically available gene expression data confirming *lac* operon proteins to be up-regulated. While **Chapter 3** is a proof-of-principle study, **Chapter 5** is focused on the implementation of the data processing pipeline for the FTICR-ion trap cluster and new ways of the data visualization. Quantitative information from ~1,000 proteins is converted to a color scale and mapped onto known metabolic pathways in Kyoto Encyclopedia of Genes and Genomes, illuminating parts of the pathway involved in the glucose metabolism. Similarly, this method can be applied system-wide to illustrate all the changes in the metabolism. Visualization of expression changes over time are here explored for 'temporal' proteomics.

Following the study of protein dynamics, as described in **Chapters 3 and 5**, the potential for studying protein expression in both time and space (cell/organelle) was investigated for a 'spatiotemporal' approach. **Chapter 6** describes the investigation of development of human stem cells into mature cardiomyocytes. Quantitative spatially and temporally resolved proteomics illuminate the mechanisms driving differentiation towards a specific end point. This knowledge can potentially be used to control the differentiation process for regenerative medicine and other purposes. In this initial study we separated time- and space-resolved proteomics. We extracted whole cell

lysates from four time points to follow the development in time and enriched for cytoplasmic, membrane, nuclear, chromatin-associated and cytoskeletal cellular components from one time point (fetal cardiomyocyte state) for the spatial aspect. In the process, >40,000 peptides from ~7,000 proteins were identified and were grouped according to their functions and cellular localization based on the gene ontology "slim" terms. As expected, proteins involved in cytoskeletal organization and motor activity were found to be upregulated towards later stages of cell differentiation. When adding an extra dimension of analysis (such as spatial components to a time course study), vast amounts of data are generated, creating a three-dimensional quantitative proteomics data cube. Unfortunately, the visualization of such data in a comprehendible manner is challenging. Protein abundances were translated into color, and mapped onto a simple representation of the cell which enables us to restrict the number of perspectives necessary for the visualization of time and space dimensions of information. In general, **Chapter 6** demonstrates the feasibility of a spatiotemporal quantitative label-free proteomics.

Moving towards proteomics which is simultaneously high-throughput, quantitative, spatiotemporal and label-free has become possible by incremental development of instrumental platforms and new ways for analysis and visualization of 'big data'. Each chapter of the current thesis highlights separate aspects and emphasizes their interdependence.

NEDERLANDSE SAMENVATTING

Een groot gedeelte van de moderne biologie is houdt zich bezig met de functionele annotatie en interpretatie van genetische informatie en de invloed daarvan op het uiteindelijke fenotype. Bij het verwerven van informatie op verschillende niveaus (metabolieten, eiwitten en genen) zijn verscheidene 'omics' velden ontstaan. Genomics en transcriptomics geven kennis over de activiteit van genen, maar de samenstelling van het genoom blijft tijdens het leven van een organisme praktisch gelijk en verschaft daarom weinig directe informatie over fenotypen. Het proteoom daarentegen is een vertaling van de genetische code en is als resultaat daarvan in continue verandering afhankelijk van zowel de interne als de omgevingstoestand van een organisme. Proteomics beschrijft daarom de staat van een systeem of organisme op het gebied van expressie, structuur, locatie, interactie en functie van de eiwitsamenstelling. Helaas is door de afwezigheid van amplificatietechnieken (zoals de PCR voor DNA) zorgvuldige planning en een efficiënte en reproduceerbare monstervoorbewerking van zeer groot belang, zeker in het geval van een label-vrije kwantitatieve strategie. De grootste uitdagingen bij het ontwerpen van een dergelijk experiment zijn beschreven in de **Inleiding**.

Proteomics is een complex onderzoeksveld dat een aantal disciplines zoals celbiologie en biochemie, analytische chemie en bio-informatica combineert voor respectievelijk, de voorbewerking, de analyse en de dataverwerking. De verschillende hoofdstukken van dit proefschrift belichten de verschillende aspecten van proteomics en benadrukken het belang van een goede combinatie van deze onderdelen. De informatie in dit proefschrift kan daarbij worden onderverdeeld in drie delen die gecorreleerd zijn aan de drie fasen in een proteomics experiment. Ten eerste is de monstervoorbewerking bekeken, in het bijzonder het verminderen van de complexiteit van een monster, met als doel het aantal geïdentificeerde eiwitten te vergroten. Het tweede deel beschrijft de techniek rond massaspectrometrie en de analyse van de verkregen data. Als laatste is een aantal applicaties van de volledig geïntegreerde aanpak beschreven.

Massaspectrometrie is een uiterst krachtig hulpmiddel voor de analyse van eiwitten. Om het aantal geïdentificeerde eiwitten in een monster zo groot mogelijk te maken hebben zelfs op massaspectrometrie gebaseerde technieken een fractionering en verdere opzuivering van het monster nodig. **Hoofdstuk 1** beschrijft de vergelijking van drie verschillende

fractioneringstechnieken voor de analyse van *Escherichia coli* en humaan plasma. Op grond van deze vergelijking werd vastgesteld wat de meest geschikte methode was voor de twee monsters, aangezien de geschiktheid van een analysemethode sterk afhankelijk is van het type en complexiteit van een monster. Verder verschaft iedere methode informatie over eigenschappen van het eiwit of peptide die ook berekend kunnen worden op grond van de aminozuursamenstelling. Deze informatie is gebruikt om de eventuele eiwit- of peptide identificaties te valideren op grond van de molecuulmassa, het iso-elektrisch punt en de peptidelading voor respectievelijk SDS-PAGE, iso-electric focusing en SCX chromatografie. Alle data analyse en verwerking is uitgevoerd in één geautomatiseerde methode binnen de 'Taverna scientific workflow manager'.

Hoofdstuk 2 introduceert een alternatieve strategie voor "de-complexering" van het monster voor eiwitanalyse van bloedplasma. Bloedplasma is een veel gebruikt klinisch monster en van groot belang voor de ontdekking van ziekte-indicatoren. Helaas vormt de hoge concentratie van maar een klein aantal eiwitten (in het bijzonder albumine) een groot probleem voor in-depth analyse. De methode die beschreven wordt in **Hoofdstuk 2** is snel en robuust en kan eenvoudig geparallelliseerd worden voor grote klinische studies. Een precipitatie-stap met een organisch oplosmiddel dat pH-gecorrigeerd is verwijderde 90% van alle albumine. Hierdoor konden minimaal 25% meer eiwitten worden gemeten, waaronder apolipoproteïnen. In vergelijking met de commercieel verkrijgbare methoden is deze methode goedkoop, reproduceerbaar en geschikt voor label-vrije kwantitatieve analyse. Een vergelijkbare methode zou toegepast kunnen worden op andere monsters die gedomineerd worden door één specifiek eiwit door de pH aan te passen aan de pI van dit eiwit.

De kwantificering van eiwitten vormt een belangrijk onderdeel van eiwitanalyses. **Hoofdstuk 3** beschrijft een massaspectrometrie platform voor grootschalige, label-vrije kwantitatieve eiwitanalyse op basis van een Fourier transform ion cyclotron resonance (FTICR)-ion trap cluster. Door het combineren van de hoge massa-accuraatheid en resolutie van de FTICR voor de kwantificering en de snelle en gevoelige MS/MS-analyses van meerdere ion traps voor peptide identificatie, kunnen tegen lagere kosten prestaties gehaald worden vergelijkbaar met hybride ion trap-FTICR instrumenten. De uitdaging bij het samenvoegen van de data gegenereerd met beide type instrumenten wordt ook besproken.

Ondanks het feit dat geautomatiseerde methoden voor data verwerking en -analyse al geïntroduceerd zijn in **Hoofdstuk 1 en 2** is de beperkende factor bij dit proces voornamelijk rekenkracht. Over het algemeen bestaat deze data verwerking en de daarop volgende analyse uit meerdere stappen, uitgevoerd in verschillende programma's en data formats. De annotatie van het fragmentatiespectrum van een peptide is hierbij over het algemeen het meest tijdrovende onderdeel, hetgeen resulteert in lange rekentijden wanneer dit wordt uitgevoerd op een "normale" PC. In **Hoofdstuk 4** wordt beschreven hoe de Taverna workflow kan worden gebruikt voor het parallel laten verlopen van meerdere spectra identificaties door middel van openbare databank (X!Tandem) of spectrum bibliotheek (SpectraST) zoekmachines. Door dit zoekproces uit te voeren op een rekencluster kan de tijd die nodig is voor de analyse van 5 humane plasma datasets met een factor 7 worden gereduceerd voor SpectraST en zelfs met een factor 30 voor X!Tandem

De verkregen kennis over monstervoorbewerking, analyse en dataverwerking kan worden toegepast op een grote verscheidenheid aan biologische vragen voor sterk verschillende monsters. **Hoofdstuk 3 en 5** beschrijven de vergelijking van de eiwitexpressie in een *Escherichia coli* glucose-arm experiment met public domain genexpressie data. Hierdoor kon bevestigd worden dat er een verhoging van de concentratie *lac*-operon eiwitten plaatsvond. Waar **Hoofdstuk 3** gericht is op de toepassing van de methode is **Hoofdstuk 5** gericht op de toepassing van de data verwerking die nodig is voor het FTICR-ion trap cluster en op nieuwe manieren om de resultaten te visualiseren. De kwantitatieve informatie van circa 1000 eiwitten is op basis van een kleurschaal gekoppeld aan de metabolisme routes gevonden in de 'Kyoto Encyclopedia of Gene and Genomes' waarbij vooral de routes die gekoppeld zijn aan het glucose metabolisme eruit sprongen. Op vergelijkbare wijze kan deze methode ook toepast worden om veranderingen in een systeem door de tijd (van het experiment) te volgen. Visualisatie van resultaten van een dusdanig experiment is in **Hoofdstuk 5** ook besproken.

In vervolg op de studies beschreven in de voorgaande hoofdstukken is in **Hoofdstuk 6** gekeken naar de mogelijkheid om de dynamiek van eiwitten te bestuderen in zowel tijd als ruimte (organellen). Deze studie beschrijft de ontwikkeling van menselijke stamcellen tot volwassen cardiomyocyten. Celcultures van vier ontwikkelingsfasen zijn gelyseerd en voor één tijdspunt (foetaal cardiomyocyten) zijn de verschillende organellen fracties geïsoleerd: cytoplasma, membraam, nucleus, chromatine-gerelateerd en

cytoskelet. De kwantitatieve eiwitdata in zowel tijd als ruimte laten de verschillende mechanismen zien die actief zijn tijdens de differentiatie. Deze kennis over de differentiatie kan in de toekomst mogelijk gebruikt worden om dit proces te kunnen controleren, bijvoorbeeld voor regeneratieve geneeskunde. Binnen de gehele studie zijn er >40.000 peptiden en ongeveer 7.000 eiwitten geïdentificeerd en deze zijn vervolgens gegroepeerd op basis van hun functie en locatie. Zoals verwacht, was te zien dat eiwitten die betrokken zijn bij de organisatie van het cytoskelet en bij de motorfunctie in de latere fasen van ontwikkeling in concentratie toenemen. Wanneer er, zoals in deze studie, een ruimtelijk aspect wordt toegevoegd aan een tijdstudie, ontstaan er immense hoeveelheden data die op een driedimensionale manier met elkaar zijn gecorreleerd. Helaas is het op een inzichtelijke manier weergeven representeren van deze data erg moeilijk. Er is hiervoor gekozen de relatieve eiwitconcentratie te vertalen naar kleuren en vervolgens te koppelen aan delen van een versimpelde representatie van een cel. Op deze manier is het mogelijk om de driedimensionale data zoals verkregen wordt bij een studie met een tijd- en ruimte-aspect weer te geven in een tweedimensionaal format. Over het geheel toont **Hoofdstuk 6** de potentie van het gebruik van label-vrije kwantitatieve eiwitanalyse voor dergelijke tijd- en ruimte-studies.

Door stapsgewijze ontwikkeling van technieken en technologie, en het weergeven van informatie uit grote datasets laat dit proefschrift zien dat het mogelijk is grootschalige, label-vrije, kwantitatieve, tijd- en ruimte-studies uit te voeren op eiwitniveau. Elk hoofdstuk in dit proefschrift beschrijft één of meerdere delen van deze ontwikkelingen met de uiteindelijke toepassing op een echt biomedisch vraagstuk.

ACKNOWLEDGEMENTS

Only a few months ago I could not even imagine myself writing the acknowledgements to my own thesis but here I am. It was a lucky chance that I got this PhD position and I am grateful for it ever since. This was an amazing time of personal and professional growth and incredible experience of living in the Netherlands.

First of all, I would like to thank my promotor, André, for giving me the opportunity to work in such friendly environment and to learn so much about proteomics and mass spectrometry in general.

My co-promotor, Magnus, I learned so much from you. You were always willing to help in any situation and to give valuable advice but at the same time I am glad that you gave me freedom. Tack för ditt stöd, förståelse och din vägledning.

I would also like to thank my co-authors and colleagues in the lab for their assistance with the experiments, fruitful discussions and useful comments.

I have been blessed with the wonderful colleagues that were very helpful in the lab and great companions outside work: Aswin, Bart, Bjorn, Dick-Paul, Frank, Hulda, Irina, Katja, Ollie, Paul and Sibel. Dank jullie wel, het was altijd echt gezellig met jullie, zowel op het werk als daarbuiten. I wish I could write something about each of you but I would need a separate book for it. I just hope you know how much I appreciate all of you: Alexandra, Alex, Axel, Benjamin, Caroline, Crina, Dana, Emanuel, Emrys, Gerhild, Guinevere, Hans, Kristell, Liam, Linda, Manfred, Marco, Martin, Maurice, Ralf, René, Ricardo, Rico, Rob, Robert, Sarantos, Sha, Suzanne, Tune, Yassene and Yuri. Anton, it has been fun sharing office with you and thank you for all the help with my text in both English and Dutch.

Irina, we have met when I experienced a very difficult time in my life and thank you so much that you always took my side and never lost faith in me. Thank you for listening to me and supporting me all these years. I am happy that you are my friend, and thank you for being my paranymph.

My dear Tiziana and Simone, we started almost at the same time and you were always there for me to listen, to support and even to give me shelter. I am very happy that I can call you my friends and I am glad that at the defense I have you Simone by my side.

I am grateful for all the amazing people that I have met during these years. Always positive and friendly Spanish girls, who taught me a lot of fun things: Alegria, Helena, Judit and Mireia. Grazias guapas, por vuestra positividad y por ser mis amigas. My dear friend, Marieth, you are such an amazing, understanding and positive person. Thank you for being there for me all these years.

I would like to thank many more other wonderful people that made Netherlands my second home: Irinka, Emanuel, Violeta, Bart, Agata, Pawel, Eleni, Eleonora, Dimitris, Ivo and Olga. Bart, you did an amazing job making the cover for this thesis, thank you so much.

My best friends Anastasia and Anastasia, we live far away from each other and cannot spend much time together but I know that any time of the day I would call, you always would be there for me to listen or even travel across the world if necessary. You are amazing, I love you and miss you a lot.

Мои дорогие мамулечка и папулечка, спасибо, что вы в меня всегда верили, всегда поддерживали и всегда мотивировали на большее. Я вас очень сильно люблю и очень по вам скучаю.

CURRICULUM VITAE

Ekaterina Mostovenko was born on 11th of September, 1987, in Tver, Russia. In 2003 she finished high school in Tver and continued her education at Faculty of Bioengineering and Bioinformatics at Lomonosov Moscow State University. In summer 2006, as part of her studies, together with a group of nine students, she spent a one-month internship in Leiden University Medical Center (LUMC), the Netherlands. She worked on comparison of several public SNP databases and identification of chromosomal regions out of Hardy-Weinberg equilibrium in the Department of Medical Statistics and Bioinformatics under supervision of Jeanine J. Houwing-Duistermaat, Irina Nischenko, Hae-Won Uh and Hans C. van Houwelingen. In 2008 she graduated from Moscow State University. The same year she started her PhD research at the Department of Parasitology/ Biomolecular Mass Spectrometry Unit, LUMC, under supervision of Prof. Dr. A.M. Deelder and Dr. M. Palmblad. Her research resulted in a number of scientific papers combined in the current thesis entitled "Towards high throughput and spatiotemporal proteomics: analytical workflows and quantitative label-free mass spectrometry". From May 2013 she continues her research as Post-Doc in the laboratory of Carol Nilsson at the Pharmacology and Toxicology Department, University of Texas Medical Branch.

LIST OF PUBLICATIONS

Mostovenko E., Hassan C., Rattke J., Deelder A.M., van Veelen P., Palmblad M. (2013) Comparison of Peptide and Protein Fractionation Methods in Proteomics. *EuPA Open Proteomics*, submitted

Heemskerk A.A.M., **Mostovenko E.**, Dalebout H., Wulff T., de Fijter J.W., Tollenaar R.A.E.M., Mayboroda O.A., Palmblad M., Deelder A.M. (2013) Complex Sample Analysis by Capillary Electrophoresis and Mass Spectrometry in Bottom-up Proteomics. *Electrophoresis*, submitted

Mostovenko E., Deelder A.M., Palmblad M. (2012) Protein Fractionation for Quantitative Plasma Proteomics by Semi-Selective Precipitation. *J. Proteomics Bioinform.*, **5**, 217-221

Mohammed Y., **Mostovenko E.**, Henneman A.A., Marissen R.J., Deelder A.M., Palmblad M. (2012) Cloud Parallel Processing of Tandem Mass Spectrometry-based Proteomics Data. *J. Proteome Res.*, **11** (10), 5101-5108

Victor B., Gabriel S., Kanobana K., **Mostovenko E.**, Polman K., Dorny P., Deelder A.M. Palmblad M. (2012) Partially Sequenced Organisms, Decoy Searches and False Discovery Rates. *J. Proteome Res.*, **11** (3), 1991-1995

Mostovenko E., Deelder A.M., Palmblad M. (2011) Protein Expression Dynamics during *Escherichia coli* Glucose-Lactose Diauxie. *BMC Microbiology*, **11** (1), 126

Palmblad M., van der Burgt Y.E.M., **Mostovenko E.**, Dalebout H., Deelder A.M. (2009) A Novel Mass Spectrometry Cluster for High-Throughput Quantitative Proteomics. *J. Am. Soc. Mass Spectrom.*, **21** (6), 1002-1011