Plant Biology Through Quantitative Proteomics

Joakim Bygdell
Faculty of Forest Sciences
Department of Forest Genetics and Plant Physiology
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2013

Acta Universitatis agriculturae Sueciae
2013:60

ISSN 1652-6880
ISBN (print version) 978-91-576-7858-4
ISBN (electronic version) 978-91-576-7859-1

Print: Print & Media Umeå 2013

Plant Biology Through Quantitative Proteomics

Abstract

Over the last decade the field of mass spectrometry based proteomics has advanced from qualitative, analyses leading to publications revolving around lists of identified proteins and peptides, to addressing more biologically relevant issues requiring measurement of the abundance of identified proteins and hence quantitive mass spectrometry.

The work described in this thesis addresses problems with quantitive proteomics in plant sciences, particularly complications caused by the complexity of plant proteomes (generated by genomic duplications), which makes mass spectrometry-based based proteomic analyses more difficult than in mammalian species. In order to understand complex biological processes it is vital to analyse the participating molecules with as little bias as possible. Strategies for minimizing and maximizing the acquired information in proteomic investigations of plants are presented in the appended papers and discussed in the thesis.

Keywords: quantitative proteomics, mass spectrometry, peptide identification, spectra matching, protein database.

Author's address: Joakim Bygdell, SLU, Department of Forest Genetics and Plant Physiology, 901 87 Umeå, Sweden
E-mail: Joakim.Bygdell@slu.se

Just because you can't see it doesn't mean it isn't there.
Claire London

Contents

List of Publications

This thesis is based on the work described in the following papers, which are referred to by the corresponding Roman numerals in the text. For convenience, the studies described in them are sometimes referred to as Studies I-IV.

I Nilsson, R., Bernfur, K., Gustavsson, N., **Bygdell, J**., Wingsle, G., & Larsson, C. (2010). Proteomics of plasma membranes from poplar trees reveals tissue distribution of transporters, receptors, and proteins in cell wall formation. *Molecular & cellular proteomics : MCP*, *9*(2), 368–87.

II **Bygdell, J**., Nilsson, R., Srivastava M. K., Srivastava, V., Quarnström, J., Sundberg, B. & Winglse, G. Protein isoforms involved in tension wood formation monitored at high tissue resolution. *(Manuscript)*.

III Srivastava, V., Obudulu, O., **Bygdell, J**., Löfstedt, T., Rydén, P., Nilsson, R., Ahnlund, M, Johansson, A., Jonsson, P., Freyhult, E., Qvarnström, J., Karlsson, J., Melzer, M., Moritz, T., Trygg, J., Hvidsten, T. & Wingsle, G. OnPLS integration of transcriptomic, proteomic and metabolomic data reveals multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. *(Manuscript submitted to Biomedcentral)*.

IV Businge, E., **Bygdell, J**., Wingsle, G., Moritz, T., & Egertsdotter, U. (2013). The effect of carbohydrates and osmoticum on storage reserve accumulation and germination of Norway spruce somatic embryos. *Physiologia plantarum*. doi:10.1111/ppl.12039

The contribution of Joakim Bygdell to the papers included in this thesis was as follows:

I All bioinformatics and sequence analysis, providing the basis for classifying detected proteins. Writing the parts related to bioinformatic analysis in the paper.

II Development of the method used for quantitative analysis of mass spectrometry data. Analysis of all the acquired data and writing the most of the manuscript.

III All quantitative data analysis, bioinformatic evaluation of the proteomic data and writing the proteomics-related sections of the manuscript.

IV All proteomics-related work including sample preparation, mass spectrometry analysis, subsequent bioinformatic evaluation and writing related parts of the paper.

Abbreviations

AMRT	Accurate Mass Retention Time Pair
DDA	Data-Dependent-Acquisition
HPLC	High Performance Liquid Chromatography
iTRAC	Isobaric Tags for Relative and Absolute Quantification
LC-MS/MS	Liquid-Chromatography Coupled Tandem Mass Spectrometry
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MS1	Survey scan
MS2	Fragment ion scan
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
TMT	Tandem Mass Tag
UPLC	Ultra Performance Liquid Chromatography
XIC	Extracted Ion Chromatogram

1 Introduction

The principle challenge of cell biology is to reveal the mechanisms and inner workings of cells. In this quest, cells are perceived as systems in which the dynamic interplay of a large number of components determines the output of many parallel biological processes. To characterize these processes and to reveal their underlying principles, one needs to evaluate the dynamic composition and localization of the molecular components. As all cellular processes involve proteins, their characterization has therefore drawn most interest over the years (Walther & Mann 2010).

The proteome is extremely multifaceted owing to splicing and post-translational modifications (PTMs). PTMs are more than just "decorations" they can affect the activity state, localization and turnover of a protein as well as its interaction with other proteins (Mann & Jensen 2003). This diversity is further amplified by the interconnectivity of proteins into complexes and signalling networks that are highly divergent in space and time (Altelaar et al. 2013).

The emergence of proteomics, the large-scale analysis of proteins (Anderson & Anderson 1998), has been inspired by the realization that the final product of a gene is inherently more complex and closer to function than the gene itself (Graves & Haystead 2002). With correlation between gene expression levels and protein abundance reported to be poor (Maier et al. 2009; Greenbaum et al. 2003), quantitative proteomics is necessary in order to determine protein abundances.

Over the last decade mass spectrometry-based proteomics has advanced from qualitative analyses, leading to publications revolving around lists of identified proteins and peptides, to addressing more biologically relevant issues

requiring measurement of the abundance of identified proteins and hence quantitive mass spectrometry.

However, despite the advances in mass spectrometry in terms of mass accuracy and resolution, as well as peptide separation by ultra performance liquid chromatography systems (UPLC) the numbers of proteins that can be identified and quantified comprise only a fraction of organisms' proteomes (Bantscheff et al. 2007). One of the problems is that proteolytic peptides have widely varying physiochemical properties, leading to large variations in their signal responses, even if they originate from the same protein. Consequently to maximize accuracy, quantification must be performed on a peptide-to-peptide basis, comparing the same ion species across all samples. Even then the quantifications are only relative, for absolute quantification a synthetic isotopically labelled standard of known concentration is needed for all peptides of interest. One method have been described that attempts to perform absolute quantification on protein level relative to a spiked protein standard (Silva et al. 2006).

All methods for quantitative proteomics have been designed and validated on model systems with few paralogous proteins, (Silva et al. 2006) used mammalian proteins spiked into a E. coli background. While this provides a convenient starting point one must also consider the complexity of a real biological sample.

From a proteomic perspective plants are one of the most difficult organisms to analyse due to the genome duplications that have occurred in their evolutionary history and the retention and subsequent modification of large portions of the duplicated genes. All these paralogs in the plant genome can make it difficult to assign a peptide to a specific protein and as a consequence quantification may have to be done on groups of proteins. *Arabidopsis* (*Arabidopsis thaliana*) for example has almost 1000 protein families with 5 or more members (Lin et al. 2008). Another problem associated with plants is that the sequenced variant are not always the variants commonly used in research. For example the sequenced variant of poplar (*Populus spp.*) is the North American Black Cottonwood (*P. trichocarpa*) (Tuskan et al. 2006) whereas most research in Sweden is carried out on wild populations of aspen (*P. tremula*) or hybrids (*P. tremula x P. tremuloides*). Furthermore, most transgenic work on poplar is performed on the hybrid variants, which contain genetic material from both parent species, thus increasing both DNA and protein level sequence variations.

1.1 Protein Quantification

Protein quantification by mass spectrometry has advanced a long way during the last ten years with the development of various methods for various types of instruments. However, all of the methods can be classified as labelled or label-free, and subdivided as relative or absolute.

1.1.1 Stable Isotope Labelling

Quantification using any of the different stable-isotope methods except isobaric-mass-tagging (see below) is based on extracted ion chromatograms (XIC) from the survey scan (MS1) data channel for each of the labelled (heavy) and unlabelled (light) peptide variants. Unless deuterium is used to create the heavy form of the peptide both the heavy and light forms co-elute from the column during the chromatographic gradient.

Stable isotope labelling can be used to generate absolute quantification. If so the heavy isotope-labelled peptides are normally synthetically produced and accurately measured.

Metabolic labelling

First described using ^{15}N-enriched media for complete labelling of bacterial proteins (Oda 1999), this approach has since been extended, notably to stable isotope-labelled essential amino acids in mammalian cell cultures (SILAC) (Ong et al. 2002). Where ^{13}C labelled arginine and lysine are incorporated into newly synthesized proteins, resulting in a 6 Da mass shift between the heavy and light peptides.

SILAC has been shown to work with both yeast and bacteria that are auxotrophic for the labelled amino acids (de Godoy et al. 2008; Soufi et al. 2010). Auxotrophic *Chlamydomonas* (*Chlamydomonas reinhardtii*) mutants are the only organism from the plant kingdom that have successfully been labelled using SILAC (Naumann et al. 2007; Terashima et al. 2010). For plants a modified variant of the SILAC method can be used (Schütz et al. 2011) which allows for quantification even though plants normally are autotrophic.

Since plants are autotrophic they can easily be labelled metabolically by using ^{15}N-enriched salts. A complication is that the isotope clusters of ^{15}N-labelled peptides are wider and directly related to the length and sequence of the peptide, so the mass difference between the labelled and unlabelled form of the peptide are not constant. Another effect of ^{15}N-labelling is that more

isobaric peptides are generated, further complicating both the sequence matching and quantification (Nelson et al. 2007; Gouw et al. 2008).

Chemical labelling

Proteins and peptides can also be labelled by chemical or enzymatic reactions that target specific groups, primarily sulfhydryl and amines. A commonly used sulfhydryl–reacting label is the isotope-coded-affinity-tag (ICAT) (Gygi et al. 1999). Since it reacts exclusively with cysteine it can be used to study the redox status of proteins. However, as a tool for more global quantification it is of little value due to the scarcity of cysteine residues in protein sequences.

The label can also be added to peptides during protein digestion in ^{18}O-enriched water (Winter et al. 2009). In contrast to ^{15}N-labelling the mass shift created between the heavy and light peptides is constant, but since the mass difference generated by ^{18}O is only 4 Da for a tryptically cleaved peptide it is only effective for relatively short peptides (Stewart et al. 2001; Yao et al. 2004).

Isobaric Mass Tags

Isobaric mass tagging (Thompson et al. 2003) differs from the methods described above in that the label added to each peptide has the same mass and chemical properties. As the peptides co-elute from the LC column only single peaks will appear in the MS1 scan and differences between the samples will only appear upon fragmentation. Isobaric tags for relative and absolute quantitation (iTRAQ) (Chong et al. 2006) and tandem mass tags (TMT) (Thompson et al. 2003) are commercially available isobaric mass tags that allow up to eight samples to be analysed simultaneously.

Spiked Standard Peptides

The combination of multiple reaction monitoring by a triple quadrupole mass spectrometer and isotope–labelled synthetic peptides has been used for absolute quantitation of proteins (AQUA) (Gerber et al. 2003). For accurate quantification at least two peptides from each protein should be used and if the peptides are synthesized independently mixing them in equimolar concentrations can be difficult.

One way to overcome the mixing problem is to generate synthetic proteins consisting of concatenated peptides (QconCat) from the proteins targeted for analysis (Simpson & Beynon 2012). These peptides will be released in a 1:1 ratio after protein digestion and if added at an early point in the sample

preparation process any losses related to sample handling and digestion will be reflected equally in the QconCat peptides.

1.1.2 Label Free

Due to the large differences in ionization efficiency between different peptides only the same species can be accurately compared between different samples. In a normal MS/MS experiment the MS1 scans are interrupted by fragment ion scans (MS2), hence the coverage of the precursor ions in the MS1 data channel is irregular. The number of data points over the chromatographic peak for any precursor ion is determined by the duty cycle of the instrument.

As large number of MS2 scans are required to identify of as many peptides as possible in a sample and continuous sampling over the chromatographic peak is necessary for accurate quantification, unless the analysis is divided into separate parts for quantification and identification one will come at the cost of the other. This problem can be overcome by using LC-MS^E, in which low and elevated energy acquisition modes are applied, thus allowing for continuous acquisition of precursor and fragment data over the entire chromatographic peaks (Geromanos et al. 2009). The accuracy of matching fragment spectra to precursor ions can be further increased by adding drift time data from Ion Mobility equipped mass spectrometers.

The latest generations of mass spectrometers are capable of providing extremely low mass errors (< 10 ppm) under routine running conditions. Ensuring high reproducibility of the LC retention time values over an extended time frame, e.g. several days of continuous running during large-scale projects, remains the biggest hurdle for label-free quantitation. However, adding an orthogonal protein to the sample can aid the retention time alignment.

Spectral counting

The fact that abundant proteins are more readily detected during a data-dependent-aquisition (DDA) has been used as the basis for protein quantification (Liu et al. 2004). The method provides results biased towards abundant proteins that generate many proteotypic peptides, but this can be at least partially corrected by dividing the number of identified peptides by the theoretically observable number of peptide to calculate protein abundance index (PAI) (Rappsilber et al. 2002). Exponentially modified indices (emPAI) have also been applied (Ishihama et al. 2005).

1.2 Plants as Model Organisms

A major difference between plants and other organisms such as yeasts and mammals is that a whole genome duplication event has occurred at least once in plants' evolutionary history, following which most of the duplicate genes were retained. Data from genetic studies suggest that about 80% of all genes in *Arabidopsis* reside in duplicated regions of the genome (Simillion et al. 2002).

The genome duplications have inevitably increased the complexity of their genomes as the retained paralogous genes have evolved different spatial and temporal expression patterns (Tuskan et al. 2006), there is also evidence of different functions for the paralogs (Pin et al. 2010). The DNA-level sequence similarities between these paralogous genes are commonly at least 80%, and for proteomic analysis this complexity is even greater as protein sequences are less variable than the corresponding DNA sequence. Pairs of paralogous proteins may differ only by a single amino acid insertion, substitution or deletion, making them extremely difficult to analyse as they will only have one unique tryptic peptide. In *Arabidopsis* the number of proteins with only one unique peptide are 14-fold higher than in humans and of the same size as the entire *E.coli* proteome.

2 Objective

The objective of the work described in this thesis was to investigate the problems associated with mass spectrometric quantification of proteins in samples derived from plants and develop solutions, focusing on label-free methods. Study I examined spectral counting methods and complications for this type of quantification arising from the genome duplications in plants. Studies II & III address quantification using peptide precursor intensities and the requirements for reliably matching quantified precursors to identified peptides. Study IV examined the Top3 method, a variant based on precursor intensities that allows semi-absolute quantification relative to spiked proteins of known concentration. Although robust experimental design and sample preparation are essential for successful proteomic analyses these are not focal concerns of this thesis and hence are not considered in detail.

3 Methods

3.1 Experimental design

When performing proteomics analysis of greenhouse–grown plants it is better to pool material from several individuals and run multiple technical replicates than to treat them as biological replicates. This is because the differences between individuals from the same line can be as large as the differences between individuals from different transgenic lines, depending on the growth conditions and placement in the greenhouse(Pinto et al. 2011).

In order to calculate proper statistics a minimum of three replicate injections are necessary for each sample. Although it might be strategic to do 4-5 replicate injections of each sample depending on system stability, as a failure of the LC or MS system during data acquisition of one replicate will have lesser impact on the overall results. Filtering out all peptides found in less than half of the replicate injections for each sample will decrease the number peptides and proteins that cannot be properly quantified.

3.2 Protein digestion

“Bottom-up” sequencing of proteins by MS/MS refers to their sequencing via the analysis of peptides generated by proteolytic digestion. One of the most widely used proteases for this purpose is trypsin, because it conveniently generates peptides with a basic residue (lysine or arginine) at the C-terminus. As arginine and lysine each constitute about 5% of the amino acids in many proteomes the typical peptide generated will be between 1000 and 3000 Da a suitable range for MS/MS analysis (Brownridge & Beynon 2011). While there are other proteases they are mainly used in targeted cases where trypsin will

generate peptides that are either to short or to long to be effectively analysed by LC-MS.

In the work described in this thesis my colleagues and I (here after we) used both gel based and other digestion methods, since the former enables the removal of detergents that are incompatible with LC-MS analysis but they are more laborious and can introduce contaminants in the form of dust (keratin) into the sample.

Importance of complete digestion for quantitative proteomics

While mis-cleaved peptides can sometimes enhance the quality of an identification, for quantitative proteomics they can cause errors if two mis-cleavage products represent parallel but different dead-end proteolytic processes as trypsin will not cleave after a lysine or arginine that is N-terminally located (Brownridge & Beynon 2011). To minimize the amount of mis-cleaved peptide products a double digest strategy can be applied using two proteases that yield overlapping products. This may also have the benefit of allowing the first round of digestion to be performed in a more denaturing environment, for example digestion using lysine-C in 8M urea followed by trypsin after dilution in ammonium bicarbonate buffer.

3.3 Nano-flow liquid-chromatography

In order to analyse the complex sample mixtures generated in these studies we employed online reversed phase separation by nano-flow liquid chromatography. The peptides where separated on a C18 column with 75 µm inner diameter using a mobile phase consisting of water, acetonitrile and 0.1 % formic acid with a flow-rate of 300 nL min^{-1}. The low flow-rates compared to those of standard HPLC, gives nano-flow systems higher sensitivity as smaller droplets are ejected from the electrospray emitter, enhancing the desorption of ions into the gas phase (Abian et al. 1999).

Peptides are separated in the column largely due to differences in the strength of interactions between their amino acids and the hydrophobic stationary phase of the column (Krokhin et al. 2004). The small diameter of the column is sub-optimal for sample loading as the UPLC system used cannot load volumes smaller than 0.1 µL. Thus, to inject volumes in the 1-10 µL range a pre-column was used to trap and wash the samples using µL min^{-1} flow-rates with the trapping valve open, then the trapping valve was closed to redirect the flow to the analytical column for sample separation at flow-rates of 200-400 nL min^{-1}.

3.4 Electrospray ionization

A major breakthrough in protein and peptide analysis came in the late 1980's with the invention of the electrospray ionization (ESI) source by John B. Fenn and co-workers (Fenn 2002). This enabled mass spectroscopic analyses of liquid samples. In ESI an aqueous solution is passed through a needle with a small diameter, and a very large voltage differential between the needle and the entrance of the mass spectrometer causes formation of a Taylor cone from which small highly charged droplets are generated. The size of the droplets is reduced, and thus surface charge density is increased, through continuous evaporation of the solvent. Finally, ions are ejected from the surface of the droplets into the gaseous phase (Bruins 1998). ESI of peptides in positive mode results in protonation of their N-terminal amines and the basic side-chains of arginine, histidine and lysine, leading to the formation of multiply charged species of tryptically digested peptides.

3.5 Time-of-flight mass spectrometry

The working principles of a time-of-flight (TOF) mass spectrometer are elegant and simple. Ions are accelerated by a fixed electric field (U) to a velocity (v) that is inversely proportional to their mass to charge (m/z) ratio (Eq. 1). The time it takes for the ions to travel through the field-free region of the flight tube can be accurately measured and thus used to calculate their m/z (Kinter & Sherman 2000).

$$v = \sqrt{\frac{2 * U * z}{m}} \qquad \text{(Eq. 1)}$$

Since velocity (v) is distance (D) over time (t) equation 1 can be rewritten.

$$\frac{D}{t} = \sqrt{\frac{2 * U * z}{m}} \qquad \text{(Eq. 2)}$$

Solving equation 2 for mass gives us equation 3.

$$m = \frac{2 * U * z}{D^2} * t^2 \qquad \text{(Eq. 3)}$$

Here, v = velocity in m/s, m = mass in kg, U = accelerating voltage, z = charge, D = distance, t = time.

The derivative of equation 3 yields equation 4:

$$\Delta m = \frac{2 * U * z}{D^2} * 2t * \Delta t \qquad \text{(Eq. 4)}$$

The resolution of a TOF instrument is defined as the measured mass divided by the width of the corresponding peak at 50% height, full width at half maximum (FWHM), which is the same as the relationship between equations 3 and 4.

$$\frac{m}{\Delta m} = \frac{\frac{2 * U * z}{D^2} * t}{\frac{2 * U * z}{D^2} * 2t * \Delta t} = \frac{t}{2 * \Delta t} \qquad \text{(Eq. 5)}$$

From Eq. 5 we can see that the resolution is directly related to the difference in flight times of two ions with a given mass difference and thus the length of the flight path in the field-free region. Modern TOF instruments use reflectrons to increase the flight path and thus the resolution while maintaining the compactness of the instrument.
From Eq. 5 we can also see that in order to resolve two peaks at FWHM a resolving power of 2*m/Δm is required.

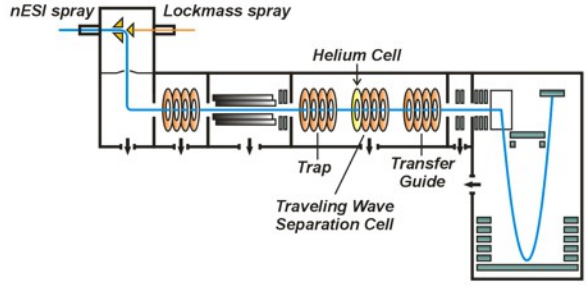

Figure 1. Schematic diagram of an orthogonal-accelerating Q-TOF mass spectrometer with a traveling wave ion mobility separation cell and a reflectron-equipped TOF chamber (Image from Waters).

In Studies I – III we used a Waters Q-TOF Ultima mass spectrometer operated at a resolution of 10,000 with a mass measurement error of less than 100 ppm. A Waters Synapt G2 HDMS capable of a resolution of 40,000 and mass

measurement error of a few ppm was used in Study IV. Both instruments are of hybrid quadrupole type (Figure 1) with a collision cell located after a quadrupole mass filter, allowing use of the instruments for tandem-in-time mass analysis. The major difference between the two instruments is that the Synapt G2 HDMS has an ion mobility (IM) drift cell, and thus adds a third dimension to the MS data by recording the IM drift time of each peptide. Peptides can be fragmented before or after IM separation, if fragmentation is performed after IM separation peptides can be sequenced and identified even if they cannot be resolved by the TOF mass analyser directly (Figure 2).

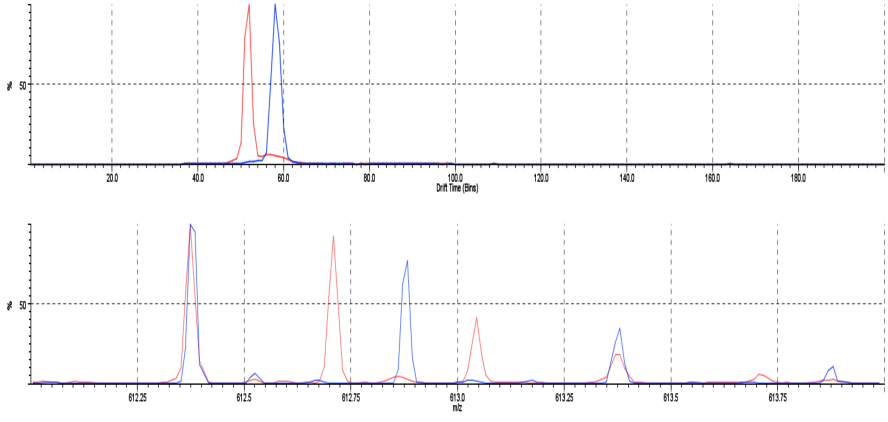

Figure 2. Illustration showing the effect of ion mobility. The two peptides are clearly separated by the difference in their drift time (top). The mono-isotopic peaks of the two peptides cannot be resolved by the mass analyser (bottom).

3.6 Peptide fragmentation

Although a number of fragmentation techniques are available today most peptide sequencing is done using low-energy collision-induced-dissociation (CID). In this procedure, peptides are allowed to collide with a chemically inert gas, usually Ar or N_2. As each peptide undergoes repeated collisions with the gas its internal energy increases until the stored energy reaches the point where a chemical bond breaks. As the peptide fragments primarily along the backbone the amino acid sequence of the peptide can be deduced from the resulting fragment ions (Figure 3). Fragments are only detected if they carry at least one charge. If this charge is located at the N-terminal side of the fragment

the ion is classified as an *a* or *b* ion. Fragments with the charge located at the C-terminal side are classified as an *x* or *y* ion (Roepstorff & Fohlman 1984).

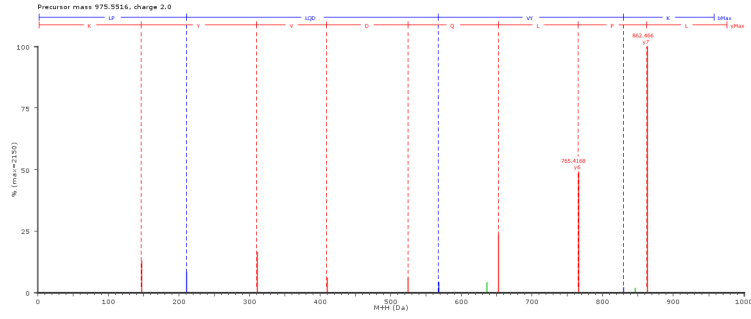

Figure 3. Showing fragment spectra from the tryptic peptide LPLQDVYK and how the peptide sequence can be derived from the fragment ion series. Red, C-terminal y-ions, blue, N-terminal b-ions, green, neutral loss ions.

3.7 Data processing

Continuous mass spectral data must be processed to generate spectra files that can be used for database searching. The general method applies to both MS1 and MS2 data and involves smoothing, background reduction, mass off-set calibration, peak integration and charge state deconvolution. For MS1 data, in addition the area under the chromatographic peak are calculated for each component. The retention-time of the peak apex for each monoisotopic ion must also be determined, combined with the calibrated measured mass and tabulated as an accurate mass retention time pair (AMRT). Unless the isotope cluster is deconvoluted the calculated charge state of the ion is added to the AMRT.

3.8 Database searching

In the post genomic era peptides are identified by comparing their experimentally derived fragment spectra to theoretical spectra for all peptides in a sequence database that are within the set tolerances of the measured mass of the intact peptide. This means that any peptide sequence that are not included in the database cannot be identified, which must be considered if the sequence database and the sample material are from different species or different ecotypes of the same species. The genome duplication events that

have occurred in plants adds problems on another level as the paralogous proteins may be almost identical, and if a group of proteins cannot be unambiguously identified with the protease used they should be merged into a single database entry. For proper calculations of false positive identification rates the databases should include all proteins that may be detected, including contaminants introduced when handling samples, e g. keratin and the protease used for digestion.

3.8.1 Mass measurement error

Mass measurement error plays a major role in database searching. With low measurement error, the tolerances for a database search can be very narrow and if the measurement error is normally distributed three standard deviations will capture > 99% of all peptides (Zubarev & Mann 2007).

3.9 Matching identifications to precursor peaks.

During a DDA experiment a peptide can be selected for fragmentation at anytime during its chromatographic elution. Thus, when separate injections are used for identification and quantification the entire width of the chromatographic peak must be considered when matching an identified peptide to a peak in the MS1 data. The matching is performed by creating mass-corrected retention-time pairs (AMRT) from the MS1 data, the identifications are then matched to the corresponding AMRT.

If the same chromatographic gradient is used in the MS and MS/MS runs the matching is quite straight-forward as the variation in retention time of modern UPLC systems is in the order of a few seconds while chromatographic peak widths are 15-30 seconds. Hence, the retention-time (RT) window used when matching need not be much larger than the average chromatographic peak width. The mass window used for matching should be the same as for the database search, but can be based on the measured masses and charge-states, or the MS1 data can be deconvoluted with respect to charge and compared to the nominal masses reported by the search engine.

During Study II no software was available that could match identifications to precursors with data from our instrument. Therefore I developed a program, described in Paper II, to generate the AMRT pairs and assign the identified peptides to the corresponding MS1 peak. This method where later also used in Study III.

4 Results and Discussion

None of the methods discussed in this thesis have been developed using material with the same complexity level as the material we have applied them to analyse. Figure 4 shows the clear difference in the amount of sequence unique peptides for some of the more common model organisms.

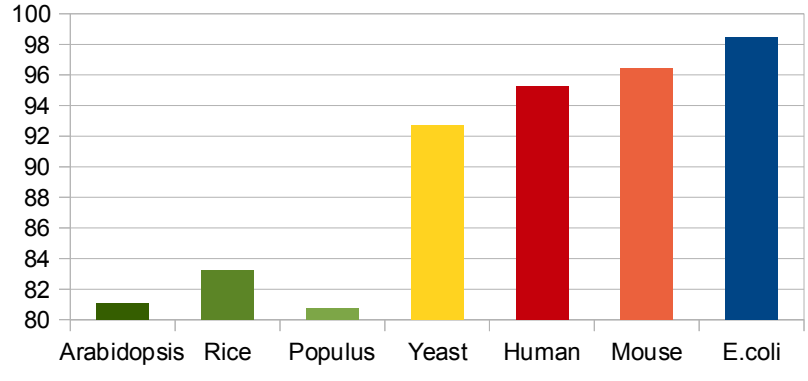

Figure 4. Comparison of percentages of sequence unique peptides in different model organisms.

Closer analysis of the Human and Arabidopsis proteomes reveals that the pools of tryptic peptides are of similar size (~600 000). However, arabidopsis has ca. 4-fold more non-unique peptides (~114 000 versus ~27 000).

4.1 Spectral counting and protein paralogs

Spectral counting is a method developed in early attempts to determine the abundance of proteins based on the number of identified peptides. Theoretically it is straightforward; a protein of high abundance will be selected for MS/MS more often than a protein of low abundance. For a purified protein analysed in different concentrations this will be true until all proteotypic peptides have been found at which point no further increase in abundance vill be detectable. For a complex sample, on the other hand, in which most proteins are identified by a few peptides an increase in the abundance of a specific protein might only mean that the quality of the spectra improves or that another peptide are identified some times at the cost of loosing an identified peptide belonging to another protein.

One problem we had to address was linked to the way that the database search engine we used, MASCOT, arranged the identified proteins. In MASCOT the protein scores are derived from the sum of the non-redundant list of scores for peptides matching to the protein. When when there is a group of proteins that share a pool of peptides and thus will have the same score, the group will then be represented by the protein with the lowest accession number. Another issue is that if the identified peptides are unique to this group of proteins they will be marked as unique even though they match more than one protein, implications of this will be discussed later on. If on the other hand the proteins share a subset of peptides the protein with the most matching peptides or the protein with a sequence unique peptide will be in the result list. For samples with many paralogous proteins this can have serious consequences as the difference of one identified peptide can change what proteins are in the result list. As such the same set and sub set result lists will show a higher degree of consistency than the primary result list.

In order to reduce the complexity of our samples we utilized SDS-PAGE to separate the proteins, as described in Paper I. As each section from the resulting gels contained proteins of a limited size range, identified proteins of the “wrong” size could be excluded from the result list. We also grouped the identified proteins according to their annotations, allowing us to estimate any differences in abundance based on the number of identified peptides for each group of proteins in the gel slices.

At best, spectral counting will give hints of differences in protein abundance in samples of similar complexity. It is not reliable or sensitive enough to be used at all for protein quantification.

4.2 Quantification by precursor intensity

When quantifying peptides by precursor intensity the number of isotopic peaks to use must be carefully chosen. This might seem trivial but it is important to keep the number consistent, i.e. use only the monoisotopic peaks or a set number of isotopic peaks. Using the monoisotopic peak will cause problems for quantifying peptides larger than ca. 2 kDa as they will not be the most abundant. On the other hand using multiple isotopic peaks will be problematic for peptides of low abundance as the third and fourth isotopic peak might not be detectable. Unless the same isotope peaks are used the differences between samples might be exaggerated, especially for large peptides, as the higher order isotope peaks constitute a larger fraction of the signal response.

An important aspect to keep in mind is that the amount of sample injected for quantitative MS1 runs needs to be low enough to avoid risks of overloading either the column or the detector. Thus, in order to maximize the number of identifications with the rather short gradients used in Studies II & III (25-30 min), we analysed the samples in fractions of several m/z intervals while maintaining the same injection volume and chromatographic gradient as in the quantitative MS1 runs. Keeping the same chromatographic gradient for all injections greatly facilitated the matching of identifications. Out of 1091 non-redundant peptides identified in Study III, 458 were considered unique and of sufficient intensity to use for quantification of 271 proteins using the three first isotopic peaks.

Although modern HPLC systems have high run-to-run stability, consecutive runs are not identical. Thus, a window defined in terms of retention-time is required when performing the matching. The easiest way to calculate such a window is to intermix the DDA injections for identification and the MS1 injections. This allows the RT drift to be calculated by examining any peptides identified in several of the DDA runs. The mass tolerances used for matching the peptide m/z value should be the same as the one used for the database search, calculated from the detected m/z for the peptide in question. Each identified peptide is then defined as an AMRT window, in which the peptide should also yield an isotopic cluster corresponding to its charge state.

Due to the genome duplications in plants the identified peptides must be sorted into pools depending on whether they are sequence-unique or may originate from more than one protein. Unique peptides will provide direct indications of differences in abundance between the samples, but shared peptides must be compared to see if the overlapping proteins have any impact on the calculated ratios when compared to the unique peptide. For proteins

identified only by shared peptides the ratios between samples should be reported for the whole group.

4.3 Top 3 Quantification

The Top3 method allows semi-absolute quantification, relative to a protein of known concentration that is added to the samples either before or after digestion. The basic quantification is done using the XIC for the peptide precursors and thus relies on the ability to match precursors to identified peptides. As the method uses the average signal response for the three most intense peptides identified for each protein, three sequence unique peptides must be identified for each protein. This is not problematic in analyses of mammalian organisms, in which most proteins have low sequence similarity to each other. However, in plants large proportion of the proteins cannot be quantified with this method as they yield less than three unique peptides upon digestion with trypsin. These paralogous proteins must be grouped so that the consensus sequence in the group yields at least three unique peptides. Since thee peptides are required for the quantification the same rule should apply for protein identification to prevent the issue where a protein is identified with one or two peptides yet cannot be properly quantified.

This quantification method was used in Study IV, together with a data independent mode of acquisition (MS^E) (Silva et al. 2005). When MS^E precursor and fragment data are acquired alternately, identification, quantification and matching can be done simultaneously. By combining MS^E with ion mobility separation we were able to increase the quality of the results significantly as co-eluting peptides could be separated in the mobility cell, thus non-chimeric spectra could be generated as fragmentation occurred after mobility separation (Figure 5). While ion mobility increases data quality it also increases the risk of saturating the detector as more ions reach at the detector in a given time span. As can be seen in Figure 2 the width of the peaks in drift time is about 6 bins, which translates to 30 ms of acquisition time. This is a 33-fold decrease compared to if the data where acquired over the entire scan time of 1 s and thus and increase in the number of ions per time unit.

As the Spruce genome had been not sequenced (Nystedt et al. 2013) during Study IV we used the available sequences belonging to the *Picea* genus compiled in the non-redundant National Center for Biotechnology Information database. This minimised the number of paralogous sequences in the databank, allowing us to use the Top3 method.

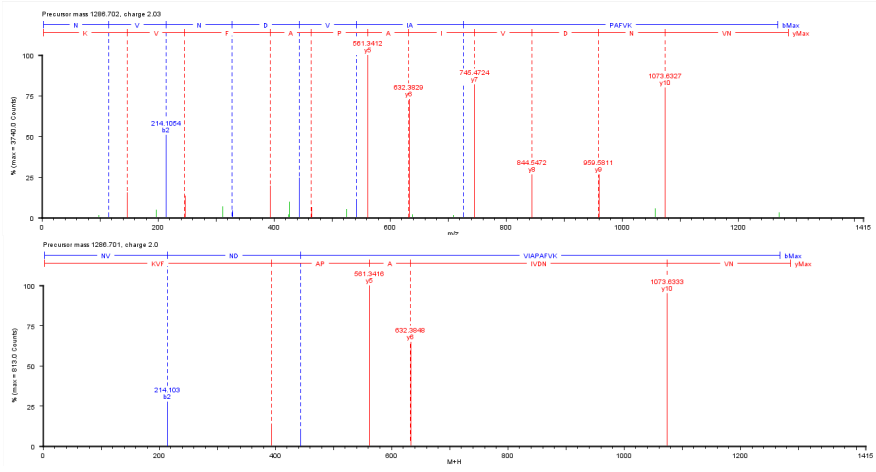

Figure 5. Illustration of the increase in data quality for a peptide identified by MS^E with (top) and without (bottom) ion mobility.

The Top3 method has both substantial weaknesses and strengths as a quantification method. The requirement for three unique peptides for each protein is more difficult to meet in proteomic analyses of plants than in analyses of other organisms, due to their large number of paralogs. On the other hand, quantification is baed on XIC of the identified peptides, which enables relative quantification of all identified peptides.

5 Conclusions and Future Perspectives

The work described in this thesis has provided insights into problems associated with quantitative proteomic analyses of plants, particularly complications arising from the complexity of plant proteomes, generated by the genome duplications for large scale quantification. The results highlight, *inter alia*, the importance of identifying sequence-unique peptides when analysing plant samples.

The method used for quantification is less important, whether it is based on reporter ions or the chromatographic profiles of precursor ions. However, the Top3 method has the particularly attractive feature of allowing the relative quantification of all identified peptides and the capacity to provide semi-absolute values for all proteins that can be identified with 3 unique peptides.

The main problem that remain to be resolved lie with the way sequence databanks currently are arranged and the search engines deal with overlapping protein sequences that complicate protein identification. Sequences are normally compared on text level, i.e. the sequence of two or more proteins are compared in terms of the similarity of the letters in the text sequence. For robust proteomic analyses sequences should ideally be compared with consideration of the mass spectrometer's limitations, notably leucine and isoleucine should not be considered to be two different amino acids as they are isobaric. In addition sequences should be compared after *in silico* digestion and removal of all peptides that are outside the mass spectrometers typical detection window, normally 600 to 4000 Da. Any proteins that cannot be uniquely identified should be grouped and annotated accordingly.

The addition of transcriptional meta data to sequence databases would allow the creation of databases that are specific to a particular tissue or developmental time period, similar to how taxonomy data are utilised in databases that contain sequences from multiple organisms.

For small-scale targeted projects a relatively simple procedure as RT-PCR might be enough to identify the subset of paralogs that are present in the focal samples. This would allow the number of paralogous sequences in the database to be reduced and thus increase the number of unique peptides for each protein, facilitating quantification of the paralogs present in the samples. Another way to generate a reduced but robust databank is to pool aliquots of all samples in each group and run a multi-dimensional experiment either through fractionation of the digested peptides or the intact proteins. Using only the proteins that can be unambiguously identified as a subset for quantification would reduce the influence of the non-identified paralogs.

Even after applying these procedures unidentified proteins may still affect quantifications. Thus, the possibility that ratios of shared peptides differ from those of sequence-unique peptides should be tested for all proteins that share identified peptides. If a difference is observed between these ratios the proteins corresponding to shared peptides should be quantified as a separate pool.

Until search engines and bioinformatic tools are developed that can properly deal with the complications caused by protein paralogs quantitative plant proteomics will rely heavily on manual procedures.

References

Abian, J., Oosterkamp, A.J. & Gelpí, E., 1999. Comparison of conventional, narrow-bore and capillary liquid chromatography/mass spectrometry for electrospray ionization mass spectrometry: practical considerations. *Journal of Mass Spectrometry*, 34(4), pp.244–254.

Altelaar, A.F.M., Munoz, J. & Heck, A.J.R., 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics*, 14(1), pp.35–48.

Anderson, N.L. & Anderson, N.G., 1998. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11), pp.1853–61.

Bantscheff, M. et al., 2007. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4), pp.1017–31.

Brownridge, P. & Beynon, R.J., 2011. The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods (San Diego, Calif.)*, 54(4), pp.351–60.

Bruins, A.P., 1998. Mechanistic aspects of electrospray ionization. *Journal of Chromatography A*, 794(1-2), pp.345–357.

Chong, P.K. et al., 2006. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *Journal of proteome research*, 5(5), pp.1232–40.

Fenn, J.B., 2002. Electrospray ionization mass spectrometry: How it all began. *Journal of biomolecular techniques : JBT*, 13(3), pp.101–18.

Gerber, S.A. et al., 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12), pp.6940–5.

Geromanos, S.J. et al., 2009. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics*, 9(6), pp.1683–95.

De Godoy, L.M.F. et al., 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217), pp.1251–4.

Gouw, J.W. et al., 2008. Optimizing identification and quantitation of 15N-labeled proteins in comparative proteomics. *Analytical chemistry*, 80(20), pp.7796–803.

Graves, P.R. & Haystead, T.A.J., 2002. Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews : MMBR*, 66(1), pp.39–63; table of contents.

Greenbaum, D. et al., 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology*, 4(9), p.117.

Gygi, S.P. et al., 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology*, 17(10), pp.994–9.

Ishihama, Y. et al., 2005. 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 : MCP*, 4(9), pp.1265–72.

Kinter, M. & Sherman, N.E., 2000. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Krokhin, O. V et al., 2004. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Molecular & cellular proteomics : MCP*, 3(9), pp.908–19.

Lin, H. et al., 2008. Characterization of paralogous protein families in rice. *BMC plant biology*, 8(1), p.18.

Liu, H., Sadygov, R.G. & Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, 76(14), pp.4193–4201.

Maier, T., Güell, M. & Serrano, L., 2009. Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24), pp.3966–73.

Mann, M. & Jensen, O.N., 2003. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3), pp.255–61.

Naumann, B. et al., 2007. Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathways under iron deficiency in Chlamydomonas reinhardtii. *Proteomics*, 7(21), pp.3964–79.

Nelson, C.J. et al., 2007. Implications of 15N-metabolic labeling for automated peptide identification in Arabidopsis thaliana. *Proteomics*, 7(8), pp.1279–92.

Nystedt, B. et al., 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451), pp.579–84.

Oda, Y., 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences*, 96(12), pp.6591–6596.

Ong, S.-E. et al., 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular cellular proteomics MCP*, 1(5), pp.376–386.

Pin, P.A. et al., 2010. An antagonistic pair of FT homologs mediates the control of flowering time in sugar beet. *Science*, 330(6009), pp.1397–1400.

Pinto, R.C. et al., 2011. Design of experiments on 135 cloned poplar trees to map environmental influence in greenhouse. *Analytica Chimica Acta*, 685(2), pp.127–131.

Rappsilber, J. et al., 2002. Large-scale proteomic analysis of the human spliceosome. *Genome research*, 12(8), pp.1231–45.

Roepstorff, P. & Fohlman, J., 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry*, 11(11), p.601.

Schütz, W. et al., 2011. Extending SILAC to proteomics of plant cell lines. *The Plant cell*, 23(5), pp.1701–5.

Silva, J.C. et al., 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Molecular & cellular proteomics : MCP*, 5(1), pp.144–56.

Silva, J.C. et al., 2005. Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical Chemistry*, 77(7), pp.2187–2200.

Simillion, C. et al., 2002. The hidden duplication past of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21), pp.13627–32.

Simpson, D.M. & Beynon, R.J., 2012. QconCATs: design and expression of concatenated protein standards for multiplexed protein quantification. *Analytical and bioanalytical chemistry*, 404(4), pp.977–89.

Soufi, B. et al., 2010. Stable isotope labeling by amino acids in cell culture (SILAC) applied to quantitative proteomics of Bacillus subtilis. *Journal of proteome research*, 9(7), pp.3638–46.

Stewart, I.I., Thomson, T. & Figeys, D., 2001. 18O labeling: a tool for proteomics. *Rapid communications in mass spectrometry : RCM*, 15(24), pp.2456–65.

Terashima, M. et al., 2010. Characterizing the anaerobic response of Chlamydomonas reinhardtii by quantitative proteomics. *Molecular & cellular proteomics : MCP*, 9(7), pp.1514–32.

Thompson, A. et al., 2003. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, 75(8), pp.1895–1904.

Tuskan, G.A., Difazio, S. & Jansson, S., 2006. The genome of black cottonwood, Populus trichocarpa (Torr Gray). *Science*, 313, pp.1596–1604.

Walther, T.C. & Mann, M., 2010. Mass spectrometry-based proteomics in cell biology. *The Journal of cell biology*, 190(4), pp.491–500.

Winter, D. et al., 2009. Simultaneous Identification and Quantification of Proteins by Differential 16O/18O Labeling and UPLC-MS/MS Applied to Mouse Cerebellar Phosphoproteome Following Irradiation. *Anticancer Res*, 29(12), pp.4949–4958.

Yao, X. et al., 2004. Proteolytic 18 O Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus. *Analytical Chemistry*, 76(9), pp.2675–2675.

Zubarev, R. & Mann, M., 2007. On the proper use of mass accuracy in proteomics. *Molecular & cellular proteomics : MCP*, 6(3), pp.377–81.

Acknowledgements

First and foremost I wish to thank my supervisor **Gunnar** for giving me this opportunity, I feel that we both learnt a lot during these years. I have really appreciated the freedom to work largely independently while knowing that you were always available for any discussion.

To the people at the **Couch** and the **Round Table** thanks for all the interesting and varied discussions over coffee. I'd like to thank **Simon** for some greatly needed breaks in the thought processes, tackling problems in other areas are sometimes refreshing and sometimes annoying, especially when experiments or instruments don't work as intended. To all the other people I had the opportunity to work with over the years at **UPSC** a big thank you: it has been fun and interesting.

I would like to give thanks to: **Anton**, **Christoffer J.**, **Mats**, **Peter** and **Christoffer H.** for fun and tasty times in and out of the kitchen; all the **KCs** and other people at **Villan** for everything during those years; **Anton**, **Johan** and **Anke** for some subsurface recreation time; my **family** for all the support over the years; and finally **Elena**, for being mine.