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